nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-019-0095-1

Increasing the specificity of CRISPR systems with engineered RNA secondary structures

D. Dewran Kocak[1,2], Eric A. Josephs[3,6], Vidit Bhandarkar[1,2], Shaunak S. Adkar[1,2], Jennifer B. Kwon[2,4] and Charles A. Gersbach[1,2,5*]

CRISPR (clustered regularly interspaced short palindromic repeat) systems have been broadly adopted for basic science, biotechnology, and gene and cell therapy. In some cases, these bacterial nucleases have demonstrated off-target activity. This creates a potential hazard for therapeutic applications and could confound results in biological research. Therefore, improving the precision of these nucleases is of broad interest. Here we show that engineering a hairpin secondary structure onto the spacer region of single guide RNAs (hp-sgRNAs) can increase specificity by several orders of magnitude when combined with various CRISPR effectors. We first demonstrate that designed hp-sgRNAs can tune the activity of a transactivator based on Cas9 from *Streptococcus pyogenes* (SpCas9). We then show that hp-sgRNAs increase the specificity of gene editing using five different Cas9 or Cas12a variants. Our results demonstrate that RNA secondary structure is a fundamental parameter that can tune the activity of diverse CRISPR systems.

CRISPR–Cas systems are adaptive immune systems in bacteria and archaea, and have proven to be robust genome editing platforms[1]. Efforts to repurpose CRISPR–Cas systems for genome editing have largely focused on class 2 CRISPR systems because of their simplicity. While class 1 systems use multi-protein complexes to target nucleic acids, class 2 systems use a single Cas protein, termed the Cas effector, which can be easily reconstituted and harnessed for a variety of applications[2].

The arms race between viruses and prokaryotes has driven immense genetic diversity of Cas effectors. Each Cas effector has unique properties (for example, nucleic acid preference, protospacer-adjacent motif (PAM) requirements, size of the Cas effector) that endow it with advantages and disadvantages for particular applications. The identification and characterization of class 2 CRISPR systems is thus an active area of research, with the overarching goal of finding Cas effectors with novel or improved properties[3–5]. Since the initial characterization of SpCas9, the number of Cas effectors active in mammalian cells has expanded to include compact Cas9 effectors from the type II CRISPR systems, Cas12a (previously Cpf1) effectors with (A+T)-rich PAMs from type V systems and RNA-targeting Cas13 variants[6–12].

Although these nucleases are versatile tools for gene editing outside of their native environments, they also have off-target effects, leading to unintended DNA breaks at sites with imperfect complementarity to the spacer sequence[13–15]. Thus, improving the specificity of these nucleases is a critical goal, especially for gene therapy applications[16]. Methods to increase the specificity of class 2 CRISPR systems through rational design have largely focused on SpCas9 and have adopted two general strategies. The first strategy is to create an AND gate that requires coordinate binding of two Cas9 molecules, imposing a stricter requirement for nuclease activity[17–20]. The second strategy is to reduce the energetics of DNA interrogation by the Cas9–single guide RNA (sgRNA) complex, which results in an overall increase in specificity[21–25]. The second strategy is particularly attractive because, unlike the first strategy, it does not increase the number of components of the gene editing system. This simplifies gene delivery, which is often a critical barrier. While previous efforts with either strategy were successful, they suffer from one or more of a variety of limitations, including incompatibility with viral packaging constraints, a greater number of components of the system and the requirement for extensive protein engineering. Recent studies that employ directed evolution rather than rational design have yielded many new variants with improved properties[26–28]. However, it remains to be seen which of these many approaches will have general applicability across CRISPR systems. Thus, there is a need for a simple method for increasing specificity of diverse CRISPR systems.

Employing rational design and adopting the second strategy, we hypothesized that engineering the sgRNA might serve as a means to regulate diverse CRISPR systems. Specifically, we engineered RNA secondary structure onto the spacer by extending a designed hairpin on the 5′ end of the sgRNA (hp-sgRNA). The resulting hairpin structure could then serve as a steric and energetic barrier to R-loop formation. We hypothesized that by adjusting the strength of the secondary structure, R-loop formation could proceed to completion at the on-target site but could be impeded at off-target sites, which have reduced energetics due to RNA–DNA mispairing. Because R-loop formation is the critical process governing the conformational change of SpCas9 to an active nuclease[29,30], this would block off-target nuclease activity and result in an increase in specificity. Since CRISPR endonucleases accommodate a nucleic acid duplex within their binding channel, we hypothesized that the RNA–RNA duplexes of hp-sgRNAs could also be accommodated without interfering with formation of the sgRNA–protein complex. Moreover, hp-sgRNAs are simple to design and produce: RNA hairpins generally follow Watson–Crick base-pairing guidelines, and sgRNA production methods are rapid and inexpensive.

[1]Department of Biomedical Engineering, Duke University, Durham, NC, USA. [2]Center for Genomic and Computational Biology, Duke University, Durham, NC, USA. [3]Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, USA. [4]University Program in Genetics and Genomics, Duke University Medical Center, Durham, NC, USA. [5]Department of Surgery, Duke University Medical Center, Durham, NC, USA. [6]Present address: Department of Nanoscience, University of North Carolina at Greensboro, Greensboro, NC, USA. *e-mail: charles.gersbach@duke.edu

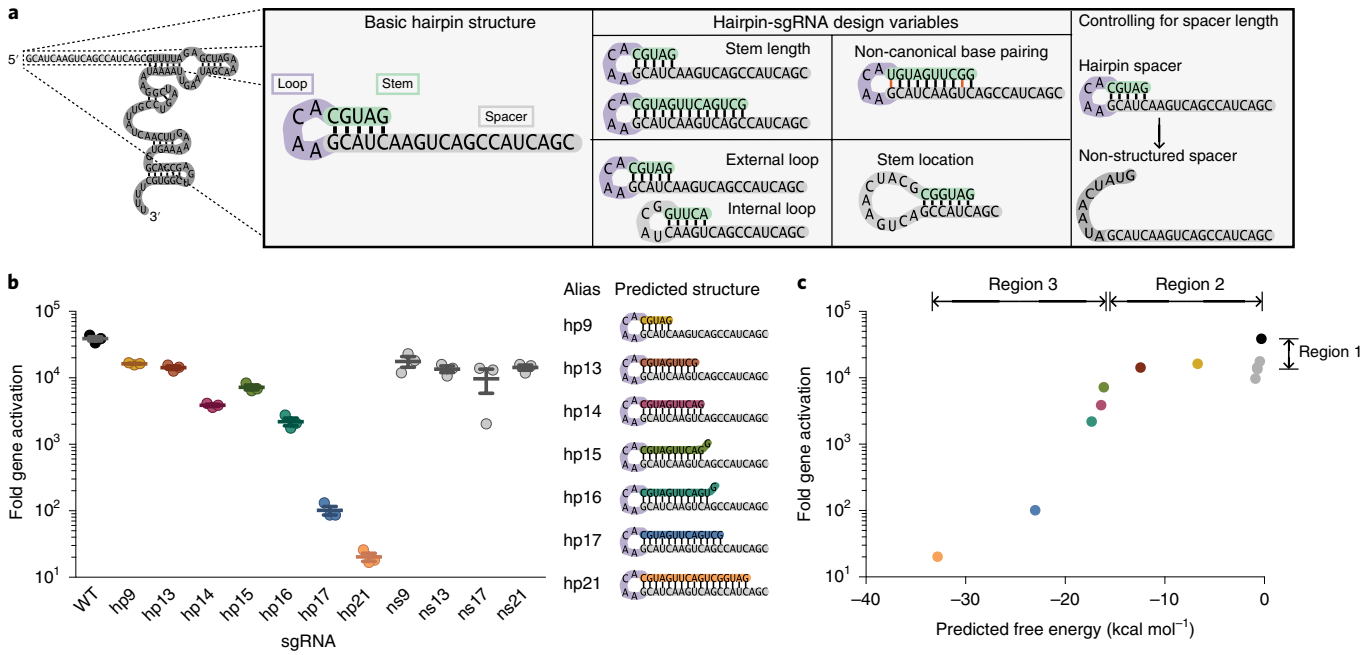

Fig. 1 | Engineered RNA secondary structures tune the activity of dCas9-P300. a, Structure of the WT-sgRNA for SpCas9 and design parameters of hp-sgRNAs. **b**, Gene activation of *IL1RN* using hp-sgRNAs with varying stem lengths, measured by qRT–PCR. Hairpin sgRNAs are abbreviated as 'hp', non-structured controls are abbreviated as 'ns' and numbers indicate the number of nucleotides added 5′ of the spacer. Data are shown as fold increase relative to the control sample, which was transfected with dCas9-P300 only. Error bars represent s.e.m. for $n = 3$. All hp-sgRNA variants show significant activation over control, $P < 0.005$ using a two-sided *t*-test after a global one-way analysis of variance (ANOVA). **c**, Replotting the mean of each group in **b** as a function of the predicted folding energy of each hp-sgRNA's engineered secondary structure. Trends in the data are annotated for clarity (for example, 'Region 1'). The sequences of all sgRNAs used are listed in Supplementary Table 1.

Results

Design considerations for hp-sgRNAs. RNA can fold into many different complex structures. For our initial engineered structures we adopted the RNA hairpin, a fundamental structural unit in many RNA molecules[31]. RNA hairpins are composed of two components, stems and loops, which we create by extending the PAM-distal end of the spacer to generate hp-sgRNAs (Fig. 1a). All designs were informed through the use of in silico structure determination, and only spacer sequences were used for these predictions (that is, structural sequences in the tracrRNA or crRNAs were excluded).

We expected thermodynamic stability of the secondary structure to be an influential characteristic of hp-sgRNAs. However, there are many variables one can use to create different structures with similar stability (Fig. 1a). The stem can be placed along any area of the 20-nucleotide spacer, which may have variable effects on R-loop formation kinetics. Stem lengths, the major determinant of hairpin stability, can also be varied. To modulate stability but not necessarily overall hp-sgRNA structure, non-canonical rG-rU base pairs can be substituted for potential rG-rC/rA-rU sites in the stems. Many RNA hairpins found in nature utilize 5′-ANYA-3′ or 5′-UNCG-3′ tetraloops, which have favorable base-stacking behavior[32]. We utilize these tetraloops for our initial structures, but one can also use part of the spacer itself for the hairpin loop. In this study, all of these variables were used to generate hp-sgRNAs. Furthermore, to control for any effects of sgRNA length, we also designed non-structured sgRNAs (ns-sgRNAs), which have extensions to the spacer but whose extensions are not predicted to form any secondary structures.

hp-sgRNAs regulate a SpCas9-based transcriptional activator. We first tested the effect of predicted hp-sgRNAs structures on Cas9 binding to DNA. Critically, we wanted to analyze this interaction in human cells, where reports have shown that extensions to the 5′ end of the sgRNA can be processed back to lengths of the native spacer[19,33]. We thus decided to utilize nuclease-inactive dCas9-based transcriptional activators[34,35], where endogenous gene activation can serve as a sensitive measure of dCas9 binding to target DNA.

For our initial hp-sgRNA designs, we used a tetraloop that is external to the 20-nucleotide spacer and placed the hairpin stems on the PAM-distal end of the spacer using canonical Watson–Crick base pairing. We used a spacer that targets the endogenous promoter of *IL1RN*, a gene we have previously activated with high efficiency[34,35]. Transfecting sgRNA variants and a dCas9-P300 transactivator into human cells, we observed that hp-sgRNAs can tune gene activation at the target locus (Fig. 1b), suggesting modulation of dCas9 binding.

We observed a generally regular relationship between length of the hp-sgRNA spacer extension and impact on dCas9 binding (Fig. 1b). The only irregularity was observed with hp15, which has an unpaired 5′ guanine, necessitated by the U6 promoter. Replotting the activity of each hp-sgRNA variant as a function of thermodynamic stability of their predicted structures, we observed a monotonic decrease of gene activation over four orders of magnitude (Fig. 1c). These data provide evidence that the predicted RNA structures form in human cells and demonstrate that the in silico predicted free energy of the structures is an accurate predictor of its regulatory effect on dCas9 binding to genomic DNA (gDNA) target sites.

Notably, use of ns-sgRNAs did not decrease transactivation to the same degree as seen with hp-sgRNAs, indicating that hairpin formation, and not simply sgRNA extension, was responsible for modulating dCas9 binding. However, on average, ns-sgRNAs caused a ~2.8-fold reduction in gene activation when compared with the unmodified guide (wild-type sgRNA (WT-sgRNA)). This is consistent with other evidence of spacer length having subtantial effects on the efficiency of dCas9-based transcriptional regulators[36], underscoring the need to control for guide length when measuring

the effects of sgRNA secondary structure. In fact, length effects may be the underlying cause for the observation that sgRNAs with guanine-dinucleotide extensions have increased specificity[37].

These data describe nonlinear effects of 5′ sgRNA extensions on SpCas9 binding to DNA, dependent on both the length and secondary structure of the spacer. This relationship is characterized by three key regions in the data (Fig. 1c). First, extensions to the 20-nucleotide spacer cause a decrease in overall binding that is independent of secondary structure (Fig. 1c, 'Region 1'). Second, extensions that form weaker predicted secondary structures do not seem to have measurable effects on SpCas9 binding beyond those caused by length effects (Fig. 1c, 'Region 2'); however, it is possible that R-loop formation is still being inhibited in this region[38,39]. Finally, more stable hairpins cause measurable decreases in Cas9 binding as a function of the strength of the hp-sgRNA's secondary structure (Fig. 1c, 'Region 3'). Further, these decreases in activity occur as the hairpin extends into the seed region of the sgRNA that is critical for initiating the interaction between Cas9 and a target. The trend of hairpin structure modulating targeted gene activation was corroborated at two additional gene targets in human cells (Supplementary Fig. 1).

Although we ascribe the changes in gene activation to modulation of R-loop formation by hp-sgRNAs, previous studies showed by northern blot that 5′ extensions to sgRNAs were efficiently processed to 20-nucleotide spacers[19,33]. To control for both processing of the hairpins and expression of sgRNA variants, we repeated this experiment, collected total RNA and performed sample-matched measurements of *IL1RN* and sgRNA expression by reverse transcription with quantitative PCR (RT–qPCR), and 5′ sgRNA processing by 5′ rapid amplification of cDNA ends (RACE) followed by RNA sequencing (Supplementary Fig. 2a,b). Patterns in *IL1RN* gene activation were faithfully replicated (Supplementary Fig. 2c,d). We observed no correlation between hp-sgRNA expression and hp-sgRNA activity (Supplementary Fig. 2e,f).

In contrast to the previous reports[19,33], we observed that hp-sgRNAs are moderately to minimally processed, with stronger predicted secondary structures undergoing less processing (Supplementary Fig. 2g, range 0.8–48% processed). The corresponding ns-sgRNAs had higher rates of processing (Supplementary Fig. 2h, range 52–79% processed). We observed no clear association between the level of hp-sgRNA processing and *IL1RN* transactivation (Supplementary Fig. 2i,j). These data suggest that hp-sgRNAs are maintained in cells and can be accommodated within the Cas9 binding pocket where they are protected from processing.

Kinetic modeling of R-loop formation. The differences in behavior between hp-sgRNAs and ns-sgRNAs indicate that the secondary structure of the spacer is a critical determinant of CRISPR activity. To gain a better understanding of how spacer secondary structure might affect SpCas9 behavior, we applied a kinetic model of R-loop formation and generalized it to accommodate any species of mismatches, an arbitrary number of mismatches and RNA secondary structure (Fig. 2a)[29]. Strand invasion is represented as a series of 20 discrete states and the probability of exchange between states is governed by 3 energetic processes: (1) hybridization or melting of the genomic target (DNA-DNA), (2) the hybridization or melting of the spacer to the genomic target (RNA–DNA) and (3) the breaking or forming of spacer secondary structure (RNA–RNA). This approach defines the kinetics of R-loop formation entirely in terms of empirically measured thermodynamic values of nucleic acid pairs (see Methods).

To test the model, we used previously reported chromatin immunoprecipitation followed by sequencing (ChIP-seq) data of 16 sgRNAs and 12,181 called binding sites of dCas9[40,41]. We simulated the mean residence time of each of the 16 sgRNAs to each of the reported binding sites, compared this simulation with the measured ChIP-seq signal and combined correlations across sgRNAs using Fisher's method. We find correlation coefficients of 0.285 (95% confidence: 0.252, 0.317) when the simulation is initiated at the PAM-proximal site and a correlation of 0.380 (95% confidence: 0.349, 0.410) if initiated with a preformed R-loop (Fig. 2b). These correlations were higher than the previously reported best performing feature, chromatin accessibility[40]. The predictive power of our model demonstrates that the dynamics of R-loop formation play an important role in Cas9 binding to DNA.

To determine the contribution of spacer secondary structure to the model's predictive power, we removed the energetic terms for RNA folding from the reaction rates. We observed a decrease in correlation from 0.285 to 0.194 (95% confidence: 0.160, 0.228) if the simulation is initiated at the PAM-proximal nucleotides or from 0.380 to 0.273 (95% confidence: 0.240, 0.305) if the simulation is initiated with the R-loop already preformed (Fig. 2c). Finally, we performed simulations to predict the behavior of the hp-sgRNA variants used to modulate the expression of the *IL1RN* promoter in Fig. 1 (Fig. 2d). We found a strong correlation, 0.915, between estimated binding lifetime and fold increase in gene expression. Collectively, these findings suggest that spacer secondary structure influences Cas9 binding activity by modulating invasion kinetics and stability of the R-loop, key determinants of nucleolytic activation of SpCas9[30].

hp-sgRNAs increase the gene editing specificity of SpCas9. We next assessed the effect of spacer secondary structure on SpCas9 nuclease activity. It was our hypothesis that hairpin structures could increase nuclease specificity by modulating R-loop formation without necessarily altering binding to target sites[29,30]. Thus, for hp-sgRNAs designed for the SpCas9 nuclease, we generally chose hairpins with predicted free energies weaker than $-15\,\mathrm{kcal\,mol^{-1}}$, that is, within Region 1 of Fig. 1c, since any further increase in hairpin stability resulted in significant decreases in SpCas9 binding to its on-target site. To assess the effects of engineered hp-sgRNAs on the nuclease activity and specificity of Cas9 in human cells, we chose spacers that have large numbers of well-characterized off-target sites[42]. We generated a variety of hp-sgRNAs for these spacers where we varied several hp-sgRNA structural characteristics, including utilizing both external and internal loops or adjusting PAM-distal and PAM-proximal stem placement. We measured indel frequency at on-target and off-target sites for each spacer and compared the activity of these hp-sgRNAs to activities of both unextended sgRNAs (WT-sgRNAs) and truncated sgRNAs (tru-sgRNAs)[23]. We observed a number of hp-sgRNA designs with on-target activities comparable to WT-sgRNAs and reduced off-target activity, comparable to tru-sgRNAs (Fig. 3a–c and Supplementary Figs. 3–7). We defined a specificity metric by dividing on-target mutation rates by the sum of all off-target mutation rates. All optimized hp-sgRNAs significantly increased the specificity of SpCas9, on par with increases observed with tru-sgRNAs (Fig. 3d and Supplementary Fig. 6e). hp-sgRNA 7 of the *EMX1.1* spacer, which had the highest fold increase in specificity, had both a spacer truncation and designed secondary structure, suggesting that these approaches may be combined in some cases (Supplementary Fig. 6e). We observed that tru-sgRNAs increase off-target activity at 8 of the 37 off-target loci (Fig. 3a–c). This increase may be due to the decreased sequence complexity of tru-sgRNAs and was not observed for any hp-sgRNA variants, consistent with hp-sgRNAs behaving in an entirely inhibitory manner (Fig. 3a–c and Supplementary Fig. 6a–c). Collectively these results show that hp-sgRNAs can increase the specificity of SpCas9 nuclease by multiple orders of magnitude.

To test whether the 5′ extensions of hp-sgRNAs might lead to any new off-target cleavage events beyond what had previously been identified for the corresponding WT-sgRNAs, we performed CIRCLE-seq (circularization for in vitro reporting of cleavage

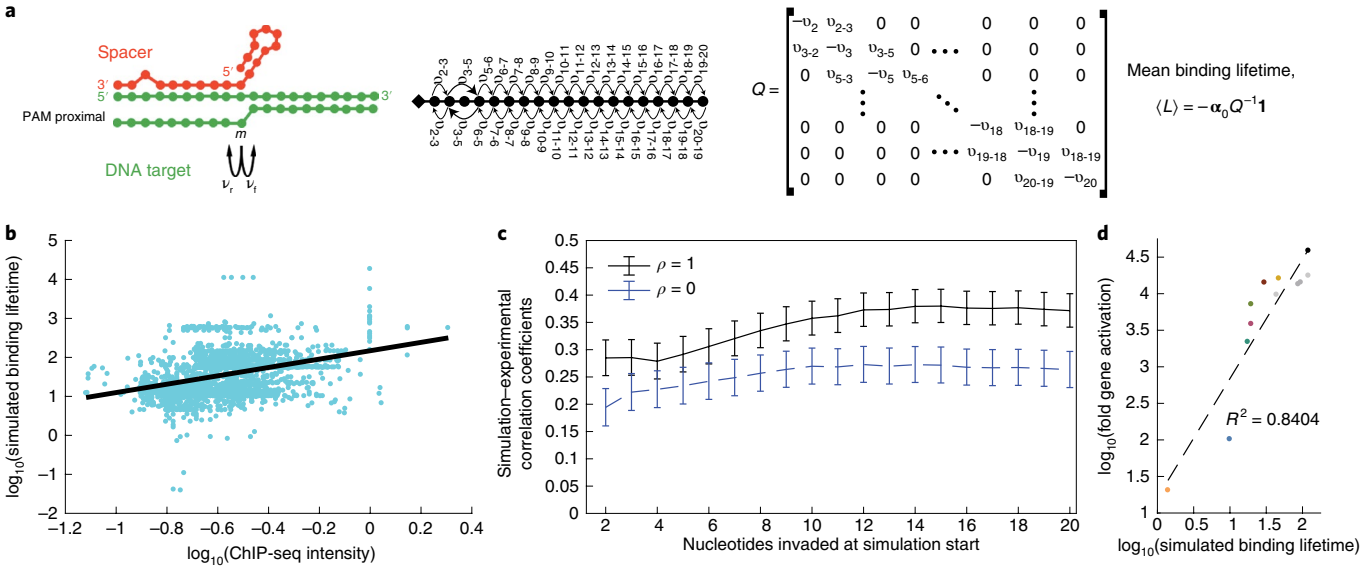

Fig. 2 | Spacer secondary structure improves the performance of a kinetic model of R-loop formation. a, Schema of kinetic model of R-loop formation. Left panel: modeled molecular interactions. The target DNA is shown in green and sgRNA spacer is shown in red with both a mismatch and RNA secondary structure. Center panel: distinct states representing degree of R-loop formation by the spacer. The forward and reverse rates between states are calculated using the free energy differences between states (see Methods). Right panel: *Q*-matrix of forward and reverse reaction rates. The starting state of the simulation is represented by vector $\boldsymbol{\alpha}_0$. **b**, Correlation between model-based predictions of binding lifetime and the ChIP-seq intensity[40,41]. Model was initiated with a preformed R-loop. For each sgRNA, log(*L*) was correlated (Pearson) with log(ChIP-seq intensity), and these correlations combined using Fisher's method, $n = 12{,}181$. **c**, Correlation coefficients with ($\rho = 1$) and without ($\rho = 0$) energetic contributions from spacer secondary structure, for various starting states. Plots show the calculated Pearson correlation coefficient, and error bars are 95% confidence intervals. **d**, Simulated values of the mean binding lifetimes for sgRNA variants, shown in Fig. 1b, plotted against their activation of the *IL1RN* gene, $n = 12$.

effects by sequencing), an unbiased in vitro method to determine genome-wide cleavage events[43]. We performed CIRCLE-seq using the *EMX1.1* spacer and used WT-, tru- and, hp-sgRNA variants; off-targets were reliably identified across replicates for each sgRNA variant (Supplementary Fig. 7a–d). Comparing with WT-sgRNA, the tru-sgRNA eliminated 77 off-target sites but also had 25 unique off-target sites that were reproducibly detected using CIRCLE-seq (Supplementary Figs. 8a and 9a, b). In contrast, the hp-sgRNA eliminated 124 off-target sites found with the WT-sgRNA and generated no unique off-target sites (Supplementary Figs. 8b and 9a, c).

We next sought insight into the mechanism of specificity increases driven by hp-sgRNAs—in particular, whether this was a result of decreased binding to DNA. We performed chromatin immunoprecipitation with quantitative PCR (ChIP-qPCR) to measure the relative enrichment of the nuclease-null dSpCas9 at on-target versus off-target sites using the same *EMX1* spacer tested with nuclease-active SpCas9. We observed that both the hp-sgRNAs and tru-sgRNA yielded similar levels of dCas9 occupancy at the on-target site (Fig. 4a). Interestingly, hp-sgRNA 2 did not measurably decrease dCas9 occupancy at any of the measured off-target sites relative to the WT-sgRNA (Fig. 4b–d), even though nuclease activity was reduced at these sites by an order of magnitude or more (Fig. 4e and Supplementary Fig. 6b). This suggests that, similar to high-fidelity Cas9 variants[24], hp-sgRNAs do not mediate specificity increases through a decrease in binding. Hp-sgRNA 7 had more variable behavior, which we attribute to the combination of a hairpin and a truncated spacer.

hp-sgRNAs increase specificity of Cas9 and Cas12a variants. We next tested whether hp-sgRNA designs can be extended to other CRISPR systems. In particular, we were interested in SaCas9 because its compact size facilitates delivery by AAV vectors and is therefore of significant interest for gene therapy applications[6,44]. While SaCas9 and SpCas9 have many analogous domains and a similar bilobed structure, they share only 17% sequence similarity[45].

Focusing on SaCas9 and SaCas9-KKH, a relaxed PAM variant, we designed hp-sgRNAs of varying stem lengths using target sites with previously characterized off-target effects[6,13]. We delivered sgRNA variants with each SaCas9 to human cells and assayed for nuclease activity at on-target and off-target loci. Similar to SpCas9, SaCas9 activity is tuned by hp-sgRNAs according to the strength of predicted secondary structure (Fig. 5a,b and Supplementary Fig. 10a–c). tru-sgRNAs of varying length were also used, though they did not eliminate off-target activity without severely impacting on-target activity; shorter truncations resulted in complete abrogation of off-target and on-target nuclease activity (Fig. 5a,b and Supplementary Fig. 10a–c; data not shown).

We next tested whether hp-sgRNAs could be applied to type V Cas12a nucleases. While SpCas9 and Cas12a share a bilobed architecture, they share no structural or sequence homology other than a single RuvC domain[46]. Cas12a nucleases are unique in that they can process their own crRNAs, and these crRNAs are sufficient for Cas12a target recognition and cleavage[47]. Cas12a recognizes its crRNA via a hairpin that is at the 5′ end of the crRNA and the spacer is at the 3′ end: the reverse orientation relative to Cas9 sgRNA structure. Target recognition by Cas12a and R-loop formation mechanisms are also reversed when comparing with that of Cas9: the PAM sequence is located at the 5′ end of the target sequence and R-loop formation of the target strand proceeds 3′ to 5′. Despite these many differences, we hypothesized that the activity of Cas12a nucleases could also be regulated by spacer secondary structure. Using a spacer with previously characterized off-target sites[14,15,48], we designed hp-crRNAs with varying structural stability. We observed that both AsCas12a and LbCas12a activity can be regulated by spacer secondary structure and that off-target activity can be reduced without altering on-target activity by tuning the strength of the secondary structure (Fig. 5c,d and Supplementary Fig. 11a–c). Truncated crRNAs did not consistently result in specificity increases for either AsCas12a or LbCas12a, indicating that this

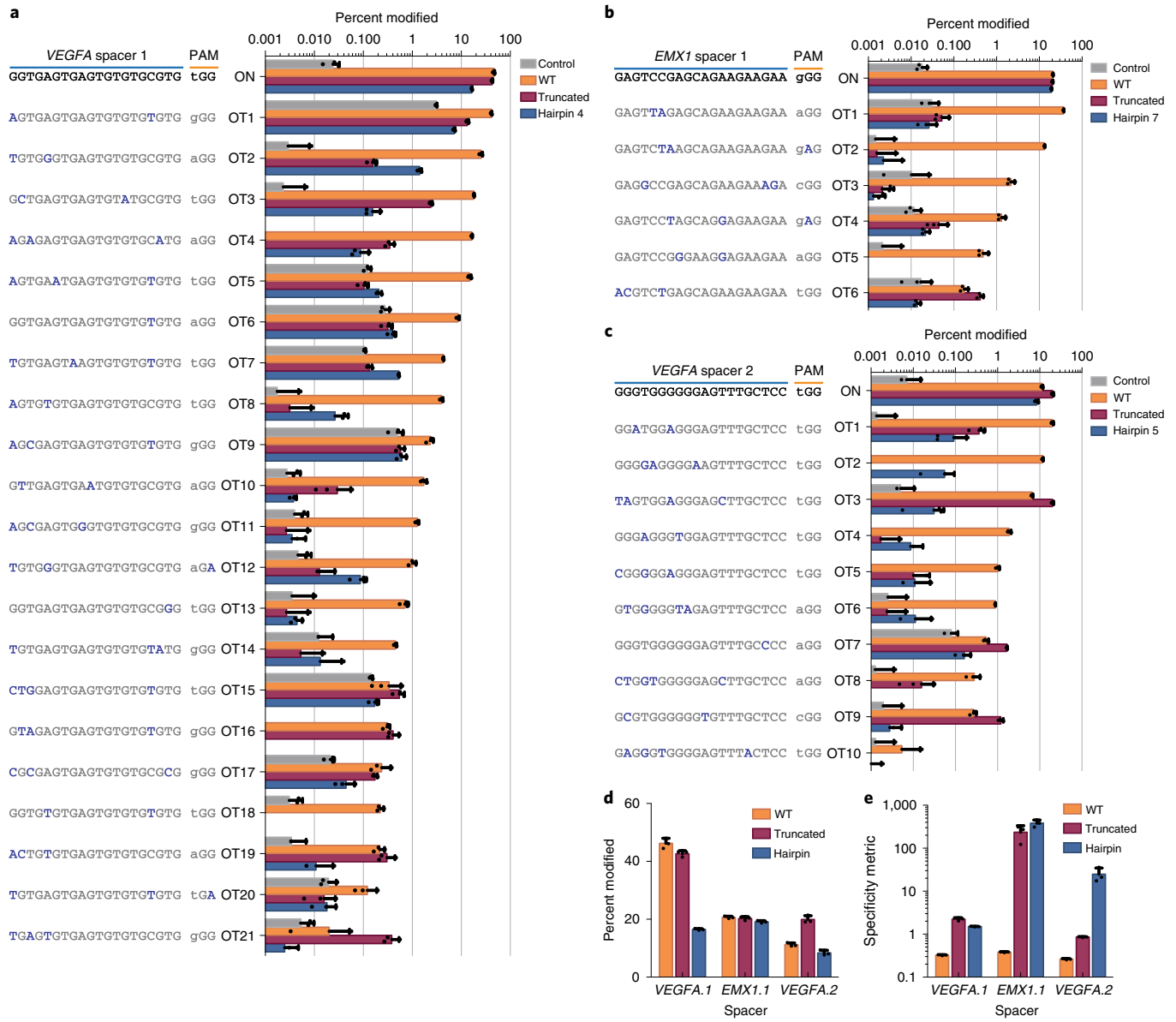

Fig. 3 | hp-sgRNAs increase the specificity of SpCas9 in human cells. a–c, On-target and off-target mutation rates for sgRNA variants targeting the *EMX1* and *VEGFA* genes, measured by deep sequencing: *VEGFA* spacer 1 (**a**), *EMX1* spacer 1 (**b**) and *VEGFA* spacer 2 (**c**). 'Percent modified' indicates percentage of reads containing indels compared with the wild-type sequence (mean + s.e.m., $n = 3$). WT-sgRNAs ('WT') generated significant editing activity at all off-target sites, except for *VEGFA* spacer 2 at OT10 ($P < 0.01$). hp-sgRNAs show significant decreases in activity at all measured off-target sites when compared with WT-sgRNA ($P < 0.05$). Hypothesis testing using a one-sided Fisher exact test with pooled read counts, adjusting for multiple comparisons using the Benjamini–Hochberg method. **d**,**e**, On-target activity (**d**) and specificity metric (**e**) for different sgRNA variants. Samples labeled as 'hairpin' use the same hairpin variant listed in panels **a–c**. The specificity metric is defined as on-target indel rate divided by the sum of all off-target indel rates (mean + s.e.m., $n = 3$). The sequences of sgRNA variants are listed in Supplementary Table 1. The predicted structures of hp-sgRNAs are displayed in Supplementary Figs. 3–5.

strategy might not be consistently translatable to Cas12a nucleases (Fig. 5c–d and Supplementary Fig. 11a–c). Shorter truncations of the spacer resulted in complete abrogation of off-target and on-target nuclease activity. We observed that hp-crRNAs influence the activity of Cas12a nucleases according to the strength of the secondary structure, consistent with the effect of hp-sgRNAs on SpCas9 and SaCas9 activity (Fig. 5c,d and Supplementary Fig. 11a–c). Significantly, as predicted folding energy increases, decreases in gene editing activity occur preferentially at off-target loci, allowing for increases in specificity (Fig. 5i).

To confirm that increases in specificity are caused by RNA secondary structures, we generated ns-sgRNAs for hp-sgRNAs used with Cas9 and Cas12a effectors. For each Cas effector we generally chose hp-sgRNA variants that maintained on-target activity but had the most stable predicted free energy. We delivered these sgRNA variants with their respective Cas nuclease and used deep sequencing to assay mutational rates at both on-target and off-target loci (Fig. 6a–e). Across 12 spacer sequences and 6 different Cas9 or Cas12a variants, hp-sgRNAs increased specificity by an average of 55-fold (median 12-fold) compared with unmodified sgRNAs and 9-fold compared with length-matched non-structured control sgRNAs (Fig. 6f and Supplementary Fig. 12). Hp-sgRNAs showed particular sensitivity to off-targets with multiple mismatches (Supplementary Fig. 13).

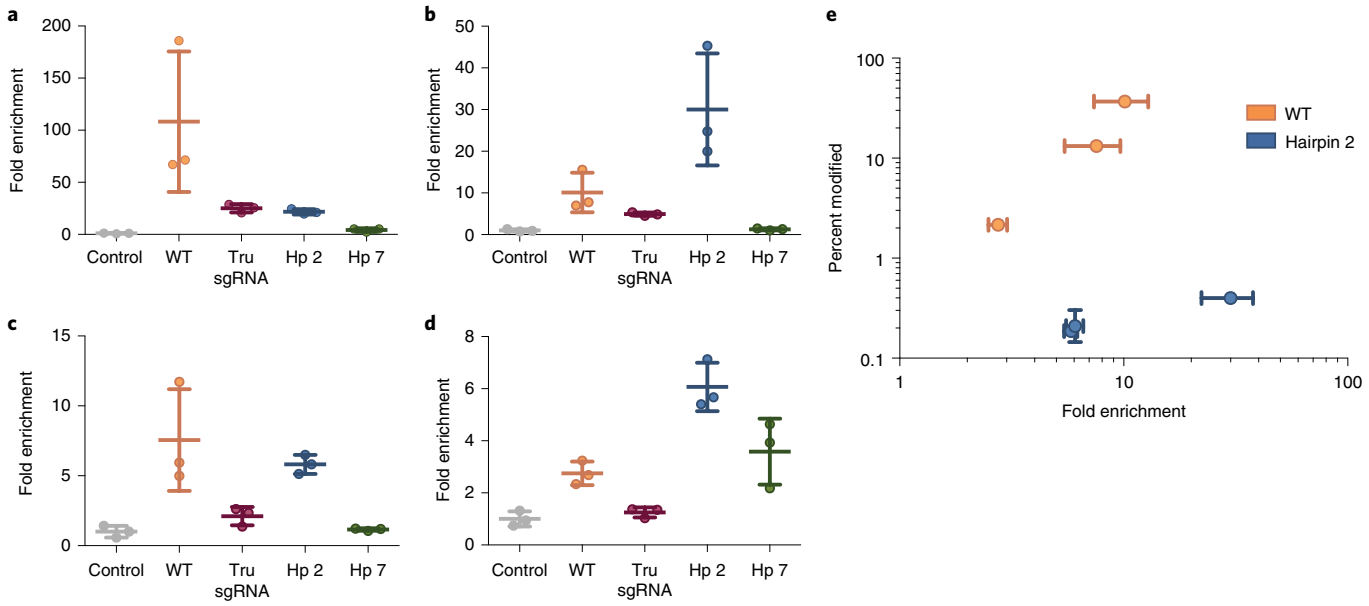

Fig. 4 | hp-sgRNAs retain binding activity at off-target loci. a, dCas9 enrichment at the on-target site using sgRNA variants containing *EMX1* spacer 1 by ChIP-qPCR. The WT-sgRNA sample had significant enrichment over control, $P < 0.001$. The tru-sgRNA and hp-sgRNAs showed a decreased enrichment relative to WT-sgRNA, $P < 0.05$. **b**–**d**, dCas9 enrichment at designated off-target sites (OT1 (**b**), OT2 (**c**), OT3 (**d**)) using sgRNA variants containing *EMX1* spacer 1 by ChIP-qPCR. hp-sgRNAs were also assayed for editing activity with nuclease-active SpCas9 (Supplementary Fig. 6b), and their predicted secondary structure is shown in Supplementary Fig. 1. **e**, Off-target editing rates, as shown in Supplementary Fig. 5b, as a function of corresponding DNA binding as measured by ChIP-qPCR. Hairpin 2, when compared with WT, showed significantly decreased editing activity at off-target sites ($P < 5 \times 10^{-20}$), but showed no significant decreases in ChIP enrichment (mean + s.e.m., $n = 3$). *P* values for ChIP-qPCR data were calculated using a post hoc Tukey test after a global one-way ANOVA. For editing activity, hypothesis testing was carried out using a one-sided Fisher exact test with pooled read counts, adjusting for multiple comparisons using the Benjamini–Hochberg method. All fold enrichments are relative to transfection of a control sgRNA plasmid targeted to the *IL1RN* promoter and normalized to a region of the β-actin locus. The sequences of sgRNA variants are listed in Supplementary Table 1.

To further ensure that the specificity increases were due to modulation of kinetics of R-loop formation, rather than changes to expression or stability that could occur within transfected cells, we completed in vitro assays for nuclease activity and DNA binding. For in vitro nuclease activity, we digested PCR amplicons containing the on-target *EMX1* spacer 1, *EMX1* spacer 2 or *DNMT1* spacer 1, by defined concentrations of purified SpCas9, SaCas9 or AsCas12a protein, respectively, complexed with corresponding chemically synthesized WT-, hp- or ns-sgRNAs (Supplementary Fig. 14). At the on-target sites, the activity of the hp-sgRNAs was reduced by 85%, 59% and 69% relative to activity of WT-gRNAs at the on-target sites for SpCas9, SaCas9 and AsCas12a, respectively, compared with a reduction of 12% and increases of 35% and 6% with the corresponding ns-sgRNAs. The significant reduction of activity of hp-sgRNAs at on-target sites in vitro, but not in cells (Figs. 3b,d and 6a,c), may be the result of the short time frame of the assay or other differences with the intracellular environment in which these particular hairpin structures were optimized. We also tested identical digestion reactions with PCR amplicons containing the corresponding off-target 1 (OT1) spacer sequence. At the off-target sites, hp-sgRNAs also showed decreases of 91%, 79% and 67% relative to WT-sgRNAs, compared with decreases of 88%, 38% and 0% for the ns-sgRNAs. To assay DNA binding, we used atomic force microscopy (AFM) to directly image and quantify interactions of the same combinations of Cas effectors and sgRNAs at on-target and off-target sequences (Supplementary Fig. 15). These analyses showed that only hp-sgRNAs, and not ns-sgRNAs, robustly and reproducibly decreased occupancy at off-target sites relative to the on-target site. Collectively, these data support that, under controlled conditions of in vitro reactions, hairpin structure—and not simply any 5′ extension—modulates CRISPR activity.

Discussion

CRISPR–Cas endonucleases did not evolve to function for highly specific gene editing of mammalian genomes, and cases of off-target activity have been reported for the majority of CRISPR endonucleases tested so far in human cells. Additionally, the discovery of novel CRISPR systems with potential biotechnological applications is occurring at a steady pace. Hence, there is a need to improve the performance of CRISPR endonucleases that is robust and can be applied easily across CRISPR systems.

The rational design of hp-sgRNAs as characterized in this study is a promising method to meet this need. For 5 of the most commonly applied Cas effectors, utilizing well-characterized off-target sites, we demonstrate that rationally designed RNA secondary structures increase specificity by an average of 55-fold. Moreover, despite the widely ranging biochemical properties of each Cas effector used, we observe consistent behavior of hp-sgRNAs, where CRISPR activity is inhibited as a function of the stability of the secondary structure.

The strategy used in this study was inspired by previous efforts, which aimed to increase nuclease specificity by weakening direct interactions between Cas9 and the DNA[21,22]. While we do not directly determine the mechanism of hp-sgRNA-driven specificity increase, we hypothesize that it occurs through inhibition of R-loop kinetics, which inhibits the structural transitions of the CRISPR endonuclease that are necessary for activity at off-target sites[30]. The evidence for this is threefold. First, using ChIP-qPCR we show that hp-sgRNAs do not decrease dCas9 binding at off-target sites, even when nuclease activity is reduced by orders of magnitude (Fig. 4e). This is evidence that nuclease activity is diminished by the inhibition of full R-loop formation. Second, because RNA–DNA duplexes are regularly accommodated in the central binding channel of CRISPR endonucleases, it is likely that RNA–RNA duplexes are similarly accommodated without interfering with RNP complex formation.

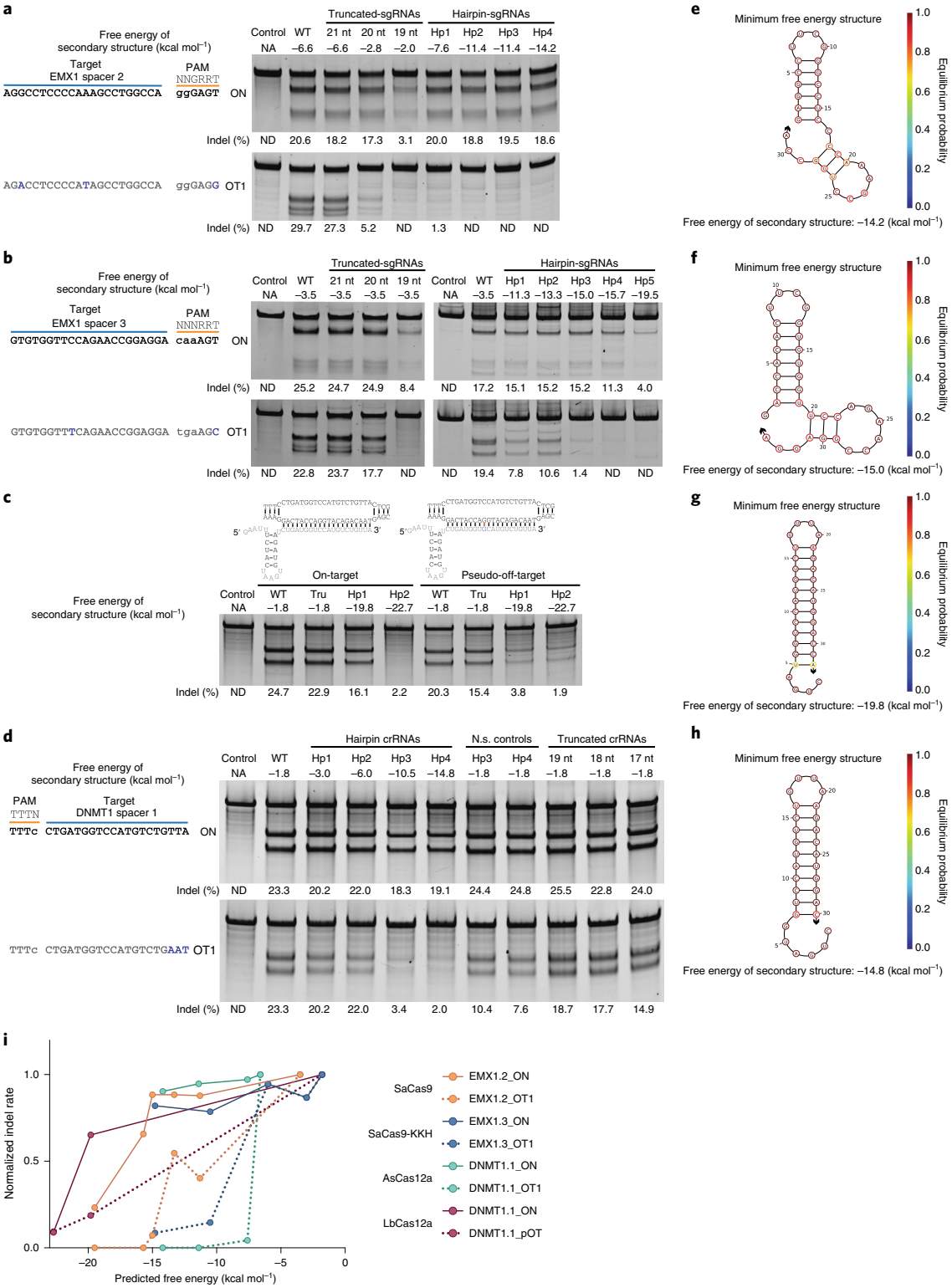

Fig. 5 | hp-sgRNAs and -crRNAs increase the specificity of various Cas effectors. a–d, On-target and off-target nuclease activity of sgRNA and crRNA variants with SaCas9 (**a**), SaCas9-KKH (**b**), LbCas12a (**c**) and AsCas12a (**d**). Plasmids encoding the Cas effector and the sgRNA or crRNA variant were transfected into human cells and mutational activity was measured using the Surveyor nuclease assay. Representative gels are shown from optimizations that were performed one to three times. Optimized structures were further investigated with deep sequencing in Fig. 6. Aliases for sgRNA variants are listed above each lane and are detailed in Supplementary Table 1. tru-sgRNAs/crRNAs are abbreviated as 'Tru', hp-sgRNAs/crRNAs are abbreviated as 'Hp' and non-structured controls are abbreviated as 'N.s.'. For LbCas12a, off-target activity was generated by introducing a mismatch in the sgRNA spacer, as shown, and is referred to as a 'pseudo-off-target'. **e–h**, Predicted structure of optimized hp-sgRNA spacers (hairpin 1 (**e**), hairpin 3 (**f**), hairpin 1 (**g**), hairpin 4 (**h**)); arrows indicate 3′ end of RNA. The sequences of the sgRNA variants are listed in Supplementary Table 1. **i**, Normalized nuclease activity of WT-sgRNAs and various hp-sgRNAs, plotted against their predicted free energy of secondary structure folding. Data from panels **a–d** were normalized to the WT-sgRNA activity at the corresponding on-target (solid line) or off-target site (dotted line).

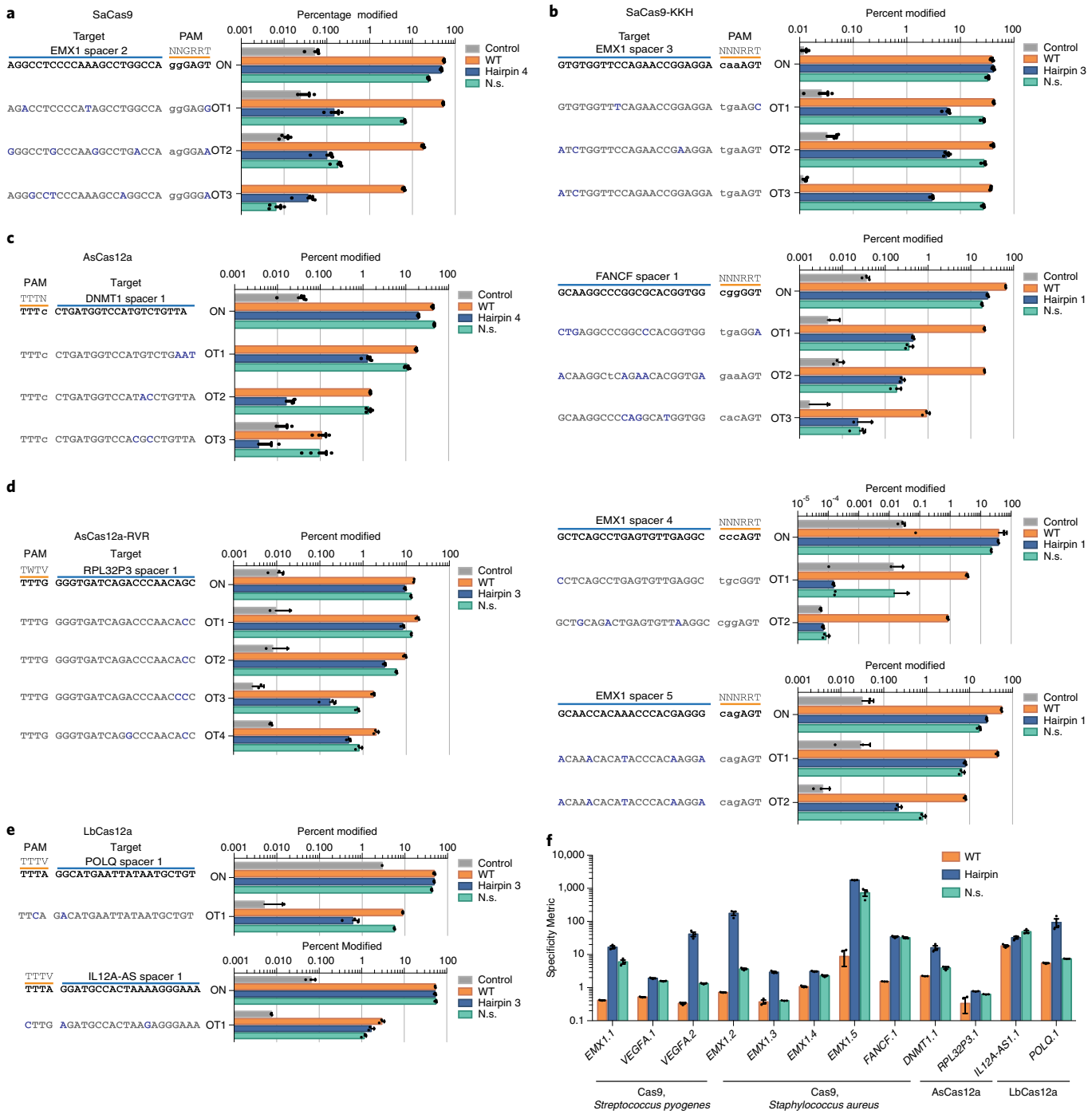

Fig. 6 | RNA secondary structure drives the specificity increases observed with hp-sgRNAs. a–e, Nuclease activity of hp-sgRNAs/crRNAs and corresponding non-structured controls in human cells, applied with SaCas9, SaCas9-KKH, AsCas12a, AsCas12a-RVR and LbCas12a, respectively. Deep sequencing was used to measure editing activity of Cas effector-sgRNA pairs. WT-sgRNAs induced significant editing activity at all off-target sites ($P < 1 \times 10^{-9}$). hp-sgRNAs/crRNAs significantly reduced editing activity at all examined off-target sites when compared with WT-sgRNA/crRNA ($P < 5 \times 10^{-11}$). Hypothesis testing was carried out using a one-sided Fisher exact test with pooled read counts, adjusting for multiple comparisons using the Benjamini–Hochberg method. **f**, Specificity metric for sgRNA variants applied with the indicated Cas effector (mean + s.e.m., $n = 3$). The gene target of each spacer is listed on the *x* axis.

This is supported by evidence that sgRNAs with significant spacer secondary structure could readily complex with SpCas9[49]. Finally, the predictive power of our kinetic model supports its principle hypothesis: that R-loop formation is a kinetic process that is modulated by RNA secondary structures. Collectively, these points suggest that sgRNA–endonuclease complex levels are maintained and that observed specificity increases are caused by secondary-structure mediated inhibition of R-loop formation, limiting the conformation change to an activated endonuclease at off-target sites.

Our study considers R-loop formation as the central process governing CRISPR nuclease activity: its modulation allows for more specific genome editing and its modeling facilitates predictions of CRISPR activity. Improvements to the modeling of this process would be broadly useful for in silico prediction of off-target effects

and for designing functional hp-sgRNAs a priori. As our model approximates this behavior using thermodynamic parameters of nucleic acids derived from in vitro data, further refinement of our understanding of RNA–DNA interactions and mispairing within the catalytic environment of different CRISPR endonucleases will probably improve its predictive and design performance. Recent methods using massively parallel assessment of CRISPR endonuclease binding and catalysis could provide attractive data sets for model refinement[50,51].

In this study, we demonstrate a method to increase specificity across diverse CRISPR systems. Future studies will be useful to determine whether hp-sgRNAs can similarly regulate new Cas12, Cas13 or Cas14 effectors[4,5,11,52,53]. The hp-sgRNA secondary structures that regulate specificity may be combined with other methods of sgRNA engineering to modulate activity, specificity and orthogonality[54–56]. sgRNA engineering, in conjunction with careful spacer choice and optimized gene delivery, could enable higher specificity of CRISPR nucleases for next-generation genome editing and facilitate realizing the potential of CRISPR for sensitive therapeutic and diagnostic applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-019-0095-1.

Received: 31 March 2018; Accepted: 11 March 2019;
Published online: 15 April 2019

References

1. Barrangou, R. & Doudna, J. A. Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.* **34**, 933–941 (2016).
2. Wright, A. V., Nunez, J. K. & Doudna, J. A. Biology and applications of CRISPR systems: harnessing Nature's toolbox for genome engineering. *Cell* **164**, 29–44 (2016).
3. Shmakov, S. et al. Discovery and functional characterization of diverse class 2 CRISPR–Cas systems. *Mol. Cell* **60**, 385–397 (2015).
4. Burstein, D. et al. New CRISPR–Cas systems from uncultivated microbes. *Nature* **542**, 237–241 (2017).
5. Yan, W. X. et al. Functionally diverse type V CRISPR–Cas systems. *Science* **363**, 88–91 (2019).
6. Ran, F. A. et al. In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
7. Hou, Z. et al. Efficient genome engineering in human pluripotent stem cells using Cas9 from *Neisseria meningitidis*. *Proc. Natl Acad. Sci. USA* **110**, 15644–15649 (2013).
8. Kim, E. et al. In vivo genome editing with a small Cas9 orthologue derived from *Campylobacter jejuni*. *Nat. Commun.* **8**, 14500 (2017).
9. Chatterjee, P., Jakimo, N. & Jacobson, J. M. Minimal PAM specificity of a highly similar SpCas9 ortholog. *Sci. Adv.* **4**, eaau0766 (2018).
10. Zetsche, B. et al. Cpf1 Is a single RNA-guided endonuclease of a class 2 CRISPR–cas system. *Cell* **163**, 759–771 (2015).
11. Abudayyeh, O. O. et al. RNA targeting with CRISPR–Cas13. *Nature* **550**, 280–284 (2017).
12. Konermann, S. et al. Transcriptome engineering with RNA-targeting type VI-D CRISPR effectors. *Cell* **173**, 665–676.e14 (2018).
13. Kleinstiver, B. P. et al. Broadening the targeting range of *Staphylococcus aureus* CRISPR–Cas9 by modifying PAM recognition. *Nat. Biotechnol.* **33**, 1293–1298 (2015).
14. Kleinstiver, B. P. et al. Genome-wide specificities of CRISPR–Cas Cpf1 nucleases in human cells. *Nat. Biotechnol.* **34**, 869–874 (2016).
15. Kim, D. et al. Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat. Biotechnol.* **34**, 863–868 (2016).
16. Maeder, M. L. & Gersbach, C. A. Genome editing technologies for gene and cell therapy. *Mol. Ther.* **24**, 430–446 (2016).
17. Tsai, S. Q. et al. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* **32**, 569–576 (2014).
18. Shen, B. et al. Efficient genome modification by CRISPR–Cas9 nickase with minimal off-target effects. *Nat. Methods* **11**, 399–402 (2014).
19. Ran, F. A. et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).
20. Guilinger, J. P., Thompson, D. B. & Liu, D. R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat. Biotechnol.* **32**, 577–582 (2014).
21. Slaymaker, I. M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
22. Kleinstiver, B. P. et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
23. Fu, Y. et al. Improving CRISPR–Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).
24. Chen, J. S. et al. Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).
25. Bolukbasi, M. F. et al. DNA-binding-domain fusions enhance the targeting range and precision of Cas9. *Nat. Methods* **12**, 1150–1156 (2015).
26. Casini, A. et al. A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat. Biotechnol.* **36**, 265–271 (2018).
27. Lee, J. K. et al. Directed evolution of CRISPR–Cas9 to increase its specificity. *Nat. Commun.* **9**, 3048 (2018).
28. Vakulskas, C. A. et al. 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 (2018).
29. Josephs, E. A. et al. Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage. *Nucleic Acids Res.* **43**, 8924–8941 (2015).
30. Sternberg, S. H. et al. Conformational control of DNA target cleavage by CRISPR–Cas9. *Nature* **527**, 110–113 (2015).
31. Bevilacqua, P. C. & Blose, J. M. Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. *Annu. Rev. Phys. Chem.* **59**, 79–103 (2008).
32. Klosterman, P. S. et al. Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns. *Nucleic Acids Res.* **32**, 2342–2352 (2004).
33. Zalatan, J. G. et al. Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell* **160**, 339–350 (2015).
34. Perez-Pinera, P. et al. RNA-guided gene activation by CRISPR–Cas9-based transcription factors. *Nat. Methods* **10**, 973–976 (2013).
35. Hilton, I. B. et al. Epigenome editing by a CRISPR–Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**, 510–517 (2015).
36. Gilbert, L. A. et al. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661 (2014).
37. Kim, D. et al. Genome-wide target specificities of CRISPR–Cas9 nucleases revealed by multiplex Digenome-seq. *Genome Res.* **26**, 406–415 (2016).
38. Dahlman, J. E. et al. Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease. *Nat. Biotechnol.* **33**, 1159–1161 (2015).
39. Kiani, S. et al. Cas9 gRNA engineering for genome editing, activation and repression. *Nat. Methods* **12**, 1051–1054 (2015).
40. Wu, X. et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–674 (2014).
41. Kuscu, C. et al. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32**, 677–683 (2014).
42. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
43. Tsai, S. Q. et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).
44. Nelson, C. E. et al. In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy. *Science* **351**, 403–407 (2016).
45. Nishimasu, H. et al. Crystal Structure of *Staphylococcus aureus* Cas9. *Cell* **162**, 1113–1126 (2015).
46. Yamano, T. et al. Crystal structure of Cpf1 in complex with guide RNA and target DNA. *Cell* **165**, 949–962 (2016).
47. Fonfara, I. et al. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517–521 (2016).
48. Yan, W. X. et al. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat. Commun.* **8**, 15058 (2017).
49. Thyme, S. B. et al. Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat. Commun.* **7**, 11750 (2016).
50. Boyle, E. A. et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl Acad. Sci. USA* **114**, 5461–5466 (2017).
51. Jung, C. et al. Massively parallel biophysical analysis of CRISPR–Cas complexes on next generation sequencing chips. *Cell* **170**, 35–47.e13 (2017).

52. Harrington, L. B. et al. Programmed DNA destruction by miniature CRISPR–Cas14 enzymes. *Science* **362**, 839–842 (2018).
53. Liu, J. J. et al. CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* **566**, 218–223 (2019).
54. Briner, A. E. et al. Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol. Cell* **56**, 333–339 (2014).
55. Yin, H. et al. Partial DNA-guided Cas9 enables genome editing with reduced off-target activity. *Nat. Chem. Biol.* **14**, 311–316 (2018).
56. Kartje, Z. J. et al. Chimeric guides probe and enhance Cas9 biochemical activity. *Biochemistry* **57**, 3027–3031 (2018).

Acknowledgements

We thank C. E. Nelson and T. S. Klann for useful discussions related to experimental design and execution. This work was supported by an Allen Distinguished Investigator Award from the Paul G. Allen Frontiers Group; a US National Institutes of Health (NIH) Director's New Innovator Award (no. DP2OD008586); NIH grant nos. R01DA036865, R01AR069085 and P30AR066527; and National Science Foundation grant nos. DMR-1709527 and EFMA-1830957.

Author contributions

D.D.K., E.A.J. and C.A.G. designed the experiments. D.D.K., E.A.J., V.B., S.S.A. and J.B.K. performed the experiments. D.D.K., E.A.J. and C.A.G. analyzed the data. D.D.K., E.A.J. and C.A.G. wrote the manuscript with input from all authors.

Competing interests

D.D.K., E.A.J. and C.A.G. have filed for a patent related to this work. C.A.G. is an advisor for Sarepta Therapeutics and a cofounder of and advisor for Element Genomics and Locus Biosciences.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0095-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.A.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Plasmids and oligonucleotides. Expression plasmids for the Cas effectors and their respective sgRNAs were obtained through Addgene (Addgene catalog nos. 41815, 47108, 65776, 70708, 70709, 78741, 78742, 78743, 78744); crRNA sequences are listed in Supplementary Table 1 and oligonucleotide sequences are found in Supplementary Table 2. To create sgRNA plasmids, oligonucleotides containing the target sequences were obtained from IDT, hybridized, phosphorylated and cloned in the appropriate plasmids using BbsI or BsmBI sites.

All hp-sgRNA designs were informed through the use of in silico structure determination and only spacer sequences were used for these predictions (that is, structural sequences in the tracrRNA or crRNAs were excluded)[57].

Human cell culture and transfection. HEK293T cells were obtained from the American Tissue Collection Center through the Duke University Cancer Center Facilities and were maintained in DMEM supplemented with 10% FBS and 1% penicillin-streptomycin at 37 °C with 5% CO_2. HEK293T cells were transfected with Lipofectamine 2000 (Invitrogen) according to manufacturer's instructions. Transfection efficiencies were routinely higher than 80%, as determined by fluorescence microscopy after delivery of a control eGFP expression plasmid. All transfections were performed in 24-well cell culture plates that were coated with a 1:10 dilution of poly-L-lysine (P8920 SIGMA). On day 1, cell culture plates were coated and 200,000 cells were seeded per well. On day 2, cells were put in Opti-MEM and transfected with 800 ng plasmid (600 ng of Cas effector, 200 ng sgRNA) and 2 µl Lipofectamine 2000. On day 3, medium was changed to DMEM supplemented with 10% FBS and 1% penicillin-streptomycin. Cells were collected for downstream analysis on day 5.

Surveyor assays. The region surrounding the sgRNA or crRNA target site was amplified by PCR with the AccuPrime PCR kit (Invitrogen) and 50–200 ng of gDNA as template using primers listed in Supplementary Table 3. The PCR products were melted and reannealed using the temperature program: 95 °C for 180 s, 85 °C for 20 s, 75 °C for 20 s, 65 °C for 20 s, 55 °C for 20 s, 45 °C for 20 s, 35 °C for 20 s and 25 °C for 20 s with a 0.1 °C s^{-1} decrease rate in between steps. This allows the formation of mutant and wild-type DNA strands with the consequent formation of distorted duplex DNA. Without purifying the PCR product, 18 µl of the reannealed duplex was combined with 2 µl of the Surveyor nuclease (IDT), which cleaves DNA duplexes at the sites of distortions created by either bulges or mismatches, and 1 µl of enhancer solution. This reaction was incubated at 42 °C for 60 min and then separated on a 10% TBE polyacrylamide gel. The gels were stained with ethidium bromide and quantified using ImageLab (Bio-Rad)[58].

Deep sequencing. gDNA was purified from cells using the DNeasy kit (Qiagen). Biological replicates were generated from three separate transfections for each experimental condition. On-target and off-target sites were amplified using 100 ng gDNA with AccuPrime polymerase (Invitrogen). Primers are listed in Supplementary Table 3. For some regions, 4% v/v dimethylsulfoxide was used in the PCR for efficient amplification. PCR primers included Nextera adapters for binding to Illumina flowcells. Using a second round of PCR, group-specific barcodes were added. The resulting PCR products were purified using Agencourt AMPure beads (Beckman coulter), quantified using Qubit Fluorimeter (Thermo Fisher), pooled and sequenced with 150-base pair (bp) paired-end reads on an Illumina MiSeq instrument. CRISPResso was used for sequence analysis[59]. Sequences were first trimmed to remove adapter sequences. Sequences were filtered using a minimum average quality score of 30. Reads were trimmed to remove adapter sequences. Paired reads were then merged using fast length adjustment of short reads (FLASH) to create a single sequence of higher quality; a minimum overlap of 40 bp was used. CRISPRessoPooled was then used to demultiplex reads and quantify non-homologous end joining rates. A minimum identity score of 80 was used for demultiplexing. Only insertions and deletions were used in calling CRISPR-generated non-homologous end joining events, since CRISPR-based gene editing largely causes indels and not substitutions. Each biological replicate had a minimum of 1,500 reads per loci; the average was approximately 20,000 reads per replicate per loci. Hypothesis testing was carried out using a one-sided Fisher exact test on the pooled read counts of three biological replicates. *P* values were adjusted for multiple comparisons using the method of Benjamini and Hochberg.

RT–qPCR. IL1RN activation experiment. Cells were transfected as described above. RNA was isolated using the RNeasy Plus RNA isolation kit (Qiagen). Complementary DNA synthesis was performed using the SuperScript VILO cDNA Synthesis Kit (Invitrogen). Real-time PCR using SYBR green Fastmix (Quanta BioSciences) was performed with the CFX96 Real-Time PCR Detection System (Bio-Rad) with oligonucleotide primers reported in Supplementary Table 3 that were designed using Primer3Plus software and purchased from IDT. Primer specificity was confirmed by agarose gel electrophoresis and melting curve analysis. Reaction efficiencies over the appropriate dynamic range were calculated to ensure linearity of the standard curve. The results are expressed as fold-increase messenger RNA expression of the gene of interest normalized to *GAPDH* expression by the $\Delta\Delta C_t$ method, whereby the difference in cycle number of the control sample is used to normalize the difference in cycle number of the experimental sample.

HBG1 and *IL1B* activation experiments. The day before transfection, HEK293T cells were plated at 10^5 cells per well in a 96-well plate coated with poly-L-lysine. The day of transfection, DMEM was aspirated and 100 µl Opti-MEM was added to each well. Each well was then transfected with 400 ng plasmid (300 ng of dCas9-P300 and 100 ng of sgRNA). Plasmids were brought to 25 µl with Opti-MEM. A separate mixture was made of 24.5 µl Opti-MEM and 0.5 µl Lipofectamine 2000, and this was combined with the 25-µl plasmid mixture. The 50-µl solution was incubated for 5 min and pipetted slowly onto each well. Media was changed the next day to DMEM + 10% FBS + penicillin-streptomycin. Cells were collected using Cells-to-CT 1-Step TaqMan Kit and TaqMan gene expression assays (Thermo Fisher).

Sample-matched 5′ RACE and sgRNA expression measurements. Cells were grown and transfected as described above. Cells were collected using the miRNeasy kit (Qiagen) and on-column DNase digestion was performed to rid the sample of any remaining plasmid DNA. RNA concentrations were then measured and normalized by dilution. For measurement of *IL1RN* gene activation and sgRNA expression, cDNA was created using SuperScript VILO cDNA Synthesis Kit. Primers for the sgRNA RT–qPCR were designed to bind the spacer region and end of the sgRNA scaffold. RT–qPCR was carried out as described above.

5′ RACE was carried out on the RNA samples using Maxima H Minus reverse transcriptase (EP0753, Thermo Fisher). Both the template-switch primer and sgRNA-specific reverse transcription primer were ordered from IDT. The reverse transcription primer included a 10-nt random barcode that serves as a unique molecular identifier (UMI). Reactions were run using the manufacturer protocol with slight modification. Specifically, 1 µg total RNA, 0.2 pmol reverse transcription primer, 50 pmol template-switch primer, 1 µl 10 mM dNTP mix and 4 µl 5× reverse transcription buffer were combined and brought to 19.5 µl with water. The mixture was incubated at 85 °C for 2 min to disrupt RNA secondary structure. The temperature was then brought down to 55 °C, 0.5 µl reverse transcriptase was added and the reaction was incubated at 55 °C for 30 min and terminated by incubating at 85 °C for 5 min. Then 1 µl of each reaction was used in a 50-µl PCR to enrich for the desired product, barcode and add i5 and i7 Illumina adapters. PCR product was run on an agarose gel to confirm expected product lengths. The desired sgRNA cDNAs were purified using a 0.9× bead cleanup (Agencourt AMPure XP Beads, Beckman Coulter), concentrations were measured using the high-sensitivity qubit assay and samples were pooled and run on an Illumina MiSeq instrument.

Samples were sequenced using 150-bp single-end reads at an average depth of approximately 100,000 reads per replicate. Any reads without an exact 76-nucleotide sgRNA scaffold sequence were discarded. UMI sequences were used to remove any events that might result from PCR duplication. After these 2 filters, each sample had an average of 47,675 reads with a minimum of 8,092. Spacer lengths were then calculated using locations of the sgRNA scaffold and template-switch sequence as anchors. Finally, the frequency of each observed spacer length was determined for each sample.

CIRCLE-seq. CIRCLE-seq libraries were generated largely as previously described[60].

Large quantities of HEK293T gDNA were collected as follows: 6 ml NK Lysis Buffer (50 mM Tris, 50 mM EDTA, 1% SDS, pH 8) and 30 µl 20 mg ml^{-1} Proteinase K (QIAGEN 19131) were used to resuspend 5×10^7 cells. This lysate was incubated at 55 °C overnight. The next day, 30 µl of 10 mg ml^{-1} RNase A (QIAGEN 19101) was added to the lysed sample. The sample was vortexed and incubated at 37 °C for 30 min. Samples were cooled on ice before addition of 2 ml prechilled 7.5 M ammonium acetate (Sigma A1542) to precipitate proteins. The samples were vortexed, centrifuged at ≥10,000*g* for 10 min, and the supernatant was carefully decanted into a new 15-ml conical tube. Then, 6 ml 100% isopropanol was added to the tube, inverted several times and centrifuged at ≥10,000*g* for 10 min. gDNA was visible as a small white pellet in each tube. The supernatant was discarded, and 5 ml freshly prepared 80% ethanol was added to wash the pellet and then centrifuged at ≥10,000*g* for 1 min. The supernatant was carefully discarded, and the pellet was air dried for 30 min and finally resuspended in TE buffer.

Approximately 50–100 µg of starting gDNA was needed to generate enough circles for each CIRCLE-seq reaction. Using a Diagenode Bioruptor XL sonicator at 4 °C, gDNA was sonicated to an average size of approximately 500 bp, with a visible range of 200–1,000 bp, as determined by agarose gel electrophoresis. The enzymatic procedure to generate circles was carried out as previously described[60]. For the in vitro digest of the circles, sgRNAs were synthesized from Synthego and SpCas9 was purchased from New England Biolabs. Library production was carried out as previously described[60]. Libraries were quantified using a Qubit Fluorimeter (Thermo Fisher), pooled and sequenced with 150-bp paired-end reads on an Illumina MiSeq instrument. CIRCLE-seq read counts were obtained using previously described methods and software[43]. The following parameters were used for running the CIRCLE-seq pipeline: read threshold of 4, window size of 3, mapq threshold of 50, start threshold of 1, gap threshold of 3 and mismatch threshold of 6.

ChIP-qPCR. Chromatin immunoprecipitation (ChIP) experiments were performed in biological triplicate, starting from independent cell transfections, and collected 3 d after transfection. For each replicate, 2×10^7 nuclei were resuspended

in 1 ml RIPA buffer (1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS in PBS at pH 7.4). Samples were sonicated using a Diagenode Bioruptor XL sonicator at 4 °C to fragment chromatin to 200–500-bp segments. Insoluble components were removed by centrifugation for 15 min at 15,000 r.p.m. Then, 5 µg of FLAG M2 antibody (F1804) was conjugated with sheep anti-mouse IgG magnetic beads (Life Technologies, 11203D/11201D). Sheared chromatin in RIPA buffer was then added to the antibody-conjugated beads and incubated on a rotator overnight at 4 °C. After incubation, beads were washed 5 times with an LiCl wash buffer (100 mM Tris, pH 7.5, 500 mM LiCl, 1% NP-40, 1% sodium deoxycholate), and remaining ions were removed with a wash in 1 ml TE (10 mM Tris-HCl, pH 7.5, 0.1 mM Na_2-EDTA) at 4 °C. Chromatin and antibodies were eluted from beads by incubation for 1 h at 65 °C in immunoprecipitation elution buffer (1% SDS, 0.1 M $NaHCO_3$) followed by overnight incubation at 65 °C to reverse formaldehyde cross-links. DNA was purified using MinElute DNA purification columns (Qiagen). qRT–PCR using SYBR green Fastmix (Quanta BioSciences) was performed with the CFX96 Real-Time PCR Detection System (Bio-Rad) and the oligonucleotide primers reported in Supplementary Table 3. A total of 100 pg ChIP DNA was loaded into each reaction. The results are expressed as a fold increase of signal at the target locus normalized to signal of a region in the β-actin locus using the $\Delta\Delta C_t$ method.

Kinetic R-loop formation simulations. A first-principles, biophysical simulation of sgRNA invasion of a DNA duplex was performed in MATLAB by modeling the processes as a one-dimensional random walk in a position-dependent potential[29]. This was formulated as a continuous-time Markov chain in MATLAB. The position-dependent potential is determined by the nearest-neighbor-dependent DNA:DNA binding free energies[61], RNA:DNA binding free energies[62] and guide RNA secondary structure free energies that are disrupted or restored as invasion progresses/recedes. Here we have generalized the model to estimate sgRNA residence time at spacers with arbitrary numbers and species of mismatches, and to account for effects of spacer secondary structure on invasion kinetics.

The sgRNA is base paired with the spacer up to spacer site m ($2 \geq m \geq 20$). At each state m, the sgRNA is assumed to be in quasi-equilibrium with the DNA, such that at perfectly matched spacer sites the forward rate (rate of additional guide RNA invasion; m to $m+1$) v_i is estimated using the symmetric approximation to be $\exp(-(\Delta G°(m+1)_{\mathrm{RNA:DNA}} - \Delta G°(m+1)_{\mathrm{DNA:DNA}} - \Delta G°(m+1)_{\mathrm{RNA,SS}})/2RT)$, where R is Boltzmann's constant, T is the temperature (here 37 °C to correspond with the parameter set we used) and the 1/2 corrective term is included to satisfy detailed balance. $\Delta G°(m+1)_{\mathrm{RNA:DNA}}$ is the free energy of the base pairing between the RNA and DNA target at site $m+1$. $\Delta G°(m+1)_{\mathrm{DNA:DNA}}$ is the free energy of the base pairing between the spacer and its complementary DNA strand. $\Delta G°(m+1)_{\mathrm{RNA,SS}}$ is the difference in free energies between the predicted structures of the $20-m-1$ uninvaded nucleotides of the sgRNA at site $m+1$ and the $20-m$ uninvaded nucleotides of the sgRNA at site m. The reverse rate v_r was calculated similarly as $\exp(-(\Delta G°(m-1)_{\mathrm{DNA:DNA}} - \Delta G°(m-1)_{\mathrm{RNA:DNA}} + \Delta G°(m-1)_{\mathrm{RNA,SS}})/2RT)$. At $m=1$, the sgRNA irreversibly falls off the DNA ($m=1$ acts as an absorbing state). RNA secondary structure free energy was calculated using the rnafold function in MATLAB[63,64].

To estimate transition rates from site m in the presence of mismatched nucleotides, the next complementary site n is identified, and $\Delta G°(n)_{\mathrm{MM}}$ is estimated from the difference in free energies between the sgRNA (R_m)–DNA target (P_m) duplex from sites *1* to m and the sgRNA–DNA target duplex from sites *1* to n. These duplex free energies were calculated using the MATLAB rnafold function using the sequence R_m–UUUU–P_m, with a minimum size of the loops (in nucleotides) set to 4. The forward rate was then calculated as $\exp(-(\Delta G°(n)_{\mathrm{MM}} - \Sigma_{(k=m+1,n)} \Delta G°(k)_{\mathrm{DNA:DNA}} - \Delta G°(k)_{\mathrm{RNA,SS}})/2RT)$ and similarly for the reverse.

The forward and reverse rates were calculated and assembled into a 19×19 Q-matrix (Q)[65], and the mean lifetimes L of the sgRNA–spacer interaction was calculated as $L = -\boldsymbol{\alpha}_0 Q^{-1}\mathbf{1}$, where $\mathbf{1}$ is a 19-element column vector with values all 1. $\boldsymbol{\alpha}_0$ is a 19-element row vector containing the fractional population of initial states ($m=2$–20). These experiments were performed for all 16 sgRNAs and 12,181 ChIP-seq hits using the published data sets from Kuscu et al.[41] and Wu et al.[40]. For each sgRNA, $\log(L)$ was correlated (Pearson) with log(ChIP-seq count normalized to on-target site), and these correlations combined using Fisher's method.

Protein purification. Plasmids encoding SpCas9 and SaCas9 were transformed into Rosetta 2 (DE3) competent cells. Clones were used to inoculate 25-ml starter cultures. Starter cultures were grown overnight, spun down and used to inoculate 1-liter cultures. Inoculated 1-liter cultures were grown for 5 h at 25 °C after which the temperature was dropped to 16 °C and expression induced using 0.1 mM isopropylthiogalactoside. Induced cultures were grown for another 12 h at 16 °C. Cells were collected by centrifugation at 4,000*g* and stored at −80 °C for long-term storage. Cell pellets were resuspended in 30 ml lysis buffer (50 mM Tris-HCl, 500 mM NaCl, 10 mM $MgCl_2$, 10% v/v glycerol, 0.2% Triton-1000, 1 mM PMSF). The cell suspension was lysed by sonication at 30% duty for 5 min. The suspension was then centrifuged for 30 min at 12,000*g*. The supernatant was then taken and incubated with Ni-NTA resin (Qiagen) for 30 min under gentle agitation. The resin was then loaded onto a column, washed with wash buffer (35 mM imidizole, 50 mM Tris-HCl, 500 mM NaCl, 10 mM $MgCl_2$, 10% v/v glycerol) and eluted with elution buffer (120 mM imidizole, 50 mM Tris-HCl, 500 mM NaCl, 10 mM $MgCl_2$, 10% v/v glycerol). Ultracel-30k centrifugal filters were then used to exchange solvents to the storage buffer (50 mM Tris-HCl, 500 mM NaCl, 10 mM $MgCl_2$, 10% v/v glycerol). The samples were then aliquoted and frozen at −80 °C.

In vitro digestion. Regions of interest were amplified using PCR from HEK293T gDNA and purified using bead purification (Agencourt AMPure XP Beads, Beckman Coulter). Cas9 and sgRNA were combined and incubated for 10 min at room temperature at a 1:1 molar ratio. The Cas9–sgRNA complex was then combined with DNA at a 10:1 molar excess of RNP in NEB buffer 2.1. The reaction was incubated at 37 °C for 1 h after which Gel Loading Dye, Purple (6×) (NEB catalog no. B7024S) was added. To fully dissociate Cas9–DNA interactions the reaction was heated to 90 °C and cooled. The reaction was then resolved on a 2% agarose gel.

AFM. AFM was performed in air as previously described; see ref. [6] for details. Imaging was performed using a Bruker Nanoscope V Multimode with RTSEP (Bruker) probes (nominal spring constant, 40 N m^{-1}; resonance frequency, 300 kHz). Before experiments, protein and guide RNAs were mixed at 1:1.5 ratio for 10 min in a buffer designed to limit DNA cleavage but not DNA binding (20 mM Tris-HCl (pH 7.5), 100 mM potassium glutamate, 5 mM $CoCl_2$ and 0.4 mM TCEP)[66]. SpCas9 and SaCas9 proteins were purified as described above, AsCas12a was purchased from IDT, and all sgRNAs/crRNAs were purchased from Synthego. Protein and DNA were mixed in a solution of working buffer for at least 10 min at room temperature, deposited for 8 s on freshly cleaved mica (Ted Pella, Inc.) that had been treated with 3-aminopropylsiloxane as previously described[67], rinsed with ultra-pure (>17 MΩ) water and dried in air. Proteins were centrifuged briefly before incubation with DNA. At least three preparations for each experimental condition were imaged and analyzed. Images were acquired with pixel resolution of 1,024 × 1,024 over 2.75-µm^2 areas or 2,048 × 2,048 over 5.5-µm^2 areas at 1.5 lines per second for each sample. Image analysis to determine the distribution of binding sites along the DNA was performed as described previously[29]. Apparent dissociation constants of CRISPR proteins were determined using the method pioneered by Yang et al.[68], adapted as previously described[29]. Consensus structures of images of CRISPR proteins were determined by performing a reference-free alignment as previously described[5].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequencing data are available through the National Center for Biotechnology Information Sequence Read Archive (SRA) database (PRJNA524383), including all deep sequencing, 5′ RACE RNA-seq and CIRCLE-seq files. All other relevant raw data are available from the corresponding author on reasonable request.

Code availability

Custom scripts used to analyze 5′ RACE experiments and conduct kinetic modeling are available upon reasonable request.

References

57. Gruber, A. R. et al. The vienna RNA websuite. *Nucleic Acids Res.* **36**, W70–W74 (2008).
58. Guschin, D. Y. et al. A rapid and general assay for monitoring endogenous gene modification. *Methods Mol. Biol.* **649**, 247–256 (2010).
59. Pinello, L. et al. Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat. Biotechnol.* **34**, 695–697 (2016).
60. Lazzarotto, C. R. et al. Defining CRISPR–Cas9 genome-wide nuclease activities with CIRCLE-seq. *Nat. Protoc.* **13**, 2615–2642 (2018).
61. SantaLucia, J., Allawi, H. T. & Seneviratne, P. A. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **35**, 3555–3562 (1996).
62. Sugimoto, N. et al. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry* **34**, 11211–11216 (1995).
63. Wuchty, S. et al. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49**, 145–165 (1999).
64. Mathews, D. H. et al. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940 (1999).
65. Colquhoun, D. H. & Hawkes, A. G. A Q-matrix cookbook. in *Single-Channel Recording* (eds Sakmann B. & Neher E.) 589–633 (Springer, 2009).
66. Dagdas, Y. S. et al. A conformational checkpoint between DNA binding and cleavage by CRISPR–Cas9. *Sci. Adv.* **3**, eaao0027 (2017).
67. Shlyakhtenko, L. S. et al. Silatrane-based surface chemistry for immobilization of DNA, protein–DNA complexes and other biological materials. *Ultramicroscopy* **97**, 279–287 (2003).
68. Yang, Y. et al. Determination of protein-DNA binding constants and specificities from statistical analyses of single molecules: MutS-DNA interactions. *Nucleic Acids Res.* **33**, 4322–4334 (2005).

nature research

Corresponding author(s): Charles A. Gersbach

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Image Lab 5.2.1,
Data analysis	Image Lab 5.2.1, Python 2.1, CRISPResso, MATLAB, GraphPad Prism 6, R Studio

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Accession codes for sequencing data will be available before publication, all other raw data files available on request

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size calculations were performed before experiments. Three biological replicates were used for all cell-based assays. This size has previously been shown as sufficiently powered to determine statistical differences in mean values of our investigated parameters.
Data exclusions	Some preliminary data are not included for clarity and conciseness.
Replication	Hairpin structure optimization was often performed only once using the Surveyor assay, and the optimized design was characterized more thoroughly afterward. All increases in specificity observed with Surveyor assays were repeated in triplicate and were again observed using deep-sequencing. There were no examples of conflicting observations between these two stages of experimentation.
Randomization	All samples were given a numerical alias and processed in parallel.
Blinding	Investigators were not blinded to sample identity.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	FLAG M2 antibody (F1804, Sigma)
Validation	Antibody was tested for efficient immunoprecipitation of DNA using a qubit for DNA quantification. Quality control of the antibody was not performed beyond what is performed by the manufacturer.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T were obtained from the American Type Culture Collection (ATCC) via the Duke University Cancer Center facilities.
Authentication	Once received, cell lines were not authenticated.
Mycoplasma contamination	Cell lines tested negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.