Physical Biology

PAPER

Complex dependence of CRISPR-Cas9 binding strength on guide RNA spacer lengths

To cite this article: Aset Khakimzhan *et al* 2021 *Phys. Biol.* **18** 056003

Physical Biology

RECEIVED
18 February 2021

REVISED
5 June 2021

ACCEPTED FOR PUBLICATION
8 June 2021

PUBLISHED
2 July 2021

PAPER

Complex dependence of CRISPR-Cas9 binding strength on guide RNA spacer lengths

Aset Khakimzhan[1,*], David Garenne[1], Benjamin Tickman[2], Jason Fontana[2], James Carothers[2,3,4] and Vincent Noireaux[1,*]

[1] School of Physics and Astronomy, University of Minnesota, 115 Union Street SE, Minneapolis, MN 55455, United States of America
[2] Molecular Engineering & Sciences Institute, University of Washington, Seattle, WA, 98195, United States of America
[3] Department of Chemical Engineering, University of Washington, Seattle, WA, 98195, United States of America
[4] Center for Synthetic Biology, University of Washington, Seattle, WA, 98195, United States of America
[*] Authors to whom any correspondence should be addressed.

E-mail: khaki005@umn.edu and noireaux@umn.edu

Keywords: CRISPR-Cas9, CRISPRi, CRISPRa, cell-free transcription–translation, synthetic biology, thermodynamics

Supplementary material for this article is available online

Abstract

It is established that for CRISPR-Cas9 applications guide RNAs with 17–20 bp long spacer sequences are optimal for accurate target binding and cleavage. In this work we perform cell-free CRISPRa (CRISPR activation) and CRISPRi (CRISPR inhibition) experiments to demonstrate the existence of a complex dependence of CRISPR-Cas9 binding as a function of the spacer length and complementarity. Our results show that significantly truncated or mismatched spacer sequences can form stronger guide–target bonds than the conventional 17–20 bp long spacers. To explain this phenomenon, we take into consideration previous structural and single-molecule CRISPR-Cas9 experiments and develop a novel thermodynamic model of CRISPR-Cas9 target recognition.

1. Introduction

CRISPR-Cas9 revolutionized molecular biology by providing an easy-to-use and programmable tool for accurate genome editing [1, 2]. What makes CRISPR-dCas9/Cas9 both accurate and programmable is its use of a guide RNA (gRNA) to bind to a complementary target DNA. While much has been achieved in CRISPR-dCas9/Cas9 biotechnologies [3], the design of guide RNAs still adheres mostly to empirical models and ad hoc rules [4]. One of these rules is that the length of a guide RNA spacer sequence should be between 17 bp and 20 bp to securely bind to the target DNA [5].

From a naïve thermodynamic perspective, the longer spacer sequences result in more RNA:DNA bonds between the Cas9/gRNA complex and the target DNA, and more RNA:DNA bonds result in a more stable bond with the target. This intuition has been confirmed by previous CRISPR-dCas9/Cas9 experiments [6–8] and used for previous CRISPR-dCas9/Cas9 binding models. Using a cell-free transcription–translation system (TXTL), we characterize the binding efficiencies of guide RNAs with significantly truncated or mismatched spacers and demonstrate that under some conditions those guide RNAs are comparable or are better at binding to the target DNA than their conventional 20 bp spacer counterparts. In a series of CRISPRi (CRISPR interference) [9] and CRISPRa (CRISPR activation) [10, 11] experiments, we demonstrate that binding strength of a CRISPR-dCas9/Cas9 system does not decrease monotonically with the length/complementarity of the guide RNA, but instead dips and peaks as a function of length/complementarity.

Since our experimental results are not explained by any of the previous models of CRISPR-Cas9 binding [12, 13], we developed a novel thermodynamic model based on previous experiments about the conformational changes of CRISRP-Cas9 during target interrogation. We built the model using energies and rates obtained from structural experiments [14, 15], single-molecule Forster resonance energy transfer (smFRET) experiments [16–20], magnetic torque measuring experiments [20, 21], atomic force microscopy experiments [22], and molecular dynamics simulations [23–26].

All of the above-mentioned experiments share a common basis: they argue that CRISPR-Cas9 target interrogation occurs in discrete steps controlled by conformational changes that are in turn coordinated with the number of formed RNA:DNA bonds. Interestingly, it is hypothesized that for purposes of increased specificity Cas9 has evolved to its cleavage active conformation to be less stable than the intermediate conformation [27]. That idea is strengthened with rotor bead tracking (RBT) experiments, where even though only a single PAM-distal mismatch is introduced the binding free energy increases as the system progresses into the final DNA unwinding conformation [20]. Therefore, one can imagine that a short spacer sequence will form a strong bond if it can also minimize the probability of being destabilized by conformational changes. This is the core of our model and we demonstrate how balancing the energies and the rates of conformational changes cause the emergence of strongly binding truncated and mismatched spacer sequences.

2. Methods and materials

2.1. Materials

DNA was purchased from Integrated DNA Technologies (Coralville, IA) and Twist Biosciences (San Francisco, CA). Unless otherwise mentioned all the other reagents were purchased from Sigma Aldrich (St. Louis, MO).

2.2. DNA constructs

For silencing experiments in TXTL, the Cas9 or dCas9 enzymes were synthesized constitutively through the endogenous Sp. pCas9 promoter [11]. CRISPRi was achieved using sgRNAs (single guide RNA), expressed from the Anderson *E. coli* promoter J23119 (http://parts.igem.org/Promoters/Catalog/Anderson). The target template for CRISPRi was the plasmid P70a-*degfp* [28, 29]. The reporter protein deGFP (enhanced green fluorescent protein) is a slightly truncated version of eGFP with the same fluorescent properties as eGFP. For CRISPR activation experiments in TXTL, we used the same plasmids as for CRISPRi to express Cas9 or dCas9. The scRNAs (scaffold RNA) gene was expressed from the Anderson *E. coli* promoter J23119 (http://parts.igem.org/Promoters/Catalog/Anderson). The activator gene was expressed constitutively from the Anderson *E. coli* promoter J23107 [30]. The reporter target template for CRISPRa were the plasmids pJF143.J2, pJF143.J3, and pJF143.J4 that contain the Anderson promoter J23117 [30] upstream of the mRFP gene. The sgRNA and scRNAs were expressed using linear templates, whose degradation in TXTL was prevented by adding the Chi6 dsDNA linear template [31, 32]. All the constructs have been sequenced. These plasmids are available on demand.

2.3. TXTL target template sequences

The target and **PAM** site sequences for the CRISPRi experiments:

(a) sg2: GTTGACAATTTTACCTCTGG**CGG**
(b) sg3: AATTTTGTTTAACTTTAAGA**AGG**
(c) sg6: GGTAAAATTGTCAACACGCA**CGG**
(d) sg9: GTCGCCCTCGAACTTCACCT**CGG**
(e) sg15: CCAGGGCACGGGCAGCTTGC**CGG**

The target and **PAM** site sequences for the CRISPRa experiments:

(a) J206: TAGTAGCCGAACACGTCCTC**AGG**
(b) J306: TTGTGTCCAGAACGCTCCGT**AGG**
(c) J406: GAACATCCTTTCACTTCCGG**AGG**

2.4. TXTL reactions

The myTXTL kit (Arbor Biosciences) was used for cell-free expression. TXTL reactions are composed of an *E. coli* lysate, an amino acid mixture, an energy buffer, and the desired DNA templates. This TXTL system has been described previously [28, 29]. All the batch mode TXTL reactions were incubated at 29 °C in either a bench-top incubator, for endpoint measurements, or in plate readers, for kinetic measurements. 29 °C is the optimum temperature of incubation for the myTXTL kit.

2.5. Quantitative measurements of fluorescence in TXTL reactions

Fluorescence from batch mode TXTL reactions was measured using the reporter protein deGFP (25.4 kDa, 1 mg ml^{-1} = 39.4 μM) and mRFP (25.4 kDa, 1 mg ml^{-1} = 39.4 μM). Fluorescence was measured at 5 min intervals using monochromators (deGFP Ex/Em 488/525 nm, mRFP Ex/Em 555/583) on a Biotek Synergy H1 plate reader in Costar polypropylene 96-well, V-bottom plates sealed with a mat. Endpoint reactions were measured after 16 h of incubation. To measure protein concentration (eGFP reporter), a linear calibration curve of fluorescence intensity versus eGFP concentration was generated using purified recombinant eGFP obtained from Cell Biolabs (STA-201), Inc or purified in the lab.

2.6. Estimation of CRISPRa binding fraction

Binding of CRISPR-dCas9-SoxS complex near the promoter site of the reporter gene activates the transcription of the reporter gene mRNA, which in turn induces the synthesis of the reporter protein observed by an increase of the fluorescence intensity of the TXTL reaction. Thus, we estimated that the fraction of bound reporters is proportional to the intensity of TXTL reaction:

$$p_{\text{bound}} \sim I_{\text{TXTL}} \qquad (1)$$

where I_{TXTL} is the fluorescence intensity of the TXTL reaction.

2.7. Estimation of CRISPRi binding fraction

The binding of CRISPR-dCas9 to the promoter or gene interferes with the transcriptional activity of the gene, which in turn decreases the amount of reporter protein synthesized and thus the fluorescence intensity of the TXTL reaction. By comparing the fluorescence intensities of the TXTL CRISPRi experiments using a non-targeting sgRNA (sg-NT) with the fluorescence intensity of TXTL experiments using a targeting sgRNA (sg6 or sg9), we can estimate the fraction of silenced P70a-degfp genes. The formula to estimate the fraction of bound genes is the following:

$$f = 1 - \frac{I_{\mathrm{ON}}}{I_{\mathrm{OFF}}} \qquad (2)$$

and f will be referred to as the fraction coefficient, I_{ON} is the fluorescent intensity of the on-target TXTL reaction, and I_{OFF} is the fluorescent intensity of the off-target TXTL reaction. The fraction coefficient of the non-target control is always $f = 0$.

2.8. Liquid handler systematic bias correction

The liquid handler (Labcyte Echo 550) used for to perform the CRISPRi experiments systematically introduced uneven expression levels on the 96-well plate. The calculation to account for the liquid handler introduced systematic error is presented in appendix A (supplementary file (https://stacks.iop.org/PB/18/056003/mmedia)). The uncorrected data is presented in the appendix in supp. figure 1.

2.9. Computer codes

Scripts are available on demand.

3. Experimental results

3.1. CRISPR-Cas9/dCas9 in TXTL

The TXTL system used in this work enables characterizing CRISPR technologies with an excellent agreement between the observations made *in vitro* and *in vivo* [29, 33]. For this work we used both CRISPRi (CRISPR interference) and CRISPRa (CRISPR activation) systems in TXTL experiments (figure 1). While the results from CRISPRi experiments would be sufficient to make the argument, we supplement them with CRISPRa experiments to verify that the results are general and not an artefact of either method. In addition, replacing the dCas9 with a Cas9 in a CRISPRa length/complementarity experiment is used to measure the guide RNA length/complementarity needed for Cas9 cleavage. The efficiency of CRISPR systems in TXTL is estimated by measuring the expression of the CRISPR targeted genes. The dCas9 necessary for CRISPR experiments was either expressed in a regular TXTL reaction or was available at a fixed concentration of approximately 20–50 nM in the dCas9 pre-expressed TXTL system [33]. Pre-synthesized dCas9 in the TXTL system is convenient because, first, it simplifies the experiment by maintaining the total concentration of dCas9 constant and, second, it helps to conserve the reaction resources (ATP, amino acids). Since we are mostly interested in the binding efficiency of CRISPR-Cas9/dCas9 systems most of the experiments involve only the use of dCas9. However, as mentioned above, when we were interested in determining the necessary length/complementarity of the guide RNA needed make Cas9 cleavage active we also performed experiments with expressing Cas9 enzymes.

In the CRISPRi experiments with pre-synthesized dCas9 we added a deGFP expressing reporter target plasmid (P70a-degfp) and sgRNA expressing linear DNA (sg2, sg3, sg6, sg9, sg15, and sg-NT) (figure 1(a)). When a regular extract (without dCas9 pre-synthesized) was used we also added a dCas9 expressing plasmid (pCD017). For the main text we either truncated or introduced mismatches into the spacers of the sg6 sgRNA to test the effects that such mutations have on the efficiency of CRISPRi systems in TXTL. And for the supplementary section we truncated the rest of the spacers (sg2, sg3, sg9, and sg15).

As opposed to sgRNAs used in CRISPRi, CRISPRa systems utilize scRNAs (scaffold RNA) for target search, which in addition to the regular single guide RNA structure also contains an MS2 RNA hairpin [30, 34] that can bind to an MS2 coat protein [35] (MCP). By expressing the activator protein SoxS fused to MCP we can use the binding of the CRISPRa complex to the DNA to localize the SoxS activator [36] near the promoter region. For the CRISPRa experiments with pre-synthesized dCas9 we added an mRFP expressing reporter target plasmid (pJF143.J3.117), the activator MCP-SoxS (pBT005), and the scRNA DNA (linear scRNA expressing segment of pJF144.206.x, pJF144.306x, pJF144.406x) in TXTL reactions, while for the regular TXTL system experiments we also expressed dCas9 or Cas9 (figure 1(b)). In the CRISPRa experiments the target sequence of the guide RNA matched with a region −81 bp from the transcription start site (TSS) [34]. The efficiency and the binding fraction of CRISPRa experiments is estimated using equation (2) from the methods section.

For the mismatch experiments, the spacer–target mismatches are always introduced consecutively from the PAM distal side. If a spacer has a maximum length of 20 bp, but only has 11 PAM-proximal matches, and 9 PAM distal nucleotides are not complementary to the target, then we refer to it as an 11–20 mm spacer/guide RNA (figure 1(c)). More generally a spacer with N PAM-proximal bonds would be referred to as an N-20 mm spacer/guide RNA. For truncation experiments, we exclusively remove the PAM-distal nucleotides of the spacer.

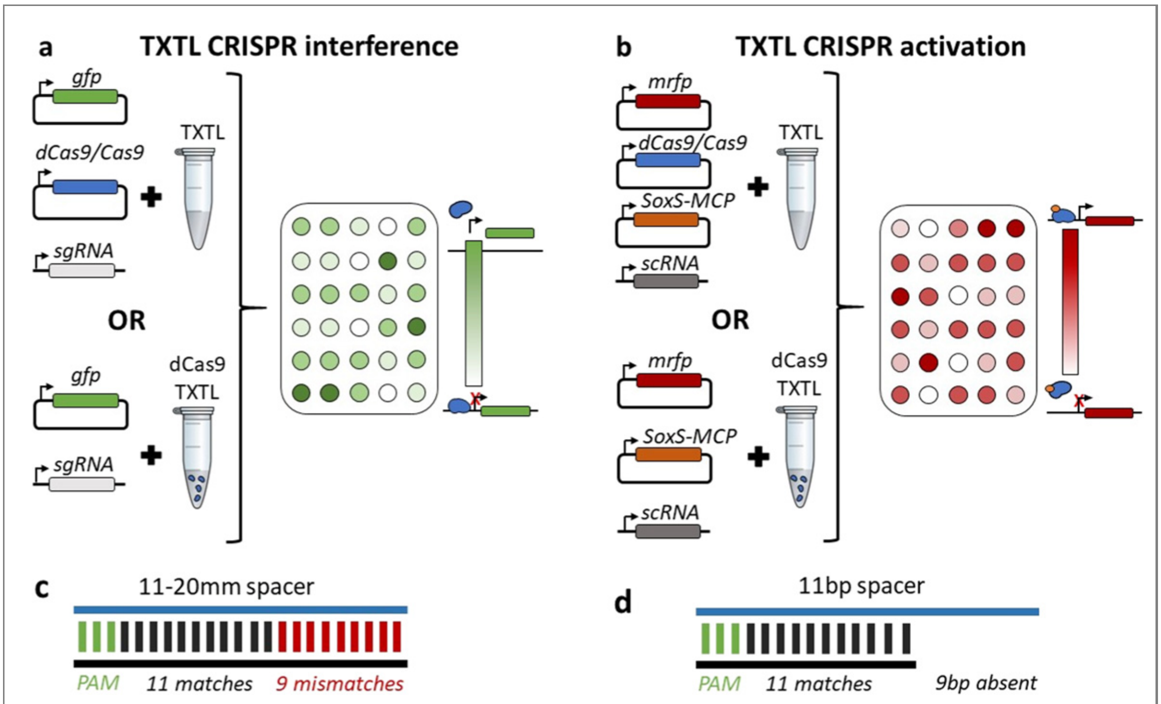

Figure 1. Schematic of the TXTL experiment. (a) The TXTL CRISPR interference (CRISPRi) experiment is composed of two plasmids to express the genes *degfp* and *dCas9*, and a linear DNA to express the *sgRNA*. The spacer sequence of the sgRNA matches with a sequence on the promoter (P70a) of the *degfp* gene (P70a-*degfp* plasmid). The TXTL reactions are incubated at 29 °C on a well plate. The measured fluorescence signals are analysed quantitatively using either kinetics or endpoints measurements. The experiment can also be performed with a TXTL system in which dCas9 is pre-synthesized (dCas9 TXTL), thus simplifying the reaction given that the concentration of dCas9 is constant and dCas9 does not need to be dynamically synthesized. All other parts of the experiment remain the same. (b) The TXTL CRISPR activation (CRISPRa) is composed of three plasmids and one linear DNA part: an mRFP target plasmid, a dCas9 (pCD017) or Cas9 (pCas9) plasmid, a SoxS-MCP plasmid (pBT005), and a linear DNA to express the *scRNA*. The scRNA spacer matches with a sequence 70 bp upstream the ribosome binding site (RBS) of the mRFP gene. The TXTL reactions are incubated at 29 °C on a well plate. The fluorescence signals are analysed quantitatively using either kinetics or endpoints measurements. As in CRISPRi experiments, we can also simplify the TXTL reaction by using TXTL system with pre-synthesized dCas9 (dCas9 TXTL), thus removing the need to express *dCas9* and maintaining its concentration constant. (c) A picture to demonstrate the naming convention of mismatched spacers. In the picture the spacer has 11 PAM-proximal matches out of 20 total, and is thus named an 11–20 mm spacer. If a spacer had *N* PAM-proximal matches, then it would an *N*-20 mm spacer. (d) A picture demonstrating the naming convention for truncater spacers. If a spacer has only 11 of the PAM-proximal matches, then we refer to it as 11 bp spacer. If the spacer was only *N* bp long, then it would be referred to as an *N* bp spacer.

If the spacer is only 11 bp long and thus can at most form 11 RNA:DNA bonds, we refer to it as 11 bp spacer/guide RNA (figure 1(d)). As in the case of mismatches, a truncated spacer with *N* PAM-proximal nucleotides would be referred to as an *N* bp spacer/guide RNA.

3.2. CRISPRa mismatch experiments

In the mismatch experiments we changed the PAM-distal nucleotides of the fully complementary guide RNA to test the effect PAM-distal mismatches have on the binding efficiency of a CRISPRa system. For the dCas9 CRISPRa mismatch experiments we used the pre-expressed *dCas9* TXTL system. We added 2 nM of the mRFP reporter plasmid (pJF143.J3.117) and added 2.5 nM of the MCP-SoxS activator plasmid (pBT005) to the TXTL reaction. We varied the concentration of the scRNA expressing DNA (pJF144.306.xmm) between 1 nM, 2 nM, and 4 nM. We observe that the activation decreases for the 15–20 mm scRNA (figures 2(a)–(c)), thus resulting in two distinct local optimal spacer complementarities (12–20 mm and fully complementary). We do not observe any clear dependence between the concentration of the scRNA expressing DNA added and the activation levels of the target plasmid. That is probably because 1 nM of the scRNA DNA provided a saturating concentration of the scRNA. However, independent of the concentration, we notice that the decrease in the activation level is not monotonic with the decrease of the number of PAM-proximal matches.

To test how many matched base pairs are required for Cas9 to enter its catalytically active state we performed the same experiment, but in a regular TXTL system with 1 nM of the Cas9 plasmid (pCas9 [11]). The concentrations of scRNAs are varied between 0.25 nM, 1 nM, and 2 nM (figures 2(d)–(f)). We observed that starting from the 16–20 mm guides the activation is small in comparison to the peaks between 10–20 mm and 13–20 mm guides, which suggests that 16 RNA:DNA bonds are sufficient for the Cas9 enzymes to become catalytically active [15] and instead of activating target plasmid, the Cas9-CRISPR complexes cut the target. Along with dCas9-CRISPR and Cas9-CRISPR experiment we performed experi-

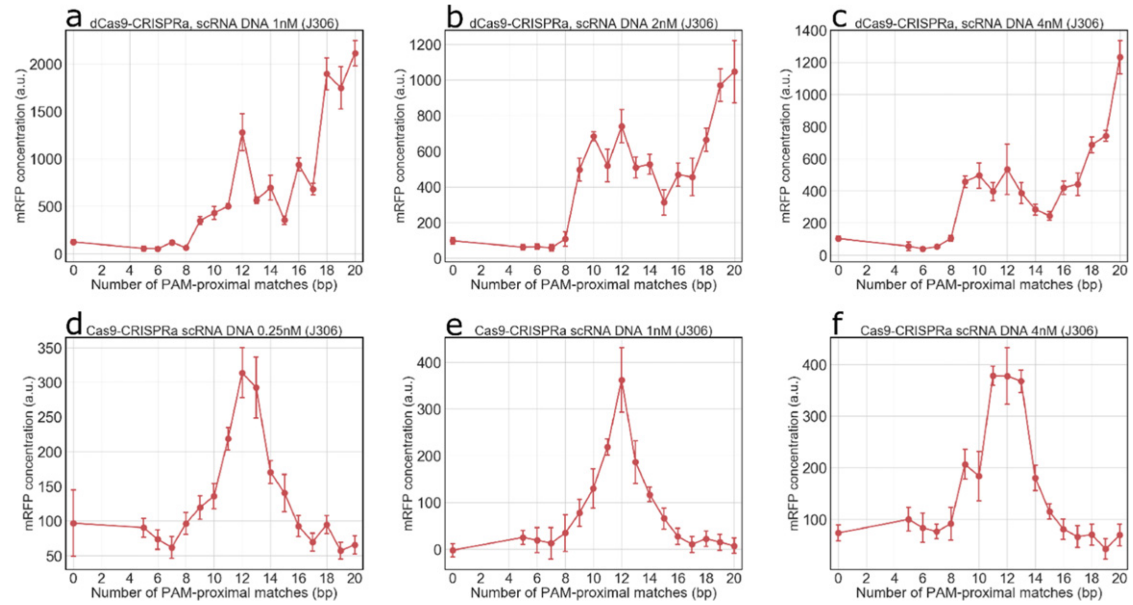

Figure 2. CRISPRa mismatch experiments in TXTL. (a)–(c) are results of pre-synthesized dCas9 CRISPRa mismatch experiments with 1 nM, 2 nM, and 4 nM of scRNA DNA, respectively. (d)–(f) repeat the conditions of (a)–(c) but with regular TXTL reaction that has 1 nM of a Cas9 plasmid. The observed decline in the expression for the 16–20 mm spacer is caused by catalytic activation of the CRISPR-Cas9 complex, which in turn cuts the reporter target plasmid instead of activating it.

ments in which we express Cas9 in the pre-expressed dCas9 TXTL and we observe similar results to the Cas9-CRISPR experiments (supp. figure 2).

3.3. CRISPRa truncation experiments

First, we performed CRISPRa truncation experiments with a pre-synthesized dCas9 TXTL. We maintained the concentration of the activator plasmid at 3 nM and the concentration of the mRFP expressing target plasmid at 2.5 nM. The concentration of the scRNA expressing linear DNA was fixed at 0.25 nM, 0.5 nM, and 1 nM (figures 3(a)–(c)). For each of the concentrations we observe drops in activation for the 11 bp and 15–16 bp spacer lengths. As we increased the concentration of the scRNA in the TXTL reaction, the relative heights of the peaks at non-standard target sequence lengths have also increased (figure 3(d)). We performed similar experiments for two more scRNA spacer sequences, J206 and J406, that targeted plasmids J2 and J4 respectively (supp. figure 3(a) and (b)). The targets were also located −81 bp away from the TSS. As in the experiment with the J306 spacer and the J3 reporter, we observed that some shorter spacer lengths can activate the reporter plasmid more efficiently than the longer spacer lengths.

The second series of experiments performed with CRISPRa truncations used the catalytically active Cas9 enzyme instead of the catalytically inactive mutant dCas9. As in the experiments described above, the concentrations of activator plasmid and mRFP reporter plasmid were maintained at 3 nM and 2.5 nM respectively. The TXTL extract with pre-synthesized dCas9 was replaced with regular TXTL extract and we added 2 nM of a Cas9 encoding plasmid to the reaction [11]. Instead of observing 3 peaks as with the catalytically inactive CRISPRa, we only observe 2, with expression drops for the same lengths of 11 bp and 15–16 bp (figures 3(e) and (f)). As in the CRISPRa mismatch experiments, when the spacer length is longer than 16 bp it can form 16 RNA:DNA bonds, which in turn allow the Cas9 to become catalytically active. As in the CRISPRa mismatch experiments, the spacer lengths at which we observed activation efficiency drops for the dCas9 experiments (figure 3(d)) is the same length for which we observe the transition to a cleavage active Cas9 enzyme in the Cas9 experiments (figures 3(e) and (f)). We performed a similar experiment (but with less different spacer lengths) for the J206 spacer and the J2 target and confirm that the effect can be observed for it as well (supp. figure 3(c)).

3.4. CRISPRi mismatch experiments

The goal of the CRISPRi mismatch experiments was to observe how decreasing the number of matching PAM-proximal base pairs affects the silencing efficiency of CRISPR-dCas9 in TXTL. For the CRISPRi mismatch experiments we used a pre-synthesized dCas9 TXTL and added mismatched sgRNA expressing DNA and a reporter target DNA plasmid (P70a-*degfp*). The concentration of the target plasmid (P70a-*degfp*) was varied between 0.33 nM, 1 nM, and 3 nM, while the concentrations of the sgRNA expressing linear DNA were varied in the range of 0.25–4 nM. In the experiments in which the concentration of the reporter target DNA (P70a-*degfp*) was maintained at 1 nM we observe two target sequence match

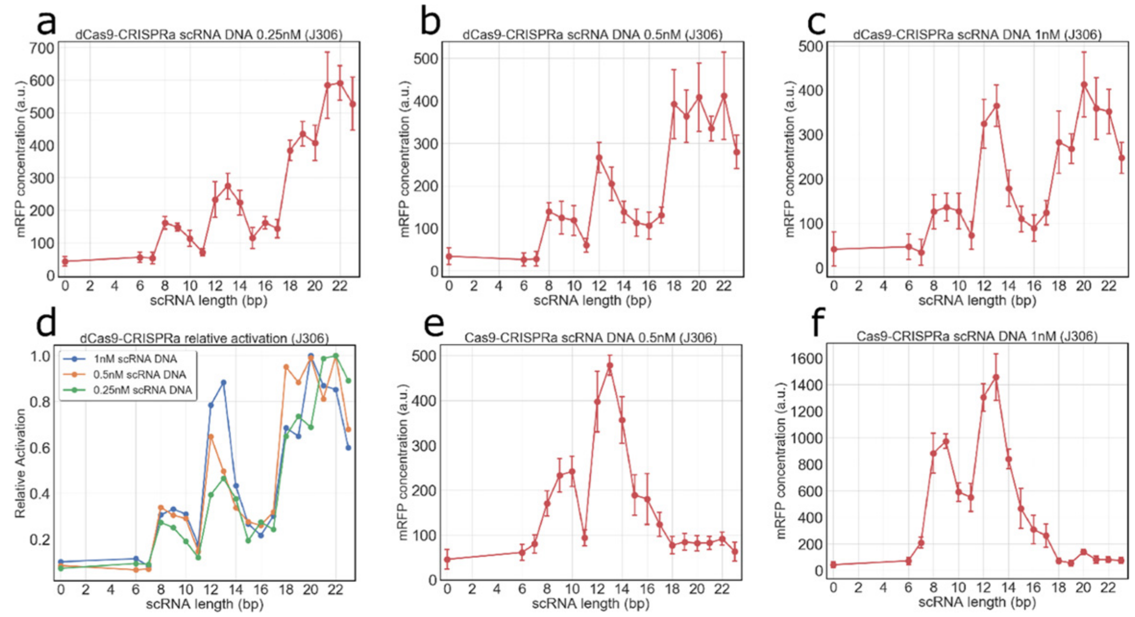

Figure 3. CRISPRa truncation experiments in TXTL. (a)–(c) are the activation levels as a function of the scRNA spacer length (scRNA length) for 0.25 nM, 0.5 nM, and 1 nM of the scRNA DNA added to the reactions respectively. We observe that besides an optimal length of 18–20 bp, truncation lengths of 12–13 bp and 8–10 bp are also capable of activating expression. (d) The relative activation levels demonstrate that as we increase the concentration of the added scRNA DNA the 12–13 bp and 8–10 bp peaks also increase, indicating that with the increase of the concentration of dCas9-CRISPRa complexes we can saturate their expression to the peak values of the 18–20 bp peak. (e) and (f) are the activation levels in regular TXTL reaction, where instead of the dCas9 enzyme we express a Cas9 enzyme. As in (c) and (b) we observe the 12–13 bp and 8–10 bp peaks.

lengths for which we observe fraction coefficient drops (figures 4(a) and (b)). The first drop is observed for the 12–20 mm guide, while the second drop occurs for an 18–20 mm guide. We notice that as we change the concentration of the sgRNA expressing DNA the drops can become prominent. As we increase the concentration of the target plasmid to 3 nM, we also increase the number of target sequences that need to be silenced, therefore the fraction coefficient drop for 18–20 mm guide is more prominent (figure 4(c)). When we lowered the concentration of the reporter target plasmid to 0.33 nM, we can see a decrease in the fraction coefficient at the 12–20 mm guide, which disappears as we increase the concentration of the added sgRNA expressing linear DNA (figure 4(d)). As in the CRISPRa mismatch and CRISPRa truncation experiments we notice multiple optimal binding lengths: one that corresponds to the lengths at which DNA unwinding occurs [20], and the other corresponds to the length at which CRISPR-Cas9 becomes cleavage active [15]. The data in figures 4(a)–(c) have been corrected as described in appendix A, while figure 4(d) was not corrected. The uncorrected data is presented in supp. figure 1.

Since the first dip for the 12–20 mm guide (figures 4(b) and (d)) is not as noticeable as the dip for the 18–20 mm guide (figures 4(a) and (c)), we additionally performed a significance measurement comparing the 11–20 mm and the 12–20 mm silencing strength. To have more data for the significance experiments we performed more CRISPRi mismatch experiments (figures 4(a), (b) and (d), and supp. figures 4(a)–(c)). We then assembled the measured fraction coefficients from figures 4(a), (b) and (d), and supp. figures 4(a)–(c) (56 measurements for 11–20 mm and 12–20 mm each) and performed 10 000 000 simulations calculating the difference for randomly split groups and plotting the distribution (supp. figure 4(d)). We determined that 0.6811% of the simulations had a larger difference between the means that the experimentally measured difference, thus resulting in a significance of $p = 0.006\,811$.

3.5. CRISPRi truncation experiments

The goal of the CRISPRi truncation experiments was to observe how shortening the target sequence on the guide RNA affects the silencing efficiency of CRISPR-dCas9. The truncated sgRNA were designed by removing PAM-distal base pairs of the spacer sequence. The CRISPRi truncation experiments were performed with pre-synthesized dCas9 TXTL and the concentration of the reporter target plasmid was maintained (P70a-*degfp*) at 0.5 nM. The concentration of the sgRNA expressing DNA was varied between 5 pM and 8 nM and the length of the spacer was varied between 5–32 bp. As we decrease the concentration of the sgRNA expressing DNA from 150 pM to 5 pM we observe the emergence of strong binding truncations that are far from the classical 20 bp length (figure 5(a)). Surprisingly, the strongest binding sgRNA truncation was the 10 bp truncation. As in the CRISPRa truncation experiments the strength

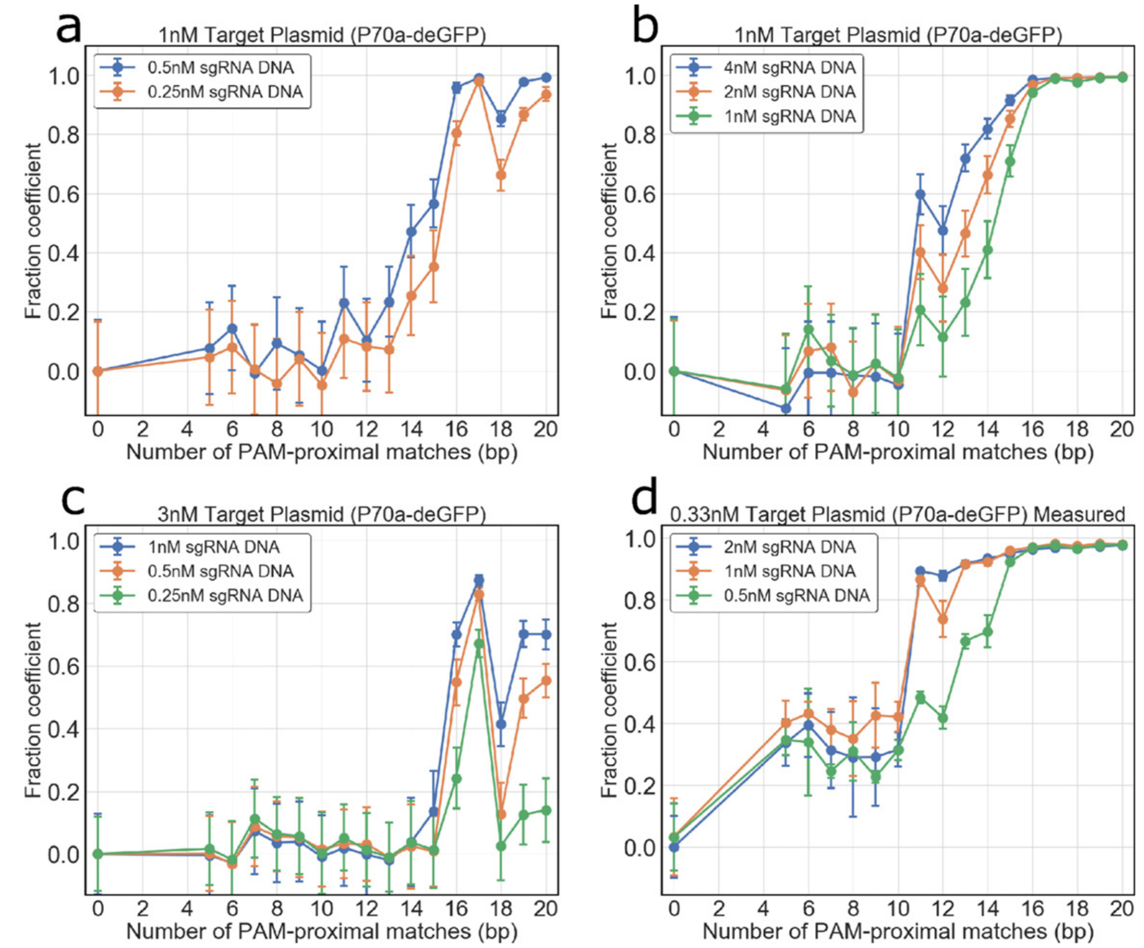

Figure 4. CRISPRi mismatch experiments in TXTL. (a) The concentration of the reporter plasmid P70a-*degfp* is maintained at 1 nM, while the concentration of the sgRNA DNA is varied between 0.25 nM and 0.5 nM. We observed a drop in binding for the 18–20 mm spacer. (b) is continuation of the (a), with sgRNA DNA concentrations of 1 nM, 2 nM, and 4 nM. As we increased the concentrations the drop in the fraction coefficient at the 18–20 mm sgRNA is no longer visible, while the binding drop for the 12–20 mm sgRNA emerges. (c) The concentration of the target sequence plasmid is maintained at 3 nM, while the sgRNA DNA concentrations are varied between 0.25 nM, 0.5 nM, and 1 nM. As we increased the concentration of target sequence DNA from 1 nM to 3 nM, the drop at the 18–20mm guide becomes more prominent. (d) For the experiments where the concentration of the target sequence plasmid is maintained at 0.33 nM the 18–20 mm drop is unnoticeable. The 12–20 mm guide drop can be controlled by varying the concentration of the sgRNA expressing DNA added to the reaction.

of binding of a CRISPR-complex does not decrease monotonically as we truncate the target sequence. We performed the same experiment with less truncations for more sgRNA sequences (sg2, sg3, sg9, and sg15) and we observe the anomalous binding patterns for those targets as well (supp. figure 5).

While it takes approximately 1 h for the concentration of an expressed RNA to reach a steady state in a TXTL reaction [37], we approximate for simplicity that the concentration of sgRNAs in the TXTL reaction is constant and is proportional to the concentration of sgRNA DNA added to the reaction. The second approximation we make, is disregarding the complex formation kinetics of CRISPR-dCas9, [Cr], and simply calculating its equilibrium value using formula (B3) described in appendix B (supplementary file). However, after approximating the concentration of CRISPR-dCas9 complexes, we noticed that the binding of CRISPR-dCas9 for some lengths was highly sensitive. To measure the sensitivity of CRISPR-dCas9 binding we fit the concentration of CRISPR-dCas9 complexes to the fraction coefficient, f, with the Hill equation:

$$f = \frac{[\mathrm{Cr}]^{n_{\mathrm{Hill}}}}{[\mathrm{Cr}]^{n_{\mathrm{Hill}}} + {K_{\mathrm{D}}}^{n_{\mathrm{Hill}}}}. \qquad (3)$$

We used equations (3) and (4) to find the apparent K_{M}, K_{D}, n^{Hill}, and α values (figure 5(b)), while keeping [dC] fixed at 50 nM. The fit showed little dependency on K_{M} if the value of K_{M} was less than 2 nM (supp. figure 6). This result agrees with *in vitro* kinetic studies that demonstrate that K_{M} values are on the order of 0.1–1 nM even for truncated guides [38, 39]. We infer that in TXTL experiments the difference in formation of CRISPR complexes between sgRNA with different target sequence truncations is negligible, since the concentration of dCas9 and sgRNAs is significantly larger than the K_{M} value. The steady state amount of sgRNAs per single sgRNA DNA was found to be $\alpha = 88.4$, which is a reasonable value considering that *degfp* mRNA has an

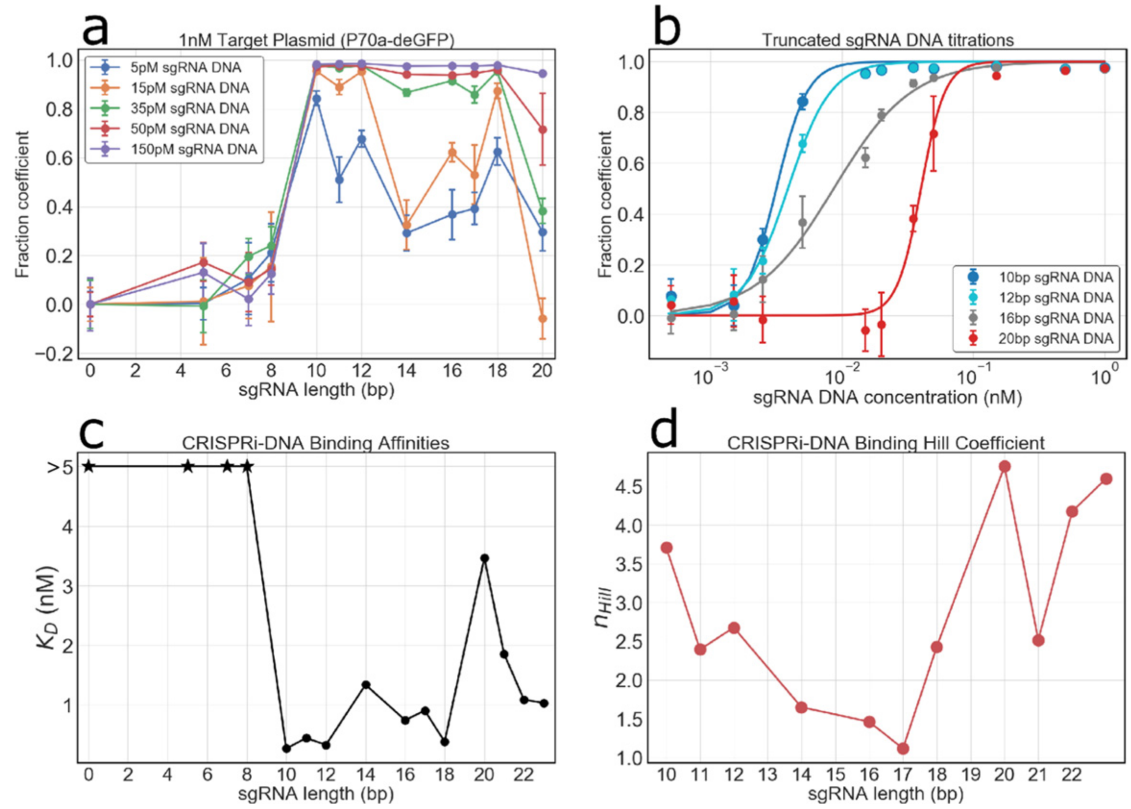

Figure 5. CRISPRi truncation experiments in TXTL. (a) This plot demonstrates the fraction coefficient as a function of the sgRNA spacer length (sgRNA length). The zero in the x-axis is the fraction coefficient of the off-target sgRNA, which is always $f = 0$ from equation [1]. (b) A selection of the binding curves of some sgRNA spacer lengths. All of the curves were obtained by fitting equations [2, 3] to the data. (c) and (d) respectively show the binding coefficients, K_d, and Hill coefficients, n_{Hill}, as functions of the sgRNA spacer length. The stars indicate that the binding affinities of those truncation lengths are significantly larger than 5 nM.

$\alpha_{\text{degfp}} \sim 30$, while the *degfp* gene is 3 times longer than an sgRNA gene [37]. The K_D values did not increase monotonically with the decrease of the length of the target sequence, showing similar conclusions to the mismatch experiments (figure 5(c)). Surprisingly, different truncations of the sgRNA target sequence also exhibited noticeably different levels of modularity in the fits (figure 5(d)). A previous work studying the engineering of CRISPR based circuit in *E. coli* has reported the development of a bistable switch using CRISPRi [40]. They hypothesized that the unspecific binding of dCas9 served as a sufficient condition for bistability, as opposed to possible modularity of CRISPRi. It is possible that the cooperativity observed in the TXTL CRISPRi truncation experiment is an artefact of unspecific binding as well.

4. CRISPR-Cas9 binding model

4.1. Evidence from previous experiments

Our model of CRISPR-Cas9 binding focuses on estimating the unbinding rate of CRISPR-Cas9 enzymes from the target DNA as a function of spacer length and matching. However, estimating the off-rate of a CRISPR-Cas9 system is not trivial since CRISPR-Cas9 undergoes multiple conformational changes during target interrogation [41]. Therefore, besides simply calculating the binding energy of the formed R-loop, it is also necessary to account for the effects conformational changes have on the stability of that R-loop.

Structural studies of CRISPR-Cas9 demonstrated that both the unwinding of the target DNA and the subsequent formation of an RNA–DNA helix induce changes in the structure of the Cas9 enzyme [14, 15]. The conformational changes position the nucleating domains of Cas9 (HNH and RuvC) near the cutting locations of the target DNA strands [42]. As in the TXTL experiments, these studies show that the catalytic activity of Cas9 is present only for spacers that can form PAM-distal RNA:DNA bonds. These experimental results have been confirmed and explained with MD simulations of CRISPR-Cas9 systems [23–26].

Multiple smFRET experiments demonstrated the existence of intermediate conformations in which CRISPR-Cas9 has formed a stable bond with the target sequence, but the Cas9 enzyme is not in its catalytically active conformation [16–19]. This intermediate step has been hypothesized as a checkpoint conformation that improves the specificity during target recognition. For a fully matched guide RNA the dwell times in the intermediate conformations range on the scales of 0.01 s to 1 s until the target DNA is cut [19], while for a mismatched or truncated guide RNA the system can be stably bound in the intermediate conformation for minutes [16]. In the qualitative models of CRISPR-Cas9 target interrogation unwinding of DNA is hypothesized to act as a proofreading mechanism. Studies reporting on the probability of off-target effects in stretched DNA strengthen that hypothesis [43].

In agreement with the smFRET experiments, a recent RBT study demonstrated that the target recognition process for CRISPR-Cas9 occurs in discrete steps [20]. The guide RNAs used in the work demonstrated that Cas9 unwinds the first ten PAM proximal nucleotides, forms a 10 RNA:DNA bond R-loop, and only then unwinds the rest of the target. The rate of unwinding the 10–20 base pairs (the 10 PAM-distal base pairs) varied when changing the applied torque. For PAM-distal mismatched guide RNAs the stability of the 'open' state decreased in comparison to the 'intermediate' state even though it is likely that more RNA:DNA bonds are formed in the more progressed states despite the mismatch. A similar clue was observed in the CRISPRa TXTL experiments. The binding efficiency decreased for the same truncations and matching lengths at which CRISPR-Cas9 underwent a conformational change to become catalytically active (figures 2 and 3). Therefore, for our model we assume that the net binding energy offered by a single RNA:DNA bond changes depending on the conformation. This difference in the net binding energy could be a difference in the mechanical strain of each CRISPR-Cas9 conformation.

4.2. Model of CRISPR-Cas9 binding

By incorporating the information from the aforementioned experiments, we construct a model of CRISPR-Cas9 binding (figure 6(a)). The model proposes that the system can inhabit 3 different 1D energy landscapes: seed (S), intermediate (I), and active (A). The landscapes (S, I, and A) represent the different states/conformations of CRISPR-Cas9 during target interrogation. As in single-molecule experiments, the conformation which the system inhabits is strongly dependent on the number of RNA:DNA bonds formed. The *x*-axis in the figure represents the number of RNA:DNA bonds formed between the spacer and the target DNA and will be referred to as sites. A similar model relying on 1D energy landscapes with transitions between conformations has been proposed by Eslami-Mossallam *et al* [44] in parallel to our work. Their model was applied to the prediction of off-target effects in CRISPR-Cas9 genome editing and outperforms previous off-target prediction models [45, 46]. An important distinction between the two approaches is that their model considers conformational changes as transitions on the energy landscape, while we consider conformational changes as transitions between energy landscapes. From a physical standpoint the conformational changes should also correspond to some transition within the energy landscape of the system, therefore our model can be considered as a 2D energy landscape: one dimension for number of RNA:DNA bonds formed and the other dimension for the conformation of the CRISPR-Cas9 complex.

For each matching RNA:DNA bond formed the binding free energy decreases by E_C. Physically $E_C$corresponds to the energy acquired by replacing a matching DNA:DNA bond with a matching RNA:DNA bond, and the value of E_C can be approximated from meltingexperiments [47–49]. We are also interested in the energy the system acquires when encountering an RNA:DNA mismatch/absence and we refer to the parameter as E_{MM}. For a mismatch E_{MM} corresponds to the energy acquired by replacing a matching DNA:DNA bond with a mismatched RNA:DNA bond and can be approximated with melting experiments [50]. For an absence E_{MM} can be modelled as a large increase in binding free energy, since there is no RNA:DNA bond to be formed and thus the system has reached its ultimate state. As in the experiments, all the mismatches are introduced serially from the PAM-distal segment of the target.

Since the binding energy of the system can decrease as more RNA:DNA bonds are formed, we need to form an assumption based on previous experiments that account for these binding energy decreases. Our model does that by assuming that the average binding free energy provided by the addition of an RNA:DNA bond decreases in the more progressed conformations/landscapes of target recognition. This is shown in the schematic of the model in which the slope of the more progressed states is flatter than the slope of the preceding states (figure 6(a)). The decrease in the average contribution of RNA:DNA bonds are scaled with factors S_1 and S_2, thus making the average RNA:DNA bond S_1E_C in the intermediate state and $S_1S_2E_C$ in the active state. The energies of an average RNA:DNA mismatch/absence, E_{MM}, are also scaled for each state. The average mismatch energy is also scaled in the other conformations with S_1E_{MM} in the intermediate state and $S_1S_2E_{MM}$ in the active state (figure 6(a)). The seed and the intermediate states have maximal lengths L_S and L_I, respectively. These lengths determine the maximum number of RNA:DNA bonds that can form for each CRISPR-Cas9 conformation. Both lengths are measured from the start of the sequence and experimentally have been observed to be 9–11 bp for L_S [20] and 16–19 bp for L_I [15]. The total length of the guide RNA spacer would determine the ultimate site of energy landscape. For the model we used a total length of 22 bp. Additionally, the energy scaling assumption requires that a CRISPR-dCas9/Cas9 system can unbind from the target DNA from both

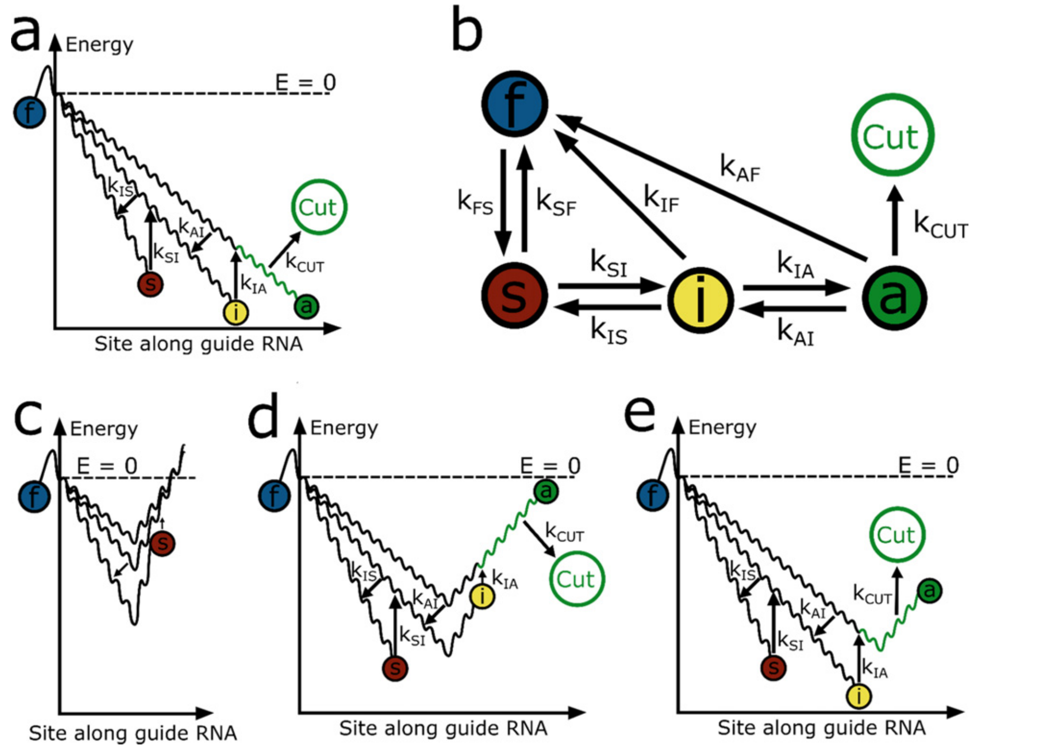

Figure 6. CRISPR-Cas9 model. (a) The model energy landscapes of CRISPR-Cas9 binding for a fully complementary target. The system can inhabit 3 energy landscapes of different lengths: seed (S)—red, intermediate (I)—yellow, and active (A)—green. The system acquires a binding energy E_C, S_1E_C, and $S_1S_2E_C$ for each formed RNA:DNA bond in landscapes S, I, and A respectively, which can be seed by the different slopes of the energy landscapes. The minimas in the landscapes correspond to sites—substates of landscapes which stand for the number of PAM-proximal RNA:DNA bonds formed. The forward transitions (k_{SI}, k_{IA}) between landscapes occur from the final site of each landscape, while the inverse transitions (k_{SI}, k_{AI}) are independent of the landscape. CRISPR-Cas9 can cut the target DNA in the active sites (painted green) of landscape A with a rate k_{CUT}. The unbinding from target DNA can occurs at a rates k_{SF}, k_{IF}, and k_{AF} for landscapes S, I, and A respectively. The unbinding rates for each landscape are calculated by multiplying the base off-rate k_{OFF} by the probability of inhabiting site 0 (0 RNA:DNA bonds formed) in that specific landsacpe. When unbound, the CRISPR-Cas9 occupies the free state labelled as f. (b) The differential equation schematic of the CRISPR-Cas9 binding with all the rates labelled. (c)–(e) are the landscapes with lengths/complementarities less than L_S, L_I, and 22 bp respectively. It can be noticed that for truncated/mismatched spacers the landscape acquires a positive absence/mismatch energy E_{MM} for each absent/mismatched RNA:DNA bond in the landscape S, and S_1E_{MM}, $S_1S_2E_{MM}$ for landscapes I, A respectively. This in turn reduced the probability of inhabiting the critical sites that control the rates of transitions and cutting, thus changing the behavior of the system.

the intermediate and active conformations without having to revert to the seed conformation. As of now there is no experimental evidence that conclusively supports or denies this possibility and thus this assumption is yet to be tested.

For this model we assume that the timescale of RNA:DNA bond formation [51] are significantly smaller than the timescales of CRISPR-Cas9 target search, the timescales of target DNA unwinding [52], and the timescale of Cas9 conformational changes [16–19]. This approximation is only performed to simplify the model and avoid additional rate parameters for both the formation and dissociation of both matched and mismatched RNA:DNA bonds. With this assumption we can approximate that the distribution of R-loop sizes is near equilibrium and thus the probability of the system being in site n in landscape X is:

$$p_{n,\mathrm{X}} = \frac{\exp\left(-\beta E_{n,\mathrm{X}}\right)}{z_x}, \tag{4}$$

where z_x is the partition function of the energy landscape X and $E_{n,\mathrm{X}}$ is the binding energy at site n [53]. The partition function does not account for degeneracies of sites in the energy landscape. The partition function is therefore calculated by adding up the exponentials of the energies of each site in the respective landscape, which 0 RNA:DNA (site 0) bonds representing the $0k_bT$ state.

In the experiments, when zero RNA:DNA bonds are formed between the target and the guide RNA, it is possible that the PAM is still bound. In our model we account for PAM binding by estimating a k_{OFF}. Unbinding of CRISPR-Cas9 from the target DNA in our model can occur in all landscapes (S, I, A) and is proportional to the probability of occupying site 0 for the respective landscape. In microscopy experiments it has been observed that the average dwell time for a CRISPR-Cas9 with a mismatched spacer is on the order of 10–100 ms [54, 55]. We use that value and approximate the dwell time of a system at site 0 (0 RNA:DNA bonds) to be the same as the dwell time of unbinding for a spacer with noncomplementary PAM-proximal nucleotides. Therefore, the off-rates k_{XF} from any of landscapes obeys the following equation:

$$k_{\mathrm{XF}} = k_{\mathrm{OFF}} p_{0,\mathrm{X}}. \tag{5}$$

That is because if we consider the landscape of a fully mismatched spacer, site 0 is the most probable site and therefore the unbinding rate of a fully mismatched guide gives a good approximation of a base k_{OFF}.

An important assumption in our model is that the system undergoes conformation changes which effect a transition to a weaker binding energy landscape. The forward transitions between landscapes (S to I, I to A) can only occur from the final site of a given landscape (when $n = L_{\mathrm{S}}$ for S, and $n = L_{\mathrm{I}}$ for I). This approximation is justified for seed to intermediate since experimentally it has been observed that the rate of unwinding of target DNA is slower with PAM proximal matches, which inhibit the ability to enter the partial R-loop state [20]. And for the transition I to A, as pointed out previously, PAM-distal mismatches prevent the Cas9 enzyme to enter its catalytically active states, which hints that the binding of a PAM-distal mismatch is necessary for the transition.

For a fully complementary spacer sequence the dwell time for both transitions were measured to be on the scale of 1 s [19, 27]. The rates of the transitions k_{SI} and k_{IA} are then approximately 1 s^{-1}. These values of the rates were used as the base rates to calculate the speed of transitions for non-complimentary spacer sequences. Forward transitions in the model can only happen from the most progressed site, so to find rates for non-complementary targets we can adjust the rates by comparing the distributions of the landscapes. For example, the transition rate of for a non-complementary spacer $k_{\mathrm{SI}}^{\mathrm{nc}}$ from seed (S) to intermediate (I) is then:

$$k_{\mathrm{SI}}^{\mathrm{nc}} = k_{\mathrm{SI}} \frac{p_{L_{\mathrm{S}},\mathrm{S}}^{\mathrm{nc}}}{p_{L_{\mathrm{S}},\mathrm{S}}}, \tag{6}$$

where k_{SI} is the experimental smFRET rate for a fully complementary spacer, $p_{L_{\mathrm{S}},\mathrm{S}}^{\mathrm{nc}}$ is the probability of being in the final site of the seed (S) landscape for a non-complementary spacer, and L_{S} is the length of the seed (S) landscape. The same principle can be applied to estimate the transition rate from the intermediate (I) landscape to the active (A) landscape. The total time between binding and cleavage has been measured to be on the order of 10 s using atomic force microscopy measurements [22].

The equilibrium constants of conformational changes, K_{SI} and K_{IA}, are the exponents of their respective Gibbs free energies changes. Therefore, the difference in the Gibbs free energies ΔG_{SI} and ΔG_{IA} is what determines the reverse rates k_{IS} and k_{AI}. The default values for both parameters were set to $\Delta G_{\mathrm{SI}} = -3k_{\mathrm{b}}T$ and $\Delta G_{\mathrm{IA}} = -3k_{\mathrm{b}}T$, which are estimations based on RBT experiments [20] and molecular dynamics simulations [23]. For our model we assume that the reverse transition rates (k_{IS} and k_{AI}) are independent of the complementarity of the spacer to the target. We made this approximation because experimentally we only have values for the reverse transition rates for the complementary spacer CRISPR-Cas9 binding. It is likely that the reverse transition rate will increase for a significantly altered spacer sequence, but it also cannot be arbitrarily large because there must be some attempt rate that serves as an upper-bound. Since we do not know this attempt rate and do not know how the system approaches the attempt rate with increased non-complementarity, we compromise by having a simpler model with landscape independent reversal rates.

The on-rate k_{FS} is inversely proportional to the average time it takes for a single CRISPR-Cas9 to form a PAM-bond at the target site. Estimating the on-rate can be a complex calculation with significant dependence on target DNA accessibility, diffusion, and enzyme structure. In our model we assume that the on-rate is independent of the spacer sequence. This assumption is justified because searching for the correct PAM sequence and forming a PAM bond is a random process that does not rely on information regarding the target sequence. We avoided a first-principles calculation of the on-rate and instead approximated the on-rate k_{FS} from the kinetics of TXTL experiment with an sgRNA that has a fully complementary spacer sequence (supp. figure 7(a)). In a TXTL experiment with pre-synthesized dCas9, nearly complete silencing of the target DNA occurs under 1 h (supp. figure 7(b)). Considering that for the sg6-10 bp guide the binding can be considered nearly irreversible and the concentration of the pre-pressed dCas9 in the TXTL reaction is approximately 20–50 nM [33], we estimate that the on-rate k_{FS} should be on the order of 0.001 s^{-1} nM^{-1}.

With all the rates and processes defined we can construct the system of differential equations that model the dynamics of CRISPR-Cas9 binding (figure 6(b)). The equations describing the system are the following:

$$\frac{\mathrm{d}f}{\mathrm{d}t} = -k_{\mathrm{FS}}\,[\mathrm{Cr}]\,[\mathrm{T}] + k_{\mathrm{SF}}s + k_{\mathrm{IF}}i + k_{\mathrm{AF}}a \tag{7}$$

$$\frac{\mathrm{d}s}{\mathrm{d}t} = k_{\mathrm{FS}}\,[\mathrm{Cr}]\,[\mathrm{T}] + k_{\mathrm{IS}}i - (k_{\mathrm{SF}} + k_{\mathrm{SI}})\,s \tag{8}$$

$$\frac{\mathrm{d}i}{\mathrm{d}t} = k_{\mathrm{SI}}s + k_{\mathrm{AI}}a - (k_{\mathrm{IS}} + k_{\mathrm{IA}} + k_{\mathrm{IF}})\,i \tag{9}$$

$$\frac{\mathrm{d}a}{\mathrm{d}t} = k_{\mathrm{IA}}i - (k_{\mathrm{AI}} + k_{\mathrm{AF}})\,a, \tag{10}$$

where [Cr] is the concentration of the CRISPR-Cas9 complexes from equation (appendix B.3) and [T] is the concentration of the target DNA. To calculate what ratio of targets are bound we set the system of

Table 1. Default values of parameters for modeling length/complementarity dependences in CRISPR-Cas9 TXTL experiments. The n/a for scales S_1 and S_2 means there are no measured values of the parameter in available literature or TXTL experiments.

Parameter	Approximate value	Description
k_{FS}	0.001 $\mathrm{nM}^{-1}\,\mathrm{s}^{-1}$ [TXTL kinetic supp. figures 7(a) and (b)] [TXTL kinetic supp. figures 7(a) and (b)]	CRISPR-dCas9 target DNA search rate/on-rate
k_{OFF}	10–100 s^{-1} [54, 55]	Off-rate of a CRISPR-dCas9 with PAM-proximal mismatches
k_{SI}, k_{IA}	1 s^{-1} [19, 27, 52]	The rates of forward conformational changes (S to I, I to A)
E_{C}	$-1k_{\mathrm{b}}T$ [47–49]	Average free energy difference between an RNA:DNA match and an DNA:DNA match
E_{MM}	$4k_{\mathrm{b}}T$ [50]	Average free energy difference between an RNA:DNA mismatch and a DNA:DNA match
ΔG_{SI}, ΔG_{IA}	$-3k_{\mathrm{b}}T$ [20, 23]	The free energy difference between conformations (S and I, I and A)
S_1, S_2	0.75 (n/a)	The scaling factors to model the binding free energy increase in the more progressed conformations
L_{S}	10 bp [20]	The maximal number of RNA:DNA bonds that can form in the seed conformation
L_{I}	18 bp [15]	The maximal number of RNA:DNA bonds that can form in the intermediate conformation
K_{M}	10 pM to 1 nM [38, 39]	The equilibrium constant of dCas9/Cas9-gRNA complex formation
[dC]	20–50 nM [33]	Concentration of dCas9 in a TXTL extract with pre-expressed dCas9
[sg]	1 nM (TXTL experimental value)	Concentration of sgRNA/scRNA expressing DNA
α	50 (CRISPRi fit, [37])	Steady state ratio of guide RNAs to guide RNA expressing DNA
[T]	0.5 nM (TXTL experimental value)	Concentration of target DNA in the TXTL experiment
k_{CutNorm}	1 s^{-1} [52]	Rate of cleavage for a fully matched target once in the active conformation

ODEs to a steady state and find the steady state concentration of all bound targets, which is the sum of all occupants of states S, I, and A. In the steady state the ratio of bound targets obeys the enzyme–ligand binding equation with a dissociation constant K_{D}:

$$K_{\mathrm{D}} = \frac{k_{\mathrm{SF}}\left(1-\gamma\delta-\delta\right)+k_{\mathrm{IF}}\delta+k_{\mathrm{AF}}\gamma\delta}{k_{\mathrm{FS}}} \tag{11}$$

$$\gamma = \frac{k_{\mathrm{IA}}}{k_{\mathrm{AF}}+k_{\mathrm{AI}}} \tag{12}$$

$$\delta = \frac{k_{\mathrm{SI}}}{k_{\mathrm{IF}}+k_{\mathrm{IS}}+k_{\mathrm{IA}}-k_{\mathrm{AI}}\gamma+k_{\mathrm{SI}}\left(1+\gamma\right)}. \tag{13}$$

Equation (11) is the final derivation of the model and we used it to demonstrate the emergence of peaks as a function of spacer length and complementarity. A summary of all the default parameters and their sources are presented in table 1.

A visual demonstration of the effect truncations or mismatches can carry on the energy landscapes are presented in the following figures: figure 6(c) for a guide with a spacer shorter than L_{S}, figure 6(d) for a guide with a spacer longer than L_{S}, but shorter than L_{I}, and figure 6(e) for a guide longer than L_{I}.

4.3. Emergence of peaks in CRISPR-Cas9 truncation and mismatch experiments

We used equation (11) to test how varying the parameters of the energy landscapes and the transition rates affected the binding efficiency of CRISPR-Cas9 with truncated or mismatched spacers (figure 7). Unless stated otherwise, each of the plots were performed by only perturbing one parameter from table 1, while keeping the rest of parameters the same.

We start with varying the parameters S_1 and S_2, the main assumption of our model. The assumptions states that the absolute values of RNA:DNA bonds and of RNA:DNA mismatches/absences decrease as CRISPR-Cas9 enters its more progressed conformations. In the seed conformation the parameters are E_{C} and E_{MM}, in the intermediate conformation the parameters are S_1E_C and S_1E_{MM}, and in the active conformation the parameters are $S_2S_1E_{\mathrm{C}}$ and $S_2S_1E_{\mathrm{MM}}$. For simplicity, we assigned S_1 and S_2 to be the same and varied their values from 0.5 to 1 (figure 7(a)). As the scaling factors decrease, we notice an emergence of optimal lengths, which appear to be 8, 15, and 22 bp. The optimal lengths are caused by the trapping of the system in a weaker energy landscape. If the deepest site in landscape S is deeper than the deepest site in landscape I, and the transition rate k_{IS} is not negligible, then the landscape I will act as an unbinding sink in the system of differential equations (7)–(10). The same principle can be applied to the transition from the landscape I to landscape A.

In the TXTL experiments we observed that the optimal lengths are different for different guide RNAs (figures 2–5). We hypothesize that

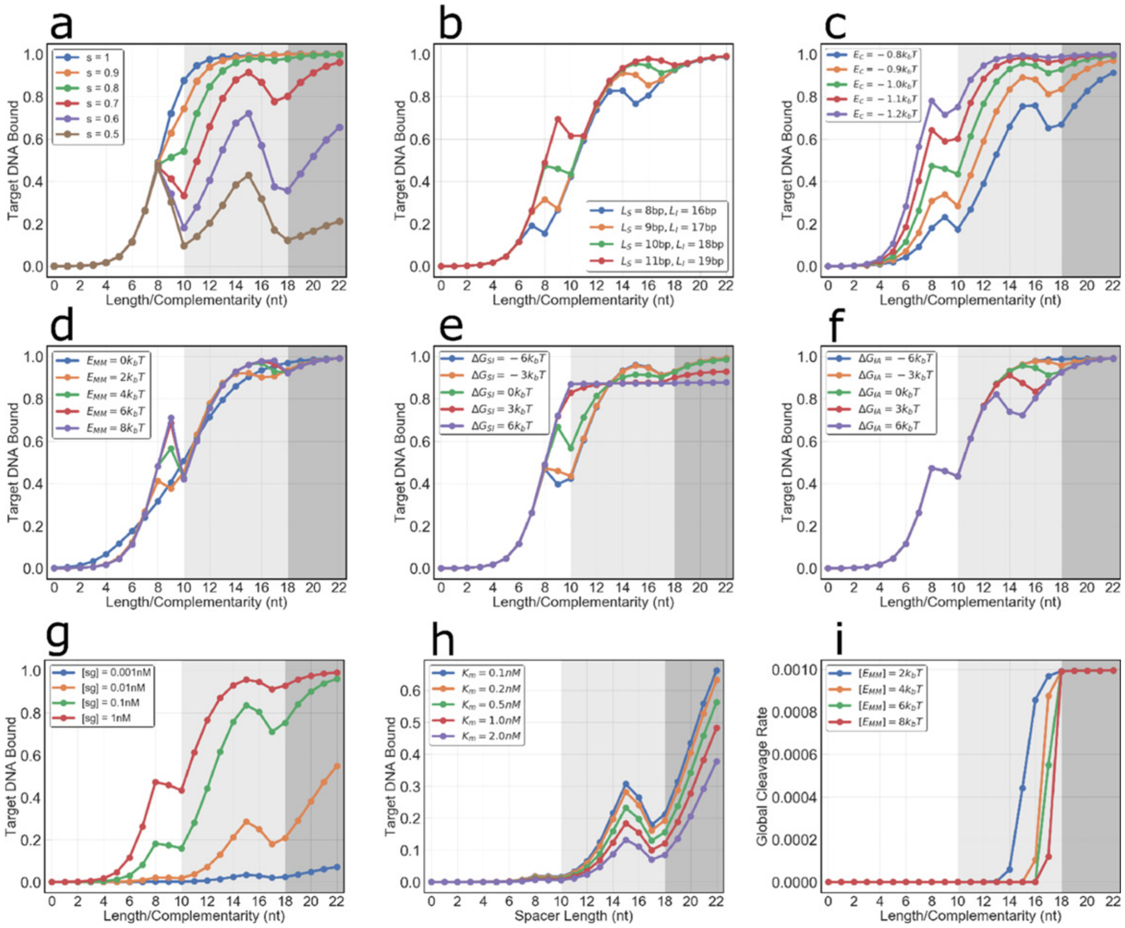

Figure 7. The emergence of multiple optimal lengths in CRISPR-Cas9 length/complementarity plots. In all plots the y-axis corresponds to the ratio of bound targets and the x-axis corresponds to the length/complementarity of the spacer sequence. The transitions to darker shaded areas keep track of the landscape lengths L_S and L_I. L_S is at the transition from white to light grey, and L_I is at the transition from light grey to dark grey. (a) Both energy scale factors S_1 and S_2 were set to the same value ranging from 0.5 to 1. The scale factor $s = 1$ means the average RNA:DNA bond E_C has the same energy in all landscapes, while $s = 0.5$ means an RNA:DNA bond in landscape (I) provides $0.5E_C$ and an RNA:DNA bond in landscape (A) provides $0.25E_C$. It can be seen that for smaller values of s the multiple optimal lengths become more noticeable. (b) To test how in experiment different targets could have different optimal lengths we tested the effect landscape lengths L_S and L_I have on the length/complementarity plot. The different conditions of L_S and L_I change the optimal values of and the overall profile of the length/complementarity plots. (c) Increasing the average RNA:DNA bond energy E_C increases the ratio of bound targets independent of the length/complementarity. (d) The length/complementarity plot for multiple values of the energetic parameter E_{MM}. The values of E_{MM} were ranged between $0k_bT$ to $8k_bT$. (e) and (f) are the length/complementarity plots for different values of ΔG_{SI} and ΔG_{IA}, which are the changes in the Gibbs free energy when the CRISPR-Cas9 system undergoes conformational changes during target interrogation. (g) The length/complementarity plot for different concentrations of the initial sgRNA/scRNA concentration added to the TXTL reaction. (h) The length dependence plot as under conditions when the concentration of dCas9 [dC] are thus the concentration of CRISPR-Cas9 complexes [Cr] are on the same order of as the CRISPR-Cas9 complex formation binding constant K_M. (i) Finally, we are interested in the global cleavage rate as a function of length/complementarity and the mismatch/absence energy E_{MM}. As expected, the target DNA is cleaved nearly immediately upon binding if its length/complementarity are larger or equal to L_I. The larger the E_{MM} the less likely a spacer with a length below L_I can cleave the target DNA.

it is caused by the sequence dependent nature of the lengths of the seed (S) and intermediate (I) landscapes. These lengths correspond to the average number of RNA:DNA bonds the CRISPR-Cas9–target system needs to form to undergo a conformational change. In this work, we set the landscape lengths to $L_S = 10$ bp and $L_I = 18$ bp. Unwinding of DNA and the conformational changes of Cas9 are complex processes that depend on many variables including the sequence. Because of the sequence dependence of DNA unwinding, there could be different options for landscape lengths L_S and L_I. We tested equation (11) with four different combinations of L_S and L_I ($L_S = 8$ bp and $L_I = 16$ bp, $L_S = 9$ bp and $L_I = 17$ bp, $L_S = 10$ bp and $L_I = 18$ bp, $L_S = 11$ bp and $L_I = 19$ bp) (figure 7(b)). Depending on the lengths of the landscapes L_S and L_I the optimal lengths can also vary. It is important to note that our model only accounts for an average amount of RNA:DNA bonds needed to progress to the next step of target recognition.

We plotted the length/complementarity dependence for different E_C values (figure 7(c)). Guides with larger average differences between RNA:DNA bond energies and the DNA:DNA bond energies expectedly had more target DNA bound for all lengths. In our model a mismatch/absence does not actively destabilize the system, but instead makes the system less likely to occupy the final sites of landscapes. That in turn decreases the transition rate to the next landscape. However, since the transition to the next landscape destabilizes the system, it might be beneficial for shorter guides to have larger E_{MM}. We confirm this idea by varying the energetic cost of an RNA:DNA mismatch/absence (figure 7(d)). For $E_{\mathrm{MM}} = 8k_{\mathrm{b}}T$ the unconventional optimal lengths (8–9 bp and 14–16 bp) have stronger bonds than for $E_{\mathrm{MM}} = 1k_{\mathrm{b}}T$ or $E_{\mathrm{MM}} = 2k_{\mathrm{b}}T$. However, when the $E_{\mathrm{MM}} = 0$ the ratio of bound targets monotonically increases with length. The larger E_{MM} values can be considered as modeling the absences, since the last RNA:DNA bond of a truncated spacer should represent the ultimate state of the landscape. The E_{MM} values that range between 2–$6k_{\mathrm{b}}T$ are better suited to demonstrate the effect of mismatches, since those are within the range of realistic mismatch energies [50].

While multiple single-molecule experiments and molecular simulations have been performed that confirm the various conformational changes that CRISPR-Cas9 undergoes during target recognition, the equilibrium constants of the conformational changes vary between experiments. We tested how changing the value of equilibrium coefficients, K_{SI} and K_{IA}, affect the binding efficiency vs complementarity/length dependence (figures 7(e) and (f)). We varied the free energies changes associated with the conformational changes from $-6k_{\mathrm{b}}T$ to $6k_{\mathrm{b}}T$. For most of the conditions we could clearly observe multiple optimal lengths, but for some of the conditions ($\Delta G_{\mathrm{SI}} = 3k_{\mathrm{b}}T$, $6k_{\mathrm{b}}T$) the unconventional optimal values disappear.

It is also of interest to understand how the stoichiometry of TXTL experiments assisted in the observation of the unconventional peak lengths. First, we consider how changing the concentration of the guide RNA expressing DNA affects the complementarity/length dependence (figure 7(g)). As the concentration of the sgRNA DNA was increased, the multiple peaks of the complementarity/length plot became more visible. Conversely, if the concentration of CRISPR-Cas9 is too high the poorly binding lengths between optimums can saturate and reduce the visibility of peaks. Therefore, revealing binding peaks requires a balancing of concentrations.

Another important factor in the stoichiometry of CRISPR-Cas9 is the dependence on K_{M}, the equilibrium constant of Cas9-sgRNA/scRNA complex formation. Previous reports have demonstrated that depending on the length of a spacer K_{M} can vary. For truncated spacers, the K_{M} can increase to the scale of 1 nM [38] and that results in a lower concentration of CRISPR-dCas9/Cas9 complexes with truncated spacers. In the conducted TXTL experiments the concentrations of the free dCas9 and free sgRNA/scRNA were significantly larger than the possible equilibrium constants K_{M} of guides with truncated spacers, therefore effects caused by CRISPR-Cas9 complex formations were unnoticed. However, for an experiment in which the concentrations of the CRISPR components are closer to the K_{M} values, the sgRNAs/scRNAs with truncated spacers might appear as poorly binding guides. Therefore, the dCas9-gRNA binding kinetics can mask the peaks under conditions that either dCas9 or the gRNA or both have comparable concentrations to the K_{M} values. We calculated the expected target bound ratio, when the concentration of the total dCas9 in the reaction is set to $[\mathrm{dCas9}] = 0.5$ nM and ranging the K_{M} values from 0.1 nM to 2 nM, and we indeed observe that if the K_{M} value increases for more truncated guide RNA, then the peaks can be masked.

The last parameter of the model, k_{CUT}, determines the rate at which the target DNA is cut when the system is in the cleavage competent state and more than L_{I} bonds are formed. Therefore, to acquire the cleavage rate for a specific target, value for each target we use the following equation:

$$k_{\mathrm{CUT}} = k_{\mathrm{CutNorm}} * p_{n>L_{\mathrm{I}}}, \tag{14}$$

where $p_{n>L_{\mathrm{I}}}$ is the probability of having more than L_{I} RNA:DNA bonds in the active conformation, which can be calculated with equation (5), and k_{CutNorm} is the approximate rate of cleavage for a fully matched target. By modeling the system as an absorbing Markov chain, we can calculate the mean time to DNA cleavage and, thus, the global cleavage rate of the system [56]. Based on AFM experiments [22], smFRET experiments [16, 27], and kinetics experiments [52] we can approximate the rate to be on the order of $k_{\mathrm{CutNorm}} = 1\ \mathrm{s}^{-1}$ and still get a good understanding of the relation between length/complementarity and the mismatch/absence energy E_{MM}. We ranged E_{MM} from $2k_{\mathrm{b}}T$ to $8k_{\mathrm{b}}T$ and plotted the global cleavage rate as a function of the length/complementarity (figure 7(i)). Expectedly, we observe that the global cleavage rate is approximately the same as the on-rate k_{FS} for targets that can form at least L_{I} RNA:DNA matches. If E_{MM} is lower, then the system can cleave the target even with spacers that have lengths/complementarities of less than L_{I}.

5. Conclusions

Predicting the binding behaviour of CRISPR-Cas9 systems to target DNA sites and non-target DNA sites is an important problem for CRISPR-Cas9 applications. As of now, much of the prediction of on-target and off-target effects is done with empirical models. It is critical to develop physics-based models that account for experimental information, especially when it comes to the development of high-risk CRISPR-Cas9 tools. Since the CRISPR genome editing is still in relatively early stages, physics-based models are being developed part-by-part. There is still important to consider a wide range of timescales, levels of detail, and focus on critical elements of the target interrogation steps, before a more sophisticated evidence model is developed. This work demonstrates the importance of considering conformational changes and DNA unwinding during to model CRISPR-Cas9 target interrogation, which we hope can assist in the development of more complete CRISPR-Cas9 models in the future.

Our model also provides ideas for new potential experiments. First, an experiment that measures whether the binding affinity does decrease in more progressed CRISPR-Cas9 conformations that is paired with a measurement that tracks the mechanical strain the Cas9 enzymes causes on the target DNA can be performed to falsify the model. Also, simulations and experiments need to be performed to analyse the effect conformational changes themselves have on the stability of the system. We can imagine a situation in which the factors S_1 and S_2 are caused by CRISPR-Cas9 undergoing frequent forward and reverse conformational changes during which the CRISPR system is in a transient unstable state. Finally, our model utilizes many parameters that were obtained for a small number of Cas9-gRNA-DNA systems in a cell-free environment. Therefore, to develop a more realistic model based on the dynamics of CRISPR-Cas9 there is a need for high throughput experiments that can track the rates of conformational changes during target interrogation both *in vitro* and *in vivo*.

Funding

This work was supported by the National Science Foundation (CBET 1844152 to VN and JC); and the Binational Science Foundation (award 2018208 to VN).

Conflict of interest

The Noireaux laboratory receives research funds from Arbor Biosciences, a distributor of the myTXTL cell-free protein synthesis kit.

Acknowledgments

AK and VN would like to thank Seth Thompson for plasmid amplifications and Diego Alba Burbano for his helpful manuscript comments.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

ORCID iDs

Vincent Noireaux https://orcid.org/0000-0002-5213-273X

References

[1] Barrangou R and Doudna J A 2016 Applications of CRISPR technologies in research and beyond *Nat. Biotechnol.* **34** 933–41

[2] Anzalone A V, Koblan L W and Liu D R 2020 Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors *Nat. Biotechnol.* **38** 824–44

[3] Pickar-Oliver A and Gersbach C A 2019 The next generation of CRISPR-Cas technologies and applications *Nat. Rev. Mol. Cell Biol.* **20** 490–507

[4] Mohr S E, Hu Y, Ewen-Campen B, Housden B E, Viswanatha R and Perrimon N 2016 CRISPR guide RNA design for research applications *Febs J.* **283** 3232–8

[5] Santos-Moreno J and Schaerli Y 2020 CRISPR-based gene expression control for synthetic gene circuits *Biochem. Soc. Trans.* **48** 1979–93

[6] Kiani S *et al* 2015 Cas9 gRNA engineering for genome editing, activation and repression *Nat. Methods* **12** 1051–4

[7] Yin H *et al* 2018 Partial DNA-guided Cas9 enables genome editing with reduced off-target activity *Nat. Chem. Biol.* **14** 311–6

[8] Fu Y, Sander J D, Reyon D, Cascio V M and Joung J K 2014 Improving CRISPR-Cas nuclease specificity using truncated guide RNAs *Nat. Biotechnol.* **32** 279–84

[9] Qi L S, Larson M H, Gilbert L A, Doudna J A, Weissman J S, Arkin A P and Lim W A 2013 Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression *Cell* **152** 1173–83

[10] Zalatan J G *et al* 2015 Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds *Cell* **160** 339–50

[11] Bikard D, Jiang W, Samai P, Hochschild A, Zhang F and Marraffini L A 2013 Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system *Nucleic Acids Res.* **41** 7429–37

[12] Klein M, Eslami-Mossallam B, Arroyo D G and Depken M 2018 Hybridization kinetics explains CRISPR-Cas off-targeting rules *Cell Rep.* **22** 1413–23

[13] 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

[14] Jinek M *et al* 2014 Structures of Cas9 endonucleases reveal RNA-mediated conformational activation *Science* **343** 1247997

[15] Sternberg S H, Lafrance B, Kaplan M and Doudna J A 2015 Conformational control of DNA target cleavage by CRISPR-Cas9 *Nature* **527** 110–3

[16] 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

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

[18] Chen J S *et al* 2017 Enhanced proofreading governs CRISPR-Cas9 targeting accuracy *Nature* **550** 407–10

[19] Osuka S, Isomura K, Kajimoto S, Komori T, Nishimasu H, Shima T, Nureki O and Uemura S 2018 Real-time observation of flexible domain movements in CRISPR–Cas9 *EMBO J.* **37** e96941

[20] Ivanov I E, Wright A V, Cofsky J C, Aris K D P, Doudna J A and Bryant Z 2020 Cas9 interrogates DNA in discrete steps modulated by mismatches and supercoiling *Proc. Natl Acad. Sci. USA* **117** 5853–60

[21] 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. USA* **111** 9798–803

[22] Shibata M, Nishimasu H, Kodera N, Hirano S, Ando T, Uchihashi T and Nureki O 2017 Real-space and real-time dynamics of CRISPR-Cas9 visualized by high-speed atomic force microscopy *Nat. Commun.* **8** 1430

[23] Palermo G, Miao Y, Walker R C, Jinek M and McCammon J A 2017 CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations *Proc. Natl Acad. Sci. USA* **114** 7260–5

[24] Ricci C G, Chen J S, Miao Y, Jinek M, Doudna J A, McCammon J A and Palermo G 2019 Deciphering off-target effects in CRISPR-Cas9 through accelerated molecular dynamics *ACS Cent. Sci.* **5** 651–62

[25] Casalino L, Nierzwicki Ł, Jinek M and Palermo G 2020 Catalytic mechanism of non-target DNA cleavage in CRISPR-Cas9 revealed by *ab initio* molecular dynamics *ACS Catal.* **10** 13596–605

[26] East K W *et al* 2020 Allosteric motions of the CRISPR-Cas9 HNH nuclease probed by NMR and molecular dynamics *J. Am. Chem. Soc.* **142** 1348–58

[27] Yang M, Peng S, Sun R, Lin J, Wang N and Chen C 2018 The conformational dynamics of Cas9 governing DNA cleavage are revealed by single-molecule FRET *Cell Rep.* **22** 372–82

[28] Shin J and Noireaux V 2012 An *E. coli* cell-free expression toolbox: application to synthetic gene circuits and artificial cells *ACS Synth. Biol.* **1** 29–41

[29] Garamella J, Marshall R, Rustad M and Noireaux V 2016 The all *E. coli* TX-TL toolbox 2.0: a platform for cell-free synthetic biology *ACS Synth. Biol.* **5** 344–55

[30] Dong C, Fontana J, Patel A, Carothers J M and Zalatan J G 2018 Synthetic CRISPR-Cas gene activators for transcriptional reprogramming in bacteria *Nat. Commun.* **9** 4318

[31] Marshall R, Maxwell C S, Collins S P, Beisel C L and Noireaux V 2017 Short DNA containing χ sites enhances DNA stability and gene expression in *E. coli* cell-free transcription–translation systems *Biotechnol. Bioeng.* **114** 2137–41

[32] Suzuki S *et al* 2020 Meganuclease-based artificial transcription factors *ACS Synth. Biol.* **9** 2851–5

[33] Marshall R *et al* 2018 Rapid and scalable characterization of CRISPR technologies using an *E. coli* cell-free transcription–translation system *Mol. Cell* **69** 146

[34] Fontana J, Dong C, Kiattisewee C, Chavali V P, Tickman B I, Carothers J M and Zalatan J G 2020 Effective CRISPRa-mediated control of gene expression in bacteria must overcome strict target site requirements *Nat. Commun.* **11** 1618

[35] Johansson H E, Liljas L and Uhlenbeck O C 1997 RNA recognition by the MS2 phage coat protein *Semin. Virol.* **8** 176–85

[36] Griffith K L and Wolf R E 2004 Genetic evidence for pre-recruitment as the mechanism of transcription activation by SoxS of *Escherichia coli*: the dominance of DNA binding mutations of SoxS *J. Mol. Biol.* **344** 1–10

[37] Marshall R and Noireaux V 2019 Quantitative modeling of transcription and translation of an all-*E. coli* cell-free system *Sci. Rep.* **9** 11980

[38] Wright A V, Sternberg S H, Taylor D W, Staahl B T, Bardales J A, Kornfeld J E and Doudna J A 2015 Rational design of a split-Cas9 enzyme complex *Proc. Natl Acad. Sci. USA* **112** 2984–9

[39] Mekler V, Minakhin L, Semenova E, Kuznedelov K and Severinov K 2016 Kinetics of the CRISPR-Cas9 effector complex assembly and the role of 3′-terminal segment of guide RNA *Nucleic Acids Res.* **44** 2837–45

[40] Santos-Moreno J, Tasiudi E, Stelling J and Schaerli Y 2020 Multistable and dynamic CRISPRi-based synthetic circuits *Nat. Commun.* **11** 2746

[41] Jiang F, Zhou K, Ma L, Gressel S and Doudna J A 2015 A Cas9-guide RNA complex preorganized for target DNA recognition *Science* **348** 1477–81

[42] Jiang F, Taylor D W, Chen J S, Kornfeld J E, Zhou K, Thompson A J, Nogales E and Doudna J A 2016 Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage *Science* **351** 867–71

[43] Newton M D, Taylor B J, Driessen R P C, Roos L, Cvetesic N, Allyjaun S, Lenhard B, Cuomo M E and Rueda D S 2019 DNA stretching induces Cas9 off-target activity *Nat. Struct. Mol. Biol.* **26** 185–92

[44] Eslami-Mossallam B, Klein M, Smagt C, Sanden K, Jones S, Hawkins J, Finkelstein I and Depken M 2020 A kinetic model improves off-target predictions and reveals the physical basis of Sp Cas9 fidelity (bioRxiv 2020.05.21.108613) (Retrieved 10 February 2021)

[45] Doench J G *et al* 2016 Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9 *Nat. Biotechnol.* **34** 184–91

[46] Zhang D, Hurst T, Duan D and Chen S-J 2019 Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design *Proc. Natl Acad. Sci. USA* **116** 8693–8

[47] Sugimoto N, Nakano S-i, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, Yoneyama M and Sasaki M 1995 Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes *Biochemistry* **34** 11211–6

[48] SantaLucia J 1998 A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics *Proc. Natl Acad. Sci.* **95** 1460–5

[49] SantaLucia J, Allawi H T and Seneviratne P A 1996 Improved nearest-neighbor parameters for predicting DNA duplex stability† *Biochemistry* **35** 3555–62

[50] Watkins N E, Kennelly W J, Tsay M J, Tuin A, Swenson L, Lee H-R, Morosyuk S, Hicks D A and Santalucia J 2011 Thermodynamic contributions of single internal rA·dA, rC·dC, rG·dG and rU·dT mismatches in RNA/DNA duplexes *Nucleic Acids Res.* **39** 1894–902

[51] Ouldridge T E, Šulc P, Romano F, Doye J P K and Louis A A 2013 DNA hybridization kinetics: zippering, internal displacement and sequence dependence *Nucleic Acids Res.* **41** 8886–95

[52] Gong S, Yu H H, Johnson K A and Taylor D W 2018 DNA unwinding is the primary determinant of CRISPR-Cas9 activity *Cell Rep.* **22** 359–71

[53] Kittel C, Kroemer H and Scott H L 1998 Thermal physics, 2nd ed *Am. J. Phys.* **66** 164–7

[54] Jones D L, Leroy P, Unoson C, Fange D, Ćurić V, Lawson M J and Elf J 2017 Kinetics of dCas9 target search in *Escherichia coli Science* **357** 1420–4

[55] Martens K J A *et al* 2019 Visualisation of dCas9 target search *in vivo* using an open-microscopy framework *Nat. Commun.* **10** 3552

[56] Barbour A D and Resnick S I 1993 Adventures in stochastic processes *J. Am. Stat. Assoc.* **88** 1474