Massively parallel CRISPRi assays reveal concealed thermodynamic determinants of dCas12a binding

David A. Specht,[1] Yasu Xu,[2] and Guillaume Lambert[1,*]

[1] School of Applied and Engineering Physics, Cornell University, Ithaca, NY 14853, USA
[2] Field of Biophysics, Cornell University, Ithaca, NY 14853, USA

The versatility of CRISPR-Cas endonucleases as a tool for biomedical research has lead to diverse applications in gene editing, programmable transcriptional control, and nucleic acid detection. Most CRISPR-Cas systems, however, suffer from off-target effects and unpredictable non-specific binding that negatively impact their reliability and broader applicability. To better evaluate the impact of mismatches on DNA target recognition and binding, we develop a massively parallel CRISPR interference (CRISPRi) assay to measure the binding energy between tens of thousands of CRISPR RNA (crRNA) and target DNA sequences. By developing a general thermodynamic model of CRISPR-Cas binding dynamics, our results unravel a comprehensive map of the energetic landscape of *Francisella novicida* Cas12a (FnCas12a) as it searches for its DNA target. Our results reveal concealed thermodynamic factors affecting FnCas12a DNA binding which should guide the design and optimization of crRNA that limit off-target effects, including the crucial role of an extended, 6-base long PAM sequence and the impact of the specific base composition of crRNA-DNA mismatches. Our generalizable approach should also provide a mechanistic understanding of target recognition and DNA binding when applied to other CRISPR-Cas systems.

INTRODUCTION

Clustered regularly interspaced short palindromic repeats (CRISPR) and its associated genes are part of an adaptive immunity system used to combat phage infections in bacteria and archaea [1]. The system consists of two main components: a CRISPR array, which contains repetitive sequences called repeats and variable sequences called spacers, and CRISPR-associated (Cas) genes, which facilitate spacer acquisition and the destruction of foreign DNA and RNA. Mature CRISPR RNAs (crRNAs) derived from the CRISPR array can in turn program Cas nucleases to recognize and cleave DNA targets whose nucleic acid sequence is complementary with the guide portion of the crRNA and proximal to a PAM (protospacer adjacent motif) site. Due to their simple and programmable nature, the nucleases of class 2 CRISPR systems, particularly Cas9 (type II) and Cas12 (type V), have been the subject of intense research interest for the purposes of genome editing [2–4], programmable gene regulation utilizing a catalytically-dead CRISPR nuclease (dCas) [5–7], and nucleic acid detection [8, 9].

While CRISPR has already revolutionized many areas of research, from fundamental biomedical sciences to synthetic biology to disease diagnostics, a fundamental understanding of the underlying factors affecting CRISPR-Cas off-target *binding* is still lacking. This is especially important in the context of CRISPR base editors [10, 11] because off-target binding, which may not entirely correlate with DNA cleavage [12–14], needs to be reduced to a minimum level to prevent unintended base changes. While several *in silico* models [15–20] have been developed to predict the binding affinity of RNA guided CRISPR-Cas proteins using data from *in vitro* biochemical assays [21–24] or *in vivo* indel frequencies [12–14, 25–27], these approaches only provide empirical interpretations of CRISPR-Cas DNA binding and often fail to yield a conceptual understanding of the underlying factors involved in CRISPR-Cas binding. Furthermore, it can be difficult to extract quantitative binding affinity measurements from *in vivo* indel frequencies due to the inherent CRISPR-Cas binding inefficiencies associated with cellular physiological factors such as cell type, chromatin state, and delivery method [28–30]. Thus, there is a critical need for fundamental models that can help unravel the sequence-dependent determinants of CRISPR-Cas target recognition and DNA binding affinity.

To elucidate determinants of CRISPR-Cas12 off-target binding, we combine a thermodynamic model of dCas12a binding with a rationally designed CRISPRi assays to map the binding energy landscape of a type V CRISPR-Cas system from *Francisella novicida* (FnCas12a) as it inspects and binds to its DNA targets. Our approach, inspired by a recent theoretical framework that employs a unified energetic analysis to predict *S. pyogenes* Cas9 (SpCas9) cleavage activity [31] and recently developed massively parallel multiplexed assays [32–35], aims to directly measure the energetic and thermodynamic determinants of CRISPR-Cas *binding*. In other words, our assays excludes sources of variation in DNA cleavage activity caused by unknown physiological factors [28–30] by only focusing on the steps *leading* to final DNA cleavage step. Furthermore, our predictive framework is not limited to FnCas12a and can be applied to any other CRISPR-Cas systems, which should in turn fa-

[*] Corresponding author

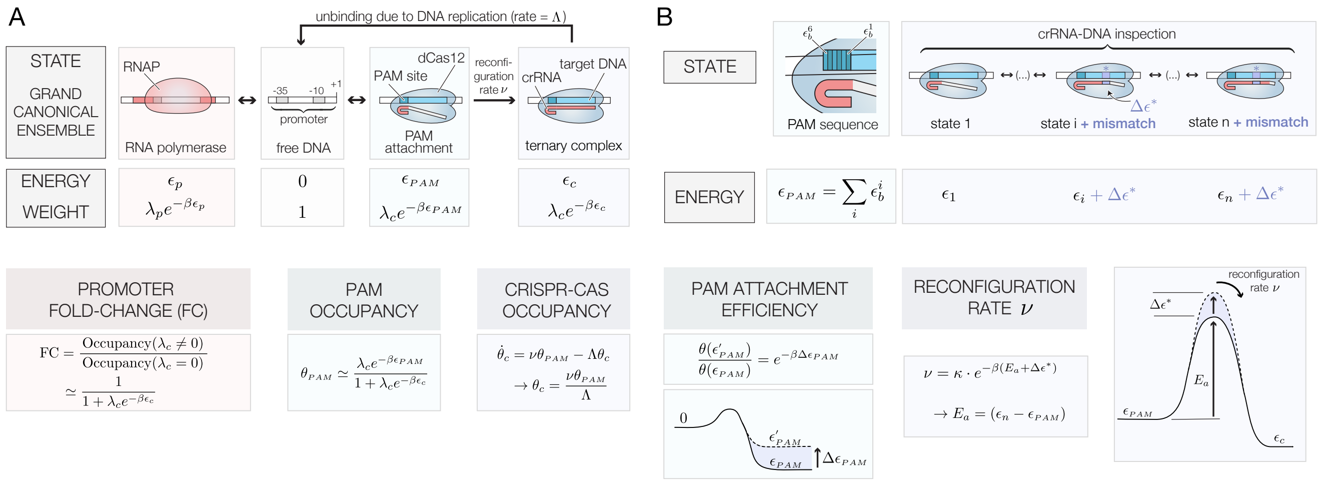

A
unbinding due to DNA replication (rate = Λ)
STATE
GRAND CANONICAL ENSEMBLE
RNAP
RNA polymerase
-35
-10
+1
promoter
free DNA
dCas12
PAM site
PAM attachment
reconfi-guration rate ν
target DNA
crRNA
ternary complex
ENERGY
WEIGHT
ϵ_p
$\lambda_p e^{-\beta\epsilon_p}$
0
1
ϵ_{PAM}
$\lambda_c e^{-\beta\epsilon_{PAM}}$
ϵ_c
$\lambda_c e^{-\beta\epsilon_c}$
PROMOTER FOLD-CHANGE (FC)
$\mathrm{FC} = \frac{\mathrm{Occupancy}(\lambda_c \neq 0)}{\mathrm{Occupancy}(\lambda_c = 0)} \simeq \frac{1}{1+\lambda_c e^{-\beta\epsilon_c}}$
PAM OCCUPANCY
$\theta_{PAM} \simeq \frac{\lambda_c e^{-\beta\epsilon_{PAM}}}{1+\lambda_c e^{-\beta\epsilon_c}}$
CRISPR-CAS OCCUPANCY
$\dot{\theta}_c = \nu\theta_{PAM} - \Lambda\theta_c \rightarrow \theta_c = \frac{\nu\theta_{PAM}}{\Lambda}$
B
STATE
ϵ_b^6
ϵ_b^1
PAM sequence
crRNA-DNA inspection
↔ (...) ↔
$\Delta\epsilon^*$
state 1
state i + mismatch
state n + mismatch
ENERGY
$\epsilon_{PAM} = \sum_i \epsilon_b^i$
ϵ_1
$\epsilon_i + \Delta\epsilon^*$
$\epsilon_n + \Delta\epsilon^*$
PAM ATTACHMENT EFFICIENCY
$\frac{\theta(\epsilon'_{PAM})}{\theta(\epsilon_{PAM})} = e^{-\beta\Delta\epsilon_{PAM}}$
0
ϵ'_{PAM}
ϵ_{PAM}
$\Delta\epsilon_{PAM}$
RECONFIGURATION RATE ν
$\nu = \kappa \cdot e^{-\beta(E_a+\Delta\epsilon^*)}$
$\rightarrow E_a = (\epsilon_n - \epsilon_{PAM})$
reconfiguration rate ν
$\Delta\epsilon^*$
E_a
ϵ_{PAM}
ϵ_c

FIG. 1. Thermodynamic model used to describe a nuclease-dead Cas12 endonuclease's "PAM attachment," "crRNA-DNA inspection,", and "reconfiguration" steps. A) Energy states, energies and Boltzmann weights of a dCas12a ($\beta = k_BT$). The fold-change, PAM occupancy and CRISPR-Cas occupancy depends on the effective PAM energy ϵ_{PAM} and CRISPR-Cas binding energy ϵ_c. All expressions assume the weak promoter ($\lambda_p e^{-\beta\epsilon_p} \ll 1$) and weak PAM binding ($\lambda_c e^{-\beta\epsilon_{PAM}} \ll 1$) limits. B) Internal base-dependent states define a PAM specific binding energy. The specific PAM sequence dictates the relative PAM attachment efficiency between two targets. The presence of crRNA-target DNA mismatches increase the effective activation energy E_a and affect the effective reconfiguration rate ν.

cilitate the development of predictive models of target recognition and binding efficiency for type II and type V RNA-guided CRISPR-Cas proteins.

RESULTS

Thermodynamic model of dCas binding

DNA cleavage by CRISPR-Cas endonucleases may be hindered by other factors [28–30] besides the specific crRNA-DNA sequence, and it is important to disentangle these effects to gain a deep understanding of off-target binding mechanisms. We thus hypothesize that the variability in indel formation observed in live cells may not entirely originate from differences in Cas12a's cleavage activity caused by the specific crRNA-DNA sequence targeted, but also from sequence-dependent PAM attachment efficiencies and the existence of crRNA-target DNA mismatches. In other words, we ask whether the steps *leading* to a ternary complex formation play a role in CRISPR-Cas off-target binding affinity.

To formalize this approach and to obtain a fundamental understanding of the energetic landscape of dCas12a as it inspects and associates with its DNA target, we developed a general thermodynamic model of CRISPR-Cas binding dynamics to determine how crRNA-DNA mismatches affect FnCas12a target recognition and binding. This model (see supplementary information and Fig. 1) is based on recent structural biology and single-molecule studies [36, 37] which revealed that DNA hydrolysis by Cas12a occurs in three discrete stages: "PAM attachment," where FnCas12a latches onto a PAM site, "crRNA-DNA inspection," where FnCas12a forms a partial crRNA-DNA hybrid, and "reconfiguration," where the protein forms a ternary complex and undergoes a conformal change that exposes its catalytic residues. While the final DNA cleaving step occurs after approximately 1 minute under the conditions tested in [36], Cas12 molecules with inactivated nuclease sites remain stably bound to their DNA target for more than 500s. Hence, the reconfiguration step effectively has no detectable *off*-rate, suggesting that DNA cleavage may be inevitable (given enough time) once Cas12a has reached this stably-bound ternary state. The same stability has also been observed in single-molecule Cas9 experiments [35].

Hence, our thermodynamic model describes the probability that FnCas12a loaded with a crRNA sequence will bind to a free, unobstructed target DNA sequence using the grand canonical ensemble [38–40] to derive an expression for θ_c, the FnCas12a occupancy, which is defined as the fraction of time a DNA target will be occupied by nuclease-dead FnCas12a endonuclease. This occupancy is given by

$$\theta_c = \nu\frac{\theta_{PAM}}{\Lambda} \tag{1}$$

where θ_{PAM} is the PAM occupancy (the attachment probability) and ν is the probability that FnCas12a will form a stable ternary complex once it encounters a PAM site (the reconfiguration rate). Since DNA replication forks appear to be the only processes that can kick nuclease-dead SpCas9 off of its DNA binding site [41], we assume dCas12a unbinding occurs through a similar process–i.e. DNA duplication machinery kicks-off FnCas12a at a rate equal to Λ (the cell's duplication rate).

We next use this approach to compare occupancies of targets that vary by a few base determinants (Fig. 1B). In this framework, the propensity of a given crRNA to target to bind to an off-target DNA region compared with its intended target is simply given by the different energetic contributions of that specific off-target location. For instance, two identical DNA targets that possess different PAM sequences have effective binding energies that differ by $\Delta\epsilon_{PAM}$, which in turn translates in a reduction of the attachment probability by a factor equal to $e^{-\beta\Delta\epsilon_{PAM}}$ (the Boltzmann factor). Similarly, the presence of mismatches may alter the crRNA-DNA duplex energy by $\Delta\epsilon^{*}$, which in turn also yields a $e^{-\beta\Delta\epsilon^{*}}$ change in relative binding probabilities. Hence, the *relative* binding affinity between two targets that have different PAM sites, or between an intended target and an off-target candidate, is simply given by the binding sites' Boltzmann weight

$$\text{Relative binding affinity} = \underbrace{e^{-\beta\Delta\epsilon_{PAM}}}_{\text{PAM}}\ \underbrace{e^{-\beta\Delta\epsilon^{*}}}_{\text{Mismatches}}. \qquad (2)$$

Our framework shares similarities with the uCRISPR model recently developed by Zhang *et al.* to describe SpCas9 cleavage activity [31]. However, instead of testing our model using *in vivo* indel measurements performed in human cells (which can be imprecise due to cellular physiological factors [28–30]), we use a massively parallel CRISPRi assay to directly measure the sequence-specific PAM binding energies and the energetic costs associated with crRNA-DNA mismatches in *E. coli* bacteria.

Context dependence of FnCas12a CRISPR interference

In order to test our thermodynamic model and further explore FnCas12a target binding in *E. coli*, we developed a highly compact, 175bp-long genetic inverter inserted into a low-copy number plasmid (pSC101) containing a catalytically-dead nuclease FnCas12a (Fig. 2A, inset). The inverter element consists of a constitutive promoter driving the expression of a crRNA followed by two rho-independent terminators. Located immediately downstream of two terminators is the output promoter, which either contains a built-in PAM site within the promoter or after the promoter's +1 location.

We first sought to investigate effectiveness of Cas12a-mediated CRISPRi by measuring protein and mRNA levels of a simple inverter driving sfGFP expression. The inverter constitutively expresses a crRNA targeting a DNA binding region located at the promoter's -19 position. Fluorescence levels for constructs containing a crRNA were 24.3 times lower than those without a crRNA (Fig. 2B) and mRNA transcript levels measured using digital droplet PCR resulted in a 123-fold reduction in mRNA transcript levels when a crRNA is expressed (Fig. 2C). Both of these results confirm that FnCas12a can repress RNA transcription [7].

Next, we tested how dCas12a interferes with RNA transcription under various configurations (Figs. 2D-E) by placing a library of up to several thousands simple inverter constructs in front of a tetA-sacB cassette. Since sacB is counterselectable genetic markers in the presence of sucrose [44] (see Fig. S2), the genetic inverters that efficiently repress RNA transcription will be enriched in the population when grown under sucrose conditions (SK). Thus, we can evaluate the ability of a RNA-guided FnCas12a to prevent transcription by comparing the number of times each construct is present in the whole population for control (K) and SK conditions using the MiSeq or iSeq100 platform from Illumina. The relative change in the population fraction is then used to find the effective growth rate Λ of every construct in each condition. While selection experiments are also performed under tetracycline-selective (TK) media, the counterselection experiment (SK) yields more useful information because the binding affinity and the dCas12 promoter occupancy is directly related to each construct's growth rate (see supplementary materials for a complete description of this method).

Fig. 2D (top) and S3 shows that CRISPRi occurs efficiently when the FnCas12a target is located after the output promoter's +1 transcription initiation site because the growth rate under SK conditions is close to its maximum value (Λ_0) regardless of the location of the DNA binding site. Interestingly, while interference measurements performed using SpCas9 (a type II CRISPR-Cas nuclease) revealed that a second binding site results in suppressive combinatorial effects that multiplicatively increases CRISPRi efficiency [6], the existence of a second PAM+target DNA sequence does not improve dCas12 CRISPRi efficiency beyond what is achieved by a single target (Figs. 2D, bottom and S3).

Next, we tested FnCas12's ability to interfere with RNA transcription initiation by introducing a PAM+target DNA sequence within the promoter sequence. In particular, we tested several inverter constructs whose PAM+target DNA sequence was located at different positions within the promoter's -35 and -1 location, testing both the coding and template strands without altering conserved promoter regions (Fig. 2E). Our results show that CRISPR interference through promoter occlusion is efficient for most targets on both the coding and template strands, although the effective repression rate is more variable than what has been reported for CRISPR-Cas9 in-

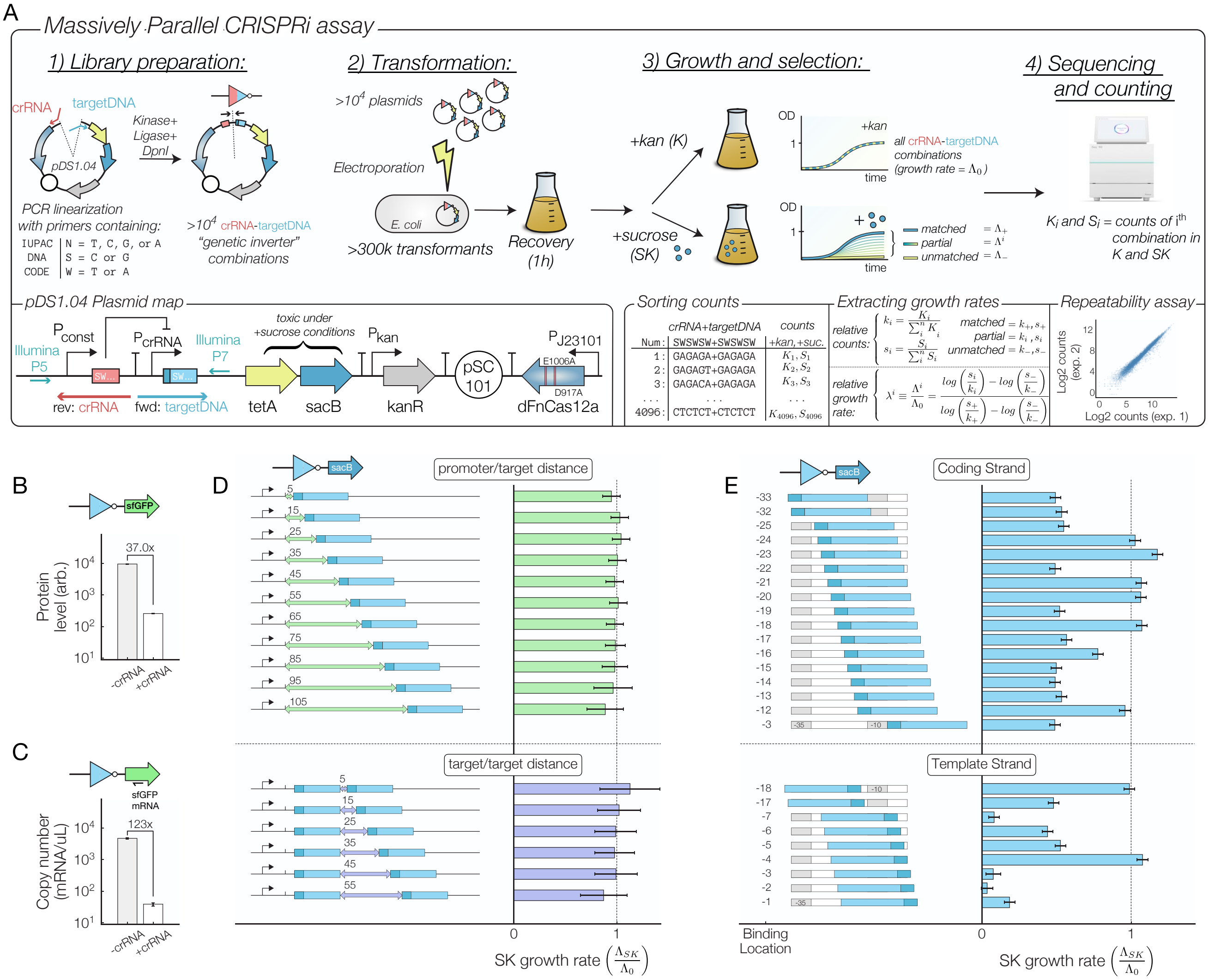
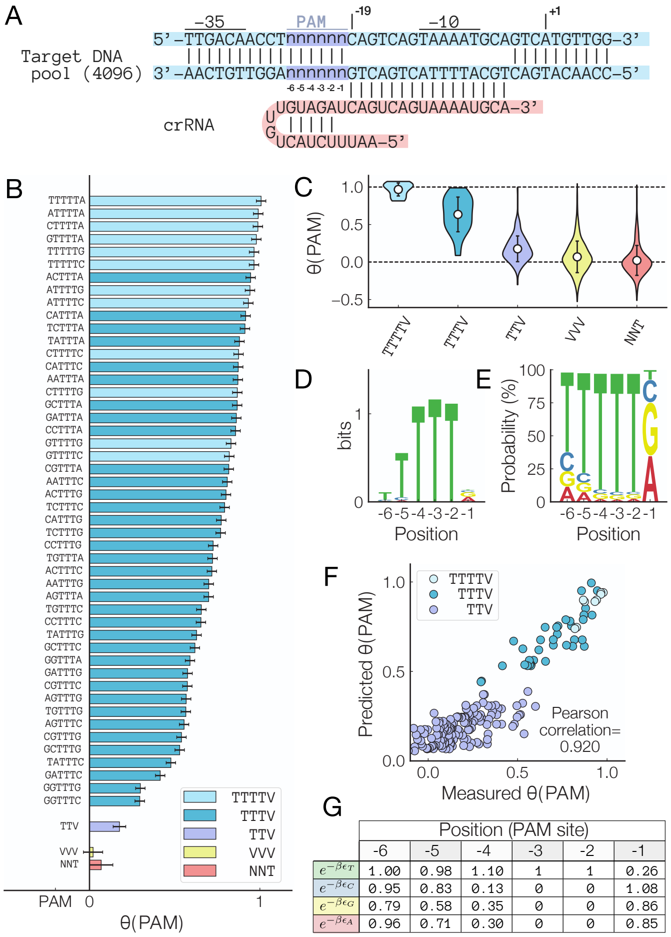


FIG. 2. A) Experimental workflow. More than 10^4 different crRNA-target DNA combinations are assembled in parallel using PCR primers containing degenerate IUPAC DNA codes (e.g. S, W, N). The ability of each construct to repress a tetA-sacB cassette is measured by comparing growth rates under control (K) and sucrose (SK) conditions. While the library construction is prone to some biases in the PCR amplification and transformation steps, a high level of repeatability is observed between experiments that started with the same assembled library (lower right). B) Protein and C) RNA level fold-change for a genetic inverter diving sfGFP expression. D) Growth rate under sucrose conditions when one or two DNA targets are located after the +1 promoter location and when E) the DNA target overlaps with the -35 and -10 regions of promoter. Λ_0 = growth rate under control (K) conditions. Error bars are calculated using a LOESS fit [42, 43] of the mean/variance relationship between experimental replicates of the fold change.

terference [6]. Growth under SK conditions is also lowest when the target DNA is located on the promoter's template strand at locations -1, -2, -3, and -7 with respect to the transcription initiation site, which suggests that RNA:DNA hybrids on the non-template strand display a decreased effectiveness in preventing RNA transcription initiation.

FnCas12a binding energies depend on an extended PAM sequence

Having demonstrated the validity of our massively parallel CRISPRi assay to test multiple genetic inverter combinations, we next investigated the impact of a PAM sequence on the binding affinity of dCas12. We first tested the sequence determinant of the PAM attachment step using an oligo pool containing a degenerate 5'-NNNNNN-3' motif for a target DNA sequence located at the promoter's -19 posi-



	Position (PAM site)					
	-6	-5	-4	-3	-2	-1
$e^{-\beta\epsilon_T}$	1.00	0.98	1.10	1	1	0.26
$e^{-\beta\epsilon_C}$	0.95	0.83	0.13	0	0	1.08
$e^{-\beta\epsilon_G}$	0.79	0.58	0.35	0	0	0.86
$e^{-\beta\epsilon_A}$	0.96	0.71	0.30	0	0	0.85

FIG. 3. A) Sequence of the PAM site occupancy library. B) Measured PAM occupancies for all 6-base PAM sites. Error bars = LOESS fit of the mean/variance relationship between experimental replicates of the fold change. C) Aggregated PAM site occupancies. Error bars = std. dev. D) Bit content and E) probability density of the SK selected PAM site libraries. F) Predicted $\theta(PAM)$ using the base-dependent binding energy expression $\epsilon_{PAM} = \sum_i \epsilon_b^i$ G) Fitted values for each position- and base-dependent binding energies.

tion (Fig. 3A) targeted by a single crRNA (targetDNA sequence=CAGTCAGTAAAATGCA). While previous work has shown that the PAM motif required for DNA cleavage are TTV or TTTV, depending on the species of Cas12 tested [3], we tested all 4,096 PAM site variants in a single experiment in order to measure the attachment efficiency of all sequences containing up to six bases of upstream context. These extra bases turn out to be very important: Fig. 3B shows that while TTV is a suitable PAM site, its attachment efficiency is approximately six times lower than the PAM site with the highest measured attachment efficiency (TTTTTA). In both individual and aggregate measurements, we observe that DNA binding to a DNA target proximal to a TTTV or TTTTV PAM site is 3.6 and 5.5 times more efficient than a TTV PAM site, respectively (Fig. 3C). This result is also confirmed by the bias towards TTTTTV PAM sites in the information content (Fig 3D) and the base-specific probability density in SK conditions (Fig. 3E).

Our results agree with recent work [45] which demonstrated that FnCas12a does exhibit activity in mammalian cells, but only when used with a TTTV PAM site. It is important to note that while Zetsche *et al.* [3] showed that a TTV PAM site appears to be sufficient to induce FnCas12a cleavage, it appears to be the *least efficient* motif that permits DNA binding (which could explain why FnCas12a was found to be ineffectual for mammalian cell editing using a TTV PAM site). Hence, our results suggest that PAM sites with an extended TTTTTV sequence should be prioritized when seeking potential FnCas12a DNA targets for CRISPRi, gene editing, nucleic-acid detection, or other applications.

Expanding on this result, we next used the measured attachment efficiencies to develop a predictive model that takes into account the full 6-base PAM site context to predict the attachment efficiency. Specifically, a natural prediction that emerges from our thermodynamics model is that the effective PAM site attachment energy is *additive*, meaning that to PAM binding energy ϵ_{PAM} of an arbitrary sequence is given by $\epsilon_{PAM} = \sum \epsilon_b^i$, where ϵ_b^i is the specific binding energy of a base of type b=(T,C,G,A) at location i=(1..6). In this case the relative PAM binding energy between two targets ($\Delta\epsilon_{PAM} = \epsilon'_{PAM} - \epsilon_{PAM}$) is related to the relative growth rate $\lambda(PAM)$ under SK condition according to $\lambda(PAM')/\lambda(PAM) = e^{-\Delta\epsilon_{PAM}}$.

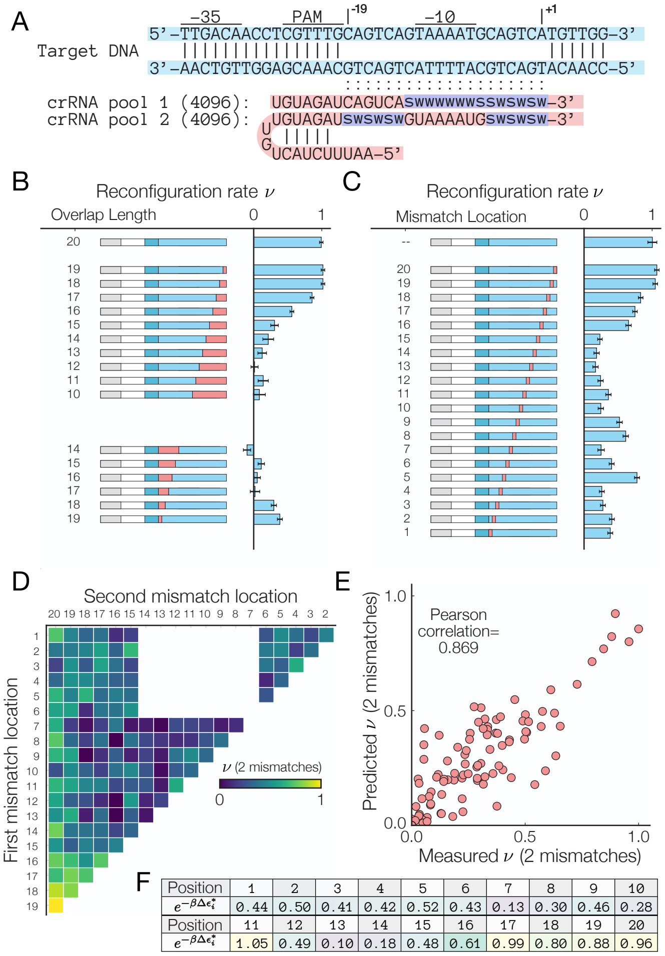
By using an initial set of values for each ϵ_b^i extracted from the PAM specific growth rates (see supplementary methods for details), we compare the accuracy of the model prediction with measured occupancies (Fig. 3F). Surprisingly, the Pearson correlation between the predicted and measured growth rate values using only the baseline value for all $e^{-\beta\epsilon_b^i}$ is fairly high at 0.911. The model was then optimized for 1,000 additional steps to minimize the measured-predicted mean square error and its final predictions agree very well with the measured attachment efficiencies (Pearson correlation=0.920). The optimized energetic contribution ϵ_b^i of each base b located at position i is shown in Fig. 3G. Hence, to ensure that the DNA target with the most efficient PAM site is selected when designing and optimizing a crRNA sequence for DNA binding or other gene editing application, we strongly recommend that the relative performance of each PAM sequence is evaluated on a sequence-specific manner using the base-dependent binding energies provided in Fig. 3G.

Off-target FnCas12a binding depends additively on mismatch energy

To better understand the impact of crRNA-DNA mismatches on dCas12 binding, we next examined how a mismatch affects the effective activation energy (Fig. 1B) that is required for FnCas12a to form a stable ternary complex. Indeed, even though a PAM site is present and dCas12 attaches itself to DNA, the additional energy associated with a crRNA-DNA mismatch can prevent DNA unzipping if insufficient homology is found. According to our model, the reconfiguration step occurs at a rate $\nu = e^{-\beta \sum_i \Delta\epsilon_i^*}$, where $\Delta\epsilon_i^*$ is the base-dependent energy cost associated with a single mismatche at location i. Thus, the location-specific energy costs associated with individual mismatches should in theory be directly obtained by measuring the reconfiguration rate ν of crRNA-DNA sequences that possess the same PAM sequence but with a crRNA that differs from the target DNA by one or more bases.

To test this, we used two different crRNA pools (Fig, 4A) to the mismatch-dependent reconfiguration rate ν. Each oligo pool consists of 4,096 different primer sequences generated by specifying degenerate DNA codes in the primer sequence (e.g. W = A or T, S = G or C), allowing us to test multiple mismatch combinations in a single experiment. Using the degenerate DNA codes S and W ensures that all crRNA sequences maintained the same GC content. In Fig. 4B, we tested the impact of "truncated" (i.e. a crRNA whose distal sequence is noncomplementary to its target DNA) and "gapped" (i.e. a crRNA whose seed sequence is noncomplementary to its target DNA) crRNAs . Consistent with other work performed in Cas12a [26, 27], our results show that optimal reconfiguration rates occur for truncated crRNAs that possess more than 15 bases of homology. Furthermore, no significant binding was detected for gapped crRNAs whose sequences that contain more than 2 mismatches.

Next, we measured the reconfiguration rate for crRNA containing a single mismatch (Fig. 4C). The presence of a single mismatch can decrease the configuration rate by up to 82% when the mismatch occurs in the first 17 bases of the crRNA. Consistent with prior observations by Kim *et al.* [19], the energy cost of a single mismatch does not increase monotonically with distance from the PAM site, suggesting that other contextual determinants besides position affects the reconfiguration rate ν. Furthermore, the presence of mismatches located in the last 3 bases of the crRNA does not impede DNA binding, confirming other works performed using *in vivo* indel measurements [26, 27] which demonstrated that crRNA-DNA mismatches negatively impact FnCas12a binding, but only in the seed and the beginning of the trunk region.



Position	1	2	3	4	5	6	7	8	9	10
$e^{-\beta\Delta\epsilon_i^*}$	0.44	0.50	0.41	0.42	0.52	0.43	0.13	0.30	0.46	0.28
Position	11	12	13	14	15	16	17	18	19	20
$e^{-\beta\Delta\epsilon_i^*}$	1.05	0.49	0.10	0.18	0.48	0.61	0.99	0.80	0.88	0.96

FIG. 4. A) Sequence of the single target mismatch libraries. Reconfiguration rates for B) truncated and gapped crRNAs and for C) single mismatch crRNAs. Error bars = LOESS fit of the mean/variance relationship between experimental replicates of the fold change. D) Experimental and E) predicted reconfiguration rates for crRNA with 2 mismatches. F) Fitted values for the location-dependent binding energies for single mismatches.

Next, we analyzed how the presence of two mismatches impacts the reconfiguration rate. Since in our model the energetic contributions $\Delta\epsilon_i^*$ of single mismatches at location i are additive, we anticipated that the 2 mismatch reconfiguration rate is related to the single mismatch energies according to $\epsilon_{2MM} = \sum_i \Delta\epsilon_i^*$. To test this, we developed a predictive model that uses the single-base mismatch energies to predict ν_{2MM}. Fig. 4D shows the experimentally measured, location-dependent reconfiguration rate ν_{2MM}. Using an approach similar to the one used to predict PAM attachment efficiencies, we derived baseline values for the location-dependent binding energy. While the initial Pearson correlation between the predicted and baseline energy values was initially fairly low (P=0.769), the predicted values for the two mismatch reconfiguration rate ν_{2MM} agree very well with the measured rates after the 1,000 optimization steps (P=0.869, Fig. 4E). Our results confirm that the en-

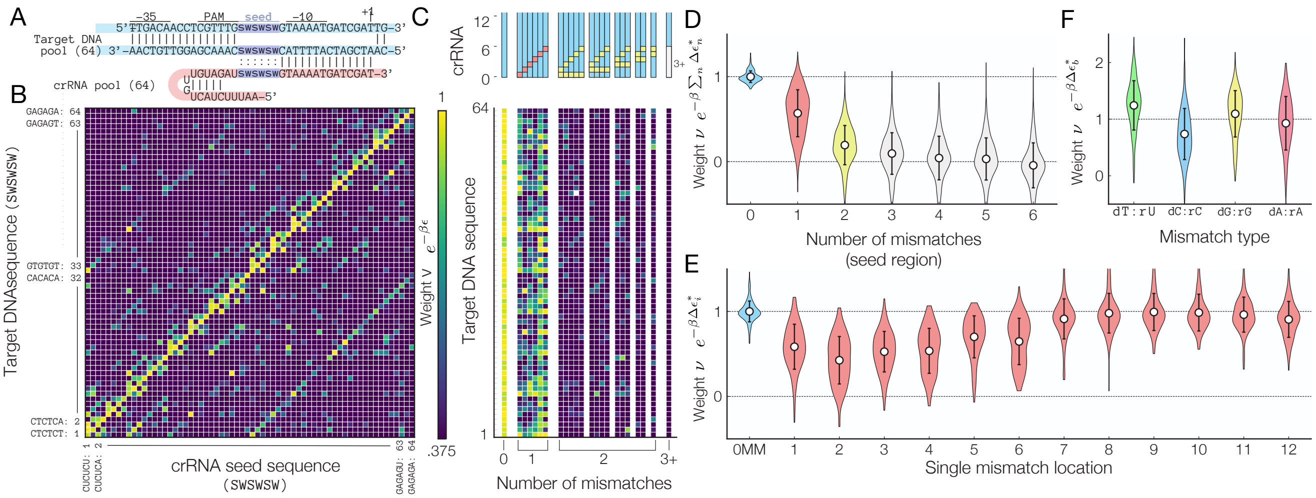


FIG. 5. A) Sequence of the multiplexed mismatch assays for Cas12's seed region. B) Cross-talk map of the reconfiguration rate for (SWSWSW)x(SWSWSW) data subset (full dataset is shown in Fig. S4). Note that the off-diagonal elements represent crRNA-DNA target that differ by a single mismatch. C) Target DNA dependent reconfiguration rate for SWSWSW sequences containing 0, 1, 2, or 3+ mismatches. D) Aggregated ν for crRNA-target DNA combinations containing between 1 to 6 mismatches. E) Aggregated ν for crRNA-target DNA combinations containing a single mismatches. F) The reconfiguration rate ν for base-specific mismatches. Note that dT:rU and dG:rG mismatches are tolerated at a higher level than dC:rC and dA:rA mismatches. Error bars = std.dev.

ergetic impacts of individual mismatches are additive, and location-dependent binding energy costs reported in Fig. 4F should be incorporated to models that aim to predict off-target binding.

High throughput cross-talk assays reveal position- and nucleotide-specific energy costs

We next asked how both crRNA and DNA variations in the first six bases of the PAM-proximal seed region affected the reconfiguration rate ν. We performed multiplexed CRISPRi assays using two oligo pools, each containing 128 different sequences, to test the pairing between all possible crRNA-DNA sequences of the form SWSWSW or WSWSWS in a single step. Once again, those pairings were chosen to maintain all crRNA-DNA sequences at a fixed GC content. This approach covers a large combinatorial space between the spacer-target sequences and produces a comprehensive cross-talk map between 16,384 possible crRNA-DNA combinations (Fig. 5A). While we also performed the same analysis on the crRNA-DNA "trunk" region (Fig. S6), only the SW quadrant of the seed region is shown in Fig. 5B (see Figs. S4-S6 for the full cross-talk maps).

The cross-talk maps show that fully matching crRNA-DNA sequences (i.e. those along the main diagonal of Fig. 5B and in the first column in Fig. 5C) have the highest ν. Interestingly, the reconfiguration rate ν for all fully-matched crRNA-DNA targets fall within a very narrow range of 1.00 ± 0.06 (mean ± std.dev.), suggesting that the specific base composition of the seed region does not have a large impact on DNA binding. This contrasts with *in vivo* multiplexed DNA cleavage assays for Cas12a variants that *do* show significant sequence dependence on cleavage activity [15, 19, 20]. In addition, while SpCas9 binding and cleavage activity has different sequence specificities [12–14], we do not observe any significant discrepancies between the binding and cleavage assays performed using catalytically-active FnCas12a nuclease (Fig. S7). Hence, we believe our approach may provide a more accurate representation of dCas12a's binding energy landscape because our approach excludes any source of variation caused by unknown cellular physiological factor by only investigating a small but comprehensive portion of all possible crRNA-target DNA sequences that possess the same GC content.

To further understand how single mismatches affect the reconfiguration rate, we considered how ν varies as a function of the number and location of mismatches present. First, we show in Fig. 5D that no significant binding observed for sequences containing more than 4 mismatches in the seed region. Our analysis, however, reveals that formation of a stable ternary complex does occur in the presence of 1, 2 or 3 mismatches (P = 1 x 10^{-232}, 8 x 10^{-94}, and 1 x 10^{-15}, respectively; null-hypothesis=no binding will occur for 1, 2, of 3 mismatches). It is important to note that by performing aggregate measurement across thousands of crRNA and DNA sequences, our results confers a much stronger statistical predictive power than other assays that only test a limited number of crRNA-DNA partners. In addition, we also show in Fig. 5E that mismatches have the greatest impact when located within the first 6 bases of the seed region. Sensitivity to a mismatch decreases with distance from the PAM site, and mismatches located in the trunk region (bases 6-12) only minimally impact DNA binding.

We next considered whether the type of mismatch affects ν in Fig. 6F. Surprisingly, we find that single crRNA-DNA mismatches of the form dC:rC decrease ν by an additional 26% on average. In contrast, dT:rU and dG:rG mismatches are tolerated and increase the reconfiguration rate by 9.5% and 24% compared to all types of single-base mismatch, respectively. This effect can be visualized in Fig. 6B, where off-diagonal elements that correspond to a single mismatch in the sixth location are more prominent in the lower right quadrant than those in the upper left quadrant (the upper left quadrant corresponds to a dC:rC mismatch while the lower right corresponds to dG:rG mismatches). Insensitivity to wobble-transition mismatch has been previously reported in SpCas9 [21, 46] and AsCas12a [19], but other work in AsCas12a found no significant effect due to a transversion mismatch [19], suggesting tolerance to transversion mismatches may be unique to FnCas12a.

DISCUSSION

We have established that massively parallel CRISPRi assays, with their ability to rapidly measure thousands of different crRNA-target DNA variants in parallel, are a viable method to assess dCas12 binding efficiencies. Our results reveal the fundamental relationship between crRNA-DNA interactions and the underlying energy landscape that dictates binding behavior of dCas12. One major outcome of this study is that binding of DNA by CRISPR-Cas12a endonuclease does not strongly depend on the specific crRNA sequence used (at least within the set of tested sequences which were kept at 50% GC content). Rather, variance in DNA binding affinities depends on the PAM sequence, the presence of mismatches, and the type of mismatch present. Indeed, the propensity of identical DNA targets to be recognized by a CRISPR-Cas nuclease matching crRNA may be significantly different depending on their respective 6-base PAM sequence. Similarly, the absolute number of mismatches in the seed region of a crRNA-DNA hybrid is more important than their specific location, and mismatches that occur in the distal region of a crRNA (i.e. after base 17) do not significantly affect binding affinity. Our results also show that dT:rU and dG:rG mismatches are tolerated at a higher level than dA:rA and dC:rC mismatch.

Beyond that, the power of our approach also resides in our ability to use a parameter-free statistical mechanics framework to extract thermodynamic determinants of dCas12a binding. Importantly, our results are not specific to nuclease-dead CRISPR-Cas endonucleases –we confirm in Fig. S7 that the same behavior is observed for *catalytically-active* Cas12a nuclease– and our approach should foster the development of predictive, parameter-free biophysical models of on- and off-target binding affinities and DNA cleavage activities. In addition, because CRISPR-Cas systems are very common amongst prokaryotes [1], there is a need for the rapid and efficient characterization of newly-sequenced CRISPR-Cas systems that may display enhanced target differentiation capabilities or alternative PAM site compositions. We anticipate that this method will also provide a mechanistic understanding of the thermodynamic determinants of DNA target recognition and binding affinities in uncharacterized CRISPR-Cas endonucleases and other nucleic-acid binding enzymes.

Because our method is applicable to both the catalytically active and dead versions of the nuclease, it should also lead to improvements in a vast range of CRISPR applications, including *in vivo* gene editing, programmable repression, and nucleic acid detection. Our multiplexed approach is particularly applicable to the advancement of dCas-based gene circuit elements, which can be been used to create complex circuits that behave orthogonally, operating independently without crosstalk [47–51]. Furthermore, our approach can expedite the rational design of enhanced CRISPR nucleases and facilitate the development of CRISPR-Cas variants with greater specificity, improved proofreading capabilities, or increased activities [52–57].

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

Conceptualization, G.L. and D.A.S.; Methodology, D.A.S. and G.L.; Formal Analysis, D.A.S. and G.L.; Investigation, D.A.S., Y.X., and G.L.; Writing – Original Draft, G.L. and D.A.S.; Writing – Review & Editing, G.L., D.A.S., and Y.X.; Supervision, G.L.;

DECLARATION OF INTERESTS

The authors declare no competing interests.

[1] Mohanraju, P. *et al.* Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science* **353**, aad5147 (2016).

[2] Barrangou, R. & Doudna, J. A. Applications of CRISPR technologies in research and beyond. *Nature Biotechnology* **34**, 933–941 (2016).

[3] Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**, 759–771 (2015).

[4] Zetsche, B. *et al.* Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. *Nature Biotechnology* **35**, 31–34 (2017).

[5] Qi, L. S. *et al.* Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152**, 1173–1183 (2013).

[6] Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Research* **41**, 7429–7437 (2013).

[7] Kim, S. K. *et al.* Efficient Transcriptional Gene Repression by Type V-A CRISPR-Cpf1 from Eubacterium eligens. *ACS synthetic biology* **6**, 1273–1282 (2017).

[8] Chen, J. *et al.* CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439 (2018).

[9] Gootenberg, J. S. *et al.* Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science* **360**, 439–444 (2018).

[10] Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016). URL `https://www.nature.com/articles/nature17946`.

[11] Gaudelli, N. M. *et al.* Programmable base editing of AT to GC in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017). URL `https://www.nature.com/articles/nature24644`.

[12] Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nature Biotechnology* **32**, 677–683 (2014).

[13] Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature Biotechnology* **32**, 670–676 (2014).

[14] O'Geen, H., Henry, I. M., Bhakta, M. S., Meckler, J. F. & Segal, D. J. A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. *Nucleic Acids Research* **43**, 3389–3404 (2015).

[15] Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nature Biotechnology* **32**, 1262–1267 (2014).

[16] Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34**, 184–191 (2016).

[17] Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology* **17**, 148 (2016).

[18] Tycko, J., Myer, V. & Hsu, P. Methods for Optimizing CRISPR-Cas9 Genome Editing Specificity. *Molecular Cell* **63**, 355–370 (2016).

[19] Kim, H. K. *et al.* *In vivo* high-throughput profiling of CRISPRCpf1 activity. *Nature Methods* **14**, 153–159 (2017).

[20] Kim, H. K. *et al.* Deep learning improves prediction of CRISPRCpf1 guide RNA activity. *Nature Biotechnology* **36**, 239–241 (2018).

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

[22] Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proceedings of the National Academy of Sciences* **111**, 9798–9803 (2014).

[23] Hua Fu, B. X., Hansen, L. L., Artiles, K. L., Nonet, M. L. & Fire, A. Z. Landscape of target:guide homology effects on Cas9-mediated cleavage. *Nucleic Acids Research* **42**, 13778–13787 (2014).

[24] Singh, D. *et al.* Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1 (Cas12a). *Proceedings of the National Academy of Sciences of the United States of America* **115**, 5444–5449 (2018).

[25] Duan, J. *et al.* Genome-wide identification of CRISPR/Cas9 off-targets in human genome. *Cell Research* **24**, 1009–1012 (2014).

[26] Kleinstiver, B. P. *et al.* Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nature Biotechnology* **34**, 869–874 (2016).

[27] Kim, D. *et al.* Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nature Biotechnology* **34**, 863–868 (2016).

[28] Singh, R., Kuscu, C., Quinlan, A., Qi, Y. & Adli, M. Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. *Nucleic Acids Research* **43**, e118 (2015).

[29] Xu, H. *et al.* Sequence determinants of improved CRISPR

sgRNA design. *Genome Research* **25**, 1147–1157 (2015).

[30] Wong, N., Liu, W. & Wang, X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biology* **16**, 218 (2015).

[31] Zhang, D., Hurst, T., Duan, D. & Chen, S.-J. Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design. *Proceedings of the National Academy of Sciences* **116**, 8693–8698 (2019).

[32] Wang, T. *et al.* Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. *Nature Communications* **9**, 2475 (2018).

[33] Guo, J. *et al.* Improved sgRNA design in bacteria via genome-wide activity profiling. *Nucleic Acids Research* **46**, 7052–7069 (2018).

[34] Marshall, R. *et al.* Rapid and Scalable Characterization of CRISPR Technologies Using an E. coli Cell-Free Transcription-Translation System. *Molecular Cell* **69**, 146–157.e3 (2018).

[35] Boyle, E. A. *et al.* High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 5461–5466 (2017).

[36] Jeon, Y. *et al.* Direct observation of DNA target searching and cleavage by CRISPR-Cas12a. *Nature Communications* **9**, 2777 (2018).

[37] Stella, S. *et al.* Conformational Activation Promotes CRISPR-Cas12a Catalysis and Resetting of the Endonuclease Activity. *Cell* **175**, 1856–1871.e21 (2018).

[38] Brewster, R. *et al.* The Transcription Factor Titration Effect Dictates Level of Gene Expression. *Cell* **156**, 1312–1323 (2014).

[39] Weinert, F. M., Brewster, R. C., Rydenfelt, M., Phillips, R. & Kegel, W. K. Scaling of Gene Expression with Transcription-Factor Fugacity. *Physical Review Letters* **113**, 258101 (2014).

[40] Landman, J., Brewster, R. C., Weinert, F. M., Phillips, R. & Kegel, W. K. Self-consistent theory of transcriptional control in complex regulatory architectures. *PLOS ONE* **12**, e0179235 (2017).

[41] Jones, D. L. *et al.* Kinetics of dCas9 target search in Escherichia coli. *Science* **357**, 1420–1424 (2017).

[42] Cleveland, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* **74**, 829–836 (1979).

[43] Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**, R29 (2014).

[44] Li, X.-t., Thomason, L. C., Sawitzke, J. A., Costantino, N. & Court, D. L. Positive and negative selection using the tetA-sacB cassette: recombineering and P1 transduction in Escherichia coli. *Nucleic Acids Research* **41**, e204 (2013).

[45] Tóth, E. *et al.* Mb- and FnCpf1 nucleases are active in mammalian cells: activities and PAM preferences of four wild-type Cpf1 nucleases and of their altered PAM specificity variants. *Nucleic Acids Research* **46**, 10272–10285 (2018).

[46] Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology* **33**, 187–197 (2015).

[47] Jusiak, B., Cleto, S., Perez-Piñera, P. & Lu, T. K. Engineering Synthetic Gene Circuits in Living Cells with CRISPR Technology. *Trends in Biotechnology* **34**, 535–547 (2016).

[48] Didovyk, A., Borek, B., Tsimring, L. & Hasty, J. Transcriptional Regulation with CRISPR-Cas9: Principles, Advances, and Applications. *Current opinion in biotechnology* **40**, 177–184 (2016).

[49] Didovyk, A., Borek, B., Hasty, J. & Tsimring, L. Orthogonal Modular Gene Repression in Escherichia coli Using Engineered CRISPR/Cas9. *ACS synthetic biology* **5**, 81–88 (2016).

[50] Nielsen, A. A. K. & Voigt, C. A. Multi-input CRISPR/Cas genetic circuits that interface host regulatory networks. *Molecular Systems Biology* **10**, 763 (2014).

[51] Cress, B. F. *et al.* Rapid generation of CRISPR/dCas9-regulated, orthogonally repressible hybrid T7-lac promoters for modular, tuneable control of metabolic pathway fluxes in Escherichia coli. *Nucleic Acids Research* **44**, 4472–4485 (2016).

[52] Chen, J. S. *et al.* Enhanced proofreading governs CRISPRCas9 targeting accuracy. *Nature* **550**, 407–410 (2017).

[53] Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).

[54] Kleinstiver, B. P. *et al.* High-fidelity CRISPRCas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).

[55] Casini, A. *et al.* A highly specific SpCas9 variant is identified by *in vivo* screening in yeast. *Nature Biotechnology* **36**, 265–271 (2018).

[56] Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).

[57] Kleinstiver, B. P. *et al.* Engineered CRISPRCas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. *Nature Biotechnology* **37**, 276 (2019).