scientific reports

Check for updates

OPEN

PAM-adjacent DNA flexibility tunes CRISPR-Cas12a off-target binding

Aleique Allen[1], Brendon H. Cooper[2,5], Jaideep Singh[1], Remo Rohs[1,2,3,4] & Peter Z. Qin[1]✉

Cas12a is a class 2 type V CRISPR-associated nuclease that uses an effector complex comprised of a single protein activated by a CRISPR-encoded small RNA to cleave double-stranded DNA at specific sites. Cas12a processes unique features as compared to other CRISPR effector nucleases such as Cas9, and has been demonstrated as an effective tool for manipulating complex genomes. Prior studies have indicated that DNA flexibility at the region adjacent to the protospacer-adjacent-motif (PAM) contributes to Cas12a target recognition. Here, we adapted a SELEX-seq approach to further examine the connection between PAM-adjacent DNA flexibility and off-target binding by Cas12a. A DNA library containing DNA-DNA mismatches at PAM + 1 to + 6 positions was generated and subjected to binding in vitro with FnCas12a in the absence of pairing between the RNA guide and DNA target. The bound and unbound populations were sequenced to determine the propensity for off-target binding for each of the individual sequences. Analyzing the position and nucleotide dependency of the DNA-DNA mismatches showed that PAM-dependent Cas12a off-target binding requires unpairing of the protospacer at PAM + 1 and increases with unpairing at PAM + 2 and + 3. This revealed that PAM-adjacent DNA flexibility can tune Cas12a off-target binding. The work adds support to the notion that physical properties of the DNA modulate Cas12a target discrimination, and has implications on Cas12a-based applications.

The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas (CRISPR-associated proteins) system is a programmable immune system natively occurring in bacteria and archaea[1,2]. Among the many types of CRISPR-Cas, the class 2 type V system cleaves DNA duplexes at specific sites using an RNA-guided endonuclease that is comprised of a single protein, Cas12a, together with a CRISPR RNA (crRNA)[3]. The Cas12a system has unique features as compared to other CRISPRs such as Cas9[4,5], and has been adapted for genome editing[6,7], gene regulation[7,8], in vivo imaging[4], and nucleic acid detection[6].

Since the discovery of Cas12a in 2015[3], enormous progress has been made in elucidating its functional mechanisms[5,7,9]. For its "cis-activity" of cleaving a double-stranded DNA, a cognate target (i.e., "on-target") must meet two requirements: (1) a segment of the DNA duplex, referred to as "protospacer", that is complementary to the crRNA guide; and (2) a short DNA sequence flanking the protospacer, which is denoted as the protospacer adjacent motif (PAM)[3]. Cas12a identifies an on-target DNA by first recognizing the PAM, then unwinds the protospacer to form a three-stranded R-loop, in which the target-strand (ts) of the DNA hybridizes with the crRNA guide, while the non-target-strand (nts) was rendered single-stranded. Studies have uncovered a number of conformational states during Cas12a target interrogation, including PAM binding and local distortion of the PAM-adjacent protospacer[5,9–13], initiation of R-loop at the PAM-proximal "seed" segment[5,9,14], propagation of the R-loop[5,9,15], and R-loop dependent DNA strand cleavage[5,9]. Dynamic equilibria between these states serve as checkpoints for Cas12a to discriminate between correct and incorrect targets[16–18]. However, despite these checkpoints, Cas12a is capable of binding and cleaving off-target DNAs that present mismatch(es) between the DNA protospacer and the crRNA guide[19–22]. These off-targets have undesired effects that can have serious implications in genetic engineering. This includes off-target cleavage resulting in gene disruption at undesired locations[19–22] as well as off-target binding disrupting CRISPR-based imaging[23–25], transcriptional regulation[26–28] and base editing[29–32]. Mechanistic investigations on Cas12a have been critical for understanding and minimizing the off-targets[33–35], and have led to the development of high fidelity variants of Cas12a[34,36,37].

Given that a crucial early step in Cas12a target interrogation is to bind the DNA by kinking and unwinding the DNA duplex[10–13], the physical properties of the unbound DNA duplex, which dictates its response to distortions,

[1]Department of Chemistry, University of Southern California, 3430 S Vermont Ave., Los Angeles, CA 90089, USA. [2]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA. [3]Department of Physics & Astronomy, University of Southern California, Los Angeles, CA 90089, USA. [4]Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA. [5]Present address: Beckman Coulter, 1584 Enterprise Blvd, West Sacramento, CA 95691, USA. ✉email: pzq@usc.edu

are expected to contribute significantly to specificity. A number of studies have investigated the impact of DNA physical properties such as topology on off-target binding[21,38–41], including a recent report on Cas12a binding and cleavage of biologically-relevant branched DNA configurations[40]. In previous work, we have shown that off-targets lacking complementarity between the crRNA guide and the DNA target-strand can be bound by Cas12a if one increases the DNA duplex flexibility by introducing unpaired nucleotides next to the PAM[42]. However, limited by the biochemical approaches used[42], the study examined a limited number of DNA sequences, and was not able to further elucidate the position and base preferences for off-target binding.

In this study, we further examined promiscuous Cas12a off-target binding of flexible DNA by adapting a Systematic Evolution of Ligands by Exponential Enrichment with DNA sequencing techniques (SELEX-seq) (Fig. 1)[43–48]. Building on constructs containing DNA-DNA mismatch(es) (i.e., "DNA bubble") that have been employed successfully to study mechanisms of Cas12a[14,42] and Cas9[49], a DNA duplex library was generated with a Cas12a PAM and randomized DNA-DNA mismatches at the PAM + 1 to + 6 segment of the protospacer (Fig. 1). The library was subjected to binding to a Cas12a/crRNA effector that lacks complementarity between its RNA guide and the target-strand of the DNAs, and bound and unbound DNAs were sequenced to determine the propensity for off-target binding for each of the individual sequences. The data revealed that favorable off-target binding depends on the presence of a properly formed PAM and unpairing of the protospacer at the PAM + 1–3 region. The positional and nucleotide dependencies of off-target binding clearly support the notion that PAM-adjacent DNA duplex flexibility facilitates Cas12a binding. The work adds support to the notion that physical properties of the DNA (in this case flexibility) influence Cas12a target discrimination. This has implications in Cas12a-based applications, particularly those relying on Cas12a binding to specific DNA sites such as imaging and transcriptional regulation.

Results

Adaptation of SELEX-seq to study off-target binding by Cas12a

A modified version of SELEX-seq[43–45] was adapted to study off-target DNA binding by Cas12a (Fig. 1, see Methods). A DNA duplex library was designed based on a Cas12a target site in the DNMT1 gene (NC_000019.10 (10133346.10194953, complement)), which contains a TTTG PAM. Building on our previous finding that off-target binding increases as the nominal PAM-adjacent DNA-DNA mismatch increases from 0 to 3 base-pairs and saturates with 4 base-pairs[42], a target library was constructed with a constant target-strand (ts) sequence, while the non-target-strand (nts) had randomized sequences presented at the 6 protospacer positions adjacent to the PAM [Fig. 1, Supporting Information (SI) Sect. S1.1]. Hybridizing ts and nts generated a library of 4,096 duplexes with variable DNA-DNA mismatch(es) presented at the PAM + 1 to + 6 positions. Note that except for the TTTG site, the DNA duplexes do not contain any TTTN sequence that may serve as an alternative PAM for FnCas12a recognition. The DNA library presents no complementarity between the crRNA guide and the protospacer at the DNA target-strand (Fig. 1). Previous studies have shown that this construct does not elicit

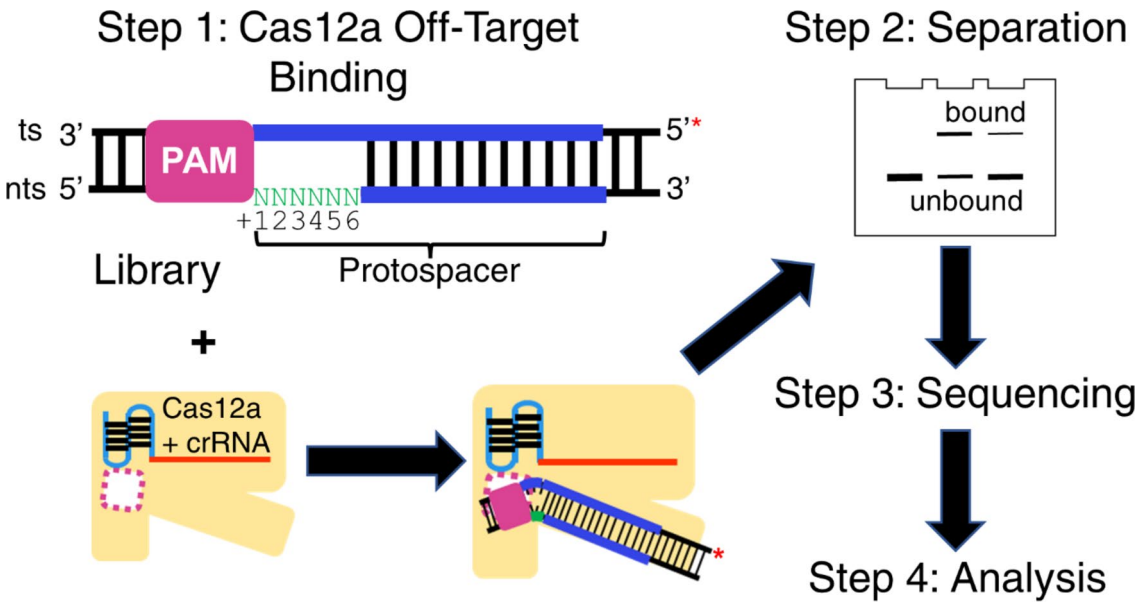

Fig. 1. A schematic of SELEX-seq adapted for analyzing Cas12a binding to a library of DNA duplexes containing DNA-DNA mismatch(es). The duplexed DNA library contains a FnCas12a PAM and randomized DNA-DNA mismatch(es) at the PAM + 1 to + 6 positions. "N" indicates the randomized nucleotides at the non-target strand and "*" indicates a 5' fluorescence tag (5'-FAM) at the target strand. The library was subjected to binding to a Cas12a effector with an RNA guide (red line) that does not complement the DNA protospacer (blue lines) (Step 1). The bound and unbound DNA populations were separated using a native gel shift assay (Step 2). The individual DNAs in the corresponding populations were identified by sequencing (Step 3). The sequencing data were analyzed to obtain the propensity of binding for each individual DNA (Step 4).

DNA cleavage by Cas12a, but instead serves as a system to investigate Cas12a off-target binding of DNA[42]. The library was validated by control experiments with catalytically active Cas12a, which demonstrated that the DNA library was cleaved with an on-target crRNA guide but not with the off-target guide (SI Sect. S1.2, Fig. S2).

To study off-target binding by Cas12a, the library was subjected to binding with a catalytically-inactive dFnCas12a effector complex. Consistent with a prior report[42], binding of the library with the off-target dFnCas12a effector (i.e., containing an RNA guide that is completely non-complementary to the target-strand of the protospacer), while measurable, was weaker than with the on-target effector (SI Sect. S1.3, Fig. S3). Selection was then carried out with 50 nM effector and 100 nM total DNA duplexes, and the bound DNA and unbound DNA populations were separated using a native gel shift assay (Fig. 2A), then recovered and sequenced (see Methods, SI Sect. S1.3 and S.2.1). In addition, the unbound DNA library was sequenced to assess variations within the starting library. In this work, three replicas were sequenced and analyzed individually, each with sufficient quality to support the analysis performed (SI Sect. S2.2).

Figure 2 shows analysis of the individual unique sequences from a representative dataset (designated as dataset 3 in "Supplementary Data.xlsx"). For each unique sequence, its respective proportions in the bound, unbound, and library dataset were used to obtain the bound enrichment (E_B^i, Eq. 2) and unbound enrichment (E_U^i, Eq. 3) (Fig. 2B), from which the corresponding relative enrichment r_i (Eq. 4) was computed (Fig. 2B, also see SI Sect. S2.3). The r_i values measure the relative proportion of a given sequence at the Cas12a-bound and unbound states, with $r_i > 0$ indicating that the bound proportion is larger than that of the unbound, thus a preference for

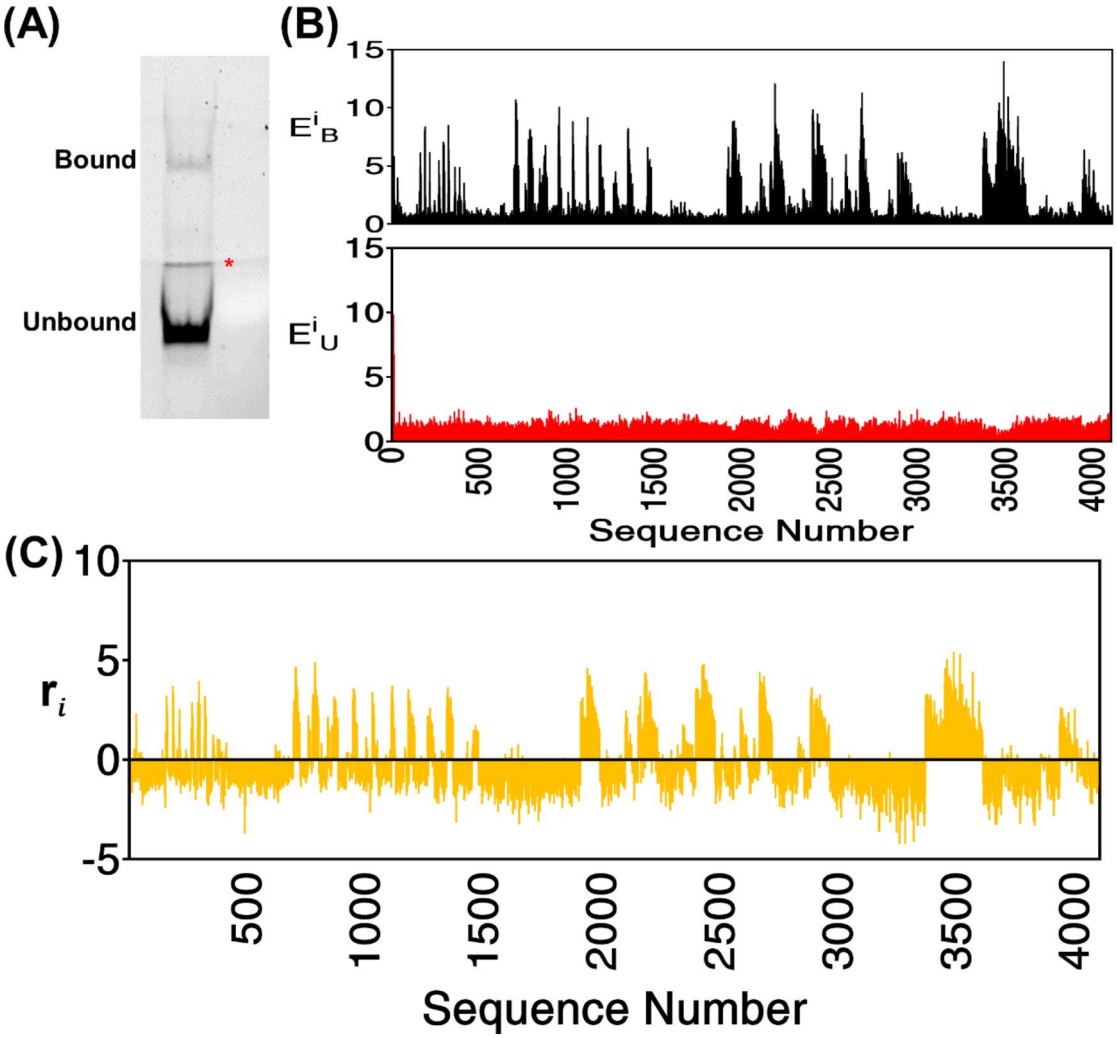

Fig. 2. Example of a SELEX-seq dataset. (**A**) An example of separation of the bound and unbound DNA. As shown binding was carried out with 100 nM duplexed DNA library and 50 nM dCas12a effector complex (formed with a protein/RNA ratio of 1:2) (see Methods). Note that to account for the weak off-target binding, the effector concentration was set much higher than the reported 12 nM on-target K_d of FnCas12a[14]. The bound and unbound DNAs were separated on a native polyacrylamide gel with a 5% and a 10% segment and visualized with fluorescence imaging. "*" indicated an imaging artifact at intersections of the gel segments. (**B**) Enrichment analysis of the bound (E_B^i, Eq. 2) and unbound (E_U^i, Eq. 3) populations recovered from the gel shown in panel (**A**). (**C**) Relative enrichment values, r_i (Eq. 4), calculated for data shown in panel (**B**).

Cas12a binding. Of the 4,095 possible unique sequences that contain DNA-DNA mismatch(es), the observed r_i for this dataset ranged from − 4.25 to 5.44 (SI Sect. S2.4, Table S8). 30.4% of the sequences preferred to be bound (i.e., $r_i > 0$), while 69.6% preferred to be unbound (i.e., $r_i < 0$). Given that the DNA target-strand was fixed and had no complementarity with the RNA guide, the smaller size of the bound population as compared to that of the unbound is consistent with the expected weak binding. Similar variations in the r_i values were observed in the other two replicas (SI Sect. S2.4). Together the data clearly indicates that the designed DNA library captures variation in Cas12a off-target binding. In addition, corresponding r_i values from different replicas show high correlations (SI Sect. S2.4, Fig. S4). Consequently, conclusions drawn from dataset 3 are consistent with those obtained from the other two datasets (see example in SI Sect. S3.2.1, Fig. S5). In the following sections, analysis and conclusions drawn from dataset 3 are presented.

Cas12a off-target binding correlates with the total number of DNA-DNA mismatches

To reveal DNA features that dictate Cas12a off-target binding, we first analyzed the correlation between the preference for binding and the total number of DNA-DNA mismatches in our construct (Fig. 3A). The DNA sequences were categorized into groups based on the total number of mismatches presented, and their relative enrichment ($r(\#MM)$) was computed as the ratio of the weighted average enrichment between the bound and unbound datasets (Eq. 5). For example, "1MM" includes sequences with one mismatch at any position along the PAM + 1 to + 6 segment, and its relative enrichment, $r\,(1\mathrm{MM})$, was found to be -0.86 (Fig. 3, also see SI Sect. S3.1, Table S9).

Figure 3 shows the dependence of $r(\#MM)$ on the total number of mismatches. The analysis revealed that $r\,(1\mathrm{MM})$ and $r\,(2\mathrm{MM})$ are both negative, indicating that as a group, sequences with 1 or 2 mismatches are not preferentially bound (Fig. 3). Sequences with 3 mismatches gave a small positive $r\,(3MM)$, indicating that binding becomes slightly favored (Fig. 3). As the number of mismatches increased to 4 and beyond, $r(\#MM)$ became clearly greater than 1, indicating binding becomes favorable (Fig. 3). This feature is also observed based on the other two datasets (SI Sect. S3.2.1, Fig. S5 and Fig. S6). Overall, this indicates that under the experimental setup reported here, a bubble size of 4 results in preferable binding, and further increasing the bubble size has only small impact on binding. This is in complete agreement with conclusions drawn from a prior biochemical study using specific sequences[42].

Off-target binding depends on mismatches located at the PAM-adjacent protospacer positions

Next, we examined the correlation between the preference for binding and the location of DNA-DNA mismatch(es) (Fig. 4). The sequences were grouped based on their DNA-DNA mismatch position(s) relative to the PAM, and the relative enrichment, $r\,(G_{feat})$, was computed for each group (see Eq. 5). For example,

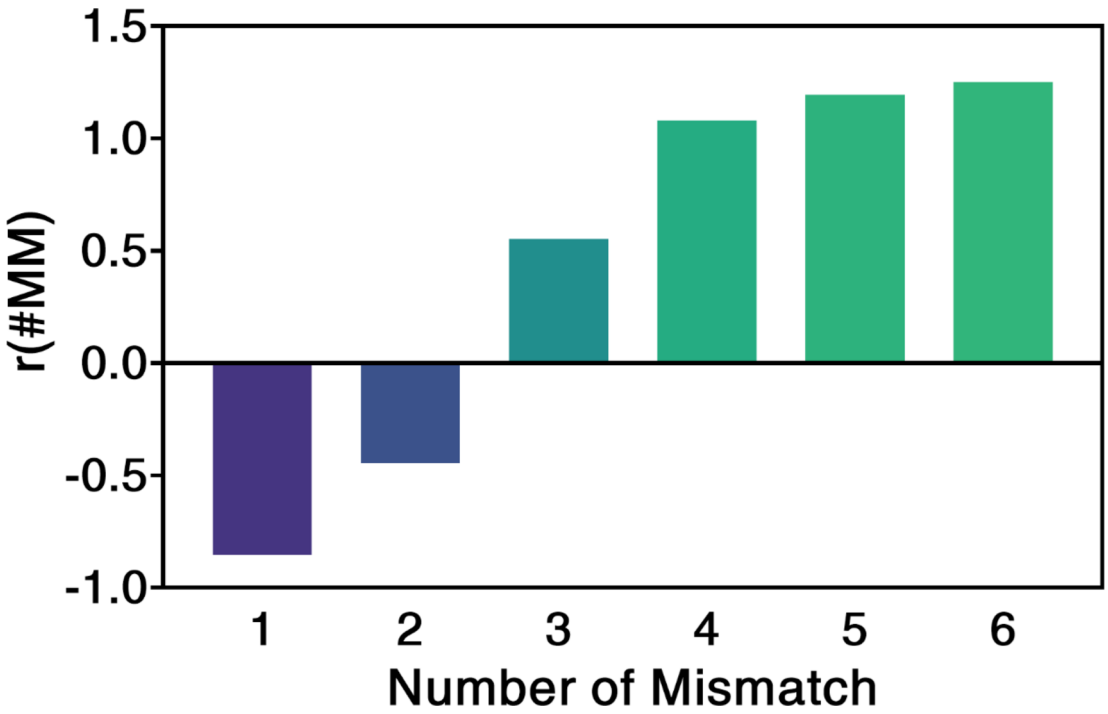

Fig. 3. Correlation between off-target binding and the total number of mismatches. The relative enrichment for groups of sequences containing a given total number of mismatches (r(#MM)) was plotted against the total number of mismatches. The data shows off-target binding increases with the total number of PAM-adjacent mismatches. See additional information in SI Sect. S3.2.1.

the group of DNA targets with 6 mismatches (i.e., 6MM, Fig. 3) included all sequences nominally containing mismatches at PAM + 1, 2, 3, 4, 5, and 6, and the resulting $r(1,2,3,4,5,6) = r(6\mathrm{MM}) = 1.25$ (Fig. 3).

Targets with five DNA-DNA mismatches (i.e., 5MM) fall into six sub-groups with different arrangements of mismatch position (Fig. 4A). As a group, 5MM sequences are favored to bind (i.e., $r(5\mathrm{MM}) > 0$, Fig. 3), however, enrichment for binding (i.e. $r(G_{feat}) > 0$) was observed for only five of the six groups that contain a DNA mismatch at the PAM+1 position (Fig. 4A). The PAM+2,3,4,5,6 group, which has the DNA-DNA paired at the PAM + 1, gave an $r(2,3,4,5,6) < 0$ (Fig. 4A, marked as "#1"), indicating that they are not favored to bind. The same feature is observed from the other two datasets (SI sect. S3.2.2., Fig. S7). Together, these indicate that DNA-DNA mismatch at the PAM + 1 position is required for favorable off-target binding.

Furthermore, amongst the five sub-groups favored for off-target binding (i.e., $r(G_{feat}) > 0$), $r(G_{feat})$ value of the PAM+1,2,3,5,6 group (Fig. 4A, "#4") is larger than that of the PAM + 1,3,4,5,6 and PAM + 1,2,4,5,6 groups (Fig. 4A, "#2" and "#3", respectively), and this is deemed significant when considering all three datasets (SI, sect. S3.2.2). Note that the number of "consecutive" PAM-adjacent mismatches are nominally 3, 2, and 1, respectively, for the PAM + 1,2,3,5,6, PAM + 1,2,4,5,6, and PAM + 1,3,4,5,6 groups. That $r(1,2,3,5,6)$ has the highest value thus indicates that increasing size of the PAM-adjacent consecutive DNA bubble favors binding. However, further increasing the nominal consecutive mismatches from 3 (i.e., PAM + 1,2,3,5,6) to 4 (i.e., PAM + 1,2,3,4,6) and 5 (i.e., PAM + 1,2,3,4,5) gives lower $r(G_{feat})$ values in dataset 3 (Fig. 4A, "#4", #5", and "#6"), and analysis of all three datasets indicate these $r(G_{feat})$ values are not significantly different (SI, Sect. S3.2.2). To understand the cause of this observation, the expected most-stable secondary structure of each DNA duplex was predicted based on its computed folding energy (ΔG_i, see Method). Analysis on the most stable secondary structures of the 5MM DNAs showed that the PAM + 1,2,3,4,6 subgroup ("#5", Fig. 4A), which nominally forms four consecutive mismatches, does not favor the 4–4 (i.e., 4-nt at ts and 4-nt at nts) PAM-adjacent bubble, but instead predominately adopts 3–3 (25.9%) and 2–4 (14.8%) bubbles (see examples in Fig. 4B). Similarly, the PAM + 1,2,3,4,5 subgroup ("#6", Fig. 4A), which nominally forms five consecutive mismatches, does not favor 5–5 bubbles, but predominately adopts 5 – 0 (19.8%) and 3 – 2 (14.4%) bubbles (Fig. 4B). As such, PAM-adjacent bubbles favored by both subgroups are comparable or slightly more constrained (i.e., smaller) than the 3–3 bubble adopted predominately by the PAM + 1,2,3,5,6 subgroup (59.3%, Fig. 4B). This likely accounts for the lack of $r(G_{feat})$ increases from PAM + 1,2,3,5,6 to PAM + 1,2,3,4,6 and PAM + 1,2,3,4,5 (Fig. 4A, Fig. S7). Overall, the analysis indicates that in addition to the PAM + 1 mismatch requirements, mismatches at PAM+2 and + 3 contribute the most to off-target binding. This is consistent with prior conclusions drawn from biochemical analysis of a set of specific off-target DNAs[42].

SELEX-seq results also revealed that a small number of 5MM DNAs with a nominal PAM + 1 mismatch has negative r_i values (see examples in Fig. 4C), indicating that they are not favored for off-target binding. Analysis of the most-stable secondary structure revealed that these DNAs fall into two classes. One class favors a dG/dT wobble at PAM + 1 (Fig. 4C), thus is analogous to the "PAM + 2,3,4,5,6" subgroup with a PAM + 1 pairing ("#1",

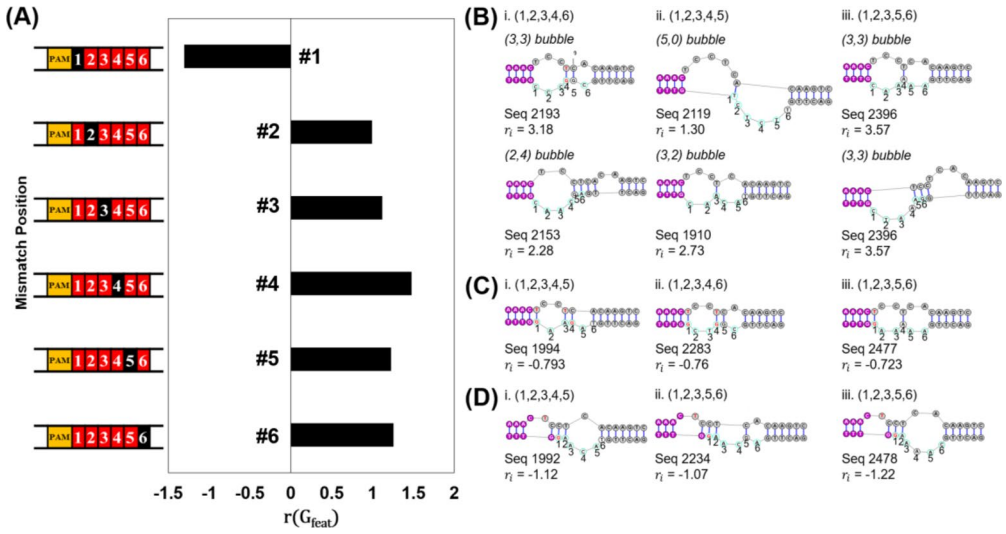

Fig. 4. Positional preference analysis of the 5MM group of sequences. (**A**) $r(G_{feat})$ for different positional mismatch arrangements, with PAM identified by a yellow box, mismatched nucleotides by a red box, and matched nucleotides by a black box. Numbers 1–6 represent the position of the randomized nucleotides. (**B**) Examples of individual 5MM sequences with positive r_i and therefore favor off-target binding. (**C**) Examples of individual 5MM sequences that form a dT/dG$_1$ wobble pair. (**D**) Examples of individual 5MM sequences with misfolded PAM. For panels (**B**), (**C**), and (**D**), the individual sequence number, nominal DNA-DNA mismatch pattern, r_i, and ΔG_i values are indicated. The predicted most stable secondary structures are shown, with magenta filled nucleotides indicating PAM, white filled nucleotides indicate the 6 randomized nucleotides, green circled nucleotides indicate mismatched nucleotides, and the PAM + 1 to + 6 positions are marked. See additional information in SI Sect. S3.2.2.

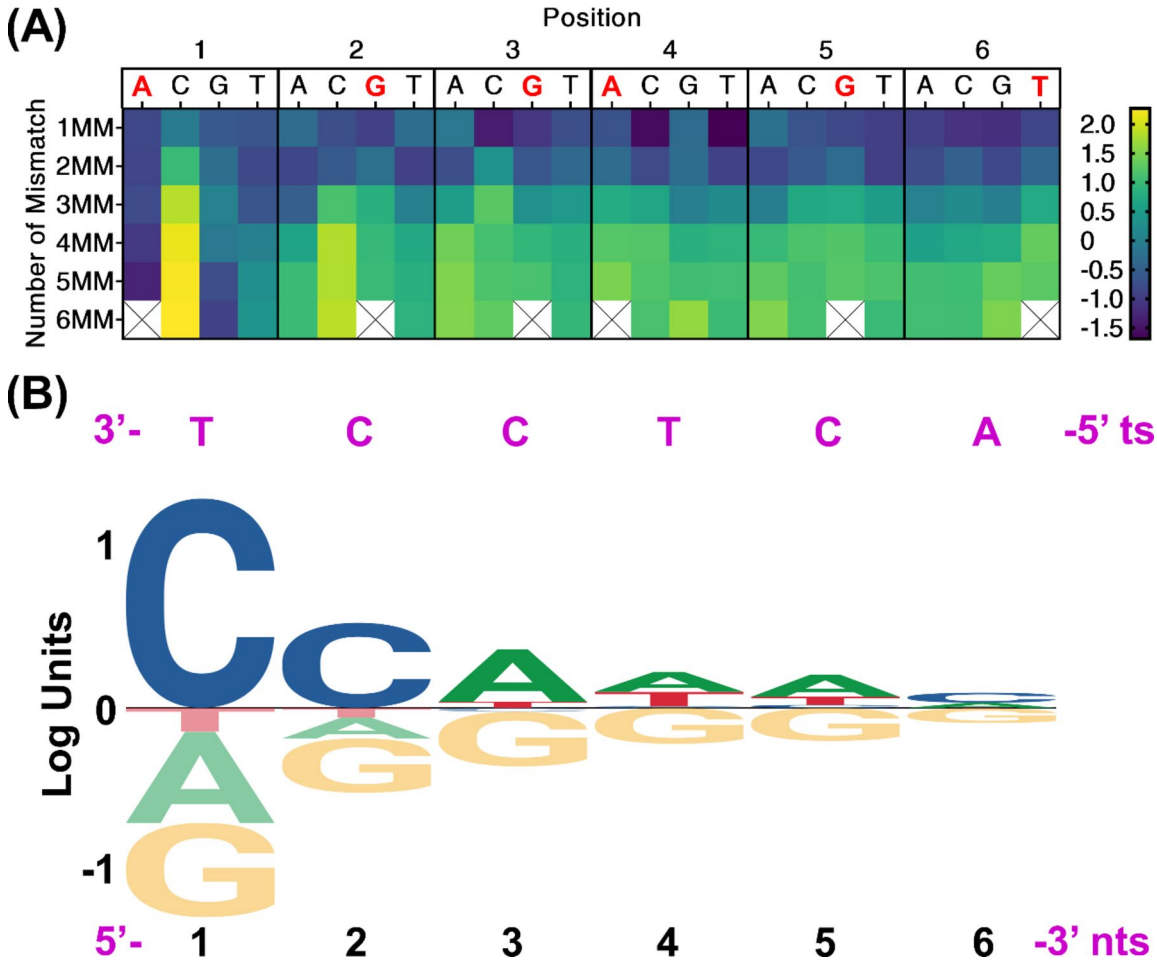

Fig. 5. Analysis of nucleotide preferences. (**A**) $r\,(G_{feat})$ for all possible nucleotides at the randomized positions for different numbers of mismatches represented as a heatmap. (**B**) Web logo plot of position-dependent nucleotide preferences obtained via multiple linear regression (MLR) analysis. Further details are described in SI. Sect. S3.4.

Fig. 4A). In the second class, the most stable fold shifts the mismatches away from the nominally PAM + 1 to + 6 segment and adopts bubble(s) at the PAM (Fig. 4D), thus altering the PAM that is required for proper PAM-Cas12a recognition[50]. The unfavorable binding of sequences with misfolded PAM strongly supports the notion that results drawn from SELEX-seq are not biased by non-specific DNA binding to Cas12a.

Analysis on the 4MM group of DNAs that nominally has a total of four DNA-DNA mismatches yielded conclusions that support all those obtained from the 5MM group (SI Sect. S3.3). The 4MM data show that favorable off-target binding requires a properly folded PAM (SI Sect. S3.3, Fig. S8B) and unpairing of protospacer at PAM + 1 (i.e., excluding dA/dT and dG/dT) (SI Sect. S3.3, Fig. S8A and S8C). In addition, off-target binding increases with DNA-DNA mismatches at PAM + 2 and + 3, with consecutive bubble being the most effective (SI Sect. S3.3, Fig. S8 and Fig. S9).

Analysis of nucleotide preferences for off-target binding

Next, we analyzed the correlation between the nucleotide identities within the PAM + 1 to + 6 randomized segment and preference for Cas12a off-target binding. Values of relative enrichments, $r\,(N_k)_{\#MM}$, were computed for groups of DNA targets with a given nominal total number of mismatches (#MM) and a particular nucleotide (N) at a given position (k) (see example of calculating $r\,(C_1)_{2MM}$ in SI Sect. S3.1, Table S10). The $r\,(N_k)_{\#MM}$ analysis clearly reveals preferences on certain nucleotides at the PAM-adjacent positions (Fig. 5A). At PAM + 1, sequences with an "A" or a "G" at the PAM + 1 position of nts all gave negative $r\,(A_1)_{\#MM}$ and $r\,(G_1)_{\#MM}$ values (Fig. 5A and SI Sect. S3.4, Table S11). For example, among sequences with a total of 5 mismatches, $r\,(A_1)_{5MM}$ = -1.30 and $r\,(G_1)_{5MM}$ = -0.76 (SI Sect. S3.4, Table S11). These sequences form either a "dT/dA" Watson-Crick pair or a "dT/dG" wobble pair at PAM + 1, and their negative r values indicate that they are disfavored for binding, even within the context of a large number of mismatches at the PAM + 2 to + 6 positions (e.g., 4MM and 5MM, Fig. 5A). On the other hand, $r\,(C_1)_{\#MM}$ and $r\,(T_1)_{\#MM}$ are all higher than the corresponding $r\,(A_1)_{\#MM}$ and $r\,(G_1)_{\#MM}$ (Fig. 5A), indicating that "C" and "T", which result in mismatch at PAM + 1, are more favorable for off-target binding. This is exactly what was revealed from analyzing

the 5MM (Figs. 4) and 4MM sequences (SI Sect. S3.3). Furthermore, $r\,(C_1)_{\#MM}$ values are consistently higher than the corresponding $r\,(T_1)_{\#MM}$ (Fig. 5A), for example, $r\,(C_1)_{5MM}$= 2.22 while $r\,(T_1)_{5MM}$= 0.26 (SI Sect. S3.4, Table S11). This reveals that "C" is more preferred than "T" at PAM+1. Similarly, the analysis reveals that "C" is the most preferred at PAM+2, while "A" is the most preferred at PAM+3 (Fig. 5A). In addition, consistent with analysis showing increasing off-target binding as the total number of mismatches increases (Fig. 3), $r\,(N_k)_{\#MM}$ increases as the total number of mismatches increases (Fig. 5A, comparing the rows). The analysis also shows that the difference between nucleotides decreases from PAM+1 to PAM+2 and beyond, and becomes similar at PAM+4, + 5, and + 6 (Fig. 5A). This is consistent with the positional analysis showing that the PAM-adjacent + 1 to + 3 positions are the most impactful for off-target binding (Fig. 4, SI sect. S3.2.2).

Going beyond analyses described above that classified the DNA library based on user defined features, multiple linear regression (MLR) was carried out to objectively quantify the correlation between nucleotide identity at each position and off-target binding (see Methods). The analysis used the entire set of unique sequences and their corresponding measured r_i values as input to yield relative feature weight coefficients, B_N^k, of each nucleotide at the PAM + 1 to + 6 position (Methods, and SI Sect. S3.4, Table S12). Figure 5B shows a logo plot representation of B_N^k. At PAM + 1, the relative coefficient of "C" is B_C^1=1.289 (Fig. 5B and SI Sect. S3.4, Table S12), indicating that it is highly preferable. "T" has a negative B_T^1 value, although is higher than those for "A" and "G", (Fig. 5B, Supplementary Table S12). This indicates that "T" is not as preferable as "C", and "A" and "G" are the least preferable. This is completely consistent with the $r\,(N_k)_{\#MM}$ analysis (Fig. 5A). Similarly, the B_N^k values indicate that "C" is the most preferable nucleotide at PAM + 2, while "A" is the most preferable at PAM + 3 (Fig. 5B), agreeing with conclusions drawn from the $r\,(N_k)_{\#MM}$ analysis (Fig. 5A). A multiple linear regression (MLR) analysis revealed smaller B_N^k variations at PAM + 4, +5, and + 6 (Fig. 5B, SI Sect. S3.4, Table S12), again consistent with prior analysis (e.g., Fig. 4) showing that these positions, which are further away from PAM, play a secondary role in off-target binding.

To further investigate the location-dependent nucleotide preferences identified above, we examined the folding energy (ΔG_i) predicted for the most stable secondary structures of the corresponding sequences (see Methods). With sequences containing only one mismatch (1MM), the one containing a dT/dC$_1$ mismatch at PAM + 1 has a $\Delta G\,(C_1)_{1MM}$ =-46.9 kcal/mol (see sequence #2, SI Sect. S1.1, Table S1), while that containing a dT/dT$_1$ mismatch at PAM + 1 has a $\Delta G\,(T_1)_{1MM}$ = -47.6 kcal/mol (see sequence#4, SI Sect. S1.1, Table S1). Similarly, for one mismatch at PAM + 2, $\Delta G\,(C_2)_{1MM}$ = -45.4 kcal/mol while $\Delta G\,(T_2)_{1MM}$ = -46.0 kcal/mol (sequences #6 and #7, respectively, SI Sect. S1.1, Table S1); and for one mismatch at PAM + 3, $\Delta G\,(A_3)_{1MM}$ = -45.3 kcal/mol, $\Delta G\,(T_3)_{1MM}$ =-45.4 kcal/mol, and $\Delta G\,(C_3)_{1MM}$ = -45.6 kcal/mol (sequences #8, #10, and #9, respectively, SI Sect. S1.1, Table S1). These indicate that the preference for "C$_1$" over "T$_1$", "C$_2$" over "T$_2$", and "A$_3$" over "T$_3$" and "C$_3$" can be correlated with higher (less negative) ΔG_i values. The same correlation holds for consecutive dinucleotides and trinucleotides: C$_1$C$_2$ is preferred over T$_1$T$_2$ while $\Delta G\,(C_1C_2)_{2MM}$ = -44.2 kcal/mol and $\Delta G\,(T_1T_2)_{2MM}$ = -44.6 kcal/mol (sequences #21 and #28, respectively, SI Sect. S1.1, Table S1); and C$_1$C$_2$A$_3$ is preferred over T$_1$T$_2$T$_3$ while $\Delta G\,(C_1C_2A_3)_{3MM}$ = -42.1 kcal/mol and $\Delta G\,(T_1T_2T_3)_{3MM}$ = -42.3 kcal/mol (sequences #158 and #181, respectively, SI Sect. S1.1, Table S1).

Discussion

Work presented in this study implemented the SELEX-seq approach to investigate the correlation between intrinsic DNA property and off-target binding by Cas12a. A DNA duplex library was designed with a PAM for Cas12a recognition and a constant target-strand that presents no complementarity between the DNA protospacer and the crRNA guide. The library contained randomized DNA-DNA mismatches at the PAM + 1 to + 6 segment of the protospacer. DNA species bound to FnCas12a in an off-target fashion were selected via native gel shift assay and identified by sequencing, and analysis uncovered position and nucleotide dependence in DNA-DNA mismatch(es) that lead to PAM-dependent Cas12a off-target binding.

This work employed a design of the DNA/crRNA constructs that have been used successfully in a prior biochemical study to examine the role of intrinsic DNA physical properties in Cas12a off-target discrimination[42]. In particular, in the current Cas12a mechanism, the steps following PAM binding are distortion of PAM-adjacent DNA duplex and attempts to form the DNA/RNA pairing. The role of DNA/RNA pairing has been investigated using a wide range of constructs with partial pairings between the RNA guide and the DNA target-strand[14,46,48,51], and the studies have revealed complex "mismatch" configurations, including slippages between RNA and DNA as well as bulge(s) or bubble(s) between the RNA/DNA pairing[48,52]. On the other hand, constructs that lack pairings between the crRNA guide and the DNA target strand (Fig. 1) allow one to exclude "interference" of RNA/DNA pairing (including RNA/DNA mismatch(es)) and zoom in on the DNA distortion aspect, which is the most relevant to intrinsic DNA properties. Our prior biochemical study revealed that PAM-adjacent DNA-DNA mismatches (i.e., DNA bubble) cause promiscuous off-target binding by Lb-, As-, and FnCas12a[42]. However, with the limited DNA sequences examined, the earlier study was not able to further elucidate the position and base preferences within the PAM-adjacent region.

In this work, analysis conducted with the DNA library confirms that nominally 3 or more total DNA-DNA mismatches result in preferable off-target binding (Fig. 3), and bubbles at PAM + 1 to + 3 are necessary and sufficient for PAM-dependent off-target binding (Fig. 5). Data obtained show a clear correlation between binding and the presence of a properly folded PAM (Fig. 4 and SI Sect. S3.3), indicating that conclusions drawn are relevant to the PAM-dependent target interrogation step in the canonical cis-cleavage pathway of the Cas12a/crRNA effector enzyme. Importantly, comprehensive examination of DNA-DNA mismatches reveals that the PAM + 1 position plays a much more dominant role than that of the PAM + 2 and + 3 positions (Figs. 4 and 5, and SI Sect. S3.3). Specifically, analysis on the 5MM showed that a dT/dA pair and a dT/dG wobble at PAM + 1 results in unfavorable binding even with the large number of mismatches at the PAM + 2–6 positions (Fig. 4 and SI Sect. S3.3). Analysis has also uncovered that PAM-adjacent nucleotide preference for off-target binding

is correlated with increasing ΔG_i for the predicted most stable secondary structure (see Fig. 5 and the last paragraph in Results). Note that with the lack of complementarity between the RNA guide and the DNA target-strand in our construct (main text Fig. 1A, SI Sect. S1.1, Fig. S1), DNA duplex unwinding (to form an RNA/DNA hybrid) is not playing a role in Cas12a binding. Therefore, instead of considering ΔG_i as an indicator of strand dissociation (i.e., duplex stability), we propose that it is more appropriate to regard it as an indicator of the flexibility of the PAM-adjacent segment of the DNA duplex, with higher ΔG_i (i.e., less negative) indicating higher flexibility. The finding thus indicates that PAM + 1 flexibility dictates off-target Cas12a binding.

The SELEX-seq data also reveal negative r_i values for sequences containing only a PAM + 1 DNA-DNA mismatch (see examples of sequence #2 to #10, SI Sect. S2.3, Table S7), indicating that flexibility at PAM + 1, while required, by itself alone is not sufficient for stable off-target binding. Positional and preferred nucleotide analyses reveal that flexibility at PAM + 2 and + 3 augments off-target binding (Figs. 4 and 5, and SI Sect. S3.3). This is consistent with a model previously proposed based on our biochemical studies[42], that even though DNA/RNA hybrid formation is not supported, sufficient flexibility at the PAM-protospacer junction allows the duplex to adopt a bent (or "kinked") configuration, thus mitigating steric collision between the DNA and the effector complex and allowing stable binding of the off-target DNA[42]. Our model is further supported by very recently reported cryo-EM structures of trapped complexes between *Acidaminococcus sp.* Cas12a-guide RNA and a PAM-containing non target DNA (i.e., completely lacking base complementarity between the DNA target-strand and the RNA guide)[10]. The report shows that, upon binding to PAM, the Cas12a PAM-interaction domain induces progressively more DNA bending between the PAM and the protospacer, ultimately leading to DNA base unstacking and flipping[10]. PAM-adjacent flexibility of the DNA likely correlates with DNA bending upon interacting with Cas12a, which is needed prior to base unstacking and flipping[10]. Furthermore, a recent study reported that a Cas12a ternary complex with on-target DNA maintains an inherent equilibrium between a DNA unwound state and a DNA-paired duplex-like state, with the DNA-paired state containing an RNA/DNA hybrid at the PAM adjacent 5–8 base-pair "seed" segment while the remaining protospacer maintains DNA-DNA pairing[18]. It seems that bubbles at PAM + 1–3, which is smaller than the "seed" segment, can provide sufficient flexibility for an off-target, which does not have RNA/DNA pairing to support DNA unwinding, to adopt a "bent" configuration and enable stable binding while the DNA remains in the duplexed-paired state.

A desirable outcome of the SELEX-seq approach is to derive a quantitative metric for predicting off-target Cas12a binding. As a first attempt towards this goal, analysis of PAM + 1–3 of the bubble DNA show a limited correlation between the preferred nucleotides with the corresponding DNA duplex folding energy, ΔG_i (see Results, last paragraph). However, further analysis indicates the ΔG_i metric cannot properly reflect the "PAM-adjacent" DNA flexibility, and consequently, when all sequences were considered, the individual r_i and ΔG_i show no correlation (SI Sect. S3.5). Further studies are needed to properly quantify position-dependent DNA flexibility in order to predict off-target Cas12a binding. Furthermore, while the Cas12a family shares the same mechanistic framework in DNA interrogation, variations in structure-dynamics and mismatch discriminations among different Cas12a orthologs have been reported[14,21,53,54]. Results reported here were obtained with FnCas12a. We expect that PAM-adjacent DNA flexibility will lead to off-target binding in other orthologs (as we previously reported[42]), although there are likely differences in the details on how flexibility correlates with off-target binding, and further studies are needed.

Genome-wide studies utilizing Cas12a directly or Cas12a in conjunction with other modules have demonstrated off-target binding independent of the RNA guide sequence and beyond those predicted by computational approaches[55,56]. The propensity of Cas12a to bind DNA with higher flexibility may be one of the reasons behind such off-target binding, although much more in depth investigations are needed. DNA flexibility, while intrinsically a collective property of the underlying sequence, is also intimately connected with the environment within the genomic context in the cell. Flexibility of the same DNA sequence can vary substantially due to protein-DNA interactions (e.g., nucleosome positioning, transcription factor binding) and changes of topological constraints (e.g., super-coiling). In addition, during biological processes such as transcription, DNA replication and DNA repair, the DNA duplex is destabilized, leading to higher flexibility. Noticeably, it has been shown that flexibility of DNA in a nucleosome impacts Cas12a accessibility and potential cleavage and binding[41]. Based on such finding, Strohkendl et al. have suggested that increasing nucleosome breathing dynamics, which increases the flexibility of the DNA, could be a strategy to improve Cas12a DNA targeting in eukaryotic cells[41]. Our work here connecting DNA flexibility to Cas12a off-target binding would be relevant in such context. Furthermore, the dynamic DNA genome can adopt a variety of non-duplex conformations. Interestingly, Bhattacharya et al. recently reported that *Acidaminococcus sp.* Cas12a (AsCas12a) can bind and cleave biologically-relevant branched DNA constructs[40]. In particular, AsCas12a/crRNA complexes were shown to bind to Holliday junction constructs with low nanomolar affinity[40]. DNA flexibility is higher within some of these non-duplex conformations, as well as at the junctions between such conformations and the regular duplex segments. It remains to be investigated whether promiscuous binding of Cas12a to flexible DNA contributes to interactions with these non-duplex conformations.

Conclusion

In summary, this study adopts a SELEX-seq approach with a mismatched DNA library to uncover positional and nucleotide-dependent DNA features in off-target Cas12a binding. The analysis supports a role of PAM-adjacent DNA flexibility in Cas12a binding. It is possible that DNA flexibility may also play a role in Cas12a interactions with fully-paired DNAs, although that remains to be investigated. The work adds support to the notion that physical properties of the DNA (in this case conformational flexibility) influences Cas12a target discrimination. This has implications in Cas12a-based applications, particularly those relying on Cas12a binding to specific DNA sites such as transcriptional repression or activation and genome scale imaging.

Materials and methods

Cas12a protein expression and purification

Studies reported in this work were carried out with *Francisella novicida* Cas12a (FnCas12a). The catalytically active FnCas12a protein was expressed using a plasmid encoding the full-length protein (residues 1 – 1300) with a fusion of N-terminal His tag followed by an MBP tag and a TEV cleavage site[42,57]. The catalytically inactive FnCas12a (designated as dFnCas12a) contained mutations D917A and E1006A6[42]. All plasmid sequences were confirmed prior to protein expression.

FnCas12a and dFnCas12a were expressed in *Escherichia coli* and purified following procedures similar to those previously reported[3,42,50,57,58]. Rosetta 2 (DE3) Competent *Escherichia coli* Cells (Novagen) were transformed with a designated plasmid through heat shock. A single colony from the transformation was inoculated into Lysogeny Broth with kanamycin antibiotic (100 μg/ml) and incubated at 37 °C overnight. The small-scale culture was added to Terrific Broth with 50 μg/ml antibiotics (approximately 14 ml culture for a 500 ml of cell growth) and incubated at 37 °C until the optical density at 600 nm (OD_{600}) reached 0.8-1.0. Then the temperature was reduced to 18 °C while overexpression was induced by adding 200 μM isopropyl β-D-1-thiogalactopyranoside (IPTG). The large-scale culture was shaken at 18 °C for 16–20 h.

The cells were harvested by centrifugation and resuspended at 4°C in lysis buffer [25 mM tris(hydroxymethyl) aminomethane (Tris), pH 8.0, 500 mM NaCl, 20 mM imidazole, 5% (v/v) glycerol] with protease inhibitor (Roche). The cells were sonicated at 4°C followed by ultracentrifugation at 4°C to remove the cell debris. The supernatant was collected and was subjected to nickel-NTA affinity chromatography at 4°C. The target protein was eluted from the nickel-NTA column in a buffer containing 20 mM Tris pH 8.0, 500 mM NaCl and 250 mM imidazole, then subjected to MBP-His tag cleavage with the TEV protease while dialyzing in against a low-salt buffer [20 mM HEPES pH 8.0, 250 mM NaCl] at 4°C. After an overnight dialysis of at least 24 hours, the resulting product was further purified by fast protein liquid chromatography (FPLC) using an ion-exchange column (Mono-S or Na-Heparin, GE Healthcare). Fractions collected from ion-exchange chromatography were analyzed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), and those containing the target protein were combined and further purified via Size Exclusion Chromatography (SEC) using a S200 column (GE Healthcare) and a buffer containing 20 mM Tris pH 8.0, 500 mM KCl, and 20% (v/v) glycerol. SEC fractions containing the target protein were identified via SDS-PAGE and combined and concentrated in the storage buffer [20 mM Tris pH 8.0, 500 mM KCl, 20% (v/v) glycerol., and 0.5 mM 3,3',3"-phosphanetriyltripropanoic acid tris(2-carboxyethyl)phosphine (TCEP)]. The purified protein was flash-frozen and stored at -80 °C.

Concentrations of either FnCas12a or dFnCas12a were determined according to absorbance at 280 nm with an extinction coefficient of 144,000 $M^{-1}{\cdot}cm^{-1}$, which was calculated based on the amino acid sequence.

DNA library preparation

All DNA oligonucleotide strands used in this work were synthesized chemically (Integrated DNA Technologies, Inc. Coralville, Iowa). Details of the library construct and DNA sequence are described in SI Sect. S1.1. During chemical synthesis, the non-target-strand (nts) was synthesized with an equal mixture of AGCT at the PAM + 1 to + 6 positions, and a 5'-FAM label was attached at the target-strand (ts).

To form a duplexed DNA (dsDNA), appropriate amount of ts- and nts strands were first mixed in a 1:1 ratio in water and heated at 95 °C for 1 min, the mixture was cooled at room temperature for two minutes, then combined in an annealing buffer [20 mM Tris pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, and 5% (v/v) glycerol] and allowed to anneal at room temperature overnight. Following annealing, the dsDNA was separated from the single-stranded DNA via SEC using a S200 column and the Cas12a reaction buffer [20 mM Tris pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 5% (v/v) glycerol, and 0.5 mM TCEP]. Product fractions were collected and concentrated then stored at -20 °C. The concentrations of the dsDNA were determined by UV-Vis absorbance at 260 nm.

RNA transcription

Two crRNAs were used in this study: the "mismatch RNA" has a guide that lacks Watson-Crick base-pairing with the protospacer of the DNA library (Fig. 1 and SI Sect. S1.1, Fig. S1); and the "match RNA", which serves as a control, contains a guide that fully complements the protospacer of the DNA library (SI Sect. S1.1, Fig. S1). Both RNAs were synthesized through T7 in vitro transcription[42]. A 400 μl transcription reaction contained 0.5 μM single-stranded DNA template (Supplementary Table S2), 1 μM T7 top-strand-primer (SI Sect. S1.1, Table S2), 1 mM each of nucleotide tri-phosphate, 0.01% Triton, ~20–30 μg T7 polymerase, 40 mM Tris pH 7.5, 15 mM $MgCl_2$, 2 mM spermidine and 5 mM dithiothreitol (DTT). The reaction mixture was incubated at 37 °C for 3 h and then quenched by adding 20 mM ethylenediaminetetraacetic acid (EDTA). The RNA products were recovered by ethanol precipitation and purified by denaturing PAGE. The final RNA product was resuspended in ME buffer [10 mM 3-(N-morpholino) propane sulfonic acid pH 6.5 and 1 mM EDTA] and stored at -20 °C. The concentrations of RNA were determined according to UV-Vis absorbance at 260 nm. The molar extinction coefficients of RNA were estimated by ϵ = # of nucleotide × 10,000 $M^{-1}{\cdot}cm^{-1}$.

SELEX identification of Cas12a bound off-targets

To implement the SELEX scheme to identify Cas12a-bound off-targets, a native gel shift assay was carried out to separate the unbound DNA populations from those bound to a dFnCas12a effector complex containing the mismatch RNA (Fig. 1). The SELEX reaction was carried out with 50 nM dFnCas12a, 100 nM mismatch RNA, and 100 nM dsDNA library. To form the Cas12a effector complex, the proper amount of the mismatch RNA was first heated at 95 °C for 1 min, then cooled at room temperature for 2 min. After incubating the RNA in a Cas12a reaction buffer at room temperature for 10 min, appropriate amount of dFnCas12a was added, the solution was adjusted to maintain the salts at the same concentrations as that in the Cas12a reaction buffer, and the mixture was incubated at room temperature for 15 min. Appropriate amount of dsDNA library in the reaction buffer was

then added, and the mixture was incubated at 37 °C for 30 min. On conclusion of the incubation, the sample was combined with an equal volume of a native loading buffer [1x Cas12a reaction buffer, 50% (v/v) glycerol], and then subjected to native PAGE gel electrophoresis (carried out at 4 °C) to separate the unbound and bound populations. The DNA species were directly visualized by FAM imaging and unbound and bound DNAs were recovered from the corresponding gel slices by eluting in the TE buffer (10 mM Tris, pH 8.0, 1 mM EDTA) at room temperature, followed by phenol/chloroform extraction to remove the Cas12a protein. The recovered DNA was buffer exchanged and concentrated into 10mM Tris pH 8.0 and used for sequencing. Note that a follow-up PCR amplification step was not included in the SELEX scheme in this study as it would lead to a fully paired DNA duplex library and destroy the PAM-adjacent DNA-DNA mismatch(es).

DNA sequencing

Sequencing of the unbound and bound DNA populations was carried out on the Illumina MiSeq platform by commercial vendors (Laragen Inc., Culver City, CA and MCLAB, San Francisco, CA). To prepare the samples for sequencing, primers containing adapter and index sequences (see details in SI Sect. S2.1) were designed following the Illumina 16 S Metagenomic Sequencing Library Preparation guide, and then added to the sample via PCR. For the adapter PCR, each reaction mixture (total 50 µl) contained 2 µl of the DNA being sequenced as template, and included 0.2 µM forward read 1 adapter primer, 0.2 µM reverse read 2 adaptor primer, 1x Pfu HF Buffer, 0.25 mM dNTP and 0.05 U/ml Pfu DNA polymerase. Typically, five of these 50-µl adapter PCR reactions were carried out simultaneously, then pooled together. The DNA products were then purified using the Genejet PCR Purification Kit (Invitrogen) and then used as the template for index PCR. Each index PCR reaction (total 50 µl) included 0.2 µM Nextera XT Index 2 (I5) Primers, 0.2 µM Nextera XT Index 1 (I7) Primers, 1x Pfu HF Buffer, 0.25 mM dNTP, 0.05 U/ml Pfu DNA polymerase, and 8 nM of amplicon from adapter PCR. Typically, five of these 50-µl index PCR reactions were carried out simultaneously, and the products were pooled and purified using a 12% native PAGE gel. The purified DNA were dissolved in a 10 mM Tris pH 8.0 buffer and sent out for sequencing.

Sequencing data analysis

Sequencing data was processed from compressed FASTQ files using custom generated Python scripts based on NumPy[59] and pandas[60]. Unique reads from each sample were counted (c_i), and used to calculate the proportion of each sequence (p_i) in the dataset as:

$$p_i = \frac{counts\ of\ the\ i-th\ sequencec_i}{total\ counts\ of\ the\ dataset \sum\ c_i} \tag{1}$$

Proportion was calculated for every unique sequence from the library data set (p_i^L), the bound data set (p_i^B), and the unbound data set (p_i^U). The bound enrichment for each unique sequence, E_B^i, was computed as:

$$E_B^i = \frac{p_i^B}{p_i^L} \tag{2}$$

The unbound enrichment for each unique sequence, E_U^i, was computed as:

$$E_U^i = \frac{p_i^U}{p_i^L} \tag{3}$$

The relative enrichment for a unique sequence was computed as:

$$r_i = \log_2\left(\frac{E_B^k}{E_U^k}\right) \tag{4}$$

In addition, for a particular group of sequences sharing a feature of interest (G_{feat}), the relative enrichment ($r\left(G_{feat}\right)$) for that feature was computed as the ratio of the weighted average enrichment between the bound and unbound data sets:

$$r\left(G_{feat}\right) = \log_2\left(\frac{\frac{\sum_{i\in G}\left(p_i^B \times E_B^i\right)}{\sum_{i\in G}\left(p_i^B\right)}}{\frac{\sum_{i\in G}\left(p_i^U \times E_U^i\right)}{\sum_{i\in G}\left(p_i^U\right)}}\right) \tag{5}$$

Secondary structure analysis of individual DNA duplexes containing DNA-DNA mismatches

Secondary structure of DNA duplexes containing various DNA-DNA mismatches were predicted using RNAStructure[61]. For each given sequence, the total free energy for a variety of secondary structures was computed using the program based on reported DNA nearest-neighbor parameters. The one with the lowest total free energy was chosen as the most stable secondary structure, and the corresponding free energy, designed as ΔG_i, was used in this work as the duplex folding free energy of the corresponding DNA sequence. To carry out the calculation, a batch file containing the PAM-12 to PAM + 24 nucleotides of each of the 4096 unique full duplex sequences in the DNA library (SI Sect. S1.1, Fig. S1 and "Supplementary Data.xlsx") was created in the

FASTA format using Python scripts described in the Biopython[62] tool box. The possible secondary structures and associated folding energies were then calculated with a Python script utilizing the DuplexFold command of RNAStructure.

Multiple linear regression analysis

A multiple linear regression (MLR) model with Lasso regularization was used to predict the binding affinity of the Cas12a complex and gauge the effect of interdependencies between nucleotide positions. Every nucleotide along the 6-base pair variable region was encoded using one hot encoding and used as features for the model. The model was trained to predict the r_i using the LassoCV function within Python's scikit-learn package[63] using five-fold cross validation for hyperparameter tuning. The model performance was then measured using the adjusted R^2. To interpret the model, the model feature weights (F_N^k) at each position along the 6-base pair variable region, normalized by the mean feature weight ($\overline{F}^k$), are then plotted using LogoMaker[64].

$$B_N^k = F_N^k - \overline{F}^k \tag{6}$$

This provides a visual representation where nucleotides that improve binding relative to the mean are positive and one that decreased binding are negative. The height of the letters (B_N^k) are proportional to the strength of the effect at each nucleotide position.

Data availability

Data is provided within the manuscript or supplementary information files. Sequencing data files from this study have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) with accession number PRJNA1091700.

Received: 21 September 2024; Accepted: 20 January 2025
Published online: 10 February 2025

References

1. Koonin, E. V. & Makarova, K. S. Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, 20180087 (2019).
2. Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr. Opin. Microbiol.* **37**, 67–78 (2017).
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. Yao, R. et al. CRISPR-Cas9/Cas12a biotechnology and application in bacteria. *Synth. Syst. Biotechnol.* **3**, 135–149 (2018).
5. Swarts, D. C. & Jinek, M. Cas9 versus Cas12a/Cpf1: structure-function comparisons and implications for genome editing. *Wiley Interdiscip. Rev. RNA* **9**, e1481 (2018).
6. Knott, G. J. & Doudna, J. A. CRISPR-Cas guides the future of genetic engineering. *Science* **361**, 866–869 (2018).
7. Paul, B. & Montoya, G. CRISPR-Cas12a: functional overview and applications. *Biomed. J.* **43**, 8–17 (2020).
8. Meliawati, M., Schilling, C. & Schmid, J. Recent advances of Cas12a applications in bacteria. *Appl. Microbiol. Biotechnol.* **105**, 2981–2990 (2021).
9. C Swarts, D. Making the cut(s): how Cas12a cleaves target and non-target DNA. *Biochem. Soc. Trans.* **47**, 1499–1510 (2019).
10. Soczek, K. M., Cofsky, J. C., Tuck, O. T., Shi, H. & Doudna, J. A. CRISPR-Cas12a bends DNA to destabilize base pairs during target interrogation. Preprint at. https://doi.org/10.1101/2024.07.31.606079 (2024).
11. Gao, P., Yang, H., Rajashankar, K. R., Huang, Z. & Patel, D. J. Type V CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. *Cell. Res.* **26**, 901–913 (2016).
12. Swarts, D. C. & Jinek, M. Mechanistic insights into the cis- and trans-acting DNase activities of Cas12a. *Mol. Cell* **73**, 589–600e4 (2019).
13. Stella, S., Alcón, P. & Montoya, G. Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage. *Nature* **546**, 559–563 (2017).
14. Singh, D. et al. Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1 (Cas12a). *Proc. Natl. Acad. Sci.* **115**, 5444–5449 (2018).
15. Swarts, D. C., van der Oost, J. & Jinek, M. Structural basis for Guide RNA Processing and seed-dependent DNA targeting by CRISPR-Cas12a. *Mol. Cell* **66**, 221–233e4 (2017).
16. Zhang, L. et al. Conformational dynamics and cleavage sites of Cas12a are modulated by complementarity between crRNA and DNA. *iScience* **19**, 492–503 (2019).
17. Son, H. et al. Mg2+-dependent conformational rearrangements of CRISPR-Cas12a R-loop complex are mandatory for complete double-stranded DNA cleavage. *Proc. Natl. Acad. Sci.* **118**, e2113747118 (2021).
18. Singh, J., Liu, K. G., Allen, A., Jiang, W. & Qin, P. Z. A DNA unwinding equilibrium serves as a checkpoint for CRISPR-Cas12a target discrimination. *Nucleic Acids Res.* **51**, 8730–8743 (2023).
19. Pattanayak, V. et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
20. Nazipova, N. N. & Shabalina, S. A. Understanding off-target effects through hybridization kinetics and thermodynamics. *Cell Biol. Toxicol.* **36**, 11–15 (2020).
21. Murugan, K., Seetharam, A. S., Severin, A. J. & Sashital, D. G. CRISPR-Cas12a has widespread off-target and dsDNA-nicking effects. *J. Biol. Chem.* **295**, 5538–5553 (2020).
22. Kang, S. H. et al. Prediction-based highly sensitive CRISPR off-target validation using target-specific DNA enrichment. *Nat. Commun.* **11**, 3596 (2020).
23. Wu, X., Mao, S., Ying, Y., Krueger, C. J. & Chen, A. K. Progress and challenges for live-cell imaging of genomic loci using CRISPR-based platforms. *Genomics Proteom. Bioinf.* **17**, 119–128 (2019).
24. Maass, P. G. et al. Spatiotemporal allele organization by allele-specific CRISPR live-cell imaging (SNP-CLING). *Nat. Struct. Mol. Biol.* **25**, 176–184 (2018).
25. Chaudhary, N., Im, J. K., Nho, S. H. & Kim, H. Visualizing live chromatin dynamics through CRISPR-Based imaging techniques. *Mol. Cells* **44**, 627–636 (2021).
26. Didovyk, A., Borek, B., Tsimring, L. & Hasty, J. Transcriptional regulation with CRISPR-Cas9: principles, advances, and applications. *Curr. Opin. Biotechnol.* **40**, 177–184 (2016).

27. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR–Cas9. *Nat. Rev. Genet.* **16**, 299–311 (2015).
28. McCarty, N. S., Graham, A. E. & Studená, L. & Ledesma-Amaro, R. Multiplexed CRISPR technologies for gene editing and transcriptional regulation. *Nat. Commun.* **11**, 1281 (2020).
29. Tang, J., Chen, L. & Liu, Y. G. Off-target effects and the solution. *Nat. Plants* **5**, 341–342 (2019).
30. Tadić, V., Josipović, G., Zoldoš, V. & Vojta, A. CRISPR/Cas9-based epigenome editing: an overview of dCas9-based tools with special emphasis on off-target activity. *Methods* **164-165**, 109–119 (2019).
31. Tong, Y. et al. Highly efficient DSB-free base editing for streptomycetes with CRISPR-BEST. *Proc. Natl. Acad. Sci.* **116**, 20366–20375 (2019).
32. Rees, H. A. et al. Improving the DNA specificity and applicability of base editing through protein engineering and protein delivery. *Nat. Commun.* **8**, 15790 (2017).
33. Chen, P. et al. A Cas12a ortholog with stringent PAM recognition followed by low off-target editing rates for genome editing. *Genome Biol.* **21**, 78 (2020).
34. Kleinstiver, B. et al. Engineered CRISPR–Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. *Nat. Biotechnol.* **37**, 276–282 (2019).
35. Paul, B., Chaubet, L., Verver, D. E. & Montoya, G. Mechanics of CRISPR-Cas12a and engineered variants on λ-DNA. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkab1272 (2021).
36. Wang, W. et al. Expanding the range of editable targets in the wheat genome using the variants of the Cas12a and Cas9 nucleases. *Plant. Biotechnol. J.* **19**, 2428–2441 (2021).
37. Liu, X. et al. Lb2Cas12a and its engineered variants mediate genome editing in human cells. *FASEB J.* **35** (2021).
38. Losito, M., Smith, Q. M., Newton, M. D., Cuomo, M. E. & Rueda, D. S. Cas12a target search and cleavage on force-stretched DNA. *Phys. Chem. Chem. Phys.* **23**, 26640–26644 (2021).
39. van Aelst, K., Martínez-Santiago, C., Cross, S. & Szczelkun, M. The effect of DNA topology on observed Rates of R-Loop formation and DNA strand cleavage by CRISPR Cas12a. *Genes* **10**, 169 (2019).
40. Bhattacharya, S., Agarwal, A. & Muniyappa, K. Deciphering the substrate specificity reveals that CRISPR-Cas12a is a bifunctional enzyme with both endo- and exonuclease activities. *J. Mol. Biol.* **436**, 168550 (2024).
41. Strohkendl, I. et al. Inhibition of CRISPR-Cas12a DNA targeting by nucleosomes and chromatin. *Sci. Adv.* **7** (2021).
42. Jiang, W. et al. CRISPR-Cas12a nucleases bind flexible DNA duplexes without RNA/DNA complementarity. *ACS Omega* **4**, 17140–17147 (2019).
43. Wang, J., Lu, J., Gu, G. & Liu, Y. In vitro DNA-binding profile of transcription factors: methods and new insights. *J. Endocrinol.* **210**, 15–27 (2011).
44. Gu, G., Wang, T., Yang, Y., Xu, X. & Wang, J. An Improved SELEX-Seq Strategy for characterizing DNA-Binding specificity of transcription factor: NF-κB as an Example. *PLoS One* **8**, e76109 (2013).
45. Wu, Y. X. & Kwon, Y. J. Aptamers: the evolution of SELEX. *Vitro Sel. Evol.* **106**, 21–28 (2016).
46. Zhang, L. et al. Systematic in vitro profiling of off-target affinity, cleavage and efficiency for CRISPR enzymes. *Nucleic Acids Res.* **48**, 5037–5053 (2020).
47. Tsai, S. Q. et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).
48. Jones, S. K. et al. Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nat. Biotechnol.* **39**, 84–93 (2021).
49. Eggers, A. R. et al. Rapid DNA unwinding accelerates genome editing by engineered CRISPR-Cas9. *Cell* **187**, 3249–3261e14 (2024).
50. Yamano, T. et al. Crystal structure of Cpf1 in Complex with Guide RNA and target DNA. *Cell* **165**, 949–962 (2016).
51. Jeon, Y. et al. Direct observation of DNA target searching and cleavage by CRISPR-Cas12a. *Nat. Commun.* **9**, 2777 (2018).
52. Zhang, W. et al. The Off-Target Effect of CRISPR-Cas12a System toward insertions and deletions between Target DNA and crRNA sequences. *Anal. Chem.* **94**, 8596–8604 (2022).
53. Cofsky, J. C. et al. CRISPR-Cas12a exploits R-loop asymmetry to form double-strand breaks. *eLife* **9**, e55143 (2020).
54. Kleinstiver, B. P. et al. Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat. Biotechnol.* **34**, 869–874 (2016).
55. Randall, L. B. et al. Genome- and transcriptome-wide off-target analyses of an improved cytosine base editor. *Plant Physiol.* **187**, 73–87 (2021).
56. Sturme, M. H. J. et al. Occurrence and nature of off-target modifications by CRISPR-Cas genome editing in plants. *ACS Agric. Sci. Technol.* **2**, 192–201 (2022).
57. Sundaresan, R., Parameshwaran, H. P., Yogesha, S. D., Keilbarth, M. W. & Rajan, R. RNA-Independent DNA cleavage activities of Cas9 and Cas12a. *Cell Rep.* **21**, 3728–3739 (2017).
58. Mohanraju, P., Oost, J., Jinek, M. & Swarts, D. Heterologous expression and purification of the CRISPR-Cas12a/Cpf1 protein. *BIO-Protoc.* **8** (2018).
59. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
60. McKinney, W. *Data Structures for Statistical Computing in Python*in 56–61 (Austin, 2010). https://doi.org/10.25080/Majora-92bf1922-00a
61. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
62. Chapman, B., Chang, J. & Biopython Python tools for computational biology. A*CM SIGBIO Newsl.* **20**, 15–19 (2000).
63. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
64. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).

Acknowledgements

This research was supported in part by grants from National Institute of General Medical Sciences (R01GM124413 and R35GM145341 to P.Z.Q.; R35GM130376 to R.R.).

Author contributions

A.A. and P.Z.Q. designed the study and wrote the manuscript; A.A. and J.S. prepared protein and RNA reagents; A.A. carried out the SELEX-seq experiments; A.A., B.H.C., R.R., and P.Z.Q. analyzed the sequencing data; B.H.C. designed and performed the MLR analysis. All authors reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-87565-9.

Correspondence and requests for materials should be addressed to P.Z.Q.

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

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