PAM-Adjacent DNA Flexibility Tunes CRISPR-Cas12a Off-Target Binding

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

[1]Department of Chemistry, University of Southern California, 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

*Corresponding Author: Peter Z. Qin, 3430 S Vermont Ave., Los Angeles, CA 90089, USA; Tel: (213) 821-2461; Fax: (213) 740-2701; E-mail: pzq@usc.edu

[†]Present address: Beckman Coulter, 1584 Enterprise Blvd, West Sacramento, CA 95691, USA

Supporting Information

Table of Contents

S1. Additional Data on Biochemical Characterization

S1.1. DNA and RNA Constructs

Figure S1 shows schematics of the library of DNA duplex construct together with the match and mismatch crRNAs. Examples of individual DNA sequences are shown in Table S1. The two crRNAs used in this study were synthesized through T7 *in vitro* transcription as described in Methods in the main text. Sequences of the DNA templates used for RNA transcription are listed in Table S2.

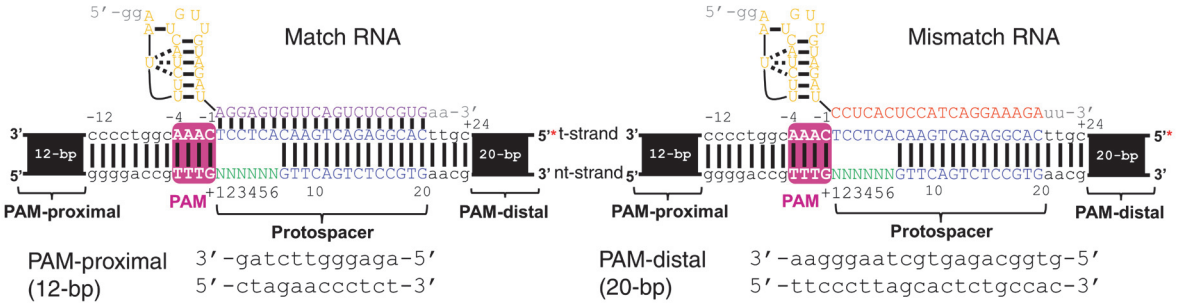


Figure S1. Schematics of the DNA and RNA constructs. The DNA duplexes show the sequence of the central core region, with the upper-case letters indicating the PAM (in pink) and the protospacer (in blue). The 6 randomized nucleotides are marked by 'N' (in green) with the positions labelled below. Sequences of the PAM-proximal and PAM-distal segments are listed below. "*" indicates a 5'-FAM label for visualization of the DNA. The crRNAs contain a 19-nt core (organ uppercase letters) followed by a 3' single-stranded segment. 20 nucleotides of the single-stranded RNA segment (capitalized) serve as the guide that interacts with the DNA protospacer,[1] while the 2 nucleotides at the 3' terminus do not interact with the DNA[1] and are included to mitigate 3'-termus heterogeneity that is known to occur in T7 *in vitro* transcription.[2] The match RNA (left) allows pairing between the RNA guide (purple uppercase letters) and the DNA target-strand (blue), while the mismatch RNA (right) does not allow Watson-Crick pairing between the RNA guide (red uppercase letters) and the DNA target-strand (blue).

Table S1. Examples of individual DNA sequences and their attributes.

Seq#	Seq[a]	MM[b] Count	MM[b] Locations	ΔG_i[c]	r_i[d]
2	3′-TCCTCA-5′ 5′-CGGAGT-3′	1	1	-46.9	-0.0597
4	3′-TCCTCA-5′ 5′-TGGAGT-3′	1	1	-47.6	-0.6595
6	3′-TCCTCA-5′ 5′-ACGAGT-3′	1	2	-45.4	-0.7568
7	3′-TCCTCA-5′ 5′-ATGAGT-3′	1	2	-46.0	-0.2825
8	3′-TCCTCA-5′ 5′-AGAAGT-3′	1	3	-45.3	-0.1180
9	3′-TCCTCA-5′ 5′-AGCAGT-3′	1	3	-45.6	-1.3851
10	3′-TCCTCA-5′ 5′-AGTAGT-3′	1	3	-45.4	-0.7513
21	3′-TCCTCA-5′ 5′-CCGAGT-3′	2	1,2	-44.2	0.5562
28	3′-TCCTCA-5′ 5′-TTGAGT-3′	2	1,2	-44.6	-0.6954
68	3′-TCCTCA-5′ 5′-ACAAGT-3′	2	2,3	-43.8	-1.0224
70	3′-TCCTCA-5′ 5′-ACTAGT-3′	2	2,3	-43.7	-0.9293
73	3′-TCCTCA-5′ 5′-ATTAGT-3′	2	2,3	-43.8	-0.8658
158	3′-TCCTCA-5′ 5′-CCAAGT-3′	3	1,2,3	-42.1	3.2186
160	3′-TCCTCA-5′ 5′-CCTAGT-3′	3	1,2,3	-42.1	2.6475
181	3′-TCCTCA-5′ 5′-TTTAGT-3′	3	1,2,3	-42.3	-0.8531
D3[e]	3′-TCCTCA-5′ 5′-AGGAGT-3′	0	---	-50.5	---

(a) The PAM+1 to PAM+6 sequence (see Figure S1) is shown from left to right, with the target-strand on the top in the 3′ to 5′ direction, and the non-target strand at the bottom in the 5′ to 3′ direction. Unpaired nucleotides are shown in red.

(b) "MM" refers to mismatch(es).

(c) Folding energy in units of kcal/mol. Computed as described in main text, Materials and Methods.

(d) Enrichment r_i computed as described in main text, Materials and Methods.

(e) The D3 DNA is a fully matched duplex and designated as sequence #1 in the extended data file "Supplementary Data.xlsx". In each sequencing dataset this particular DNA accounted for ~ 50% of the total counts due to bias arisen from the PCR step during preparation for sequencing. Therefore, bind enrichment r_i cannot be determined.

Table S2. DNA sequences used as templates for *in vitro* transcription.

Name	**Sequences (5'-3')**[a]
D3a.1 template (match RNA)	TTCACGGAGACTGAACACTCCTATCTACAACAGTAGAAATTCcta Tagtgagtcgtatta
D3a.2 template (mismatch RNA)	AATCTTTCCTGATGGAGTGAGGATCTACAACAGTAGAAATTCcta Tagtgagtcgtatta
T7 top-strand primer	Gcgcgctaatacgactcactatag

(a) Lowercase letters indicate the sequences that form a duplex with the T7 top-strand primer to serve as the T7 promoter.

S1.2. Biochemical Validation of the DNA Library

The DNA library was validated by a cleavage assay. Experiments were carried out with Cas12a effector complexes formed between a catalytically-active FnCas12a and either the match or the mismatch RNA (Figure S1). To form the effector complex, 120 nM of the proper RNA was first heated at 95°C for 1 minute, then cooled at room temperature for 2 minutes. After incubating the RNA in the Cas12a reaction buffer [20 mM Tris pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 5% (v/v) glycerol, 0.5 mM TCEP] at room temperature for 10 minutes, 100 nM of FnCas12a was added, and the solution was adjusted to maintain the salts at the same concentrations as that in the Cas12a reaction buffer. The RNA/FnCas12a mixture was incubated at room temperature for 15 minutes. Following incubation, 10 nM of DNA duplex substrate was added, and the reaction was allowed to proceed at 37°C for 30 minutes. Upon conclusion of the reaction, an equal volume of a denaturing loading buffer was added [8M urea, 20 mM EDTA, 20% glycerol, 0.1% bromophenol blue, 0.1% xylene cyanol], and the reaction species were resolved using denaturing polyacrylamide gel electrophoresis. Figure S2 shows an example of the cleavage experiments. DNA cleavage occurred with only the match RNA, thus validating the library.

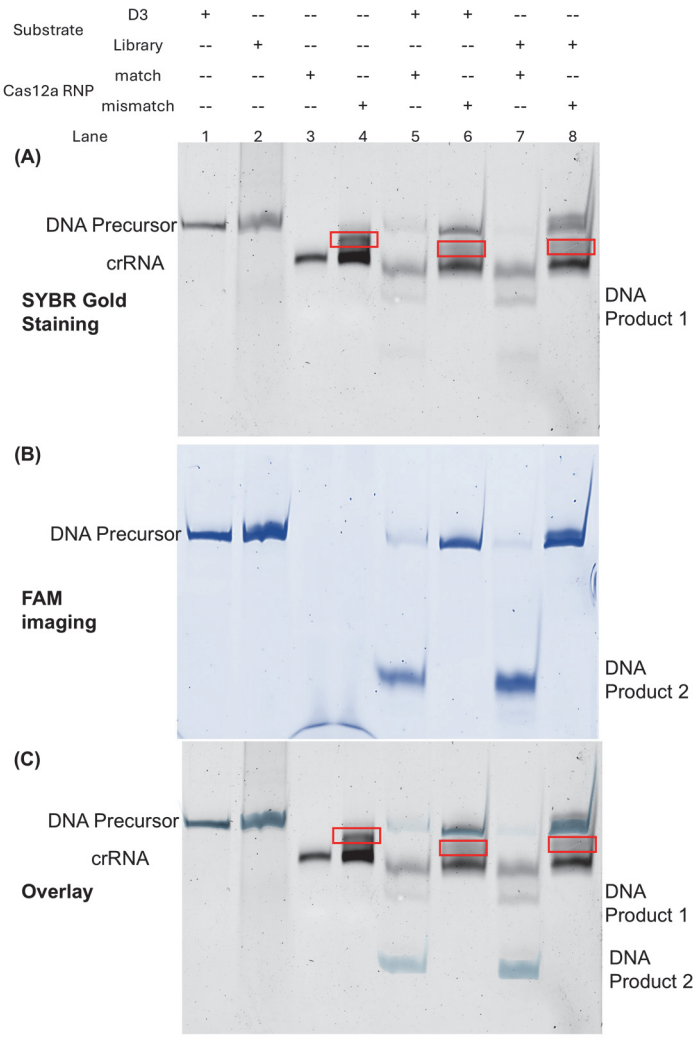


Figure S2. Validation of DNA cleavage with FnCas12a effectors by denaturing gel. (A) SYBR Gold staining that revealed all DNA and RNA species; (B) FAM imaging that revealed only the FAM-labeled DNA target; and (C) overlay of (A) and (B). Detailed sequence information can be found in Figure S1 and Table S1. Both the D3 duplex (full complementary strands) and the library duplex are cleaved with the match RNP (Lanes 5 and 7, respectively) but not with the mismatch RNP (Lanes 6 and 8, respectively). Also note that the mismatch RNA showed an extra band (red box) that may indicate heterogeneous length or folding, but this does not impact the conclusion that the mismatch RNP cannot cleave either the D3 or the library DNA.

S1.3. Example of Separation of Bound and Unbound DNA for Sequencing

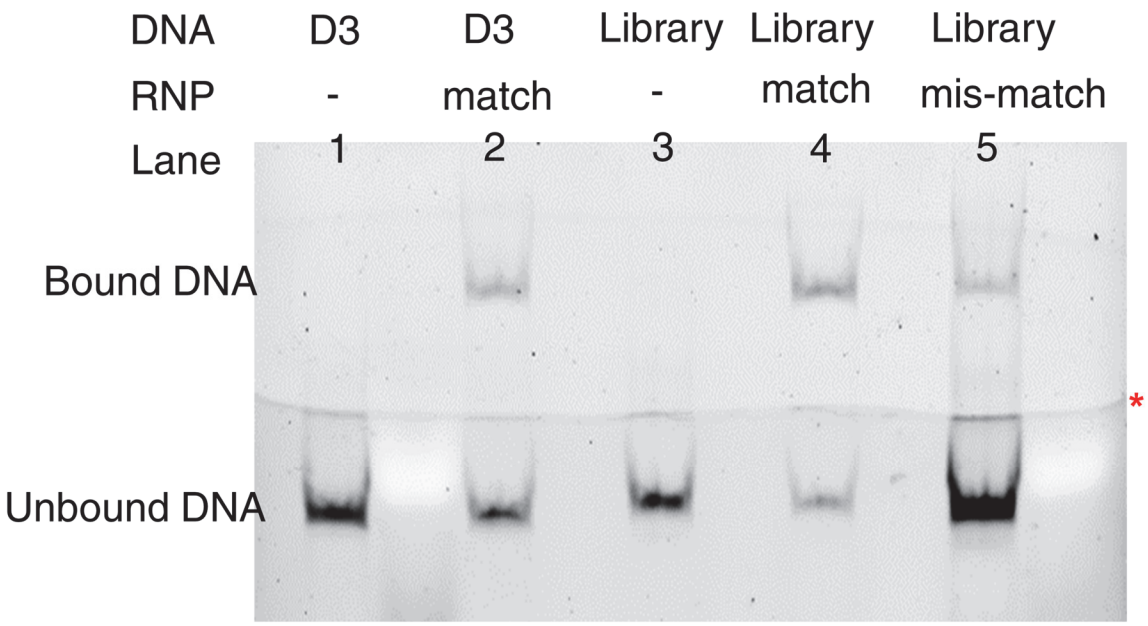


Figure S3. An example of SELEX identification of bound and unbound DNAs. The experiment was carried out with 100 nM DNA and 50 nM dFnCas12a RNP as described in the main text, and the species were identified by FAM imaging. Lanes 1 – 4 were included to mark the respective DNA species. Bound and unbound DNAs from lane 5, which is also shown in main text Figure 2A, were recovered to obtain sequencing dataset 3. When comparing bound and unbound DNA bands, binding of the library with the dFnCas12a effector containing the mismatch RNA (lane 5), while measurable, was clearly weaker than that with the match RNA (lane 4). The red star indicates the gel interface when transitioning from 5% to 10%.

S2. Additional Data on Sequencing

S2.1. Information on Primers for Preparing Sequencing Samples

Primers used in the PCR reactions for adding adapters and indexes required for sequencing (see main text, Materials and Methods) are listed in Table S3. Table S4 shows the details of the 8-nucleotide indexes used for the pooling of samples for sequencing.

Table S3. Sequences of index and adapter primers.

Name	Sequences (5'-3')[a][b][c]	T_m (°C)
Forward Read 1 Adapter primer	**TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**gtggcaga gtgctaagggaaCGTT	71.4
Reverse Read 2 Adapter primer	**GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG**ctagaac cctctggggaccgT	71.4
Nextera XT Index 2 I5 primers	**AATGATACGGCGACCACCGAGATCTACAC (I5) TCGTCGGC AGCGTC**	-
Nextera XT Index 1 I7 primers	**CAAGCAGAAGACGGCATACGAGAT (I7) GTCTCGTGGGCTCGG**	-

(a) Adapter sequence is bolded and colored. Lower case letters are sequence from library DNA while upper case letters are read and adapter sequence.

(b) Read sequence is underlined.

(c) "(I5)" and "(I7) indicate the position of index.

Table S4. Sequences of I5 and I7 indices.

Name	Bases in Adapter	Bases for Sample Sheet
S501	TAGATCGC	TAGATCGC
S502	CTCTCTAT	CTCTCTAT
S503	TATCCTCT	TATCCTCT
S504	AGAGTAGA	AGAGTAGA
N701	TCGCCTTA	TAAGGCGA
N702	CTAGTACG	CGTACTAG
N703	TTCTGCCT	AGGCAGAA
N704	GCTCAGGA	TCCTGAGC

S2.2. Assessment of Quality of the Sequencing Data

Table S5 shows the quality control (QC) parameters obtained using cutadapt.[3] While the three datasets obtained differed in depth, all of them had reads-written > 85%, indicating the majority of the data met the stringent requirements for further analysis.

Table S5. Summary of quality control parameters.[(a)]

Parameters	Library	Dataset 1		Dataset 2		Dataset 3	
		Bound	Unbound	Bound	Unbound	Bound	Unbound
Total reads processed	1,137,790	123,972	304,632	195,795	305,001	337,555	1,078,411
Reads with adapters[(b)]	1,119,729 (98.4%)	122,106 (98.5%)	299,386 (98.3%)	192,540 (98.3%)	298,374 (97.8%)	318,301 (94.3%)	1,032,410 (95.7%)
Reads that were too short[(b)]	5,345 (0.5%)	356 (0.3%)	1,084 (0.4%)	587 (0.3%)	1,128 (0.4%)	596 (0.2%)	2,667 (0.2%)
Reads that were too long[(b)]	121,543 (10.7%)	14,753 (11.9%)	36,033 (11.8%)	26,709 (13.6%)	42,703 (14.0%)	42,839 (12.7%)	149,043 (13.8%)
Reads written (passing filters[(b)]	1,010,902 (88.8%)	108,863 (87.8%)	267,515 (87.8%)	168,499 (86.1%)	261,170 (85.6%)	294,120 (87.1%)	926,701 (85.9%)
Total basepairs processed	171,806,290	18,719,772	45,999,432	29,565,045	46,055,151	25,128,775	78,463,503
Total written basepairs (filtered) [(c)]	6,065,412 (3.5%)	653,178 (3.5%)	1,605,090 (3.5%)	1,010,994 (3.4%)	1,567,020 (3.4%)	1,764,720 (7.0%)	5,560,206 (7.1%)

(a) The initial library and dataset 3 were collected using available reads in a lane of 1 million for bound and unbound data each (Laragen Inc., Culver City, CA). Dataset 1 and dataset 2 were collected from available reads in a lane of 1 million per dataset. (MCLAB, San Francisco, CA).

(b) Percentages obtained from the ratio of the corresponding reads over the total reads processed in the dataset.

(c) Percentages obtained from the ratio of the corresponding number of base-pairs over the total base-pairs processed in the dataset.

S2.3. Examples of Calculation of Enrichment

Table S6 lists reads of selective individual sequences obtained in dataset 3. Table S7 lists the values of the weight of the corresponding sequence from the library dataset (p_i^L), the bound dataset (p_i^B), the unbound dataset (p_i^U), the bound enrichment (E_B^i), and unbound enrichment (E_U^i), and the relative enrichment (r_i).

Table S6. Reads of Selective Individual Sequences.

Seq#	Seq[(a)]	MM Count	MM Locations	Bound Count	Unbound Count	Library Count
2	3′-TCCTCA-5′ 5′-CGGAGT-3′	1	1	74	243	350
4	3′-TCCTCA-5′ 5′-TGGAGT-3′	1	1	86	428	508
6	3′-TCCTCA-5′ 5′-ACGAGT-3′	1	2	71	378	262
7	3′-TCCTCA-5′ 5′-ATGAGT-3′	1	2	590	2261	620
8	3′-TCCTCA-5′ 5′-AGAAGT-3′	1	3	93	318	361
9	3′-TCCTCA-5′ 5′-AGCAGT-3′	1	3	48	395	273
10	3′-TCCTCA-5′ 5′-AGTAGT-3′	1	3	942	4996	552
21	3′-TCCTCA-5′ 5′-CCGAGT-3′	2	1,2	35	75	72
28	3′-TCCTCA-5′ 5′-TTGAGT-3′	2	1,2	49	250	267
68	3′-TCCTCA-5′ 5′-ACAAGT-3′	2	2,3	25	160	150
70	3′-TCCTCA-5′ 5′-ACTAGT-3′	2	2,3	33	198	184
73	3′-TCCTCA-5′ 5′-ATTAGT-3′	2	2,3	120	689	401
158	3′-TCCTCA-5′ 5′-CCAAGT-3′	3	1,2,3	130	44	91
160	3′-TCCTCA-5′ 5′-CCTAGT-3′	3	1,2,3	175	88	121
181	3′-TCCTCA-5′ 5′-TTTAGT-3′	3	1,2,3	94	535	447

(a) The PAM+1 to PAM+6 sequence of the protospacer(see Figure S1) is shown from left to right, with the target-strand on the top in the 3′ to 5′ direction, and the non-target-strand at the bottom in the 5′ to 3′ direction. Unpaired nucleotides are shown in red.

Table S7. Examples of Enrichment Calculation.

Seq#	Seq[a]	$p_i^B\ (\times 10^{-4})$[b]	$p_i^U\ (\times 10^{-4})$[b]	$p_i^L\ (\times 10^{-4})$[b]	E_B^i[c]	E_U^i[c]	r_i[d]
2	3′-TCCTCA-5′ 5′-CGGAGT-3′	2.52	2.62	3.46	0.7267	0.7574	-0.0597
4	3′-TCCTCA-5′ 5′-TGGAGT-3′	2.92	4.62	5.03	0.5819	0.9191	-0.6595
6	3′-TCCTCA-5′ 5′-ACGAGT-3′	2.41	4.08	2.59	0.9314	1.5738	-0.7568
7	3′-TCCTCA-5′ 5′-ATGAGT-3′	20.1	24.4	6.13	3.2707	3.9781	-0.2825
8	3′-TCCTCA-5′ 5′-AGAAGT-3′	3.16	3.43	3.57	0.8854	0.9609	-0.1180
9	3′-TCCTCA-5′ 5′-AGCAGT-3′	1.63	4.26	2.70	0.6043	1.5784	-1.3851
10	3′-TCCTCA-5′ 5′-AGTAGT-3′	32.0	53.9	5.46	5.8654	9.8731	-0.7513
21	3′-TCCTCA-5′ 5′-CCGAGT-3′	1.19	0.809	0.712	1.6708	1.1363	0.5562
28	3′-TCCTCA-5′ 5′-TTGAGT-3′	1.67	2.70	2.64	0.6308	1.0214	-0.6954
68	3′-TCCTCA-5′ 5′-ACAAGT-3′	0.850	1.73	1.48	0.5728	1.1636	-1.0224
70	3′-TCCTCA-5′ 5′-ACTAGT-3′	1.12	2.14	1.82	0.6164	1.1739	-0.9293
73	3′-TCCTCA-5′ 5′-ATTAGT-3′	4.08	7.43	3.97	1.0285	1.8743	-0.8658
158	3′-TCCTCA-5′ 5′-CCAAGT-3′	4.42	0.475	0.900	4.9101	0.52745	3.2186
160	3′-TCCTCA-5′ 5′-CCTAGT-3′	5.95	0.950	1.20	4.9709	0.7934	2.6475
181	3′-TCCTCA-5′ 5′-TTTAGT-3′	3.20	5.77	4.42	0.7228	1.3056	-0.8531

(a) The PAM+1 to PAM+6 sequence of the protospacer (see Figure S1) is shown from left to right, with the target-strand on the top in the 3′ to 5′ direction, and the non-target-strand at the bottom in the 5′ to 3′ direction. Unpaired nucleotides are shown in red.

(b) Values computed according to eq 1 in main text.

(c) Values computed according to eq 2 and eq 3 in main text.

(d) Values computed according to eq 4 in main text.

S2.4. Assessment of Consistency Between Datasets Obtained

The three datasets analyzed show similar key metrics (Table S8) and good correlations between the computed relative enrichment (r_i) of individual sequences (Figure S4). This indicates a high level of consistency between the three datasets that supports the conclusions drawn.

Table S8. Comparison of Key Metrics Between the Three Datasets.

Parameters	Library	Dataset 1		Dataset 2		Dataset 3	
		Bound	Unbound	Bound	Unbound	Bound	Unbound
Minimum Count	7	0	1	0	0	0	1
Maximum Count[(a)]	730	311	478	261	570	942	4996
Minimum Enrichment		0.00	0.137	0.00	0.00	0.00	0.091
Maximum Enrichment		21.0	4.32	11.5	3.98	14.0	9.87
Minimum r_i[(b)]		-4.06		-4.32		-4.25	
Maximum r_i[(b)]		6.82		4.30		5.44	
$r_i > 0$		29.2%		36.0%		30.4%	
$r_i < 0$		70.8%		64.0%		69.6%	

(a) Excludes the match sequence, which represents ~50% of the dataset due to the bias during PCR for sequencing preparation.

(b) Exclude sequences with zero counts in either the bound or unbound that render undefined enrichment values.

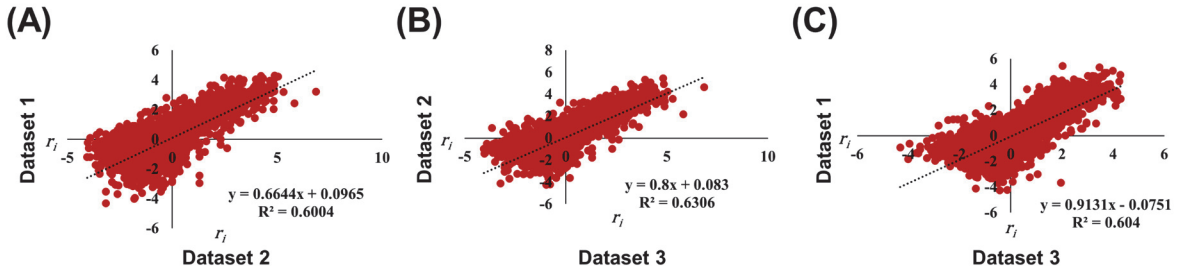


Figure S4: Correlation of r_i for all 4095 unique mismatching sequences in different data sets. (a) Correlation between dataset 1 and dataset 2 gives an R^2 of 0.6004. (b) Correlation between dataset 1 and dataset 3 gives an R^2 of 0.6306. (c) Correlation between dataset 2 and dataset 3 gives an R^2 of 0.604.

S3. Additional Data on Characterization of Mismatches

S3.1. Examples of Computing Relative Enrichment, $r(G_{feat})$, for a Group of Sequences

Table S9 shows an example of calculating *r(1MM)*, the relative enrichment for the group of sequence containing a total of one mismatch. Table S10 shows an example of calculating $r(C_1)_{2MM}$, the relative enrichment for the group of sequence containing a total of two mismatches and with a dT/dC$_1$ mismatch at the PAM+1 position.

Table S9. Example of Calculating *r*(1MM).[a]

Seq#	Seq[b]	$p_i^B(\times 10^{-4})$	E_B^i	$p_i^B \times E_B^i(\times 10^{-4})$	$p_i^U(\times 10^{-4})$	E_U^i	$p_i^U \times E_U^i(\times 10^{-4})$
2	CGGAGT	2.561	0.7267	1.828	2.622	0.7574	1.986
3	GGGAGT	3.910	0.7021	2.745	5.946	1.068	6.348
4	TGGAGT	2.924	0.5819	1.701	4.619	0.9191	4.245
5	AAGAGT	2.924	0.5912	1.729	3.626	0.7331	2.658
6	ACGAGT	2.414	0.9314	2.248	4.079	1.574	6.420
7	ATGAGT	20.06	3.271	65.61	24.40	3.978	97.06
8	AGAAGT	3.162	0.8854	2.800	3.432	0.9609	3.297
9	AGCAGT	1.632	0.6043	0.9862	4.262	1.578	6.728
10	AGTAGT	32.03	5.865	187.9	53.91	9.873	532.3
11	AGGCGT	4.692	0.9210	4.321	14.17	2.781	39.41
12	AGGGGT	5.472	1.218	6.541	6.788	1.538	10.44
13	AGGTGT	12.85	2.106	27.06	41.46	6.793	281.6
14	AGGAAT	3.468	0.9082	3.150	4.090	1.071	4.380
15	AGGACT	0.748	0.2886	0.2159	1.198	0.4622	0.5536
16	AGGATT	1.768	0.3546	0.6270	3.453	0.6926	2.392
17	AGGAGA	1.156	0.2342	0.2707	2.201	0.4460	0.9817
18	AGGAGC	1.020	0.2261	0.2306	2.212	0.4904	1.085
19	AGGAGG	3.808	0.6331	2.411	8.460	1.407	11.90
	Sum	106.5	--	312.3	190.9	--	1014

$$r(1MM) = log_2\left[\frac{\frac{\sum_{i\in G}(p_i^B \times E_B^i)}{\sum_{i\in G}(p_i^B)}}{\frac{\sum_{i\in G}(p_i^U \times E_U^i)}{\sum_{i\in G}(p_i^U)}}\right] = log_2\left(\frac{312.3\times10^{-4}/106.5\times10^{-4}}{1014\times10^{-4}/190.9\times10^{-4}}\right) = -0.8558$$

(a) Computed following eq. 5 described in main text Methods.

(b) The red nucleotide indicates position of mismatch.

Table S10. Example calculation of $r(C_1)_{2MM}$.[a]

Seq#	Seq[b]	$p_i^B(\times 10^{-4})$	E_B^i	$p_i^B \times E_B^i(\times 10^{-4})$	$p_i^U(\times 10^{-4})$	E_U^i	$p_i^U \times E_U^i(\times 10^{-4})$
20	CAGAGT	0.7140	0.7001	0.500	1.025	1.006	1.031
21	CCGAGT	1.190	1.670	1.988	0.809	1.136	0.920
22	CTGAGT	1.156	0.8172	0.945	1.209	0.8544	1.033
29	CGAAGT	1.258	1.285	1.616	0.885	0.904	0.800
30	CGCAGT	3.366	4.003	13.47	0.669	0.796	0.532
31	CGTAGT	3.910	2.972	11.62	1.263	0.960	1.212
38	CGGCGT	0.4420	0.7325	0.3238	0.5072	0.8405	0.4263
39	CGGGGT	0.5440	0.9648	0.5248	0.6798	1.206	0.8197
40	CGGTGT	1.292	0.9970	1.288	1.888	1.457	2.752
47	CGGAAT	0.3740	0.4396	0.1644	0.8849	1.040	0.9204
48	CGGACT	0.3060	0.4549	0.1392	0.7014	1.043	0.7314
49	CGGATT	0.5440	0.4135	0.2249	1.586	1.206	1.913
56	CGGAGA	0.2380	0.4221	0.1005	0.4424	0.7847	0.3472
57	CGGAGC	0.1020	0.2022	0.0206	0.4964	0.9839	0.4884
58	CGGAGG	0.4420	0.8593	0.3798	0.3993	0.7762	0.3099
	Sum	15.88	--	33.31	13.45	--	14.24

$$r(C_1)_{2MM} = log_2\left[\frac{\frac{\sum_{i\in G}\left(p_i^B \times E_B^i\right)}{\sum_{i\in G}(p_i^B)}}{\frac{\sum_{i\in G}(p_i^U \times E_U^i)}{\sum_{i\in G}(p_i^U)}}\right] = log_2\left(\frac{33.31\times10^{-4}/15.88\times10^{-4}}{14.24\times10^{-4}/13.45\times10^{-4}}\right) = 0.9866$$

(a) Computed following eq. 5 described in main text Methods.

(b) The red nucleotide indicates the dT/dC$_1$ mismatch at the PAM+1 position. The blue nucleotide indicates position of remaining mismatch within sequence.

S3.2. Examining Consistency of $r(G_{feat})$ Between Multiple Datasets

S3.2.1. Consistency of $r(\#MM)$ Values Between Datasets

$r(\#\mathrm{MM})$ values were computed for all three datasets obtained (see extended data file "Supplementary Data.xlsx"). Figure S5 shows the comparisons in a pair-wised fashion, which indicate a high-degree of correlation among all three datasets. In addition, averaged <$r(\#\mathrm{MM})$> values and the corresponding standard deviations were computed with all three datasets. As shown in Figure S6A, the averaged <$r(\#\mathrm{MM})$> values show exactly the same feature as those shown in main text Figure 3, that: (i) $r(1\mathrm{MM})$ and $r(2\mathrm{MM})$ are negative; (ii) $r(3\mathrm{MM})$ is slightly positive; and (iii) $r(4\mathrm{MM})$, $r(5\mathrm{MM})$, and $r(6\mathrm{MM})$ are positive. Furthermore, t-test analysis indicates that $r(4\mathrm{MM})$, $r(5\mathrm{MM})$ and $r(6\mathrm{MM})$ are significantly larger than $r(1\mathrm{MM})$ and $r(2\mathrm{MM})$ (Fig. S6B).

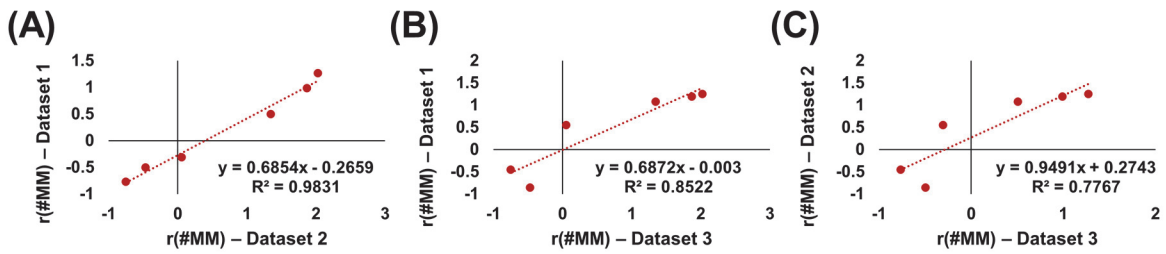


Figure S5: Correlation between the datasets on the analysis of the total number of mismatches. The enrichment values (*r*(#MM)) when the sequences are grouped based on the total number of mismatches were computed as described in eq. 5 in the main text. (A) Correlation between dataset 1 and dataset 2 gives an R^2 of 0.9831. (B) Correlation between dataset 1 and dataset 3 gives an R^2 of 0.8522. (C) Correlation between SELEX dataset 2 and dataset 3 gives an R^2 of 0.7767.

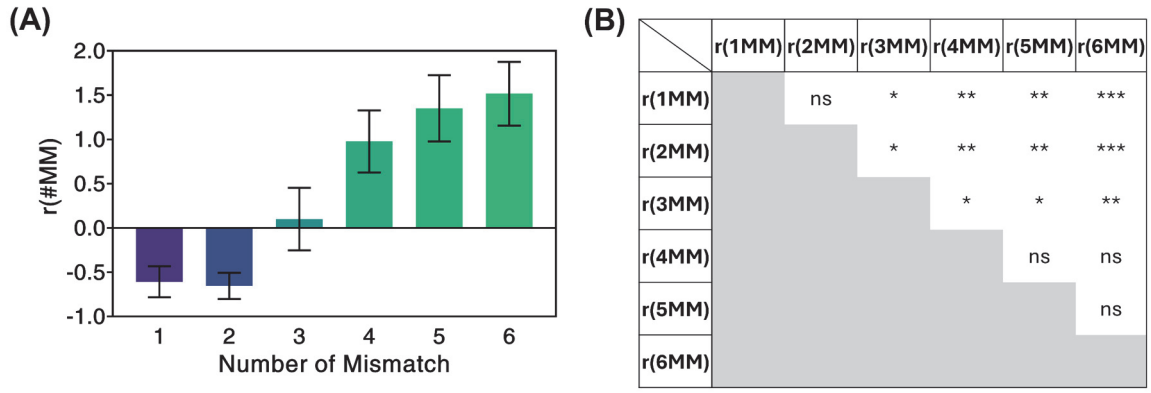
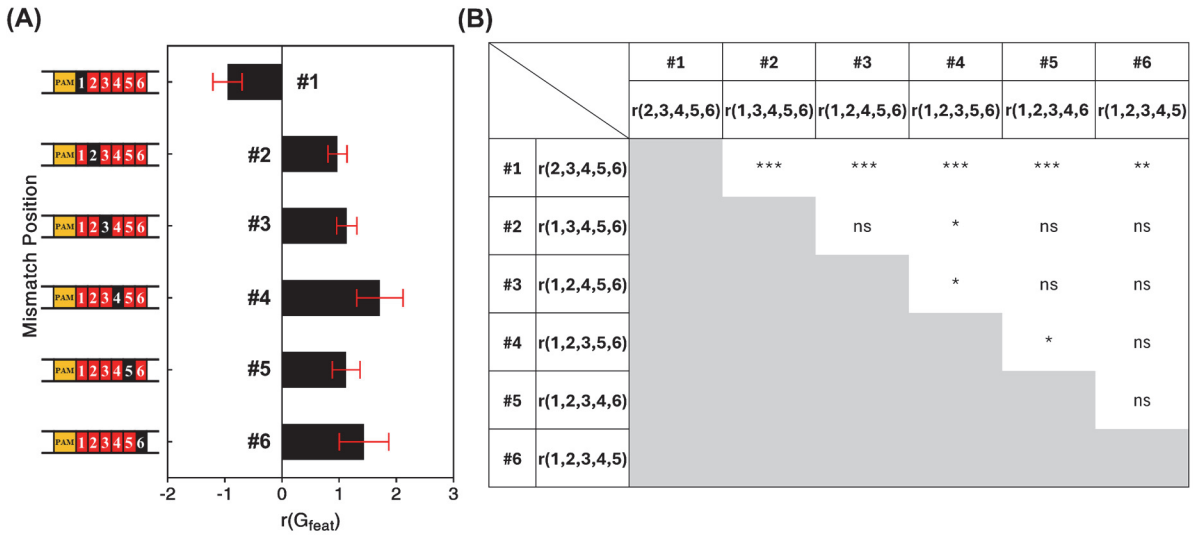


Figure S6: (A) Averages of *r*(#MM) obtained from all three datasets with error bars representing the corresponding standard deviations. (B) Unpaired *t*-test regarding significances of pair-wise differences between <*r*(#MM)> values. Results were obtained with Excel, and *p* (one-tail) values were represented as: **** $p \leqslant 0.0001$, *** $0.0001 \leqslant p \leqslant 0.001$, ** $0.001 \leqslant p \leqslant 0.01$, * $0.01 \leqslant p \leqslant 0.1$, ns $p > 0.1$.

S3.2.2. Consistency of $r(G_{feat})$ Values of the 5MM Subgroups Among the Three Datasets

Averaged <$r(G_{feat})$> values and the corresponding standard deviations were computed for the 5MM subgroup with all three datasets. As shown in Figure S7A, the averaged <$r(G_{feat})$> values show the same features as those shown in main text Figure 4A, that <$r(2,3,4,5,6)$> (Fig. S7A, "#1") is negative while all others are positive, and <$r(1,2,3,5,6)$> (Fig. S7A, "#4") has the highest positive value. Furthermore, t-test analysis (Fig. S7B) indicates that <$r(2,3,4,5,6)$> (Fig. S7B, "#1") is significantly smaller than all the others. Among the subgroups with positive <$r(G_{feat})$>, <$r(1,2,3,5,6)$> (Fig. S7, "#4") can be considered larger than <$r(1,3,4,5,6)$> (Fig. S7, "#2") and <$r(1,2,4,5,6)$> (Fig. S7, "#3"), but differences between <$r(1,2,3,5,6)$> ("#4"), <$r(1,2,3,4,6)$> ("#5"), and <$r(1,2,3,4,5)$> ("#6") are not significant. This supports analyses presented in the main text.



		#1 r(2,3,4,5,6)	#2 r(1,3,4,5,6)	#3 r(1,2,4,5,6)	#4 r(1,2,3,5,6)	#5 r(1,2,3,4,6	#6 r(1,2,3,4,5)
#1	r(2,3,4,5,6)		***	***	***	***	**
#2	r(1,3,4,5,6)			ns	*	ns	ns
#3	r(1,2,4,5,6)				*	ns	ns
#4	r(1,2,3,5,6)					*	ns
#5	r(1,2,3,4,6)						ns
#6	r(1,2,3,4,5)						

Figure S7: (A) Averages of $r(G_{feat})$ values of the 5MM subgroup obtained from all three datasets with error bars representing the corresponding standard deviations. (B) Unpaired *t*-test regarding significances of pair-wise differences between <$r(G_{feat})$> values. Results were obtained with Excel, and *p* (one-tail) values were represented as: **** $p \leqslant 0.0001$, *** $0.0001 \leqslant p \leqslant 0.001$, ** $0.001 \leqslant p \leqslant 0.01$, * $0.01 \leqslant p \leqslant 0.1$, ns $p > 0.1$.

S3.3. Analyzing Preference of the Location of Mismatches for the Group of Sequences Containing a Total of Four Mismatches

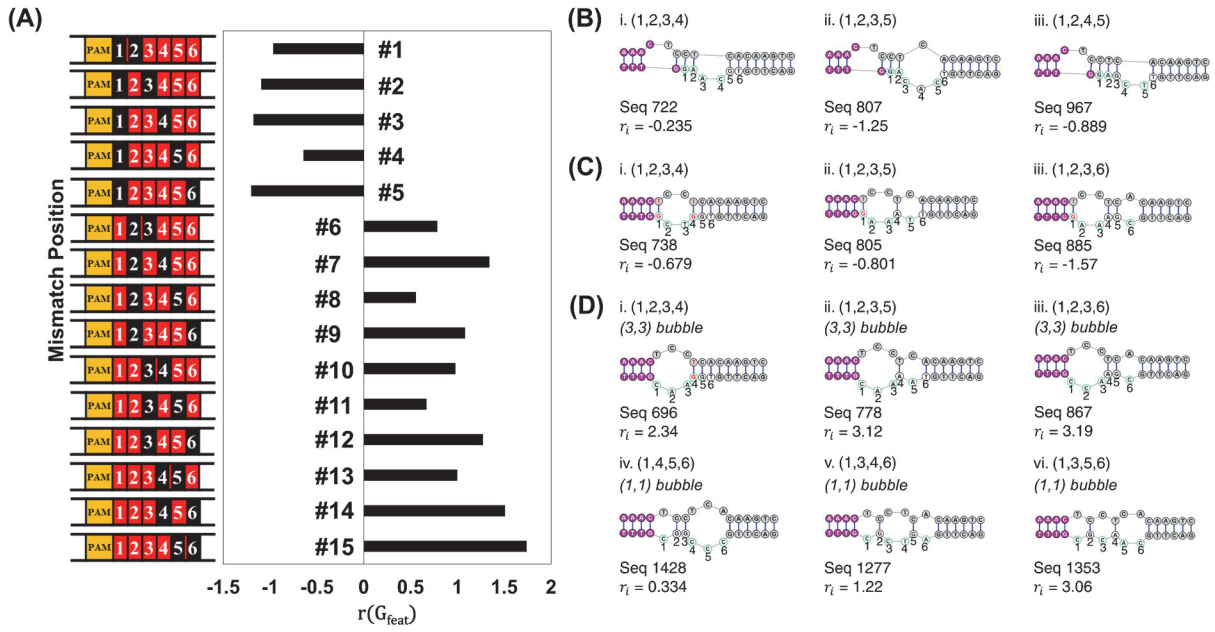


Figure S8: Analysis of mismatch locations of 4MM sequences. (A) $r(G_{feat})$ obtained with dataset 3 for different mismatch arrangements. Numbers 1-6 represent the position of the randomized nucleotides. The mismatch arrangements are represented as: PAM, yellow box; mismatch nucleotide(s), red box(es); and match nucleotide(s), black box(es). (B) Examples of sequences with PAM disrupted. (C) Examples of sequences with a PAM+1 dG/dT wobble pair. (D) Examples illustrating impacts of PAM-adjacent bubble sizes. For panels (B), (C) and (D), pink filled nucleotides indicate the PAM, white filled nucleotides indicate the 6 randomized nucleotides, green circled nucleotides indicate mismatch nucleotides and red text nucleotides show the possible dG/dT pair.

The group of DNA target duplexes containing a total of 4 mismatches (i.e., 4MM) overall is favored for off-target binding (i.e., $r(4\mathrm{MM}) > 0$, main text Fig. 3A and Fig.S6). Detailed analysis on the subgroups with different mismatch position arrangements (Fig. S8 and Fig. S9) revealed features that are completely consistent with those drawn from the 5MM analysis (main text, Figure 4 and related Results). DNA sequences that cause mis-folding of the PAM gave negative r_i (Figure S8B), again supporting the notion that the off-target binding studied in this work are PAM-dependent and therefore is Cas12a specific. Furthermore, negative averaged $r(G_{feat})$ values were observed for the five subgroups with a PAM+1 dT/dA$_1$ paired ("#1" to "#5", Fig. S8A and Fig. S9A), as well as sequences with a PAM+1 dT/dG$_1$ wobble pair (Fig. S8C). This indicates that PAM+1 pairing is not favorable for binding, which is also observed for the 5MM sequences (main text, Figure 4A and 4C). Note that previous biochemical studies with Lb- and AsCas12a observed off-target binding with several sequences with a dT/dG$_1$ wobble pair when the PAM-adjacent bubble is 3 or 4 base-pair,[2] while this work with FnCas12a shows that 4MM sequences with dT/dG$_1$ wobble pair are not favored for binding. This likely arises due to the different Cas12a effector and DNA concentrations used, and also may reflect differences between Cas12a orthologs, which was also reported in the prior work.[2] Overall, all data reported in this work indicate that DNA-DNA pairing at PAM+1 is detrimental for off-target binding by FnCas12a.

Among the other ten subgroups, DNA sequences with a larger PAM-adjacent bubble overall gave a higher degree of binding. For example, the PAM+1,2,3,4 sub-group ("#15", Fig. S8A and Fig. S9A) forms predominately (32%) PAM-adjacent 3-3 bubble (i.e., 3-nt at TS and 3- nt at NTS, Fig. S8D), the PAM+1,2,3,5 ("#14") and PAM+1,2,3,6 ("#13") sub-groups favor the nominally expected 3-3, 1-1 bubble (48.0% and 54.0%, respectively, Fig. S8D), the PAM+1,2,4,5 ("#12") sub-group favors the nominally expected 2-2, 2-2 (30.0%) bubble. All four of these sub-groups show clearly positive $r(G_{feat})$ (Fig. S8A and Fig. S9), indicating they are favorable for binding. On the other hand, the PAM+1,4,5,6 sub-group ("#6", Fig. S8 and Fig. S9), which preferably forms 1-1, 3-3 bubbles (32.0%) (Fig. S8D), has a slightly negative $<r(1,4,5,6)>$ with large variations among the three datasets (Fig. S9), indicating this subgroup has near equal chance of being free or bound. Also note that PAM+1,3,5,6 ("#7") and PAM+1,3,4,6 ("#8") show different $r(G_{feat})$ (Fig. S8A, Fig. 9), although the predicted most populated class for either subgroup adopts a three-bubble pattern (Fig. S8D). With the "three-bubble" patent, the boundaries between bubbles are fluid, and the discrepancy likely indicates limitation on correlating $r(G_{feat})$ with only the most populated class of bubble.

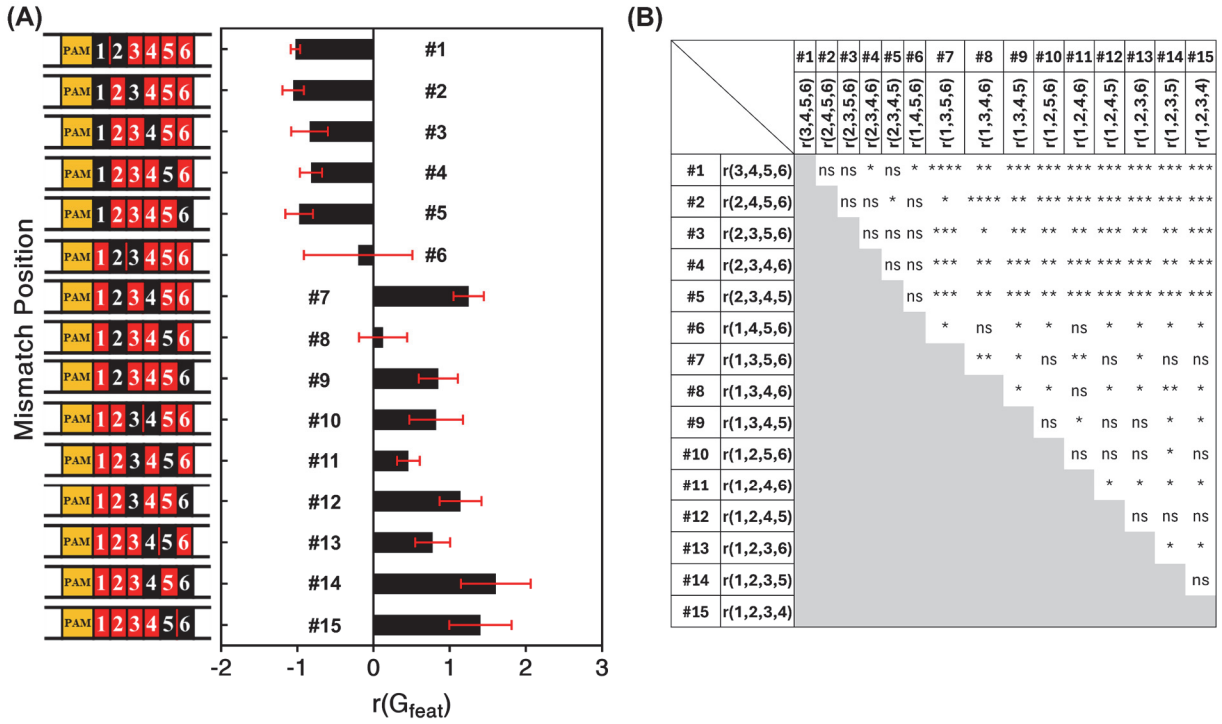


Figure S9: (A) Averages of $r(G_{feat})$ values of the 4MM subgroup obtained from all three datasets with error bars representing the corresponding standard deviations. (B) Unpaired *t*-test regarding significances of pair-wise differences between $<r(G_{feat})>$ values. Results were obtained with Excel, and *p* (one-tail) values were represented as: **** $p \leqslant 0.0001$, *** $0.0001 \leqslant p \leqslant 0.001$, ** $0.001 \leqslant p \leqslant 0.01$, * $0.01 \leqslant p \leqslant 0.1$, ns $p > 0.1$.

Overall, analysis of the 4MM group shows that a favorable off-target: (i) requires a proper PAM as well as unpairing of protospacer at PAM+1; and (ii) increases with unpairing at PAM+2 and +3, with consecutive bubble being the most effective. These are completely consistent with conclusions drawn from the 5MM analysis (main text, Fig. 4).

S3.4. Additional Data on Analysis of Position-Dependent Nucleotide Preferences

Table S11 shows values of the relative enrichment, $r(N_k)_{\#MM}$, for groups of DNA targets with a given nominal total number of mismatches (#MM) and a particular nucleotide (N) at a given position (k). The data were plotted as a heatmap in Figure 5A of the main text.

Table S12 shows the coefficients obtained from a multiple linear regression analysis (main text, Materials and Methods, eq. 6). The data was used to generate the web-logo plot of the position-dependent nucleotide preference for off-target binding shown as main text Figure 5B.

Table S11. $r(N_k)_{\#MM}$ values.

Number of Mismatch (#MM)	Nucleotide (N)	Position (k)					
		1	**2**	**3**	**4**	**5**	**6**
1	A	-0.8377	-0.3104	-0.1180	-0.6927	-0.2379	-0.9293
	C	-0.0597	-0.7568	-1.3850	-1.5944	-0.6793	-1.1169
	G	-0.6047	-0.9322	-1.0727	-0.3374	-0.8480	-1.1517
	T	-0.6595	-0.2825	-0.7513	-1.6897	-0.9658	-0.8665
2	A	-0.8860	-0.8787	-0.7518	-0.3086	-0.8284	-0.7917
	C	0.9866	-0.6011	0.3151	-0.8147	-0.6057	-0.4921
	G	-0.2746	-0.2593	-0.6371	-0.2992	-0.3346	-0.8205
	T	-0.8255	-0.9305	-0.3617	-0.9209	-0.9609	-0.3966
3	A	-0.8034	-0.5234	0.4852	0.7288	-0.0124	0.0549
	C	1.8427	1.1275	1.2449	0.6366	0.6946	0.2018
	G	0.0588	0.8191	0.2983	0.0029	0.7076	-0.0137
	T	-0.6538	0.0279	0.4362	0.1989	0.4735	0.7098
4	A	-1.0427	0.6186	1.4158	1.2163	0.9710	0.5490
	C	2.1536	1.8170	1.0933	1.2100	1.1468	0.6646
	G	-0.1075	0.9586	0.9120	0.8064	1.2050	0.7342
	T	0.0305	0.6923	0.8013	0.8674	0.9998	1.3298
5	A	-1.3034	1.0051	1.5118	1.4801	1.2824	1.0457
	C	2.2253	1.8051	1.1613	1.1809	1.0833	1.0237
	G	-0.7646	0.9918	1.1180	1.0579	1.2271	1.3410
	T	0.2592	0.8039	0.8446	1.0600	1.1524	1.2566
6[(a)]	A		1.0097	1.5189		1.5241	1.0667
	C	2.2712	1.8705	1.2995	1.1157	1.0963	1.0471
	G	-0.9171			1.6330		1.4982
	T	0.3601	0.8464	0.9338	1.0098	0.9886	

(a) Empty cells indicate nucleotides that match the target-strand and therefore cannot fit the criterion of 6 total mismatches.

Table S12. Coefficients (B_N^k) determined for each nucleotide for the 6 positions from the entire dataset.

Position \ Nucleotide	A	C	G	T
1	-0.5608	1.289	-0.5801	-0.1481
2	-0.1282	0.5236	-0.3351	-0.0604
3	0.3243	-0.0245	-0.3364	0.03654
4	0.1226	0.009135	-0.2215	0.08972
5	0.1364	0.01809	-0.2040	0.04960
6	0.03136	0.06194	-0.0898	-0.003504

S3.5. Examining the Correlation Between Binding Preference and Folding Energy.

In this work, for each DNA duplex sequence the relative enrichment (r_i) for Cas12a off-target binding was measured, and the duplex folding free energy for the predicted most stable secondary structure (ΔG_i) was computed (main text, Methods, see examples in Table S1). When analyzing the position-dependent nucleotide preference for off-target binding, it appears that the degree of preferrable binding is higher when the DNA has a higher (less negative) ΔG_i (main text, Figure 5 and related Results section). To further explore whether the duplex folding free energy can serve as a quantitative predictor of Cas12a binding, we computed an average <ΔG(#MM)> for each group of sequences with a particular total number of mismatches (#MM).

$$\Delta G(\#MM) = \frac{1}{n}\sum_{i\in\#MM}^{n} \Delta G_i \qquad \text{(S1)}$$

where *n* is total number of sequences in the group and ΔG_i is the predicted folding energy for the corresponding individual sequence in the group.

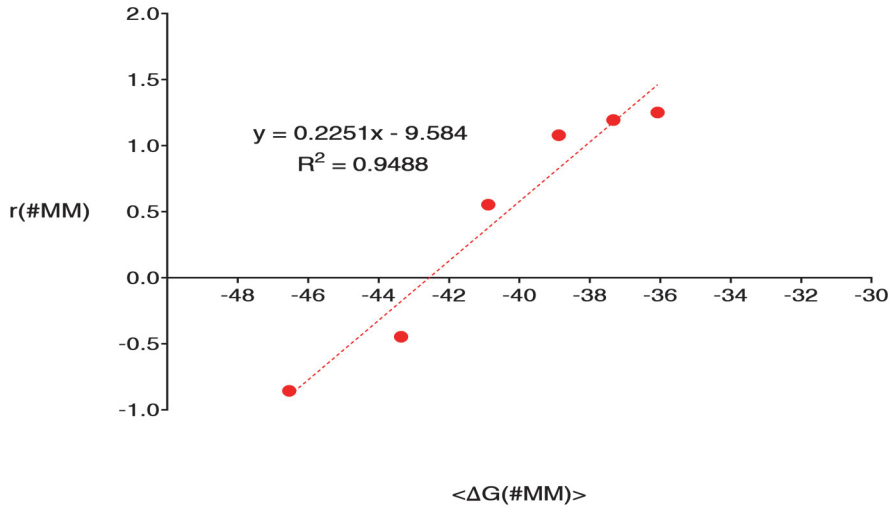


Figure S10. Correlation between $r(\#MM)$ (from dataset 3) and <ΔG(#MM)>.

We then examined possible correlation between <ΔG(#MM)> and $r(\#MM)$, the weighted average enrichment for the groups of sequences containing a particular number of total mismatches (main text, Fig. 3). Interestingly, $r(\#MM)$ was found to show a high degree of positive linear correlation with <ΔG(#MM)> (Fig. S10). Note that with the lack of complementarity between the RNA guide and the DNA target strand in our construct (main text Fig. 1A, Fig. S1), DNA duplex unwinding (to form an RNA/DNA hybrid) is not playing a role in Cas12a binding. Therefore, instead of considering <ΔG(#MM)> as an indicator of strand dissociation, it is more appropriate to regard it as an indicator of the flexibility of the DNA duplexes. The high degree of correlation between $r(\#MM)$ and <ΔG(#MM)> therefore is consistent with the notion that DNA flexibility is correlated to Cas12a off-target binding.

However, when all sequences were considered together without sub-group classification, no correlation was found between the individual r_i and ΔG_i (Fig. S11), indicating that ΔG_i cannot serve as a quantitative predictor for Cas12a off-target binding. With further consideration, it becomes clear that without any classification, ΔG_i, which measures individual duplex stability,

cannot capture the requirements for Cas12a off-target binding revealed in this work. Specifically, as revealed in location preference analyses of 5MM and 4MM, sequences with disrupted PAM (main text, Fig. 4D; Fig. S8B) or PAM+1 pairing (main text Fig. 4C; Fig. S8C) are not favors for off-target binding, but they can have high ΔG_i (i.e., overall flexible). Furthermore, even if one

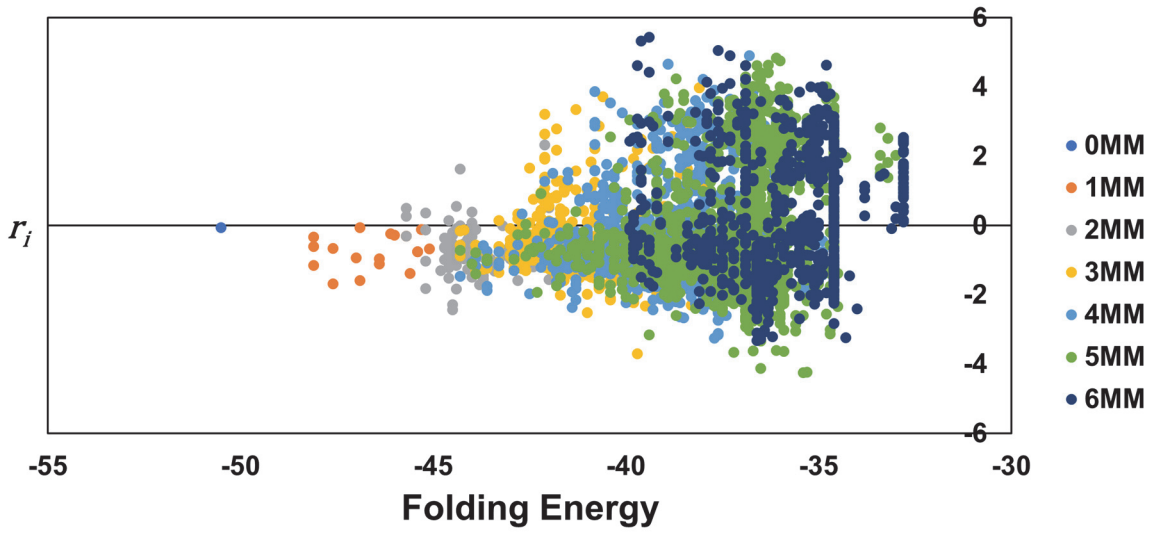


Figure S11. Plot of r_i vs. ΔG_i for each unique sequence in the DNA library. The sequences are color identified by the number of mismatches (#MM) in the sequence. As the #MM increases the sequences shifts to the right correlating with increasingly positive ΔG_i. However, for the entire group, there is no correlation between individual r_i and ΔG_i.

excludes those sequences with a disrupted PAM and PAM+1 pairing, ΔG_i still cannot serve as a general predictor to properly capture the PAM-adjacent bubble feature. For example, the 4MM DNA Seq 1428 has a higher ΔG_i than that of Seq 696 (-39.0 kcal/mol vs. -40.8 kcal/mol, respectively) and is less stable (more flexible) in the context of the overall duplex. However, Seq 696 has a PAM-adjacent 3-3 bubble, while Seq 1428 has the 3-3 bubble located away from PAM (Fig. S8D). Consequently, Seq 696 has a higher r_i than Seq 1428 and is more preferrable for off-target binding (Fig. S8D).

Overall, the SELEX data demonstrates that "PAM-adjacent flexibility" is the factor that dictates off-target binding. This includes features of (i) an intact PAM, (ii) flexibility at the PAM+1 position and (iii) contributions of PAM+2 and +3 to topology. The ΔG_i metric as computed in this work cannot properly reflect such positional dependent DNA flexibility, and therefore does not serve as generalizable quantitative indicator of off-target binding. Further work is needed to develop a quantitative metric for predicting of off target binding.

Appendix: Original Gel images

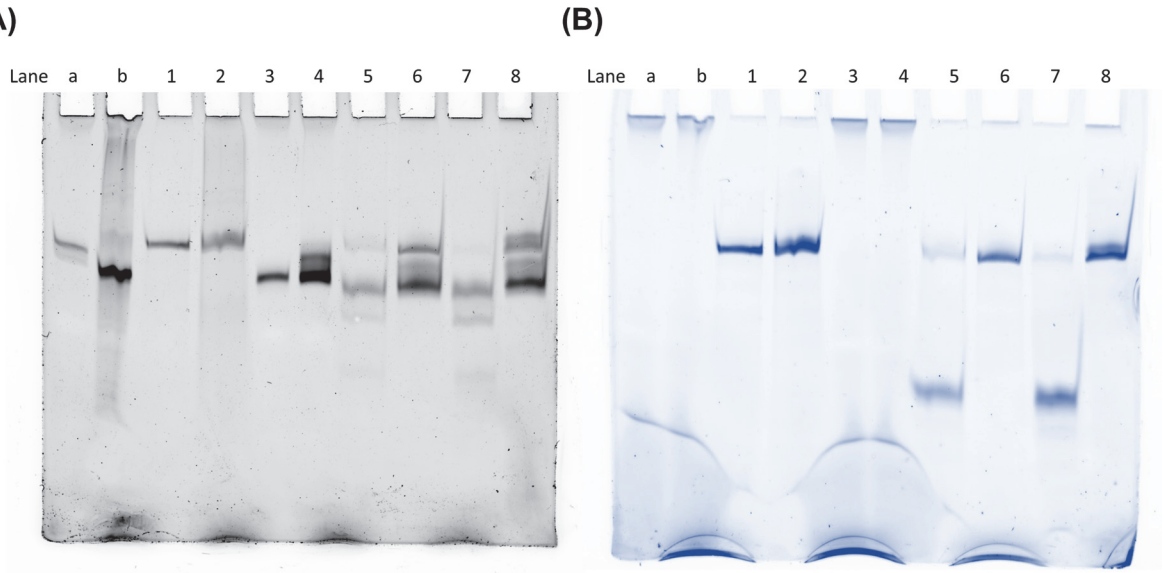


Figure S12. Original gel images for SYBR Gold (A) and FAM (B) imaging panels shown in Figure S2. Lanes numbered as 1 to 8 correspond to the lanes with the same number shown in Figure S2. Lanes "a" and "b" are extra lanes for samples not related to this experiment. In lanes a, b, 3 and 4, inclusion of dyes (Bromophenol Blue and Xylene Cyanol) with the loaded samples gives rise to a visible uneven dye front near the bottom of the gel, as well as possible artifact bands in the well.

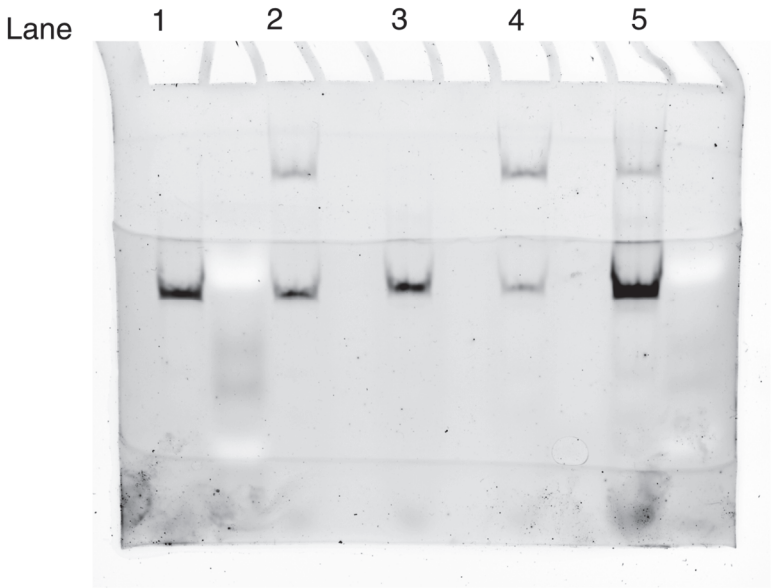


Figure S13. Original gel images for Figure S3. Lanes are numbered exactly the same as that in Figure S3.

References

1. Yamano, T. *et al.* Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* **165**, 949–962 (2016).
2. Jiang, W. *et al.* CRISPR-Cas12a Nucleases Bind Flexible DNA Duplexes without RNA/DNA Complementarity. *ACS Omega* **4**, 17140–17147 (2019).
3. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).