

Supplemental Information

**Rational and computation-assisted engineering
of a compact and efficient
CRISPR-Cas12f genome editor**

Jingtong Liu, Shengzhou Wang, Bei Wang, Tonglu Yu, Xuan Liu, Miaomiao Li, Min-Chi Ku, Rui Wang, Renjie Chai, Wei V. Zheng, and Yongming Wang

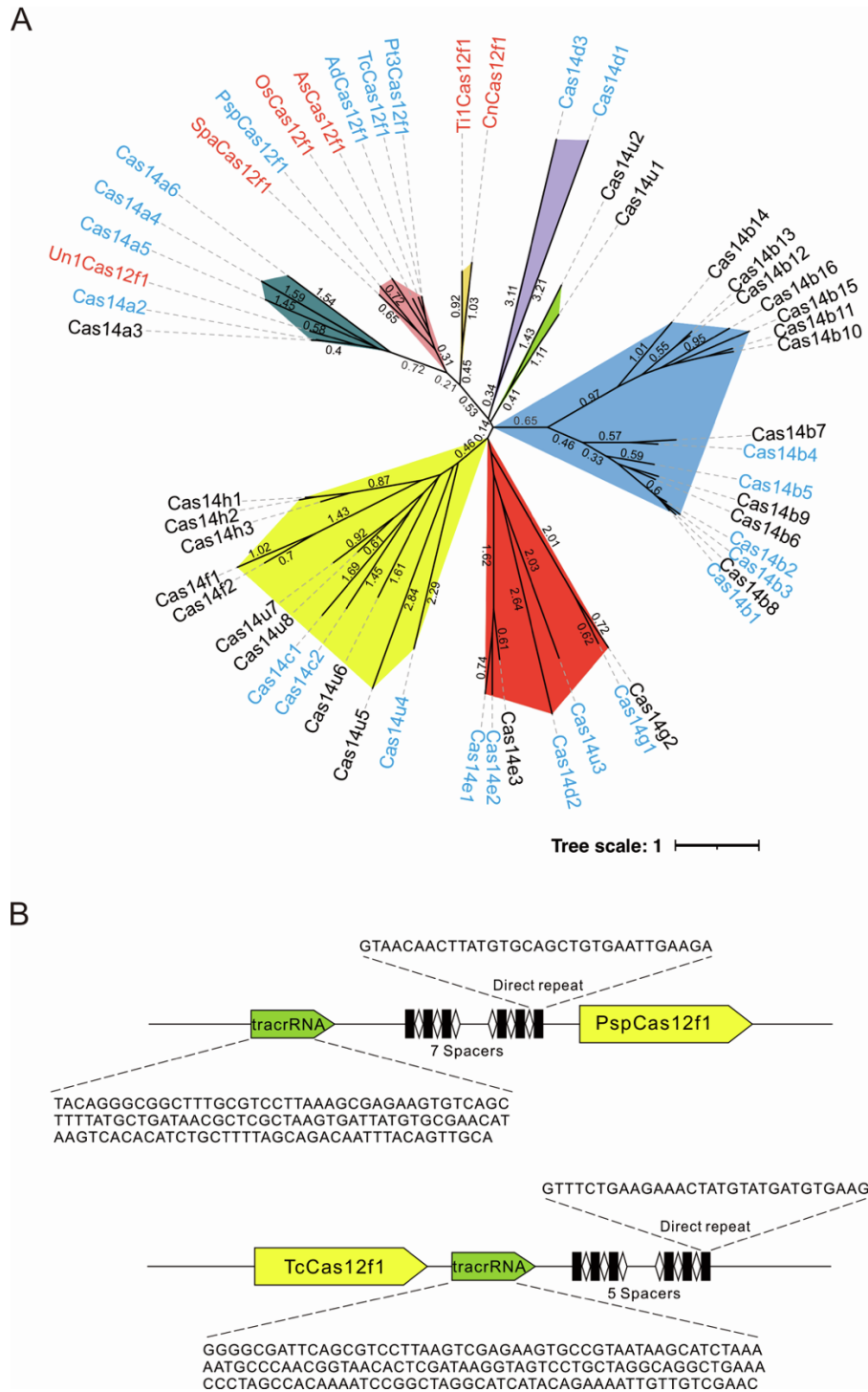


Figure S1. Analysis of newly identified Cas12f orthologs. (A) Phylogenetic tree of newly identified and previously reported Cas12f orthologs. Previously characterized Cas12f orthologs are highlighted in red, whereas the Cas12f orthologs characterized in this study are highlighted in blue. (B) Representative genomic loci and predicted tracrRNA sequences of PspCas12f1 and TcCas12f1, shown from top to bottom. For each locus, repeat sequences are displayed above, and corresponding tracrRNA sequences are shown below.

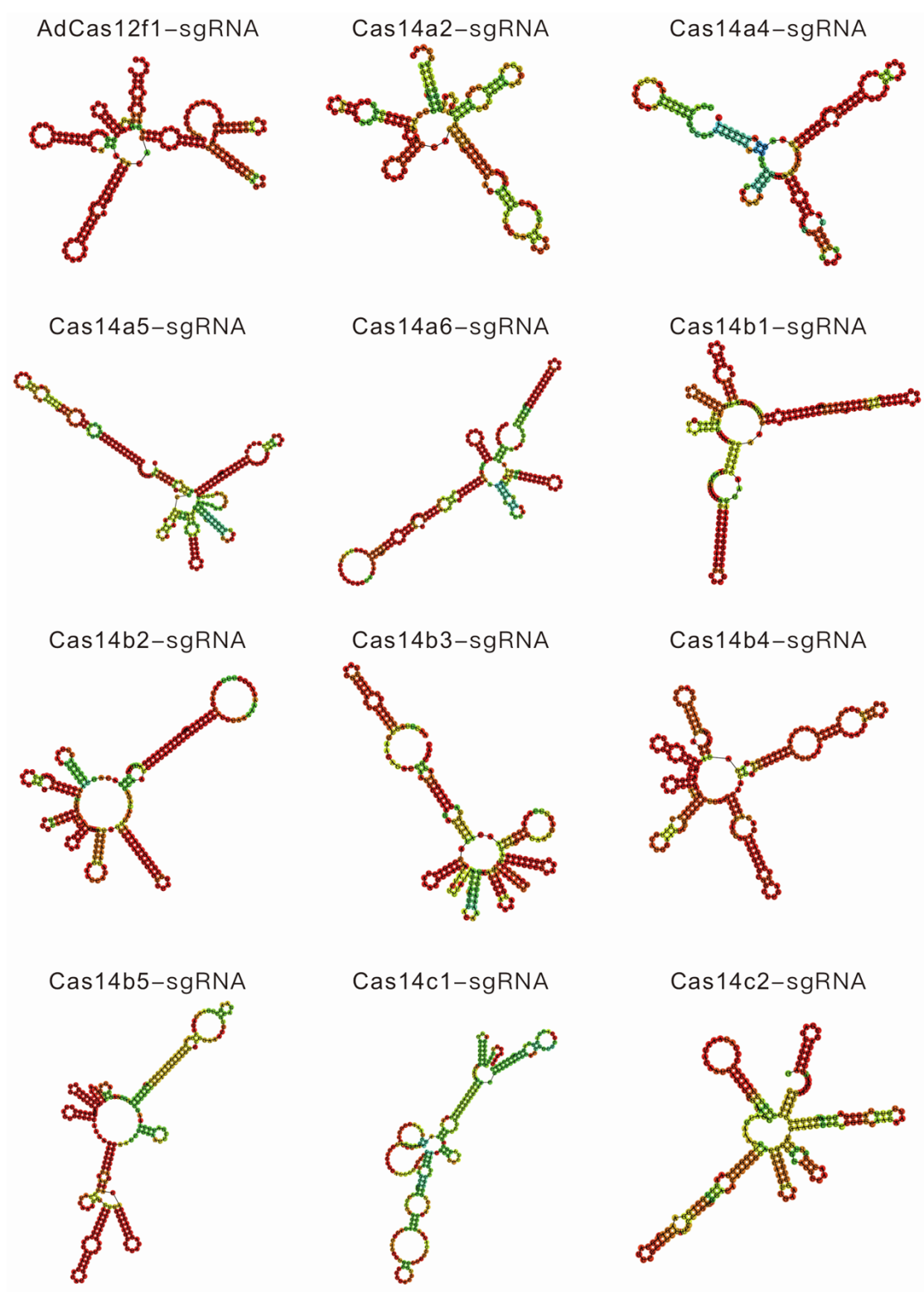


Figure S2. Predicted sgRNA secondary structures of 12 Cas12f orthologs. Secondary structures of sgRNAs from 12 Cas12f orthologs were predicted using the RNAfold Web Server (<http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>).

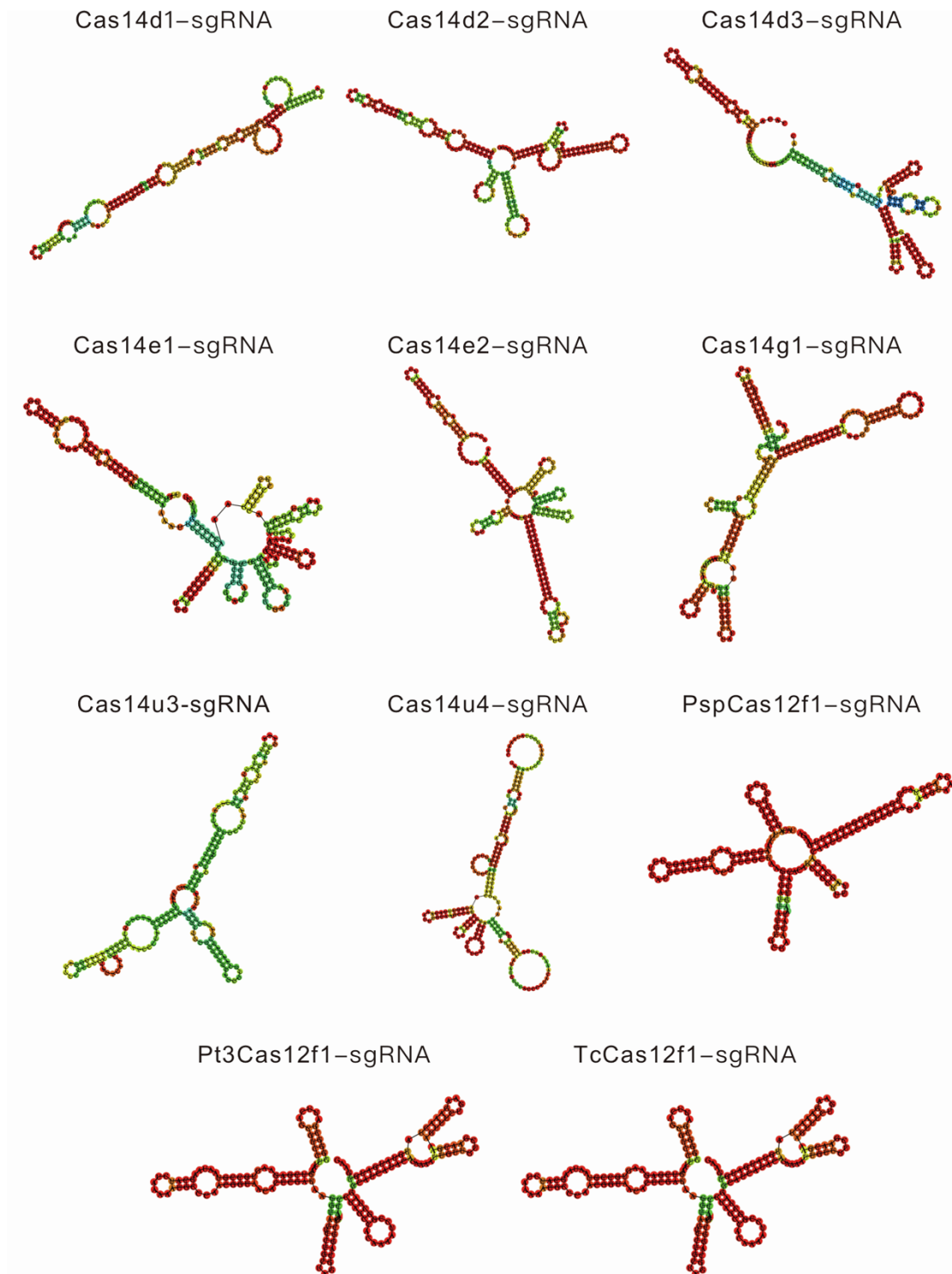


Figure S3. Predicted sgRNA secondary structures of 11 Cas12f orthologs. Secondary structures of sgRNAs from 11 Cas12f orthologs were predicted using the RNAfold Web Server (<http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>).

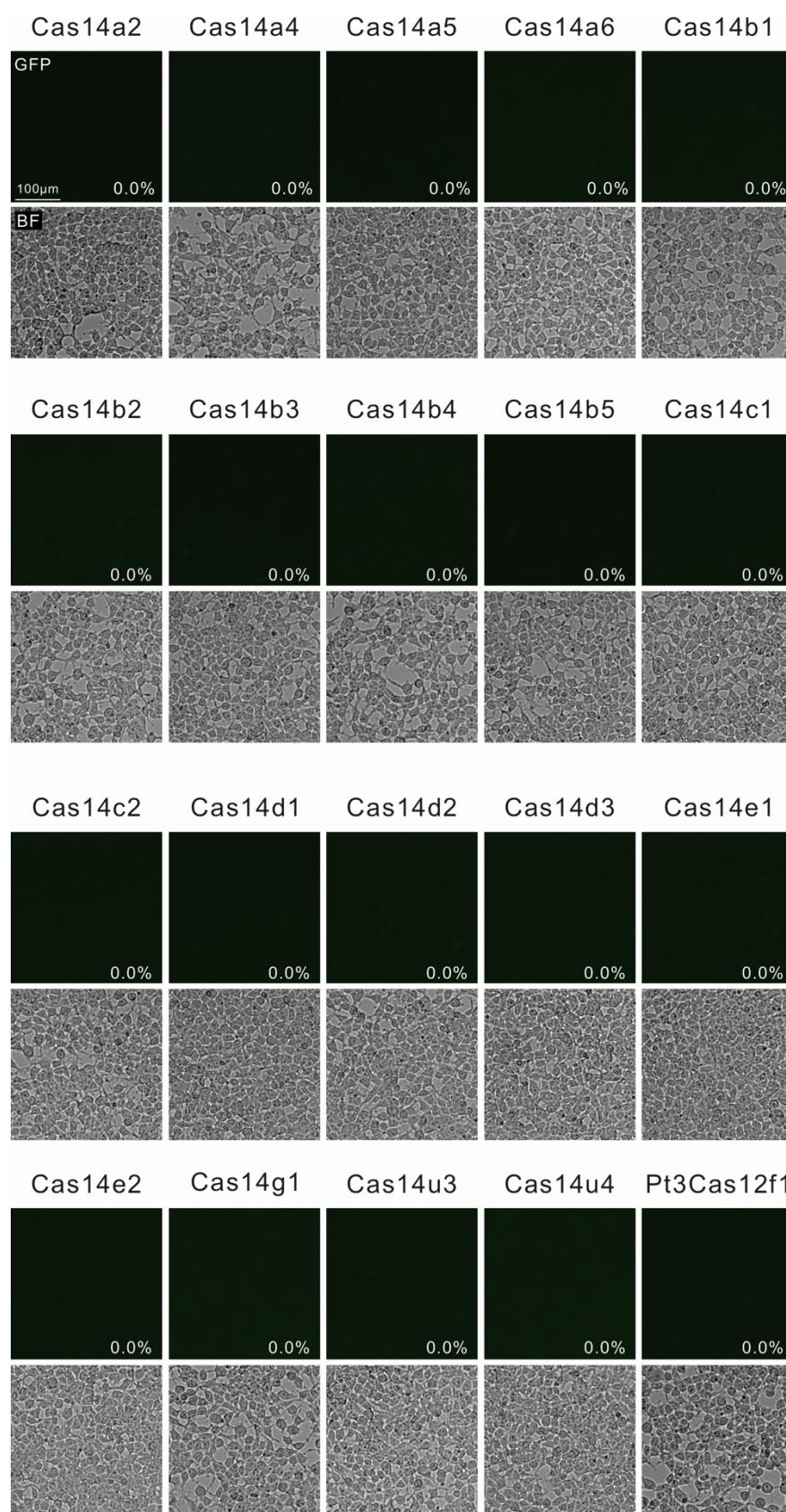


Figure S4. Test of Cas12f ortholog activity using a GFP-activation reporter system. Twenty Cas12f orthologs did not induce GFP expression upon transfection. Percentages of GFP-positive cells are shown.

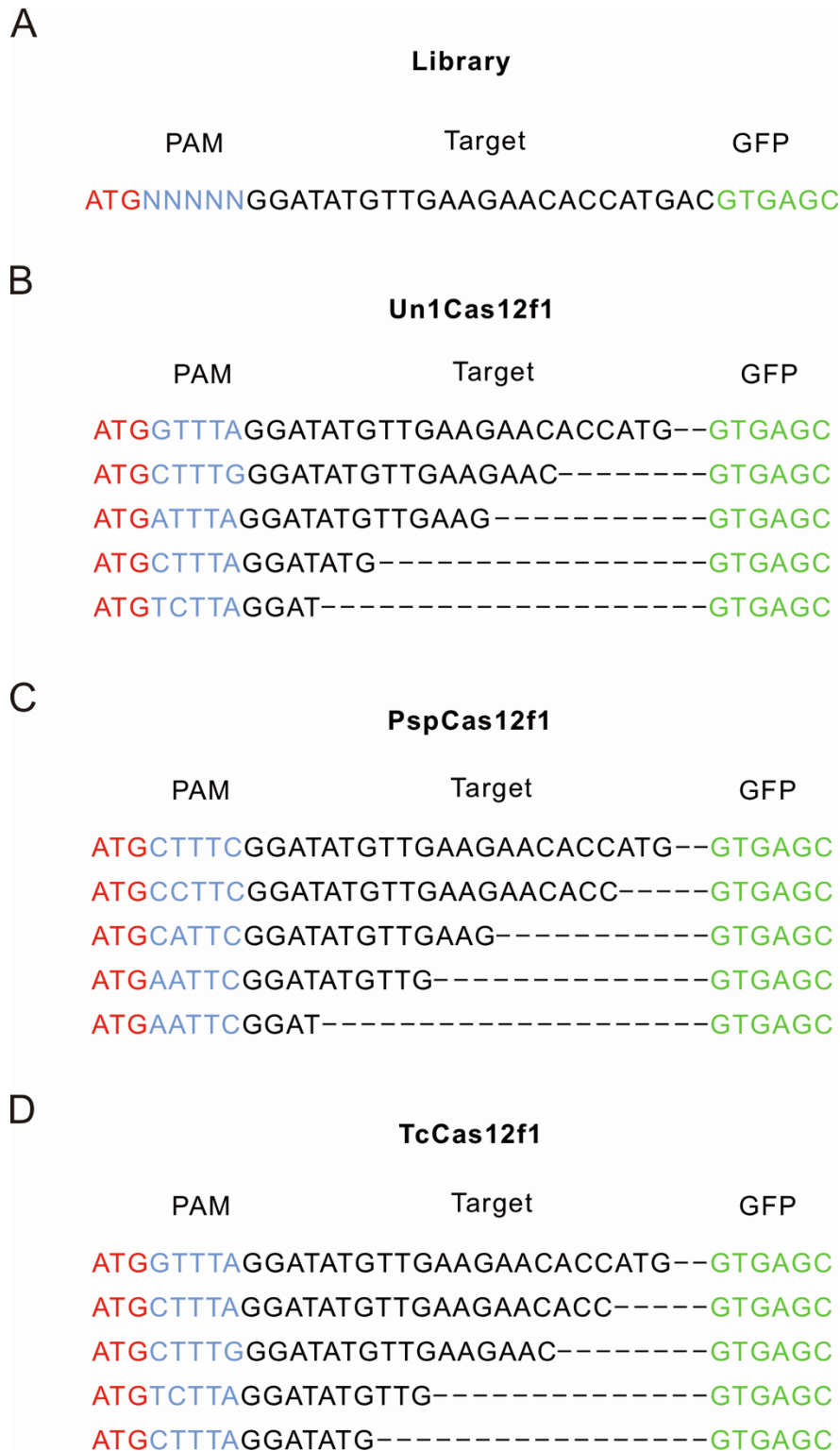


Figure S5. Deep sequencing analysis of indel formation at target sites by Cas12f orthologs. (A) Schematic of the reporter library sequence, with the ATG start codon shown in red, the PAM sequence in blue, and the GFP coding region in green. (B) Indel profiles at target sites for Un1Cas12f1. (C) Indel profiles at target sites for PspCas12f1. (D) Indel profiles at target sites for TcCas12f1.

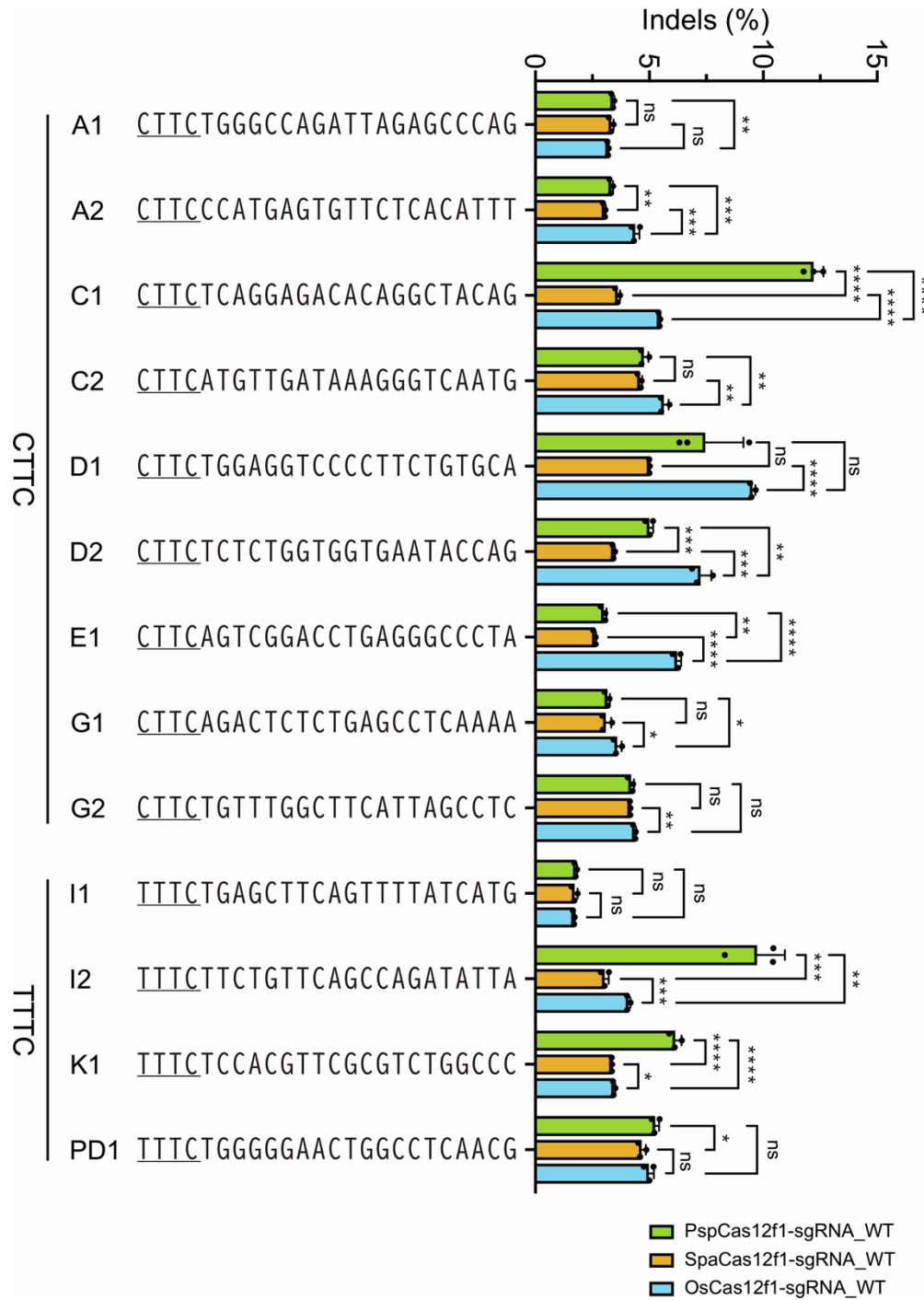


Figure S6. Comparison of the genome-editing efficiencies of PspCas12f1, SpaCas12f1, and OsCas12f1. Genome-editing efficiencies of PspCas12f1, SpaCas12f1, and OsCas12f1 were evaluated at 13 endogenous genomic loci in HEK293T cells, including nine loci with a CTTC PAM and four loci with a TTTC PAM. Indel frequencies were quantified by targeted deep sequencing. Data are presented as mean $\pm$ standard deviation (s.d.) from three independent biological replicates ($n = 3$). Statistical significance was determined with a threshold of $P < 0.05$ (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$).

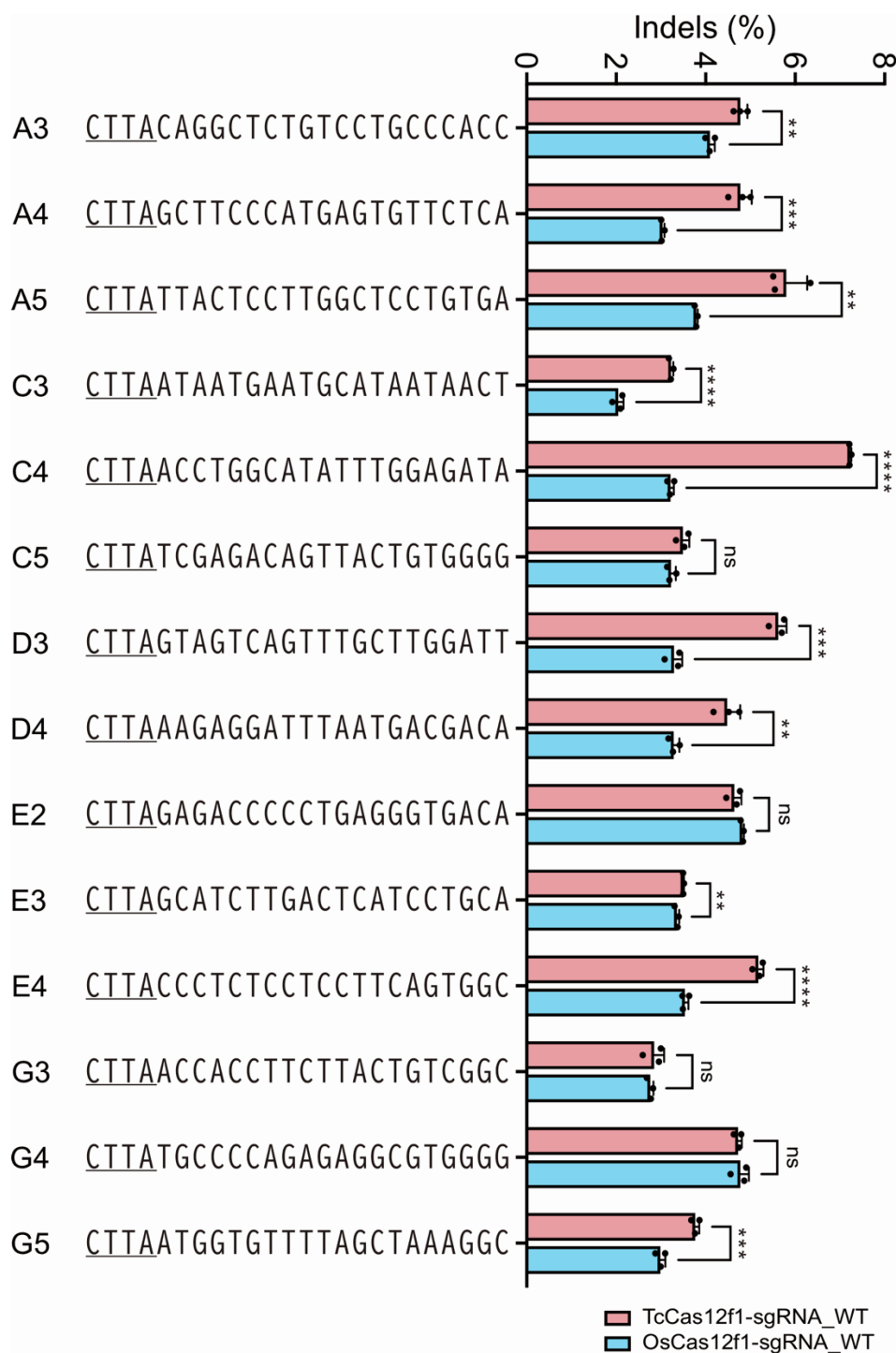


Figure S7. Comparison of the genome-editing efficiencies of TcCas12f1 and OsCas12f1. Genome-editing efficiencies of TcCas12f1 and OsCas12f1 were evaluated at 14 endogenous genomic loci in HEK293T cells. Indel frequencies were quantified by targeted deep sequencing. Data are presented as mean $\pm$ standard deviation (s.d.) from three independent biological replicates ($n = 3$). Statistical significance was determined with a threshold of $P < 0.05$ (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$).

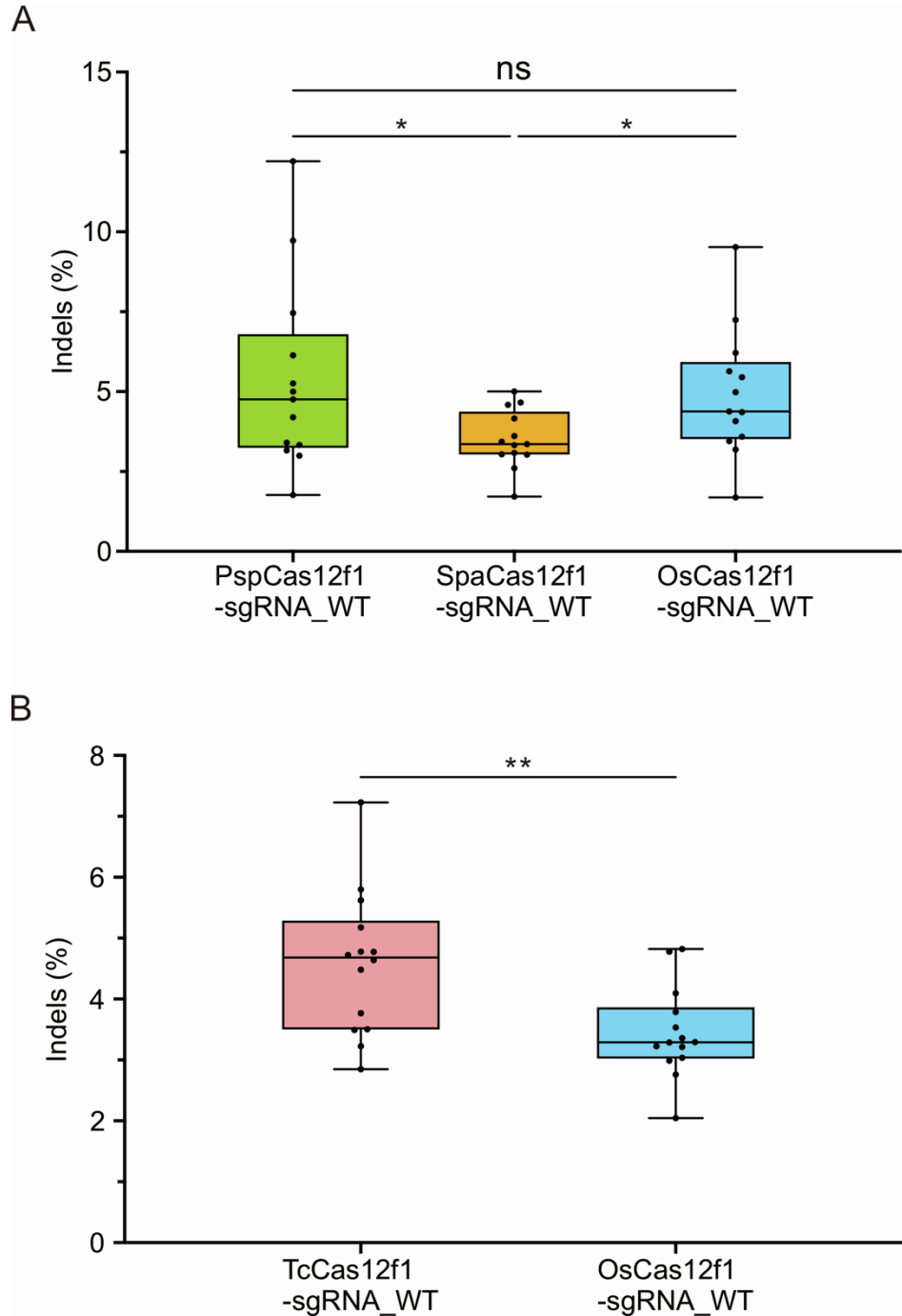


Figure S8. Comparison of the genome-editing efficiencies of PspCas12f1, SpaCas12f1, TcCas12f1 and OsCas12f1. (A) Boxplots of indel frequencies generated by PspCas12f1, SpaCas12f1, and OsCas12f1. (B) Boxplots of indel frequencies generated by TcCas12f1 and OsCas12f1. Boxes represent the interquartile range, the center line indicates the median, and whiskers denote the minimum and maximum values. Each data point represents the mean indel frequency at one endogenous target site from three independent biological replicates. Statistical significance was determined with a threshold of $P < 0.05$ (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$).

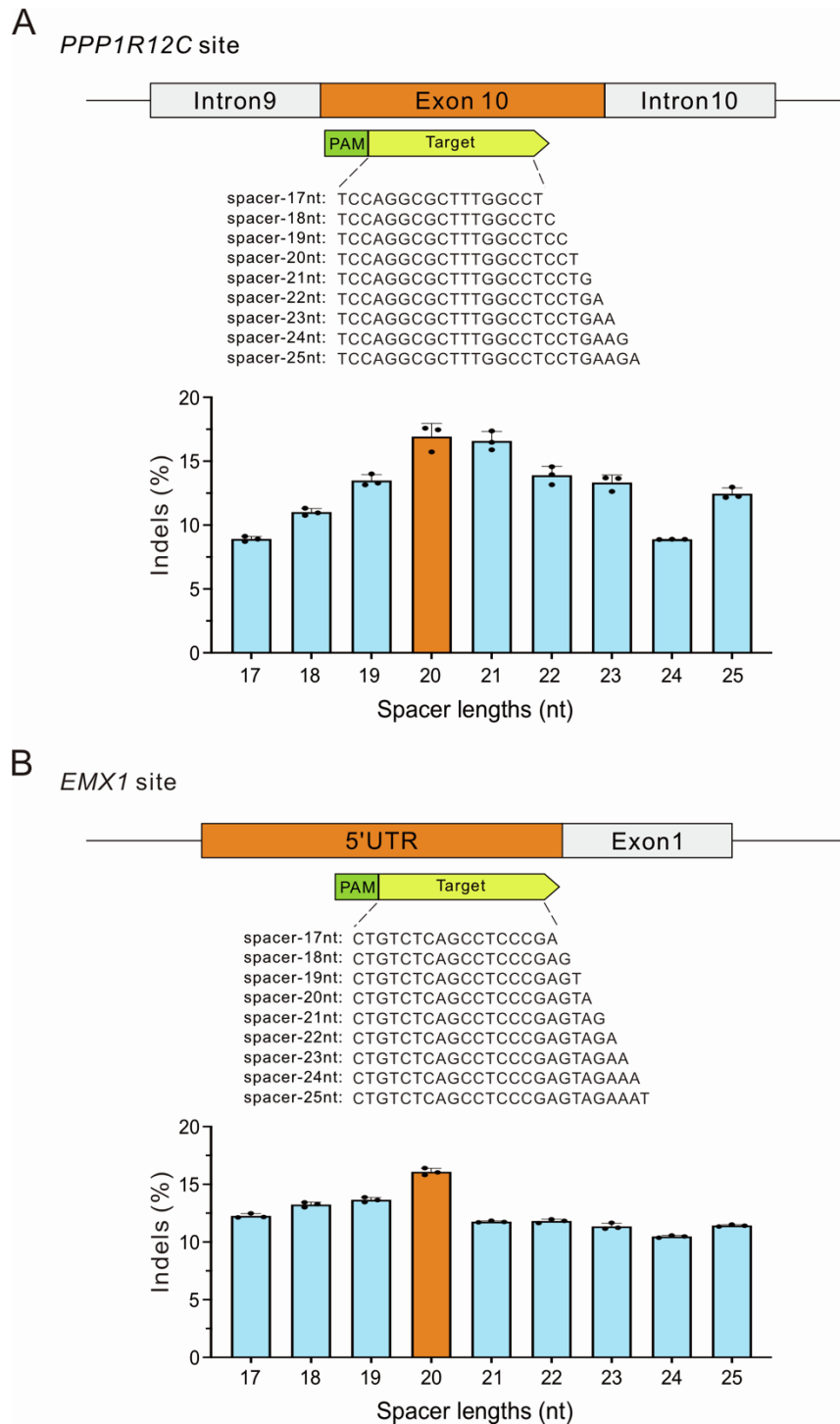


Figure S9. Optimization of spacer length. (A–B) Schematics of the target sequences and PAMs for the *PPP1R12C* (A) and *EMX1* (B) loci. Spacer sequences ranging from 17 to 25 nucleotides were designed for each target site and are shown below the corresponding target. Editing efficiencies associated with different spacer lengths are displayed in the lower panels, with the most efficient spacer length highlighted in orange. Data are presented as mean $\pm$ standard deviation (s.d.) from three independent experiments ($n = 3$).

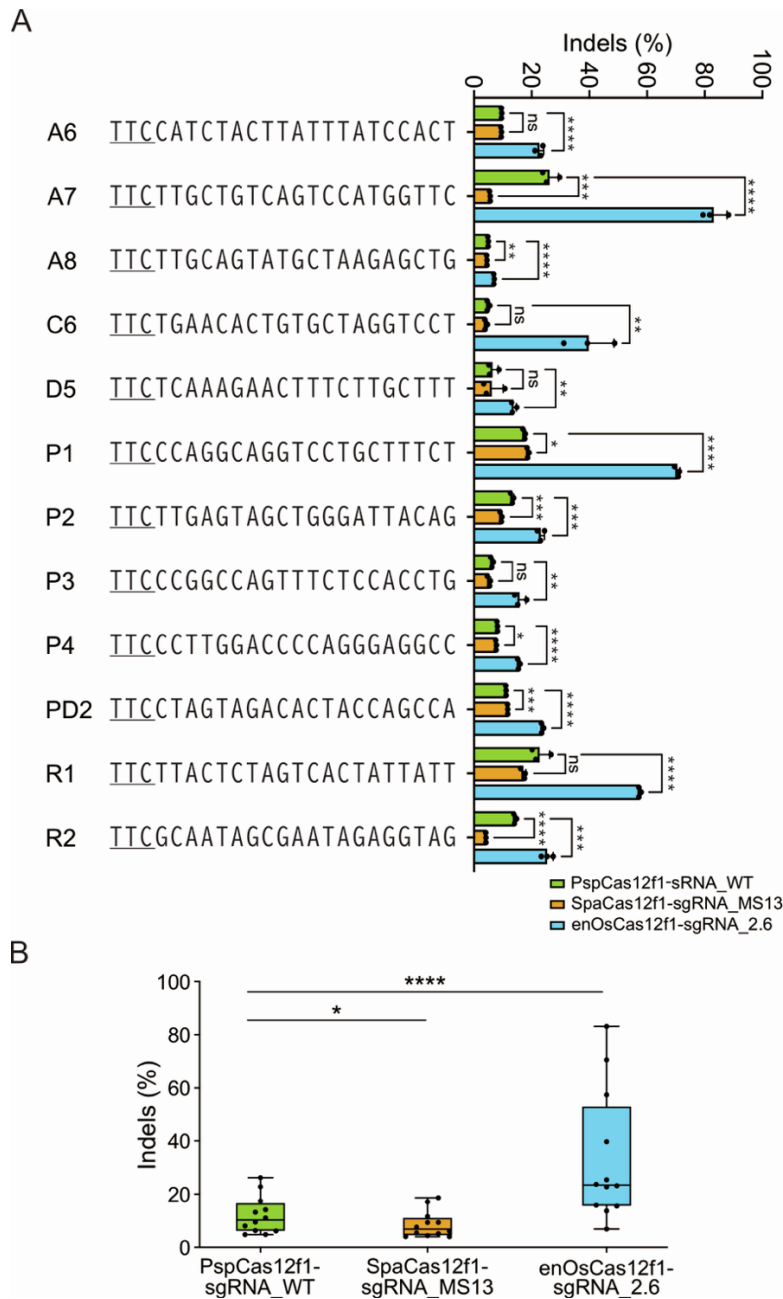


Figure S10. Comparison of genome-editing efficiencies of PspCas12f1, SpaCas12f1, and enOsCas12f1. (A) Editing efficiencies of PspCas12f1, SpaCas12f1, and enOsCas12f1 at 12 endogenous loci in HEK293T cells. Indel frequencies were quantified by targeted deep sequencing. Data are presented as mean $\pm$ standard deviation (s.d.) from three independent experiments ($n = 3$). (B) Boxplots of indel frequencies generated by PspCas12f1, SpaCas12f1, and enOsCas12f1. Boxes represent the interquartile range, the center line indicates the median, and whiskers denote the minimum and maximum values. Each data point represents the mean indel frequency at one endogenous target site from three independent biological replicates. Statistical significance was determined with a threshold of $P < 0.05$ (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$).

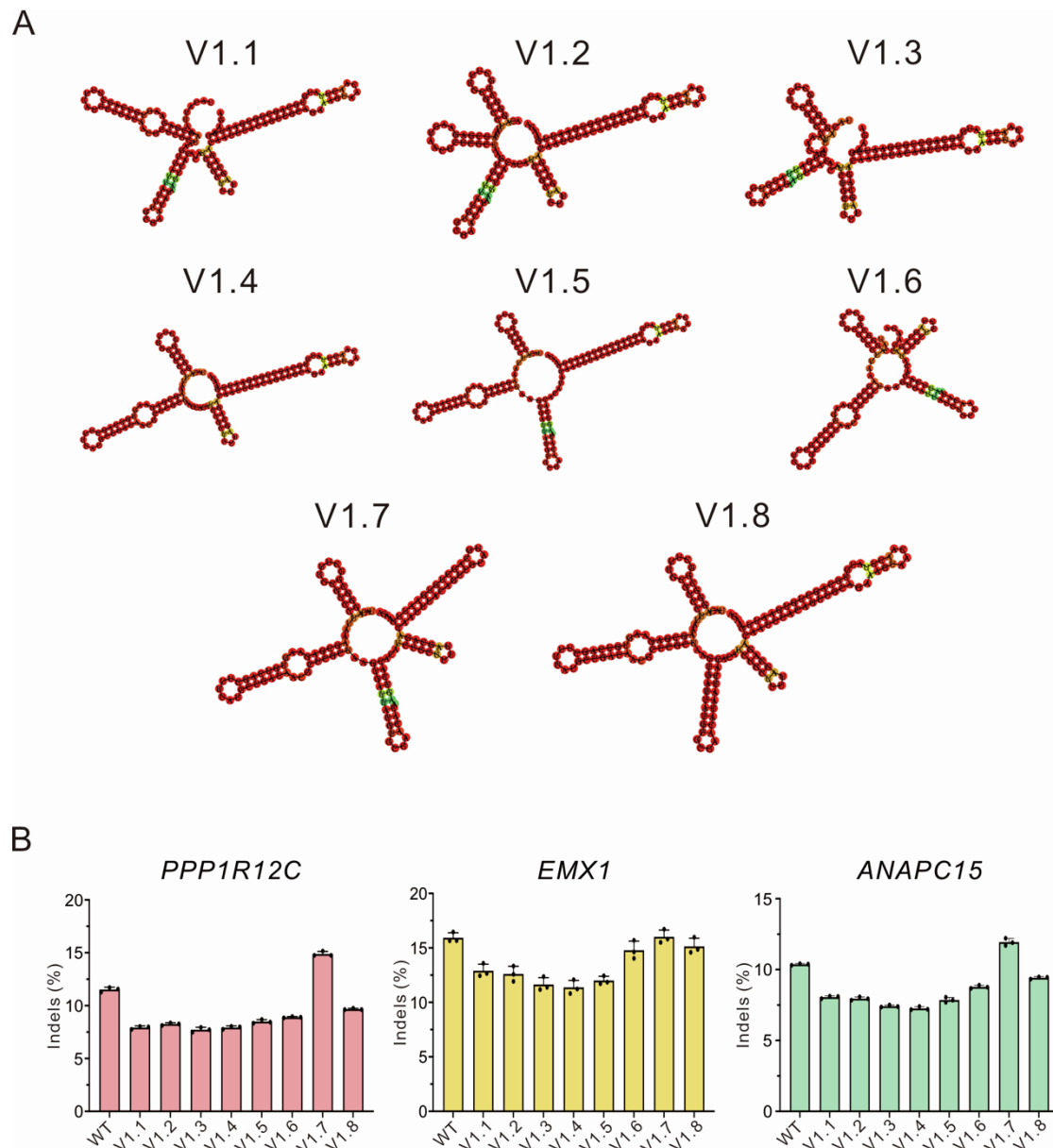


Figure S11. Comparison of genome-editing efficiencies of eight truncated sgRNA variants. (A) Predicted secondary structures of the eight truncated sgRNA variants. All secondary-structure predictions were generated using the RNAfold Web Server (<http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>). (B) Editing efficiencies of the wild-type sgRNA and the eight truncated sgRNA variants at three endogenous loci in HEK293T cells. Editing efficiencies were quantified as indel frequencies by targeted deep sequencing. Data are presented as mean $\pm$ standard deviation (s.d.) from three independent experiments ($n = 3$).

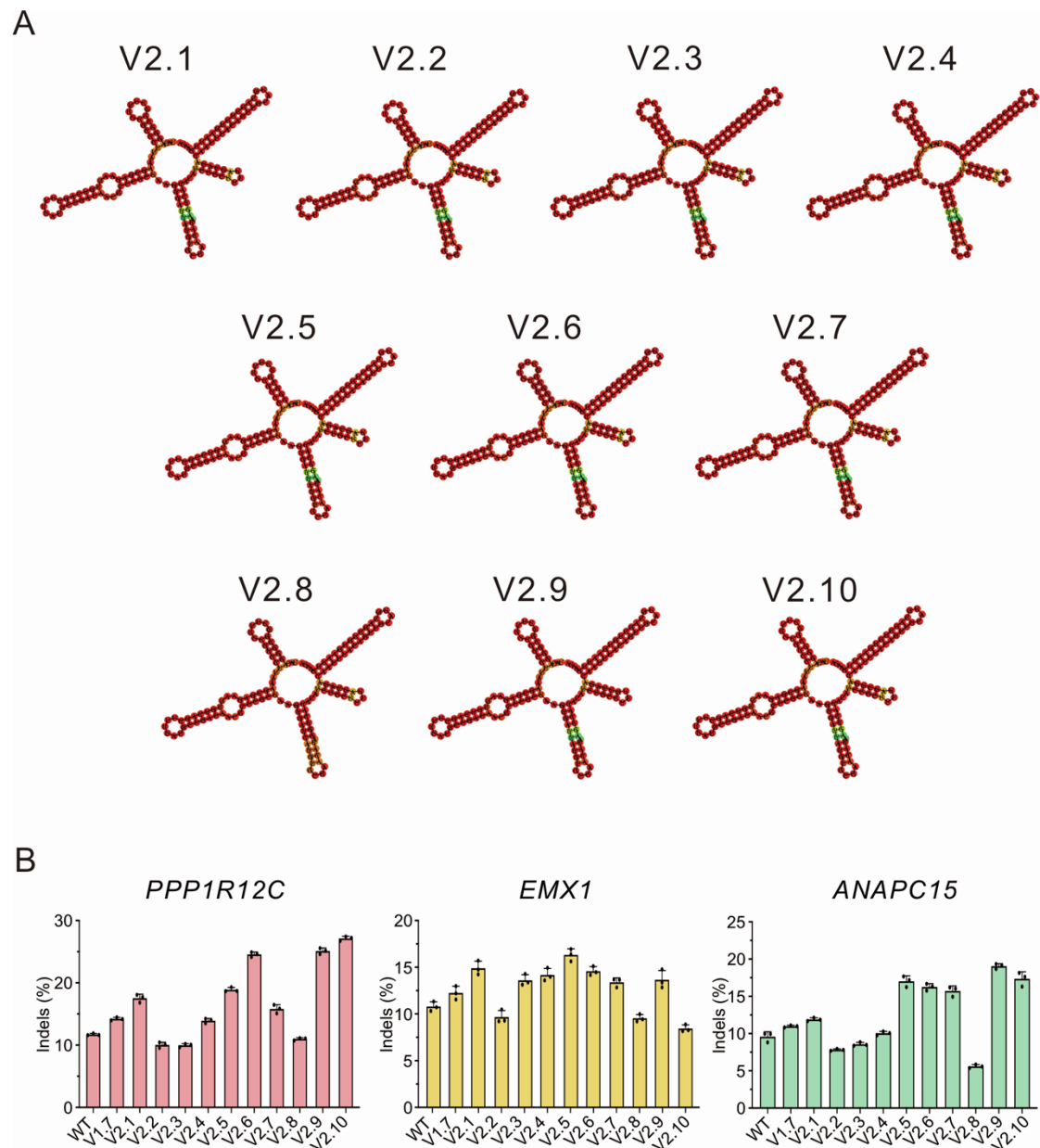


Figure S12. Comparison of genome-editing efficiencies of ten sgRNA variants containing nucleotide substitutions. (A) Predicted secondary structures of ten sgRNA variants carrying nucleotide substitutions. All secondary-structure predictions were generated using the RNAfold Web Server (<http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>). (B) Editing efficiencies of the wild-type sgRNA, sgRNA_V1.7, and the ten-nucleotide-substitution variants at three endogenous loci in HEK293T cells. Editing efficiencies were quantified as indel frequencies by targeted deep sequencing. Data are presented as mean $\pm$ standard deviation (s.d.) from three independent experiments ($n = 3$).

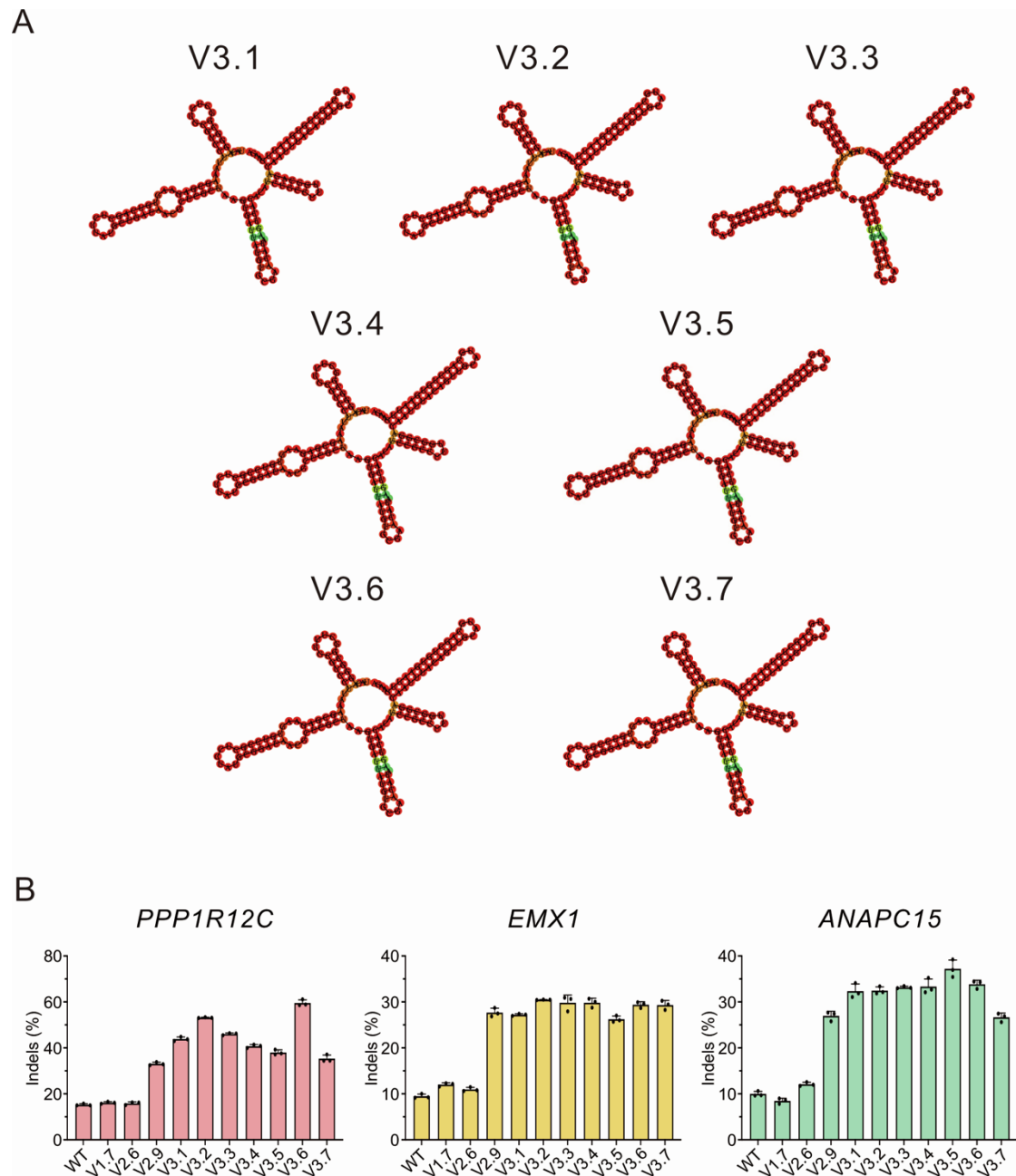


Figure S13. Comparison of genome-editing efficiencies of seven sgRNA variants containing combinatorial substitutions. (A) Predicted secondary structures of seven sgRNA variants carrying combinatorial nucleotide substitutions. All secondary-structure predictions were generated using the RNAfold Web Server (<http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>). (B) Editing efficiencies of the wild-type sgRNA, selected high-activity variants (sgRNA_V1.7, sgRNA_V2.6, and sgRNA_V2.9), and the seven combinatorial-substitution variants at three endogenous loci in HEK293T cells. Editing efficiencies were quantified as indel frequencies by targeted deep sequencing. Data are presented as mean $\pm$ standard deviation (s.d.) from three independent experiments (n = 3).

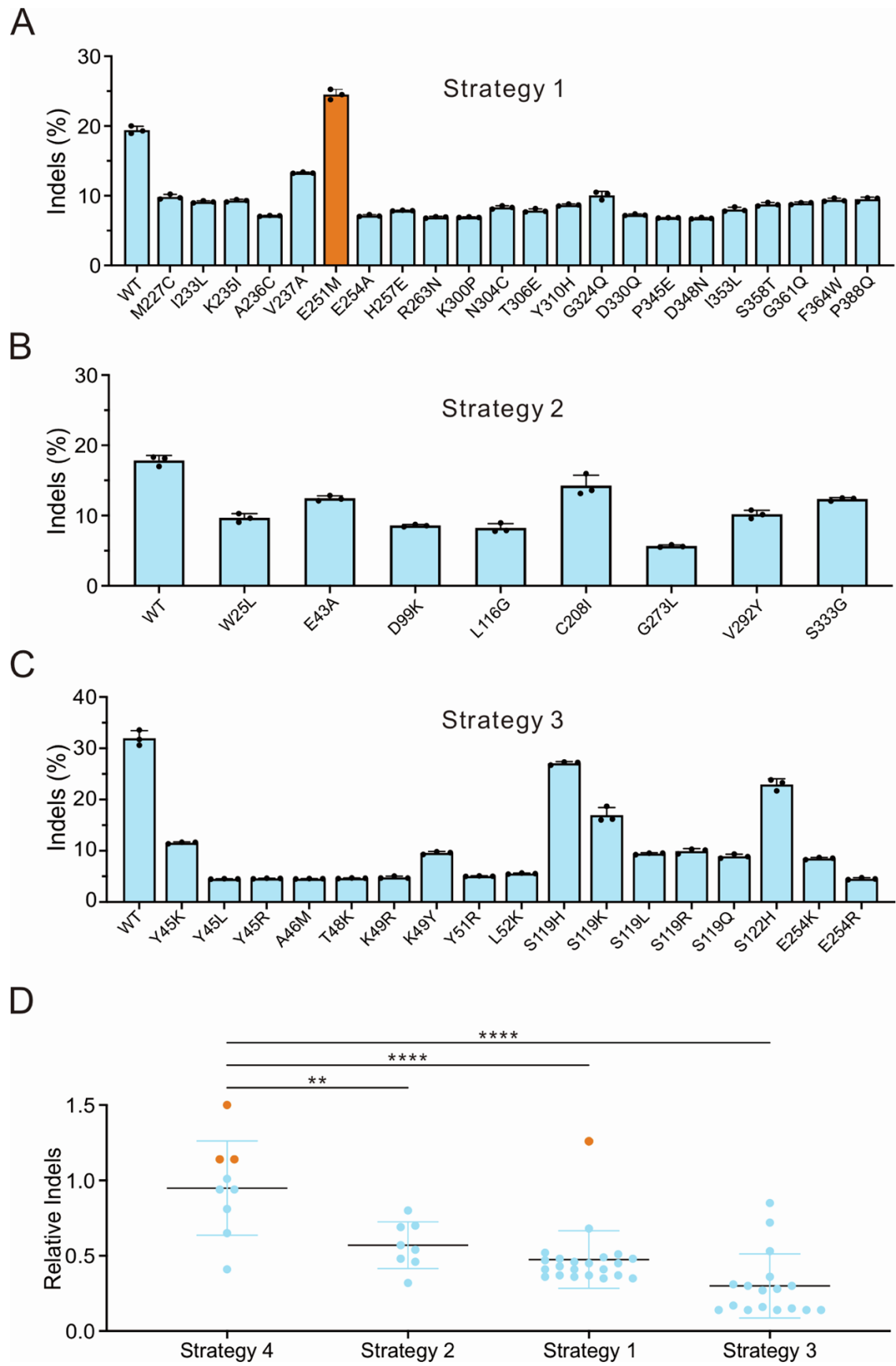


Figure S14. Comparison of editing efficiencies of PspCas12f1 single-point mutants predicted by different strategies. (A–C) Editing efficiencies of wild-type PspCas12f1 and single-point mutants predicted by Strategy 1 (A, 22 mutants), Strategy 2 (B, 8 mutants) and Strategy 3 (C, 17 mutants) at a single endogenous locus in HEK293T cells. Indel frequencies were quantified by targeted deep sequencing. Blue bars denote the wild type and individual single-point mutants, whereas orange bars indicate mutants exhibiting enhanced editing activity. (D) Quantitative comparison of editing efficiencies for all single-point mutants predicted across Strategies 1-4, normalized to wild-type PspCas12f1. Data are presented as mean $\pm$ standard deviation (s.d.) from three independent experiments ($n = 3$). Statistical significance was determined with a threshold of $P < 0.05$ (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$).

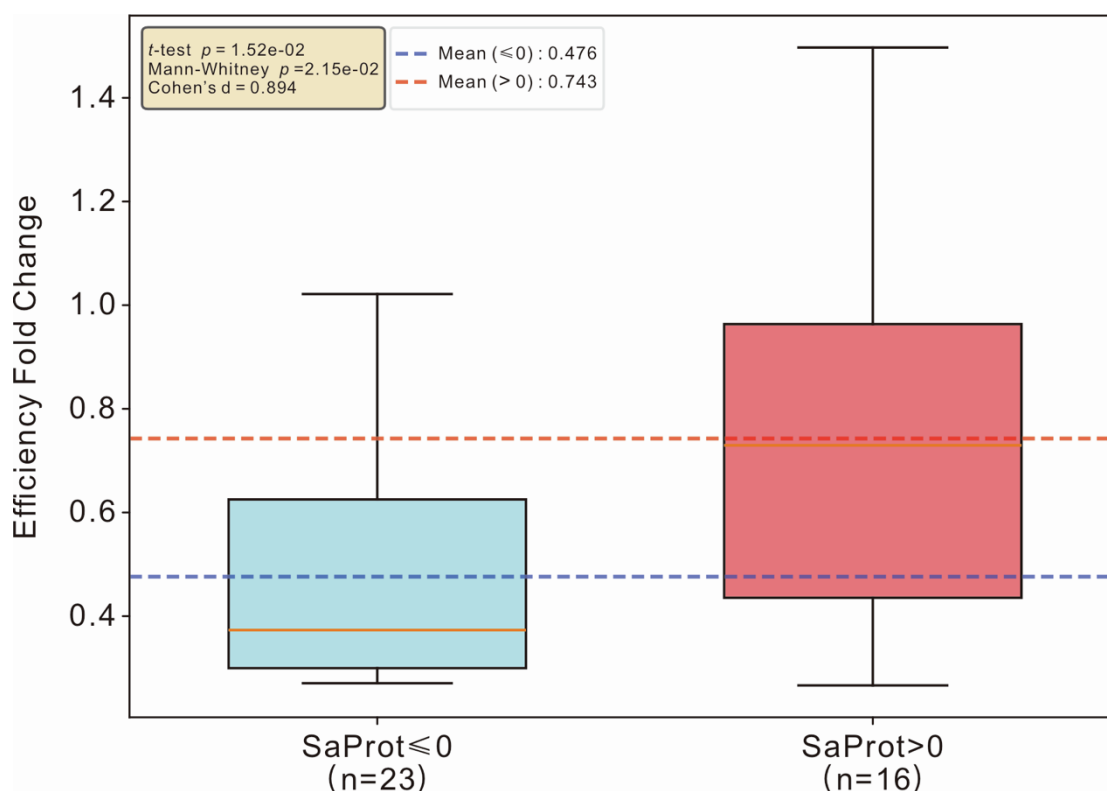


Figure S15. Distribution of efficiency fold changes stratified by a SaProt score threshold of zero. Boxplots showing the distribution of efficiency fold changes for mutants with SaProt scores ≤ 0 and > 0 . Sample sizes (n), group means, and P values from Student's t -test and Mann–Whitney test are indicated. Cohen's d is shown as a measure of effect size.

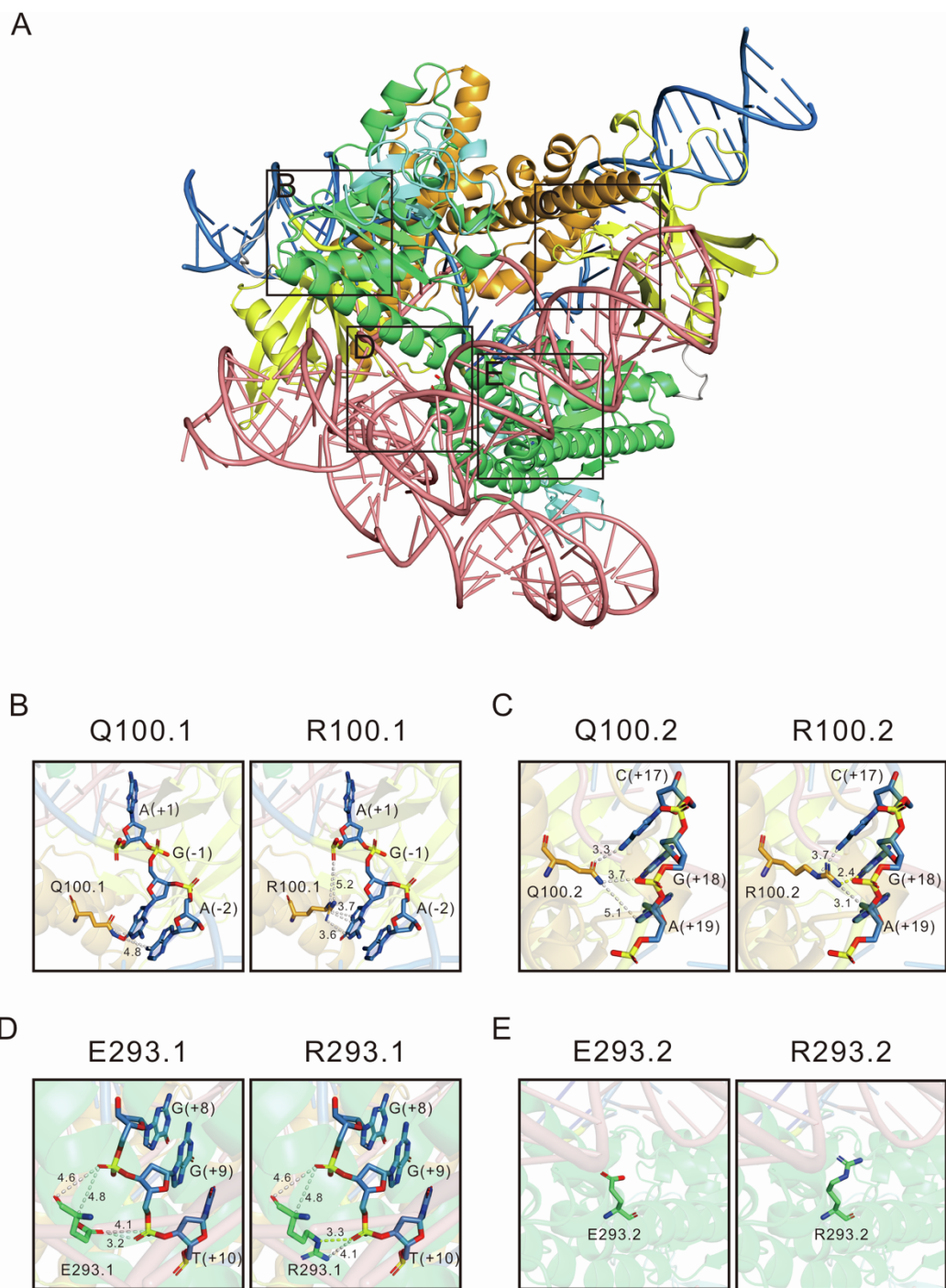


Figure S16. Predicted structural basis for the enhanced activity of the Q100R and E293R variants. (A) Overall predicted structure of the PspCas12f1–sgRNA–DNA complex. (B, C) Predicted structural model and close-up view of local residue–DNA interactions for the Q100R variant. (D, E) Predicted structural model and close-up view of local residue–DNA interactions for the E293R variant. Protein structure models were generated using the AlphaFold3 server. Gray dashed lines indicate predicted residue–DNA distances, whereas yellow dashed lines indicate predicted hydrogen bonds. Numbers indicate the corresponding predicted distances in Å.

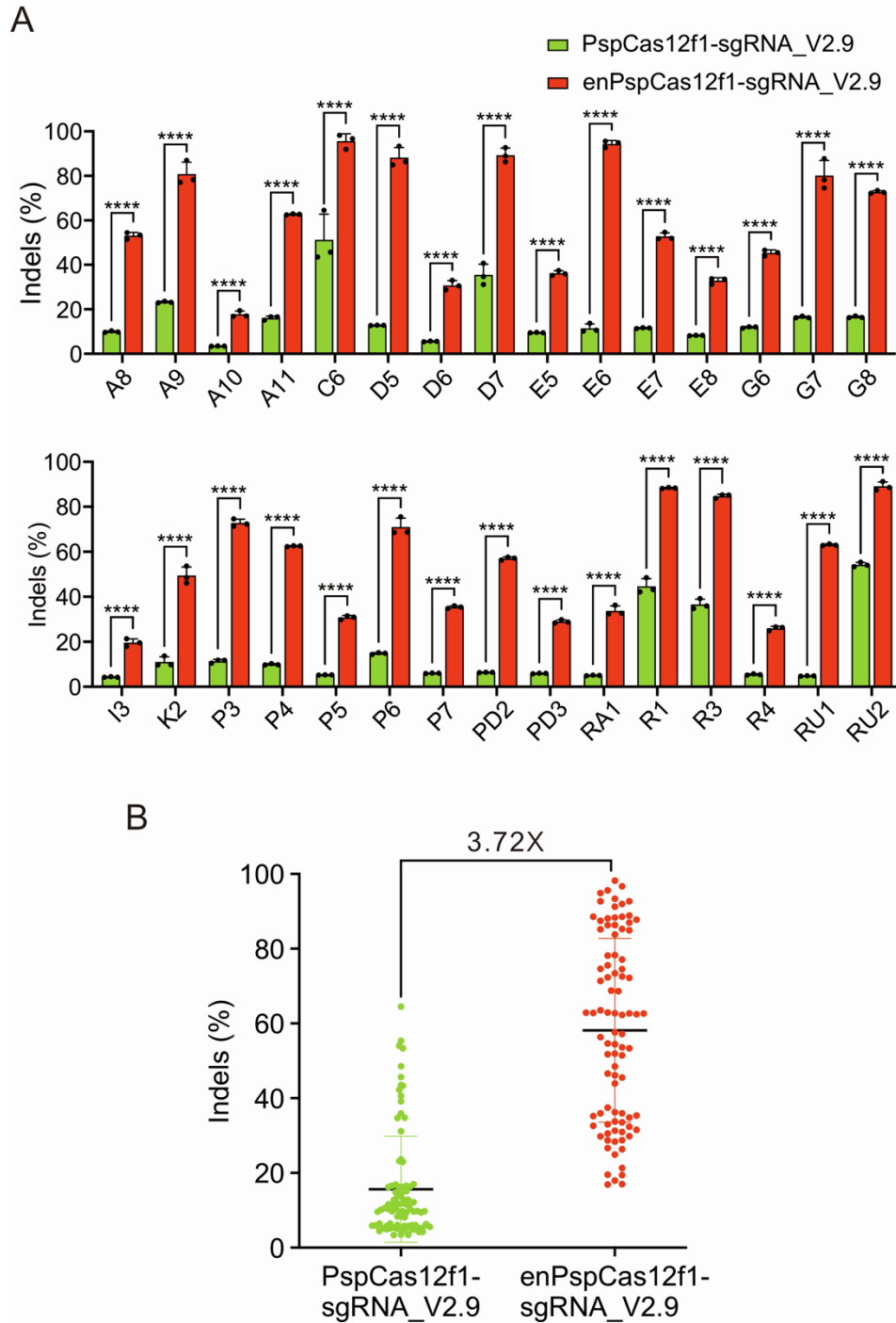


Figure S17. Comparative analysis of the genome-editing efficiency of PspCas12f1 and enPspCas12f1. (A) Editing efficiencies of PspCas12f1 and enPspCas12f1 across 30 endogenous loci in HEK293T cells. Indel frequencies were quantified by targeted deep sequencing. (B) Quantitative summary and statistical analysis of the editing efficiencies shown in panel (A). Data are presented as mean $\pm$ standard deviation (s.d.) from three independent experiments ($n = 3$). Statistical significance was determined with a threshold of $P < 0.05$ (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$).

ANAPC15-2

		GUIDE-seq reads	
		PspCas12f1	enPspCas12f1
1	T T C T C C G C G A T G C T T T G G A G C T G	284	6,849
10		0	3
20	 T . A A C G . A		

DYRK1A

		GUIDE-seq reads	
		PspCas12f1	enPspCas12f1
1	T T C T T C C C A T G T T C T C T A C T C	4,239	9,529
10		0	23
20	 A A A T . T A . .		

EMX1

		GUIDE-seq reads	
		PspCas12f1	enPspCas12f1
1	T T C T G C C C T T G T G G T C C A G G C A A	150	504
10			
20			

RNF2-1

		GUIDE-seq reads	
		PspCas12f1	enPspCas12f1
1	T T C G C A A T A G C G A A T A G A G G T A G	9,052	170,560
10			
20			

RNF2-2

		GUIDE-seq reads	
		PspCas12f1	enPspCas12f1
1	T T C A T G A G T G A A G C T G T G A A A A G	1,638	120,347
10		188	0
20		0	6

Figure S18. Genome-wide off-target analysis of PspCas12f1 and enPspCas12f1 using GUIDE-seq. Target sequences for five additional endogenous sites are shown on the left, and corresponding read counts for on-target and off-target sites are shown on the right.

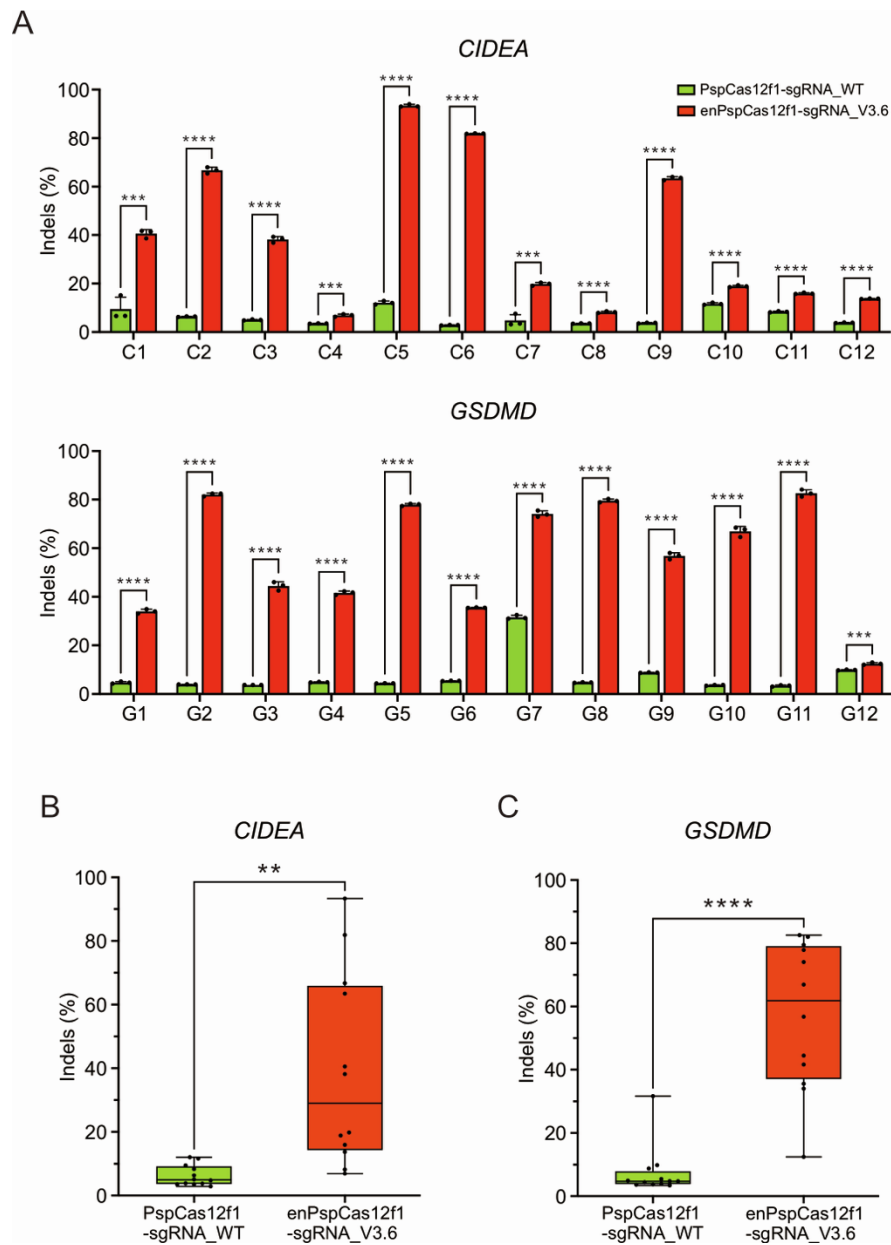


Figure S19. Enhanced genome-editing activity of enPspCas12f1 at hearing loss-associated therapeutic target genes. (A) Comparison of genome-editing efficiencies of PspCas12f1-sgRNA_WT and enPspCas12f1-sgRNA_V3.6 at 12 endogenous target sites within each of the *CIDEA* and *GSDMD* genes in HEK293T cells. Indel frequencies were quantified by targeted deep sequencing. (B, C) Distribution of genome-editing efficiencies across the 12 endogenous target sites in *CIDEA* (B) and *GSDMD* (C) for PspCas12f1-sgRNA_WT and enPspCas12f1-sgRNA_V3.6. Boxes represent the interquartile range, the center line indicates the median, and whiskers denote the minimum and maximum values. Each data point represents the mean indel frequency at one endogenous target site from three independent biological replicates. Statistical significance was determined with a threshold of $P < 0.05$ (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$).

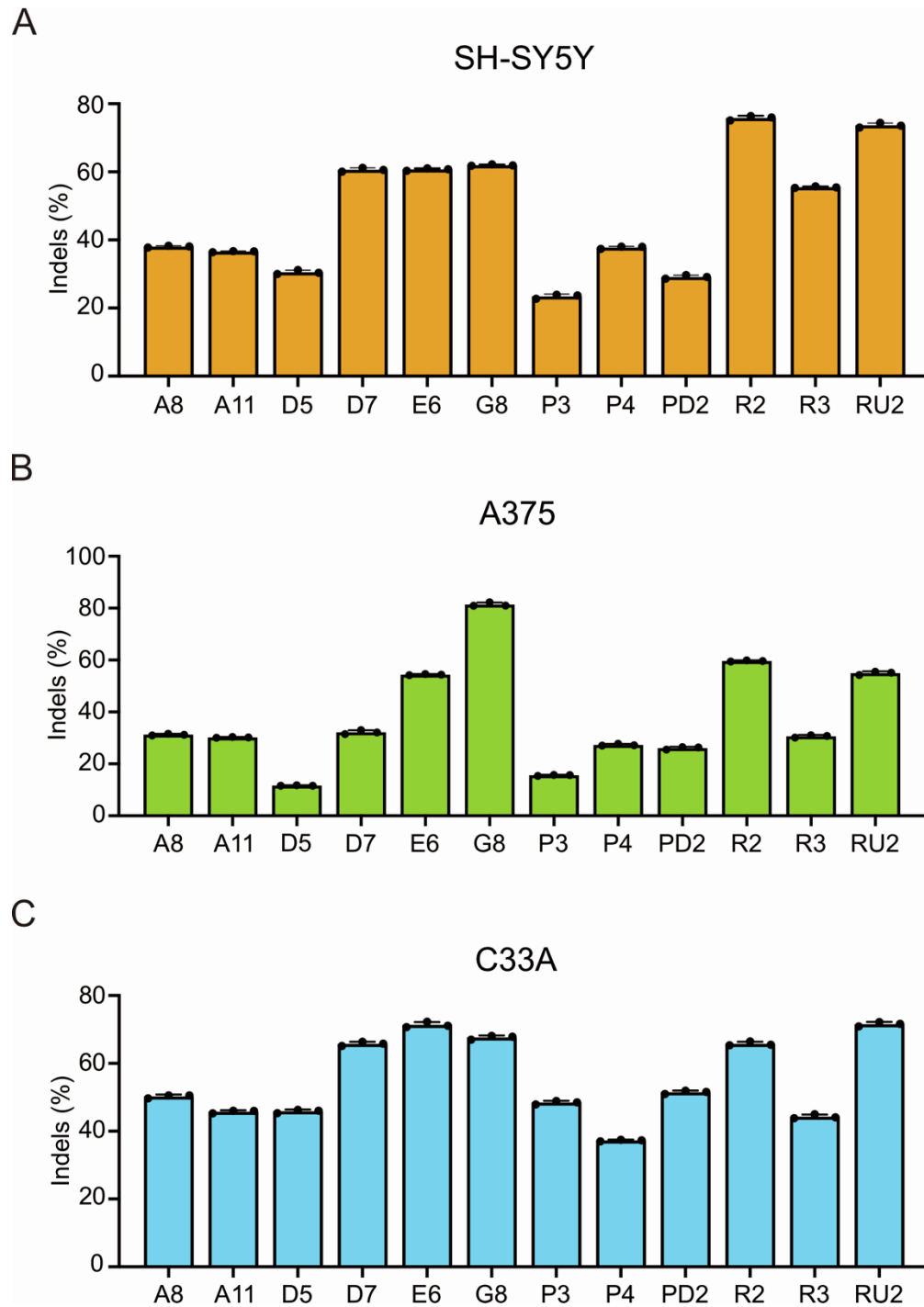


Figure S20. Genome-editing activities of AAV-delivered enPspCas12f1 in different cell lines. (A–C) Genome-editing efficiencies of an all-in-one AAV vector expressing enPspCas12f1-sgRNA_V3.6 at 12 endogenous genomic target sites in SH-SY5Y (A), A375 (B), and C33A (C) cells. Indel frequencies were quantified by targeted deep sequencing. Data are presented as mean $\pm$ standard deviation (s.d.) from three independent experiments ($n = 3$).

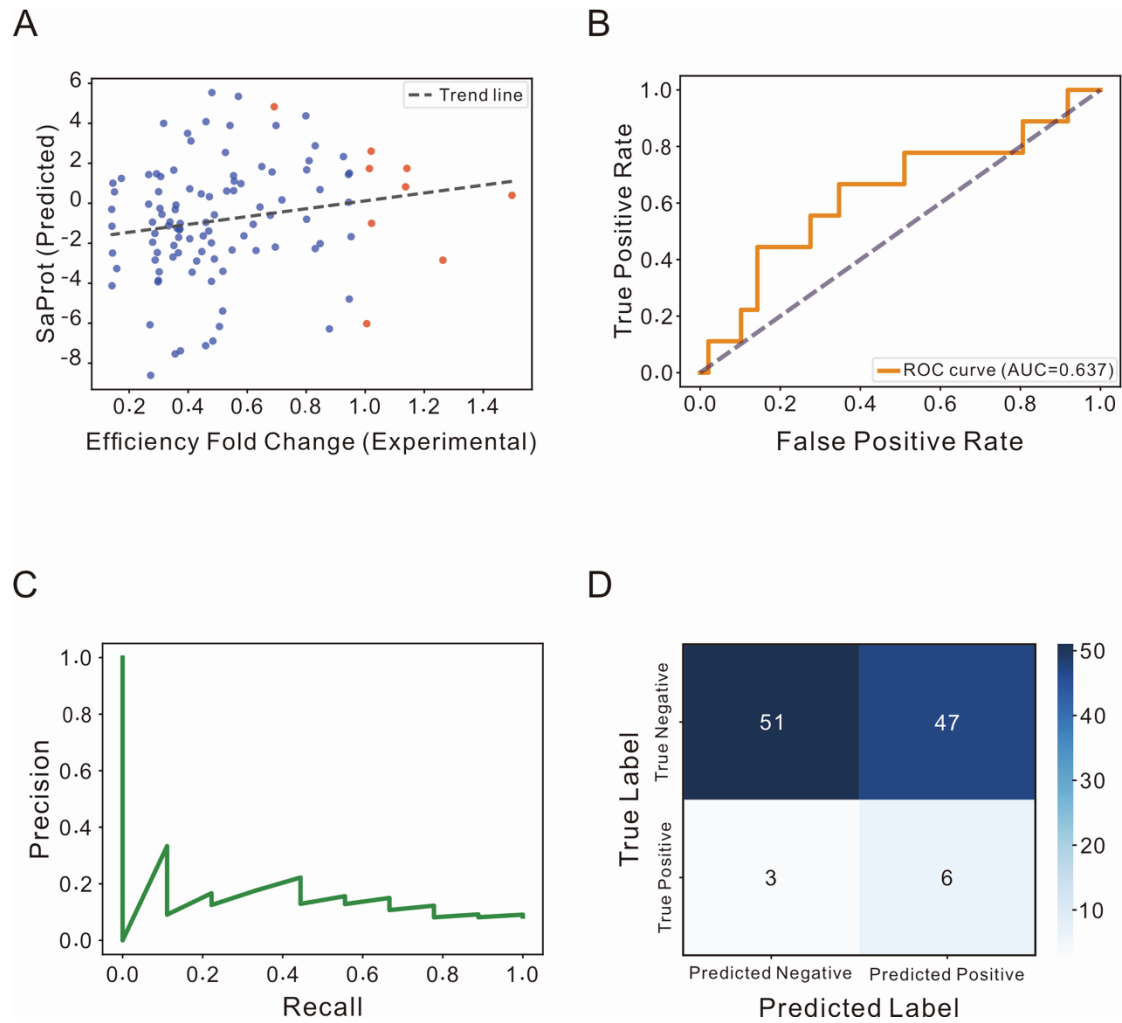


Figure S21. Evaluation of the relationship between SaProt protein-language-model predictions and experimentally measured Efficiency Fold Change. (A) Scatter plot comparing SaProt-predicted scores with experimentally determined Efficiency Fold Change. Correlation strength was assessed using Spearman's rank correlation ($\rho = 0.245$). (B) Receiver operating characteristic (ROC) curve with an area under the curve (AUC) of 0.637. (C) Precision–Recall curve. (D) Confusion matrix generated by dichotomizing variants using the median Efficiency Fold Change.

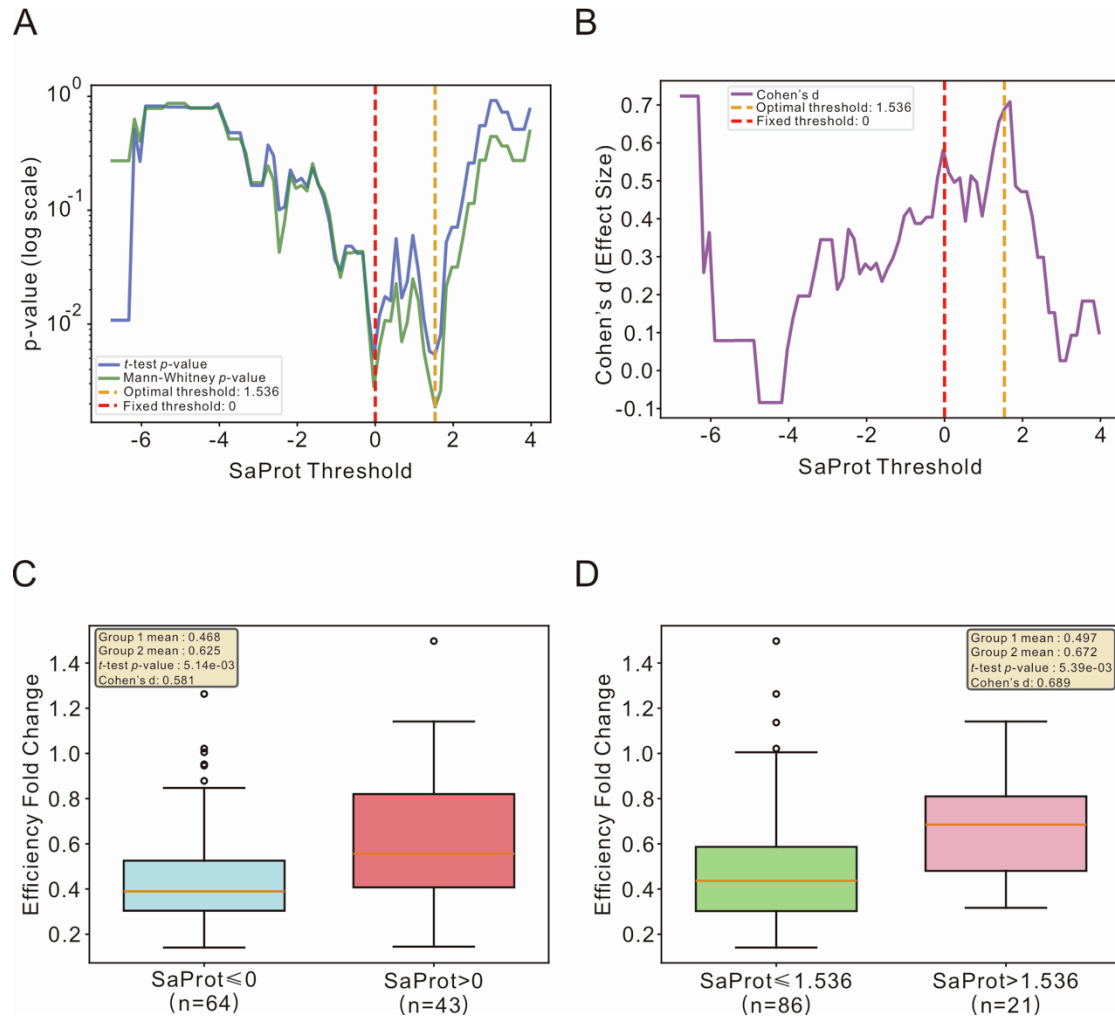


Figure S22. Comparison of PspCas12f1 mutant-screening performance across different SaProt prediction thresholds. (A) Statistical significance (*p*-values) for between-group comparisons across a range of prediction thresholds. The dashed line indicates the significance cutoff ($p = 0.05$). Thresholds of 0 and 1.536 are highlighted. (B) Between-group effect sizes (Cohen's *d*) across prediction thresholds, with values at thresholds 0 and 1.536 marked for comparison. (C) Boxplot showing the distribution of Efficiency Fold Change when mutants are grouped using a threshold of 0 (default grouping). Sample sizes (*n*), group means, and *p* values from statistical tests are indicated. (D) Boxplot showing the distribution of Efficiency Fold Change when mutants are grouped using a threshold of 1.536. Sample sizes (*n*), group means, and *p* values from statistical tests are indicated.

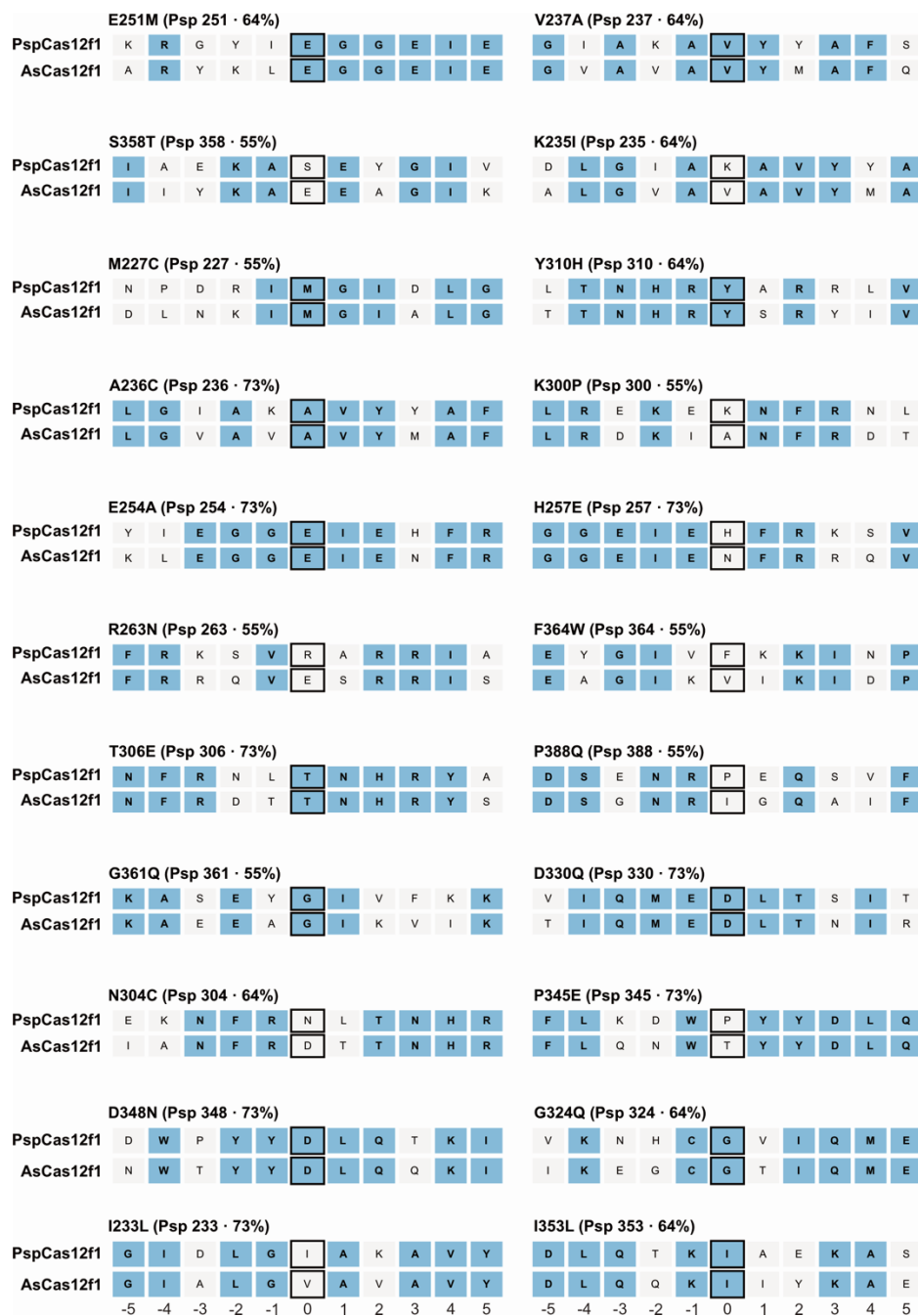


Figure S23. Local sequence alignments of candidate mutation sites between PspCas12f1 and AsCas12f1 for Strategy 1. Local sequence alignments of PspCas12f1 and AsCas12f1 surrounding each candidate residue selected for homology-based mutagenesis. Each panel displays an 11-amino-acid sliding window centered on the candidate residue (± 5 residues), the corresponding residue positions in the two proteins, and the sequence similarity within the window. Black boxes indicate the residues selected for mutagenesis. Dark and light shading indicate identical and non-identical amino acid residues, respectively.

Supplemental Tables

Table S1. The sequences of the Cas12f orthologs and sgRNA scaffolds

Table S2. Primers used in this study

Table S3. The sequences of sgRNA variants

Table S4. Target sites used in this study

Table S5. Single-point mutations predicted by Strategy 1

Table S6. Single-point mutations predicted by Strategy 2

Table S7. Candidate mutations at residues within 6 Å of DNA with corresponding SaProt scores and efficiency fold changes

Site	Distance_to_DN Å	Mutant	SaProt_score	Group	Efficiency Fold Change	Efficiency Fold Change_bin	$\Delta\Delta G$
84	4.3	T84R	3.116438866	Saprot>0	0.40931617	0	-1.53523
100	1.3	Q100R	0.822889686	Saprot>0	1.13677876	1	-0.885744
81	4.7	S81K	1.744920731	Saprot>0	1.14115416	1	-0.379745
342	1.4	K342R	0.051902309	Saprot>0	0.94418469	0	-0.148718
152	5.8	N152K	2.127695322	Saprot>0	0.80983626	0	0.0102383
204	5.8	Q204K	1.737418056	Saprot>0	1.0141937	1	0.0704506
165	3.4	T165R	1.837360144	Saprot>0	0.64946323	0	0.359428
290	3.2	Q290R	1.439507484	Saprot>0	0.94309789	0	0.494522
293	3.2	E293R	0.396849573	Saprot>0	1.49713423	1	0.537373
164	2	T164H	0.468631804	Saprot>0	0.44417511	0	1.50858
333	4.1	S333H	1.505684733	Saprot>0	0.94657662	0	1.56078
80	3	Y80K	1.435004592	Saprot>0	0.2663605	0	1.64646
78	3.7	D78R	0.630534291	Saprot>0	0.554069	0	1.77896
78	3.7	D78K	2.542242765	Saprot>0	0.52628629	0	1.96074
80	3	Y80R	1.33555162	Saprot>0	0.30688767	0	2.19598
154	2.6	N154K	1.475261092	Saprot>0	0.29307832	0	3.25806
74	5	K74R	-2.267720461	Saprot<0	0.83055919	0	
84	4.3	T84Y	-1.281552672	Saprot<0	0.37227576	0	
96	2.8	N96Y	-0.25500688	Saprot<0	0.30437361	0	
97	3.2	Q97R	-1.004883766	Saprot<0	0.37310642	0	
100	1.3	Q100L	-0.578225017	Saprot<0	0.48872006	0	
100	1.3	Q100Y	-0.600442588	Saprot<0	0.67857672	0	
100	1.3	Q100H	-0.9549914	Saprot<0	0.45989182	0	
108	1.1	N108R	-0.18908295	Saprot<0	0.64335141	0	
133	2.2	N133K	-1.052599669	Saprot<0	0.62050849	0	
133	2.2	N133R	-1.981839538	Saprot<0	0.47875468	0	

154	2.6	N154H	-1.507827401	Saprot<0	0.28731759	0
157	3.4	S157N	-1.129147649	Saprot<0	0.33309074	0
157	3.4	S157H	-3.930332661	Saprot<0	0.29766023	0
157	3.4	S157L	-6.078481197	Saprot<0	0.27061904	0
165	3.4	T165M	-2.338695288	Saprot<0	0.54887565	0
191	1.8	P191R	-2.362500668	Saprot<0	0.62988592	0
192	2.3	K192H	-0.800340772	Saprot<0	0.80133919	0
192	2.3	K192R	-1.001986504	Saprot<0	1.02139669	1
193	1.7	G193R	-2.840811491	Saprot<0	0.28819899	0
193	1.7	G193K	-3.428968906	Saprot<0	0.30166449	0
194	3.9	I194R	-8.607885361	Saprot<0	0.27286329	0
195	3.4	Q195R	-1.332035422	Saprot<0	0.36955302	0
211	3.9	P211R	-1.950201511	Saprot<0	0.27959604	0

Mutations with SaProt > 0 and $\Delta\Delta G < 1.5$ kcal/mol are highlighted in red;
mutations with SaProt > 0 and $\Delta\Delta G > 1.5$ kcal/mol are highlighted in blue;
all remaining mutations with SaProt ≤ 0 are shown in black.

Table S8. Genome-editing efficiency fold changes of PspCas12f1 mutants relative to wild-type

Table S9. Raw data in this study