nature communications

Article

https://doi.org/10.1038/s41467-025-57150-9

Highly parallel profiling of the activities and specificities of Cas12a variants in human cells

Received: 3 April 2024

Accepted: 11 February 2025

Published online: 28 March 2025

Check for updates

Peng Chen[1,2,5], Yankang Wu[2,5], Hongjian Wang[2,5], Huan Liu[2,5], Jin Zhou[1,2,3], Jingli Chen[4], Jun Lei[1,2], Zaiqiao Sun[2], Chonil Paek[2] & Lei Yin[1,2]

Several Cas12a variants have been developed to broaden its targeting range, improve the gene editing specificity or the efficiency. However, selecting the appropriate Cas12a among the many orthologs for a given target sequence remains difficult. Here, we perform high-throughput analyses to evaluate the activity and compatibility with specific PAMs of 24 Cas12a variants and develop deep learning models for these Cas12a variants to predict gene editing activities at target sequences of interest. Furthermore, we reveal and enhance the truncation in the integrated tag sequence that may hinder off-targeting detection for Cas12a by GUIDE-seq. This enhanced system, which we term enGUIDE-seq, is used to evaluate and compare the off-targeting and translocations of these Cas12a variants.

The CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR-associated protein) system, an adaptive immune system prevalent in prokaryotes[1–6], has recently been exploited for a wide range of applications in eukaryotic genome manipulation[7–11]. Besides the widely used SpCas9, many other programmable endonucleases have been repurposed for genome manipulation[12–23]. Several distinct features distinguish Cas12a from other proteins. These features, such as (i) recognition of T-rich protospacer adjacent motif (PAM) sequences[24–28]; (ii) utilization of a single short ~40 nt CRISPR RNA[28–30]; (iii) production of double-strand breaks distal to the PAM[28]; (iv) processing of multiple functional CRISPR RNAs (crRNAs) from a single long transcript[29,30]; (v) non-specific single-stranded DNAse activity (cis-activity) following crRNA-guided target DNA binding and cleavage (trans-activity)[31,32], have accelerated the utilization of Cas12a for multi-gene editing, gene regulation and pathogen detection[29,33–36]. Alternatively, the intrinsic high fidelity of Cas12a may facilitate its use in clinical translation. For example, Cas12a was successfully used to generate patient-derived corrected iPSC clones for type 1 ocular albinism, restoring normal splicing and increasing GPR143 expression to normal levels[37]. In addition, Cas12a edited patient HSPCs showed reversal of aberrant splicing and restoration of b-globin expression[38]. Collectively, these results suggest that Cas12a-based therapeutics have the potential to offer improvements in a wide range of genetic diseases.

Although several groups have identified a number of Cas12a orthologs, such as AsCas12a[28], LbCas12a[28], FnCas12a[28,39], Lb2Cas12a[28,40–42] and EbCas12a[43], and verified their gene editing function in cells, limited genomic targeting coverage and relatively low editing efficiency have precluded the practical utility of Cas12a. To address the above limitations, several Cas12a variants with extended targeting capabilities or increased editing efficiency, such as enAsCas12a-HF1[44], iCas12as[45], AsCas12a Ultra[46], HyperLbCas12a[47] have been generated. Additionally, high fidelity orthologs or variants such as CeCas12a[48], AsCas12a-Plus[49], LbCas12aK538R[42], HyperFi-AsCas12a[50] have also been identified or engineered. However, selecting the appropriate Cas12a among the many orthologs for target-specific gene manipulation is time-consuming and laborious.

Here, we use high-throughput analysis to evaluate the activity as well as the compatibility with dozens of popular PAMs of 24 Cas12as: AsCas12a, LbCas12a, FnCas12a, Lb2Cas12a, CeCas12a, EbCas12a, enAsCas12a-HF1, AsCas12a Ultra, HyperLbCas12a, AsCas12aRR[51], AsCas12aRVR[51], AsCas12a-Plus, HyperFi-AsCas12a,

[1]Department of Pediatric Research Institute; Ministry of Education Key Laboratory of Child Development and Disorders, National Clinical Research Center for Child Health and Disorders, China International Science and Technology Cooperation Base of Child Development and Critical Disorders, Children's Hospital of Chongqing Medical University, School of Basic Medical Sciences, Chongqing Medical University, Chongqing, China. [2]State Key Laboratory of Virology, Hubei Key Laboratory of Cell Homeostasis, College of Life Sciences, Wuhan University, Wuhan, China. [3]Wuhan Biorun Biosciences Co., Ltd., Wuhan, China. [4]School of Medicine, Wuhan University of Science and Technology, Wuhan, China. [5]These authors contributed equally: Peng Chen, Yankang Wu, Hongjian Wang, Huan Liu. e-mail: yinleiwh@163.com

https://doi.org/10.1038/s41467-025-57150-9

a

Name	Engineered domain(s)	Mutations	PAM	Biological significance
AsCas12a			TTTV	
enAsCas12a-HF1	REC1, REC2, PI	E174R/S542R/K548R/N282A	TTYN, VTTV, TRTV	Expanding targeting range, increasing genome editing activity
AsCas12aRR	WED-II, PI	S542R/K607R	TYCV	Expanding targeting range
AsCas12aRVR	WED-II	S542R/K548V/N552R	TYCV, TWTV	Expanding targeting range
AsCas12a Ultra	WED-II, WED-III	M537R/F870L	TTTN	Increasing genome editing activity
AsCas12a-Plus	REC1, Bridge helix	E174R/R951K/R955A	TTTV	Increasing genome editing specificity
HyperFi-AsCas12a	REC1, REC2, RuvC-II	S186A/R301A/T315A /Q1014A/K414A	TTTV	Increasing genome editing specificity
LbCas12a			TTTV	
LbCas12aRR	WED-II, PI	K538V/K595R	TTTV, TYCC	Expanding targeting range
LbCas12aRVR	WED-II	G532R/K538R/Y542R	TTTV, TYCC,TATC	Expanding targeting range
LbCas12aRVRR	WED-II, PI	G532R/K538R/Y542R/K595R	TTTV, TYCC,TMTC	Expanding targeting range
LbCas12a-Plus	REC1, Bridge helix	D156R/R883K/R887A	TTTV	Increasing genome editing specificity
LbCas12a-K538R	WED-II	K538R	TTTV	Increasing genome editing specificity
HyperLbCas12a	REC1, REC2	D156R/D235R/D292R/D350R	TTTV	Increasing genome editing activity
iCas12a(mut2C-W)	Nuc, Ruvc, Bridge helix	F863V/F884L/D952Y/C965Y /V1011A/Q1108L/A1113V /S1132T/S1214P	TTTV	Increasing genome editing activity
iCas12a(mut2C-WF)	Nuc, Ruvc, Bridge helix	F863V/D952Y/C965Y/V1011A /Q1108L/A1113V/S1132T /S1214P	TTTV	Increasing genome editing activity
FnCas12a			YTV	
eaFnCas12a	REC1	Q125R	YTV	Increasing genome editing activity
FnCas12aRVR	WED-II	N607R/K613V/N617R	YTV, TATM	Expanding targeting range
Lb2Cas12a			BYYV	
Lb2Cas12aK518R	WED-II	K518R	TTTV	Increasing genome editing specificity
CeCas12a			TTTV	
EbCas12a			TTTV	
enEbCas12a	REC1	D141R	TTTV	Increasing genome editing activity

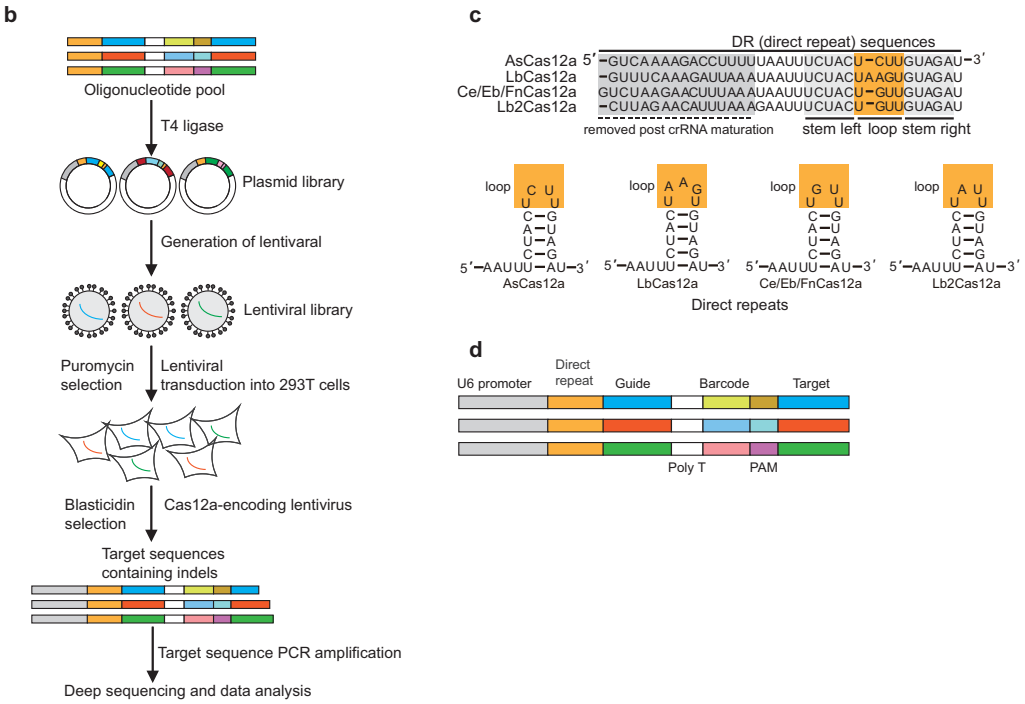


Fig. 1 | Summary of the 24 Cas12a variants and an overview of the high throughput experiment. a The engineered domains, mutation residues, PAM and biological significance of each Cas12a. N = A/G/C/T; V = A/C/G; Y = C/T; B = C/G/T; M = A/C. **b** Four crRNAs that used by Cas12a variants. The difference between crRNAs lies in the loop region. **c** Schematic representation of oligonucleotides containing pairs of target and guide RNA sequences. **d** Schematic representation of the high-throughput experiments.

LbCas12aRR[52], LbCas12aRVR[52], LbCas12aRVRR[52], LbCas12a-Plus[49], LbCas12aK538R, iCas12as (mut2C-W, mut2C-WF), eaFnCas12a[53], FnCas12aRVR[54], Lb2Cas12aK518R[42], enEbCas12a at thousands of target sequences in human cells (Fig. 1a). In addition, we present an enhanced version of the GUIDE-seq method, which improves the capture of tag sequence and compensates for the omissions in the original method due to the incomplete tag, and used it to evaluate the off-targeting and translocations of 24 Cas12a orthologs on 31 targets with a large number of homopolymeric sequences in human cells. Together, these analyses will greatly facilitate the use of these Cas12 nucleases in genome editing applications.

Results

High throughput activity assessment of 24 Cas12a orthologs

To compare the nuclease activities of 24 Cas12a variants (Fig. 1a), we used a method similar to a previously described high-throughput approach[55] (Fig. 1b–d). Four lentiviral libraries, named Library As, Lb, Ce/Fn/Eb and Lb2 were generated and transduced into the HEK239T cells at amultiplicity of infection (MOI) of 0.4. Each library pool contains 11,968 pairs of guide sequences and corresponding target sequences oligonucleotide. These 24 Cas12a variant coding sequences were then cloned into lentiviral vectors together with BSD (a gene conferring blasticidin resistance) and expressed under the control of the CMV promoter. To assess whether the multiplicity of infection (MOI) of Cas12a affects indel frequency, we used different Cas12a lentiviruses with different MOIs (0.1, 0.5, 1, 2) to transduce HEK293T cells integrated with the target sequence. At an MOI of 0.1, the indel frequency of the six Cas12as were low, averaging 10%; at an MOI of 1, the indel frequency increased to 45%. As the MOI continued to increase, the indel frequency was not significantly improved (Supplementary Fig. 1a–c). To evaluate the protein expression levels of different Cas12 variants, we performed Western blotting assay. The results showed that the expression levels were overall comparable to each other except that Lb2Cas12a and Lb2Cas12aK518R showed relatively lower expression levels (Supplementary Fig. 1d, e). In order to obtain the optimal indel frequencies of various Cas12a on target sequence libraries, we chose an MOI of 1 for the following experiments. Thus, 24 large datasets in HEK293T cells were obtained to show the activity of these Cas12as on a total of 11968 lentiviral integrating target sequences containing all 15 PAMs currently known to be recognized by Cas12a (Supplementary Fig. 2 and Supplementary Datasets 1, 2). And the correlation between indel frequencies in experimental replicates is very high (Supplementary Fig. 3). We first identified which Cas12a variant had the highest activity on the target sequences with the given PAM sequences (Fig. 2a, Supplementary Fig. 4 and Supplementary Datasets 3, 4). For the TATV PAM, AsCas12aRVR had the highest indel frequency with 53%. For the CTCV, GTTV, TCTV, ATTV, CTTV PAMs, enEbCas12a had the highest frequency of indels with 54.3%, 48.7%, 47.5%, 42.8%, and 54.3%, respectively. For the GTCV, TCCV, TTCV, TTAV, CCCV PAMs, LbCas12aRR had the highest frequency of indels with 22.7%, 59.5%, 64%, 16.6%, and 53.4%, respectively. For the TGTV, TGCV and TACV PAMs, enAsCas12a-HF1 had the highest indel frequency with 32.4%, 23.8% and 32%, respectively. Furthermore, the general activities of these Cas12a variants were compared at the target sequences containing the classical TTTV PAM. Cas12a variants can be ordered as follows: mut2C-W (74.5%), mut2C-WF (73.3%), enAsCas12a-HF1 (70.2%), enEbCas12a (67%), AsCas12a Ultra (64.4%), HyperLbCas12a (64%), LbCas12a (63%), LbCas12aRVR (62%), LbCas12aRR (60.7%), AsCas12a (60.3%), eaFnCas12a (60.2%), Lb2Cas12a (58.8%), AsCas12a Plus (57.2%), LbCas12a Plus (56%), CeCas12a (55.5%), AsCas12aRVR (55%), EbCas12a (54%), LbCas12a K538R (53.4%), Lb2Cas12a K518R (52%), LbCas12aRVRR (50.7%), FnCas12a (50.6%), HyperFi-AsCas12a (49%), FnCas12aRVR (44.7%), AsCas12aRR (31.8%) (Fig. 2a, Supplementary Fig. 4 and Supplementary Datasets 3, 4). Similarly, the activity of these Cas12a variants was validated against a number of endogenous targets (Supplementary Fig. 5).

We then investigated whether the different activities of these Cas12a variants were affected by the composition of the target sequences. When we compared the indel frequencies induced by each of the Cas12a variants at the targeted sequence, we found very high correlations between derivatives of the same Cas12a ortholog (Fig. 2b, c). The results of the analysis showed that all of the Cas12a variants had a high indel frequency in the TTTA, TTTG, TTTC PAMs (Fig. 2d). However, the correlations between different Cas12a orthologs were poor (Fig. 2b, c). These results suggest that the relative activity of Cas12a variants at specific target sequences does not necessarily correspond to the general activity ranks described above. These poor correlations suggest that the composition of target sequences associated with high nuclease activity may differ between Cas12a orthologs.

PAM compatibility analysis for Cas12a variant

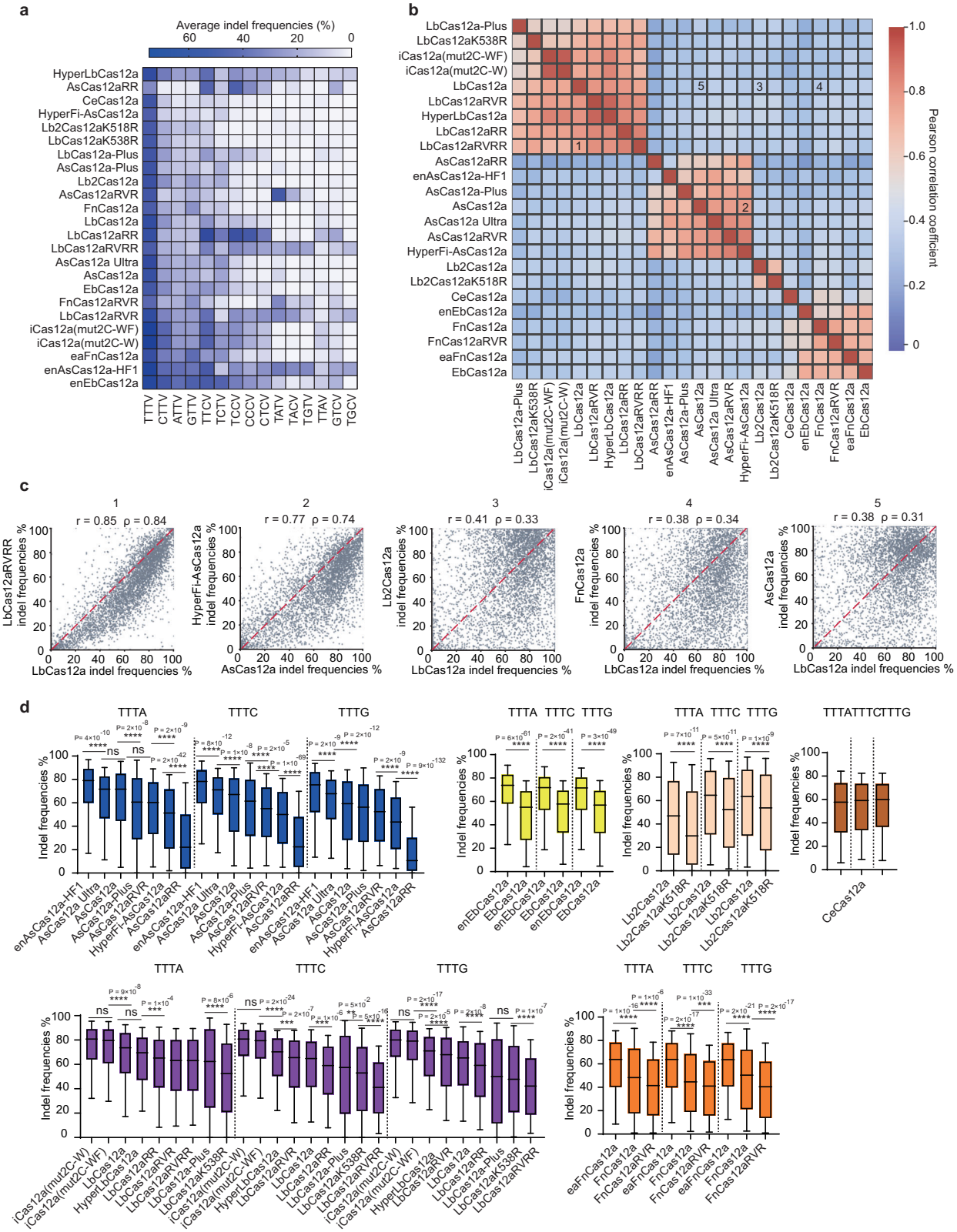
The PAM compatibility for Cas12a variants were then determined. The PAMs of the target sequences with more than 5% cleaving efficiencies were used for further analysis[56] (Fig. 2a and Supplementary Fig. 4 and Figs. 6–11 and Supplementary Datasets 3, 4). As expected, TTTV can be used as a PAM for all Cas12a variants (average indel frequencies range 31.8–74.5%) (Fig. 2a and Supplementary Fig. 12). In addition, CTTV, TTCV, GTTV, TCTV and ATTV can be used as the PAMs for AsCas12a, LbCas12a, FnCas12a, EbCas12a, Lb2Cas12a, enAsCas12a-HF1, enEbCas12a, eaFnCas12a, LbCas12aRVR, LbCas12aRVRR, LbCas12aRR, AsCas12a Ultra, FnCas12aRVR, mut2C-W, mut2C-WF, AsCas12a Plus, LbCas12a Plus (average indel frequencies range 14.2–63.9%). TCCV, CCCV, CTCV, GTCV, TTAV and TATV can also be used as PAMs for enEbCas12a and enAsCas12a-HF1 (average indel frequencies range 11.2–38.2%). TGTV, TACV, TGCV can be used as PAMs for enAsCas12a-HF1, LbCas12aRVR, LbCas12aRVRR (average indel frequencies range 15.8–32%). ATTV, TCCV, CCCV, CTCV, GTCV can be used as PAMs for LbCas12aRVR, LbCas12aRR and LbCas12aRVRR (average indel frequencies range 11.8–59.4%). TATV can be used as a PAM for LbCas12aRVR, FnCas12aRVR, LbCas12aRVRR and AsCas12aRVR (average indel frequencies range 20.2–52.9%). Furthermore, all fifteen known PAMs can be used as PAMs for enAsCas12a-HF1. Except for the TTAV PAM, the remaining fourteen PAMs can be used as PAMs for LbCas12aRVRR. In addition, AsCas12aRR had a significantly higher indel frequency at TCCV PAM (48.7%) and TTCV PAM (46.7%) than at TTTV PAM (31.8%). As the Cas12a ortholog with stringent PAM recognition, CeCas12a could hardly use any other PAM except TTTV.

Next, we tested whether the 5′ terminal nucleotides have an effect on the cleavage of Cas12a towards the target sequences that containing the 15 PAMs (Fig. 3 and Supplementary Figs. 13–16). We observed that AsCas12a variants exhibit a significantly higher indel frequency in target sequences with the 5′ end of the PAM being A, G, or T (DCTTV, DATTV, DGTTV, DTTCV, DTCTV, DTCCV, DCCCV, DCTCV, DTATV, DTACV, DTGTV, DTTAV, DGTCV, DTGCV, D = A/G/T, V = A/G/C) compared to those with the 5′ end of the PAM being C (Fig. 3a and Supplementary Fig. 16). LbCas12a, LbCas12aRR, LbCas12aRVR, mut2C-W and mut2C-WF had a higher indel frequency when the nucleotide at the 5′ end of CTTV, CCCV and TTAV was A/G/C (VCTTV, VCCCV, VTTAV) (Fig. 3b and Supplementary Fig. 16). In addition, LbCas12a, LbCas12aRR, LbCas12aRVR, mut2C-W, and mut2C-WF show higher activity in target sequences where the 5′ end of TACV, TATV, TCCV, TGCV, and TTCV is C/G/T (BTACV, BTATV, BTCCV, BTGCV, DCCCV, B = C/G/T) compared to those where the 5′ end of the PAM is A (Fig. 3b and Supplementary Fig. 16).

Activity prediction of Cas12a variants

Selecting the most appropriate candidate for a given target from the large number of Cas12a proteins is difficult since there is one DeepCpf1 algorithm for the activity prediction of AsCas12a crRNAs[57] and differences are big between Cas12a orthologs and variants. We initially

https://doi.org/10.1038/s41467-025-57150-9



grafted the DeepCpf1 algorithm onto other variants of Cas12a, but the results showed that it did not perform as well as expected (Supplementary Fig. 17). The main reason for this might be the poor correlation between different Cas12a on the given target sequences. Using these activities of Cas12a variants measured against thousands of target sequences, we developed the deep learning model based on convolutional neural network that predict the activities of the other Cas12a variants (Supplementary Fig. 18). With the training datasets, we performed 5-fold cross-validation to optimize the hyperparameter of model (named seq-DeepCpf1variants) (Fig. 4a and Supplementary Dataset 5). We next assessed seq-DeepCpf1variants using the test datasets that had not been used for training. As shown in Fig. 4b, the Pearson correlation coefficients ranged from 0.80 to 0.92 (average 0.86, median 0.86) and the Spearman correlation coefficients ranged

https://doi.org/10.1038/s41467-025-57150-9

Fig. 2 | Massively parallel evaluation of general activities and PAM compatibilities of Cas12a variants. a PAM compatibilities of the Cas12a variants in human cells. A heat map showing the average indel frequencies induced by the Cas12a variants at the 15 PAMs that are currently known to be recognized by Cas12a. **b** Pearson correlation coefficients between the Cas12a variant induced indel frequencies measured at the target sequences containing the same protospacers 7 days after transduction of Cas12a variant lentivirus into libraries integrated HEK293T cells. **c** Scatter plots showing the correlations of the five cases that are labeled as 1, 2, 3, 4 and 5 in the **b**. The Spearman correlation coefficient (ρ) and the Pearson correlation coefficient (r) are shown. In **b** and **c**, 15 PAM sequences were used for the analyses. **d** Comparison of the frequency of indels induced by the Cas12a variants at TTTA, TTTC, TTTG PAM; The boxes represent the 25th, 50th, and 75th percentiles, and the whiskers represent the 10th and 90th percentiles. The numbers of analyzed target sequences (n) are as follows: n = 1195, 1223, 1224, 1157, 1191, 1178, 1171 for TTTA, n = 1198, 1642, 1621, 1560, 1591, 1585, 1569 for TTTC, n = 1711, 1753, 1729, 1682, 1706, 1694, 1687 for TTTG (As variants); n = 1240, 1232 for TTTA, n = 1649, 1645, n = 1762, 1775 for TTTG (Eb variants); n = 1295, 1283 for TTTA, n = 1675, 1674, n = 1809, 1805 for TTTG (Lb2 variants); n = 1250 for TTTA, n = 1653 for TTTC, n = 1780 for TTTG (CeCas12a); n = 1202, 1216, 1244, 1300, 1243, 1199, 1210, 1165, 1174 for TTTA, n = 1629, 1623, 1682, 1615, 1646, 1647, 1597, 1596, 1623 for TTTC, n = 1753, 1755, 1809, 1738, 1780, 1778, 1699, 1697, 1748 for TTTG (LbCas12a variants); n = 1205, 1224, 1201 for TTTA, n = 1622, 1632, 1618 for TTTC, n = 1742, 1761, 1745 for TTTG (FnCas12a variants); paired t-test, two tailed. Source data are provided as a Source Data file.

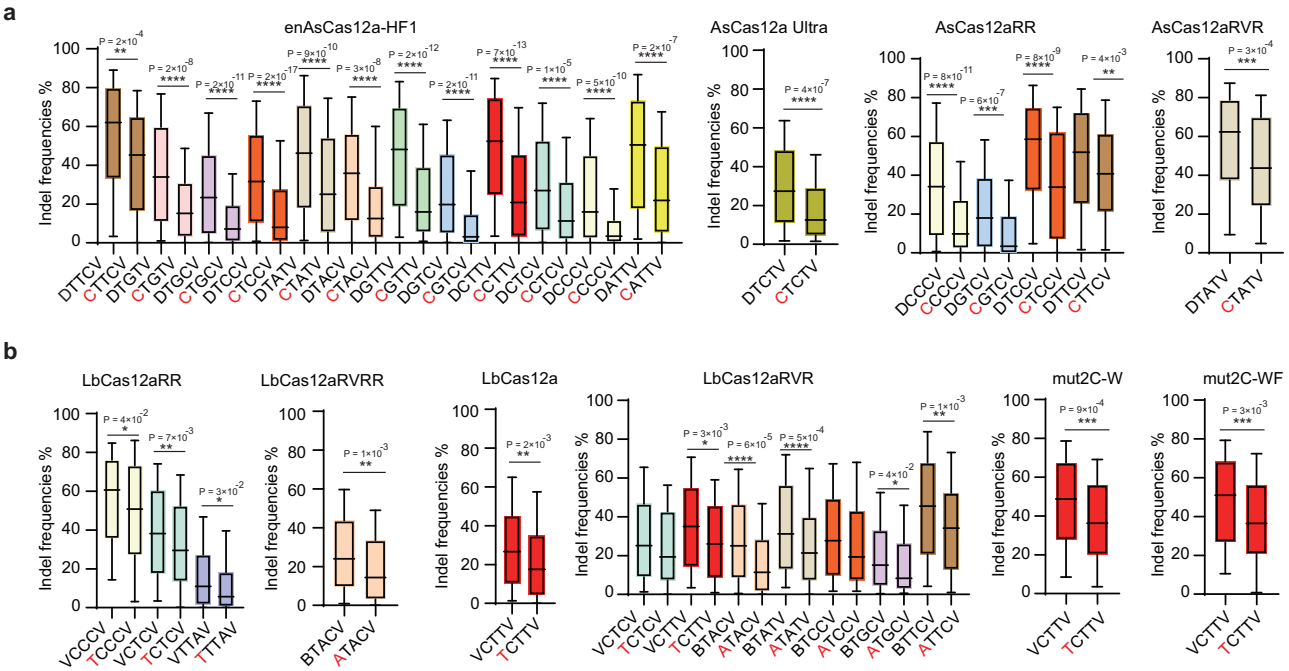


Fig. 3 | Effect of PAM 5′ terminal nucleotides on Cas12a cleavage to target sequences. Indel frequencies for different PAM sequences for enAsCas12a-HF1, AsCas12a Ultra, AsCas12aRR, AsCas12aRVR (**a**) and LbCas12aRR, LbCas12aRVRR, LbCas12a, LbCas12aRVR, mut2C-W, mut2C-WF (**b**). V = A/C/G; D = A/G/T; B = C/G/T. The boxes represent the 25th, 50th, and 75th percentiles, and the whiskers represent the 10th and 90th percentiles. The numbers of analyzed target sequences (n) are as follows: n = 359 for DTTCV, 124 for CTTCV, 366 for DTGTV, 127 for CTGTV, 374 for DTGCV, 126 for CTGCV, 372 for DTCCV, 125 for CTCCV, 237 for DTATV, 125 for CTATV, 363 for DTACV, 118 for CTACV, 372 for DGTTV, 122 for CGTTV, 367 for DGTCV, 125 for CGTCV, 363 for DCTTV, 122 for CCTTV, 369 for DCTCV, 120 for CCTCV, 374 for DCCCV, 125 for CCCCV, 358 for DATTV, 121 for CATTV (enAsCas12aHF1); n = 374 for DTCTV, 127 for CTCTV (AsCas12aUltra); n = 374 for DCCCV, 125 for CCCCV, 367 for DGTCV, 125 for CGTCV, 373 for DTCCV, 125 for CTCCV, 358 for DTTCV, 122 for CTTCV (AsCas12aRR); n = 351 for DTATV, 120 for CTATV (AsCas12aRVR); n = 380 for VCCCV, 127 for TCCCV, n = 377 for VCTCV, 122 for TCTCV, 369 for VTTAV, 126 for TTTAV (LbCas12aRR); n = 365 for BTACV, 123 for ATACV (LbCas12aRVRR); n = 374 for VCTTV, 124 for TCTTV (LbCas12a); n = 376 for VCTCV, 124 for TCTCV, n = 365 for VCTTV, 120 for TCTTV, 362 for BTACV, 125 for ATACV, 353 for BTATV, 117 for ATATV, 375 for BTCCV, 127 for ATCCV, 358 for BTGCV, 124 for ATGCV, 368 for BTTCV, 125 for ATTCV (LbCas12aRVR); 365 for VCTTV, 121 for TCTTV (mu2c-W); n = 367 for VCTTV, 123 for TCTTV (mu2c-WF); paired t-test, two tailed. Source data are provided as a Source Data file.

from 0.79 to 0.91 (average 0.86, median 0.86) (Supplementary Fig. 19 and Supplementary Dataset 6).

To validate the accuracy of predictions between different models, we determined the Spearman correlation coefficients on different test datasets, and found that the corresponding test sets did not perform as well in each other's models as they did in their own models (Fig. 4c). To address this issue, we merged the training sets used in the two models to generate a larger training set for training the model to obtain enseq-DeepCpf1 and tested it with different test sets (Fig. 4c). The results show that enseq-DeepCpf1 has high accuracy in test sets (Fig. 4c, d). In addition, chromatin accessibility has been reported to affect the activities of Cas12a on endogenous targets[55,57]. To further improve prediction accuracy, we developed enDeepCpf1 by fine-tuning the enseq-DeepCpf1 using chromatin accessibility information[58]. When evaluated using HEK-Plasmid and HCT-Plasmid as test datasets, enDeepCpf1 showed significantly better performance than enseq-DeepCpf1 (Fig. 4e). To verify the generalization ability of the enDeepCpf1, three new sets of independent tests were performed[59,60] and showed that the Pearson correlation coefficients were 0.68, 0.64, and 0.74, and the Spearman correlation coefficients were 0.68, 0.73, and 0.83, respectively (Fig. 4f). All of these tests showed excellent generalization performance. To compare our models, named enseq-DeepCpf1 for integrated sites and enDeepCpf1 for endogenous sites, among the current models with good performance, we evaluated the performance of four deep learning or machine learning models[55,57,61,62] as indicated by the correlation coefficients

https://doi.org/10.1038/s41467-025-57150-9

between measured and predicted indel frequencies on several test datasets. As shown in Fig. 4g, h, the Pearson correlation coefficients ranged from 0.62 to 0.78 and the Spearman correlation coefficients ranged from 0.65 to 0.83. Compared to other models, our model showed excellent performance on all test datasets.

Developing enGUIDE-seq and evaluating off-targeting and translocations of Cas12a variants

Genome-wide unbiased identification of DSBs by sequencing (GUIDE-seq) is one of the most popular methods to detect true breakpoints in living cells after genome editing[63–65]. We found that turn-over cleavage of a marker oligodeoxynucleotide (ODN) by the Cas12a-crRNA complex results in a truncated tag (Fig. 5a), which in turn leads to an incomplete matching of the tag-specific primer with the tag sequence in GUIDE-seq (Fig. 5b). As a result, some integration sites are missed (Fig. 5c and Supplementary Fig. 20). Therefore, we present an enhanced version of the GUIDE-seq method (enGUIDE-seq) that improves the amplification of the tag and compensates for the omissions in the original method due to the incomplete tag (Fig. 5d and Supplementary Figs. 21, 22). It was used to evaluate the off-targeting of 24 Cas12a variants, on 31 targets with a large number of homopolymeric sequences in human cells (Supplementary Fig. 23). We

Fig. 4 | Development of computational models to predict the activities of Cas12a variants. a 5-fold cross-validation of seq-DeepCpf1variants models on training data sets. The Pearson correlation coefficients (left) and the Spearman correlation coefficients (right) were shown. Boxes represent the 25th, 50th, and 75th percentiles, and whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. **b** Evaluation of seq-DeepCpf1variants computational models predicting the activities of Cas12a variants on data sets of indel frequencies that were never used for the training. The Spearman correlation coefficient (ρ) and the Pearson correlation coefficient (r) are shown. **c** Evaluation of enseq-DeepCpf1 computational models predicting the activities of AsCas12a on test data sets As HT and HT 1-2; As HT test data set was split from AsCas12a induced high-throughput data set; HT 1-2 test data set and seq-DeepCpf1 computational models are derived entirely from the Kim 2017[57]; The Spearman correlation coefficients are shown. Data are presented as mean values +/− SEM. **d** Correlation between enseq-DeepCpf1 prediction scores and measured indel frequency ranks at an independent test Kim 2017[55] ($n = 82$, guide RNA-target sequence pairs, 293 T). **e** Development and evaluation of enDeepCpf1 computational models after consideration of chromatin accessibility. Performance comparison of enDeepCpf1 with enseq-DeepCpf1 in HCT116 cells and HEK293T cells; The training data set of HEK-lenti, the test data sets of HCT plasmid and HEK plasmid are derived entirely from the Kim 2017[57]. Data are presented as mean values +/− SEM. **f** Correlation between enDeepCpf1 prediction scores and measured indel frequency ranks at three independent tests (Yin ($n = 15$, transfection, 293 T); Chari 2017[59] ($n = 38$, transfection, 293T); Kim 2016[60] ($n = 10$, transfection, 293T)). The Spearman correlation coefficients are shown. **g, h** Model comparation at integrated and endogenous sites on test data sets. Pearson and Spearman correlation coefficients between different models and data sets of integrated target sites (**g**) and data sets of endogenous target sites (**h**). The test data sets are arranged vertically, whereas the prediction models are placed horizontally. Correlation coefficient values are listed in boxes. $^{****}P < 0.0001$ by Two-tailed paired t-test. Source data are provided as a Source Data file.

calculated an activity and specificity score for each variant, and found that a general trade-off between activity and specificity among all variants (Fig. 5e). The general specificities of Cas12a variants were as follows: HyperFi-AsCas12a (0.7), LbCas12a Plus (0.67), LbCas12a K538R (0.63), CeCas12a (0.62), EbCas12a (0.59), FnCas12a (0.59), Lb2Cas12a K518R (0.57), AsCas12a (0.56), AsCas12a Plus (0.54), Lb2Cas12a (0.54), enEbCas12a (0.52), LbCas12aRVRR (0.52), eaFnCas12a (0.50), LbCas12aRVR (0.49), LbCas12a (0.48), AsCas12aRR (0.45), FnCas12aRVR (0.44), AsCas12aRVR (0.44), AsCas12a Ultra (0.42), LbCas12aRR (0.38), mut2C-W (0.32), enAsCas12a-HF1 (0.30), mut2C-WF (0.26), HyperLbCas12a (0.14). Furthermore, the frequency of translocations was detected among these 31 targets and the correlations were poor between translocation frequency and either activity or specificity across all variants (Fig. 5f–h). This suggests that there could be additional factors within the cell that influence translocation. Overall, AsCas12a variants exhibit relatively low translocation frequencies compared to LbCas12a variants despite big differences in activity and off-targeting of these variants, which might suggest that the mechanism of translocation is complex and might be influenced by the intrinsic properties of Cas12a orthologues themselves in addition to their activity and off-targeting.

Discussion

Although Cas12a nuclease offers a powerful potential for genome engineering, it is relatively time-consuming and laborious to select the appropriate Cas12a for a given sequence among the many variants. In this study, we generated 24 large datasets in HEK293T cells to analyze the activity of these Cas12as on a total of 11968 lentiviral integrating target sequences, including all 15 PAMs currently known to be recognized by Cas12a. We extensively characterized the PAM compatibilities and measured the editing activities at thousands of target sequences for 24 Cas12a variants. We also developed deep learning models based on convolutional neural network to predict the activities of Cas12a variants. In addition, we developed an enhanced version of GUIDE-seq, termed enGUIDE-seq, and evaluated the off-targeting and translocations of these Cas12a variants. This comprehensive evaluation, together with the developed enGUIDE-seq and deep learning methods for predicting activity, will greatly facilitate the selection of guide RNA sequences and appropriate Cas12a candidates for gene editing.

It is known that crRNA design, target site content and structure, chromatin niche, and other currently unknown factors may affect editing activity[55,56,66]. Therefore, comparing the efficiency of gene editing of these variants on a small number of targets may not be optimal. High-throughput analysis of CRISPR/Cas nuclease activity using a guided RNA-target pairing strategy minimizes these effects and ensures reliable results. Another way to obtain high-throughput data is to introduce Cas12a and thousands of crRNAs into cells by lentiviral transduction or plasmid transfection, followed by PCR amplification of all editing targets and high-throughput sequencing. This approach is more labor-intensive than the RNA-target pairing strategy. However, it has the advantage of being able to determine in situ editing efficiency while also obtaining the effect of chromatin accessibility on the editing efficiency of these Cas proteins.

In the future, it will be necessary to develop methods that can accurately analyze the editing efficiency of Cas proteins at endogenous targets across the genome on a large scale. This will facilitate the confirmation of the activity of various Cas proteins for any given cell types as well as the construction of better models.

Methods

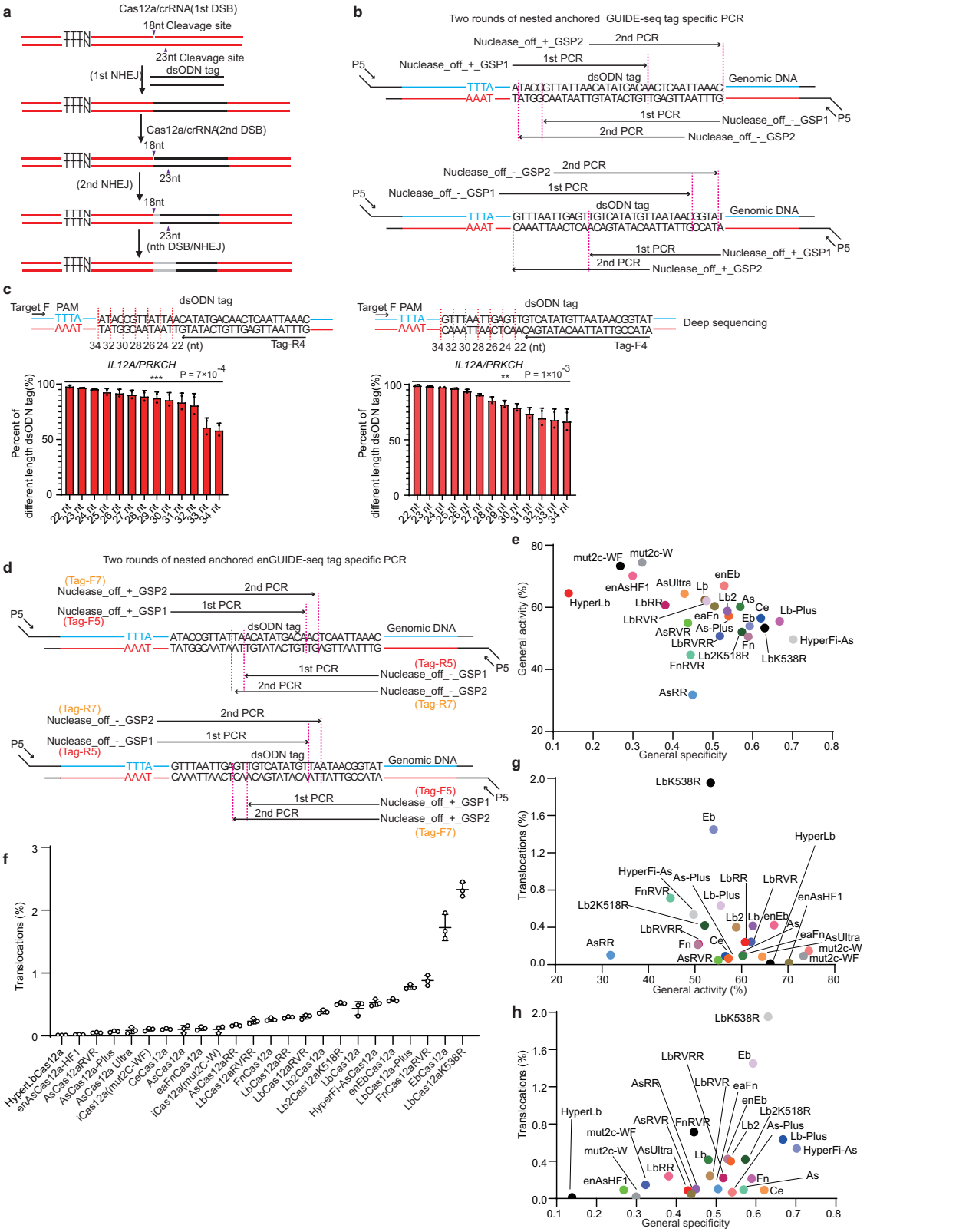
Oligonucleotides

To construct the plasmid library, four oligonucleotide pools were array-synthesized by Azenta Life Sciences, namely the library As, library Lb, library Lb2, and library Ce/Eb/Fn. Each pool contains 11,968 pairs of guide sequences and corresponding target sequences oligonucleotide. Fifteen PAM sequences were used for each oligonucleotide pool: TTTV (4800 pairs), CTTV (512 pairs), ATTV (512 pairs), GTTV (512 pairs), TTCV (512 pairs), TCTV (512 pairs), TCCV (512 pairs), CCCV (512 pairs), CTCV (512 pairs), TATV (512 pairs), TACV (512 pairs), TGTV (512 pairs), TTAV (512 pairs), GTCV (512 pairs), TGCV (512 pairs). Each oligonucleotide was designed to include the 20-nt guide-RNA-encoding sequence, 19-nt barcode, and the corresponding 31-nt target sequence with a PAM sequence in a total length of 135 nt nucleotides. For each PAM, we designed a sequence that is at least one nucleotide (nt) longer than the previously characterized PAM sequences. For example, NTTTV (ATTTV, 1200 pairs; GTTTV, 1200 pairs; CTTTV, 1200 pairs; TTTTV, 1200 pairs); NCTTV (ACTTV, 128 pairs; GCTTV, 128 pairs; CCTTV, 128 pairs; TCTTV, 128 pairs); NATTV (AATTV, 128 pairs; NATTV (AATTV, 128 pairs; GATTV, 128 pairs; CATTV, 128 pairs; TATTV, 128 pairs); NGTTV (AGTTV, 128 pairs; GGTTV, 128 pairs; CGTTV, 128 pairs; TGTTV, 128 pairs) and so on. The list of nucleotides can be found in Supplementary Dataset 1.

Plasmid library preparation

pRC11-U6-DR-crRNA-BsmbI(x2); EFS-Puro-WPRE plasmid (Addgene, #123360) was linearized with SE BsmBI restriction enzyme (SibEnzyme) at 55 °C for 2 h. The linearized plasmids were separated on a 1% agarose gel and purified using a SanPrep Column DNA Gel Extraction Kit (Sangon Biotech). The pooled oligonucleotides were PCR amplified using Phanta Super-Fidelity DNA Polymerase (Vazyme). The PCR products were separated on a 2% agarose gel and purified using a SanPrep Column DNA Gel Extraction Kit (Sangon Biotech) and then treated with SE BsmBI restriction enzyme (SibEnzyme) at 55 °C for 2 h. The purified amplicons and the linearized pRC11-U6-DR-crRNA-BsmbI(x2); EFS-Puro-WPRE were ligased using T4 DNA ligase (Rapid) (Vazyme) at 16 °C

https://doi.org/10.1038/s41467-025-57150-9



overnight. After incubation, the products were transformed into electrocompetent cells via a MicroPulser (Bio-Rad). Transformed cells were then spread on Luria-Bertani (LB) agar plates supplemented with 50 μg/mL carbenicillin and incubated for 16 h at 37 °C. Before harvest, the library coverage was calculated as (total number of colonies/total number of crRNA-target pairs in sample)[57]. The resulting library coverage ranged from 40× to 50×. Total colonies were collected and the plasmids were extracted using a EndoFree Maxi Plasmid Kit (TIANGEN).

Construction of plasmids encoding Cas12a variants

All plasmids and guide RNAs used in this study can be found in Supplementary Dataset 7. AsCas12a (Addgene, #69982), LbCas12a (Addgene, #69988), FnCas12a (Addgene, #69976), Lb2Cas12a

Fig. 5 | Assessing the specificity of 24 Cas12a variants using an enhanced vision of GUIDE-seq. a A schematic illustration showing turn-over dsDNA cleavage and NHEJ cycles for the Cas12a/crRNA complex. Cleavage sites are marked with triangles and ash-gray line indicate changes in the dsODN tag sequences. **b** Two rounds of nested anchored GUIDE-seq tag specific PCR. The primers Nuclease_off_+/−_GSP1 were used to the first PCR; The primers Nuclease_off_+/−_GSP2 were used to the second PCR; P5 means the Illumina P5 adapter primers. **c** Incomplete dsODN tag sequences induced by Cas12a/crRNA complex in HEK293T cells at *IL12A* and *PRKCH* targets. The bar graph shows the percentage of different lengths of dsODN tag that can be detected after 48 h transfection of AsCas12a and crRNA expression plasmid as well as 34 nt dsODN tag. The primers Target F/Tag-R4 and Target F/Tag-F4 were used for the PCR amplification of the sites integrated with the dsODN tag. The percentage of different lengths of the dsODN tag was determined by deep sequencing. Data are presented as mean values +/− SEM. ****$P < 0.0001$ by Two-tailed paired *t*-test. **d** Two rounds of nested anchored enhanced vision of the GUIDE-seq tag specific PCR. Notably, the primers Tag-F5 (Nuclease_off_+_GSP1) and Tag-R5 (Nuclease_off_-_GSP1) were used to the first PCR; The primers Tag-F7 (Nuclease_off_+_GSP2) and Tag-R7 (Nuclease_off_-_GSP2) were used to the second PCR; P5 means the Illumina P5 adapter primers. **e** Comparison of the general activities and specificities of the Cas12a variants. The general activity score of each Cas12a was calculated as the average indel frequency at the target sequences containing the same protospacers 7 days after transduction of Cas12a variant lentiviruses into library integrated HEK293T cells. The general specificity score of each Cas12a calculated from the enhanced vision of GUIDE-seq identified off-targets sites. **f** Total translocations induced by Cas12a variants. Data are presented as mean values +/− SEM. **g** Comparison of the general activities and translocations of the Cas12a variants. **h** Comparison of the general specificities and translocations of the Cas12a variants. Source data are provided as a Source Data file.

(Addgene, #69983) human expression plasmids were purchased from the non-profit plasmid repository Addgene. To prepare the lentiviral vectors, we removed the Cas9-encoding sequence from the lentiCas9-Blast plasmid (Addgene, #52962), and cloned Cas12a-encoding sequences between XbaI and BamHI. In all initial studies of CRISPR/Cas nucleases for high-throughput evaluation in mammalian cells, the EF1a core promoter was commonly used in lentiviral vectors to transcribe Cas mRNA. We first constructed expression cassettes using the EF1 promoter, validated the reliability of the system in HEK293T cells, and found that As, Lb showed very modest editing effects against two endogenized crRNA-target pairs (Supplementary Fig. 24a–d). However, when we used the CMV promoter, there was a significant increase in the editing activity of both Cas12as (Supplementary Fig. 24f). This finding is also consistent with recent reports that promoters can influence the editing activity of Cas12a in cells[67]. Therefore, we replaced the EF1 promoter upstream of the Cas12a-encoding DNA sequences with CMV promoter to obtained lentiCMV-As/Lb/Fn/Lb2/Ce/Eb-Blast. All Cas12a variants were generated by standard site-directed mutagenesis. Briefly, using lentiCMV-As/Lb/Fn/Lb2/Ce/Eb-Blast plasmid as a template and a pair of primers, one primer carrying site mutation nucleotide to amplify 500 bp fragments by PCR (Phanta MAX Super-Fidelity DNA Polymerase P505, Vazyme). After confirming the correct bands by agarose gel electrophoresis, the PCR products were purified. Next, 1000 ng of PCR products were used as circular mutation primers and 100 ng of lentiCMV-As/Lb/Fn/Lb2/Ce/Eb-Blast plasmid as template to perform PCR again. PCR products were purified and DpnI was used to digest the original template at 37 °C for 1 h, inactivated at 80 °C for 20 min, take 10 μL of digested products for transformation and single colonies were sent for sequencing in the next day. Plasmids were extracted using an EndoFree Maxi Plasmid Kit (TIANGEN).

Construction of plasmids expressing Cas12a crRNA

To perform GUIDE-seq, we constructed several crRNA expression plasmids. We performed PCR to linearize the pRC11-U6-DR-crRNA-BsmbI(x2); EFS-Puro-WPRE plasmid (Addgene, #123360). The oligonucleotide duplexes corresponding to the spacer sequences were also amplified by PCR and then cloned into the linearized pRC11 for human U6 promoter-driven transcription Cas12a crRNAs.

Lentivirus production

HEK293T (Cat. No.CRL-3216) cells were obtained from the ATCC and maintained in DMEM medium supplemented with 10% fetal bovine serum and 100 units mL^{-1} penicillin, 100 μg mL^{-1} streptomycin sulfate (all cell culture products were obtained from Gibco) at 37 °C in 5% CO_2. For lentivirus production, transfer plasmids (containing the gene of interest), psPAX2, and pMD2.G were mixed at a ratio of 5:3:2, and a total of 20 μg of the plasmid mixture was transfected into 70–80% confluent HEK293T cells (ATCC) per 100 mm plate using Hieff Trans Liposomal Transfection Reagent (CAT:40802ES03, Yeasen, Shanghai). After 8 h of transfection, the culture medium was changed with 15 mL of growth medium. Virus-containing supernatants were collected 48 and 72 h after the initial transfection. The first and second batches of virus-containing media were combined and centrifuged at 1500 g for 5 min at 4 °C and then the supernatants were filtered through a Millex-HV 0.45-μm low-protein-binding membrane (Millipore) and stored at −80 °C until use. To determine lentiviral titers, serial dilutions of the lentivirus were transduced into HEK293T cells in the presence of 10 μg mL^{-1} polybrene (Yeasen, Shanghai), and the culture medium was replaced with fresh growth medium after 12 h of lentivirus treatment. After another 36 h incubation, both virus-treated and untreated cells were cultured in medium containing 10 μg mL^{-1} blasticidin S (Yeasen, Shanghai) or 2 μg mL^{-1} puromycin (Yeasen, Shanghai) until no viable cells were present in the untreated cell population. The viral titer was estimated by counting the number of surviving cells in the virus-treated population when almost all of the untransduced cells had died[68,69].

Cell library generation

Four cell libraries were constructed independently as described below. HEK293T cells were seeded into four 150-mm tissue culture dishes (2.0×10^7 cells per dish) and grown overnight. The library-lentivirus were transduced into the cells at a multiplicity of infection (MOI) of 0.4 in the presence of 10 μg mL^{-1} polybrene (Yeasen, Shanghai), and the culture medium was replaced with fresh growth medium after 12 h of lentivirus treatment. After another 36 h incubation, the cell populations were trypsinized and split into quartet wells, and cultured in the presence of 2 μg mL^{-1} puromycin to remove the non-transduced cells for the following 4 d.

Cas12a delivery into the cell library

For transduction of Cas12a-expressing lentiviral vectors, approximately 4.0×10^6 cells were seeded and transduced with Cas12a-encoding lentiviral vectors at a MOI of 1 in the presence of 10 μg mL^{-1} polybrene (Yeasen, Shanghai), and the culture medium was replaced with fresh growth medium after 12 h of lentivirus treatment. After another 36 h incubation, the cell populations were trypsinized and split into quartet wells, and cultured in the presence of 10 μg mL^{-1} blasticidin S (Yeasen, Shanghai) to remove the non-transduced cells for the following 7 d. All steps were repeated equivalently for each cell library.

Deep sequencing

Genomic DNA was extracted using FastPure Cell/Tissue DNA Isolation Mini Kit DC102 (Vazyme). The integrated target sequences were PCR amplified using 2×Hieff Canace Plus PCR Master Mix (With Dye) (Yeasen, Shanghai). To achieve >1000× coverage over the library (assuming 10 μg of genomic DNA for 1.0×10^6 293 T cells)[55], we used

100 μg genomic DNA of each Cas12a transduced cell library as template for the first PCR. We performed 20 separate 100-μL reactions for each cell library, using an initial genomic DNA concentration of 5 μg per reaction, and then pooled all of the resulting products. For the second PCR, 2 μL the first PCR products were used to anneal with both Illumina adaptor and barcode sequences. After confirming the correct bands by agarose gel electrophoresis, the PCR products were purified and analysed using HiSeq (Illumina). To generate the test data Yin ($n = 15$, transfection, 293 T), approximately 1.2×10^5 cells were seeded into per well of 24-well plate a day before transfection, 500 ng Cas12a and 250 ng crRNA expression plasmids were transfected into cells using Hieff Trans Liposomal Transfection Reagent (CAT:40802ES03, Yeasen, Shanghai). 48 h post-transfection, cells were collected by centrifugation and the supernatants were removed. 50 μL Lysis buffer and 0.5 μL Proteases were mixed with cells and incubated at 55 °C for 30 min, 95 °C for 30 min (Animal Tissue Lysis Component, CAT: 19698ES70, Yeasen, Shanghai). The genomic region flanking the CRISPR target site for each gene was amplified by PCR with Phanta MAX Super-Fidelity DNA Polymerase P505 (Vazyme) using 1 μL cell lysis as template. 2 μL the first PCR products were used to anneal with both Illumina adaptor and barcode sequences. The primers used for PCR are shown in Supplementary Dataset 7.

Analysis of indel frequencies from deep-sequencing data

Pair-end deep sequencing data were first merged using FLASH2 (v2.2.0) with default parameters. Indel frequencies were then analyzed using custom Python scripts. Briefly, the data were demultiplexed by 19 bp target site barcodes. Pairwise global alignment was then performed between the merged reads and the original target sequences using Biopython. An 8 bp region centered in the middle of the expected cleavage site was selected for statistical insertions and deletions. Single base indels or substitutions were excluded from analysis. Reads with indels located within the selected region were considered to be edited by Cas12a. Indel frequencies were calculated as the proportion of edited reads for each Cas12a variant. The background indel frequency in the cell library in the absence of Cas12a delivery was subtracted from the observed indel frequency.

Datasets used for the development and evaluation of deep learning models

The target sequences and indel frequencies measured by deep sequencing were used to generate the data sets. The 31 bp target sequences consisted of a 3 nt 5′ flanking sequence, a 4 nt PAM, a 20 nt protospacer and a 4 nt 3′ flanking sequence. For each Cas12a variant model, 1000 target sequences (approximately 10%) were randomly selected as test data sets and the remaining data were used as training data sets. For the EnDeepCpf1 model, the training set contains two parts: 1) 90% target sequences (3502 sequences) with TTTV PAM of AsCas12a and 2) HT1-1 data set with 14139 sequences with TTTN PAM from Kim et al. The remaining 10% target sequences (389 sequences) with TTTV PAM of AsCas12a (referred to as As HT) together with HT1-2 from Kim et al.[57] were used as test sets. In addition, we evaluated enseq-DeepCpf1 using the published data set that generated by Kim 2017[55] ($n = 82$, guide RNA-target sequence pairs, 293 T) as test data. The training data sets and test data sets can be found in Supplementary Datasets 8, 9.

Convolutional Neural Network for Deep Learning

seq-DeepCpf1-variants and enseq-DeepCpf1 are regression models developed using a convolutional neural network (CNN). The model takes a 31 bp target sequence as input and returns a predicted indel frequency. The nucleotide sequences are first transformed into a 4-by-31-dimensional binary matrix via one-hot encoding. The matrices are then fed into a one-dimensional convolutional layer with 128 filters of length 5 and activated with the Rectified Linear Unit (ReLU) function. This is followed by another one-dimensional convolutional layer with 128 filters of length 5 and a ReLU function. No pooling layers were used in our model. The matrices are then flattened and entered sequentially into two fully connected layers of 128 units each. A ReLU function and a dropout rate of 0.3 are used in each fully connected layer. The output dense layer has a single unit with a sigmoid function. 5-fold cross validation was implemented on training data sets to select the better model hyperparameters. Then, the model was trained on training data sets and tested on test data sets. The mean squared error (MSE) loss function was used during training. We optimized our models using the Adam optimizer with a learning rate of 1e-4 and a batch size of 64. We use early stopping with a patience of 10 epochs to avoid overfitting. The model parameters of the best epoch were stored. We implemented our models using Python (v3.10.6) and Pytorch (v1.12.1). The following packages were also used Pandas (v1.4.3), Numpy (v1.21.5), Scipy (v1.7.3), Cudatoolkit (11.6.0).

Fine-tune model using endogenous target data with chromatin accessibility information

The HEK-lenti dataset ($n = 148$) obtained by Kim et al.[57] was used as a training set to fine-tune the model trained on the high-throughput deep sequencing data. To incorporate chromatin accessibility information, an additional 64-unit fully connected layer is added to the model to transform the binary chromatin accessibility into a 64-dimensional numerical vector. This layer is activated by the ReLU function after a dropout layer with a dropout rate of 0.3. The 64-dimensional vector and the output of the last fully connected layer are concatenated and then fed into the output layer. During fine-tuning, convolutional layers were fixed so that only fully connected layers and output layer weights were updated. Models were tested on HEK plasmid ($n = 55$) and HCT plasmid ($n = 66$) datasets[57]. In addition, we generated a data set named Yin ($n = 15$, transfection, 293 T) and used it as test data. We also used other published data sets of different studies from independent laboratories (Chari 2017[59] ($n = 38$, transfection, 293 T), Kim 2016[60] ($n = 10$, transfection, 293 T)) as test data. The training data sets and test data sets can be found in Supplementary Datasets 9.

Comparison with other models

To compare our models, named enseq-DeepCpf1 for integrated sites and enDeepCpf1 for endogenous sites, among the current models with good performance, we evaluated the performance of four deep learning or machine learning models, indicated by correlation coefficients between measured indel frequencies and predicted indel frequencies on several test datasets. DeepGuide[61] is a deep learning model trained on genome-wide data from the oleaginous yeast Yarrowia lipolytica. We tested the original DeepGuide model on test datasets. Given that the test datasets were obtained from experiments using HEK293T cells, we retrained DeepGuide on the same training dataset as our enseq-DeepCpf1 and enDeepCpf1 to avoid bias between the training and test datasets. For integrated site prediction, DeepGuide was trained on 80% sites of the total training dataset (17508 target sites). The remaining 20% sites were used as validation set. An early termination criterion is used to stop training, i.e., the loss of the model on the validation set does not decrease for 10 epochs. For endogenous sites, the retrained DeepGuide with integrated sites was fine-tuned with endogenous training sets (148 sites). To be consistent with enDeepCpf1, chromatin accessibility was also added as an input. Seq-DeepCpf1 and DeepCpf1[57] were accessed directly via the web service (https://deepcrispr.info). For the integrated sites data set, only Seq-DeepCpf1 was evaluated. For endogenous sites, both Seq-DeepCpf1 and DeepCpf1 were evaluated. CINDEL[55] and CRISPR-DT[62] are both machine learning based models that use sequence features including position-independent features, position-dependent

features, melting temperature, GC count, and minimum free energy of the guide RNA sequence. Since the web services in the original paper are not available and no model codes were provided, we retrained CINDEL and CRISPR-DT using features selected from the original paper. The number of features is fifty-seven for CINDEL and one hundred and five for CRISPR-DT. The original CINDEL (logistic regression model) and CRISPR-DT (support vector machine) are both classification model, however, classification model has worse performance than regression models on Cpf1 activity prediction[57]. So, linear regression model (CINDEL retrain) and support vector regression model (CRISPR-DT retrain) were developed. For consistency, chromatin accessibility was added as an additional feature for endogenous sites. DeepGuide was trained using Keras, and CINDEL and CRISPR-DT were trained using scikit-learn. For endogenous sites, 27 bases of the protospacer adjacent motif (PAM) plus protospacer sequence were aligned to the GRCh38 human reference genome using bowtie2[70]. Target sites overlapping with the DNase-seq narrow peak obtained from the Encyclopedia of DNA Elements (ENCODE)[58] were considered chromatin accessible. The training data sets and test data sets can be found in Supplementary Datasets 9.

GUIDE-seq

HEK293T cells were cultured in DMEM medium supplemented with 10% fetal bovine serum and 100 units mL^{-1} penicillin, 100 μg mL^{-1} streptomycin sulfate (all cell culture products were obtained from Gibco) at 37 °C in 5% CO_2. One day prior to transfection, approximately 5×10^5 cells were seeded per well in a 6-well plate. The next day, 500 μL OptiMEM (Thermo) were mixed with 4 μg Cas12a expression plasmid, a total of 3 μg of 1–31 different crRNA expression plasmids (sequences in Supplementary Dataset 7), and 60 pmol of the double stranded oligodeoxynucleotide (dsODN) GUIDE-seq tag. In parallel, 500 μL OptiMEM were mixed with 20 μL Hieff Trans Liposomal Transfection Reagent (Yeasen, Shanghai). After mixing all components and incubating for 20 min, 250 μL added dropwise per cell well, totaling two wells per condition. 48 h after transfection, the genomic DNA was harvested and purified using FastPure Cell/Tissue DNA Isolation Mini Kit DC102 (Vazyme). 1 μg genomic DNA was fragmented, end repaired, A-tailing by using Hieff NGS Fast-Pace DNA Fragmentation Reagent (12609ES96, Yeasen, Shanghai). Two rounds of nested anchored PCR, with primers complementary to the oligo tag, were used for target enrichment. Notably, the only difference between the revised version of the GUIDE-seq method and original GUIDE-seq method is the genomic-specific primers (GSP). The primers used for PCR are shown in Supplementary Dataset 7. Specificity scores were calculated by subtracting from the percent of GUIDE-seq reads that corresponds to off-targets[71].

T7E1 assay

250 ng of purified PCR products were mixed with 1 μL 10×T7E1 buffer (Vazyme) and ultrapure water to a final volume of 10 μL, and subjected to a re-annealing process to enable heteroduplex formation: 95 °C for 3 min, 95 °C for 30 s, 90 °C for 30 s, 85 °C for 30 s, 80 °C for 30 s, 75 °C for 30 s, 70 °C for 30 s, 65 °C for 30 s, 60 °C for 30 s, 55 °C for 30 s, 50 °C for 30 s, 45 °C for 30 s, 40 °C for 30 s, 35 °C for 30 s, 30 °C for 30 s, and 25 °C for 1 min. After re-annealing, products were treated with T7 Endonuclease I (EN303-01/02, Vazyme) for 15 min at 37 °C. The reaction mixtures were run on 2% agarose gels. The primers used for PCR are shown in Supplementary Dataset 7.

Western blotting

To detect the expression of Cas12a variants, cells were harvested and lysed after 7 days of transduction. Lysates were resolved by SDS-PAGE electrophoresis and transferred to a polyvinylidene fluoride membrane (Millipore, USA). Membranes were blocked with blocking buffer (5% non-fat milk in Tris buffered saline with Tween20 (TBST)) for 2 h and then incubated with the following primary antibodies: HA Mouse primary antibody (CAT#901515, Biolegend, USA) at 1:20000 dilution and β-actin Rabbit primary antibody (CAT#AC026, ABclonal, China) at 1:20000 dilution for 3 h at room temperature, respectively. After washing steps in TBST, the membranes were incubated with HRP Goat anti-Mouse IgG (H + L) for HA (CAT#AS003, ABclonal, China) and HRP Goat anti-Rabbit for β-actin (CAT#AS014, ABclonal, China) at 1:50000 dilution for 1 h.

Statistics and reproducibility

No statistical method was used to predetermine sample size. Statistical analysis was performed using Microsoft Excel (2016) and GraphPad Prism 10 (version 10.1.2). Significance testing was conducted using a two-tailed *t*-test, with a significance level set at $P < 0.05$. The bar and dot plots were presented as the mean ± standard deviation (s.d.). The boxes were presented as interquartile ranges with the median indicated by a line and whiskers extending from the minimum to the maximum values. Gene editing experiments were performed at least in independent biological duplicates.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The authors declare that all data supporting the results in this study are available within the paper and its Supplementary Information. The deep-sequencing data from this study are available at the NCBI Sequence Read Archive under the accession number PRJNA1074843. Source data are provided with this paper.

Code availability

The custom Python scripts used for the indel-frequency calculations and the source code for seq-DeepCpf1variants, enseq-DeepCpf1, enDeepCpf1 are available at https://codeocean.com/capsule/9398276/tree/v1[72]. All source code is released under the MIT licence. Source data are provided with this paper.

References

1. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
2. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. USA* **109**, E2579–2586 (2012).
3. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167–170 (2010).
4. Huang, C. J., Adler, B. A. & Doudna, J. A. A naturally DNase-free CRISPR-Cas12c enzyme silences gene expression. *Mol. cell* **82**, 2148–2160.e2144 (2022).
5. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
6. Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evolution* **60**, 174–182 (2005).
7. Qi, L. S. et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **184**, 844 (2021).
8. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
9. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
10. Zhao, J., Lai, L., Ji, W. & Zhou, Q. Genome editing in large animals: current status and future prospects. *Natl. Sci. Rev.* **6**, 402–420 (2019).

11. Chang, P. C. et al. Generation of antigen-specific mature T cells from RAG1(-/-)RAG2(-/-)B2M(-/-) stem cells by engineering their microenvironment. *Nat. Biomed. Eng.* **8**, 461–478 (2023).
12. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
13. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
14. Gaudelli, N. M. et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
15. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
16. Nakamura, M., Gao, Y., Dominguez, A. A. & Qi, L. S. CRISPR technologies for precise epigenome editing. *Nat. Cell Biol.* **23**, 11–22 (2021).
17. Yin, S. et al. Engineering of efficiency-enhanced Cas9 and base editors with improved gene therapy efficacies. *Mol. Ther.: J. Am. Soc. Gene Ther.* **31**, 744–759 (2023).
18. Chen, L. et al. Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing. *Nat. Biotechnol.* **41**, 663–672 (2023).
19. Kim, J. S. & Chen, J. Base editing of organellar DNA with programmable deaminases. *Nat. Rev. Mol. Cell Biol.* **25**, 34–45 (2024).
20. Zhang, H., Li, T., Sun, Y. & Yang, H. Perfecting Targeting in CRISPR. *Annu. Rev. Genet.* **55**, 453–477 (2021).
21. Yin, H., Xue, W. & Anderson, D. G. CRISPR-Cas: a tool for cancer research and therapeutics. *Nat. Rev. Clin. Oncol.* **16**, 281–295 (2019).
22. Song, C. Q. et al. Adenine base editing in an adult mouse model of tyrosinaemia. *Nat. Biomed. Eng.* **4**, 125–130 (2020).
23. Qiu, H. Y., Ji, R. J. & Zhang, Y. Current advances of CRISPR-Cas technology in cell therapy. *Cell Insight* **1**, 100067 (2022).
24. 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).
25. Yamano, T. et al. Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* **165**, 949–962 (2016).
26. Stella, S., Alcón, P. & Montoya, G. Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage. *Nature* **546**, 559–563 (2017).
27. Dong, D. et al. The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature* **532**, 522–526 (2016).
28. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–771 (2015).
29. Zetsche, B. et al. Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. *Nat. Biotechnol.* **35**, 31–34 (2017).
30. Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A. & Charpentier, E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517–521 (2016).
31. Chen, J. S. et al. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439 (2018).
32. Li, S. Y. et al. CRISPR-Cas12a has both cis- and trans-cleavage activities on single-stranded DNA. *Cell Res.* **28**, 491–493 (2018).
33. Gootenberg, J. S. et al. Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science* **360**, 439–444 (2018).
34. Broughton, J. P. et al. CRISPR-Cas12-based detection of SARS-CoV-2. *Nat. Biotechnol.* **38**, 870–874 (2020).
35. Lu, S. et al. Fast and sensitive detection of SARS-CoV-2 RNA using suboptimal protospacer adjacent motifs for Cas12a. *Nat. Biomed. Eng.* **6**, 286–297 (2022).
36. Yan, H. et al. A one-pot isothermal Cas12-based assay for the sensitive detection of microRNAs. *Nat. Biomed. Eng.* **7**, 1583–1601 (2023).
37. Torriano, S., Baulier, E., Garcia Diaz, A., Corneo, B. & Farber, D. B. CRISPR-AsCas12a Efficiently Corrects a GPR143 Intronic Mutation in Induced Pluripotent Stem Cells from an Ocular Albinism Patient. *CRISPR J.* **5**, 457–471 (2022).
38. Xu, S. et al. Editing aberrant splice sites efficiently restores β-globin expression in β-thalassemia. *Blood* **133**, 2255–2262 (2019).
39. Tu, M. et al. A 'new lease of life': FnCpf1 possesses DNA cleavage activity for genome editing in human cells. *Nucleic Acids Res.* **45**, 11295–11304 (2017).
40. Liu, X. et al. Lb2Cas12a and its engineered variants mediate genome editing in human cells. *FASEB J.: Off. Publ. Federation Am. Societies Exp. Biol.* **35**, e21270 (2021).
41. Tran, M. H. et al. A more efficient CRISPR-Cas12a variant derived from Lachnospiraceae bacterium MA2020. *Mol. Ther. Nucleic Acids* **24**, 40–53 (2021).
42. Zhou, J. et al. Cas12a variants designed for lower genome-wide off-target effect through stringent PAM recognition. *Mol. Ther.: J. Am. Soc. Gene Ther.* **30**, 244–255 (2022).
43. Wang, H. et al. Engineering of a compact, high-fidelity EbCas12a variant that can be packaged with its crRNA into an all-in-one AAV vector delivery system. *PLoS Biol.* **22**, e3002619 (2024).
44. Kleinstiver, B. P. 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).
45. Ma, E. et al. Improved genome editing by an engineered CRISPR-Cas12a. *Nucleic Acids Res.* **50**, 12689–12701 (2022).
46. Zhang, L. et al. AsCas12a ultra nuclease facilitates the rapid generation of therapeutic cell medicines. *Nat. Commun.* **12**, 3908 (2021).
47. Guo, L. Y. et al. Multiplexed genome regulation in vivo with hyper-efficient Cas12a. *Nat. Cell Biol.* **24**, 590–600 (2022).
48. 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).
49. Huang, H. et al. Engineered Cas12a-Plus nuclease enables gene editing with enhanced activity and specificity. *BMC Biol.* **20**, 91 (2022).
50. Chen, P. et al. Engineering of Cas12a nuclease variants with enhanced genome-editing specificity. *PLoS Biol.* **22**, e3002514 (2024).
51. Gao, L. et al. Engineered Cpf1 variants with altered PAM specificities. *Nat. Biotechnol.* **35**, 789–792 (2017).
52. Tóth, E. et al. Improved LbCas12a variants with altered PAM specificities further broaden the genome targeting range of Cas12a nucleases. *Nucleic Acids Res.* **48**, 3722–3733 (2020).
53. Liu, X. et al. Engineered FnCas12a with enhanced activity through directional evolution in human cells. *J. Biol. Chem.* **296**, 100394 (2021).
54. Tóth, E. et al. Mb- and FnCpf1 nucleases are active in mammalian cells: activities and PAM preferences of four wild-type Cpf1 nucleases and of their altered PAM specificity variants. *Nucleic Acids Res.* **46**, 10272–10285 (2018).
55. Kim, H. K. et al. In vivo high-throughput profiling of CRISPR-Cpf1 activity. *Nat. Methods* **14**, 153–159 (2017).
56. Kim, N. et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* **38**, 1328–1336 (2020).
57. Kim, H. K. et al. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nat. Biotechnol.* **36**, 239–241 (2018).
58. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

59. Chari, R., Yeo, N. C., Chavez, A. & Church, G. M. sgRNA Scorer 2.0: A Species-Independent Model To Predict CRISPR/Cas9 Activity. *ACS Synth. Biol.* **6**, 902–904 (2017).
60. Kim, D. et al. Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat. Biotechnol.* **34**, 863–868 (2016).
61. Baisya, D., Ramesh, A., Schwartz, C., Lonardi, S. & Wheeldon, I. Genome-wide functional screens enable the prediction of high activity CRISPR-Cas9 and -Cas12a guides in Yarrowia lipolytica. *Nat. Commun.* **13**, 922 (2022).
62. Zhu, H. & Liang, C. CRISPR-DT: designing gRNAs for the CRISPR-Cpf1 system with improved target efficiency and specificity. *Bioinformatics* **35**, 2783–2789 (2019).
63. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
64. Tsai, S. Q., Topkar, V. V., Joung, J. K. & Aryee, M. J. Open-source guideseq software for analysis of GUIDE-seq data. *Nat. Biotechnol.* **34**, 483 (2016).
65. Malinin, N. L. et al. Defining genome-wide CRISPR-Cas genome-editing nuclease activity with GUIDE-seq. *Nat. Protoc.* **16**, 5592–5615 (2021).
66. Seo, S. Y. et al. Massively parallel evaluation and computational prediction of the activities and specificities of 17 small Cas9s. *Nat. Methods* **20**, 999–1009 (2023).
67. Li, J., Liang, Q., Zhou, H., Zhou, M. & Huang, H. Profiling the impact of the promoters on CRISPR-Cas12a system in human cells. *Cell. Mol. Biol. Lett.* **28**, 41 (2023).
68. Kim, H. K. et al. High-throughput analysis of the activities of xCas9, SpCas9-NG and SpCas9 at matched and mismatched target sequences in human cells. *Nat. Biomed. Eng.* **4**, 111–124 (2020).
69. Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
71. Schmid-Burgk, J. L. et al. Highly Parallel Profiling of Cas9 Variant Specificity. *Mol. Cell* **78**, 794–800.e798 (2020).
72. Peng, C. et al. Deep learning models to predict the activity of guide RNAs for Cas12a variants [Source Code]. https://doi.org/10.24433/CO.2898556.v1 (2024).

Acknowledgements

We thank all the members of our laboratory for the fruitful discussions and support. This work was supported by the National Key R&D Program of China (2022YFA1303500, L.Y.), the National Natural Science Foundation of China (32171210, 32371271, L.Y., 32101196, P.C.), the Fundamental Research Funds for the Central Universities (2042022kf1189, L.Y.), the China Postdoctoral Science Foundation (2021TQ0253, 2022M712468 to P.C., 2022M722473, J.Z.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

P.C., Y.K.W., H.J.W. and H.L. contributed equally to this work. L.Y., P.C., conceptualized the study. P.C., designed the experiments and performed the majority of experiments, including the high-throughput evaluation of the activities of the Cas12a variants and GIUDE-seq. Y.K.W. processed the sequencing data. Y.K.W. and J.L.C. assisted with the analysis and developed seq-DeepCpf1variants, enseq-DeepCpf1, enDeepCpf1. P.C., H.L. and H.J.W. established all plasmids constructs. H.L. and H.J.W. assisted with library preparation. J.Z., Z.Q.S., J.L. and C.P assisted with the preparation of genomic DNA. P.C. wrote the manuscript. L.Y. supervised the study. The authors reviewed and approved the final manuscript.

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/s41467-025-57150-9.

Correspondence and requests for materials should be addressed to Lei Yin.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.