Prediction of activity of CRISPR-Cpf1 guide RNA via in vivo high-throughput screening

gancheng wang[1], dan zhu[1], juan li[1], junyi wang[1], and jianzhong xi[1]

[1]Peking University College of Engineering

November 30, 2021

Abstract

Background: CRISPR-cpf1 is a single RNA-guided endonuclease system, becoming a promising tool in both prokaryotic and eukaryotic genome engineering. The editing efficiency of Cpf1 based engineering still requires improvements. However, limited information regarding the relationship between guide RNA sequence and on-target activity is available. To address these challenges, we developed a screening platform based on the association of Acidaminococcus sp. Cpf1(AsCpf1) DNA cleavage with cellular lethality. Major results: In total, we measured the activities of 12,544 guide RNAs, and observed a substantial variation of the editing efficiency depending on the design of the sequence. Based on this large-scale dataset, we designed and implemented a comprehensive computational model to predict activities of guide RNAs. Through comparison using simulated and experimental data, our approach outperformed existing algorithms, enabling selection of efficient guide RNAs. Conclusions: We refine on-target design rules and isolate the important sequence features that contribute to DNA cleavage, that is, AH dimers at position1-8 of protospacer promoting Cas12a activity while TK, GB dimer playing an inhibitory role. We validate guide RNA affinities designed by our optimized rules in both E.coli and 293T cells.

Introduction

The class 2 clustered regularly interspaced short palindromic repeats (CRISPR)-CRISPR-associated proteins (Cas), which are derived from prokaryotic immune system, were identified as programmable, RNA-guided nucleases.[1-7] Generally, each CRISPR-Cas system is composed of Cas proteins and a guide RNA. In a broad spectrum of eukaryotic and prokaryotic species, CRISPR/Cas9 and CRISPR/Cpf1 could be expressed heterologously with relative guide RNAs to target complementary DNA sequences, exhibiting many advantages as powerful genome editing tools.[8-11] Cpf1 was reported with several differences from Cas9: first, Cpf1 processes its own guide RNAs and does not require a tracrRNA; second, there is a longer distance between the seed sequence and cleavage site; third, Cpf1 recognizes thymidine-rich PAM sequence; fourth, Cpf1 generates cleavage with 5'overhangs.[12,13,14] These features make Cpf1 expand the toolkit for genome editing.[15,16,17]

A general issue for the application of Cpf1 appears to be the unpredictable success of guide RNA design.[18,19] However, limited information of the relationship between guide RNAs sequence and activity is available. There is a number of tools and applications developed to predict guide RNA performance of Cas9.[20-28] It may seem that the guide RNA design for Cpf1 would benefit from these information and strategies. Recent studies for Cpf1 attempted to describe the guide RNA sequence-activity relationship and present algorithms to predict the activity of Cpf1 guide RNAs.[20-22]

Nevertheless, such approaches were developed in mammalian cell lines where Cpf1 activities at endogenous sites were found to be affected by chromatin accessibility as well as target sequence composition. And the known nonhomologous end-joining (NHEJ) pathway preference for different DSB substrates may also reshape the guide RNA activity landscape

To exclude these factors and gain more general insights into the relationship of guide RNA sequence and activity, we launched high-throughput screening experiments and collect large-scale datasets in *E.coli* cells, in which NHEJ molecular machinery is entirely absent.

In this paper, we described a library of >12,500 target sequence and guide RNA pairs and evaluated guide RNA activity in*E. coli* by associating CRISPR/Cpf1-induced DNA cleavage with cellular lethality. The guide RNA activity revealed significant diversity. It's worth noting that the current guide RNA activity prediction models showed Spearman correlations of only 0.56 when tested with our data. We therefore proposed a computational approach to design Cpf1 guide RNAs allowing the prediction of efficient and inefficient guide RNAs with an improved performance with Spearman correlation of 0.80. Lastly, our model identified important guide RNA sequence features that contribute to DNA cleavage.

Materials and Methods

Strains and media

Escherichia coli Trans1-T1 (Transgene) was used as host for cloning, plasmid propagation and library construction. *Escherichia coli* BL21(DE3) was used as host for library screening and validation. Bacteria were generally cultured on Luria–Bertani (LB) liquid medium (10 g/L Tryptone, 5 g/L Yeast extract, 10 g/L NaCl, PH adjusted to 7.0) , incubated at 37 °C.

When required, media were supplemented with chemical substrates, and the working concentrations of which were as follows: ampicillin (100 mg/l), chloramphenicol (35 mg/l), anhydrotetracycline (30 mg/l), L-arabinose(10 mM).

Oligo pool design and library

We designed each the oligonucleotide to contain a guide RNA encoding sequence and its target sequence flanked by the correct PAM(TTTV) , in a total length of 140 bases(Fig1.a) 12,044 guide RNA sequences were generated targeting human genes, and 500 were generated targeting*E. coli* genome.

Each oligonucleotide included two constant 20-base sequences at either end for PCR amplification; A unique 8-base barcode sequence was inserted in each oligonucleotide.

Oligonucleotides were synthesized electrochemically on arrays (Custom Array, GenScript). These oligonucleotides (140 bases each) were amplified by PCR using Q5 Polymerase (NEB) and gel-purified by using a QIAquick Purification Kit (QIAGEN). Purified PCR products were assembled with the PUC19 vector using a NEBuidler HiFi DNA assembly kit (NEB). Assembly reaction products were transformed into Trans1-T1 competent cells (Transgene). Transformed cells were cultured onto Luria–Bertani (LB) agar plate with 100 μg/ml ampicillin. The resulting number of colonies yielded more than 60× library coverage. The total plasmids were extracted with a Plasmid Maxiprep Kit (Biomed).

Library sequences are given in TABLE.(excel)

Electrocompetent cell preparation and transformation

Electrocompetent cells are prepared with the following protocol: (1) inoculate a single colony in a 10 mL tube for overnight growth; (2) transfer 1% V/V strains into a 500 mL shake flask with 100 mL LB liquid medium; (3) when optical density at 600 nm (OD600) of the bacterial culture reaches 0.4-0.6, place it on ice and chill for 30 min; (4) spin in centrifuge at 4000 rpm for 10 min, discard the supernatant, and add 40 ml ddH20 to wash the cell pellet; (5) spin in centrifuge at 4000 rpm for 10 min, discard the supernatant, add 40 ml 10% V/V glycerol to wash the pellet, and place it on ice for 10 min; (6) repeat step (5) with 20 ml then 10 ml 10% V/V glycerol, successively; (7) spin in centrifuge at 6000 rpm for 10 min, discard the supernatant, and add 2 ml 10% V/V glycerol to resuspend the pellet; (8) distribute the final dilutions (100 μl each) into the sterile 1.5 ml tubes, freeze it immediately with liquid nitrogen and store at - 80 °C.

About 250ng of plasmid was pipetted into 100 μl thawed electrocompetent cells, and transferred to a pre-cooled 1 mm cuvette (Bio-Rad micro pulser, 2.5 kV, 100 Ω, 25 μF). After pulse, we added 600 μL pre-warmed

LB liquid medium (containing 30 μg/ml anhydrotetracycline or 10mM L-arabinose if necessary) and recovered it in tube at 37 °C for 1.5 hours. Finally, we spread this culture onto LB agar plate containing 100 μg/ml ampicillin and incubated it at 37 °C overnight. The transformation efficiency was calculated as the number of colony forming units per 1 μg of plasmid (cfu/μg).

Screening experiments in *E coli*

For the guide RNA activity screening experiment, competent cells were prepared as described above (Electrocompetent cell preparation and transformation). We collected clones per sample to achieve a proper (>60×) coverage over the design library.

Because the DSB lethality caused by Cpf1 cleavage made the direct calculation of transformation efficiency impossible, the respective control group without inducer was used to evaluate the transformation efficiency for Cpf1 transformed with gene-targeting crRNA library.

Two screening experiments (SE1 and SE2) were conducted parallelly. In each experiment, this library was delivered to the competent cells expressing Cpf1 driven by an inducible promoter to evaluated guide RNA activity. After the *E.coli* cells were cultured with/without inducer, the remaining cells were collected as select/control. In SE1, Cpf1 was driven by a strong promoter (tet promoter) while in SE2 by a weak promoter. Through a synthetical consideration of two groups of results, we could eliminate the influence of the known leakage expression of tet promoter.

Next-Generation Sequencing (NGS) and pooled screen analysis

Plasmid DNA was isolated using a Plasmid Maxiprep Kit (Biomed). Integrated guide RNA target pair sequences were PCR-amplified using Q5 polymerase (NEB) for frequency analysis. PCR products were purified using a QIAquick Purification Kit (QIAGEN). For each sample, 1 microgram of purified PCR amplicons were used as a library template. Sequencing libraries were obtained from the replicates using a NEBNext Ultra II RNA Kit (NEB)and purified with SPRI beads(Beckman). The sequencing libraries were prepared following the manufacturer's protocol(Illumina).Library size and purity was verified by Agilent 2100(Agilent) before sequencing on a Nova seq (Illumina) using a Reagent Kit S2 (Novaseq) (2 × 150 bp).

Statistical analysis

The whole collection of NGS data-set were showing in Table1.a.While the control1 and selective1 were referring to data generated from SE1 control group and SE1 experiment group; the control2 and selective2 were referring to data generated from SE2 control group and SE2 experiment group.The reads were the raw sequencing data and the counts were the filtered data.

For a specific crRNA sequence, its abundance (crRNA frequency) in a certain group could be calculated by its divided by the summation of all crRNA counts.(Table1.b)Further, the fold-change of a specific crRNA sequence across a screening experiment (SE1 or SE2), representing the activity of a crRNA, could be calculated the ratio between the frequency in the post selection sample (selective) and the no-selection sample (control). (Table1.b)

T7 endonuclease I assay for genome modification

293T Cells were collected after 48 h post-transfection for genomic DNA extraction. The genomic region flanking the target site of each gene was PCR-amplified, and products were purified using DNA Clean Kit following the manufacturer's protocol. A total of ˜ 200 ng purified PCR amplicons was mixed with 1 μl NEBuffer 2 and diluted in ddH2O to 10 μl, then subjected to a re-annealing process to form a heteroduplex according to the reported procedure [44]. After re-annealing, the products were treated with T7EI following recommending protocol, and 2.5% agarose gels were used for further analysis. Indels were calculated via band intensities based on previously reported method [45].

Results

Design of a library of guide RNA target sequence pairs library and screening experiments

Biotechnologies based on CRISPR/Cpf1 have greatly facilitated the genetic manipulation on model and non-model bacteria for higher editing efficiency and specificity.[29-32]However, Cpf1 editing in mammalian cells were still challenging.[34-35]We ascribed it to the following factors. (i) NHEJ pathway is absent in most bacteria, such as*E.coli,* once a single-copy gene was destroyed by Cpf1-induced DNA cleavage, it can lead to cell death for lack of redundancy.(ii) The binding of Cpf1-crRNA complex to target gene can interfere with gene expression. Consequently, we intended to avoid these effects by choosing sequences non-existing in *E.coli* genome when designing guide RNAs and using multi-copy plasmids instead of genome sites as targets.

A pool of 12,544 oligonucleotides containing the target sequence and the corresponding guide RNA sequence was synthesized by a DNA microarray, amplified by PCR, and cloned into a guide RNA expression plasmid backbone using Gibson assembly (Fig.1a,b).Among this library, 12,044 pairs were designed for human genes and 500 pairs with targets across*E. coli* genome as internal controls.

Next generation sequencing (NGS) data showed that 12.434(99.12%) pairs were included in the bacteria library among the 12,544 designed pairs (Table1.a). An activity score for each member of the library was calculated by quantifying the representation of each sequence with and without Cpf1 expression (Table1.b). The activity of a guide RNA may be related to cellular survival for the successful DNA cleavage leads to the loss of resistance to antibiotics that a more positive score indicates lower activity.

The screening experiments and data processing were showing in Methods and Materials.

The diversity of activity among guide RNAs

The guide RNA activity distributions of negative control(selective1) and screening experiment(selective2) were shown (Fig.1c). As there is almost no selection pressure of DNA cleavage, the guide RNA activity in the SE1 nearly obeyed a normal distribution with mean approximately equal to 1 as expected, while in SE2 there was significant variability of guide RNA activities. Concretely, the majority of guide RNAs in the designed library exhibited high activities that the best of which resulted in 1000-fold slowdown in growth rate of E. coli cells. To validate such results from the screening experiment, we chose a series of guide RNAs consisting of 9 with high scores and 2 with low scores to retest their activities individually by measuring colony number after transformation, and the results were observed in a very good consistency with which in the screening experiment (Fig.1d). Thus, it suggested that our screening method to evaluate guide RNA activity was reliable and such activity dataset could be used for further analysis.

A computational method predicting guide RNA activities

According to our results above, quite a fraction of guide RNAs showed moderate or no activity that about 50% of guide RNAs in the library got activity scores > 0.1 (lgscore>-1), which indicates guide RNAs designed without a rational method could not be successfully used in genome editing. It is time and labor consuming to test each guide RNA before a gene-editing experiment, thus an in silico method for guide RNA efficacy prediction is in need, and recently several works have been developed to facilitate it.[ ]As we firstly created a dataset of Cpf1 guide RNA activity in prokaryotes, a novel predicting method based on our results could be a supplement to current studies and test the generalization ability of those methods based on eukaryotic datasets.

We filtered our results by removing data of low quality, and established a dataset based on selective2. We randomly separated the dataset (90% for training and 10% for testing) in order to avoid overfitting, and used 10-fold cross-validation to retest the capacity of the trained model (Fig.2a). Then, we defined a series of featurization considering the DNA sequence of the protospacer, PAM and fraction of base that convert each sequence in our library into more than 350 binary and continuous feature information as inputs. Using features extracted from each guide RNA, we built a regression model to predict its activity. (Fig.2b) We re-separated the dataset after randomly shuffling to do the training 10 times in total, and there was no significant difference at the performance of all trained models, suggesting that our model is robust and no bias have been introduced by separating the dataset (Fig.2c Fig2d).

We compared our predictive results with deepCpf1, the most-cited work, to evaluate the performance of our

model. [36] We found weak correlation between experiment results and the predictions from deepCpf1, while our model is much more predictive with Spearman correlation coefficient of 0.80 on average(Fig.2c Fig2d). It indicated that the models trained with data from mammalian cells provided limited comprehension of the guide RNA sequence features contributing to cleavage activity on the Impact of chromatin structures and the NHEJ repair pathway.

To verify our founding and explore the mechanism of different predictive power between two models, we further tested their performance on forecasting the most efficient guide RNAs as well as the inefficient ones, since the selection of guide RNA for efficient genome editing is in critical demand in research and clinical. After we have processed the data of every sequence in our library using our model and deepCpf1, each sequence got a prediction score corresponding to its experimental activity. Then, we accessed the ability of each model to distinguish efficient guide RNAs from inefficient ones individually, by comparing the experimentally measured activity scores of a group of sequence in high prediction scores with which in low scores (Fig2.e). According to the scores predicted by our model, there was significant difference of measured activities between predictive high-score group and low-score group, while no significant difference of which by deepCpf1 in contrary, confirming that our model has a better predictive capacity Furthermore, we investigated the reason that the improvement of our model may be attributed to. This time we divided the test library into high-score group and low-score group according to their experimental activity and compared those prediction scores processed by each model (Fig2.f) As expected, our model performed much better than deeCpf1 in both efficient and inefficient guide RNA predictions. Interestingly, the deepCpf1 model was proven to be of almost no ability to predict efficient guide RNAs, but of weak ability to predict inefficient ones accurately. It revealed the disability in characterization of high-activity guide RNA as a primary reason why deepCpf1 underperformed in prokaryotic. We tried to make an explanation for why deepCpf1 is not sensitive when working with data from more efficient guide RNA, later. Overall, our model makes predictions of guide RNA activity better than current approaches and is extremely good at predicting efficient ones, at least in prokaryotic where developed.

We next investigated the sequence features contributing to guide RNA activity. We mainly focus on the sequence composition of protospacer besides some other factors reported including GC content, melting temperature. A linear model was used here to plot the coefficients of position-dependent dimers and trimers respectively(Fig.3a,b,c). Results of T7 endonuclease I (T7EI) assay showed that guide RNA sequence features could affect genome editing in both human cells (Fig. 3d) . It is notable that approximately equal effect of dimer/trimer each position in seed region of protospacer was observed considering their distance from the PAM, while the first single nucleotide was used to be known as a stronger factor. We also observed the promotional effect of AH dimers, AHN trimers and the inhibitory role of GB, TK dimers, GBN trimers at certain positions. Our findings were in consistence with the results of another activity profiling screening independent, although we obtained a larger scale library and provided a more comprehensive and convincing model.[33] These effects may be attributed to the expression level or stability of guide RNA as well as the interaction of Cpf1-crRNA complex with its DNA substrate.

Discussion

CRISPR-Cpf1-based genome editing technology was invented as another promising tool following the success of CRISPR-Cas9 and have been applied for a broad spectrum of species including eukaryotic and prokaryotic organisms. Cpf1 has several advantages over Cas9 such as guided by a single and shorter RNA, requires a T-rich PAM and exhibits better specificity, making it a complementary and alternative gene editing tool.[37] However, predicting the activity of guide RNA is a challenge due to the lack of the knowledge.

In this study, we made a large-scale library containing 12544 designed guide RNAs .Take advantage of the library, we carry out screening experiments. and presented an efficient and comprehensive computational model for prediction of Cpf1 guide RNA activity trained by the results of which. Consequently, we systematically studied the guide RNA activity in bacteria for the first time, broadening the application of Cpf1 in genome editing. We observed various activities of guide RNAs and tested the results by independent validation experiments. Our trained model was developed to select optimized guide RNAs better-performing

than previous approaches and analyzed the features contributed to guide RNA activities. We propose that deepCpf1 and other models trained in mammalian cells shows a higher false negative rate processing data from in prokaryotic organisms on account of misattributing influences of NHEJ repairs, chromatin structure, or lentivirus transduction to guide RNA incapacity, hence our model may be applied to other bacteria or unexplored host, potentially. Furthermore, our unbiased screening methods may be applied in other developing CRISPR system utilization scenario

However, there may be still several defects in our model that the mechanism of Cpf1-RNA complex interacting with DNA target is unclear and the generalization capacities are unproven. To further improve the accuracy and efficiency of guide RNA design, we expect that more datasets from well-designed experiments across species could be generated, and more comprehensive algorithms such as machine and deep learning (MDL) methods could be applied.[38-40]In addition, improvement of the specificity of guide RNA to reduce off-target effects is a clinical demand, thus our platform combining high-throughput screening and analysis of computational methods may be used to address this issue.

Currently, novel cpf1 orthologs and several other CRISPR nucleases were harnessed for genome editing in mammalian and bacteria cells, while relationship between guide RNA sequence and on-target activity was underexplored. [41]Moreover, several Cpf1 variants of AsCpf1 were described with increased activity and expanded protospacer-adjacent motif (PAM) preferences.[42,43]The method described in this paper could deal with such issues and make progress towards optimizing the design of guide RNAs.

Acknowledgments

We thank Luo yufeng for suggestions on the experiments.

Data Availability Statement: The data sets used in this study are provided as supplementary table 'guide RNA activities data-1' and 'guide RNA activities data-2'. All custom scripts used are provided as a .txt file'Code'. DeepCpf1 codes are available on GitHub (at https://github.com/MyungjaeSong/Paired-Library).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research was supported by the National Natural Science Foundation of China (81827809 to Jianzhong Xi, SQ2018YFA010021 to Jianzhong Xi).

References

1. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science 2014 46:1258096.

2. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell 2015 163, 759–771.

3. Kim,H. and Kim,J.S. (2014) A guide to genome engineering with programmable nucleases. Nat. Rev. Genet., 15, 321¨C334.

4. Sander,J.D. and Joung,J.K. (2014) CRISPR¨CCas systems for editing, regulating and targeting genomes. Nat. Biotechnol., 32, 347¨C350.

5. Cox,D.B.T., Platt,R.J. and Zhang,F. (2015) Therapeutic genome editing: prospects and challenges. Nat. Med., 21, 121¨C131.

6 Mohanraju,P., Makarova,K.S., Zetsche,B., Zhang,F., Koonin,E. V. and Van Der Oost,J. (2016) Diverse evolutionary roots and mechanistic variations of the CRISPR¨CCas systems. Science, 353,

556¨C568.

7 Hart,T., Chandrashekhar,M., Aregger,M., Steinhart,Z., Brown,K.R., MacLeod,G., Mis,M., Zimmermann,M., Fradet-Turcotte,A., Sun,S. et al. (2015) High-Resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell, 163, 1515¨C1526.

8 Shalem,O., Sanjana,N.E. and Zhang,F. (2015) High-throughputfunctional genomics using CRISPR¨CCas9. Nat. Rev. Genet., 16,299¨C311.

9. Wang,W., Ye,C., Liu,J., Zhang,D., Kimata,J.T. and Zhou,P. (2014)CCR5 gene disruption via lentiviral vectors expressing Cas9 andsingle guided RNA renders cells resistant to HIV-1 infection. PLoS

One, 9, e115987.

10. Zhou,H., Liu,B., Weeks,D.P., Spalding,M.H. and Yang,B. (2014)Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice. Nucleic Acids Res., 42,

10903¨C10914.

11. Wu,W.Y., Lebbink,J.H.G., Kanaar,R., Geijsen,N. and Van Der Oost,J. (2018) Genome editing by natural and engineered CRISPR-associated nucleases. Nat. Chem. Biol., 14, 642¨C651.

12 Nakade S, Yamamoto T, Sakuma T. Cas9, Cpf1 and C2c1/2/3-What's next ? Bioengineered. 2017;8(3):265–273.

13. Tang X, Liu G, Zhou J, Ren Q, You Q, Tian L, Xin X, Zhong Z, Liu B, Zheng X, Zhang D, Malzahn A, Gong Z, Qi Y, Zhang T, Zhang Y. A large-scale whole-genome sequencing analysis reveals highly specific genome editing by both Cas9 and Cpf1 (Cas12a) nucleases in rice. Genome Biol. 2018 Jul 4;19(1):84.

14. Lee K, Zhang Y, Kleinstiver BP, Guo JA, Aryee MJ, Miller J, Malzahn A, Zarecor S, Lawrence-Dill CJ, Joung JK, Qi Y, Wang K. Activities and specificities of CRISPR/Cas9 and Cas12a nucleases for targeted mutagenesis in maize. Plant Biotechnol J. 2019 Feb;17(2):362-372.

15. Safari F, Zare K, Negahdaripour M, Barekati-Mowahed M, Ghasemi Y. CRISPR Cpf1 proteins: structure, function and implications for genome editing. Cell Biosci. 2019;9:36

16 Bayat H, Modarressi MH, Rahimpour A. The Conspicuity of CRISPR-Cpf1 System as a Significant Breakthrough in Genome Editing. Curr Microbiol. 2018 Jan;75(1):107-115.

17 Ding D, Chen K, Chen Y, Li H, Xie K. Engineering Introns to Express RNA Guides for Cas9- and Cpf1-Mediated Multiplex Genome Editing. Mol Plant. 2018 Apr 2;11(4):542-552.

18 Creutzburg SCA, Wu WY, Mohanraju P, Swartjes T, Alkan F, Gorodkin J, Staals RHJ, van der Oost J. Good guide, bad guide: spacer sequence-dependent cleavage efficiency of Cas12a. Nucleic Acids Res. 2020 Apr 6;48(6):3228-3243.

19 Kim H, Lee WJ, Oh Y, Kang SH, Hur JK, Lee H, Song W, Lim KS, Park YH, Song BS, Jin YB, Jun BH, Jung C, Lee DS, Kim SU, Lee SH. Enhancement of target specificity of CRISPR-Cas12a by using a chimeric DNA-RNA guide. Nucleic Acids Res. 2020 Sep 4;48(15):8601-8616.

20. Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,I., Tothova,Z., Wilen,C., Orchard,R. et al.(2016) Optimized sgRNA design to maximize activity and minimize

off-target effects of CRISPR–Cas9. Nat. Biotechnol., 34, 184–191.

21. Wang,T., Wei,J.J., Sabatini,D.M. and Lander,E.S. (2014) Genetic screens in human cells using the CRISPR–Cas9 system. Science(80-.)., 343, 80–84.

22. Doench,J.G., Hartenian,E., Graham,D.B., Tothova,Z., Hegde,M., Smith,I., Sullender,M., Ebert,B.L., Xavier,R.J. and Root,D.E. (2014)Rational design of highly active sgRNAs for CRISPR–Cas9-mediatedgene inactivation. Nat. Biotechnol., 32, 1262–1267.

23. Ren,X., Yang,Z., Xu,J., Sun,J., Mao,D., Hu,Y., Yang,S.J., Qiao,H.H., Wang,X., Hu,Q. et al. (2014) Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila. Cell Rep., 9, 1151–1162.

24. Malina,A., Katigbak,A., Cencic,R., Ma¨ýga,R.I., Robert,F., Miura,H. and Pelletier,J. (2014) Adapting CRISPR/Cas9 for functional genomics screens. Methods Enzymol., 546, 193–213.

25. Moreno-Mateos,M.A., Vejnar,C.E., Beaudoin,J.D., Fernandez,J.P.,Mis,E.K., Khokha,M.K. and Giraldez,A.J. (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR–Cas9 targeting in

vivo. Nat. Methods, 12, 982–988.

26. Xu,H., Xiao,T., Chen,C.H., Li,W., Meyer,C.A., Wu,Q., Wu,D., Cong,L., Zhang,F., Liu,J.S. et al. (2015) Sequence determinants of improved CRISPR sgRNA design. Genome Res., 25, 1147–1157.

27. Wong,N., Liu,W. and Wang,X. (2015) WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. Genome Biol., 16, 218.

28. Guo J, Wang T, Guan C, Liu B, Luo C, Xie Z, Zhang C, Xing XH. Improved sgRNA design in bacteria via genome-wide activity profiling. Nucleic Acids Res. 2018 Aug 21;46(14):7052-7069.

29 .Yao R, Liu D, Jia X, Zheng Y, Liu W, Xiao Y. CRISPR-Cas9/Cas12a biotechnology and application in bacteria. Synth Syst Biotechnol. 2018 Oct 3;3(3):135-149.

30 Ungerer J., Pakrasi H.B. Cpf1 is a versatile tool for CRISPR genome editing across diverse species of cyanobacteria. Sci Rep. 2016;6:39681.

31 Zhang X., Wang J., Cheng Q., Zheng X., Zhao G., Wang J. Multiplex gene regulation by CRISPR-ddCpf1. Cell Discov. 2017;3:17018

32 Yan M.Y., Yan H.Q., Ren G.X., Zhao J.P., Guo X.P., Sun Y.C. CRISPR-Cas12a-Assisted recombineering in bacteria. Appl Environ Microbiol. 2017;83

33 Kim,H.K., Song,M., Lee,J., Menon,A.V., Jung,S., Kang,Y.M.,Choi,J.W., Woo,E., Koh,H.C., Nam,J.W. et al. (2017) In vivo high-throughput profiling of CRISPR-Cpf1 activity. Nat. Methods,14, 153¨C159.

Kleinstiver BP, Sousa AA, Walton RT, Tak YE, Hsu JY, Clement K, Welch MM, Horng JE, Malagon-Lopez J, Scarfò I, Maus MV, Pinello L, Aryee MJ, Joung JK. Engineered CRISPR-Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. Nat Biotechnol. 2019 Mar;37(3):276-282.

Liu P, Luk K, Shin M, Idrizi F, Kwok S, Roscoe B, Mintzer E, Suresh S, Morrison K, Frazão JB, Bolukbasi MF, Ponnienselvan K, Luban J, Zhu LJ, Lawson ND, Wolfe SA. Enhanced Cas12a editing in mammalian cells and zebrafish. Nucleic Acids Res. 2019 May 7;47(8):4169-4180.

36. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nat Biotechnol. 2018;36(3):239©\241.

37. Creutzburg SCA, Wu WY, Mohanraju P, Swartjes T, Alkan F, Gorodkin J, Staals RHJ, van der Oost J. Good guide, bad guide: spacer sequence-dependent cleavage efficiency of Cas12a. Nucleic Acids Res. 2020 Apr 6;48(6):3228-3243

38. Luo J, Chen W, Xue L, Tang B. Prediction of activity and specificity of CRISPR-Cpf1 using convolutional deep learning neural networks. BMC Bioinformatics. 2019;20(1):332.

39.Zhang G, Zeng T, Dai Z, Dai X. Prediction of CRISPR/Cas9 single guide RNA cleavage efficiency and specificity by attention-based convolutional neural networks. Comput Struct Biotechnol J. 2021 Mar 7;19:1445-1457.

40.Wang J, Zhang X, Cheng L, Luo Y. An overview and metanalysis of machine and deep learning-based CRISPR gRNA design tools. RNA Biol. 2020 Jan;17(1):13-22.

41.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).

42.DeWeirdt PC, Sanson KR, Sangree AK, et al. Optimization of AsCas12a for combinatorial genetic screens in human cells. Nat Biotechnol. 2021

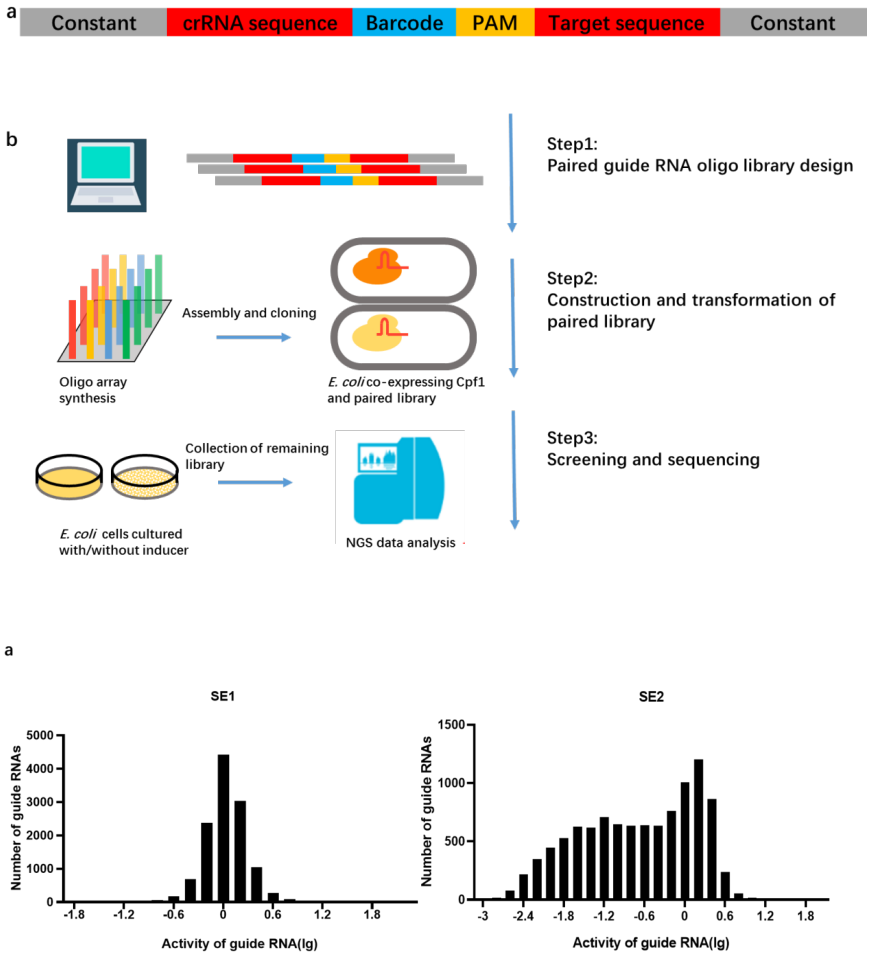
43.Zetsche B, Abudayyeh OO, Gootenberg JS, Scott DA, Zhang F. A Survey of Genome Editing Activity for 16 Cas12a Orthologs. Keio J Med. 2020

44.Li W, Teng F, Li T, Zhou Q. Simultaneous generation and germline transmission of multiple gene mutations in rat using CRISPR-Cas systems. Nat Biotechnol. 2013

45.Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013

Figure and Table

Fig1



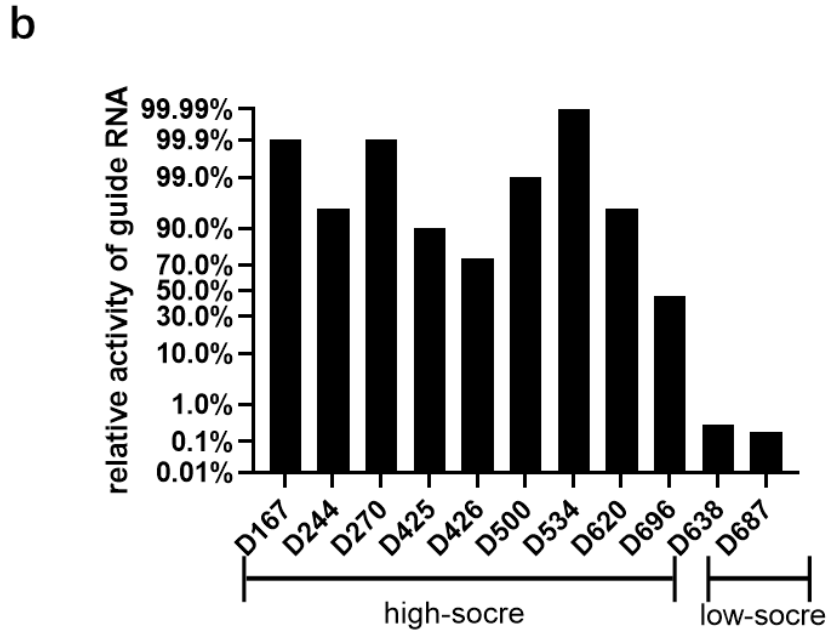


Figure 1| Library preparation and high-throughput evaluation of Cpf1 guide RNA activity.

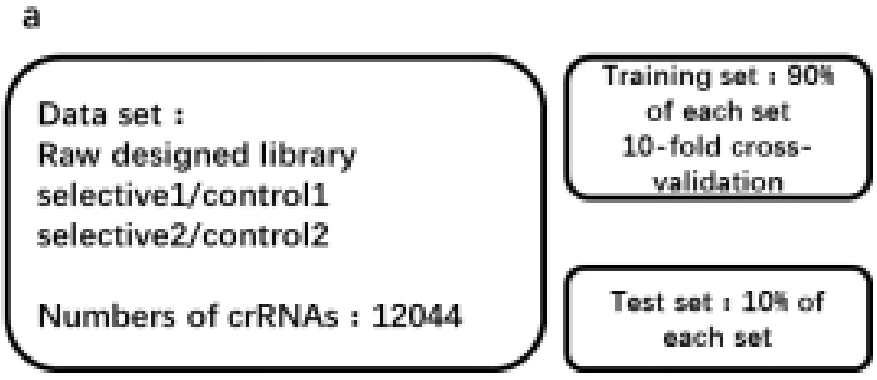
(a) Representation of an oligonucleotide containing pairs of target and guide RNA sequence.

(b) A schematic of the high-throughput evaluation of Cpf1 guide RNA activity.

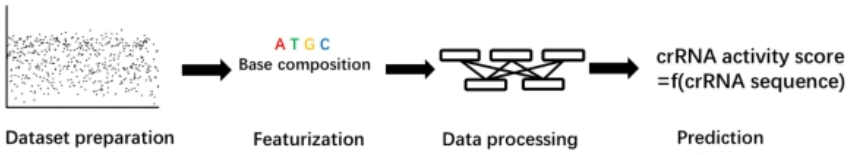
(c) The distribution of guide RNA activity in screening expriment1 and screening expriment2

(d) The pooled screening results were confirmed by cloning 11 guide RNAs (9 of high activity and 2 of low activity) individually and measured via transformation assay.

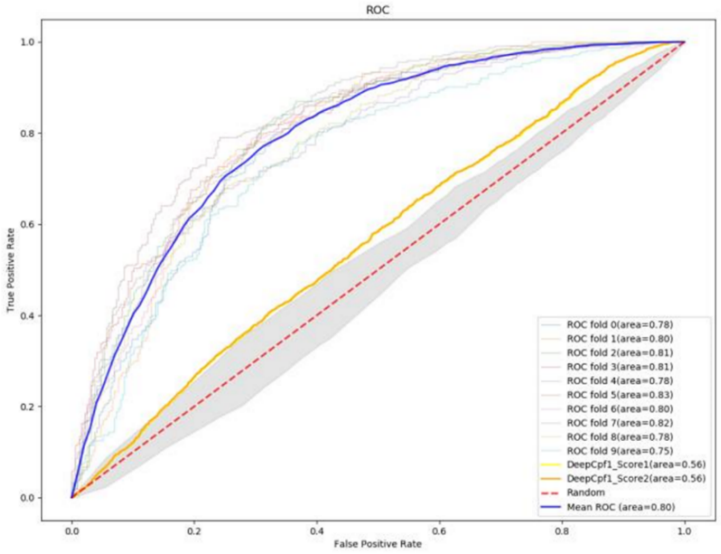
Fig2

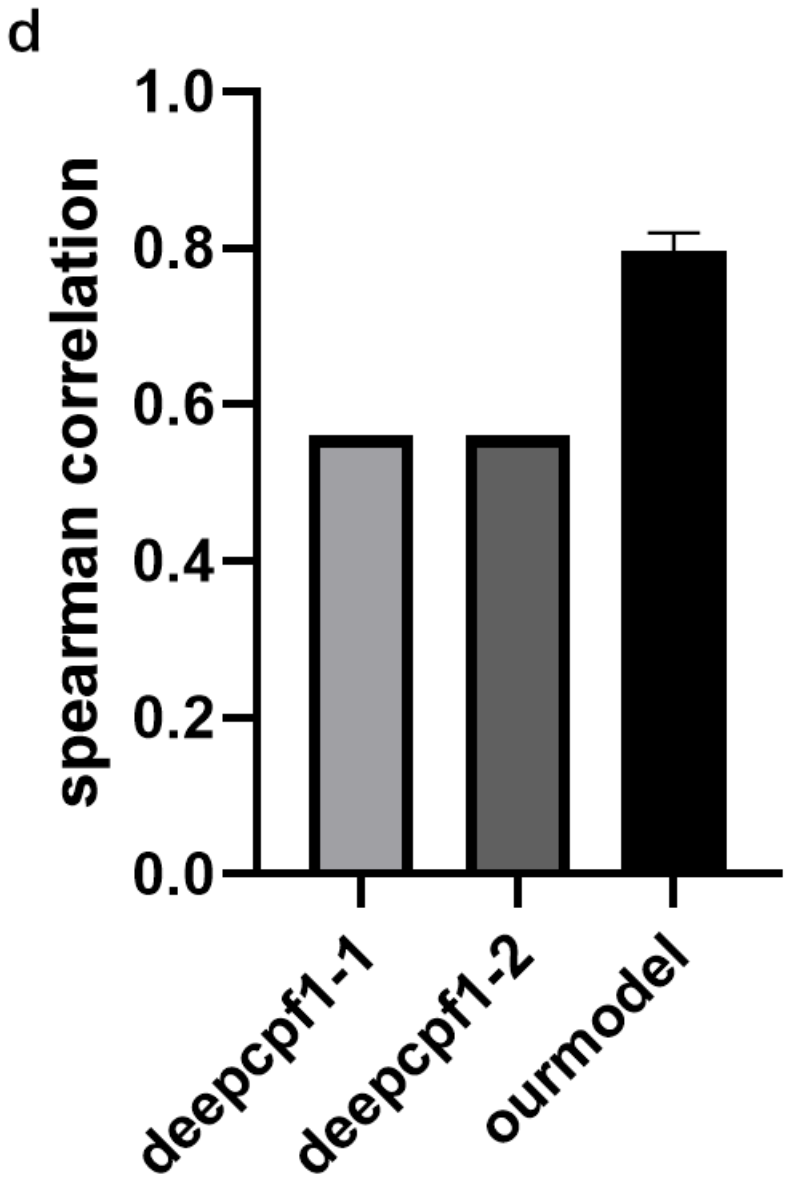


b



c





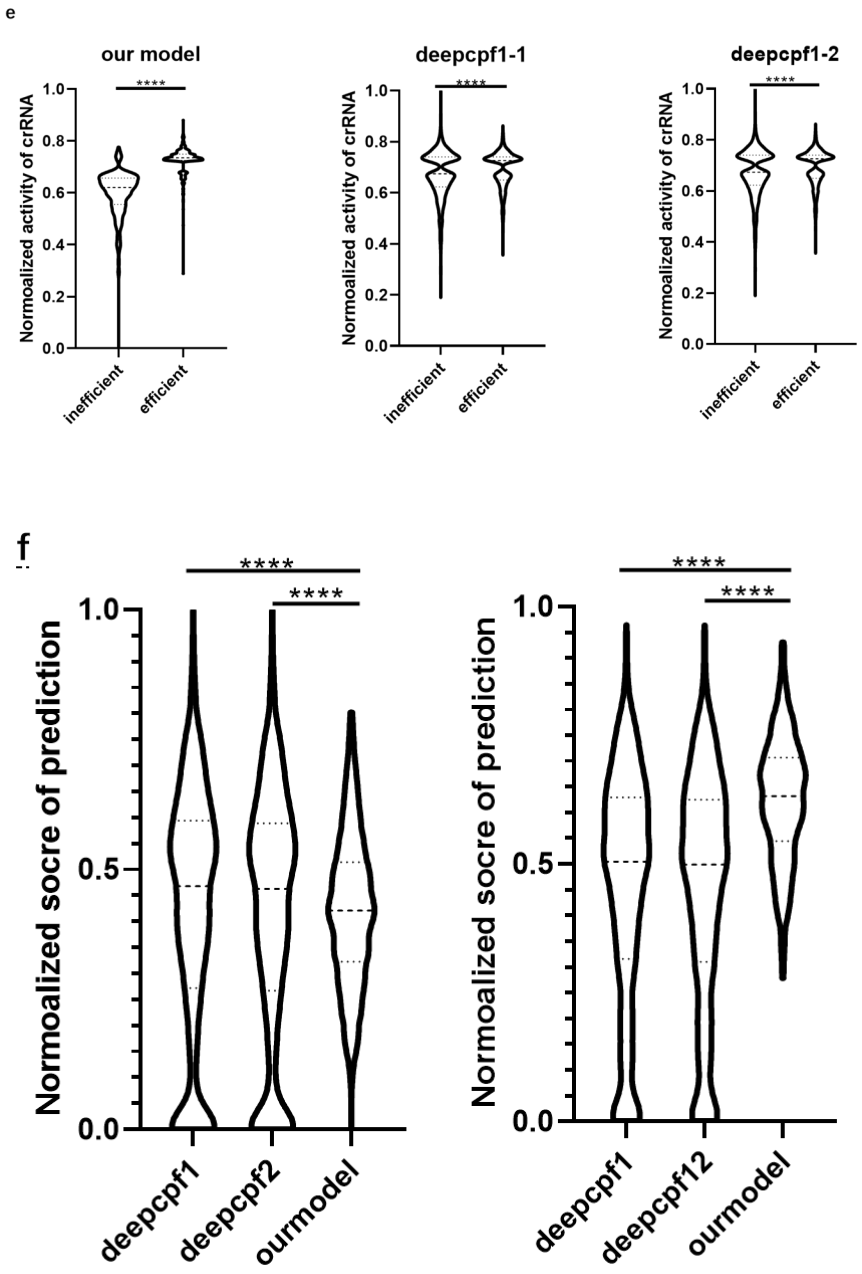


Figure 2| Computational model predicting guide RNA activities

(a) (b) The schematic of datasets and algorithm used

(c) ROC curves comparing the predictive power of our model and deepCpf1.

(d) A quantization of (c) and the bar shows the mean ±s.d. between predicted and measured guide activity scores (n=10)

(e) Comparison of the different models. A smaller scale library of 6,265 guide RNAs was processed by each model, and the top 1,000 scored ones were chosen as an efficient group, in comparison to bottom 1,000 ones

as an inefficient group. Their corresponding experimental activity scores were compared.

(f) The former library was re-chosen into high score and low score group by measured guide RNA activities (1,000 members from 6,265).

Fig3

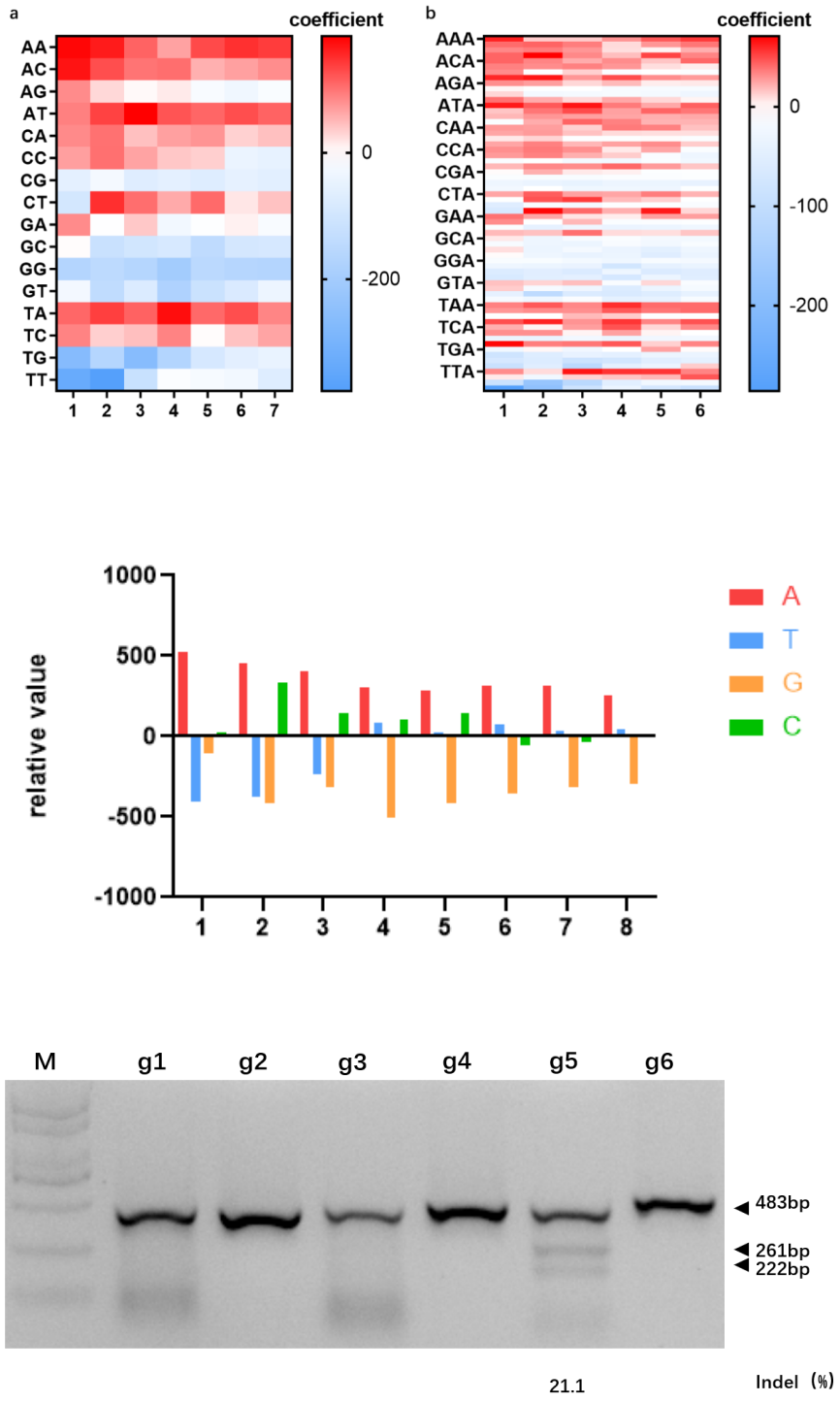


Figure 3 |The top sequence features for predicting guide RNA activities

(a)The features for adjacent dimers of guide RNAs are plotted as heatmaps based on their position in seed region, and positive coefficients suggest positive contributions to guide RNA activity.

(b)The features for adjacent trimers of guide RNAs are plotted as heatmaps based on their position in seed region, and positive coefficients suggest positive contributions to guide RNA activity.

(c) The features for adjacent nucleotides of guide RNAs are plotted as columns based on their position in seed region, and positive coefficients suggest positive contributions to guide RNA activity.

(d) T7EI analysis of targeted indel frequencies induced by the 6 guide RNAs with different sequence features.

FigS1

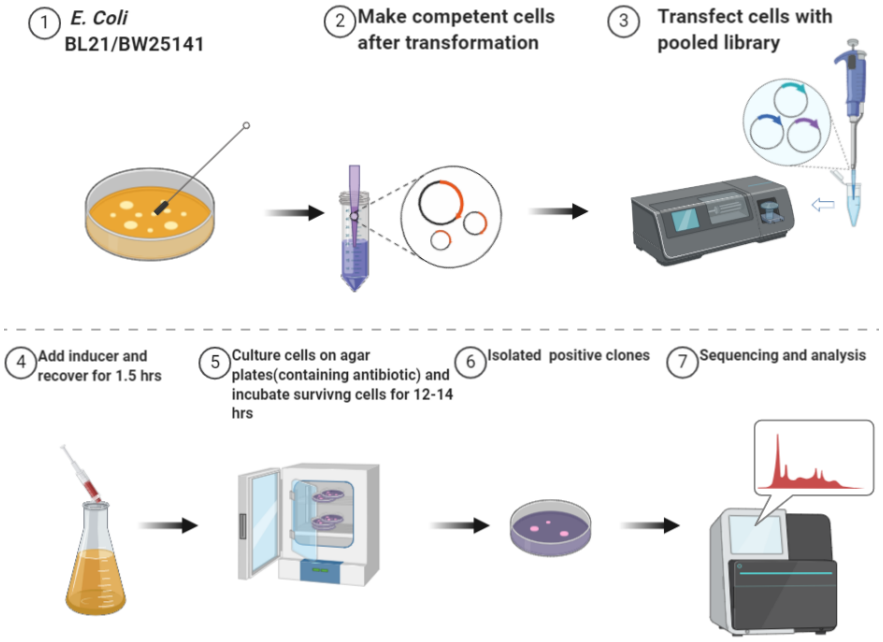


Figure s1 Workflow of the screening experiments in *E.coli*

Table 1| The numbers of pairs in the oligonucleotide pool and NGS data processing.

a

NGS data	Raw designed library	control1	selective1	control2	selective2
reads	7.2×10^7	5.4×10^7	1.3×10^8	6.0×10^7	5.9×10^7
counts	3.1×10^7	2.5×10^7	5.5×10^7	2.6×10^7	1.6×10^7

b

$$crRNA\ frequency = Read\ count_i \,/\, \sum_{i=1}^{n} read\ count_i$$

$$activity_{crRNA} = crRNA\ frequency_{selective} \,/\, crRNA\ frequency_{control}$$

(a) Raw NGS data from designed library and each screening experiment.

(b) For each condition, the activity of each guide RNA was calculated via these equations