bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

BEdeepon: an *in silico* tool for prediction of base editor efficiencies and outcomes

Chengdong Zhang[1+], Daqi Wang[1+], Tao Qi[1+], Yuening Zhang[2], Linghui Hou[1], Feng Lan[3], Jingcheng Yang[1], Leming Shi[1], Sang-Ging Ong[4,5], Hongyan Wang[1], Yongming Wang[1,6]

[1]State Key Laboratory of Genetic Engineering, School of Life Sciences, Zhongshan Hospital, Obstetrics and Gynecology Hospital, Fudan University, Shanghai 200011, China.

[2]SJTU-Yale Joint Center for Biostatistics and Data Science, (Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology) Shanghai Jiao Tong University; Shanghai, China, 200240

[3]Beijing Anzhen Hospital, Beijing Institute of Heart Lung and Blood Vessel Disease, Capital Medical University, Beijing, 100029, China.

[4]Department of Pharmacology and Regenerative Medicine, University of Illinois College of Medicine, USA.

[5]Division of Cardiology, Department of Medicine, University of Illinois College of Medicine, USA.

[6]Shanghai Engineering Research Center of Industrial Microorganisms, Shanghai, China.

[+]Equal contributors

*Correspondence: ymw@fudan.edu.cn

Abstract

Base editors enable direct conversion of one target base into another in a programmable manner, but conversion efficiencies vary dramatically among different targets. Here, we performed a high-throughput gRNA-target library screening to measure conversion efficiencies and outcome product frequencies at integrated genomic targets and obtained datasets of 60,615 and 73,303 targets for ABE and CBE, respectively. We used the datasets to train deep learning models, resulting in ABEdeepon and CBEdeepon which can predict

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

on-target efficiencies and outcome sequence frequencies. The software is freely accessible via online web server http://www.deephf.com/#/bedeep.

Introduction

Base editors are fusion of catalytically impaired Cas9 nuclease and adenosine deaminase (ABE) or cytosine deaminase (CBE) that introduce desired point mutations in the target region enabling precise editing of genomes [1-5]. Base editors have been successfully used in diverse organisms including prokaryotes, plants, fish, frogs, mammals and human embryos [6-9]. Though simple in concept, the success of base editing depends on the choice of the target sequence. First, efficiency of base conversion varies dramatically among different target sequences, and users need to select a target sequence with high efficiency. Second, there are often more than one editable nucleotide in the editing window and unwanted concurrent mutations should be avoided. Experimental evaluation of a target sequence is time-consuming, prompting us to develop in silico tools for target sequence evaluation.

We have previously established a library containing over 80,000 gRNA-target sequence pairs, covering ~20,000 human genes [10]. In this study, we made use of this library to screen both base editors and obtained the outcome product frequencies of 60,615 and 73,303 target sequences for ABE and CBE, respectively. The resulting outcomes were used to train a deep learning model, resulting in ABEdeepon and CBEdeepon which can predict efficiency and outcome frequency distributions of ABE and CBE, respectively. These models were accessible via http://www.deephf.com/#/bedeep, which will greatly facilitate base editor application.

Results

Guide RNA-target pair strategy for test of conversion efficiency and outcomes

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

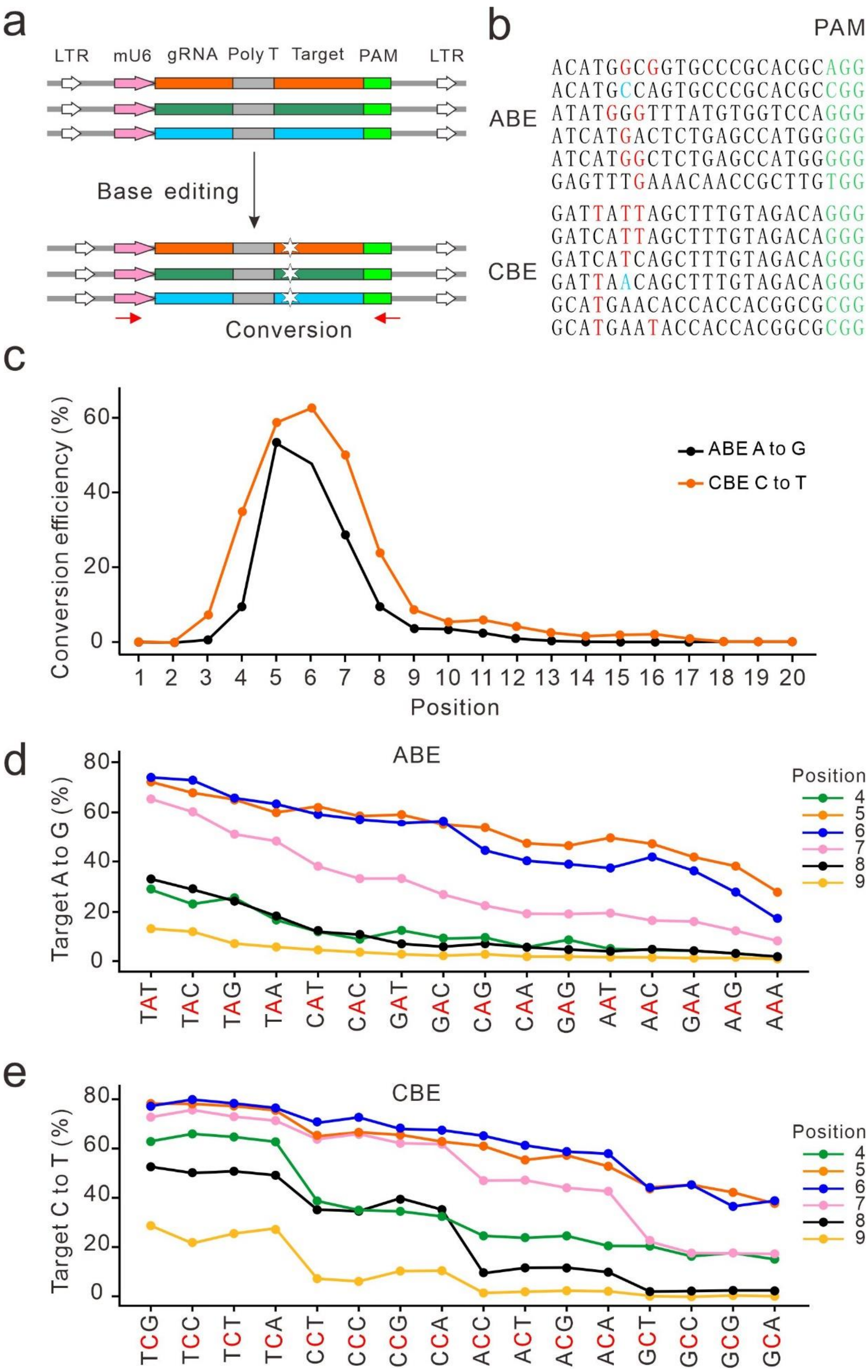


Figure 1. High-throughput screen of on-target efficiencies and

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

outcomes

(**a**) Schematic of the pooled pairwise library screen. A library of 80,263 gRNA-target pairs was packed into viruses and transduced into cells expressing ABE or CBE base editors. The integrated target sites were PCR-amplified for deep-sequencing analysis. Red arrows indicate primers for target site amplification; stars indicate nucleotide conversion. (**b**) Deep-sequencing results revealed that A to G conversion (red letter) for ABE and C to T conversion for CBE occurred. A to non-G and C to non-T conversions are shown in blue. (**c**) C to T or A to G conversion efficiency at different position. (**d, e**) Effect of the sequence context surrounding the target nucleotide (red) on the conversion efficiency at protospacer positions 4 to 9.

A previously generated library [10] containing over 80,000 gRNA-target sequence pairs was used for ABE and CBE screening. Optimized version of ABE (ABEmax) and CBE (AncBE4max)[11] was used in this study. Because of the large size, we used the Sleeping Beauty (SB) transposon [12-14] to integrate each base editor into the genome and generated single cell-derived cell lines. Next, we evaluated the nucleotide conversion efficiency with three targets at different time points. As expected, the efficiency increased over time for all three targets (Supplementary Fig. 1). One target displayed ~100% gene conversion at day 7. We selected day 5 to measure editing efficiency for the library.

We packaged the gRNA-target library into lentiviruses and transduced them into recipient cells. Five days after transduction, genomic DNA was extracted, and synthesized targets were PCR-amplified for deep-sequencing (Fig. 1a). The reads containing canonical base editing (A to G for ABE, C to T for CBE) and unedited reads were used for outcome frequency distribution. A target efficiency can be calculated by 1 subtract unedited outcome frequency (see **Data analysis** section of **Methods** part). The screening assay was experimentally repeated twice, and conversion efficiency in two independent

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

replicates showed high correlation for both ABE and CBE (Spearman correlation, 0.891 for ABE and 0.934 for CBE). Data from the two replicates were combined together for subsequent analysis. We obtained valid efficiencies (reads number>100) of 60,615 targets with 5,761,833 outcomes for ABE and 73,303 targets with 5,513,919 outcomes for CBE.

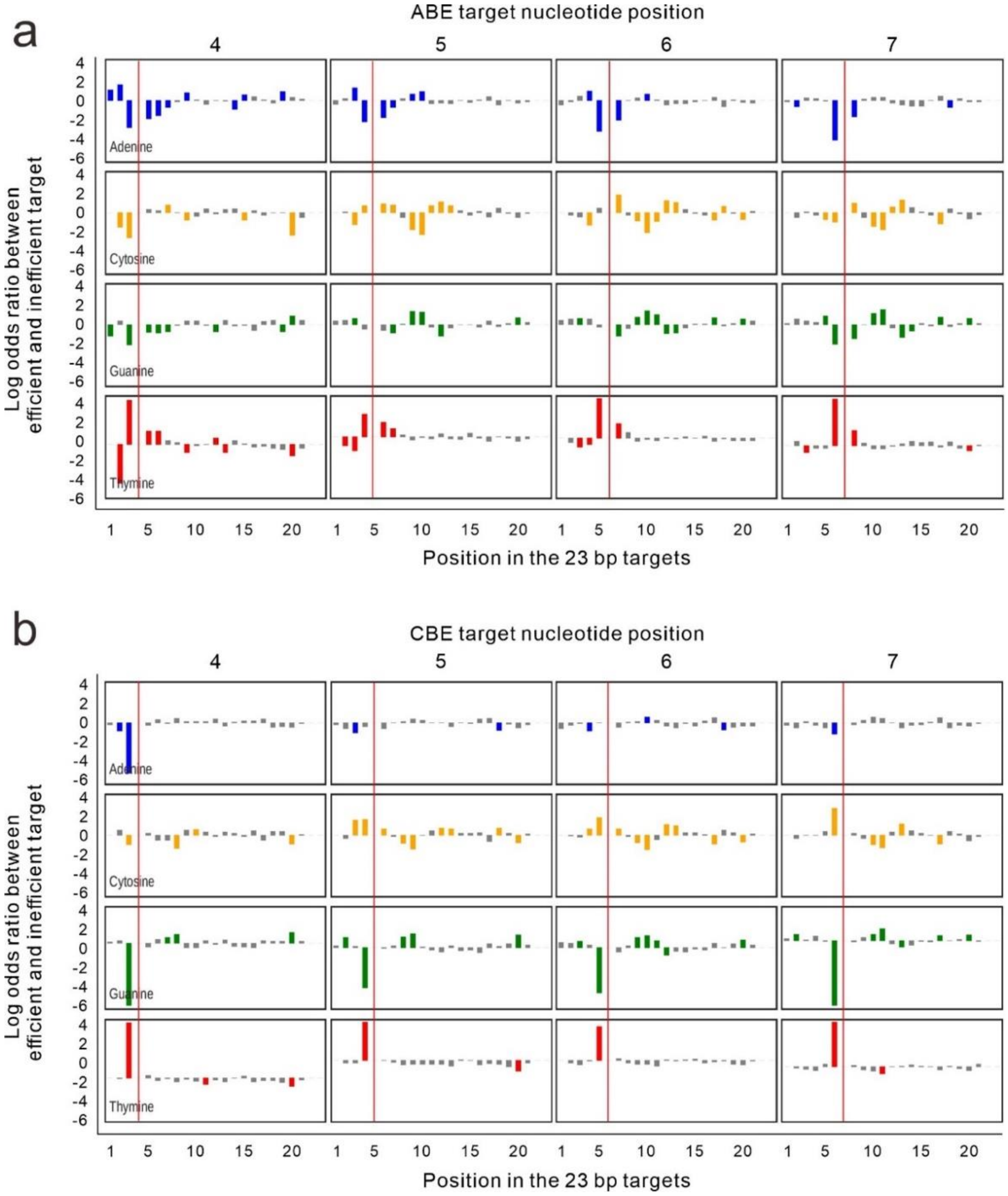


Figure 2. Conversion efficiency influenced by nucleotides on the target sequence.

Sequence preferences at each position in efficient (top 20%) vs. inefficient (bottom 20%) targets for ABE (**a**) and CBE (**b**). The target nucleotide position is shown on the top. Nucleotide position on the targets is shown below. The log odds ratios of nucleotide frequencies between efficient and inefficient target sequences are represented on the y axis. Target nucleotide position is

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

indicated by red lines.

Sequencing results revealed that A to G conversion for ABE and C to T conversion for CBE occurred (Fig. 1b). Although infrequent, C to non-C conversion for CBE and A to non-A conversion for ABE could also be observed (Fig. 1b). The ABE products (purity ranging from 98.43% to 99.69%) were purer than CBE (purity ranging from 95.55% to 97.78%, Supplementary Fig. 2), consistent with a previous report [2]. We also observed indels with mean efficiency of 0.13 for ABE and mean efficiency of 0.16 for CBE, whereas mean frequency of background indels in the plasmid library was 0.06.

Positional effects on nucleotide conversion efficiency

The large-scale dataset generated here allowed us to analyze the positional effects on nucleotide conversion efficiency. The five most efficient positions for nucleotide conversion are 4-8 for both ABE and CBE with PAM at position 21-23 (Fig. 1c), consistent with previously reported editing windows [2, 11]. Next, we investigated the influence of nearby nucleotides on conversion efficiency for position 4-9. Both nucleotides surrounding the target A influenced conversion efficiency with the upstream nucleotide having a stronger influence for ABE. The upstream nucleotide preference followed the order T>C>G>A (Fig. 1d). In contrast, only upstream nucleotide surrounding the target C influenced conversion efficiency for CBE following the preference order T>C>A>G (Fig. 1e).

Next, we investigated nucleotide preferences at each position in the target sequences associated with high editing efficiencies. The results revealed that ABE editing efficiencies were strongly influenced by both upstream and downstream nucleotides immediately adjacent to the target nucleotide (target nucleotide, ±1 bp, Fig. 2). The nearby nucleotides A and G have negative influence on editing efficiency, while the nearby T have positive influence on editing efficiency. The upstream C has positive influence on editing efficiency for target nucleotide positions 5 and 6, but has negative influence on target

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

nucleotide positions 4 and 7. The downstream C has positive influence on editing efficiency. CBE editing efficiencies were only strongly influenced by the upstream nucleotide immediately adjacent to the target nucleotide (target nucleotide, +1 bp, Fig. 2). The upstream nucleotides A and G have negative influence on editing efficiency, while the upstream T have positive influence on editing efficiency. Upstream C has positive influence on editing efficiency for target nucleotide positions 5-7, but has negative influence on target nucleotide position 4.

Correlation of editing efficiency between SpCas9 and base editors

The gRNA-target library in this study was recently used to screen for SpCas9 nuclease [10], allowing us to investigate the efficiency correlation between SpCas9 and base editors. We calculated the correlations between SpCas9 nuclease and base editors for each quartile of base editor efficiencies. The correlation (R<0.15) was low for both ABE and CBE (Supplementary Fig. 3a-b). We also calculated the correlations between SpCas9 nuclease and base editors for each quartile of SpCas9 efficiencies. Although the corrections were low, quartile 1 achieved much higher correlation than other quartiles for both ABE and CBE (R=0.21 for ABE, R=0.23 for CBE, Supplementary Fig. 3c-d).

We further investigated the relationship between SpCas9 mean efficiencies and base editor mean efficiencies for each quartile of base editor efficiencies. The results revealed that SpCas9 efficiency was similar in each quartile of base editor efficiency (Supplementary Fig. 4a-b). We also investigated the relationship between SpCas9 mean efficiencies and base editor mean efficiencies for each quartile of SpCas9 efficiencies. The results revealed that quartile 2-4 achieved similar mean efficiency, while quartile 1 achieved a little lower mean efficiency (Supplementary Fig. 4c-d). Collectively, these data suggested SpCas9 efficiency and base editor efficiency had a very low correlation for a target.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

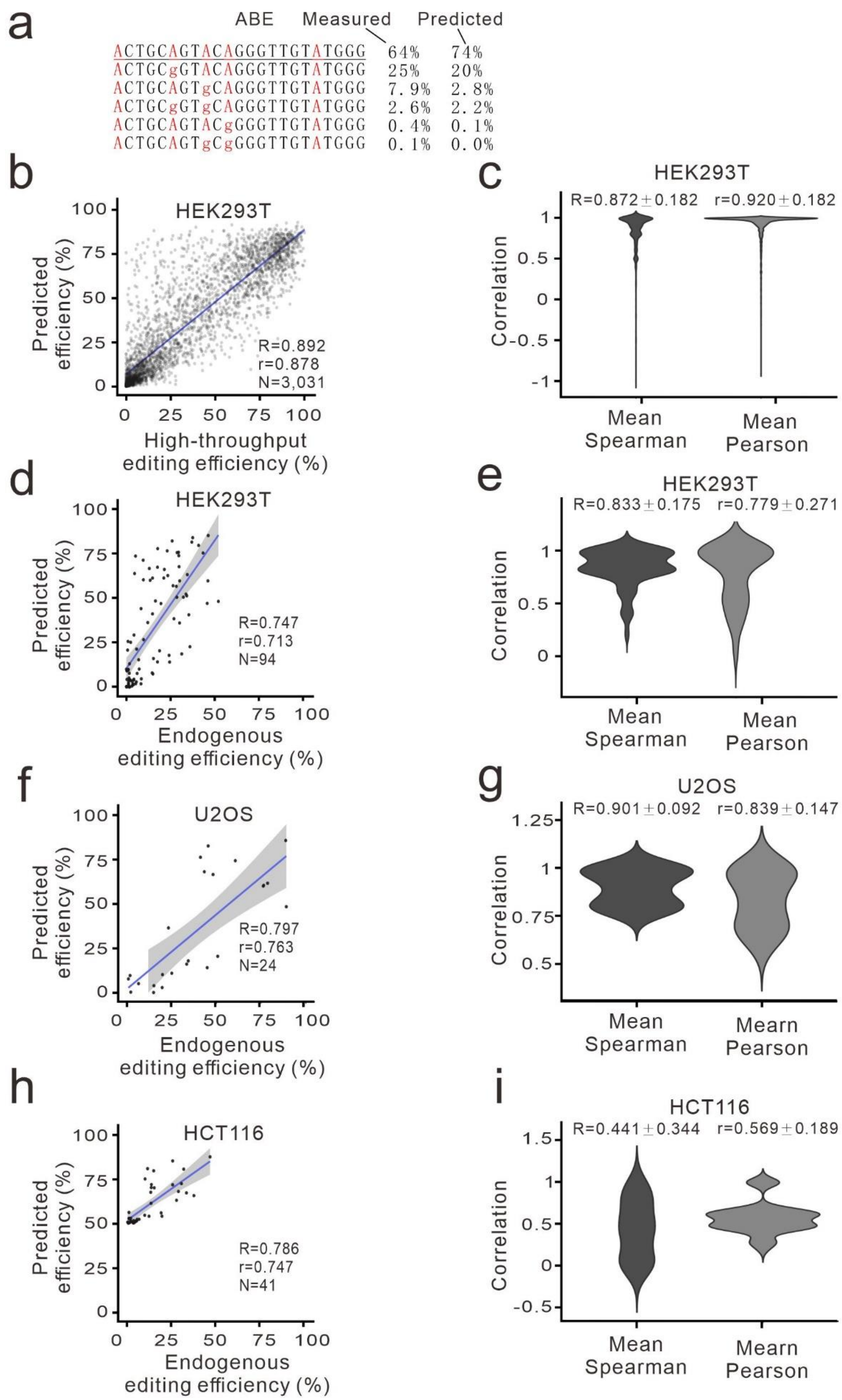


Figure 3. Evaluation of ABEdeepon for editing efficiency and outcome prediction

(**a**) Example of ABEdeepon prediction for a given target. Measured

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

efficiencies are listed for comparison. Original target sequence is underlined. (**b**) Evaluation of ABEdeepon for conversion efficiency with testing dataset. (**c**) Evaluation of ABEdeepon for conversion outcome sequence frequencies with testing dataset. (**d, f, h)** Evaluation of ABEdeepon for conversion efficiency with endogenous editing dataset generated in HEK293T, U2OS and HCT116 cells, respectively. (**e, g, i**) Evaluation of ABEdeepon for conversion outcome sequence frequencies with endogenous editing dataset generated in HEK293T, U2OS and HCT116 cells, respectively.

Developing models for prediction of on-target efficiencies and outcomes

Next, we developed models for prediction of on-target efficiencies and outcome sequence frequencies. Since a target generally contains multiple editable nucleotides, there might exit multiple editing outcome sequences. We first programmed this conversion scheme to obtain all possible editing outcome sequences, and then assigned the measured editing frequency to each outcome sequence. We used the ABE editing datasets to train a shared embedding based deep learning model. The targets were randomly split to 56,432 (5,390,492 outcomes) for training, 1,152 (98,958 outcomes) for internal validation, and 3,031 (272,383 outcomes) held out for testing. The resulting model was named “ABEdeepon” which can predict editing efficiencies and outcome sequence frequencies (Fig. 3a). ABEdeepon achieved a Spearman correlation of 0.892 (r=0.878, MSE=0.024) for prediction of editing efficiency with testing dataset (Fig. 3b). To evaluate the performance for prediction of outcome frequency distribution, we tested 2,342 targets with at least 4 outcomes and achieved mean Spearman correlation of 0.872 (r=0.920, MSE=0.011, KL divergence=0.003, Fig. 3c). To test whether ABEdeepon works in other cell types, we used the same library to generate datasets in HeLa cells and obtained 58,445 valid targets with 6,292,774 outcomes (reads number≥100). We extracted 2,636 targets present in HeLa testing dataset as external testing dataset. The editing efficiency prediction achieved Spearman

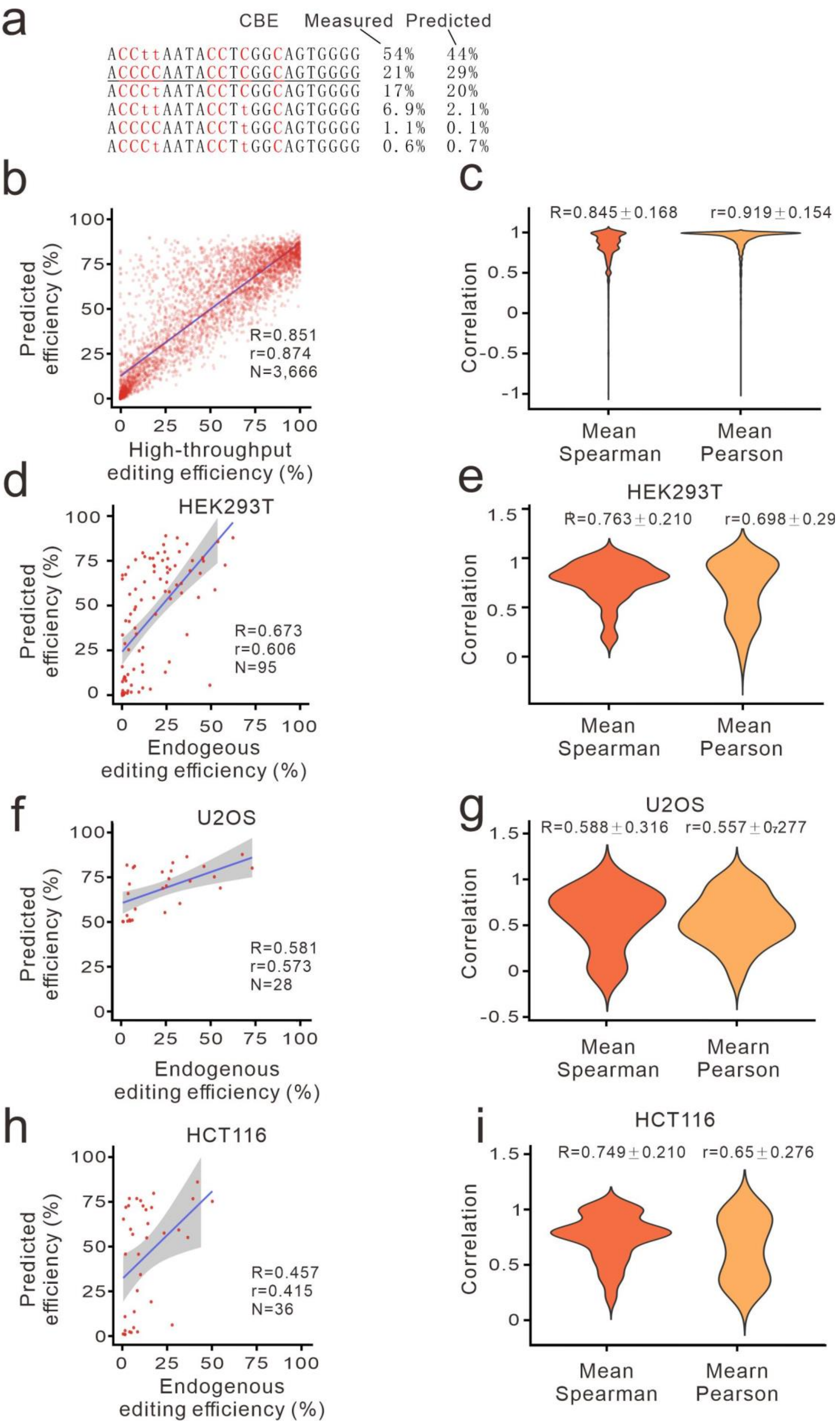
bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correlation of 0.886 (r=0.830, MSE=0.024) and the outcome frequency distribution prediction achieved mean Spearman correlation of 0.857 (r=0.883, MSE=0.020, KL divergence=0.004, Supplementary Fig. 5a-b).

To evaluate the performance of ABEdeepon for endogenous targets, we collected endogenous editing data generated in HEK293T, U2OS and HCT116 cells from literatures [15] and evaluated these datasets with our model, achieving high correlation for efficiencies (Spearman correlation: 0.747-0.797) and mild or high correlation for outcomes (mean Spearman correlation: 0.441-0.901, Fig. 3d-i). To evaluate the performance of the ABEdeepon in induced pluripotent stem cells (iPSCs), we generated dataset for 23 endogenous targets in human iPSCs. ABEdeepon achieved weak Spearman correlation of 0.227 for efficiency prediction probably due to the very low base editing efficiency in iPSCs, and mild mean Spearman correlation of 0.574 for outcome frequency prediction (Supplementary Fig. 5c-d).

Parallelly, we used the CBE editing datasets to train a shared embedding based deep learning model. The targets were randomly split to 68,244 (5,140,905 outcomes) for training, 1,393 (97,841 outcomes) for internal validation, and 3,666 (275,173 outcomes) held out for testing. The resulting model was named “CBEdeepon” which can predict editing outcome sequence frequencies (Fig. 4a). CBEdeepon achieved a Spearman correlation of 0.851 (r=0.874, MSE=0.027) for prediction of editing efficiency with testing dataset (Fig. 4b). To evaluate the performance for prediction of outcome frequency distribution, we tested 3,179 targets with at least 4 outcomes and achieved mean Spearman correlation of 0.845 (r=0.919, MSE=0.009, KL divergence=0.005, Fig. 4c). To test whether CBEdeepon works in other cell types, we used library to generate datasets in HeLa cells and obtained 56,529 valid target with 4,213,361

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4. Evaluation of CBEdeepon for editing efficiency and outcome prediction

(**a**) Example of CBEdeepon prediction for a given target. Measured efficiencies are listed for comparison. Original target sequence is underlined. (**b**) Evaluation of CBEdeepon for conversion efficiency with testing dataset. (**c**) Evaluation of CBEdeepon for conversion outcome sequence frequencies with testing dataset. (**d, f, h**) Evaluation of CBEdeepon for conversion efficiency with endogenous dataset generated in HEK293T, U2OS and HCT116 cells, respectively. (**e, g, i**) Evaluation of CBEdeepon for conversion outcome sequence frequencies with endogenous dataset generated in HEK293T, U2OS and HCT116 cells, respectively.

outcomes (reads number≥100). We extracted 2,827 targets present in HaLa testing dataset as external testing dataset and achieved Spearman correlation of 0.864 (r=0.761, MSE=0.110) for efficiency prediction, and mean Spearman correlation of 0.788 for outcome distribution prediction (r=0.736, MSE=0.042, KL divergence=0.012, Supplementary Fig. 6a-b).

To evaluate the performance of CBEdeepon for endogenous targets, we collected endogenous editing data generated in HEK293T, U2OS and HCT116 cells from literature [15] and evaluated with our model, achieving mild correlation for efficiencies (Spearman correlation: 0.457 to 0.673) and high correlation for outcomes (mean Spearman correlation: 0.588-0.763, Fig. 4d-i). To evaluate the performance of the CBEdeepon in induced pluripotent stem cells (iPSCs), we generated dataset for 24 endogenous targets in human iPSCs. CBEdeepon achieved mild Spearman of 0.525 for efficiency prediction and mean Spearman correlation of 0.558 for outcome frequency prediction (Supplementary Fig. 6c-d).

We finally investigated the positional effects of target nucleotides on prediction. ABEdeepon achieved a good Spearman correlation (>0.5) from position 4 to 12, whereas CBEdeepon achieved a good Spearman correlation (>0.5) from

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

position 3 to 18 (Supplementary Fig. 7a-b). Altogether, our results demonstrated that both ABEdeepon and CBEdeepon models perform well for prediction of conversion efficiency and outcome sequence frequencies.

Discussion

Successful application of base editors requires a rigorous understanding of intended and unintended genome editing. Several groups have offered online tools or software for gRNA design, including iSTOP [16], beditor [17], BE-Designer [18] and BEable-GPS [19]. These design rules conform to base editing requirements that the editable base falls within maximum activity window, but none of them can predict the editing activity and outcome frequencies. To solve this problem, we proposed shared embedding based deep learning models, ABEdeepon and CBEdeepon, to predict editing efficiencies and outcomes. We achieved high performance for prediction of base editing outcomes (mean R: 0.79-0.87 for high-throughput datasets) and efficiencies (R: 0.85-0.89 for high-throughput datasets). During our manuscript revision, two groups developed tools named BE-Hive [13] and DeepBaseEditor [14] which achieved comparable performance to our models for prediction of editing efficiencies and outcomes. The BE-Hive used carefully designed hand-crafted features to develop machine learning models for prediction of base editing outcomes (median r: 0.86-0.99) and efficiency (r: 0.53-0.80) [13]. The DeepBaseEditor followed the author's previous method [20], extracting features using convolutional neural networks, which also achieved high performance for prediction of base editing outcomes (pooled r: 0.95 for high-throughput datasets) and efficiency (R: 0.72-0.79 for high-throughput datasets) [14]. The application of these models together will enhance the prediction ability and facilitate selection of targets with high efficiency and desirable outcome sequences.

The shared embedding based deep learning models used here has a simple yet elegant structure which unifies on-target and off-target input in form. It can be seamlessly extended to other versions of base editors, such as

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SauriABEmax, SauriBE4max, SaKKH-BE3, BE4-CP, dCpf1-BE and eA3A-BE3 [21-25]. It may also be used to generate models to predict editing outcomes for Cas9 and Cas12 nucleases.

Methods

Cell culture and transfection

HEK293T cells and HeLa cells (ATCC) were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% FBS (Gibco), while human iPSCs were cultured on Matrigel-coated plates (ESC qualified, BD Biosciences, San Diego, CA) using hESC mTeSR-1 cell culture medium (StemCell Technologies, Vancouver, Canada) at 37 °C and 5% CO2. All media contained 100 U/ml penicillin and 100 mg/ml streptomycin. For transfection, HEK293T cells and HeLa cells were plated into 6-well plates, DNA mixed with Lipofectamine 2000 (Life Technologies) in Opti-MEM according to the manufacturer's instructions, while iPSCs using Lipofectamine 3000 (Life Technologies). Cells were tested negative for mycoplasma.

Plasmid construction

SB transposon (pT2-SV40-BSD-ABEmax and pT2-SV40-BSD-BE4max) was constructed as follows: first, we replaced the Neo^R gene (AvrII-KpnI site) on pT2-SV40-Neo^R with BSD, resulting in pT2-SV40-BSD vector; second, backbone fragment of pT2-SV40-BSD was PCR-amplified with Gibson-SV40-F and Gibson-SV40-R, and ABEmax fragment was PCR-amplified from pCMV_ABEmax_P2A_GFP (Addgene#112101) with Gibson-ABE/BE4-F and Gibson-ABE/BE4-R, and BE4max fragment was PCR-amplified from pCMV_AncBE4max (Addgene#112094) with Gibson-ABE/BE4-F and Gibson-ABE/BE4-R; third, the backbone fragments were ligated with ABEmax and BE4max using Gibson Assembly (NEB), resulting in pT2-SV40-BSD-ABEmax and pT2-SV40-BSD-BE4max, respectively.

Generation of cell lines expressing ABEmax or BE4max

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HEK293T cells and HeLa cells were seeded at ~40% confluency in a 6-well dish the day before transfection, 2 μg of SB transposon (pT2-SV40-BSD-ABEmax or pT2-SV40-BSD- BE4max) and 0.5 μg of pCMV-SB100x were transfected using 5 μl of Lipofectamine 2000 (Life Technologies). After 24 h, cells were selected with 10 μg/ml of blasticidin for 10 days. Single cells were sorted into 96-well plates for colony formation. Conversion efficiency was performed to screen cell clones with high levels of ABEmax and BE4max expression.

The gRNA-target library construction

The on-target library design and construction were described previously[10]. Briefly, full-length oligonucleotides were PCR-amplified and cloned into Lentiviral vector by Gibson Assembly (NEB). The Gibson Assembly products were electroporated into MegaX DH10B^{TM} T1^{R} Electrocomp^{TM} Cells (Invitrogen) using a GenePulser (BioRad) and grown at 32 °C, 225 rpm for 16 h. The plasmid DNA was extracted from bacterial cells using Endotoxin-Free Plasmid Maxiprep (Qiagen).

Lentivirus production

Lentivirus production was described previously[10]. Briefly, 12 μg of plasmid library, 9 μg of psPAX2, and 3 μg of pMD2.G (Addgene) were transfected into a 10-dish HEK293T cells with 60 μl of Lipofectamine 2000. Virus were harvested twice at 48 h and 72 h post-transfection. The virus was concentrated using PEG8000 (no. LV810A-1, SBI, Palo Alto, CA), dissolved in PBS and stored at -80 °C.

Screening experiments in HEK293T and HeLa cells

HEK293T or HeLa cells expressing ABEmax or BE4max were plated into 15 cm dish at ~30% confluence. After 24 h, cells were infected with gRNA library with at least 1000-fold coverage of each gRNAs. After 24 h, the cells were cultured in the media supplemented with 2 μg/ml of puromycin for 5 days. Cells

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were harvested and the genomic DNA was isolated using Blood & Cell Culture DNA Kits (Qiagen). The integrated region containing the gRNA coding sequences and target sequences were PCR-amplified using primers Deep-seq-library-F/R with Q5 High-Fidelity 2X Master Mix (NEB). We performed 60-70 PCR reactions using 10 μg of genomic DNA as template per reaction for deep sequencing analysis. The PCR conditions: 98 °C for 2 min, 25 cycles of 98 °C for 7 s, 67 °C for 15 s and 72 °C for 10 s, and the final extension, 72 °C for 2 min. The PCR products were mixed and purified using Gel Extraction Kit (Qiagen). The purified products were sequenced on Illumina HiSeq X by 150-bp paired-end sequencing.

Data analysis

FASTQ raw sequencing reads were processed to identify gRNA editing activity and editing outcomes. The nucleotides in a read with quality score < 10 was masked with a character "N". Due to the integrated design strategy, we first separated a read to designed gRNA region, scaffold region, and target region to extract the corresponding sequence. The designed gRNA was then aligned to the reference gRNA library to mark the reads. The target sequence was compared to the designed gRNA to mark all types of conversion (i.e., canonical A-G, C-T conversions and non-canonical A-C/T, C-A/G, G-A/T, T-A/C conversions at each position). We screened out gRNAs with a total valid reading of less than 100. Then, the efficiency for a specific conversion type at a position can be calculated by the following formula:

$$\text{conversion efficiency at spesific position} = \frac{\text{NO. of converted reads in a conversion type at spesific position}}{NO.of\ total\ valided\ reads}$$

However, the conversion efficiency was very low for non-canonical conversions (mean efficiency < 0.005). Thus, we only consider the canonical conversions for the outcome frequency distribution analysis. Theoretically, the canonical editing combinations of 20 bases is at most 2^20. However, we found that positions 1-2 and 18-20 always contain only one conversion. If there exist

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

canonical bases, positions 3-17 may have multiple conversions at the same time. Therefore, we programmed this conversion scheme to obtain all possible editing outcomes and assign true editing frequencies to the outcomes that exist in the sequencing data and assign 0 frequencies to the outcomes not found in the sequencing data. Then, the editing frequency of a gRNA-outcome can be described as:

$$\text{specific gRNA\_outcome frequency} = \frac{\text{NO. of reads in a specific editing outcome}}{NO.of\ total\ valided\ reads}$$

Note that the non-converted targets were also considered as an editing outcome. So, the editing efficiency for a target can be simply calculated by the following formula:

$$\text{conversion efficiency} = 1 - \text{non-converted outcome frequency}$$

Encoding

Drawing on concepts from the field of natural language processing (NLP), nucleotides A, C, G, and T can be regarded as words in a DNA sequences. Therefore, we can widely use algorithms in the NLP field to solve prediction tasks in the CRISPR field, especially the use of embedding algorithms to get the continuous representation of discrete nucleotide sequences [26]. Unlike the common efficiency prediction that only needs to input one single sequence for regression models, in this research, both the gRNA-outcome pairs (for BEdeepon) and gRNA-target pairs (for BEdeepoff) has two different sequences as inputs. For gRNA-outcome pairs, there are four words in the index vocabulary (i.e., A, C, G, and T). So, the vocabulary can be described as:

$$D = \{1{:}A, 2{:}C, 3{:}G, 4{:}T\}$$

So, an input sequence can be described as:

$$\boldsymbol{x_i} = \{x_{i0}, x_{i1} \cdots x_{it}, \cdots x_{i(T-\mathbf{1})}\},$$

where $i \in \{1,2\}$ denotes the i-th sequence in an gRNA-outcome pair or an gRNA-target pair, x_{it} is the t-th element of the i-th sequence, T is the sequence length. For example, for a gRNA-outcome

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pair GTGGAACATCCACTTGACCTAGG (seq1, gRNA + NGG) and GTGGAGCGTCCACTTGACCTAGG (seq2, one outcome) can be encoded as:

$\boldsymbol{x_1} = [3,4,3,3,1,1,2,1,4,2,2,1,2,4,4,3,1,2,2,4,1,3,3]$ (i.e., seq1)

and

$\boldsymbol{x_2} = [3,4,3,3,1,1,2,3,4,2,2,1,2,4,4,3,1,2,2,4,1,3,3]$ (i.e., seq2),

respectively.

Shared embedding

Inspired by the algorithms in the recommender system [27] and click-through rate (CTR) [28] prediction modeling, both the generalization capacity and training speed will benefit from the sharing of the same embedding matrix instead of training independent embedding matrices for each input. In this research, a discrete nucleotide encoding x_{it} is projected to the dense real-valued space $\mathbf{E}_{\boldsymbol{i}} \in \mathbb{R}^{T\times m}$ (m is a hyperparameter corresponds to the embedding dimension) to get the embedding vector $\mathbf{e}(x_{it})$. Then a final embedding matrix $\mathbf{E}$ is needed to get the combined information from those two embedding matrices by:

$$\mathbf{E} = g(\mathbf{E_1}, \mathbf{E_2}) \quad (1)$$

where g can be sum, mean, or even a simple concatenate function. However, the sum or mean function is more suitable because it can reduce the redundant features in $\mathbf{E_1}$and $\mathbf{E_2}$. We choose the sum function here for simplicity.

Feature extraction and model prediction

Long-short-term memory network (LSTM) and gated recurrent unit network (GRU) are a type of recurrent neural networks (RNN) algorithms used to address the vanishing gradient problem in modelling time-dependent and sequential data tasks [29]. Usually, a bidirectional manner was used to capture the information from the forward and backward directions of a sequence, which is biLSTM or biGRU. Our work and others' work have shown that, as an important component, biLSTM can be used alone or with convolutional neural

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

network (CNN) to achieve good performances in various regression and classification tasks involving biological sequences [10, 30-32]. Here, we tried biLSTM, biGRU, and the newly proposed transformer structure [33], and found biLSTM had the fastest convergence speed. The input and output of biLSTM can be described by the following equations:

$$\overrightarrow{\boldsymbol{h}_t} = \overrightarrow{LSTM}(\boldsymbol{e}(\overrightarrow{x_t}), \overrightarrow{\boldsymbol{h}_{t-1}}) \quad (2)$$

$$\overleftarrow{\boldsymbol{h}_t} = \overleftarrow{LSTM}\left(\boldsymbol{e}(\overleftarrow{x_t}), \overleftarrow{\boldsymbol{h}_{t-1}}\right) \quad (3)$$

Thus, the output context vectors of biLSTM are $\boldsymbol{h}_0 = [\overrightarrow{\boldsymbol{h}_0}; \overleftarrow{\boldsymbol{h}_{L-1}}]$, $\boldsymbol{h}_1 = [\overrightarrow{\boldsymbol{h}_1}; \overleftarrow{\boldsymbol{h}_{L-2}}]$, etc. Thus, we can concatenate the forward and backward hidden state as $\mathbf{H} = \{\boldsymbol{h}_0, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_{L-1}\}$, which contains the bidirectional information in the shared embedding feature matrix. Before the fully connected layers, we tried different input features based on the trade-off of the convergence speed and the performance of the model. The aforementioned features are last hidden unit, max pooling operation on $\mathbf{H}$, and average pooling on $\mathbf{H}$. The equations are as following:

$$\boldsymbol{h}_{Last} = [\overrightarrow{\boldsymbol{h}_{L-1}}; \overleftarrow{\boldsymbol{h}_{L-1}}] \quad (4)$$

$$\mathbf{F}_{MaxPool} = \max_{t=1}^{L-1} \boldsymbol{h}_t \quad (5)$$

$$\mathbf{F}_{MeanPool} = \mathrm{mean}_{t=1}^{L-1} \boldsymbol{h}_t \quad (6)$$

We observed that the last hidden state is the only need to obtain an optimal performance, i.e.,

$$s = \sigma(f(\boldsymbol{h}_{Last})) \quad (7)$$

where, f is fully connected layers, σ is the *leaky_relu* activation function and o is the output score for a specific gRNA-outcome pair.

It should be noted that the gRNA + NGG in a set of gRNA-outcome pairs (a gRNA batch with K samples) are all the same, and the outcome sequences are converted from the same target, so the output scores of a gRNA-batch can be denoted as $\boldsymbol{s} = [s_1, s_2, \cdots, s_K]$. Then, a *softmax* activation function can be

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

applied to $\boldsymbol{s}$ to get the predicted frequency distribution with a sum of 1 for the gRNA-outcome pair (Eq. 8, 9).

$$q_i = softmax(\boldsymbol{s})_i = \frac{e^{s_i}}{\sum_{j=1}^{K} e^{s_j}} \ (8)$$

$$\boldsymbol{q} = \sum_{j=1}^{K} q_i = 1 \ (9)$$

We have described in the **Data Analysis** section, that the conversion efficiency of a gRNA can be calculated by the formula:

$$\text{conversion efficiency} = 1 - \text{non-converted outcome frequency}$$

If we let q_0 be non-converted outcome frequency in $\boldsymbol{q}$. Then, the predicted efficiency $\tilde{y}$ will be $1 - q_0$.

Combined weighted loss function

For the on-target models, a calculated true gRNA-outcome frequency distribution can be denoted as $\boldsymbol{p} = [p_1, p_2, \cdots, p_K]$. So, it's naturally to apply Kullback–Leibler (KL) divergence loss function to minimize the difference between the predicted frequency distribution $\boldsymbol{q}$ and the true frequency distribution $\boldsymbol{p}$. The standard KL divergence loss function $D_{KL}(\boldsymbol{p}||\boldsymbol{q})$ is defined as:

$$D_{KL}(\boldsymbol{p}||\boldsymbol{q}) = \sum_{j=1}^{K} p_i \log\left(\frac{p_i}{q_i}\right) \ (10)$$

Basically, we wanted that a gRNA-outcome pair with a large number of reads in a gRNA batch has a more accurate prediction value. The weight of a sample loss was re-assigned depending on its corresponding read counts w_i. The modified loss function of Eq.10 is:

$$L_1 = D_{KL}(\boldsymbol{p}||\boldsymbol{q}) = \sum_{j=1}^{K} w_i p_i \log\left(\frac{p_i}{q_i}\right) \ (11)$$

A number of studies [34] have shown that multi-task learning architecture can significantly improves the stability and generalization capacity of the model. In addition to KL divergence loss (which measures the difference between the two probability distributions as a whole), the mean squared error (MSE) loss

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

function also helps to minimize the difference between each of the p_i, q_i pairs individually. So, we adopted the MSE loss in a weighted manner:

$$L_2 = \frac{1}{2K}\sum_{j=1}^{K} w_i(\, p_i - q_i)^2 \;\; (12)$$

Then, the final loss function can be denoted as follows in Eq. 13:

$$L_{total} = \sum_{j=1}^{K} w_i p_i \log\left(\frac{p_i}{q_i}\right) + \frac{1}{2K}\sum_{j=1}^{K} w_i(\, p_i - q_i)^2 \;\; (13)$$

The loss function for the off-target model is simply the ordinary MSE loss.

Training setting

Both the on-target and off-target datasets were randomly split into two parts, one for training, and the other for holdout testing. Since there are a large number of gRNA-outcome pairs (over 5.5 million) in the on-target dataset, we did not conduct cross-validation training for this dataset but used more external datasets to test the generalization capacity of the models. The sample size of off-target datasets is around 50,000, so the stability of model performance was estimated by a 10-fold shuffled validation together with the external integrated and endogenous datasets. The ABEdeepon and CBEdeepon models share the following hyperparameters: embedding dimension, 64; LSTM hidden unites, 128; LSTM hidden layers, 1; dropout rate, 0.5; fully connected layers,1 (2*128->1). The ABEdeepoff and CBEdeepoff models share the following hyperparameters: embedding dimension, 256; LSTM hidden unites, 512; LSTM hidden layers, 1; dropout rate, 0.5; fully connected layers, 2 (6*256->3*256->1). For all the models, the Adam optimizer was used with a customized learning rate decay strategy that gradually reduce the learning rate from 0.001, 0.0001, 0.00005 to 0.00001.

Tools used in the study

BWA-0.7.17 was used to identify the designed gRNA[35]. PyTorch 1.6 [36] was used for building deep learning models.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (81870199, 81630087), the Foundation for Innovative Research Group of the National Natural Science Foundation of China (31521003), Science and Technology Research Program of Shanghai (Grant Number 19DZ2282100).

References

1. 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).
2. 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).
3. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).
4. Wang, B. et al. krCRISPR: an easy and efficient strategy for generating conditional knockout of essential genes in cells. *Journal of biological engineering* **13**, 35 (2019).
5. Xie, Y. et al. An episomal vector-based CRISPR/Cas9 system for highly efficient gene knockout in human pluripotent stem cells. *Scientific reports* **7**, 2320 (2017).
6. Zhang, Y. et al. Programmable base editing of zebrafish genome using a modified CRISPR-Cas9 system. *Nature communications* **8**, 118 (2017).
7. Kim, K. et al. Highly efficient RNA-guided base editing in mouse embryos. *Nature biotechnology* **35**, 435-437 (2017).
8. Liu, Z. et al. Highly efficient RNA-guided base editing in rabbit. *Nature communications* **9**, 2717 (2018).
9. Zeng, Y. et al. Correction of the Marfan Syndrome Pathogenic FBN1 Mutation by Base Editing in Human Cells and Heterozygous Embryos. *Molecular therapy : the journal of the American Society of Gene Therapy* **26**, 2631-2637 (2018).
10. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nature communications* **10**, 4284 (2019).
11. Koblan, L.W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nature biotechnology* **36**, 843-846 (2018).
12. Mates, L. et al. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nature genetics* **41**, 753-761 (2009).
13. Wang, Y. et al. Regulated complex assembly safeguards the fidelity of Sleeping Beauty transposition. *Nucleic acids research* (2016).
14. Wang, Y. et al. Suicidal autointegration of sleeping beauty and piggyBac transposons in eukaryotic cells. *PLoS genetics* **10**, e1004103 (2014).
15. Song, M. et al. Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. *Nature biotechnology* **38**, 1037-1043 (2020).
16. Billon, P. et al. CRISPR-Mediated Base Editing Enables Efficient Disruption of Eukaryotic Genes through Induction of STOP Codons. *Molecular cell* **67**, 1068-1079 e1064 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

17. Dandage, R., Despres, P.C., Yachie, N. & Landry, C.R. beditor: A Computational Workflow for Designing Libraries of Guide RNAs for CRISPR-Mediated Base Editing. *Genetics* **212**, 377-385 (2019).

18. Hwang, G.H. et al. Web-based design and analysis tools for CRISPR base editing. *BMC bioinformatics* **19**, 542 (2018).

19. Wang, Y. et al. Comparison of cytosine base editors and development of the BEable-GPS database for targeting pathogenic SNVs. *Genome biology* **20**, 218 (2019).

20. Kim, H.K. et al. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nature biotechnology* **36**, 239-241 (2018).

21. Hu, Z. et al. A compact Cas9 ortholog from Staphylococcus Auricularis (SauriCas9) expands the DNA targeting scope. *PLoS biology* **18**, e3000686 (2020).

22. Kim, Y.B. et al. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nature biotechnology* **35**, 371-376 (2017).

23. Li, X. et al. Base editing with a Cpf1-cytidine deaminase fusion. *Nature biotechnology* **36**, 324-327 (2018).

24. Huang, T.P. et al. Circularly permuted and PAM-modified Cas9 variants broaden the targeting scope of base editors. *Nature biotechnology* **37**, 626-631 (2019).

25. Wang, X. et al. Efficient base editing in methylated regions with a human APOBEC3A-Cas9 fusion. *Nature biotechnology* **36**, 946-949 (2018).

26. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado & Dean., a.J. Distributed Representations of Words and Phrases and their Compositionality. *Preprint at: https://arxiv.org/abs/1310.4546* (2013).

27. Maja R. Rudolph, Francisco J. R. Ruiz, Stephan Mandt & Blei., a.D.M. Exponential Family Embeddings. *Preprint at: https://arxiv.org/abs/1608.00778* (2016).

28. Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li & He., a.X. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. *Preprint at: https://arxiv.org/abs/1703.04247* (2017).

29. Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau & Bengio., a.Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *Preprint at: https://arxiv.org/abs/1409.1259* (2014).

30. Rowe, G.G., Stenlund, R.R., Thomsen, J.H., Terry, W. & Querimit, A.S. Coronary and systemic hemodynamic effects of cardiac pacing in man with complete heart block. *Circulation* **40**, 839-845 (1969).

31. Effects on surface waters. *J Water Pollut Control Fed* **42**, 1084-1088 (1970).

32. Harris, A. Pacemaker 'heart sound'. *Br Heart J* **29**, 608-615 (1967).

33. Ashish Vaswani et al. in Proceedings of the 31st International Conference on Neural Information Processing Systems 6000–6010 (2017).

34. Alex Kendall, Yarin Gal & Cipolla., a.R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *Preprint at: https://arxiv.org/abs/1705.07115* (2018).

35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

36. Paszke, A. et al. in Advances in neural information processing systems 8026-8037 (2019).

37. Cancellieri, S., Canver, M.C., Bombieri, N., Giugno, R. & Pinello, L. CRISPRitz: rapid, high-throughput and variant-aware in silico off-target site identification for CRISPR genome editing.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.14.435303; this version posted March 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bioinformatics **36**, 2001-2008 (2020).