BEdeepoff: an *in silico* tool for off-target prediction of ABE and CBE base editors

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

[1]State Key Laboratory of Genetic Engineering, School of Life Sciences, Zhongshan 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, including adenine base editors (ABEs) and cytosine base editors (CBEs), are valuable tools for introducing point mutations, but they frequently induce unwanted off-target mutations. Here, we performed a high-throughput gRNA-target library screening to measure editing efficiencies at integrated genomic off-targets and obtained datasets of 48,632 and 52,429 off-targets for ABE and CBE, respectively. We used the datasets to train deep learning models, resulting in ABEdeepoff and CBEdeepoff which can predict editing efficiencies at off-targets. These tools are freely accessible via online web server

http://www.deephf.com/#/bedeep.

Introduction

Base editors enable programmable conversion of a single nucleotide in the mammalian genome and have a broad range of research and medical applications. They are fusion proteins that include a catalytically impaired Cas9 nuclease ($Cas9^{D10A}$) and a nucleobase deaminase [1, 2]. $Cas9^{D10A}$ nuclease and a ~100 nucleotide guide RNA (gRNA) forms a $Cas9^{D10A}$-gRNA complex, recognizing a 20 nucleotide target sequence followed by a downstream protospacer adjacent motif (PAM) [3-5]. Once the $Cas9^{D10A}$-gRNA complex binds to target DNA, it opens a single-stranded DNA loop [3]. The nucleobase deaminase modifies the single-stranded DNA within a small ~5 nucleotide window at the 5′ end of the target sequence [1, 2]. Two classes of base editors have been developed to date: cytidine base editors (CBEs) convert target C:G base pairs to T:A [1], and adenine base editors (ABEs) convert A:T to G:C [2]. Base editors have been successfully used in diverse organisms including prokaryotes, plants, fish, frogs, mammals, and human embryos [6-9].

As a major concern of base editing, off-target editing can occur on genomic sequences that are similar to the 20-nucleotide target sequence [10, 11]. Experimental evaluation of a target sequence is time-consuming, prompting us to develop *in silico* tools for target sequence evaluation. We recently used a high-throughput strategy for gRNA-target library screening for SpCas9 activity [12]. In this study, we designed libraries of gRNA-off-target sequence pairs and performed high-throughput screen, obtaining 48,632 and 52,429 valid off-targets for ABE and CBE, respectively. The resulting datasets were used to train deep learning models, resulting in ABEdeepoff and CBEdeepoff which can predict editing efficiency at potential off-targets for ABE and CBE, respectively.

Results

A guide RNA-target pair strategy for test of editing efficiency at off-targets

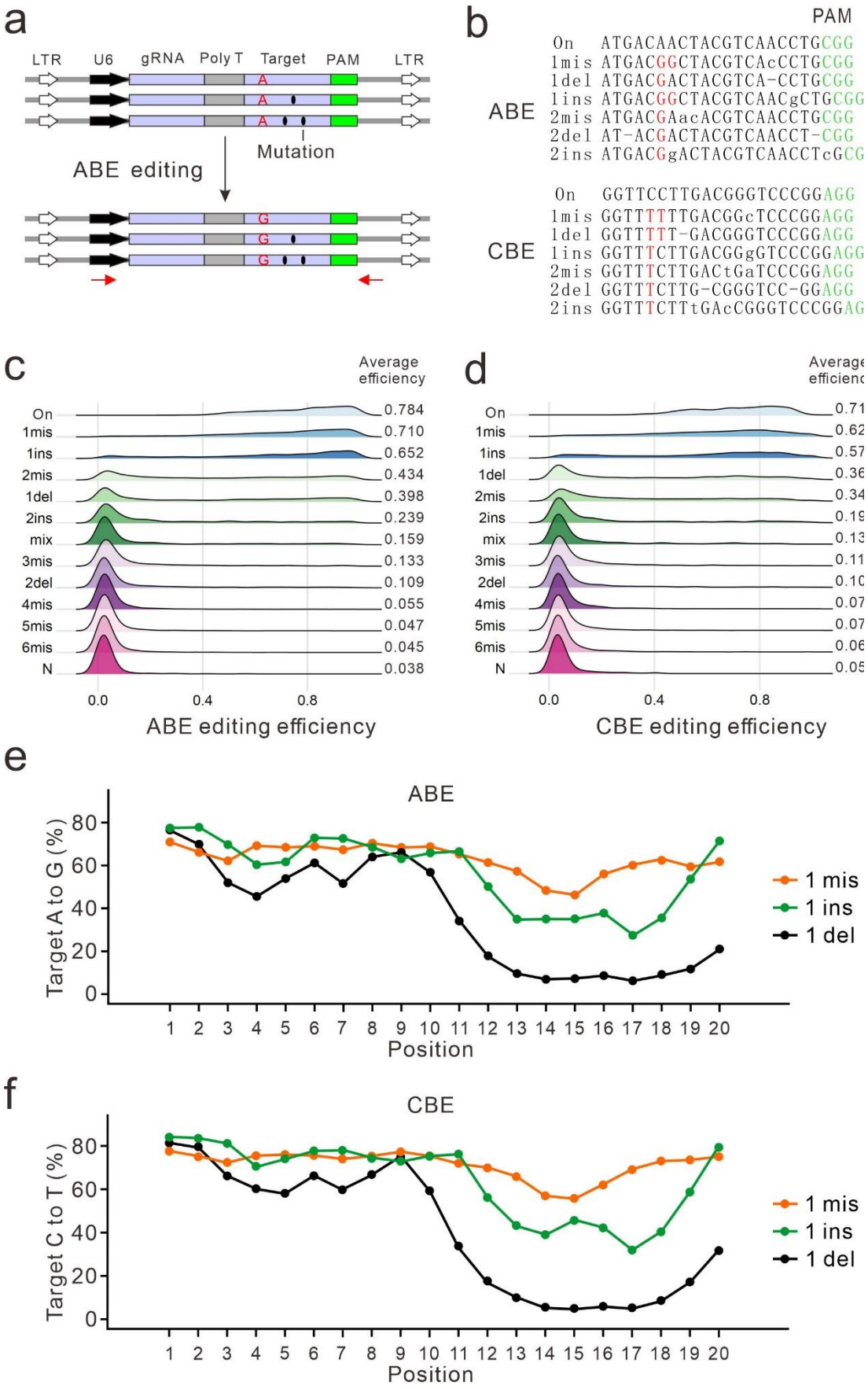

Figure 1. High-throughput screen of conversion efficiencies at off-

targets

(**a**) A gRNA group of pairwise library for off-target screening. Black ellipses indicate mutations; red arrows indicate primers for target site amplification. (**b**) Deep-sequencing results reveal that A to G conversion for ABE and C to T conversion (red letter) for CBE occurred. Mismatched nucleotides and inserted nucleotides are indicated by lowercase letters; deletion is indicated by "-". On: on-target sequence; 1mis: 1bp mismatch; 1del: 1 bp deletion; 1ins: 1 bp insertion. (**c, d**) Effects of the mutation type on nucleotide conversion efficiency. Mix: mismatches mixed with indels; On: on-target; N: 23 bp random sequence as negative control. (**e, f**) Positional effect on conversion efficiency for 1 bp mismatch, 1 bp insertion and 1 bp deletion.

To investigate editing efficiency at off-targets, we designed two gRNA-target pair libraries with one for ABE and one for CBE (Fig. 1a). Each library contains 91,566 gRNA-target pairs, distributed among 1,383 gRNA groups for ABE and 1,378 gRNA groups for CBE. Each group contains one gRNA-on-target pair and multiple gRNA-off-target pairs. The mutation type included mismatches (1-6 bp mismatches per off-target, 71,099 for ABE and 70,594 for CBE), deletions (1-2 bp per target, 6,521 for ABE and 6,759 for CBE), insertions (1-2 bp per target, 11,396 for ABE and 11,562 for CBE) and mismatches mixed with insertions/deletions (indels, 1-2 bp mismatches plus 1-2 bp indels, total mutant nucleotides are 2-3 bp, 888 for ABE and 881 for CBE). We also designed non-target sequences as control (279 for ABE and 392 for CBE).

To make the high-throughput data more reproducible, we first generated a panel of single cell-derived clones that stably express ABE or CBE base editors. Optimized version of base editors (ABEmax for ABE; AncBE4max for CBE) [13] were used in this study. Because of the large size, we integrated base editors into the genome by using the Sleeping Beauty (SB) transposon system (Supplementary Fig. 1a) [14-16]. We tested conversion efficiency in each clone

and selected an efficient clone for each base editor (Supplementary Fig. 1b-c).

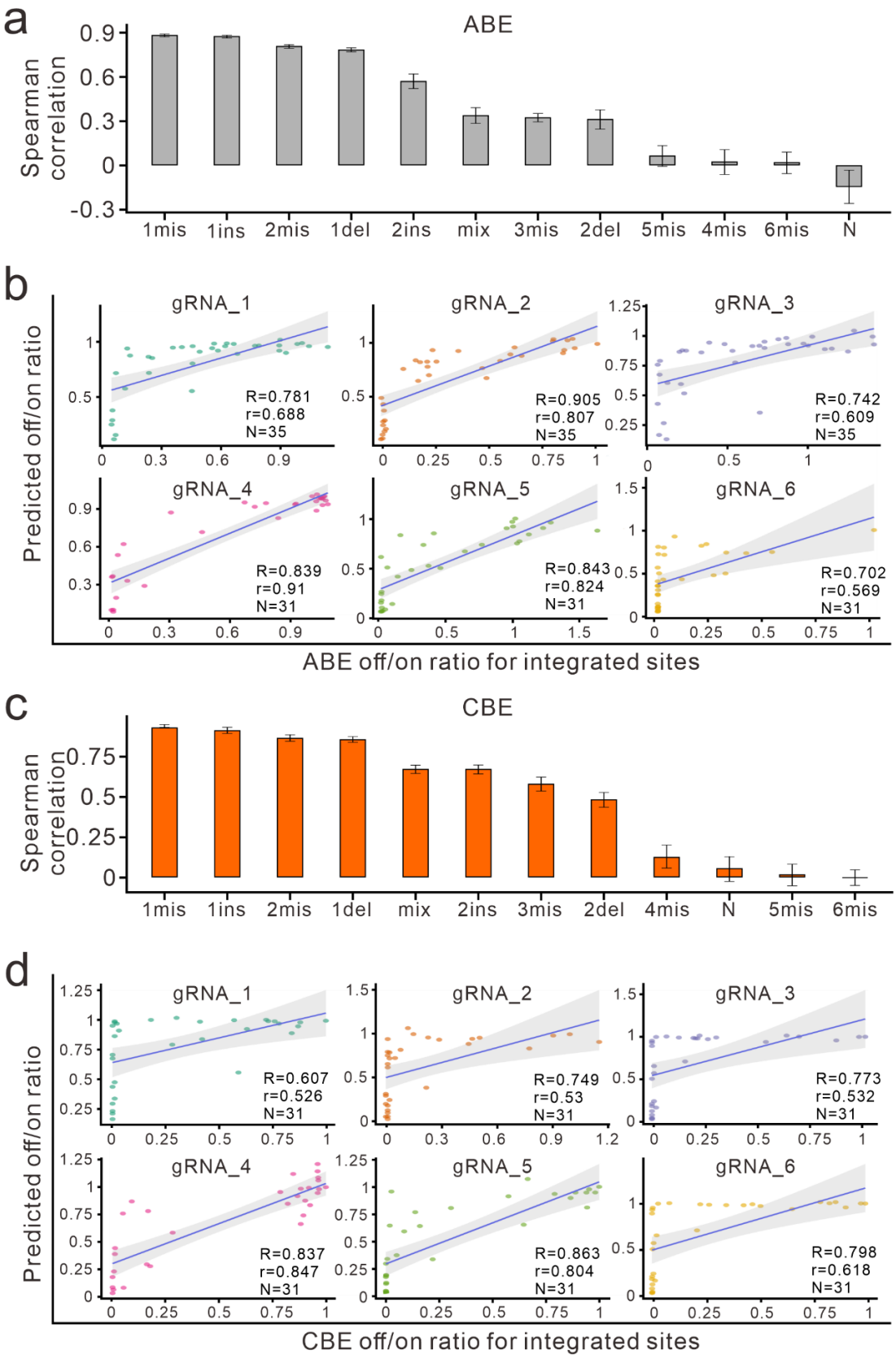

Figure 2. Evaluation of ABEdeepoff/CBEdeepoff for efficiency prediction at off-targets

(**a**) Evaluation of ABEdeepoff prediction for different mutation types with testing dataset. N: 23 bp random sequence as negative control. (**b)** Evaluation of ABEdeepoff prediction for conversion efficiency at six groups of integrated off-targets. (**c)** Evaluation of CBEdeepoff prediction for different mutation types with testing dataset. (**d)** Evaluation of CBEdeepoff prediction for conversion efficiency at six groups of integrated off-targets. R: Spearman correlation, r: Pearson correlation.

We packaged the gRNA-target pair library into lentiviruses and transduced them into recipient cells. Five days after transduction, genomic DNA was extracted, and synthesized off-targets were PCR-amplified for deep-sequencing (Fig. 1a). Deep sequencing results revealed that A to G conversion for ABE and C to T conversion for CBE occurred (Fig. 1b). In this study, an off-target efficiency was defined as No. of edited reads divided by the No. of total valid reads. The screening assay was experimentally repeated twice, and editing efficiency in two independent replicates showed high correlation (Spearman correlation, 0.956 for ABE and 0.965 for CBE). Data from the two replicates were combined together for subsequent analysis. We obtained valid ABE efficiencies (reads number>100) of 48,632 for off-targets, 1,052 for on-targets and 153 for non-target controls. We obtained valid CBE efficiencies (reads number>100) of 52,429 for off-targets, 1,030 for on-targets and 220 for non-target controls.

The large-scale datasets generated here allowed us to analyze the effects of mutation type on editing efficiency. For both base editors, editing efficiencies decreased with an increasing number of mismatches (Fig. 1c-d). Deletions had a stronger influence than insertions and mismatches. The overall influence of mutation type on ABE could be ranked as 6mis > 5mis > 4mis > 2del > 3mis > mix > 2ins > 1del > 2mis > 1ins > 1mis (Fig. 1c); the overall influence of mutation type on CBE could be ranked as 6mis > 5mis > 4mis > 2del > 3mis > mix > 2ins 2mis > 1del > 1ins > 1mis (Fig. 1d).

Next, we investigated the positional effects of mutation on editing efficiency. For both base editors, one mismatch only had minimal influence on editing efficiency at positions 14 and 15 but not at other positions; one insertion had minimal influence at the PAM-distal region but had stronger influence at positions 12-19; one deletion had mild influence at the PAM-distal region but had a stronger influence at PAM-proximal region (Fig. 1e-f). Interestingly, mutations at positions 19-20 had less influence than those at other seed region. Two mismatches had a stronger influence on editing efficiency when they both occurred on the seed region for both base editors (Supplementary Fig. 2a-b).

Developing models for prediction of editing efficiency at off-targets

Next, we used ABE off-target editing datasets to train a shared embedding based deep learning model where a gRNA-off-target pair was considered as a unique sequence and used as input (Supplementary Fig. 3). The shared embedding design encoded the input gRNA and off-target sequences in the same scheme, and then embedded the representation vectors in the same matrix space. Both generalization ability and training speed will benefit from sharing the same embedding matrix instead of training an independent embedding matrix for each input. The gRNA-off-target pairs and non-target controls (48,632+153=48,785) were randomly split to two parts: one contained 41,906 for tenfold cross-validation with shuffling, and the other contained 4,879 for holdout testing (which were never seen by the model). The resulting model was named “ABEdeepoff” which can predict editing efficiency at a potential off-target. The model achieved high correlation for 1-2 bp mismatch, 1 bp insertion and 1 bp deletion ($R>0.7$), mild correlation for 2 bp insertions ($R=0.57$), weak correlation for 2 bp indels, 3 bp mismatches and mix mutations ($R>0.3$) and very weak correlation for the remaining mutations ($R<0.03$, Fig. 2a). These results suggest that ABEdeepoff performed well for off-targets with high editing efficiency but not with low editing efficiency.

Next, we evaluated the performance of the model with six groups of integrated

off-target datasets collected from literature [10, 17], and achieved Spearman correlation values that varied from 0.702 to 0.905 (Fig. 2b). Finally, we evaluated the performance of our model with 14 groups of endogenous off-target datasets collected from literature [17]. We achieved a mild correlation (R>0.4) for seven targets and weak correlation for the remaining targets probably due to the low editing efficiencies measured at these off-targets (Supplementary Fig. 4).

Parallelly, we used the CBE off-target editing datasets to train a shared embedding based deep learning model. The gRNA-off-target pairs and non-target controls (52,429+220=52,649) were randomly split to two parts: one contained 47,384 for tenfold cross-validation with shuffling, and the other contained 5,265 for holdout testing. The resulting model was named "CBEdeepoff" which can predict editing efficiencies at potential off-targets. The model achieved high correlation (R>0.6) for 1-2 bp mismatch, 1-2 bp insertion, 1 bp deletion and mix mutations, mild correlation (R>0.4) for 3 bp mismatches and 2 bp deletions, and very weak correlation (R<0.2) for the remaining mutations (Fig. 2c). These results can be explained by that CBEdeepoff performed well for off-targets with high editing efficiency but not with low editing efficiency.

Next, we evaluated the performance of the model with six integrated groups of off-target datasets collected from literature [11, 17], and achieved Spearman correlation values varied from 0.607 to 0.863(Fig. 2d). Finally, we evaluated the performance of the model with seven groups of endogenous off-target datasets collected from literature [17]. We achieved mild correlation (R>0.4) for four targets and weak correlation for the remaining targets probably due to the very low editing efficiencies measured at these off-targets (Supplementary Fig. 5).

ABEdeepoff /CBEdeepoff can be used together with Cas-OFFinder [18] and CRISPRitz [19]. Both of the software can rapidly identify potential off-targets

based on sequence similarity in the whole genome, and ABEdeepoff /CBEdeepoff can calculate editing efficiency for each sequence to define the potential off-targets (Fig. 3). The output file of Cas-OFFinder and CRISPRitz can be used as input file for ABEdeepoff or CBEdeepoff. We finally provided an online webserver named BEdeep, which is available at http://www.deephf.com/#/bedeep.

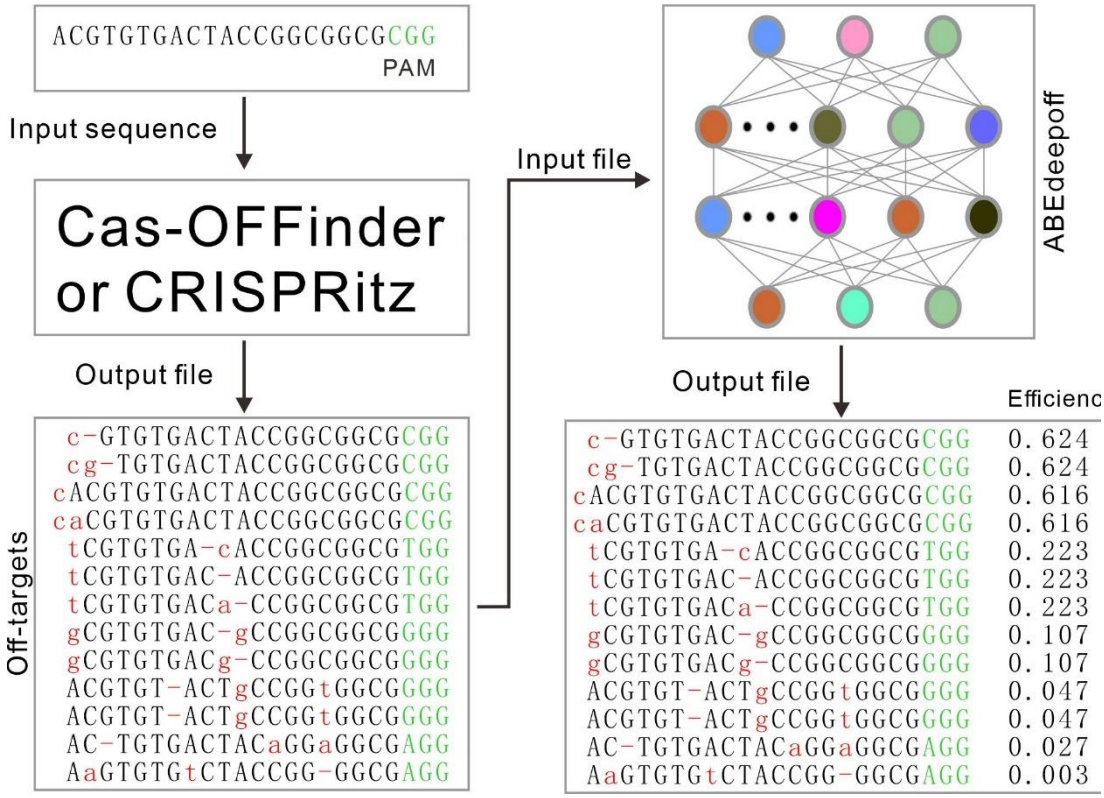

Figure 3. Example of ABEdeepoff for off-target identification.

A target sequence is input into Cas-OFFinder or CRISPRitz to identify off-targets in human genome based on sequence similarity. Here we set up the parameters as maximal 2 mismatches, 1 deletion and 1 insertion. The output file containing potential off-targets is used as input file for ABEdeepoff. ABEdeepoff calculates editing efficiency for each potential off-targets.

Discussion

Off-target editing is always a major concern of genome editing. As Cas9-deaminase fusion proteins, both ABE and CBE base editors can induce sequence independent and dependent off-target mutations [10, 11, 20-22]. By

engineering deaminases, sequence independent off-target mutations can be minimized to background level [20, 23]. However, how to minimize sequence dependent off-target mutations is a challenge because eukaryote genomes contain hundreds of sequences similar to the target sequences. In this study, we developed models to predict editing efficiencies on potential off-targets for both ABE and CBE, facilitating users to select highly specific targets.

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 SauriABEmax, SauriBE4max, SaKKH-BE3, BE4-CP, dCpf1-BE and eA3A-BE3 [24-28]. It may also be used to generate models to predict editing outcomes for Cas9 and Cas12 nucleases.

Methods

Cell culture and transfection

HEK293T cells (ATCC) were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% FBS (Gibco) at 37 °C and 5% CO2. All media contained 100 U/ml penicillin and 100 mg/ml streptomycin. For transfection, HEK293T cells were plated into 6-well plates, DNA mixed with Lipofectamine 2000 (Life Technologies) in Opti-MEM according to the manufacturer's instructions. 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

HEK293T 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-off-target library construction

The off-target library was designed as follows: both ABEs and CBEs libraries contained 91,556 oligonucleotides, distributed among 1,383 and 1,378 gRNA groups for ABEs and CBEs, respectively; for ABEs, the off-target library included 51,170 1bp-mismatch, 10,889 2bp-mismatch, 10,310 1bp-insertion, 4,960 1bp-deletion, 2,261 3bp-mismatch, 2,264 4bp-mismatch, 2,255 5bp-mismatch, 2,260 6bp-mismatch, 1,086 2bp-insertion, 1,561 2bp-deletion, 888 mix, 279 non-target and 1,383 on-target; for CBEs, the off-target library included 50,676 1bp-mismatch, 11,054 2bp-mismatch, 10,481 1bp-insertion, 5,128 1bp-deletion, 2,221 3bp-mismatch, 2,226 4bp-mismatch, 2,203 5bp-mismatch, 2,214 6bp-mismatch, 1,081 2bp-insertion, 1,631 2bp-deletion, 881 mix, 392 non-target and 1,378 on-target; each oligonucleotide chain contained left sequence (tgtggaaaggacgaaacacc), gRNA sequence (NNNNNNNNNNNNNNNNNNNN), BsmBI site (gttttgagacg), Barcode 1 (NNNNNNNNNNNNNNNNNNNN), BsmBI site (cgtctcgctcc), Barcode 2 (NNNNNNNNNNNNNN), target sequence (gtactNNNNNNNNNNNNNNNNNNNNgg), and right sequence

(cttggcgtaactagatct). The off-target library was constructed as follows: first, full-length oligonucleotides were PCR-amplified and cloned into BsmBI site of LentiGuide-U6-del-tracRNA vector by Gibson Assembly (NEB), named LentiGuide-U6-gRNA-target ; second, the tracRNA were PCR-amplified and cloned into BsmBI site of LentiGuide-U6-del-tracrRNA vector by T4 DNA ligase (NEB).The Gibson Assembly products or T4 ligation 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[12]. Briefly, for individual sgRNA packaging, 1.2 μg of gRNA expressing plasmid, 0.9 μg of psPAX2 and 0.3 μg of pMD2.G (Addgene) were transfected into HEK293T cells by Lipofectamine 2000 (Life Technologies). Medium were changed 8 hours after transfection. After 48h, virus supernatants were collected. For library packaging, 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.

Analysis of individual gRNA conversion efficiency

HEK293T cells expressing ABEmax or BE4max were seeded at ~40% confluency in a 24-well dish the day before infection. After the infection of gRNA-target paired lentivirus, genomic DNA was extracted at suitable time points using QuickExtract DNA Extraction Solution (Epicentre). We amplified the target sequence by PCR with Q5 High-Fidelity 2X Master Mix (NEB). PCR products were purified with Gel Extraction Kit (Qiagen) and Sanger-sequenced. According to ".ab1" sequencing file, the conversion efficiency was compared.

Screening experiments in HEK293T

HEK293T cells expressing ABEmax or BE4max were plated into 15 cm dish at ~30% confluence. After 24 h, cells were infected with 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 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 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-off-target editing activity. 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 if the target was edited. We screened out gRNAs with a total valid reading of less than 100. Then, the efficiency for a specific gRNA-off-target pair can be calculated by the following formula:

$$\text{editing efficiency} = \frac{\text{NO. of edited reads}}{NO.\,of\ total\ valided\ reads}$$

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 [29]. Unlike the common efficiency prediction that only needs to input one single sequence for regression models, in this research, the gRNA-off-target pair has two different sequences as inputs. For gRNA, there are four words in the index vocabulary (i.e., A, C, G, and T); but after alignment, there might exist DNA bulge or RNA bulge and thus lead to a "-" word to represent the insertion or deletion in the gRNA or off-target sequence. So, the vocabulary can be described as:

$$D = \{1{:}\,A, 2{:}\,C, 3{:}\,G, 4{:}\,T, 5{:}\,-\}$$

The input sequence can be described as:

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

where $i \in \{1,2\}$ denotes the i-th sequence in an gRNA-off-target pair, x_{it} is the t-th element of the i-th sequence, T is the sequence length. For example, ACGCTTCATCA-ATGTTGGGATGG (seq1, gRNA + NGG) and ACGC-TCATCAaAaGTT-GGATGG (seq2, off-target) can be encoded as:

$\boldsymbol{x_1} = [1, 2, 3, 2, 4, 4, 2, 1, 4, 2, 1, \mathbf{5}, 1, 4, 3, 4, 4, 3, 3, 3, 1, 4, 3, 3]$ (i.e., seq1)

and

$\boldsymbol{x_2} = [1, 2, 3, 2, \mathbf{5}, 4, 2, 1, 4, 2, 1, 1, 1, 1, 3, 4, 4, \mathbf{5}, 3, 3, 1, 4, 3, 3]$ (i.e., seq2),

respectively.

Shared embedding

Inspired by the algorithms in the recommender system [30] and click-through rate (CTR) [31] 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 $\mathbf{E_i} \in \mathbb{R}^{T \times m}$ (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 [32]. 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 network (CNN) to achieve good performances in various regression and classification tasks involving biological sequences [12, 33-35]. Here, we tried biLSTM, biGRU, and the newly proposed transformer structure [36], 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:

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

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

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

A max pooling and average pooling function were applied separately to $\mathbf{H}$ to obtain more useful features, and combine them with the final hidden state to produce the concatenated feature: $\boldsymbol{c} = [\boldsymbol{h}_{Last}; \mathbf{F}_{MaxPool}; \mathbf{F}_{MeanPool}]$, the predicted efficiency o for a specific gRNA-off-target pair can be obtained by:

$$o = \sigma\big(f(\boldsymbol{c})\big) \ (7)$$

where, f is fully connected layers, σ is the *sigmoid* activation function.

Considering that the sample sizes of off-target types vary greatly, in order to get better prediction results for the types with small sample sizes, we use the weighted MSE as the loss function.

Training setting

Off-target datasets were randomly split into two parts, one for training, and the other for holdout testing. The stability of model performance was estimated by a 10-fold shuffled validation together with the external integrated and endogenous datasets. 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[37]. PyTorch 1.6 [38] was used for building deep learning models. The global algorithm of pairwise2 in BioPython 1.7.8 were used for getting the alignment result of off-target sequence.

Data and code availability

The raw sequencing data have been submitted to the NCBI Sequence Read Archive ('SRA PRJNA587328 [https://dataview.ncbi.nlm.nih.gov/object/19730669]'). The other data for the present study is available from the corresponding author upon reasonable request.

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. *Nat Commun* **8**, 118 (2017).
7. Kim, K. et al. Highly efficient RNA-guided base editing in mouse embryos. *Nat Biotechnol* **35**, 435-437 (2017).
8. Liu, Z. et al. Highly efficient RNA-guided base editing in rabbit. *Nat Commun* **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. *Mol Ther* **26**, 2631-2637 (2018).
10. Liang, P. et al. Genome-wide profiling of adenine base editor specificity

by EndoV-seq. *Nat Commun* **10**, 67 (2019).
11. Kim, D. et al. Genome-wide target specificities of CRISPR RNA-guided programmable deaminases. *Nat Biotechnol* **35**, 475-480 (2017).
12. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat Commun* **10**, 4284 (2019).
13. Koblan, L.W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nature biotechnology* **36**, 843-846 (2018).
14. Mates, L. et al. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet* **41**, 753-761 (2009).
15. Wang, Y. et al. Regulated complex assembly safeguards the fidelity of Sleeping Beauty transposition. *Nucleic Acids Res* (2016).
16. Wang, Y. et al. Suicidal autointegration of sleeping beauty and piggyBac transposons in eukaryotic cells. *PLoS Genet* **10**, e1004103 (2014).
17. Kim, D., Kim, D.E., Lee, G., Cho, S.I. & Kim, J.S. Genome-wide target specificity of CRISPR RNA-guided adenine base editors. *Nat Biotechnol* **37**, 430-435 (2019).
18. Bae, S., Park, J. & Kim, J.S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473-1475 (2014).
19. 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. *Bioinformatics* **36**, 2001-2008 (2020).
20. Zhou, C. et al. Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis. *Nature* **571**, 275-278 (2019).
21. Zuo, E. et al. Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science* **364**, 289-292 (2019).
22. Jin, S. et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292-295 (2019).
23. Jin, S. et al. Rationally Designed APOBEC3B Cytosine Base Editors with Improved Specificity. *Mol Cell* **79**, 728-740 e726 (2020).
24. Hu, Z. et al. A compact Cas9 ortholog from Staphylococcus Auricularis (SauriCas9) expands the DNA targeting scope. *PLoS biology* **18**, e3000686 (2020).
25. 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).
26. Li, X. et al. Base editing with a Cpf1-cytidine deaminase fusion. *Nat Biotechnol* **36**, 324-327 (2018).
27. Huang, T.P. et al. Circularly permuted and PAM-modified Cas9 variants broaden the targeting scope of base editors. *Nat Biotechnol* **37**, 626-631

(2019).

28. Wang, X. et al. Efficient base editing in methylated regions with a human APOBEC3A-Cas9 fusion. *Nat Biotechnol* **36**, 946-949 (2018).
29. 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).
30. 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).
31. 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).
32. 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).
33. 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).
34. Effects on surface waters. *J Water Pollut Control Fed* **42**, 1084-1088 (1970).
35. Harris, A. Pacemaker 'heart sound'. *Br Heart J* **29**, 608-615 (1967).
36. Ashish Vaswani et al. in Proceedings of the 31st International Conference on Neural Information Processing Systems 6000–6010 (2017).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
38. Paszke, A. et al. in Advances in neural information processing systems 8026-8037 (2019).