nature biotechnology

BRIEF COMMUNICATION

https://doi.org/10.1038/s41587-020-0509-0

Base editors for simultaneous introduction of C-to-T and A-to-G mutations

Rina C. Sakata[1,2,13], Soh Ishiguro[1,3,4,13], Hideto Mori[1,3,4,13], Mamoru Tanaka[1], Kenji Tatsuno[5], Hiroki Ueda[6], Shogo Yamamoto[5], Motoaki Seki[1], Nanami Masuyama[1,3,4], Keiji Nishida[7,8], Hiroshi Nishimasu[9], Kazuharu Arakawa[3,4,10], Akihiko Kondo[7,8,11], Osamu Nureki[9], Masaru Tomita[1,3,4,10], Hiroyuki Aburatani[5] and Nozomu Yachie[1,2,3,4,9,12] ✉

We describe base editors that combine both cytosine and adenine base-editing functions. A codon-optimized fusion of the cytosine deaminase PmCDA1, the adenosine deaminase TadA and a Cas9 nickase (Target-ACEmax) showed a high median simultaneous C-to-T and A-to-G editing activity at 47 genomic targets. On-target as well as DNA and RNA off-target activities of Target-ACEmax were similar to those of existing single-function base editors.

CRISPR–Cas9 is a genome editing tool in which Cas9 is recruited by a guide RNA (gRNA) to its target DNA region upstream of a 3′ protospacer adjacent motif (PAM) to induce a double-stranded DNA break (DSB)[1,2]. This method has rapidly expanded our ability to knock genes out through error-prone DNA repair and insert transgenes into chromosomes through DSB-induced homologous recombination. In contrast, base editors derived by tethering deoxynucleoside deaminase to a nuclease-deficient or nickase Cas9 (dCas9 or nCas9, respectively)–gRNA complex induce efficient and direct base substitutions in the genomic sequence[3]. Among the available base editors, cytosine base editors (CBEs)[4,5] and adenine base editors (ABEs)[6] enable highly efficient and precise base substitutions in a narrow window of gRNA-targeting sites. CBEs consist of a cytidine deaminase (such as rAPOBEC1 used in base editors[5] and PmCDA1 used in Target-AID[4]) that converts cytidines of the non-gRNA bound DNA strand into uridines, and a uracil glycosylase inhibitor (UGI) that inhibits base excision repair, allowing uracils to be replaced with thymines through DNA replication. Similarly, ABEs use a heterodimer complex of WT and engineered TadA adenosine deaminases that convert adenines to inosines that are then replicated as guanines[6]. Currently, single-function base editors enable only two transition mutations, C·G→T·A and A·T→G·C, and have limited diversity of editing patterns that they can generate at a target site. A single base editor with both C→T and A→G base substitution activities would thus broaden the capabilities of base editing for various applications.

To this end, we developed and tested three dual-function base editors, Target-ACE, Target-ACEmax and ACBEmax, which have both cytidine deaminase and adenosine deaminase fused to a single nCas9 (D10A) (Extended Data Fig. 1). Target-ACE consists of nCas9 fused to PmCDA1 from Target-AID[4] and the TadA heterodimer from ABE7.10 (ref. [6]) at its C and N termini, respectively, along with other functional domains present in the constituent single-function base editors. Because GenScript codon optimization and the addition of an N-terminal bipartite nuclear localization signal (NLS) have previously led the development of enhanced base editors BE4max and ABEmax[7], we applied the same optimizations to Target-ACE to derive Target-ACEmax (Fig. 1a). ACBEmax was constructed by replacing the codon-optimized PmCDA1 domain of Target-ACEmax with the codon-optimized cytidine deaminase domain rAPOBEC1 from BE4max. As single-function base editor controls, in addition to Target-AID, BE4max, ABE and ABEmax, we constructed codon-optimized Target-AIDmax and BE4max(C) in which the C-terminal PmCDA1 domain of Target-AIDmax was replaced with the rAPOBEC1 domain from BE4max.

To test the base-editing activities of the single- and dual-function base editors in living cells, we constructed C→T and A→G base-editing reporter cells in which the corresponding base substitutions were designed to activate EGFP protein expression (Extended Data Fig. 2a,b). All single-function CBEs were only able to activate C→T reporter cells, whereas ABEs were only able to activate A→G reporter cells (Fig. 1b, Extended Data Fig. 2c and Supplementary Data 1). Both C→T and A→G reporter cells were activated by all three dual-function base editors, Target-ACE, Target-ACEmax and ACBEmax, as well as their corresponding enzyme mix controls, Target-AID+ABE, Target-AIDmax+ABEmax, BE4max(C)+ABEmax, and BE4max+ABEmax. These results were confirmed by amplicon sequencing of the gRNA target regions in the reporter cells (Extended Data Fig. 2d–g and Supplementary Data 2 and 3).

To characterize the C→T and A→G base-editing activities of different base-editing methods, we analyzed the base-editing spectra at 47 genomic target sites in human embryonic kidney (HEK293Ta) cells by amplicon sequencing in triplicate (1,833 assays) (Supplementary Data 4). By taking the average C→T or A→G editing frequencies at each cytosine or adenine position relative to the

[1]Synthetic Biology Division, Research Center for Advanced Science and Technology, University of Tokyo, Tokyo, Japan. [2]College of Arts and Sciences, University of Tokyo, Tokyo, Japan. [3]Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan. [4]Graduate School of Media and Governance, Keio University, Fujisawa, Japan. [5]Genome Science Division, Research Center for Advanced Science and Technology, University of Tokyo, Tokyo, Japan. [6]Biological Data Science Division, Research Center for Advanced Science and Technology, University of Tokyo, Tokyo, Japan. [7]Engineering Biology Research Center, Kobe University, Kobe, Japan. [8]Graduate School of Science, Technology and Innovation, Kobe University, Kobe, Japan. [9]Department of Biological Sciences, School of Science, University of Tokyo, Tokyo, Japan. [10]Faculty of Environment and Information Studies, Keio University, Fujisawa, Japan. [11]Department of Chemical Science and Engineering, Graduate School of Engineering, Kobe University, Kobe, Japan. [12]PRESTO, Japan Science and Technology Agency (JST), Tokyo, Japan. [13]These authors contributed equally: Rina C. Sakata, Soh Ishiguro, Hideto Mori
✉e-mail: yachie@synbiol.rcast.u-tokyo.ac.jp

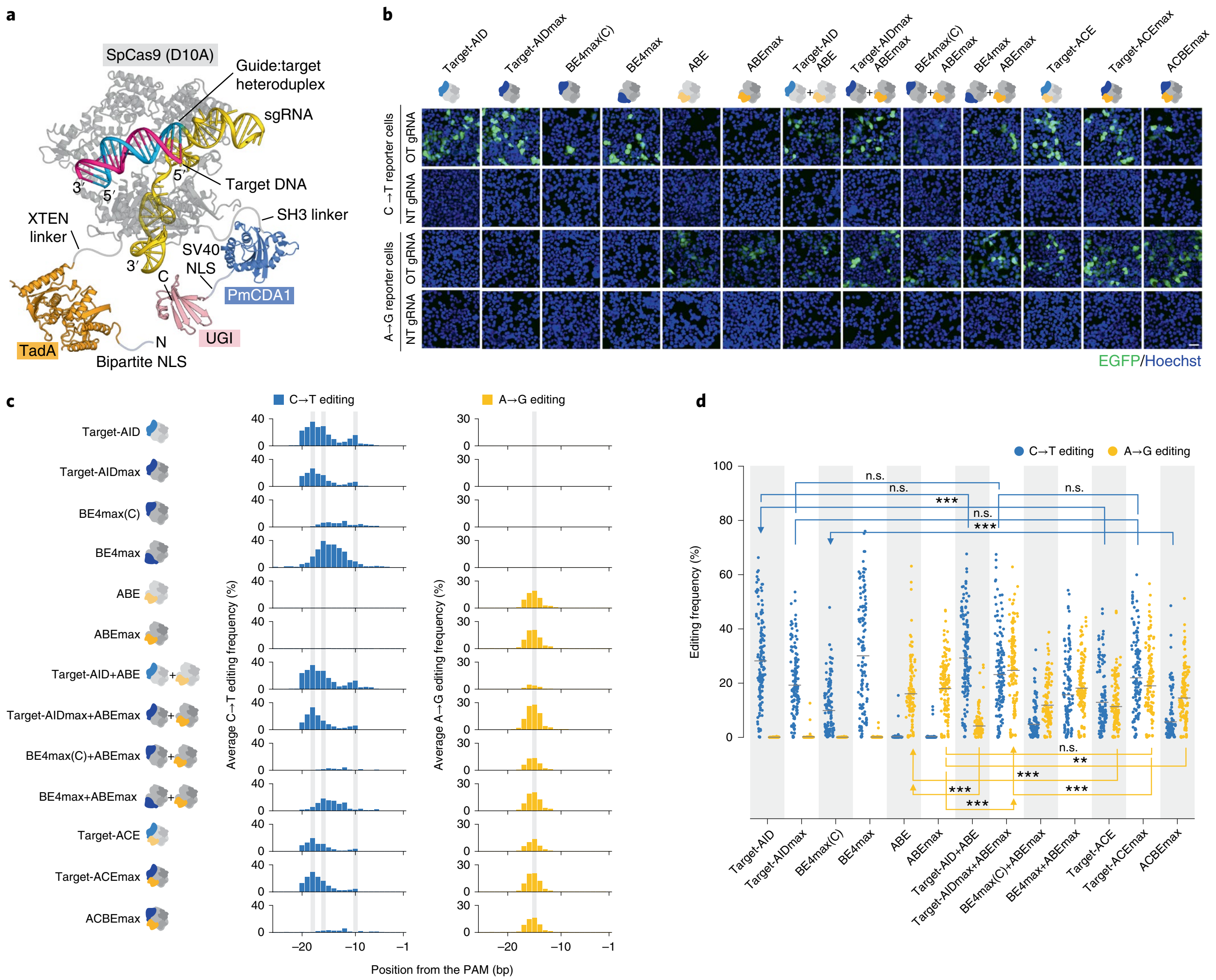


Fig. 1 | C→T and A→G base-editing activities of single- and dual-function base editors. a, Structural overview of Target-ACEmax. **b**, Microscopy images of C→T and A→G base-editing reporter cells transiently transfected with different base editor reagents and corresponding on-target (OT) or non-targeting (NT) gRNAs. Scale bar, 40 µm. Consistent results were obtained independently for four cell culture replicates. **c**, Average C→T and A→G base-editing spectra of a genomic target site. Vertical grey lines indicate the peak editing positions of the different base-editing conditions. **d**, C→T and A→G editing frequencies of 47 different endogenous target sites in the human genome. Horizontal black bars represent the average C→T or A→G editing frequencies. Two-sided Mann–Whitney's *U*-test was performed to compare arbitrary pairs of two datasets from three cell culture replicates. Arrowheads indicate datasets with a higher average editing frequency than the other in the pair (*$P<0.05$; **$P<0.01$; ***$P<0.001$).

PAM, we found that dual-function base editors commonly inherited similar base-editing characteristics from their corresponding single-function base editors (Fig. 1c and Supplementary Data 5). The C→T and A→G editing frequencies of 47 endogenous target sites containing different numbers of cytosines and adenines varied widely, but were overall consistent with the base-editing spectrum data (Fig. 1d and Supplementary Data 6).

In the amplicon sequencing data of 47 genomic target sites, 722 editing outcome patterns with both C→T and A→G edits were observed at read frequencies of 0.1% or more by the dual-function base editors and four combinations of base editor mix controls (Fig. 2a). Overall, we found that Target-ACEmax, Target-ACE and their corresponding enzyme mixes displayed a cluster of multi-base-editing patterns distinct from those of ACBEmax and its corresponding enzyme mix. To observe the co-editing patterns in the 47 different target sites, we next investigated dinucleotide homologous and heterologous co-editing spectra for each of the 13 base editor conditions. For each cytosine–cytosine, adenine–adenine or cytosine–adenine pair in different positions relative to the PAM, the average co-editing frequency was calculated using amplicon sequencing results of the relevant target sites (Supplementary Fig. 1). Among the different base editor reagents, the multiple C→T and A→G edit spectra recaptured the same trends of the average C→T and A→G frequencies of the 47 target sites (Supplementary Fig. 2 and Supplementary Data 7). For simultaneous C→T and A→G editing, Target-ACEmax and Target-AIDmax+ABEmax showed similarly efficient co-editing spectra around cytosine at −18 bp and adenine at −15 bp relative to the PAM with peak heights of 19.2% and 21.0%, respectively (Fig. 2b,c and Supplementary Fig. 3). ACBEmax and BE4max+ABEmax demonstrated another efficient co-editing spectrum: their peak frequencies at cytosine–adenine positional combinations at −12 and −15 bp were the two highest among all of the base editor reagents (23.2% and 23.6%, respectively).

To examine DNA off-target effects of the different base-editing methods caused by non-specific binding of the gRNA, we used three gRNAs and performed amplicon sequencing of their on-target and commonly observed off-target sites[8,9] (Extended Data Fig. 3). Based on the amplicon sequencing data, the off-target risk scores were calculated for the different base-editing methods (Fig. 2d and Supplementary Data 8). The off-target risk scores of dual-function base editors were within the range of those for the single-function base editors, whereas the off-target risk scores of base editor mix controls were markedly higher than those of their corresponding dual-function base editors. We also extensively surveyed genome-wide DNA and RNA off-target activities of the 13 base editor conditions by whole exome sequencing (WES) and transcriptome sequencing (RNA-seq) (Supplementary Table 1). While no elevated level of single-nucleotide variation (SNV) induction was detected in the WES dataset (Supplementary Fig. 4), the numbers of C→U RNA edits found using rAPOBEC1-related, but not PmCDA1-related, base-editing methods were strikingly higher (an average of 3,403) than those of the other samples (322), which was consistent with previous reports[10,11] (Fig. 2e and Supplementary Fig. 5). The non-specific A→I RNA-editing activities were overall significantly higher using ABEmax and base editor mixes containing ABE or ABEmax compared with the other methods ($P=0.00019$), as reported previously[10,12,13]. Notably, non-specific A→I editing activity of Target-ACEmax (an average of 3,359 across two replicates) was relatively lower than that of Target-AIDmax+ABEmax (average of 4,179).

Similar to the recent machine learning approaches to predict WT Cas9-mediated genome editing outcomes[14–16], we developed a base-editing prediction method that trains amplicon sequencing data and predicts base-editing patterns and their frequencies for a given target sequence (Extended Data Fig. 4a). In brief, we found that our method successfully predicted base-editing outcomes of untrained targets with Pearson's correlation coefficients of 0.70 and 0.71 for Target-ACEmax and ACBEmax, respectively (Extended Data Fig. 4b). We also demonstrated that the amplicon sequencing data obtained in this study were sufficient for the training procedure (Supplementary Fig. 6). Because the machine learning method enabled prediction of multi-nucleotide co-editing (Extended Data Fig. 5 and Supplementary Figs. 7 and 8), we used it to predict the frequencies of all possible codon conversion patterns in the human genome obtained by the different base-editing methods. When bystander mutations were not allowed to occur, this analysis showed that Target-ACEmax and its corresponding base editor mix Target-AIDmax+ABEmax had the highest potentials for diversifying genomic codons (Extended Data Figs. 6 and 7, Supplementary Fig. 9 and Supplementary Data 9). We then repeated the same analysis by allowing bystander mutations to occur (Extended Data Figs. 8 and 9) and estimated bystander risks of generating unwanted mutations for all of the base-editing methods (Fig. 2f). The bystander mutation risks of Target-ACEmax and ACBEmax were within the risk range of commonly used single-function base editors, of which BE4max was the highest. Although the bystander mutation risk scores of ACBEmax were significantly lower than those of Target-ACEmax, this was largely consistent with ACBEmax not showing marked improvement in expanding genome-wide codon conversion patterns. Finally, our model predicted that Target-ACEmax had the highest potentials to correct pairs of heterologous disease mutations reported in the ClinVar database[17] as a single base editor enzyme (Supplementary Fig. 10).

Target-ACEmax and ACBEmax both showed high heterologous co-editing efficiencies at their favorable positional combinations upstream of the PAM. While the dual-function base editors were evaluated using HEK293Ta cells in this study and remain to be tested in other systems, they could complementarily serve as genome editing tools to induce heterologous multiple base edits in various settings including therapeutics with the advantages of high delivery efficiency considering their compact gene sizes (Supplementary Note 1). CRISPR–X[18] involving a mutant human AID has been developed as a sequence diversification tool to induce C→A/G/T substitutions at a multiple hundred base pair region surrounding a gRNA target site, but showed low editing efficiency per position. Target-ACEmax induces a higher efficiency of heterologous base editing than CRISPR–X. Thus, Target-ACEmax could be applied as a complementary tool for in vivo diversification of targeted sequences for mutational scanning analysis of protein functions and directed protein evolution as examples.

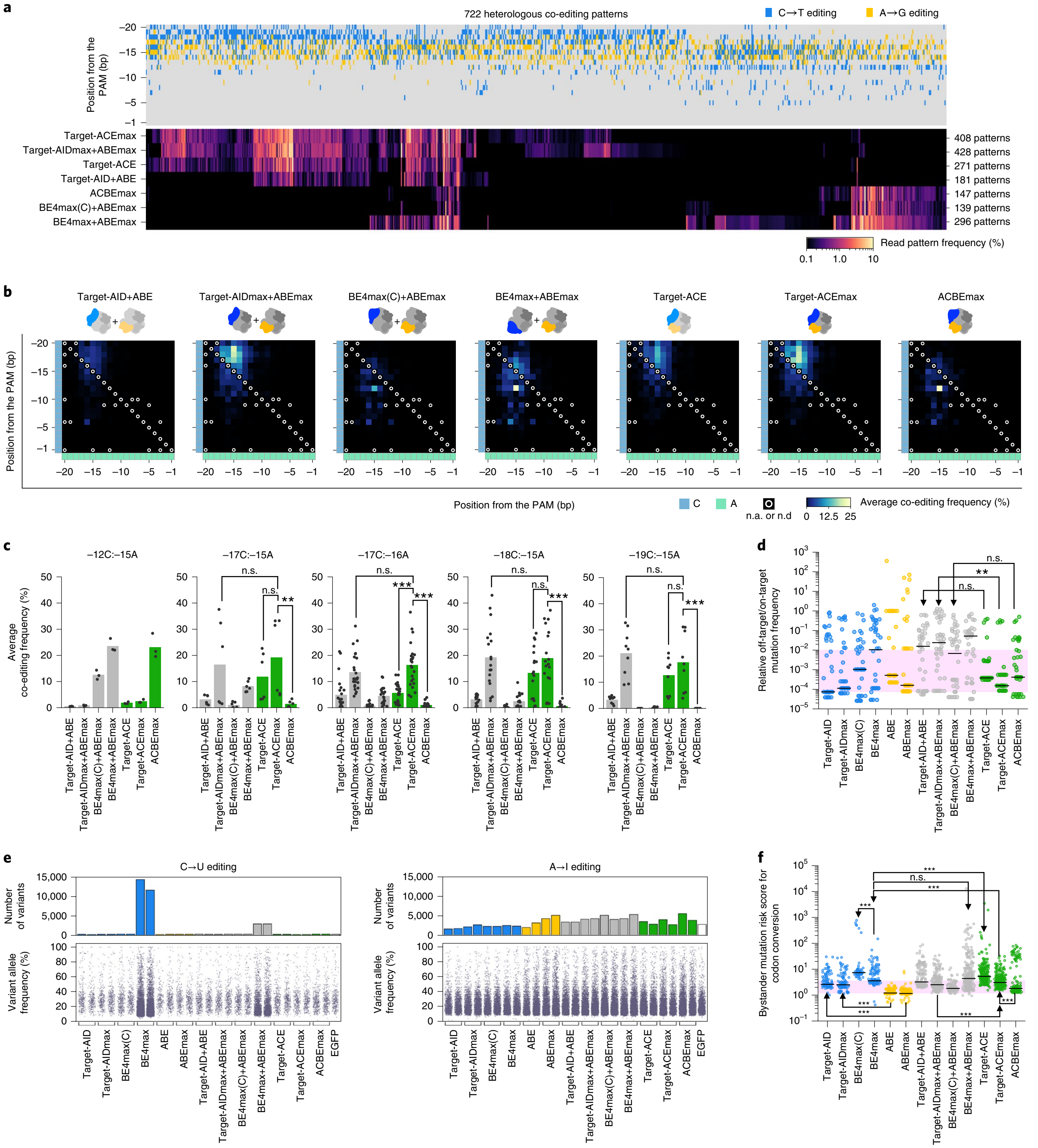
Target-ACEmax could also be a powerful tool for recent cell lineage tracing using CRISPR genome editing[19,20]. Most of the current implementations employ WT Cas9 that induces DSBs. These DSBs result in cytotoxicity and rapidly saturate the mutation patterns in DNA barcodes following target site deletions, which theoretically limit the resolution of cell lineage reconstruction[21]. While base editors could minimize these detrimental effects caused by DSBs, unidirectional

Fig. 2 | Simultaneous C→T and A→G base-editing activities and unwanted mutation risks of dual-function base editors. a, Editing outcome patterns with both C→T and A→G edits. A total of 722 heterologous co-editing patterns obtained by either base-editing condition with a read frequency threshold of 0.1% (top) were hierarchically clustered based on the frequency produced by the different base-editing conditions (bottom). The total number of editing outcome patterns observed in three independent replicate experiments are shown on the right for each base-editing condition. **b**, Average frequencies of simultaneous C→T and A→G editing at different positional combinations in the −20 to −1 bp region upstream of the PAM. **c**, Co-editing frequencies of dual-function base editors and single-function enzyme mixes for cytosines and adenines located at specific combinatorial positions relative to the PAM. Positional combinations with an average co-editing frequency within the top five ranked for any one base-editing condition are shown. Each bar shows the average co-editing frequencies measured for target sites with cytosine and adenine in the respective combinatorial positions in three cell culture replicates (dots). Two-sided Mann–Whitney's *U*-test was performed to compare Target-ACEmax with its corresponding enzyme mix and the other two dual-function base editors for positional combinations with sufficient sample sizes (*$P<0.05$; **$P<0.01$; ***$P<0.001$). **d**, DNA off-target risk estimation. The jitter plot shows relative off-/on-target editing frequencies measured for three gRNAs in three cell culture replicates experiments and horizontal black bars represent medians as their DNA off-target risk scores. The pink-shaded area shows the score range for single-function base editors. Two-sided Welch's *t*-test was performed to compare the dual-function base editors with their corresponding single-function enzyme mixes. Arrowhead indicates a dataset with a higher average score than the other (**$P<0.01$). **e**, Genome-wide C→U and A→I RNA variants detected in cells that were subjected to different base-editing conditions. Each bar shows the number of variants identified by RNA-seq and the jitter plot shows their variant allele frequencies. Experiments were performed in duplicate for each base-editing condition. **f**, Bystander mutation risk scores of codon conversion for the different base-editing methods. Horizontal bars represent the median of risk scores for different codon conversion types represented by dots. The pink shaded area shows the range of average risk scores for the single-function base editors excluding BE4max(C). Two-sided Mann–Whitney's *U*-test was performed to compare arbitrary pairs of two datasets. Arrowheads indicate the dataset with a higher average score than the other in the pair (***$P<0.001$).

mutations induced by single-function base editors would also cause saturation in the diversity of mutated DNA barcode patterns. In contrast, Target-ACEmax could alleviate this saturation issue because of its reversible C·G↔T·A activity and contribute to high-resolution cell lineage tracing. Dual-function base editors with their expanded base conversion ability and further improvements would have the potential to promote development in therapeutics and biotechnology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0509-0.



Received: 7 August 2019; Accepted: 3 April 2020;
Published online: 01 June 2020

References

1. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
2. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
3. Rees, H. A. & Liu, D. R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **19**, 770–788 (2018).
4. Nishida, K. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353**, aaf8729 (2016).
5. 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).
6. 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).
7. Koblan, L. W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).
8. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
9. Kleinstiver, B. P. et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
10. Grunewald, J. et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **569**, 433–437 (2019).
11. Grunewald, J. et al. CRISPR DNA base editors with reduced RNA off-target and self-editing activities. *Nat. Biotechnol.* **37**, 1041–1048 (2019).
12. Zhou, C. et al. Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis. *Nature* **571**, 275–278 (2019).
13. Rees, H. A., Wilson, C., Doman, J. L. & Liu, D. R. Analysis and minimization of cellular RNA editing by DNA adenine base editors. *Sci. Adv.* **5**, eaax5717 (2019).
14. Shen, M. W. et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646–651 (2018).
15. Allen, F. et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.* **37**, 64–72 (2019).
16. Chen, W. et al. Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Res.* **47**, gkz487 (2019).
17. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
18. Hess, G. T. et al. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods* **13**, 1036–1042 (2016).
19. Masuyama, N., Mori, H. & Yachie, N. DNA barcodes evolve for high-resolution cell lineage tracing. *Curr. Opin. Chem. Biol.* **52**, 63–71 (2019).
20. Woodworth, M. B., Girskis, K. M. & Walsh, C. A. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017).
21. Salvador-Martinez, I., Grillo, M., Averof, M. & Telford, M. J. Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *eLife* **8**, e40292 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Oligonucleotides. The oligonucleotides used in this study are listed in Supplementary Table 2.

Plasmid preparation. *Base editor expression plasmids.* All base editor expression plasmids were prepared with the same backbone sequence used in pCMV-BE3 (Addgene 73021). BE4 (pCMV-BE4), BE4max (pCMV-BE4max), ABE7.10 (pCMV-ABE7.10), and ABEmax (pCMV-ABEmax) plasmids with the same backbone were obtained from Addgene (100802, 112093, 102919, and 112095, respectively) and the other single- and dual-function base editors were constructed by Gibson assembly of PCR fragments as follows. The Target-AID plasmid (pCMV-Target-AID) was constructed by assembling two fragments encoding the N- and C-terminus halves of Target-AID, which were both amplified from pcDNA3.1_pCMV-nCas-PmCDA1-ugi pH1-gRNA(HPRT) (Addgene 79620) using primer pairs RS045/HM129 and HM128/RS046, respectively, with a backbone fragment amplified from pCMV-ABE7.10 using RS047/RS048. To construct the Target-AIDmax plasmid (pCMV-Target-AIDmax), the pUC-optimized-PmCDA1-ugi plasmid encoding the codon-optimized C-terminal region of Target-AIDmax was first constructed by the gene synthesis service of GenScript. This C terminus fragment was then amplified with primer pair SI1304/SI1307 and assembled with a nCas9 fragment amplified from pCMV-BE4max using SI945/SI1308 and a backbone fragment amplified from pCMV-ABEmax using SI1310/SI1309. The BE4max(C) plasmid (pCMV-BE4max(C)) was constructed to replace the C-terminal region of Target-AIDmax with the codon-optimized rAPOBEC1 and 2×UGI domains of BE4max. To this end, an nCas9 fragment obtained from pCMV-Target-AIDmax using SI447/SI1105 was assembled with rAPOBEC1 and 2×UGI fragments obtained from BE4max using SI1352/SI1357 and SI1359/SI1350, respectively, and a backbone obtained from pCMV-BE4max using SI1351/SI448. The Target-ACE plasmid (pCMV-Target-ACE) was constructed with a fragment encoding a plasmid backbone as well as ABE7.10 amplified from pCMV-ABE7.10 using RS047/RS052 and a fragment encoding the C-terminus region of Target-AID amplified from pcDNA-pCMV-nCas9 using RS051/RS046. The Target-ACEmax plasmid (pCMV-Target-ACEmax) was constructed by assembling an ABEmax fragment obtained from pCMV-ABEmax using SI945/SI1305, a fragment encoding the C-terminus region of Target-AIDmax obtained from pUC-optimized-PmCDA1-ugi using SI1304/SI1307, and a plasmid backbone obtained from pCMV-ABEmax using SI1310/SI1309. Finally, the ACBEmax plasmid (pCMV-ACBEmax) was constructed by assembling an ABEmax fragment obtained from pCMV-Target-ACEmax using SI447/SI1105 with the three fragments encoding the rAPOBEC1 domain, 2×UGI domain, and the two backbone fragments that were prepared to construct pCMV-BE4max(C). Note that the 2×UGI domain used in this study had a non-synonymous nucleotide substitution in the GS linker between the tandem UGIs (SGGSG[G > E]SGGS).

gRNA expression plasmids. gRNA spacer inserts were prepared by a single pot reaction to phosphorylate and anneal ssDNA pairs. To prepare each spacer fragment (Supplementary Table 2), a T4 polynucleotide kinase reaction sample was prepared with two ssDNAs in accordance with the manufacturer's protocol (Takara) and placed in a thermal cycler with the following conditions: 37 °C for 30 min; 95 °C for 5 min; 70 cycles of 12 s starting with 95 °C and −1 °C per cycle, and then maintained at 25 °C. The annealed spacer inserts were then ligated into a pU6-gRNA cloning backbone (pSI-356) by Golden Gate Assembly using BsmBI (NEB) and T4 DNA ligase (NEB). The assembly was performed under the following thermal cycler conditions: 15 cycles of 37 °C for 5 min and 20 °C for 5 min, 55 °C for 30 min, and then maintained at 4 °C.

Lentiviral base-editing reporter plasmids. The C→T reporter circuit was designed to restore a mutated GTG start codon to ATG by C→T base editing of its antisense strand (Extended Data Fig. 2a). In the A→G reporter circuit, EGFP translation was designed to be released by destruction of a TAA stop codon to CAA by A→G base editing of the antisense strand (Extended Data Fig. 2b). To construct the lentiviral C→T base-editing reporter plasmid (pLV-SI-112) and its positive control plasmid (pRS112), reporter cassette fragments were amplified from pLV-eGFP (Addgene 36083) using primer sets 112-V4-BC2-FW/SI680 and RS204/SI627, respectively, and cloned into the EcoRI and BamHI sites of the pLVSIN-CMV-Puro backbone vector (Takara) using T4 DNA ligase (NEB). The lentiviral A→G base-editing reporter plasmid (pLV-SI-121) and its positive control plasmid (pLV-SI-122) were constructed similarly, where reporter cassettes were amplified using primer sets SI760/SI680 and SI761/SI680, respectively.

The plasmid sequences were confirmed by Sanger sequencing. Other than the gRNA plasmids for genomic target sites, we submitted the newly constructed plasmids to Addgene. The list of plasmids used in this study is shown in Supplementary Table 3.

Selection of gRNA target sites. To select gRNA target sites with cytosines and adenines for amplicon sequencing assays, we first searched for target sites with poly-cytosine repeats (poly-C), poly-adenine repeats (poly-A), alternating poly-adenine/cytosine repeats (poly-AC), and alternating poly-cytosine/adenine repeats (poly-CA) in the human genome (hg19). Poly-C target sites were required to have cytosines filling a 7-bp sliding window for which the 5′ end shifted at intervals of 2 bp from −24 bp to −16 bp relative to the PAM. Poly-A target sites were required to have adenines filling a 6-bp sliding window for which the 5′ end shifted at intervals of 2 bp from −21 bp to −13 bp relative to the PAM. Poly-AC and poly-CA target sites were required to have the corresponding patterns in a 6-bp sliding window for which the 5′ end shifted at intervals of 2 bp from −24 bp to −14 bp relative to PAM. The candidate target sites that contained a homopolymer of ≥4 bp long within a gRNA seed region spanning from −8 bp to −1 bp to the PAM and those overlapping with annotated exons were excluded. For each sliding window position of poly-C and poly-A, we selected two candidate sites each with the highest predicted gRNA activity scores[22]. Similarly, for each sliding window position of poly-AC and poly-CA, we selected one candidate site each. Using amplicon sequencing primers designed for these target sites, we further screened seven poly-C, seven poly-A, six poly-AC, and four poly-CA target sites that were robustly amplified by our amplicon sequencing library preparation protocol. Note that poly-C and poly-A target sites that were screened also contained both cytosine and adenine bases. In addition to these 24 target sites, we screened 24 target sites from our gRNA library collections that we previously prepared for other assays. Each of these target sites was required to contain one or more cytosines and one or more adenines in a region spanning from –20 bp to –14 bp. We failed to obtain the EGFP control data for the amplicon sequencing assays of CUL3-NGG-site 2 and therefore analyzed a total of 47 on-target sites. For off-target site analysis by amplicon sequencing, we selected two on-target sites for *FANCF* and one on-target site for *EMX1* genes which had both cytosine and adenine bases in the region spanning from –20 bp to –14 bp and many off-target sites that were commonly identified by GUIDE-seq in multiple previous studies[8,9]. While we selected five of the previously reported off-target sites for each on-target site, one of the *EMX1* off-target sites was omitted from the analysis because it was not amplified by our amplicon sequencing library preparation protocol. All on-target and off-target sites analyzed by amplicon sequencing assays are shown in Supplementary Table 2.

Cell culture. HEK293Ta cells were purchased from GeneCopoeia and maintained in Dulbecco's Modified Eagle's Medium (DMEM) (Sigma) supplemented with 10% fetal bovine serum (FBS) (Thermo Fisher Scientific) and 1% penicillin–streptomycin (Sigma) at 37 °C with 5% CO_2. Cells were routinely tested for mycoplasma contamination by nested PCR using culture medium as a template.

Base-editing reporter cell lines. The C→T and A→G reporter circuits and their respective positive control circuits, containing the expected mutations, were introduced into human embryonic kidney HEK293Ta cells by lentiviral transduction. For lentiviral packaging, ~2×10^5 cells per well were seeded in a 6-well plate 1 day before transfection. In each packaging reaction, 489 ng of lentiviral plasmid was cotransfected with two helper plasmids, psPAX2 (Addgene 12260) and pMD2.G (Addgene 12259), at 366 ng and 122 ng, respectively, and 9.38 µl of 1 mg ml^{-1} polyethyleneimine MAX (PEI) (Polysciences) in 300 µl phosphate-buffered saline (PBS). The next day, the culture medium was changed to fresh medium, and 2 days later the culture supernatant containing lentiviral particles was harvested and aliquoted into 1.5-ml tubes. The viral sample was then stored at −80 °C before infection. For lentiviral infection, ~2×10^5 cells per well were seeded on a 6-well plate with 2 ml DMEM and incubated for 24 h. The viral supernatant was then thawed at room temperature, mixed with 1 µl of 8 mg ml^{-1} polybrene (Sigma), and added to each cell sample. One day after infection, ~5×10^3 infected cells were reseeded on a 96-well culture plate for functional titer measurement by the CellTiter-Glo assay (Promega). Two days after infection, 2.0 µg ml^{-1} puromycin (Thermo Fisher Scientific) was added to the culture medium, followed by incubation for 3 days to select successfully transduced cells. After puromycin selection, absence of background fluorescence or EGFP expression was confirmed for the reporter cell lines and their corresponding positive control cell lines, respectively.

Transfection. *EGFP reporter activation assays.* Base-editing reporter cells and corresponding control cells were seeded in a collagen-I-coated 24-well plate (Asone) with 500 µl DMEM at a density of ~5×10^4 cells per well. The next day, 1.2 µl of 1 mg ml^{-1} PEI, 50 µl PBS, 300 ng base editor plasmid, and 100 ng gRNA expression plasmid were mixed and then incubated at room temperature for 20 min before application to each well for transfection. For each base editor mix experiment, two base editor plasmids were mixed at a 1:1 mass ratio and 300 ng was used (see Supplementary Note 2 for detail discussion about of the base editor mix experiments). Fluorescence imaging was performed 3 days after transfection using a confocal microscopy InCellAnalyzer6000 (GE Healthcare) with a ×20 objective lens. The experiment was performed in quadruplicates. For one of the replicates, cell nuclei were stained with 10 mg ml^{-1} Hoechst 33342 (Thermo Fisher Scientific).

Genomic on-target and off-target assays. HEK293Ta cells were seeded in a collagen-I-coated 96-well plate (Asone) with 200 µl DMEM at a density of ~5×10^3 cells per well. The next day, 0.48 µl PEI, 50 µl 1× PBS, 120 ng base editor expression or control EGFP expression (pLV-eGFP) plasmid, and 40 ng gRNA expression plasmid were mixed and then incubated at room temperature for 15 min before application to each well for transfection. The experiments were performed as three independent replicates.

Genome-wide DNA and RNA off-target assays. HEK293Ta cells were seeded in a collagen-I-coated 6-well plate (Asone) with 2 ml DMEM at a density of ~2×10^5 cells per well. The next day, 3.0 µl of 1 mg ml^{-1} PEI, 200 µl 1× PBS, 666 ng base editor expression or control EGFP expression (pLV-eGFP) plasmid, and 333 ng EMX1-targeting gRNA plasmid were mixed and then transferred to each well after 15 min of incubation at room temperature. The experiments were performed as two independent replicates.

Amplicon sequencing. *EGFP reporter activation assays.* After confocal imaging, the culture medium was aspirated and 200 µl of freshly prepared 50 mM NaOH was added to each cell sample in a 24-well plate. Then, 100 µl of the sample was transferred to a 96-well PCR plate (Nippon Genetics) for cell lysis and heated at 95 °C for 15 min and cooled to 4 °C, followed by neutralization with 20 µl of 1 M Tris-HCl (pH 8.0). From each sample, the target regions were amplified using the cell lysate as the PCR template with its corresponding first HTS primer pair (Supplementary Table 2). The PCR was performed in a 20 µl volume including 1 µl template, 1 µl of 10 µM each primer, 0.2 µl Phusion DNA Polymerase, 5× Phusion HF Buffer (NEB), and 1.6 µl of 2.5 mM dNTPs with the following thermal cycler conditions: 98 °C for 30 s, 30 cycles of 98 °C for 10 s, 60 °C for 10 s, and 72 °C for 10 s, and then 72 °C for 5 min for final extension. The PCR product was then electrophoresed on a 2% agarose gel and reamplified using 1 µl of a 10-fold dilution as the template and custom Illumina index primers (Supplementary Tables 2 and 4) in a 20 µl volume reaction with the following thermal cycler conditions: 98 °C for 30 s, 15 cycles of 98 °C for 10 s, 65 °C for 10 s, and 72 °C for 30 s, and then 72 °C for 5 min for final extension. Each indexed library was electrophoresed in a 2% agarose gel and the expected band was extracted using FastGene Gel/PCR Extraction Kit (Nippon Genetics).

Genomic on-target and off-target assays. Three days after transfection, the culture medium was removed and 50 µl of freshly prepared 50 mM NaOH was added to each cell sample in a 96-well plate. The samples were transferred to a 96-well qPCR plate (BioRad), sealed with an optically clear adhesive PCR seal (BioRad), and centrifuged at 2,400 rpm for 2 min, heated at 95 °C for 15 min, and cooled to 4 °C, followed by neutralization with 5 µl of 1 M Tris-HCl (pH 8.0). The cell lysate plates were centrifuged again and stored at −20 °C. Each target region was amplified with its corresponding first HTS primer pair (Supplementary Table 1). The PCR was performed in a 20 µl volume including 2 µl genomic DNA template, 1.20 µl of 8.3 µM each primer, 0.2 µl Phusion DNA Polymerase, 5× Phusion HF Buffer, and 1.6 µl of 2.5 mM dNTPs with the following thermal cycler conditions: 98 °C for 30 s, 30 cycles of 98 °C for 10 s, 60 °C for 10 s, and 72 °C for 60 s, and then 72 °C for 5 min for final extension. For each replicate experiment, 3 µl of each PCR product of the same base editor reagent was pooled and purified using a 1.8× volume of Agencourt AMPure XP magnetic beads (Beckman Coulter). The purified product was reamplified using 1 µl of 10 ng µl^{-1} of the first PCR product as the template and custom Illumina index primers (Supplementary Table 4) in a 20 µl volume reaction with the following thermal cycler conditions: 98 °C for 30 s, 15 cycles of 98 °C for 10 s, 65 °C for 10 s, and 72 °C for 90 s, and then 72 °C for 5 min for final extension. Each indexed library was electrophoresed on a 2% agarose gel and the expected band was extracted using FastGene Gel/PCR Extraction Kit.

The sequencing libraries were quantified by qPCR using KAPA Library Quantification Kit Illumina (KAPA Biosystems) for multiplexing. The multiplexed libraries were quantified by the same qPCR protocol and sequenced with 20–30% PhiX control using Illumina HiSeq2500 (TruSeq rapid SBS kit; 2 × 151 bp paired end) or MiSeq (MiSeq v3 kit; 2 × 200 bp paired end).

Preparation of RNA-seq and WES libraries. Cells were trypsinized 3 days after transfection and divided into two 1.5-ml tubes for transcriptome sequencing (RNA-seq) and whole exome sequencing (WES). For RNA-seq library preparation, cells were centrifuged at 1,000 rpm for 5 min and the culture supernatant was removed. Total RNA was then extracted by ISOSPIN Cell & Tissue RNA (Nippon Gene) and an RNA-seq library was prepared with TruSeq Stranded mRNA Library Prep Kit (Illumina). For WES library preparation, genomic DNA was extracted using NucleoSpin Tissue (Macherey Nagel). We then sheared 500 ng of genomic DNA in a 50 µl volume to an average size of 150–300 bp using Covalis E-220, followed by end-repair, A-tailing, and SureSelect adapter ligation (Agilent). The adapter-ligated DNA was enriched using KAPA HyperPrep kit (KAPA Biosystems). We then hybridized 750 ng of the pre-amplified DNA to SureSelectXT Human All Exon V3 kit probes (Agilent) for 20 h. Post-capture DNA library amplification was then performed using KAPA DNA Polymerase and SureSelect Indexing Post-Capture Polymerase Chain Reaction Primers for library indexing. The library was finally purified with Agencourt AMPure XP beads. The library index information of RNA-seq and WES is shown in Supplementary Table 4. The fragment size distributions and yields of the RNA-seq and WES libraries were quantified using the LabChip GX electrophoresis system (Perkin Elmer). After multiplexing, the final library was sequenced by Illumina NovaSeq 6000 (S2 Flow Cells; 2 × 101 bp paired end).

Amplicon sequencing analysis. Common adapter sequences were first mapped onto the amplicon sequencing reads using NCBI BLAST+ (version 2.7.0)[23] with the blastn-short option to identify custom sample indices and demultiplex paired-end reads. The paired-end reads of each sample were then merged using FLASH (version 1.2.0)[24] to generate merged sequencing reads that were further mapped to the corresponding reference sequence of the target region using EMBOSS needle package (version 6.6.0)[25] with an identity threshold of 80%. For single and multiple editing spectrum analyses using the genomic on-target and off-target amplicon sequencing data, the EGFP transfection control data were always treated in the same manner to subtract background signals.

RNA and DNA variant calling pipelines. RNA-seq and WES base call files were demultiplexed using bcl2fastq2 (version v2.20.0). To eliminate sequencing coverage bias for variant calling, all RNA-seq and WES libraries were randomly subsampled to 74 million and 94 million reads per sample, respectively, using seqtk (version 1.3-r107-dirty) as described previously[11]. The subsampled RNA-seq reads were then mapped to the reference human genome GRCh38.d1.vd1 using STAR (version 2.7.3a)[26] with a transcript annotation GTF (GENCODE Release 22 GRCh38.p2) and deduplicated using Picard MarkDuplicates (version 2.0.1). The subsampled DNA reads were also processed for mapping according to the National Cancer Institute Genomic Data Commons DNA-Seq analysis pipeline. In brief, the reads were aligned by BWA-MEM (version 0.7.15)[27] with the reference human genome GRCh38.d1.vd1 and PCR duplicates were removed using Picard MarkDuplicates (version 2.0.1). GATK HaplotypeCaller (version 4.1.4.1)[28] was used to call both DNA and RNA variants for the reference human genome GRCh38.d1.vd1. The mutation positions called by GATK HaplotypeCaller were further filtered as described previously[11].

DNA off-target risk estimation. To examine the DNA off-target effects of the different base-editing methods caused by non-specific binding of targeting gRNA, we used three gRNAs targeting *EMX1* and *FANCF* genes, whose off-target sites are commonly observed by GUIDE-seq[8]. After transfection of base editors and on-target gRNA reagents into HEK293Ta cells, we analyzed the base-editing efficiencies of five off-target sites for the *FANCF* target site 1, five for the *FANCF* target site 2, and four for an *EMX1* target site by amplicon sequencing (Extended Data Fig. 3 and Supplementary Data 8). Among all tested off-target sites, the median of their relative off-target activities normalized to the corresponding on-target activities were calculated as the DNA off-targeting risk score for each base-editing method.

Base-editing prediction model. *Model training.* Amplicon sequencing datasets of different target sites were used to train the conditional probability model. To minimize the effects of potential sequencing errors in the training procedure, observed editing outcomes with relative read frequencies of less than 1×10^{-4} were first eliminated from the dataset. Let s_i be the nucleotide base transition status at i bp position relative to the PAM at the target site and $P(s_i)$ be the probability of s_i. For each target region, $P(s_i)$ and $P(s_j|s_i)$ were calculated for each combination of i and j in a given area ($i \neq j$). The training model was finally constructed by the average of $P(s_i)$ and $P(s_j|s_i)$ across different training target sites for which s_i is observed in 100 reads and more, which is represented by $\bar{P}(s_i)$ and $\bar{P}\left(s_j|s_i\right)$, respectively.

Base-editing outcome prediction. Let $S_{m,n}$ be a base-editing pattern in a window spanning from m bp n bp relative to the PAM, which can be alternatively represented by a string of transition statuses, $s_m, s_{m+1}, \ldots, s_{n-1}, s_n$. Using the training model, the frequency of a given outcome $S_{m,n}$ in a test target site was predicted using the following equation:

$$P(S_{m,n}) = \left(\prod_{i \in E} \left(\bar{P}(s_i) \prod_{j \in R} \bar{P}\left(s_j|s_i\right) \right) \right)^{\frac{1}{|E|}}$$

where $R := \{x \in Z | m \leq x \leq n\}$, $E := \{x \in \text{positions with base transitions}\}$, $\bar{P}(s_i) = 0$ unless defined, and $\bar{P}\left(s_j|s_i\right) = 1$ unless defined.

Essentially, to predict the frequency of a given editing outcome with multiple base transitions, this prediction model calculates a geometric mean of probabilities of base transitions at all edited positions, each considering the other independent base transition patterns. Any specific $P(s_j|s_i)$ that was devoid of training data was ignored (treated as 1) and $P(s_i)$ that was devoid of training data was 0.

Validation of base-editing prediction model. The base-editing prediction model was evaluated by 5-fold, 15-fold and leave-one-out cross-validation experiments. For a test target site, the frequencies of all of the base-editing outcome patterns detected in the amplicon sequencing dataset for a window of −25 bp to −5 bp relative to the PAM were predicted by training the amplicon sequencing data of other target sites that did not overlap with the test target site. In k-fold cross-validation experiments, $47/k$ target sites were randomly selected as test samples from the 47 target sites and their editing outcomes were predicted by training the amplicon sequencing data of the remaining target sites, which was iterated 100 times by randomly changing the test samples. For the leave-one-out cross-validation, we predicted the editing outcomes of each target site using the amplicon sequencing data of the other 46 target regions. The prediction performance was measured by first transforming the predicted editing frequencies to relative editing frequencies among all edited reads, randomly selecting one prediction result for each editing outcome pattern if there were multiple, and calculating the Pearson's correlation coefficient between the prediction and experimental measurement.

Simulation of co-editing frequencies on synthetic target sequences. To predict the multidimensional co-editing spectra of different base-editing methods using the base-editing prediction model, we simulated 100 synthetic target sequences consisting of only cytosine and/or adenine bases in the region from −20 bp to −1 bp relative to the PAM. For each target region, all possible outcomes with C→T and A→G edits (2^{20} outcomes in total) were predicted using the base-editing prediction model trained from all 47 amplicon sequencing data. The average homologous and heterologous dinucleotide-editing spectra were then calculated using all predicted frequencies. The trinucleotide-editing spectra were also predicted for a simulated sequence in which poly-AC stretched from −20 bp to −1 bp relative to the PAM.

Codon convertibility matrix and bystander mutation risk. To estimate the codon conversion potential and bystander mutation risk of each base-editing method, codon convertibility matrices (CCMs) were generated with and without allowing bystander mutations to generate unwanted mutations alongside the target codon conversion. First, for each of the 11,250,496 source codons in the human genome (hg38), possible gRNA target sites were screened in the area of ±25 bp from a target codon. For all gRNAs, base-editing outcome probabilities of all possible C→T and/or A→G editing patterns in the ±15 bp region of the target codon triplet were predicted using the base-editing prediction model trained by the amplicon sequencing data of all 47 genomic sites. The conversion potential of the target source codon to each destination codon without bystander mutations was then defined as the maximum probability of generating the target outcome among those generated by all possible gRNAs. When bystander mutations were allowed to occur, all predicted probabilities of base-editing outcomes with the target source codon conversion were summed for each gRNA, and the maximum integrated probability among the possible gRNAs was defined as the conversion potential. Conversion potentials for codons that had no targetable gRNA were defined as 0 for any destination codon type. After calculating the conversion potentials to different destination codons for all genomic codons, a CCM was finally generated to show the frequency of each source-destination codon conversion type with a conversion potential threshold of 5%. The bystander mutation risk scores for different source-destination codon types were calculated by dividing the CCM frequencies allowing bystander mutations by those not allowing bystander mutations.

ClinVar analysis. For pathological C·G→T·A and A·T→G·C SNVs reported in the ClinVar database, the possible gRNA target sites to correct each mutation were first screened within the ±25 bp region from the target mutation. The probabilities of correcting mutations by these different gRNA target sites without inducing unwanted bystander mutations in the ±15 bp region from the target mutation were then predicted using the base-editing prediction model trained by all of the amplicon sequencing datasets. For each mutation, its correction potential was defined as the maximum probability of the target codon conversion among those induced by different gRNAs. Finally, to estimate the global potential of the dual-function base editors to correct two heterologous disease mutations simultaneously, we counted the combinations of two heterologous mutations that were both predicted to be correctable with a correction potential threshold of 5%. We limited the combinatorial heterologous mutation space to ones in the same genes to reduce the calculation cost.

Statistical analysis. All of the genome editing experiments were repeated at least three times with independent cell culture samples. The statistical test was performed by Scipy (version 1.4.1.)[29] on Python (version 3.7.4) or R (version 3.6.0.). Statistical methods, exact sample sizes and *P* values are available in Supplementary Table 5.

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

Data availability

The high-throughput sequencing data of this study are available at the Sequence Read Archive (PRJNA596330) of the NCBI. The original fluorescent microscopy image data are available at https://doi.org/10.6084/m9.figshare.12016785.v1.

Code availability

The source codes for the base-editing prediction model are available at https://github.com/yachielab/base-editing-prediction. The other codes used in this study are available upon request.

References

22. Doench, J. et al. Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
23. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
24. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
25. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
26. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
27. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
28. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
29. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

Acknowledgements

We thank members of the Yachie lab for useful discussions and critical assessment of this work, especially A. Adel for reviewing the manuscript. We also thank K. Shiina, Y. Takai and N. Ishii for technical supports of high-throughput sequencing. This study was mainly funded by the Uehara Memorial Foundation (to N.Y.), the NOVARTIS Foundation (Japan) for the Promotion of Science (to N.Y.), and the Japan Agency for Medical Research and Development (AMED) Platform Project for Supporting Drug Discovery and Life Science Research (to N.Y., H.N. and O.N.), and partly supported by the New Energy and Industrial Technology Development Organization (NEDO), AMED PRIME program (17gm6110007), the Japan Science and Technology Agency (JST) PRESTO program (10814), the Naito Foundation, the SECOM Science and Technology Foundation (all to N.Y.), the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (16J06287) (to S.I.) and research funds from the Yamagata Prefectural Government and Tsuruoka City, Japan (to K.A. and M. Tomita). S.I. was supported by a JSPS DC1 Fellowship; S.I., H.M. and N.M. were supported by TTCK Fellowships; H.M. and N.M. were supported by the Mori Memorial Foundation; and N.M. was supported by the Yamagishi Student Project Support Program of Keio University.

Author contributions

R.C.S., S.I., H.M. and N.Y. conceived and designed the study. R.C.S., S.I. and M. Tanaka constructed the plasmids. R.C.S., S.I., M. Tanaka and N.M. performed the base editor assays and the library construction for high-throughput sequencing. S.I. established the base editor reporter cell lines. R.C.S. and S.I. performed the fluorescence microscopy imaging. S.I. and H.M. performed most of the data analysis. K.T., H.U., S.Y., K.A., M.S. and H.A. performed the high-throughput sequencing and data analysis. K.N., A.K., H.N. and O.N. supported the design of Target-ACE and provided materials. M. Tomita helped the computational analyses. R.C.S., S.I., H.M. and N.Y. wrote the manuscript.

Competing interests

K.N. and A.K. are shareholders and board members of BioPalette Co., Ltd.

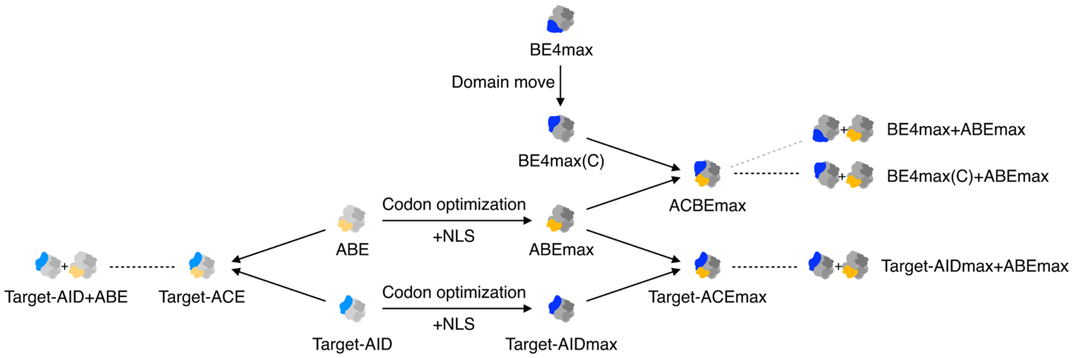
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-020-0509-0.

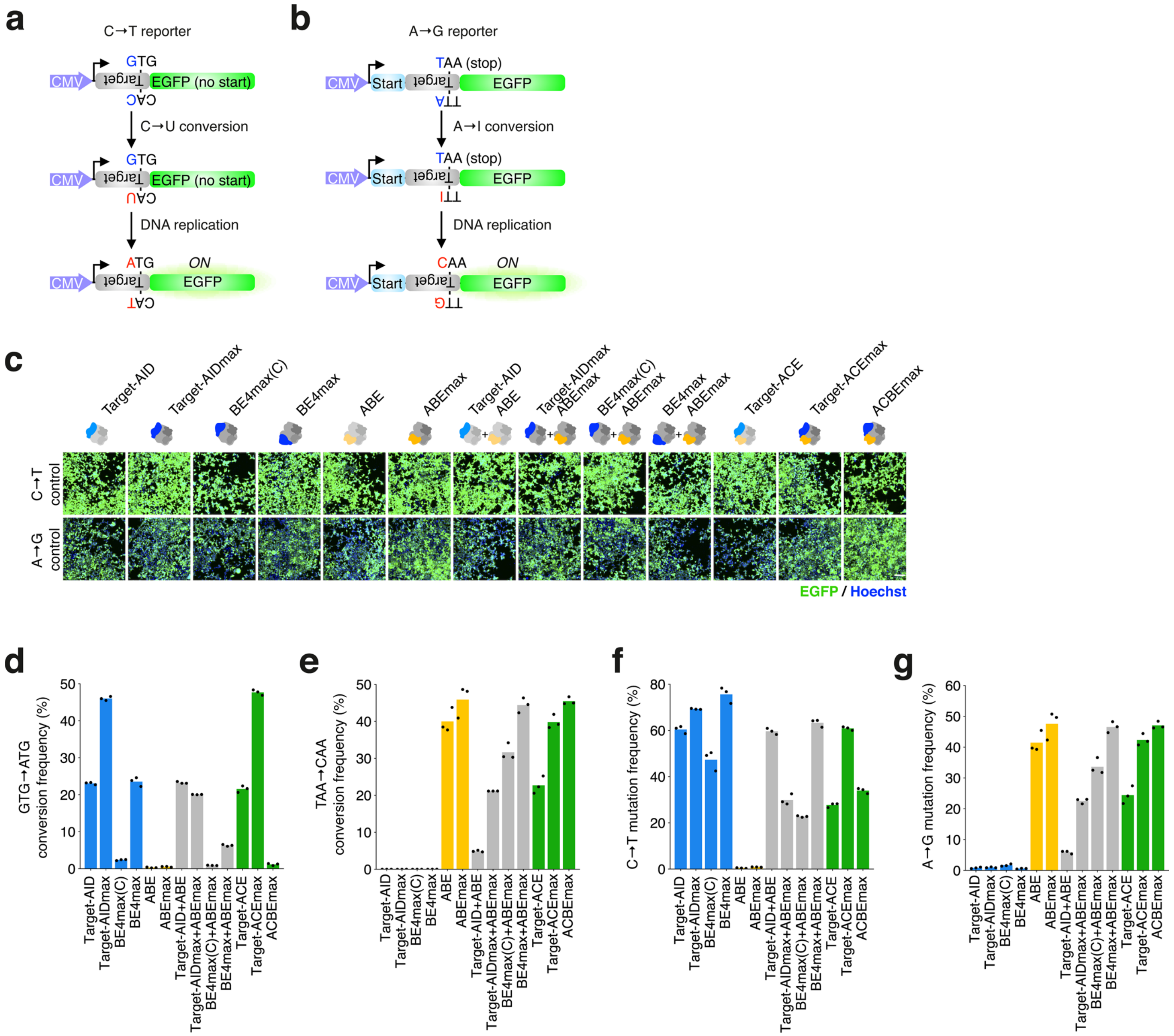
Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-0509-0.

Correspondence and requests for materials should be addressed to N.Y.

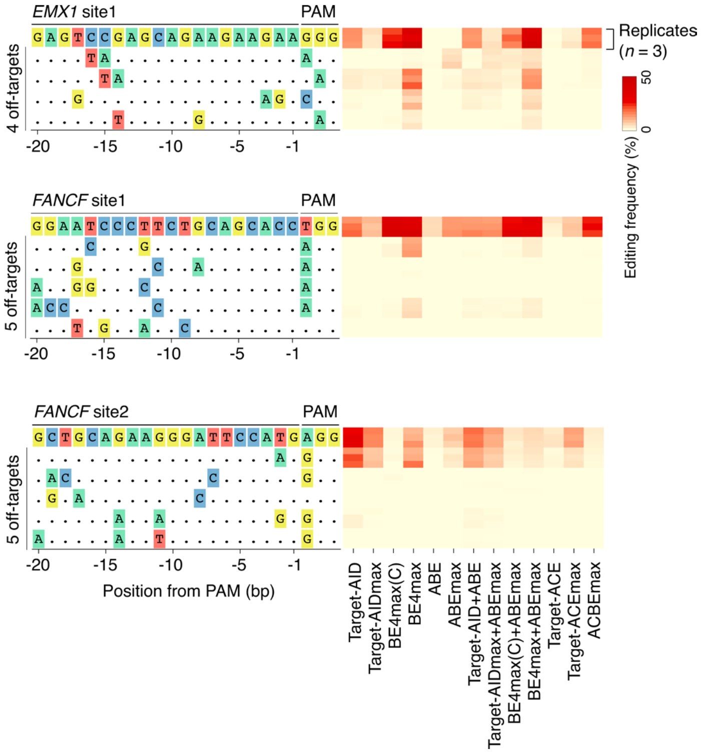
Reprints and permissions information is available at www.nature.com/reprints.



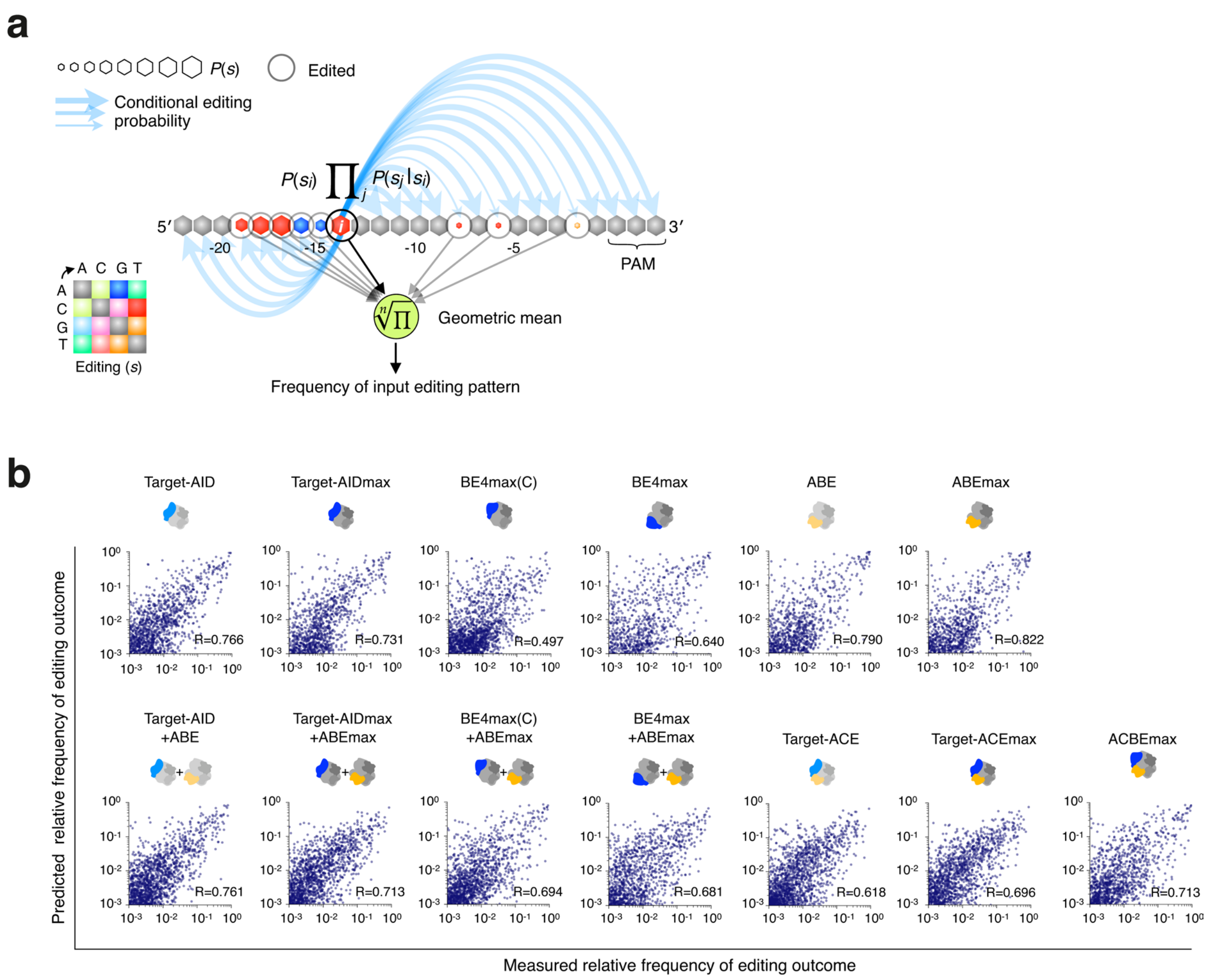
Extended Data Fig. 1 | Single- and dual-function base editors used in this study. Developmental lineages of single- and dual-function base editors used in this study are represented by arrows. Base editor mix controls for dual-function base editors are indicated by dashed lines.



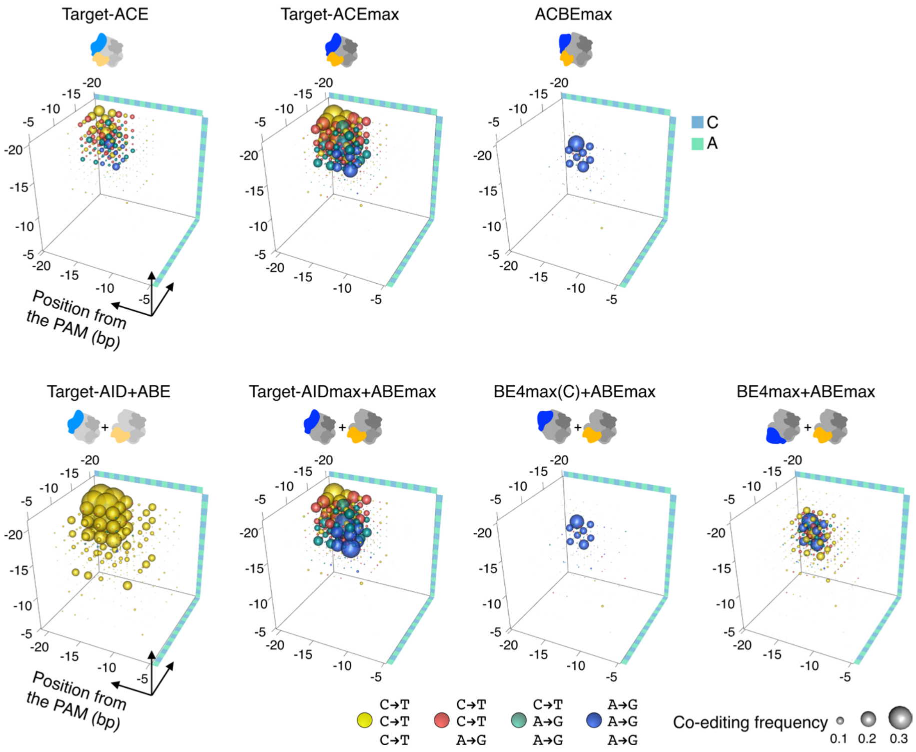
Extended Data Fig. 2 | Base-editing activity in base-editing reporter cells. a, Schematic representation of the C→T base-editing reporter. C→T base editing of the antisense strand followed by DNA replication restores the translation of EGFP by converting a mutated start codon GTG (valine) to ATG (methionine). **b**, Schematic representation of the A→G base-editing reporter. A→G base editing of the antisense strand followed by DNA replication converts the stop codon, TAA, to CAA (glutamine) releases the translation of its downstream EGFP. **c**, Microscopy images of the positive control cells for C→T and A→G base-editing reporters transiently transfected with different base editor reagents and non-targeting (NT) gRNAs. Scale bar, 40 μm. **d**, Frequency of start codon restoration in C→T editing reporter cells. Each bar shows the mean of three independent transfection experiments represented by dots. **e**, Frequency of stop codon destruction in A→G editing reporter cells. **f**, Frequency of amplicon sequencing reads showing C→T editing at any position of the gRNA target site of C→T editing reporter cells (from −30 to +10 bp relative to the PAM). **g**, Frequency of amplicon sequencing reads showing A→G editing at any position of the gRNA target site of A→G editing reporter cells (from −30 to +10 bp relative to the PAM).



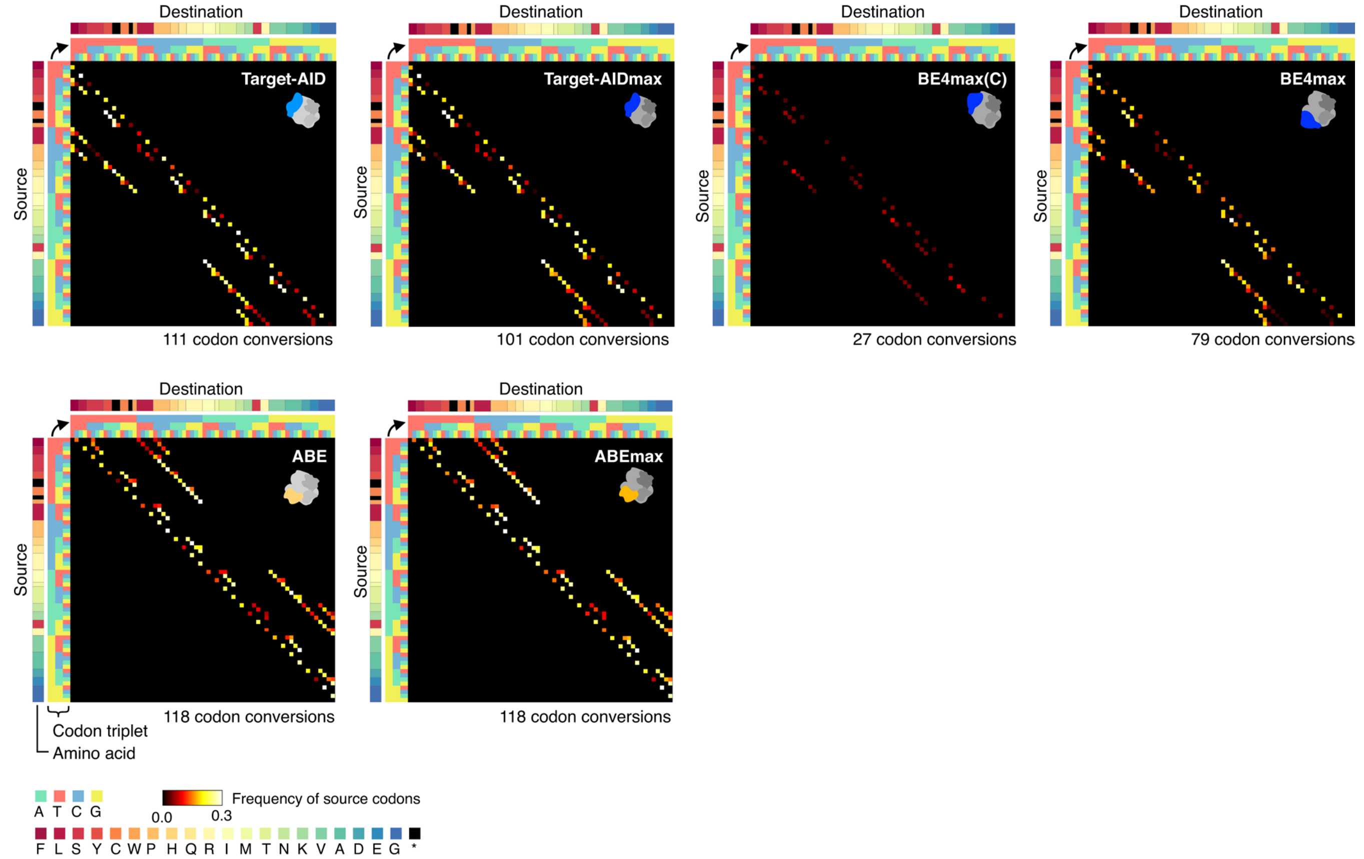
Extended Data Fig. 3 | DNA off-target editing activity. Editing frequencies of *EMX1* site 1 and *FANCF* site 1 and site 2 and their corresponding off-target sites. Amplicon sequencing experiments were performed in triplicate.



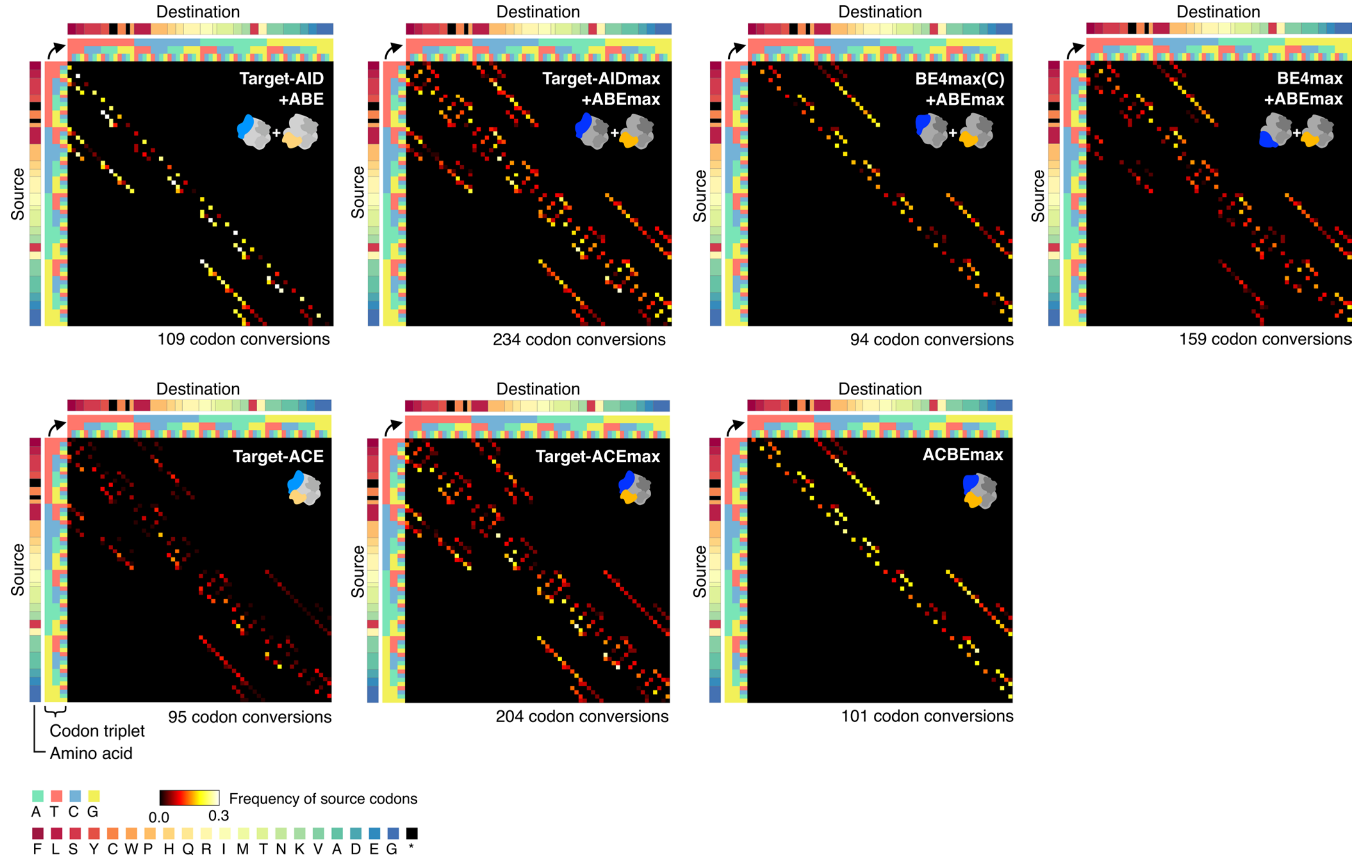
Extended Data Fig. 4 | Prediction of base-editing outcome frequencies. a, Schematic diagram of the model to predict the frequencies of each base-editing outcome. In brief, to train a given base editor model using a training amplicon sequencing dataset for different target sites, probabilities of single base transition events and their conditional probabilities given each of the other single events are thoroughly calculated for different positions relative to the PAM. The frequency of a given editing outcome in a new test target site is then predicted as a geometric mean of probabilities of base transitions at all edited positions, each given by the other independent base transition patterns. **b**, Correlation of measured and predicted relative editing outcome frequencies in the 5-fold cross-validation experiment.



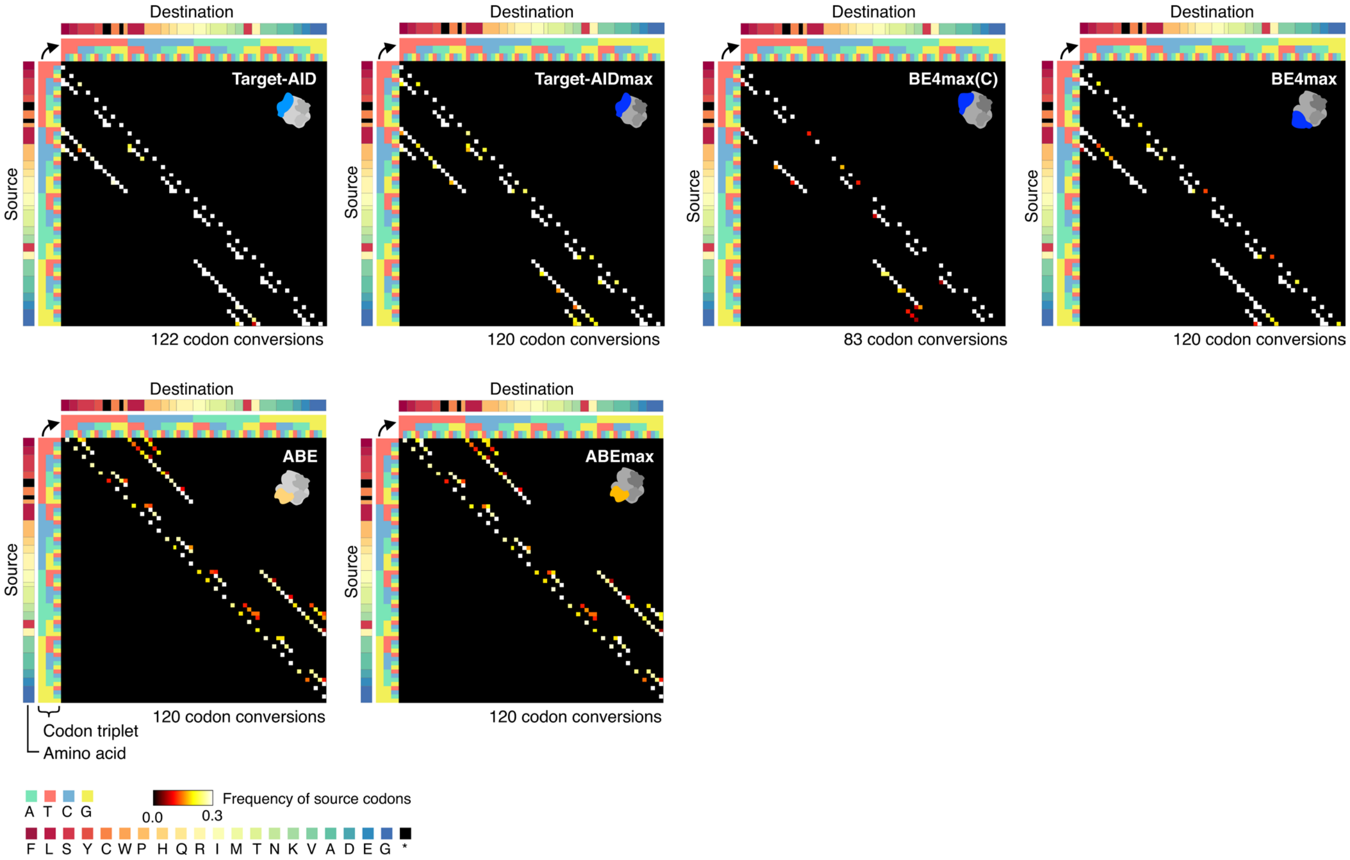
Extended Data Fig. 5 | Heterologous trinucleotide co-editing frequencies predicted by the computational model. To predict the multidimensional co-editing spectra of the different base-editing methods using the base-editing prediction model, 100 synthetic target sequences consisting of only cytosine and/or adenine bases in the region from −20 to −1 bp relative to the PAM were generated *in silico*. For each target sequence, all possible outcomes with C→T and/or A→G edits (2^{20} outcomes in total) were predicted using the base-editing prediction model trained from all 47 amplicon sequencing data. The average homologous trinucleotide-editing spectra shown by the bubble charts were then calculated using all predicted frequencies.



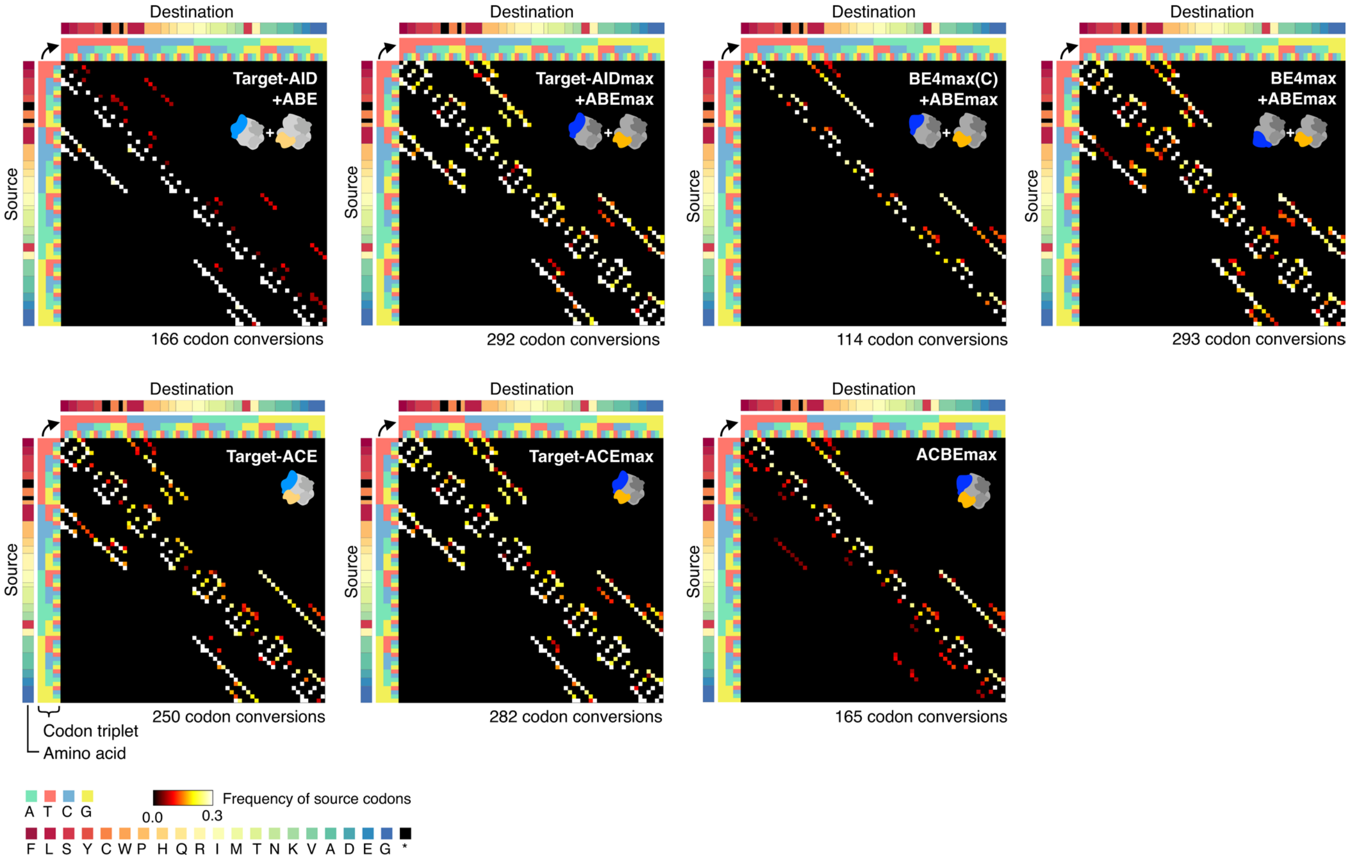
Extended Data Fig. 6 | Codon convertibility matrices (CCMs) of single-function base editors without allowing bystander mutations to occur. For each codon in the human genome (hg38), possible gRNA target sites were first screened in the area of ±25 bp. For all gRNAs, base-editing outcome probabilities of all possible C→T and/or A→G editing patterns in the ±15 bp region of the target codon were predicted using the base-editing prediction model trained by the amplicon sequencing data for all 47 genomic sites. The conversion potential of the target source codon to each destination codon without allowing bystander mutations to occur was then defined as the maximum probability of generating the target outcome among those induced by all possible gRNAs. After calculating conversion potentials to different destination codons for all genomic codons, a CCM was generated to show the genome-wide frequency of each source-destination codon conversion type with a conversion potential threshold of 5%.



Extended Data Fig. 7 | Codon convertibility matrices (CCMs) of base editor mixes and dual-function base editors without allowing bystander mutations to occur.



Extended Data Fig. 8 | Codon conversion matrices (CCMs) of single-function base editors with allowing bystander mutations to occur.



Extended Data Fig. 9 | Codon conversion matrices (CCMs) of base editor mixes and dual-function base editors with allowing bystander mutations to occur.

nature research

Corresponding author(s):	Nozomu Yachie (The University of Tokyo)
Last updated by author(s):	Mar 25, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For data acquisition, we used software tools supplied by the manufacturers for InCellAnalyzer6000 and Illumina MiSeq, HiSeq 2500 and NovaSeq6000.
Data analysis	For computational data analysis, we used NCBI Blast 2.7.0, EMBOSS 6.0.0, FLASH 1.2.0, bcl2fastq2 (version v2.20.0), seqtk version 1.3-r107-dirty, STAR version 2.7.3a, Picard MarkDuplicates version 2.0.1, BWA-MEM version 0.7.15, and GATK HaplotypeCaller version 4.1.4.1 as clarified in the manuscript. We also shared the source codes of the base editing prediction model developed in this study on GitHub (https://github.com/yachielab/base-editing-prediction)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The information of all oligonucleotides, plasmids and Illumina indices used for multiplexing of sequencing libraries are available in Supplementary Tables 2-4. Most of the plasmids constructed in this study will be all available at Addgene (depositor: Nozomu Yachie) and the other plasmids are available upon request. All the high-throughput sequencing datasets obtained in this study will be uploaded the NCBI Sequence Read Archive. Other processed datasets are also presented in Supplementary Data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The amplicon sequencing datasets were obtained for 47 genomic loci each with 13 different base editor reagents in triplicate (1,833 datasets in total). The number of data points generated from this data for different analyses are clarified in Supplementary Fig. 2. The base editing reporter cell assays were performed in quadruplicate. The whole exome sequencing and RNA-seq were performed in duplicate.
Data exclusions	One of the amplicon sequencing datasets was excluded from the analyses because the control sequencing data was not obtained successfully (clarified in the manuscript). The data exclusion criteria was preliminary established according to our previous experiences.
Replication	The replicate assays were all performed independently as different experimental batches.
Randomization	The assay samples for different combinations of base editors and guide RNAs were explicitly identifiable in the course of the experiments.
Blinding	The assay samples for different combinations of base editors and guide RNAs were explicitly identifiable in the course of experiments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293Ta (Genecopia)
Authentication	Provided by the vendor (Genecopia).
Mycoplasma contamination	Tested negative by genotyping PCR for every experiment.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell line was used.