nature biotechnology

Article

https://doi.org/10.1038/s41587-022-01532-7

# Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing

Received: 4 May 2022

Accepted: 28 September 2022

Liang Chen[1,6], Biyun Zhu[1,6], Gaomeng Ru[1,6], Haowei Meng[2,6], Yongchang Yan[2,6], Mengjia Hong[1], Dan Zhang[1], Changming Luan[1], Shun Zhang[1], Hao Wu[2], Hongyi Gao[1], Sijia Bai[1], Changqing Li[1], Ruoyi Ding[1], Niannian Xue[1], Zhixin Lei[2], Yuting Chen[3], Yuting Guan[1], Stefan Siwko[4], Yiyun Cheng[1], Gaojie Song[1], Liren Wang[1], Chengqi Yi[2] ✉, Mingyao Liu[1,5] ✉ and Dali Li[1] ✉

Cytosine base editors (CBEs) efficiently generate precise C·G-to-T·A base conversions, but the activation-induced cytidine deaminase/apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like (AID/APOBEC) protein family deaminase component induces considerable off-target effects and indels. To explore unnatural cytosine deaminases, we repurpose the adenine deaminase TadA-8e for cytosine conversion. The introduction of an N46L variant in TadA-8e eliminates its adenine deaminase activity and results in a TadA-8e-derived C-to-G base editor (Td-CGBE) capable of highly efficient and precise C·G-to-G·C editing. Through fusion with uracil glycosylase inhibitors and further introduction of additional variants, a series of Td-CBEs was obtained either with a high activity similar to that of BE4max or with higher precision compared to other reported accurate CBEs. Td-CGBE/Td-CBEs show very low indel effects and a background level of Cas9-dependent or Cas9-independent DNA/RNA off-target editing. Moreover, Td-CGBE/Td-CBEs are more efficient in generating accurate edits in homopolymeric cytosine sites in cells or mouse embryos, suggesting their accuracy and safety for gene therapy and other applications.

Base editors are composed of a nuclease-impaired Cas9 and a deaminase module to generate site-specific base conversions without inducing DNA double-stranded breaks in the absence of donor templates[1]. There are two major types of base editors, the cytosine base editors (CBEs)[2] and the adenine base editors (ABEs)[3], which catalyze C·G-to-T·A and A·T-to-G·C transitions, respectively. Through fusion of Cas9 nickase (Cas9n) with the activation-induced cytidine deaminase/apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like (AID/APOBEC) protein family of natural cytosine deaminases, CBEs usually generate considerable C-to-G or C-to-A byproducts because the U:G mismatch is recognized and excised by uracil DNA N-glycosylase (UNG) to create an abasic intermediate that initiates base excision repair to induce

[1]Shanghai Frontiers Science Center of Genome Editing and Cell Therapy, Shanghai Key Laboratory of Regulatory Biology, Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai, China. [2]School of Life Sciences, Peking University, Beijing, China. [3]CAS Key Laboratory of Quantitative Engineering Biology, Center for Genome Engineering and Therapy, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. [4]Institute of Biosciences and Technology, Texas A&M University, Houston, TX, USA. [5]BRL Medicine, Inc., Shanghai, China. [6]These authors contributed equally: Liang Chen, Biyun Zhu, Gaomeng Ru, Haowei Meng, Yongchang Yan. ✉e-mail: chengqi.yi@pku.edu.cn; myliu@bio.ecnu.edu.cn; dlli@bio.ecnu.edu.cn

unpredictable base conversions. Fusion with or simultaneous expression of a uracil glycosylase inhibitor (UGI) dramatically increases the CBE-induced efficiency and purity of C·G-to-T·A transition[4]. If UGI is replaced with UNG[5,6] or a DNA repair protein[7,8] in the CBE backbone, the C-to-G base editors (CGBEs) are developed to mainly induce cytosine base transversions. CBE/CGBEs are promising tools for a broad range of applications, while they generate considerable indels, bystander edits, and Cas9-independent DNA and RNA off-target edits[5,9–13], which raise safety concerns, especially for clinical applications. Several studies have developed more accurate CBE variants, such as BE4max-YE1/YEE, eA3A-BE4max and so on, through the engineering of APOBEC enzymes[13,14]. To improve the performance of CBE/CGBE technology, we thought to develop base editors with a cytosine deaminase that has potential distinct features superior to those of the AID/APOBEC family enzymes.

Different from CBE/CGBEs, ABEs use an unnatural adenine deaminase that was evolved from TadA, a transfer RNA (tRNA) adenine deaminase in *Escherichia coli*, to induce A-to-G conversions in DNA with very high product purity (over 99.9%), minimal indels and a relatively condensed editing window[3]. Recent studies have shown that ABEs also induce cytosine substitutions in a defined TC*N motif (where the asterisk denotes the target cytosine (C) to edit and 'T' and 'N' as adjacent bases) independently of adenine conversions, suggesting that the evolved TaDA has cytosine deamination capability of evolved[15,16]. In our recent study, we also noticed that ABE8e, a superactive ABE variant containing the TadA-8e deaminase, displayed increased cytosine deaminase activity[17]. Inspired by the identification of its unexpected cytosine deaminase activity, we attempted to repurpose adenine deaminase TadA-8e, the most efficient TadA variant, into a purely unnatural cytosine deaminase (Fig. 1a). We assumed that the TadA-8e-derived CBE would have potential advantages over AID/APOBEC-based editors because the original ABEs exhibited a minimal indel frequency and undetectable Cas9-independent DNA off-target editing[3,9,10].

Here we report the engineering of adenine deaminase TadA-8e with a substituted type of catalytic substrate, which not only eliminates intrinsic adenine activity but also enables efficient cytosine editing. Because this unnatural cytosine deaminase-derived Td-CGBE induces highly efficient and precise C·G-to-G·C transversion in human cells and rodent embryos, a series of CBEs (Td-CBEs) with distinct features was developed via further molecular evolution and UGI fusions. A low level of indels and background level of either DNA or RNA off-target events were observed in Td-CGBE/Td-CBE-treated cells. Moreover, we applied Td-CGBE/Td-CBEs to install the desired mutations in pathogenic intricately homopolymeric cytosine sequences for generation or correction of disease models.

## Results

### Structure-guided molecular engineering of TadA-8e

To evolve TadA-8e into a cytosine deaminase, we speculated that the purine ring (A) had a relatively bigger size compared to the pyrimidine ring (C); therefore, it might be more fragile and less tolerant of mutation(s) around the pocket of TadA-8e. According to recently published structures of ABE8e in complex with its substrates[18], 14 residues were selected and substituted with distinct amino acids to change the side chain size, polarity or hydrophilic–hydrophobic property (Fig. 1b). When testing on an endogenous target site in HEK293T cells, we found that several variants, such as V28G, N46A, N46G, N46L and N108G, markedly reduced A-to-G activity but kept a high C-to-D (where D represents G, T or A) editing activity (Fig. 1c). The ABE8e-N46 mutants exhibited high efficiency (up to 57.1%) and selectivity for cytosine deamination, suggesting that this residue was a key position for substrate base selectivity. Thus, we individually tested substitution of all the other amino acids at position N46 (Supplementary Fig. 1a). Although N46P and N46L variants showed similar activity and substrate selectivity at three targets, ABE8e-N46L was chosen for further investigations because it exhibited a slightly condensed editing window and relatively high C-to-G activity (Supplementary Fig. 1a,b). Additional V28G or N108G variants introduced into ABE8e-N46L did not further improve its performance (Supplementary Fig. 1c). While we were completing this project, Bae and colleagues reported that ABE7.10 containing a P48R variant displayed increased cytosine editing with reduced but not eliminated adenosine deaminase activity[19]. However, in contrast to ABE7.10, we found that the protien generated by introducing the A48R variant (A48 in TadA-8e) in ABE8e still retained a high A-to-G efficiency and a much lower rate of cytosine edits compared to ABE8e-N46L, suggesting that N46 in TadA-8e was more critical than A48 for substrate selectivity (Supplementary Fig. 1d).

Given the essential role of the N46L variant for the discrimination of adenine and cytosine, we questioned whether this substitution had a similar ability in previous ABE versions. Thus, ABEmax-N46L/N46L (with an N46L variant in both the TadA and TadA* domains) and miniABEmax-N46L variants were constructed and it was found that neither ABE variant exhibited adenine edits but both showed considerable cytosine edits at the target with a TC*N motif, with ABE8e-N46L displaying much higher activity (up to 2.2-fold; Supplementary Fig. 2a–c). For t *FANCF* site 1, the introduction of an N46L variant also increased the cytosine editing efficiency in ABEmax (34.3% versus 16.4%; Supplementary Fig. 2d), indicating that this variant might also improve cytosine catalytic ability in previous ABE versions. Importantly, ABE8e-N46L but not the other two variants (ABEmax-N46L/N46L and miniABEmax-N46L) was able to edit cytosines in the other three sequence contexts (CCN, GCN and ACN), indicating an expanded targeting scope (Supplementary Fig. 2). These data suggested that N46L was critical for the discrimination of adenines and cytosines both in ABE8e and previous ABE variants.

### TadA-derived editor induces efficient editing

Because ABE8e-N46L mainly induced C-to-G transversion (Fig. 1c and Supplementary Fig. 2), we named it TadA-derived C-to-G base editor (Td-CGBE) and compared it with previously reported representative CGBEs[5,8], CGBE1 and rAPOBEC-Cas9n-rXRCC1 (hereafter termed CGBE-XRCC1), and ABE8e at 23 target sites. Td-CGBE showed very efficient C-to-G editing (up to 72.8%) similar to CGBE1, while the activity was much higher than that of CGBE-XRCC1 on most of the targets. Importantly, Td-CGBE exhibited a very steep and narrow editing window (C5–C6) even corresponding to single cytosines at 17 of the 21 sites containing cytosines on positions 5–6 (Fig. 1d and Supplementary Fig. 3a,b), while the other two editors showed a broader window (Fig. 1e). Td-CGBE was very efficient on TC*N motifs as previously observed in ABE7.10 (ref. [15]), but unexpectedly, it had a much higher efficiency of editing CC*N (average 1.9-fold increase compared to CGBE1 at the C5 position) and GC*N (average 5.6-fold increase compared to CGBE1 at the C5 position) motifs than the other two editors (Supplementary Fig. 3c). Td-CGBE exhibited up to 94.4% C-to-G product purity, which was very comparable to CGBE1 at all tested sites and superior to CGBE-XRCC1 (up to 3.1-fold improvement) at CC*N-motif sites (Supplementary Fig. 3d). Importantly, Td-CGBE induced much fewer indels in almost all tested targets (mean indel frequencies of 18.6%, 28.6% and 9.5% for CGBE1, CGBE-XRCC1 and Td-CGBE, respectively), which was 51.1% and 33.2% of the indel rates induced by CGBE1 and CGBE-XRCC1, respectively (Supplementary Fig. 3e). To investigate whether decreased indels were attributed to lower protein levels, western blotting assays were performed. To our surprise, the protein level of Td-CGBE was much higher than those of ABE8e and the other two CGBEs, suggesting that the N46L variant somehow enhanced protein synthesis (Supplementary Fig. 3f). However, the elevated expression did not induce A-to-G conversion by Td-CGBE at all tested targets and four additional well-used and efficient ABE targets, suggesting the elimination of adenine deaminase activity (Fig. 1f,g). These data suggest that Td-CGBE is a pure cytosine editor without recruitment of natural AID/APOBEC family enzymes,

https://doi.org/10.1038/s41587-022-01532-7

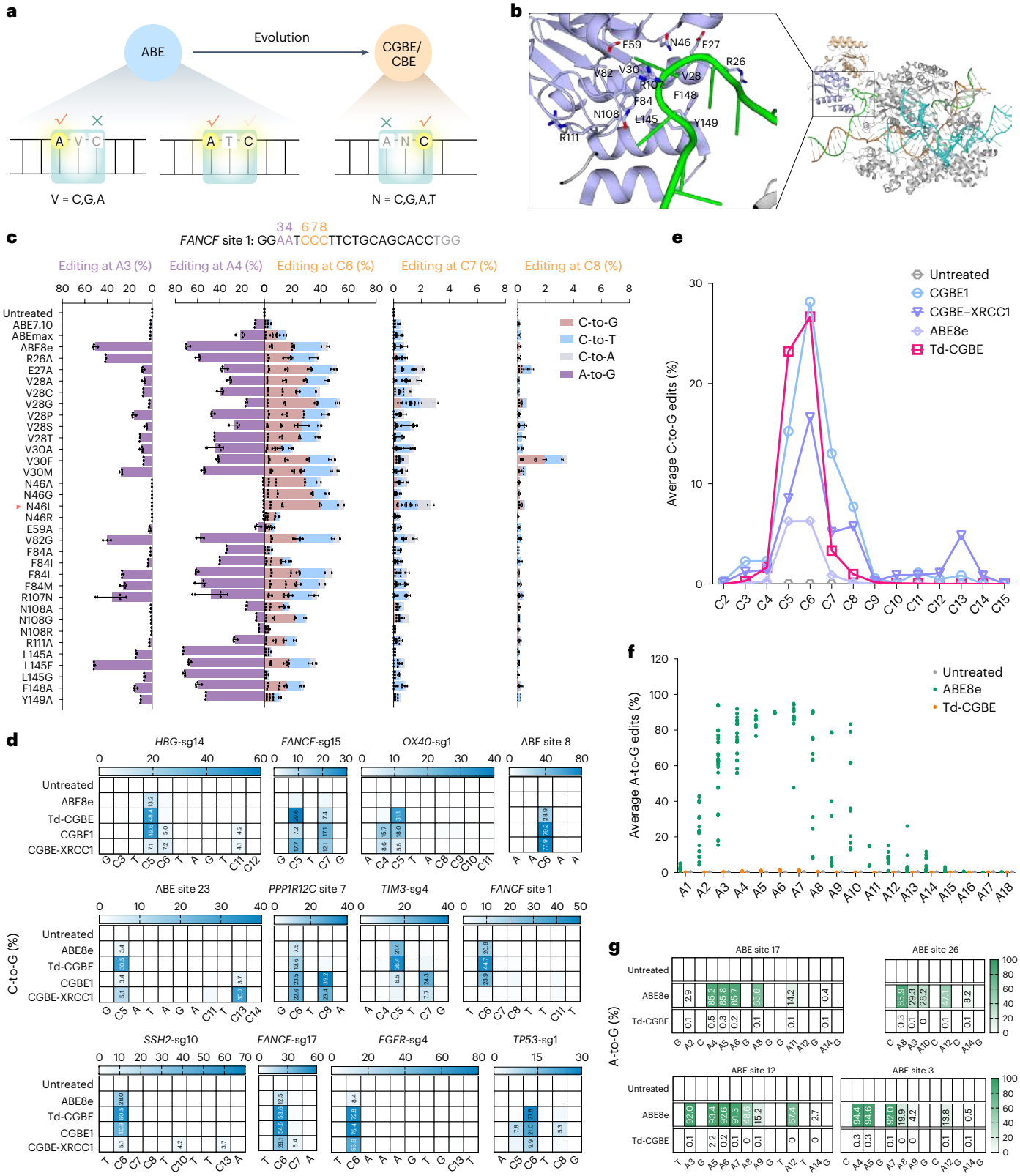


**Fig. 1 | Engineering of TadA-8e-derived CGBEs. a**, A schematic illustration of the conceptual design to evolve an ABE into pure CBEs. **b**, Overview of the interaction of TadA-8e (light purple) with the single-stranded DNA substrate (green sticks) (Protein Data Bank (PDB): 6VPC). Cas9n is in gray, sgRNA is in cyan, complementary strand DNA is in orange and noncomplementary strand DNA is in green. Amino acids spatially contacting or adjacent to the substrate DNA are labeled on the enlarged image. **c**, Base editing efficiencies of ABE7.10, ABEmax, ABE8e and ABE8e variants at *FANCF* site 1 in HEK293T cells. The N46L variant (red arrowhead) was chosen for further evaluation. **d**, Heatmaps showing the on-target C-to-G editing efficiencies of ABE8e, Td-CGBE, CGBE1 and CGBE-XRCC1 at 12 endogenous target sites in HEK293T cells. **e**, Average C-to-G editing efficiencies of each editor at 23 endogenous target sites in **d** and Supplementary Fig. 3b. **f**, Dot graph showing the A-to-G editing frequencies of ABE8e and Td-CGBE at 23 endogenous sites shown in **d** and Supplementary Fig. 3b. Each data point represents a biological replicate at each target site. **g**, Heatmaps showing the on-target A-to-G editing efficiencies of ABE8e and Td-CGBE at four endogenous targets containing multiple adenines in HEK293T cells. **d**,**e**,**g**, Data represent the mean of three independent experiments.

inducing highly efficient C-to-G conversions in a very narrow window with the fewest byproduct indels.

Because previous studies showed that YE1/YEE variants in APOBECs or shorter linkers between Cas9n and the deaminase reduced the editing window[14,20], we next compared Td-CGBE with CGBE variants (potentially with condensed editing windows) that were generated through the above strategies. However, CGBE-YEE and CGBE-nl (no linker) had very limited C-to-G editing after evaluation of six cytosine-rich targets. CGBE-YE1 showed comparable activity but a wider editing window compared to Td-CGBE. Consistent with a previous study showing that BE4max was inefficient at the target within GC*N motifs[21], we found that the three APOBEC1-derived CGBE variants showed poor C-to-G editing frequencies (0.4–4.5%) when the target cytosines were in a GC*N motif (ABE site 23), while Td-CGBE edited this target with a frequency of 29.6% (up to 74-fold higher; Supplementary Fig. 4). These data further highlighted the advantages of an unnatural cytosine deaminase-derived CGBE. In addition, we further fused the UNG element to the Td-CGBE backbone, but no marked improvement was observed (Supplementary Fig. 5). Moreover, when fusing TadA-8e-N46L with the photospacer adjacent motif (PAM)-relaxed SpCas9-NG variant[22], Td-CGBE-NG induced precise C5-to-G editing with a frequency of up to 48.9% at non-NGG PAM targets, suggesting that this TadA-8e-derived cytosine deaminase was compatible to engineered Cas9 variants (Supplementary Fig. 6).

### Characterization of TadA-derived CBE variants

The distinct precision of Td-CGBE encouraged us to investigate whether Td-CGBE could be transformed into a Td-CBE. Thus, two copies of UGI were linked to ABE8e-N46L with the P2A peptide to generate a Td-CBE construct (Supplementary Fig. 7). Through evaluation of editing efficiency on a poly(C) target, we found that Td-CBE efficiently generated C-to-T conversion at C6–C8 at a frequency of up to 84.6%, suggesting that it was highly efficient (Fig. 2a). However, its editing window was not narrow as expected. We introduced further variants in or near the active pocket of TadA-8e to reduce the editing window. The majority of additional variants reduced the editing window and efficiency simultaneously, but unexpectedly an additional E27R variant increased both the activity and window of Td-CBE (Fig. 2a). Then, we thought to shorten the linker between the deaminase and Cas9n to narrow the window because our previous study showed that using a single-stranded DNA binding domain to elongate the linker dramatically increased the editing window of CBEs[23]. Interestingly, reducing the 32-residue XTEN sequence to 3- to 7-residue linkers, which was reported to reduce the editing window[20], resulted in highly efficient editing within a 2-nucleotide (nt) window (Td-CBE-linker13, Td-CBE-linker15 and Td-CBE-linker18 in Fig. 2a; Supplementary Table 3). When all the linker residues were removed, the variant preferentially edited C6 within adjacent cytosines at this target (Fig. 2a) and was named enhanced Td-CBE (eTd-CBE). Moreover, we also noticed that although the efficiency of Td-CBE-P29A and Td-CBE-A48M was reduced, these variants showed very narrow editing windows. Thus, the linkers of these two constructs were removed to generate eTd-CBE-P29A (eTd-CBEa) and eTd-CBE-A48M (eTd-CBEm), which showed single-cytosine edits at this target as well as at an additional target site (Fig. 2a and Supplementary Fig. 8), suggesting that these two variants were very efficient and precise.

To further investigate the performance of Td-CBE with an E27R variant (named Td-CBEmax), multiple endogenous targets were tested. As Td-CBEmax was very efficient, we compared it with BE4max side by side using 13 additional gRNA sites with scattered cytosines. The activity (defined as the activity at the position with the highest activity in each target) of Td-CBEmax at these targets ranged from 57.7% to 94.9%, which was comparable to BE4max (44.9–93.2%; Fig. 2b and Supplementary Fig. 9a). We also noticed that the major editing window of Td-CBEmax was reduced to 3nt (positions 5–7) compared to the 5-nt window (positions 5–9) of BE4max (Fig. 2c). Consistent with Td-CGBE, Td-CBEmax also induced a steady, low rate of indels ranging from 1.6% to 7.5% (4.6% on average), but BE4max had a 1.8 fold higher indel rate (8.5% on average) of up to 24.9%, suggesting that Td-CBEmax induced a low level of severe DNA damage (Supplementary Fig. 9b). The data demonstrate that Td-CBEmax is a highly efficient CBE inducing fewer bystander variants and indels at the evaluated target sites. We also noticed that the protein level of Td-CBEmax was higher than that of BE4max (Supplementary Fig. 9c). As the above data showed eTd-CBE variants displaying high efficiency and a narrow window, we next compared them with other accurate CBEs with either a narrow window (BE4max-YE1 and BE4max-YEE) or preferential editing in a defined sequence context, such as eA3A-BE4max and A3G-BE5.13, which prefer TC*N and CC*N motifs, respectively[13,14,24]. Through examination of 12 endogenous targets including cytosine-rich sites, we demonstrated that eTd-CBE had comparable activity but a more condensed window in comparison to BE4max-YE1 (C5–C6 versus C4–C8). Although eTd-CBEm and eTd-CBEa showed a little bit less efficiency than BE4max-YEE, these two variants edited a single cytosine at C5 or C6 in 9 of 12 target sites with up to 48.3% efficiency, while BE4max-YEE only induced single-cytosine conversion at two sites (Fig. 2d,e and Supplementary Fig. 10a). eTd-CBE variants showed 12.5-fold (ranging from 1.8- to 131.4-fold) higher precision (determined by dividing the efficiency at the highest position with the highest activity by that at the position with the second highest activity) compared to BE4max-YEE, suggesting that eTd-CBE variants were more accurate and induced less bystander editing than BE4max-YE1 or BE4max-YEE (Supplementary Fig. 10b). The eTd-CBE variants also showed comparable activity but a more condensed editing window and higher precision (up to 90-fold higher) in comparison to other accurate CBEs, such as eA3A-BE4max and A3G-BE5.13 (Fig. 2e and Supplementary Fig. 10b). Consistent with Td-CGBE, eTd-CBEs made no adenine edits and only induced less than 2% indels on average, which was a much lower rate than with BE4max-YE1, eA3A-BE4max and A3G-BE5.13 (Supplementary Fig. 10c,d). The reduction of indels by Td-CBE variants was not due to its lower expression level (Supplementary Fig. 9c).

### Off-target evaluation of Td-CGBE and Td-CBEs

Similar to CGBE1, Td-CGBE induced background levels of cytosine mutations in analysis of all 36 in silico-predicted Cas9-dependent off-target sites[25], while CGBE-XRCC1 induced low-level off-target editing at 2 sites (Supplementary Fig. 11a). Moreover, in an enhanced orthogonal R-loop assay[26,27], CGBE1 generated mild rates of off-target effects (~3% on average and up to 5.7%), but CGBE-XRCC1 (~8.4% on average and up to 17.5%) generated much higher Cas9-independent cytosine off-target editing (Fig. 3a). Although ABE8e did not induce cytosine editing, it generated more severe Cas9-independent adenine editing. In contrast, Td-CGBE induced both cytosine and adenine off-target conversions at background levels (0.1–0.9%), suggesting that it did not generate deaminase-induced random editing (Fig. 3a,b and Supplementary Fig. 11b). Taking the data together, compared with AID/APOBEC-based CGBEs, Td-CGBE is very efficient and induces fewer bystander edits and minimal indels. As Td-CGBE also displays diminished Cas9-independent DNA off-target effects, it demonstrates that high-quality cytosine base conversions could be achieved by non-APOBEC family enzymes.

To extensively investigate the off-target effects of Td-CBEs, several strategies were used. When evaluating 29 predicted off-target sites from three loci, we found that Td-CBE variants exhibited much lower Cas9-dependent off-target editing than APOBEC family CBEs. No obvious increased off-target editing was found in Td-CBEmax-treated sites, but in 10 of these 29 sites Td-CBEmax showed a dramatic decrease compared to BE4max. For the more accurate variants, eTd-CBEm and eTd-CBEa had much lower off-target editing efficiency (average <1%) compared to BE4max-YE1 (average 6.8%) and similar to BE4max-YEE (average 1.3%) at most of the off-target sites (Supplementary Figs. 12 and 13). Using the enhanced orthogonal R-loop assay, we found that Td-CBEmax induced much fewer edits compared to BE4max in all

https://doi.org/10.1038/s41587-022-01532-7

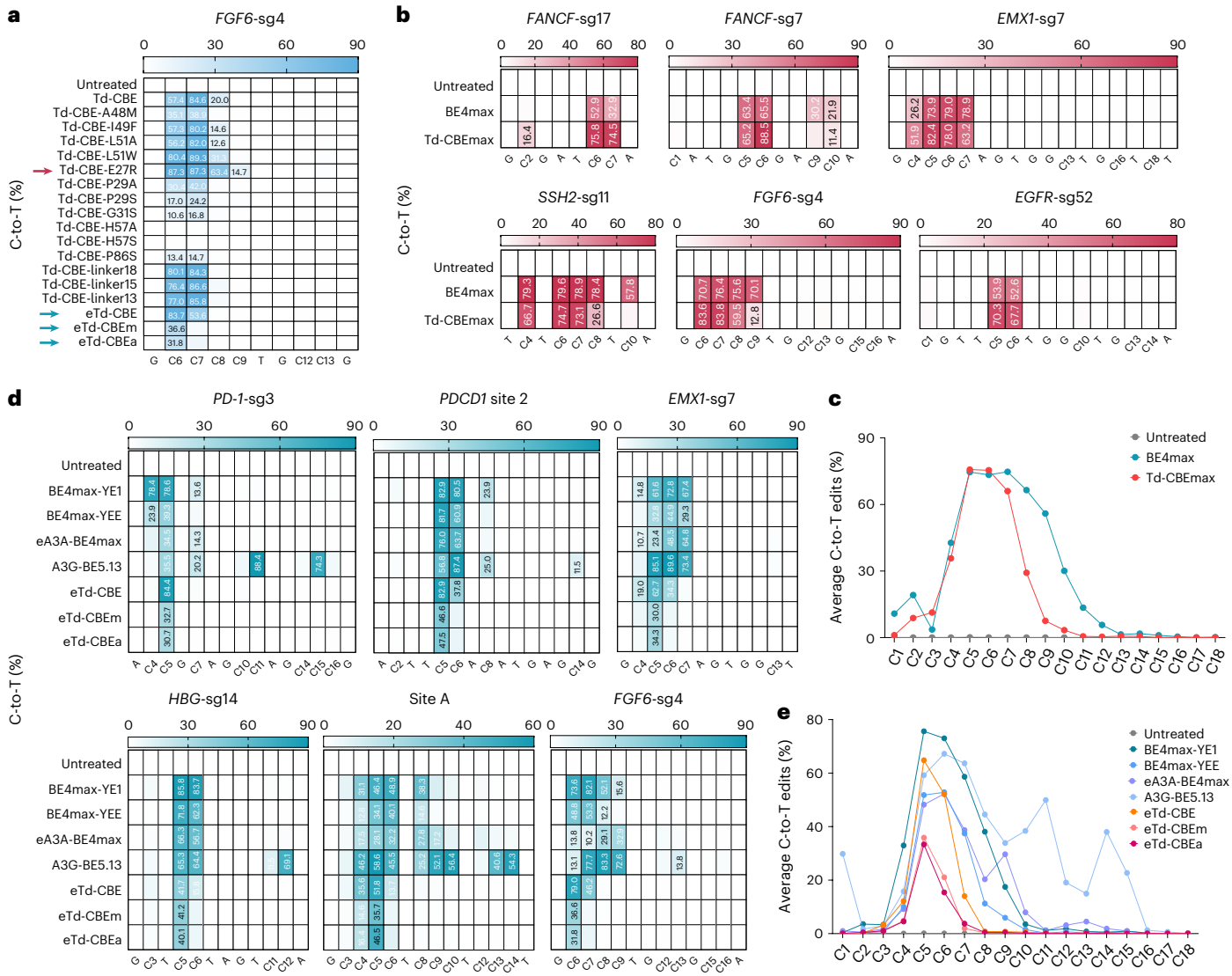


**Fig. 2 | Evolution and characterization of Td-CBEs in mammalian cells.** **a**, Comparison of C-to-T editing efficiencies of diverse Td-CBE variants at the *FGF6*-sg4 site in HEK293T cells. Data represent the mean of three independent experiments. **b**, Evaluation of the C-to-T editing efficiencies of BE4max and Td-CBEmax at six representative endogenous genomic loci in HEK293T cells. **c**, Average C-to-T editing efficiencies of BE4max and Td-CBEmax at 13 target sites in **b** and Supplementary Fig. 8a. **d**, The C-to-T editing efficiencies of the indicated CBEs at six representative endogenous genomic loci in HEK293T cells. **e**, Average C-to-T editing efficiencies of the indicated CBEs at 12 target sites in **d** and Supplementary Fig. 9a. **b**–**e**, Data represent the mean of three independent experiments except for BE4max in *FANCF*-sg17 and BE4max-YE1 in *EMX1*-sg7 ($n = 2$).

six sites. Compared to the accurate CBEs, eTd-CBEm and eTd-CBEa induced only background editing, which was slightly lower than that of BE4max-YEE at three of the six tested sites and much lower than that of BE4max-YE1, eA3A-BE4max and A3G-BE5.13 (Fig. 3c).

Recently, we developed an unbiased Detect-seq method for genome-wide assessment of CBE off-target effects and revealed that previous CBEs also generated unexpected edits outside the protospacer region and on the target strand[28,29]. We then performed Detect-seq experiments to test BE4 family and Td-CBE variants with sgRNA targeting the promiscuous *VEGFA* site 2. Through this evaluation, we found that BE4max (946 sites) induced 2.1-fold more off-target edits compared to Td-CBEmax (446 sites), although the editors had similar on-target efficiencies. Additionally, eTd-CBEm and eTd-CBEa induced similar numbers of off-target edits (37 and 32 sites, respectively) compared to BE4max-YEE (33 sites), which were much fewer than with BE4max-YE1 (387 sites), BE4max and Td-CBEmax (Fig. 3d). In addition, eTd-CBEm and eTd-CBEa had a narrow editing window as well as rigorous sequence context requirements in comparision to BE4max-YE1 and BE4max-YEE, as determined by analysis of off-target editing events (Supplementary Fig. 14a,b). Although Td-CBEs were not derived from APOBEC family enzymes, Td-CBEmax, eTd-CBEm and eTd-CBEa did not cause any de novo off-target sites compared to BE4max (Supplementary Fig. 14c). Notably, in contrast to previous results in which BE4max caused unexpected out-of-protospacer and target-strand edits, no such off-target events were observed in Td-CBE variant-treated cells (Fig. 3e). This suggests that Td-CBEs have the marked advantages of lower genome-wide off-target effects than most BE4max serial editors and are comparable to BE4max-YEE.

To evaluate Cas9-independent RNA off-target effects, transcriptome profiling was used. Consistent with previous reports[11,12], BE4max induced numerous C-to-U edits and BE4max-YE1 produced fewer off-target events. Td-CBEmax only induced 0.25% of the RNA off-target edits of BE4max, although we determined that the editors

https://doi.org/10.1038/s41587-022-01532-7

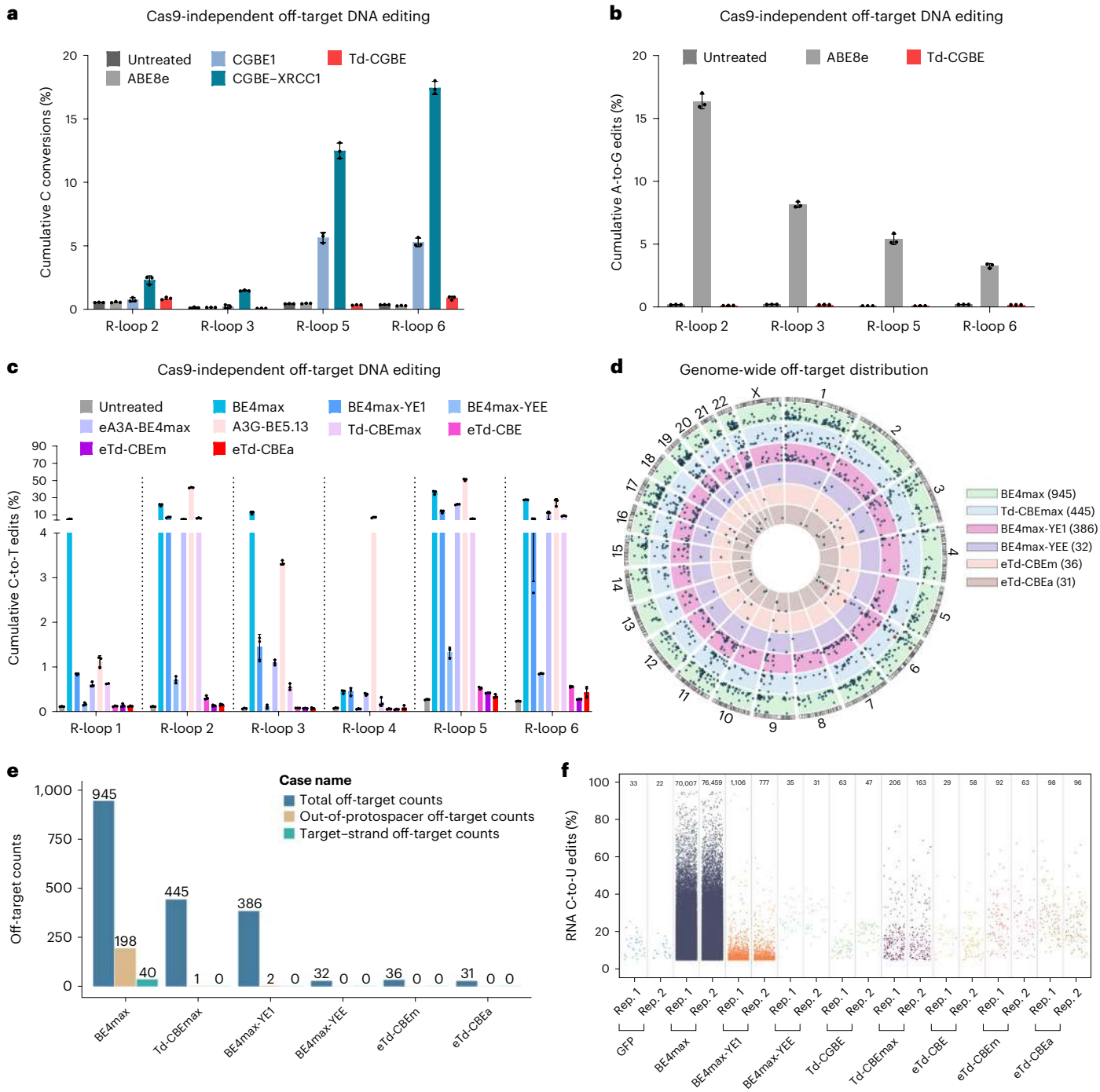


**Fig. 3 | Off-target assessment of Td-CGBE and Td-CBEs. a**, Cas9-independent DNA off-target analysis of the cumulative cytosine edits induced by ABE8e, Td-CGBE, CGBE1 and CGBE-XRCC1 using the modified orthogonal R-loop assay. **b**, Cas9-independent DNA off-target analysis of A-to-G edits induced by ABE8e and Td-CGBE using the modified orthogonal R-loop assay. **c**, Cas9-independent DNA off-target analysis of cumulative C-to-T edits induced by the indicated CBEs using the modified orthogonal R-loop assay. Data are mean ± s.d. ($n = 3$ independent experiments except for Td-CBEmax in R-loop 1 and eTd-CBEa in R-loop 6 with two biological replicate experiments). **d**, Genome-wide distribution of off-target effects determined by Detect-seq on each chromosome for the indicated CBEs with an sgRNA targeting *VEGFA* site 2. On- and off-target edits are indicated by red squares and blue circles, respectively. The number of off-target sites is in parentheses. **e**, Counts of out-of-protospacer editing, target-strand editing and all identified off-target events for the indicated CBEs. **f**, Jitter plots showing the ratio of RNA C-to-U editing (*y* axis) from the RNA-seq experiments. The total number of modified bases is listed on the top. Each dot represents an edited cytosine position in RNA. Each biological replicate is listed on the bottom. In **a** and **b**, data are mean ± s.d. ($n = 3$ independent experiments).

had a similar DNA editing activity (Fig. 3f). Moreover, Td-CGBE and eTd-CBEs induced background levels of C-to-U off-target edits (Fig. 3f and Supplementary Fig. 15a,b). Additionally, only background levels of A-to-I RNA edits were observed for BE4max and TadA-derived editors, further confirming that the variants induced in TadA-8e fully abolished its adenine deaminase activity (Supplementary Fig. 15c). These results demonstrate that Td-CGBE and Td-CBE variants have almost eliminated Cas9-independent RNA off-target effects.

https://doi.org/10.1038/s41587-022-01532-7

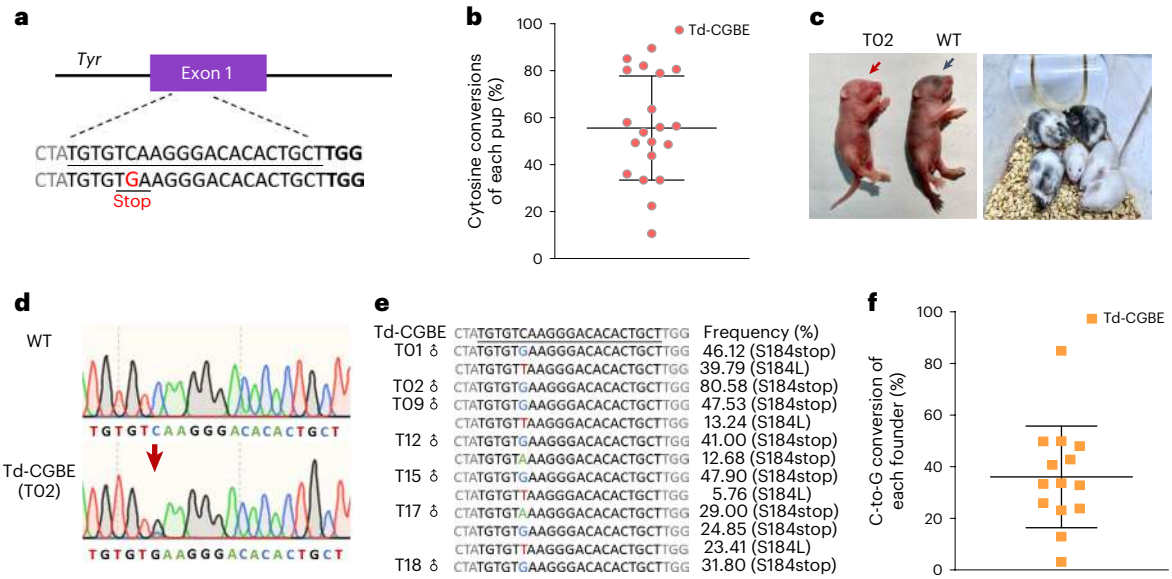


**Fig. 4 | Examination of mouse embryos with Td-CGBE. a**, Schematic of the target sequence in the exon 1 locus of the mouse *Tyr* gene. The sgRNA target sequence is in black, and the PAM is in bold. The desired C6-to-G transversion causing the premature stop codon TGA is in red. **b**, Cytosine conversion frequencies in mutant $F_0$ mice ($n = 20$). **c**, Phenotype of $F_0$ mice generated by Td-CGBE injection. The picture on the left was taken when the mice were 7 days old, while the right one was taken when mice were at 21 days old. WT, wild type. **d**, Sanger sequencing chromatograms of DNA from representative $F_0$ mice (T02) and WT mice injected with Td-CGBE mRNA. **e**, Genotyping of representative $F_0$ pups treated with Td-CGBE mRNA. The frequencies of mutant alleles were determined by high-throughput sequencing. **f**, C6-to-G editing frequencies in C-to-G founders generated by Td-CGBE ($n = 14$). In **b** and **f**, data are mean ± s.d. and each data point represents an individual mouse.

## Efficient and accurate editing in mouse embryos by Td-CGBE

To evaluate the potential of TadA-derived base editors for application, we tested their performance in mouse embryos. When injecting Td-CGBE mRNA and an sgRNA to target *Tyr* gene exon 1 to create a premature stop codon in mouse embryos (Fig. 4a), 20 of the 21 $F_0$ pups obtained had cytosine conversions with an average of 55.6% efficiency (Fig. 4b and Supplementary Fig. 16a). An Albino phenotype was observed in $F_0$ founders, suggesting ablation of tyrosinase function (Fig. 4c,d). In total, 67% of the pups had the desired C-to-G edits with 36% average efficiency (up to 84.9%), which was much higher than in a previous report, which used an optimized CGBE1 to edit the same target without observing an albino phenotype in founders[30] (Fig. 4e,f and Supplementary Fig. 16b,c). Similar to the data obtained in cell lines, Td-CGBE induced few indels in mouse embryos (Supplementary Fig. 16d). Our results demonstrated that the TadA-derived cytosine deaminase was efficient not only in cell lines but also in mouse embryos and was likely even more efficient than APOBEC1-derived CGBE1 in vivo.

## Precise editing of pathogenic single-nucleotide variants by Td-CGBE and Td-CBEs

As Td-CGBE and Td-CBE showed higher precision than other typical editors, we tried to compare them with representative base editors to edit pathogenic single nucleotide variants (SNVs) in homopolymeric cytosine sites. To create pathogenic C-to-G SNVs, Td-CGBE was delivered with individual sgRNAs targeting cytosine-rich sites, including the *MPZ* gene (causing Charcot Marie Tooth disease type 2I[31]) and the *PTEN* gene (causing macrocephalus[32]) in HEK293T cells. Compared with CGBE1 and CGBE-XRCC1, Td-CGBE was very efficient and predominantly edited the desired single cytosine, which was occured at a 1.5- and 8-fold higher rate than CGBE1-induced precise editing at two sites (Fig. 5a). Additionally, eTd-CBEa and eTd-CBEm introduced a pathogenic C-to-T mutation in the *KCNA2* gene (causing epileptic encephalopathies[33]) with 61.6% and 78.7% efficiency, respectively, frequencies that were much higher than with BE4max-YE1 (10.9%) and BE4max-YEE (8.4%) (Fig. 5b). To test the potential for correction of pathogenic SNVs, stable cell lines containing pathogenic variants were generated. The data showed that Td-CGBE generated much higher C-to-G correction ratios and fewer indels than CGBE1 and CGBE-XRCC1 in two cell lines containing G-to-C variants (*CELA2A* c.639+1G>C causing early-onset atherosclerosis[34] or *HBB* c.328G>C causing Hemoglobin Johnstown[35]; Fig. 5c and Supplementary Fig. 17a,b). In cell lines bearing T·A-to-C·G variants (*TUBB6* c.1181T>C causing congenital nonprogressive bilateral facial palsy[36] or *PFN1* c.350A>G causing amyotrophic lateral sclerosis[37]), BE4max-YE1 and BE4max-YEE mainly induced simultaneous double- or triple-cytosine transitions, but eTd-CBEm (2.9- and 1.9-fold-higher rate of correction than with BE4max-YEE) and eTd-CBEa (4.1- and 1.9-fold-higher rate of correction than with BE4max-YEE) induced much higher rates of precise corrections and generated fewer indels (<1% on average; Fig. 5d and Supplementary Fig. 17c,d). These data suggested that Td-CGBE and eTd-CBEs were efficient for precise generation or correction of pathogenic SNVs, especially for precise editing of single nucleotides in polycytosine sites.

## Evaluation of eTd-CBEs by target library analysis

To unbiasedly evaluate the precision of eTd-CBEs, the gRNA–target pair strategy[38] was adapted to generate a library of 9,120 oligonucleotides individually composed of all possible 6-mers, where cytosines (number ≥1) were distributed in positions 4–9 of a protospacer (Methods). The Tol2 transposon was leveraged to stably integrate our library into the genome of HEK293T cells, before stable transfection of candidate CBEs. An average 96% coverage of greater than 300× per guide-target pair was maintained throughout the culturing process (Supplementary Table 6). The editing efficiency at the position with the highest activity in each target was defined as 100%, and the relative activity of other positions was determined through comparison with this position. The editing activity analyzed from three CBEs showed that BE4max-YE1 (evaluated for 8,949 sgRNAs) had a major editing window (>40%) ranging from positions 4–8, whereas eTd-CBE (evaluated for 8,737

https://doi.org/10.1038/s41587-022-01532-7

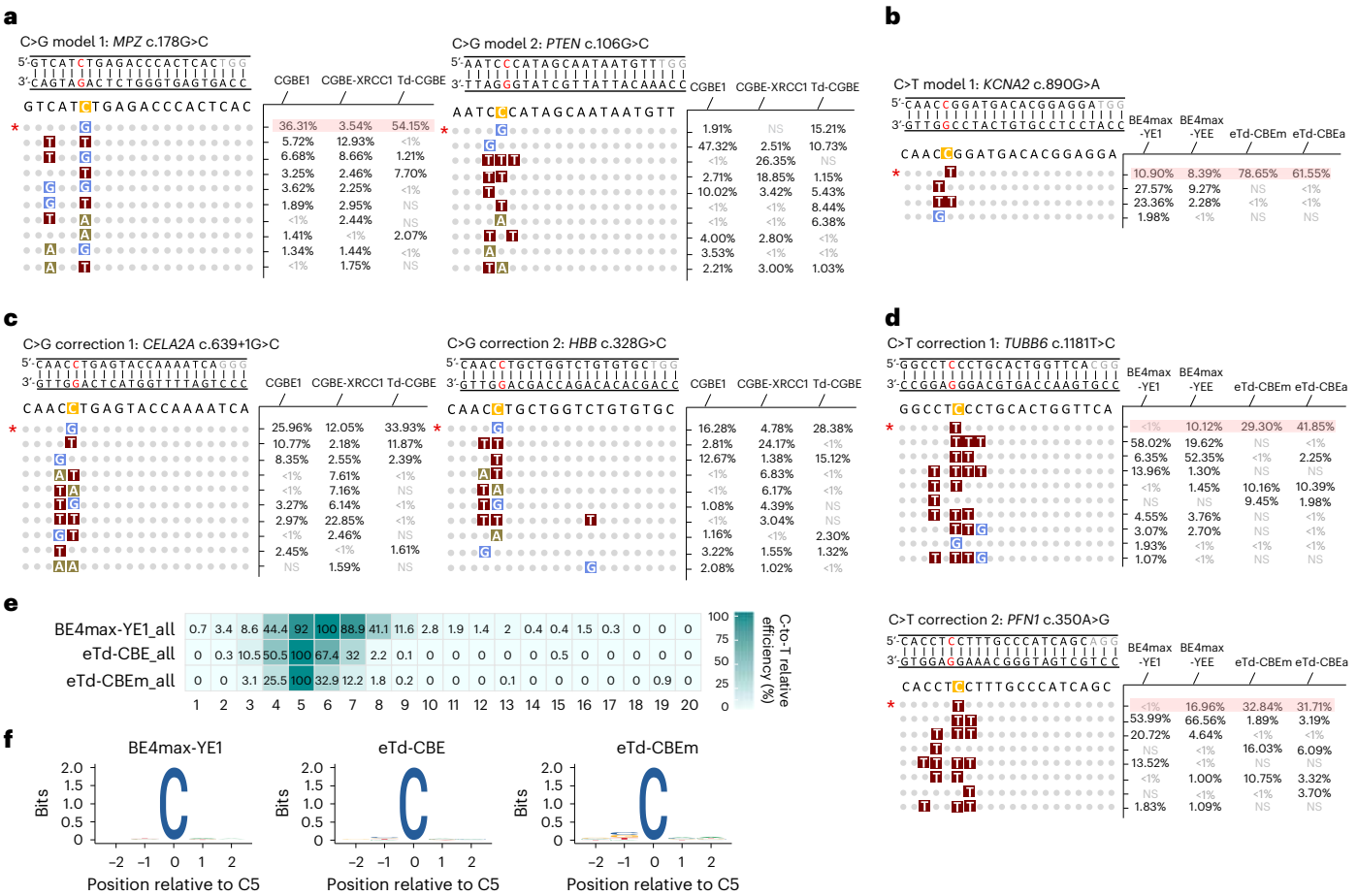


**Fig. 5 | Precise editing of pathogenic SNVs by TadA-derived base editors and target library analysis for eTd-CBEs. a**, Generation of C-to-G conversion of disease-relevant cell models (*MPZ* c.178G>C and *PTEN* c.106G>C) by CGBE1, CGBE-XRCC1 and Td-CGBE. **b**, Generation of C-to-T conversion of a disease-relevant cell model (*KCNA2* c.890G>A) by BE4max-YE1, BE4max-YEE, eTd-CBEm and eTd-CBEa. **c**, Correction of pathogenic mutations by the indicated CGBEs in stable cell lines (*CELA2A* c.639+1G>C and *HBB* c.328G>C). **d**, Correction of pathogenic mutations by the indicated CBEs in stable cell lines (*TUBB6* c.1181T>C and *PFN1* c.350A>G). **e**, Unbiased sgRNA target pair library analysis of the indicated CBEs. The heatmap represents the relative editing efficiency computed using the highest C-to-T base editing efficiency as 100%. Positions of the protospacer are shown at the bottom of the heatmap. **f**, Motif visualization of the indicated CBEs based on the data in **e** using sgRNAs with cytosine at position 5 of the protospacer. In **a**–**d**, protospacers of models or corrections are in black with the PAM in gray. Desired cytosines to be edited are indicated with yellow squares. The first line marked with an asterisk indicates desired alleles. The allele frequencies were determined by high-throughput sequencing. Data represent the mean of three independent experiments. NS, not significant.

sgRNAs) narrowed the window to positions 4–6 (Fig. 5e). As expected, an extremely condensed 1-nt window was observed for eTd-CBEm (evaluated for 8,522 sgRNAs) with the highest efficiency on position 5. The motif preferences of the eTd-CBEs were characterized in further analysis of ~2,700 targets containing C5. Similar to BE4max-YE1, these editors had the capability of a wide range of accurate C-to-T editing without a strict sequence context requirement (Fig. 5f). As eTd-CBEm preferentially edits cytosine in position 5 of protospacers without motif restrictions, it potentially corrects the majority of A-to-G pathogenic SNVs upon fusion to PAM-relaxed Cas9 variants, such as SpNG and SpRY[22,39], as we demonstrated for the Td-CGBE-NG variant.

## Discussion

Traditional CBEs are all based on AID/APOBEC family natural cytosine deaminases that endow CBE variants with distinct features. In this study, we developed an unnatural cytosine deaminase evolved from adenine deaminase TadA-8e and generated a series of TadA-derived base editors to enable highly efficient and accurate C-to-G or C-to-T conversions. Because TadA-8e is a DNA adenine deaminase variant evolved from the tRNA adenine deaminase TadA in *E. coli*, this study further demonstrates the potential of molecular engineering to advance genome editing tools.

A recent report demonstrated that a P48R variant in ABE7.10 could increase its cytosine editing efficiency in a restricted TC*N sequence context and decrease but not eliminate its adenine conversion activity[19]. Moreover, when the variant is introduced in ABE8e, substrate selectivity is severely impaired (Supplementary Fig. 1c). In this study, we have successfully converted the TadA enzyme into a pure cytosine deaminase through the introduction of an N46L variant in TadA-8e, suggesting that N46 is a very conserved residue critical for adenine deaminase activity. This is consistent with a previous study that reported that the N46A variant in miniABEmax fully abolished adenine editing activity[16], but the variant resulting from introduction of N46A in wild-type TadA of ABEmax still showed compromised activity[40]. In the TadA structure, N46 forms a weak hydrophilic interaction with the purine group or analogs of its tRNA or DNA substrates[18], whereas in cytosine deaminase structures the corresponding asparagine residue of APOBEC3A or APOBEC3B mainly forms polar contacts with the ribose group of its single-stranded DNA substrate[41]. Thus, from a structural point of view, the N46 residue contributes distinctly to recognition of different substrates. Moreover, based on the massive mutagenesis studies on N46, we found that editing adenine is highly fragile by variations on N46 and only the substitution to a similar residue (N46D) could preserve

the activity, suggesting that the hydrophilic interaction contributed by the side chain of N46 is essential for the deaminase activity of A-to-G conversion while the editing of cytosine is not sensitive to the majority of N46 substitutions. Substitution with several amino acids with small side chains (for example, cysteine, glycine, serine, threonine, valine or proline) led to even higher activity, while substitutions to amino acids with big side chains (for example, histidine, phenylalanine, arginine or tryptophan) mostly abrogates the cytosine deaminase activity, indicating that bulky residues may push the substrate cytosine away from the activation pocket (Supplementary Fig. 1a). Compared to other small residues, the N46L variant would provide an optimal conformation to allow deamination of cytosines but not adenines by disrupting the original adenosine deaminase structure to a dead enzyme.

We also found that the N46L variant could be partially applied to older versions (ABEmax or miniABEmax) to increase the cytosine editing efficiency within a TC*N motif, and we showed that the N46L variant in ABEmax and miniABEmax also abolished their A-to-G editing efficiency (Supplementary Fig. 2). In TadA-8e, the newly introduced eight substitutions are located far away from the N46 residue, and these substitutions may not affect the substrate selectivity determined by N46. Moreover, we found the N46L together with this ABE8e variant further enhanced the cytosine editing efficacy as well as expanded the editing scope (Supplementary Fig. 2). Further investigation of the substrate-bound structure for ABEmax as well as our TadA-derived base editors may help us to understand the exact mechanisms of how these mutations synergistically turned ABE8e into highly efficient TadA-derived base editors. The work described here extends our understanding of the versatile potential of the tRNA deaminase TadA, which has been evolved to efficiently catalyze adenine deamination in single-stranded DNA substrates and evolved here into a cytosine deaminase.

In this study, we have developed a distinct series of cytosine editors, Td-CGBE and Td-CBEs, derived from the TadA-8e adenine deaminase. They not only are cytosine editors without leveraging the AID/APOBEC family of deaminases, but also have distinct superiorities such as the lowest indel rate, greatly reduced bystander mutations and a background level of Cas9-independent DNA and RNA off-target effects. As they are able to efficiently generate single-cytosine conversion at homopolymeric cytosine sites without sequence context requirements in vitro and in vivo, Td-CGBE and Td-CBEs are promising accurate editors, expanding the targeting scope of precision base conversions when fused with PAM-relaxed Cas9 variants, for a wide range of applications.

## 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/s41587-022-01532-7.

## References


1. 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).
2. 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).
3. 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).
4. Wang, L. et al. Enhanced base editing by co-expression of free uracil DNA glycosylase inhibitor. *Cell Res.* **27**, 1289–1292 (2017).
5. Kurt, I. C. et al. CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells. *Nat. Biotechnol.* **39**, 41–46 (2021).
6. Zhao, D. et al. Glycosylase base editors enable C-to-A and C-to-G base changes. *Nat. Biotechnol.* **39**, 35–40 (2021).
7. Koblan, L. W. et al. Efficient C·G-to-G·C base editors developed using CRISPRi screens, target-library analysis, and machine learning. *Nat. Biotechnol.* **39**, 1414–1425 (2021).
8. Chen, L. et al. Programmable C:G to G:C genome editing with CRISPR-Cas9-directed base excision repair proteins. *Nat. Commun.* **12**, 1384 (2021).
9. Zuo, E. et al. Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science* **364**, 289–292 (2019).
10. Jin, S. et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292–295 (2019).
11. Zhou, C. et al. Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis. *Nature* **571**, 275–278 (2019).
12. Grunewald, J. et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **569**, 433–437 (2019).
13. Gehrke, J. M. et al. An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities. *Nat. Biotechnol.* **36**, 977–982 (2018).
14. Kim, Y. B. et al. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat. Biotechnol.* **35**, 371–376 (2017).
15. Kim, H. S., Jeong, Y. K., Hur, J. K., Kim, J. S. & Bae, S. Adenine base editors catalyze cytosine conversions in human cells. *Nat. Biotechnol.* **37**, 1145–1148 (2019).
16. Grunewald, J. et al. CRISPR DNA base editors with reduced RNA off-target and self-editing activities. *Nat. Biotechnol.* **37**, 1041–1048 (2019).
17. Chen, L. et al. Engineering precise adenine base editor with infinitesimal rates of bystander mutations and off-target editing. *Nat. Chem. Biol.* https://doi.org/10.1038/s41589-022-01163-8 (2022).
18. Lapinaite, A. et al. DNA capture by a CRISPR-Cas9-guided adenine base editor. *Science* **369**, 566–571 (2020).
19. Jeong, Y. K. et al. Adenine base editor engineering reduces editing of bystander cytosines. *Nat. Biotechnol.* **39**, 1426–1433 (2021).
20. Tan, J., Zhang, F., Karcher, D. & Bock, R. Engineering of high-precision base editors for site-specific single nucleotide replacement. *Nat. Commun.* **10**, 439 (2019).
21. Thuronyi, B. W. et al. Continuous evolution of base editors with expanded target compatibility and improved activity. *Nat. Biotechnol.* **37**, 1070–1079 (2019).
22. Nishimasu, H. et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262 (2018).
23. Zhang, X. et al. Increasing the efficiency and targeting range of cytidine base editors through fusion of a single-stranded DNA-binding protein domain. *Nat. Cell Biol.* **22**, 740–750 (2020).
24. Lee, S. et al. Single C-to-T substitution using engineered APOBEC3G-nCas9 base editors with minimum genome- and transcriptome-wide off-target effects. *Sci. Adv.* **6**, eaba1773 (2020).
25. 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).
26. Doman, J. L., Raguram, A., Newby, G. A. & Liu, D. R. Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors. *Nat. Biotechnol.* **38**, 620–628 (2020).
27. Wang, L. et al. Eliminating base-editor-induced genome-wide and transcriptome-wide off-target mutations. *Nat. Cell Biol.* **23**, 552–563 (2021).
28. Lei, Z. et al. Detect-seq reveals out-of-protospacer editing and target-strand editing by cytosine base editors. *Nat. Methods* **18**, 643–651 (2021).

29. Lei, Z. et al. Mitochondrial base editor induces substantial nuclear off-target mutations. *Nature* **606**, 804–811 (2022).
30. Yuan, T. et al. Optimization of C-to-G base editors with sequence context preference predictable by machine learning methods. *Nat. Commun.* **12**, 4902 (2021).
31. Auer-Grumbach, M., Strasser-Fuchs, S., Robl, T., Windpassinger, C. & Wagner, K. Late onset Charcot-Marie-Tooth 2 syndrome caused by two novel mutations in the *MPZ* gene. *Neurology* **61**, 1435–1437 (2003).
32. Rodriguez-Escudero, I. et al. A comprehensive functional analysis of *PTEN* mutations: implications in tumor- and autism-related syndromes. *Hum. Mol. Genet.* **20**, 4132–4142 (2011).
33. Syrbe, S. et al. De novo loss- or gain-of-function mutations in *KCNA2* cause epileptic encephalopathy. *Nat. Genet.* **47**, 393–399 (2015).
34. Esteghamat, F. et al. *CELA2A* mutations predispose to early-onset atherosclerosis and metabolic syndrome and affect plasma insulin and platelet activation. *Nat. Genet.* **51**, 1233–1243 (2019).
35. Ropero, P. et al. Hb Johnstown [β 109 (G11) Val–>Leu]: second case described and associated for the first time with $\beta_0$-thalassemia in two Spanish families. *Am. J. Hematol.* **65**, 298–301 (2000).
36. Fazeli, W. et al. A *TUBB6* mutation is associated with autosomal dominant non-progressive congenital facial palsy, bilateral ptosis and velopharyngeal dysfunction. *Hum. Mol. Genet.* **26**, 4055–4066 (2017).
37. Wu, C. H. et al. Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. *Nature* **488**, 499–503 (2012).
38. Arbab, M. et al. Determinants of base editing outcomes from target library analysis and machine learning. *Cell* **182**, 463–480 (2020).
39. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**, 290–296 (2020).
40. Li, J. et al. Structure-guided engineering of adenine base editor with minimized RNA off-targeting activity. *Nat. Commun.* **12**, 2287 (2021).
41. Shi, K. et al. Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. *Nat. Struct. Mol. Biol.* **24**, 131–139 (2017).

# Methods

## Plasmid construction

The PCR primers and DNA sequences used in this research are listed in Supplementary Tables 1–4 and Supplementary Sequences 1–3. ABE8e (138489), pCMV_BE4max (112093), CGBE1 (140252) and lentiCRISPR v2 (52961) were purchased from Addgene. A3G-BE5.13 was a gift from the laboratory of Erwei Zuo (Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China). PCR was performed using PrimeSTAR Max DNA polymerase (TaKaRa, R045A) or KOD-Plus-Neo DNA polymerase (Toyobo, KOD-401). A series of ABE or CBE variants were constructed using the ClonExpress MultiS One Step cloning kit (Vazyme). Td-CGBE-NG with non-NGG PAM compatibility (Cas9-NG) was generated by introducing corresponding mutations via PCR. Human codon-optimized XRCC1 (named CGBE-XRCC1) used in plasmid construction was synthesized by GENEWIZ Biotechnology and was cloned into the backbone of a digested pCMV_BE4max vector. Two copies of UGI with the P2A peptide were cloned into the C-terminal end of serial Td-CBEs, and canonical CBEs were also repurposed using the same method for a fair comparison. gRNA expression plasmids were constructed by using the previous method[23]. In brief, a pair of oligonucleotide duplexes was ligated and annealed into U6-sgRNA(sp)-EF1α-GFP linearized with BbsI (Thermo Fisher Scientific). Plasmid DNA was isolated using the TIANprep Mini Plasmid Kit (TIANGEN Biotech, DP103-03) according to the manufacturer's instructions.

## Cell culture

HEK293T (ATCC, CRL-3216) cells were cultured and expanded in DMEM (Gibco) supplemented with 10% (vol/vol) FBS (Gibco) and 1% (vol/vol) penicillin-streptomycin (Gibco) in a humidified incubator at 37 °C with 5% $CO_2$.

## Cell transfection and genomic DNA extraction

For DNA on- or off-target base editing experiments, HEK293T cells were seeded into 24-well plates (Corning) at approximately 80% confluency per well. After 12 h, cells were cotransfected with 750 ng of base editor plasmids and 250 ng of gRNA expression plasmids using PEI (Polysciences) following the manufacturer's recommended protocol. After 3 d, cells were washed with PBS and digested with 0.25% trypsin (Gibco) for sorting. Then, cell genomic DNA was isolated using the QuickExtract DNA Extraction Solution (QE09050, Epicentre), and mouse tail tip genomic DNA was isolated using the One Step Mouse genotyping kit (Vazyme) according to the manufacturer's instructions.

## Western blotting assay

The western blot assays were performed as previously described[42]. HEK293T cells were lysed 3 d after transfection using RIPA buffer complemented with proteinase and phosphatase inhibitors (Calbiochem). The total protein concentrations of cell lysate supernatants were quantified using a BCA protein assay kit (Thermo Fisher Scientific). In total, 12 μg of total protein per well was submitted to electrophoresis using a 15-well 10% SDS–polyacrylamide gel and transferred to nitrocellulose membranes (Millipore) for 150 min at 100 V. Subsequently, the membranes were blocked with 5% BSA for 1 h at room temperature and then divided and processed with different primary antibodies separately overnight, including anti-tubulin (1:2,000 dilution; Abcam, ab210797) and the anti-CRISPR–Cas9 (1:10,000 dilution; Abcam, ab189380). Then, the membranes were incubated with goat anti-rabbit IgG H&L (IRDye 800CW, 1:10,000 dilution; Abcam, ab216773) for 1 h and visualized using the Odyssey infrared imaging system (Odyssey, LI-COR). Uncropped blots are shown in Supplementary Fig. 18.

## Orthogonal R-loop assay

The modified orthogonal R-loop assay at each R-loop site with a nSaCas9-sgRNA plasmid was used in the analysis of Cas9-independent DNA off-target editing. Three hundred nanograms of an SpCas9 gRNA plasmid, 400 ng of a base editor plasmid and 300 ng of a nSaCas9-sgRNA plasmid were cotransfected into HEK293T cells using polyethyleneimine (PEI). After 3 d, cells were washed with PBS and digested with 0.25% trypsin (Gibco). Genomic DNA was extracted using the QuickExtract DNA Extraction Solution (QE09050, Epicentre) according to the manufacturer's instructions.

## Detect-seq experiments

HEK293T cells seeded on six-well culture plates (Corning) were transfected at approximately 70% confluency with 4 μg of the base editor and 2.72 μg of sgRNA plasmids using Lipofectamine LTX (Thermo Fisher Scientific). Transfected cells were collected after 5 d, and genomic DNA was isolated using the CWBIO universal genomic DNA kit (CWbiotech, CW2298M) for Detect-seq. This study performed Detect-seq experiments as previously described[28]. DNA damage that may interfere with true signals was repaired or protected to reduce the background noise. DNA fragments containing CBE-induced deoxyuridine bases were recognized by UDG and labeled via nick translation and subsequent chemical reactions. The biotin-labeled fragments were enriched by streptavidin C1 beads (Invitrogen) and ligated with Y adaptors. Detect-seq libraries were sequenced on Illumina Hiseq X Ten and MGISEQ-2000. The efficiency and specificity of Detect-seq were evaluated by qPCR and Sanger sequencing on spike-in molecules.

## Detect-seq mapping and analysis

Detect-seq data were mapped and processed with the steps described in the previous paper[28]. In brief, the adaptor was removed from sequencing reads by Cutadapt (version 1.18), and then, reads were mapped with Bismark (version 0.22.3) with default settings to the human reference genome (hg38). The reads with MAPQ lower than 20 were collected and remapped by BWA MEM (version 0.7.17) using default parameters. The mapping results were merged and sorted with samtools sort command (version 1.9). The PCR duplications were removed by Picard MarkDuplicates (version 2.23.9). For assessment of genome-wide off-target sites, we first processed alignment results into mpileup files by samtools mpileup command (version 1.9) with -q 20 -Q 20 parameters and then converted into.pmat files by parse-mpileup and bmat2pmat commands with default settings. We next searched genome-wide tandem C-to-T signals by pmat-merge command and filtered by mpmat-select script with settings as -m 4 -c 6 -r 0.01—RegionPassNum 1 to RegionToleranceNum 3. Those filtered signals were collected to run a Poisson one-sided test by find-significant-mpmat script. *P* values were adjusted with the Benjamini and Hochberg method to control the false discovery rate. Signal regions that complied with the following criteria were considered a candidate Cas-dependent off-target site: FDR < 0.05; fold change of normalized mutation count in the CBE-PD sample to normalized mutation count in the mCherry sample larger than 2; mutation count in the mCherry sample no larger than 1; and a mutation count in the All-PD sample no less than 5. All scripts used in this step were collected into the Detect-seq tools. Detect-seq genome-wide distribution off-target sites are listed in Supplementary Table 5.

## Total mRNA preparation and sequencing assays

Total mRNA preparation and RNA-seq assays were performed as previously described[23]. For the RNA-seq experiment, HEK293T cells were seeded into 10-cm dishes (Corning). Twenty-five micrograms of Cas9n-T2A-GFP, BE4max-T2A-GFP, BE4max-YE1-T2A-GFP, Td-CGBE-T2A-GFP, Td-CBEmax-T2A-GFP, eTd-CBE-T2A-GFP, eTd-CBEm-T2A-GFP and eTd-CBEa-T2A-GFP was transfected into HEK293T cells at approximately 80% confluency using PEI following the manufacturer's instructions. After 3 d, transfected cells were washed with PBS and digested with 0.25% trypsin (Gibco) for fluorescence-activated cell sorting (FACS) using FACSDiva version 8.0.2 (BD Biosciences). Approximately, $4 \times 10^5$ GFP-positive cells of the top 15% (Supplementary Note) were collected and RNA was extracted according to the standard protocols. A total of 3 μg RNA per sample

was used as input material to prepare a library. The resulting library was sequenced on an Illumina Hiseq platform, and 125-bp/150-bp paired-end reads were gathered.

### Transcriptome-wide off-target analysis

For the RNA-seq analysis, sequencing reads were first analyzed with Trim Galore to remove adaptor sequences (version 0.6.6) and then aligned with STAR (version 2.7.1a)[43] to the human reference genome (hg38). Aligned BAM files were sorted with samtools (version 1.9)[44], and duplications were removed with Picard MarkDuplicates (version 2.23.9). Then, the clean BAM files were converted to the mpileup format with samtools (version 1.9). The variant information was extracted based on mpileup files, and only sites with coverage higher than 30 and variant count over 15 were collected in the next step. Only sites that did not exist in Cas9n samples were reported as final mutation sites.

### Generation of stable cell line disease models

The HEK293T stable cell lines were established by cloning a 150-bp disease-associated gene fragment of G·C-to-C·G or T·A-to-C·G from the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) by assembling the fragments into a modified lentivector from lentiCRISPR v2 (52961), obtaining lentivector plasmids (Lenti *CELA2A*-EF1α-DsRed-P2A-puro, Lenti *HBB*-EF1α-DsRed-P2A-puro, Lenti *TUBB6*-EF1α-DsRed-P2A-puro or Lenti *PFN1*-EF1α-DsRed-P2A-puro). HEK293T cells were seeded into 24-well plates (Corning) at approximately 80% confluency per well and were cotransfected with 300 ng of the lentivector plasmid, 300 ng pMD2.G (12259) and 300 ng psPAX2 (12260) using PEI following the manufacturer's instructions. Virus-containing supernatant was collected after 48 h of transfection and filtered with a 0.45-μm polyvinylidene difluoride filter membrane (Millipore); then, 30 μl filtered virus-containing supernatant was added to HEK293T cells at approximately 40–50% confluency cultured in 12-well plates. After 24 h of transduction with lentivirus, cells were split into wells of a new plate supplemented with puromycin (1 μg $ml^{-1}$). Seventy-two hours after the puromycin selection, cells were collected with the fewest surviving colonies to ensure single-copy integration and then expanded for further transfection.

### Animal care and microinjection of zygotes

Animal manipulation was accomplished as previously described[45]. In brief, C57BL/6J and ICR mouse strains purchased from the Shanghai Laboratory Animal Center were caged in a specific pathogen-free facility on a 12-h light and 12-h dark cycle with ample access to food and water. C57BL/6J and ICR mouse strains were reared as embryo donors and foster mothers, separately. All animal experiments complied with draft regulations by the Association for Assessment and Accreditation of Laboratory Animal Care in Shanghai and were ratified by the East China Normal University Center for Animal Research. Chemically modified sgRNA was synthesized by Genscript. mRNA preparation was performed as previously described. The T7 promoter to the coding region was introduced into the Td-CGBE template by PCR amplification using primers T7-mRNA (Td-CGBE)-F/R (Supplementary Table 2), and its purified PCR product was used as the transcription template using an in vitro RNA transcription kit (mMESSAGE mMACHINE T7 Ultra Kit, Ambion). For microinjection, solutions containing complexes of Td-CGBE mRNA (100 ng $\mu l^{-1}$) and sgRNA (200 ng $\mu l^{-1}$) were diluted with nuclease-free water and the mixture was Injected into the cytoplasm using an Eppendorf TransferMan NK2 micromanipulator. Injected zygotes were transferred into ICR pseudopregnant females immediately after injection.

### Design and cloning of the library

The composition of the oligonucleotides used to express gRNA pair libraries was designed as previously described[38]. Briefly, each oligo contains a full-length oligonucleotide followed by a pairing cassette targeted by the sgRNA. The spacers of the sgRNAs were designed to satisfy the following principles: first, positions 4–9 of the spacer consist of all possible 6-mers with at least one cytidine. Random 2-mers and 11-mers flanked the 6-mers at the 5′ and 3′ end, respectively. Second, each spacer is initiated with a guanine. Last, spacers with four consecutive thymine bases were avoided because this might hinder transcription. Each targeted cassette consists of a 20-bp target sequence preceding an NGG PAM. The target sequence was flanked by randomly selected wild-type human genomic sequences. Libraries were cloned into a modified pBlueScript backbone supporting sgRNA expression, hygromycin selection and Tol2 transposon-mediated genomic integration. The library construction was accomplished by GENEWIZ Biotechnology.

### Transduction of the library and cell culture

HEK293T cells were seeded into 10-cm plates (Corning) at approximately 90% confluency; the Tol2 transposase plasmid (10 μg) and library mixture (10 μg) were cotransfected using PEI. Cells were cultured with hygromycin B (25 μg $ml^{-1}$) (Thermo Fisher Scientific, 10687010) beginning 1 d after transfection to expedite sgRNA-stable selection and lasting for >2 weeks, during which timeover 90% of cells were screened out. When the library cell lines reached approximately 90% confluency again, we performed the second round of genomic integration by cotransfection of a Tol2 transposase plasmid (10 μg) and a base editor plasmid (BE4max-YE1, eTd-CBE or eTd-CBEm; 10 μg) that contains a blasticidin resistance gene and Tol2 transposase-binding sites. For the second-round selection, 10 μg $ml^{-1}$ blasticidin S HCl (Thermo Fisher Scientific, A11139-03) was used the day after transfection, lasting for >2 weeks. When the density of cells was up to 90%, to maintain at least a 300-fold average coverage of each gRNA per library cassette, transfected cells were washed with PBS and digested with 0.25% trypsin (Gibco) for genomic DNA extraction by using phenol-chloroform extraction according to the manufacturer's instructions. The target regions of libraries were amplified from genomic DNA (100–200 ng) by PCR with the primers listed in Supplementary Table 2. The amplified products were submitted to the high-throughput sequencing (HTS) platform of GENEWIZ Biotechnology. The complete sequences of the plasmid constructs are listed in Supplementary Sequence 3.

### Editing efficiency calculation of the library and motif visualization

The JavaScript version of fastq-join (https://github.com/brwnj/fastq-join) first joins two fastq files from HTS. To determine the connection between the amplicons and the sequenced reads, the combined fastq files were aligned to all of the amplicons in the library using BWA-mem (0.7.17-r1188) and the reads were divided for each amplicon. The random mode was used for reads with many equally plausible alignments. To minimize the impact of PCR amplification, targets with sequencing depth more than 10 times higher than the average depth of the library were discarded for every library. Library sequencing depth and coverage are shown in Supplementary Table 6. Then, EMBOSS needle (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) was used to align all of the reads to their corresponding amplicon pairwise. Only the reads that matched the following criteria were included for analysis: 10-bp sequences upstream and downstream of the 20-bp target sites that perfectly matched the consensus sequences and no indels or degenerate base Ns detected in the target sites. The total number of reads aligned to amplicons, the editing types and the number of edited reads at each position were then calculated for each type's absolute editing efficiency at each site. The relative editing efficiency was then computed relative to the highest absolute editing efficiency. The effectiveness of sgRNA varied greatly, and the matching sgRNA was accumulated once for each edited read when enriching motifs. The motifs edited at C5 were tallied, and the pattern was visualized using the ggseqlogo package in R.

### Targeted deep sequencing and data analysis

On- and off-target genomic regions were amplified from genomic DNA (~100 ng) from three biological replicates for each condition by PCR with the primers listed in

Supplementary Tables 2 and 4. HTS amplicon libraries containing an adaptor sequence (forward, 5′-GGAGTGAGTACGGTGTGC-3′; reverse 5′-GAGTTGGATGCTGGATGG-3′) at the 5′ end were prepared by PCR using KOD-Plus-Neo DNA Polymerase (Toyobo, KOD-401). The above products were subjected to another round of PCR amplification with different barcode sequences in the primers, and then, the resulting libraries were mixed and sequenced with 150-bp paired-end reads on an Illumina HiSeq platform. The A-to-G or C-to-G, C-to-T and C-to-A conversions and indels in the HTS data were analyzed using BE-Analyzer[46].

### Statistical analysis and reproducibility

Data presentations were generated by using GraphPad Prism v.9.3.1. Data were calculated as the mean ± s.d. from three biologically independent replicates unless stated otherwise in the figure captions.

### Reporting summary

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

## Data availability

HTS data have been deposited in the NCBI Sequence Read Archive database under accession codes PRJNA822038, PRJNA871961, PRJNA855334, PRJNA835691, PRJNA835701 and PRJNA882574 (refs. [47–52]). RNA-seq data have been deposited in the NCBI Sequence Read Archive database under accession codes PRJNA871962 and PRJNA830998 (refs. [53,54]). There are no restrictions on data availability. Source data are provided with this paper.

## Code availability

The relevant codes of analysis for Detect-seq data were deposited in GitHub (https://github.com/menghaowei/Detect-seq)[55].

## References


42. Zhang, X. et al. Dual base editor catalyzes both cytosine and adenine base conversions in human cells. *Nat. Biotechnol.* **38**, 856–860 (2020).
43. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
44. Kruger, S. et al. Sensorimotor polyneuropathy and systemic amyloidosis as paraneoplastic symptoms of a carcinoid-like well differentiated carcinoma of the breast. *Dtsch. Med. Wochenschr.* **123**, 179–184 (1998).
45. Li, D. et al. Heritable gene targeting in the mouse and rat using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 681–683 (2013).
46. Hwang, G. H. et al. Web-based design and analysis tools for CRISPR base editing. *BMC Bioinformatics* **19**, 542 (2018).
47. Chen, L. et al. Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing. NCBI SRA, BioProject PRJNA822038 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA822038 (2022).
48. Chen, L. et al. Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing. NCBI SRA, BioProject PRJNA871961 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA871961 (2022).
49. Chen, L. et al. Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing. NCBI SRA, BioProject PRJNA855334 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA855334 (2022).
50. Chen, L. et al. Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing. NCBI SRA, BioProject PRJNA835691 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA835691 (2022).
51. Chen, L. et al. Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing. NCBI SRA, BioProject PRJNA835701 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA835701 (2022).
52. Chen, L. et al. Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing. NCBI SRA, BioProject PRJNA882574 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA882574 (2022).
53. Chen, L. et al. Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing. NCBI SRA, BioProject PRJNA871962 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA871962 (2022).
54. Chen, L. et al. Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing. NCBI SRA, BioProject PRJNA830998 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA830998 (2022).
55. Chen, L. et al. Td-BE-Detect-seq analysis. GitHub https://github.com/menghaowei/Detect-seq (2022).


## Acknowledgements


We are grateful to the East China Normal University Public Platform for Innovation (011). We thank Y. Zhang from the Flow Cytometry Core Facility of the School of Life Sciences in ECNU and H. Jiang from the Core Facility and Technical Service Center for the SLSB of the School of Life Sciences and Biotechnology in SJTU. We thank L. Ji (MedSci) for designing schematic diagrams. This work is partially supported by grants from the National Key R&D Program of China (2019YFA0802800 to M.L., 2019YFA0110802 to D.L., 2019YFA0802200 to C.Y., and 2019YFA0110900 to C.Y.), the National Natural Science Foundation of China (32025023 to D.L., 32230064 to D.L., 31971366 to L.W., 82230002 to M.L., 21825701 to C.Y., 91953201 to C.Y. and 92153303 to C.Y.), the Shanghai Municipal Commission for Science and Technology (21CJ1402200 to D.L. and 20140900200 to D.L.), and the Innovation Program of the Shanghai Municipal Education Commission (2019-01-07-00-05-E00054 to D.L.), the Fundamental Research Funds for the Central Universities (NK2022010207 to D.L.), the State Key Laboratory of Drug Research (SIMM2205KF-01 to C.Y.) and support from the East China Normal University Outstanding Doctoral Students Academic Innovation Ability Improvement Project (YBNLTS2021-026 to L.C.).


## Author contributions

L.C. and D.L. designed the experiments. L.C., B.Z., G.R., H.M., Y.Y., M.H., C.L., S.Z., H.G., S.B., C.L., R.D. and N.X. performed the experiments. L.C., B.Z., G.R., H.M., Y.Y., M.H., D.Z., C.L., H.W., C.L., Z.L., Y.C., Y.G., S.S., C.Y., G.S., L.W., C.Y., M.L. and D.L. analyzed the data. D.L., L.C., B.Z., G.R., H.M., C.Y. and G.S. wrote the manuscript with inputs from all the authors. D.L. supervised the research.

## Competing interests

The authors have submitted patent applications based on the results reported in this study (L.C., D.L., G.R., C.L., H.G., B.Z., J.Y., S.B., R.D. and M.L.). The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01532-7.


**Correspondence and requests for materials** should be addressed to Chengqi Yi, Mingyao Liu or Dali Li.



**Peer review information** *Nature Biotechnology* thanks Francisco Sanchez-Rivera and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature research

Corresponding author(s): Chengqi Yi, Mingyao Liu and Dali Li

Last updated by author(s): Sep 26, 2022

# 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 our Editorial Policies 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<br>*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<br>*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 | Targeted amplicons sequencing data were collected and demultiplexed by an Illumina HiSeq X Ten instrument.<br>RNA-seq data were collected and demultiplexed by an Illumina NovaSeq 6000 instrument.<br>Detect-seq data were collected and demultiplexed by Illumina Hiseq X Ten and MGISEQ-2000.<br>Library amplicons data were collected and demultiplexed by an Illumina NovaSeq 6000 instrument.<br>FACS gating data were collected on a FACSAria III (BD Biosciences) using FACSDiva version 8.0.2 (BD Biosciences). |
| Data analysis | High-throughput sequencing data were analyzed by BE-Analyzer (http://www.rgenome.net/be-analyzer/#!) (Hwang G-H et al, BMC Bioinformatics, 2018) for base editing (A>G, C>T, C>G and C>A) and indels efficiencies.<br>Potential DNA off-target sites for indicated base editors were predicated using Cas-OFFinder web (http://www.rgenome.net/cas-offinder/).<br>RNA-seq data were analyzed using Trim Galore (version 0.6.6), STAR (version 2.7.1a), SAMtools (version 1.9) and Picard MarkDuplicates module (version 2.23.9) software.<br>Detect-seq were analyzed using cutadapt (version 1.18), Bismark (version 0.22.3), BWA MEM (version 0.7.17), samtools (version 1.9) and Picard (version 2.23.9). All custom code used for analysis are available at https://github.com/menghaowei/Detect-seq.<br>Library amplicons data were analyzed using BWA (version 0.7.17-r1188) and EMBOSS needle software.<br>GraphPad Prism 9.3 was also used to analyze data. |

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 and 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

High–throughput sequencing reads have been deposited in the NCBI Sequence Read Archive database under accession code PRJNA822038, PRJNA871961, PRJNA855334, PRJNA835691, PRJNA835701, PRJNA882574. RNA-seq data have been deposited in the NCBI Sequence Read Archive database under accession code PRJNA871962, PRJNA830998. All custom codes of analysis for Detect-seq data were deposited on GitHub (https://github.com/menghaowei/Detect-seq). All Plasmids sequences or target site sequences mentioned in the manuscript are provided in supplementary information file. There are no restrictions on data availability.

# 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 | No statistical methods were used to predetermine sample size. Experiments were performed in biological triplicate n=3 unless otherwise noted. Sample sizes were opted to display the range and consistency of differences and three biological replicates made it sufficient to support the conclusions in this research. |
| Data exclusions | No data were excluded from the analysis. |
| Replication | Three independent biological replicates were performed on different days unless stated otherwise. All replications were successful. |
| Randomization | Samples were randomly distributed into groups. |
| Blinding | Investigators were not blinded to group allocation in this research since experimental conditions were evident and all samples of treatment were consistent throughout 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 and archaeology |
| ☐ | ☒ Animals and other organisms |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

### Methods

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

## Antibodies

| | |
|---|---|
| Antibodies used | a 1:2000 dilution of anti-Tubulin (Abcam, ab210797)<br>a 1;10,000 dilution of anti-CRISPR-Cas9 (Abcam, ab189380). |
| Validation | The Tubulin (ab210797, Abcam) antibody has been validated by western blot in HEK293T cells lysate (hhttps://www.abcam.cn/tubulin-antibody-epr13796-loading-control-ab210797.html).<br>The Cas9 (ab189380, Abcam) antibody has been validated by western blot in HEK293T cells transfected with CRISPR-Cas9 |

(Q99ZW2, Streptococcus pyogenes serotype M1) with GFP-Myc tag (https://www.abcam.com/crispr-cas9-antibody-epr18991-ab189380.html).

## Eukaryotic cell lines

Policy information about cell lines

| | |
|---|---|
| Cell line source(s) | HEK293T cells (purchased from ATCC CRL-3216). |
| Authentication | HEK293T cells are from authenticated manufacturers. |
| Mycoplasma contamination | All cell lines used were tested negative for mycoplasma contamination. |
| Commonly misidentified lines (See ICLAC register) | No commonly misidentified cell lines were used. |

## Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

| | |
|---|---|
| Laboratory animals | C57BL/6J, ICR mouse strain purchased from Shanghai Laboratory Animal Center were housed in standard cages at 20-22°C with 40-60% humidity in a specific pathogen-free facility on a 12 h light/dark cycle with ad libitum access to food and water. 4-6 week-old female wild-type C57BL/6J mouse and 6-12 week-old female wild-type ICR mouse strain were used as embryo donors and foster mothers, respectively. |
| Wild animals | No studies with wild animals were performed. |
| Field-collected samples | No studies with field-collected samples were performed. |
| Ethics oversight | All animal experiments conformed to the regulations drafted by the Association for Assessment and Accreditation of Laboratory Animal Care in Shanghai and were approved by the East China Normal University Center for Animal Research. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Flow Cytometry

### Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

| | |
|---|---|
| Sample preparation | Cell culture and transfection procedures are described in the methods. Cells were washed and filtered through a 45µm cell strainer cap before sorting (72h after transfection). |
| Instrument | FACSAria III (BD Biosciences) |
| Software | BD FACSDiva Software Diva 8.0.2 |
| Cell population abundance | The cells were subjected to FACS sorting step via GFP signals. The density of collected cells are about 500,000 cells. HEK293T cell numbers gated for target populations were similar in different biology replicates. |
| Gating strategy | For HEK293T cells, gates were established using uninfected control cells and GFP positive control. Gates were drawn to collect subsets of GFP- expressing cells. And for specified transcriptome profiling, cells with top 15% of GFP signal were sorted, after gating for the cell population (~15% of parent). Detailed gating strategy is provided in the Supplementary Note. |

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.