LETTER

https://doi.org/10.1038/s41586-019-1314-0

Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis

Changyang Zhou[1,2,9], Yidi Sun[2,3,4,9], Rui Yan[5,9], Yajing Liu[2,6,9], Erwei Zuo[1,7,9], Chan Gu[5], Linxiao Han[1], Yu Wei[1], Xinde Hu[1,2], Rong Zeng[3,6], Yixue Li[5,6,8]*, Haibo Zhou[1]*, Fan Guo[5]* & Hui Yang[1]*

Recently developed DNA base editing methods enable the direct generation of desired point mutations in genomic DNA without generating any double-strand breaks[1–3], but the issue of off-target edits has limited the application of these methods. Although several previous studies have evaluated off-target mutations in genomic DNA[4–8], it is now clear that the deaminases that are integral to commonly used DNA base editors often bind to RNA[9–13]. For example, the cytosine deaminase APOBEC1—which is used in cytosine base editors (CBEs)—targets both DNA and RNA[12], and the adenine deaminase TadA—which is used in adenine base editors (ABEs)—induces site-specific inosine formation on RNA[9,11]. However, any potential RNA mutations caused by DNA base editors have not been evaluated. Adeno-associated viruses are the most common delivery system for gene therapies that involve DNA editing; these viruses can sustain long-term gene expression in vivo, so the extent of potential RNA mutations induced by DNA base editors is of great concern[14–16]. Here we quantitatively evaluated RNA single nucleotide variations (SNVs) that were induced by CBEs or ABEs. Both the cytosine base editor BE3 and the adenine base editor ABE7.10 generated tens of thousands of off-target RNA SNVs. Subsequently, by engineering deaminases, we found that three CBE variants and one ABE variant showed a reduction in off-target RNA SNVs to the baseline while maintaining efficient DNA on-target activity. This study reveals a previously overlooked aspect of off-target effects in DNA editing and also demonstrates that such effects can be eliminated by engineering deaminases.

To evaluate the off-target effects of base editors at the RNA level, we transfected one type of CBE (BE3; APOBEC1–nCas9–UGI) or ABE (ABE7.10; TadA–TadA*–nCas9), together with GFP and with or without a single guide RNA (sgRNA) into cultured HEK293T cells (Fig. 1a, Extended Data Fig. 1). First, we validated the high on-target efficiency of DNA editing by both BE3 and ABE7.10 in these HEK293T cells using Sanger sequencing (Fig. 1b–e). We next performed RNA sequencing (RNA-seq) at an average depth of 125× on these samples (Supplementary Table 1). RNA SNVs were called from the RNA-seq data in each replicate separately, and any SNV identified in a wild-type sample was filtered out (Fig. 1a).

We found 742 ± 113 (mean ± s.e.m.) RNA SNVs in the GFP-alone control cells (Fig. 1f–h, Supplementary Table 2), but observed notably higher numbers of RNA SNVs in cells from the following sample groups: APOBEC1, BE3 without sgRNA, and BE3 with either site 3 (see Methods) or ring finger protein 2 (RNF2) sgRNA (5–40 times that in GFP-only cells; Fig. 1f, h, Extended Data Fig. 2, Supplementary Tables 2, 3). Similarly, large numbers of RNA SNVs (5–10 times those in GFP-only cells) were also found in cells expressing only TadA–TadA*, ABE7.10 without sgRNA, and ABE7.10 with either site 1 or site 2 sgRNA (Fig. 1g, h, Extended Data Fig. 2, Supplementary Tables 2, 3). Notably, transfection of APOBEC1 or TadA–TadA* induced the greatest numbers of RNA SNVs compared to other transfected groups, which implies that increased SNVs in CBE- or ABE-treated cells are likely to be caused by overexpression of the deaminase APOBEC1 or TadA (Fig. 1f, g, Extended Data Fig. 2, Supplementary Tables 2, 3). Moreover, the number of off-target RNA SNVs was increased when higher levels of CBEs or ABEs were expressed (Extended Data Fig. 2).

Notably, nearly 100% of the RNA SNVs identified in BE3-treated cells were mutations from G to A or C to U; this level is significantly higher than that in the GFP-alone control cells ($P = 2.03 \times 10^{-10}$ for BE3, $P = 0.017$ for BE3–site 3 and $P = 5.90 \times 10^{-10}$ for BE3–RNF2) (Fig. 2a, c, Extended Data Fig. 3). This mutation bias was the same as that of APOBEC1 itself[2], which indicates that these mutations were not spontaneous but rather were induced by BE3 or APOBEC1. Similarly, 95% of the ABE7.10-induced mutations were A to G or U to C, consistent with the action of TadA (Fig. 2b, c, Extended Data Fig. 3). We noted that the GFP group also exhibited bias towards A to G and U to C mutations (Fig. 2c), probably owing to innate mutation preferences[17–19]. We observed 27.7 ± 3.6% (mean ± s.e.m.) or 51.0 ± 3.3% (mean ± s.e.m.) overlap between any two samples from the BE3- or ABE7.10-transfected groups, respectively, and these overlapping SNVs were substantially enriched for genes with high levels of expression (Fig. 2d, Extended Data Fig. 3). Additionally, we found that a consensus motif ACW (W = A or U) or TAW (W = C or T) typically occurred in BE3- or ABE7.10-induced RNA SNVs, respectively (Fig. 2e). However, none of the off-target sites overlapped with predicted off-target mutations and we observed no similarities between the off-target and on-target sequences (Fig. 2d, Extended Data Fig. 4). Thus, the off-target RNA SNVs induced by the CBE and ABE were independent of sgRNA and caused by overexpression of APOBEC1 and TadA–TadA*, respectively. By Sanger sequencing validation, we found that these SNVs could be detected only in RNAs and not in DNA (Extended Data Fig. 4). Moreover, the off-target RNA SNVs were found in both coding and non-coding sequences (a substantial percentage in 3′ UTRs and exonic regions for BE3 and ABE7.10, respectively; Extended Data Fig. 5). In addition, ABE7.10 induced 56 and 12 non-synonymous RNA SNVs in oncogenes and tumour suppressor genes, respectively, and many of these showed an editing rate higher than 40%, raising concern about the oncogenic risk of DNA base editing (Extended Data Fig. 5, Supplementary Tables 4, 5).

Bulk RNA-seq is based on large pools of cells with variable editing. Thus, we performed single-cell RNA-seq to avoid the loss of random

[1]Institute of Neuroscience, State Key Laboratory of Neuroscience, Key Laboratory of Primate Neurobiology, CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. [2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. [3]CAS Key Laboratory of Systems Biology, CAS Center for Excellence in Molecular Cell Science, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China. [4]Bio-Med Big Data Center, Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China. [5]Center for Translational Medicine, Ministry of Education Key Laboratory of Birth Defects and Related Diseases of Women and Children, Department of Obstetrics and Gynecology, West China Second University Hospital, College of Life Sciences, Sichuan University, Chengdu, China. [6]School of Life Science and Technology, Shanghai Tech University, Shanghai, China. [7]Center for Animal Genomics, Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. [8]Shanghai Jiao Tong University, Fudan University, Shanghai Academy of Science & Technology, Shanghai, China. [9]These authors contributed equally: Changyang Zhou, Yidi Sun, Rui Yan, Yajing Liu, Erwei Zuo. *e-mail: yxli@sibs.ac.cn; hbzhou@ion.ac.cn; guofan@scu.edu.cn; huiyang@ion.ac.cn

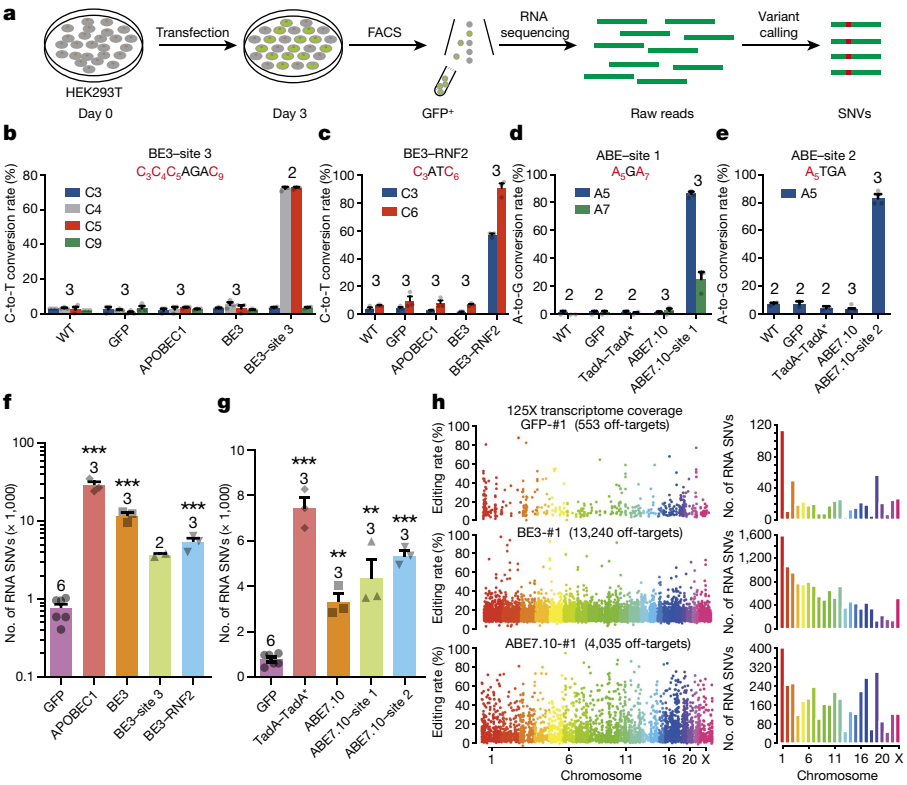


Fig. 1 | Base editors induce numerous off-target RNA SNVs. a, Scheme of the experimental design. **b**, **c**, DNA on-target efficiency of BE3–site 3 and BE3–RNF2. Note that APOBEC1 is the cytosine deaminase of BE3. **d**, **e**, DNA on-target efficiency of ABE7.10–site 1 and ABE7.10–site 2. Note that TadA–TadA* (wild-type TadA–evolved TadA heterodimer) is the adenine deaminase of ABE7.10, and evolved TadA is indicated by TadA*. **f**, **g**, Comparison of the off-target RNA SNVs for BE3 and ABE7.10 groups. **h**, Representative distributions of off-target RNA SNVs on human chromosomes for GFP, BE3 and ABE7.10. Chromosomes are indicated with different colours. Right, number of RNA SNVs for each chromosome. GFP group serves as control for all comparisons. WT, wild-type; GFP, GFP only; APOBEC1, APOBEC1 only; BE3, BE3 only; BE3–site 3, BE3 with sgRNA targeting site 3; BE3–RNF2, BE3 with sgRNA targeting RNF2; TadA–TadA*, TadA–TadA* only; ABE7.10, ABE7.10 only; ABE7.10–site 1, ABE7.10 with sgRNA targeting site 1; ABE7.10–site 2, ABE7.10 with sgRNA targeting site 2. All values are presented as mean ± s.e.m. Number above the bar indicates the number of biologically independent samples. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, two-sided unpaired *t*-test. Exact *P* values are provided in Supplementary Table 3.

off-target signals due to population averaging, on four groups of cells (wild-type, GFP-alone, BE3–site 3 and ABE7.10–site 1) (Fig. 3a). Consistently, we observed severe RNA off-target effects and similar mutation patterns in cells with high expression of deaminases, but not in cells with low deaminase expression (Fig. 3b–d, Extended Data Figs. 6, 7). Therefore, only cells with high expression of the indicated deaminase were used for further analysis (Extended Data Fig. 6). Notably, the percentage of off-target sites (4.5 ± 1.0%, mean ± s.e.m.) shared by any of the BE3- or ABE7.10-edited cells was much lower than that of the cell populations (40.8 ± 3.7%, mean ± s.e.m.), which indicates that BE3- or ABE7.10-induced off-target SNVs were essentially random and independent of sgRNA (Extended Data Fig. 8). Notably,

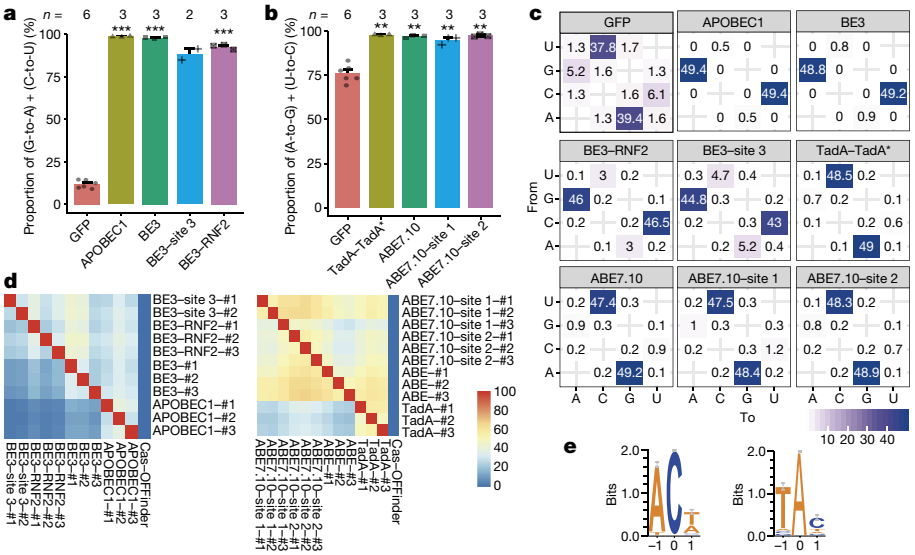


Fig. 2 | Characterization of off-target RNA SNVs. a, Proportion of G-to-A and C-to-U mutations for GFP, APOBEC1, BE3, BE3–site 3 and BE3–RNF2. **b**, Proportion of A-to-G and U-to-C mutations for GFP, TadA–TadA*, ABE7.10, ABE7.10–site 1 and ABE7.10–site 2. **c**, Distribution of mutation types in each group. The number indicates the percentage of a certain type of mutation among all mutations. **d**, The ratio of shared RNA SNVs between any two samples in the APOBEC1, BE3, TadA and ABE7.10 groups or with predicted off-target sites by Cas-OFFinder. The proportion in each cell is calculated by the number of overlapping RNA SNVs between two samples divided by the number of RNA SNVs in the row. **e**, Sequence logos derived from off-target RNA SNVs of BE3 (left) and ABE7.10 (right). Analysis was performed on generated RNA-seq data using cDNA, and thus every T depicted should be considered a U in RNA. Bits account for how much each column is conserved and how much the nucleotide frequencies obtained in the profile differ from those that would have been obtained by aligning oligonucleotides chosen at random. The GFP group serves as the control for all comparisons. All values are presented as mean ± s.e.m. *n* = biologically independent samples. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, two-sided unpaired *t*-test. Exact *P* values are provided in Supplementary Table 13.

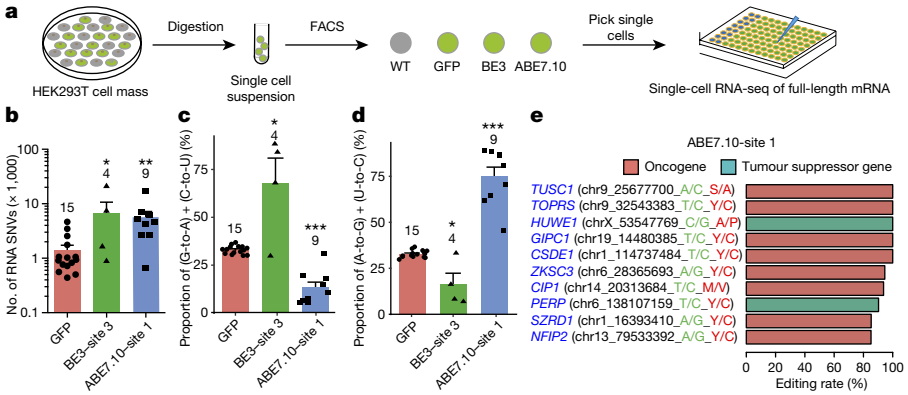


Fig. 3 | Single-cell RNA SNV analysis of cells transfected with base editors. a, Diagram of SNVs analysed by single-cell RNA-seq method. **b**, Number of off-target RNA SNVs detected in single cells transfected with GFP, BE3–site 3 or ABE7.10–site 1. **c**, **d**, Proportion of G-to-A and C-to-U mutations or A-to-G and U-to-C mutations for GFP, BE3–site 3 and ABE7.10–site 1 groups. **e**, Non-synonymous mutations located on the oncogenes and tumour suppressors with the highest editing rates from ABE7.10-treated single cells. Gene names, amino acid mutations and single nucleotide conversions are indicated in blue, red and green, respectively. GFP group serves as control for all comparisons. All values are presented as mean ± s.e.m. Number above the bar indicates the number of cells. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, two-sided unpaired *t*-test. Exact *P* values are provided in Supplementary Table 14.

some oncogene and tumour suppressor sites remained highly edited at specific sites, as in the bulk RNA-seq datasets, which implies that the editing might be directed to specific sequence motifs (Fig. 3e, Extended Data Fig. 8, Supplementary Tables 6, 7).

To further explore experimental approaches that may eliminate the RNA off-target activity of base editors, we examined the potential effect of de-stabilizing the RNA binding capacity of APOBEC1 and TadA (Extended Data Fig. 9). Specifically, we introduced a point mutation W90A to the predicted hydrophobic region in APOBEC1[20,21], and found that although BE3^{W90A} eliminated the RNA off-target effect, the on-target DNA editing activity of BE3^{W90A} was essentially absent (Fig. 4a, b, Extended Data Fig. 9, Supplementary Tables 8, 9). A previous study has shown that double mutations to BE3 (W90Y and R126E) can increase the editing specificity by reducing the hydrophobicity and binding affinity for DNA[22], which implies that BE3$^{W90Y/R126E}$ might also show reduced RNA-binding activity. The RNA off-target effect of BE3$^{W90Y/R126E}$ was reduced to a base level, but it maintained BE3-like DNA on-target efficiency. In an alternative approach, we tested whether replacing APOBEC1 with human APOBEC3A (hA3A)—which is reported to have DNA but not RNA binding activity[5,23]—could eliminate the RNA off-target activity of BE3 (Extended Data Fig. 9). Indeed, BE3(hA3A)-transfected HEK293T cells showed significantly

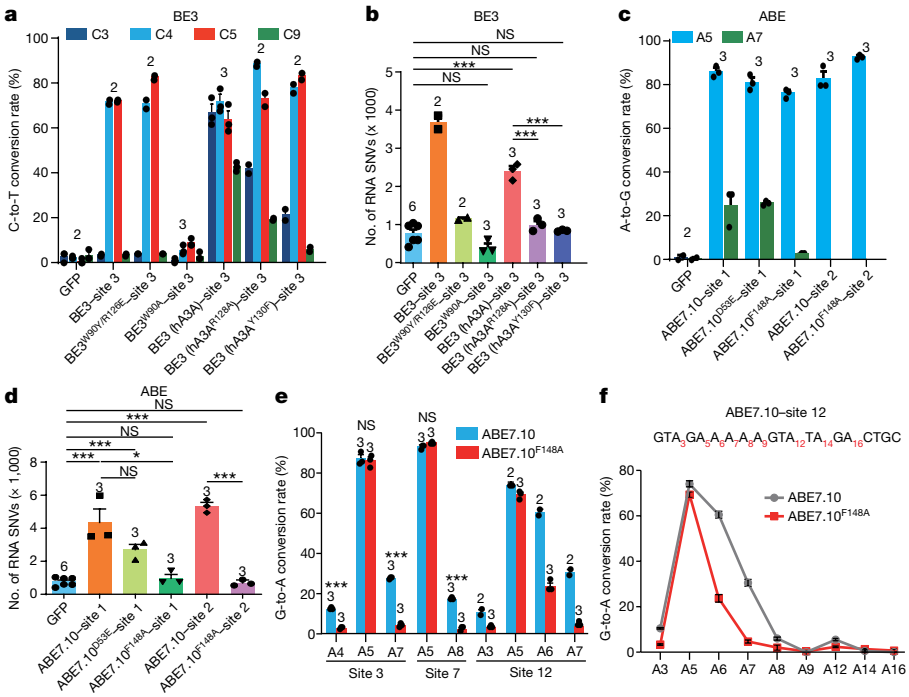


Fig. 4 | Elimination of off-target RNA SNVs by engineering of deaminases. a, Frequency of C-to-T conversion for GFP, BE3–site 3, BE3$^{W90F/R126E}$–site 3, BE3^{W90A}–site 3, BE3(hA3A)–site 3, BE3(hA3A^{R128A})–site 3, and BE3(hA3A^{Y130F})–site 3 groups. **b**, Comparison of the off-target RNA SNVs among BE3-treated groups. **c**, Frequency of A-to-G conversion for GFP, ABE7.10–site 1, ABE7.10^{D53E}–site 1, ABE7.10^{F148A}–site 1, ABE7.10–site 2 and ABE7.10^{F148A}–site 2 groups. Note that site 2 does not have A7. **d**, Comparison of off-target RNA SNVs among ABE7.10-treated groups. **e**, Comparison of editing efficiency between ABE7.10 and ABE7.10^{F148A} on four different sites. **f**, A representative editing site shows that ABE7.10^{F148A} narrows the width of the editing window. $n = 2$ and 3 biologically independent samples for ABE7.10 and ABE7.10^{F148A}, respectively. All values are presented as mean ± s.e.m. Number above the bar indicates the number of biologically independent samples. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, two-sided unpaired *t*-test. Exact *P* values are provided in Supplementary Table 9.

fewer off-target RNA SNVs than BE3(APOBEC1)-transfected cells (Fig. 4a, b, Extended Data Fig. 9, Supplementary Tables 9, 10). To further reduce off-target effects, we introduced point mutations R128A[24] and Y130F[23,25] into the predicted RNA and single-stranded DNA binding domains of hA3A, respectively, and found that the number of off-target RNA SNVs in both variants was decreased to the base level (Fig. 4a, b, Extended Data Fig. 9). Notably, the mutation patterns for three high-fidelity variants—BE3$^{W90Y/R126E}$, BE3(hA3A^{R128A}) and BE3(hA3A^{Y130F})—were similar to those found in cells transfected with GFP alone (Extended Data Fig. 9).

For ABE engineering, previous studies have shown that a D53E mutation can reduce the RNA activity of TadA in vitro, and an F148A mutation completely abolished the activity in *Escherichia coli*[9,11,26,27]. We therefore introduced a D53E or F148A mutation into both TadA and TadA* (Extended Data Fig. 9). Notably, both ABE7.10^{D53E} and ABE7.10^{F148A} maintained high DNA on-target efficiency, and only ABE7.10^{F148A} showed a complete absence of RNA off-target effects (Fig. 4c, d, Extended Data Fig. 9, Supplementary Tables 8, 9). Moreover, the remaining SNVs in ABE7.10^{F148A}-transfected cells in both sites were similar to those found in cells transfected with GFP alone (Extended Data Fig. 9). We further confirmed that the DNA on-target activity of ABE7.10^{F148A} was similar to that of ABE7.10 on three additional sites (Fig. 4e). The editing window of ABE7.10^{F148A} was substantially narrowed (Fig. 4f, Extended Data Fig. 10), which indicates increased precision of DNA base editing. To determine whether the off-target RNA editing was due to the wild-type TadA monomer, we examined the editing activities by catalytic inactivation of only the wild-type monomer (via an F148A mutation), and found that this variant maintained DNA on-target activity but could not decrease the number of RNA SNVs (Extended Data Fig. 10).

We have shown that BE3 and ABE7.10 generated substantial off-target RNA SNVs, consistent with three recent studies[28–30]. Although RNA off-target mutations could exist for only a short period of time by transient expression of base editors via ribonucleoprotein or nucleofection, in vivo genetic correction of the most common inherited diseases depends greatly on the delivery system: adeno-associated viruses, which maintain long-term gene expression[14,16]. Thus, continuous induction of tens of thousands of off-target RNA SNVs for months or even years could be highly risky in gene therapies. Here, we introduced point mutations to the deaminases and obtained high-fidelity variants for both CBEs and ABEs. Notably, recent reports have shown that CBEs, but not ABEs, induce substantial DNA off-target effects[7,8]. Thus, ABE7.10^{F148A} could potentially be used for highly specific DNA base editing without off-target effects on DNA or RNA.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1314-0.

Received: 5 April 2019; Accepted: 30 May 2019;
Published online 10 June 2019.

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. Kim, D., Kim, D. E., Lee, G., Cho, S. I. & Kim, J. S. Genome-wide target specificity of CRISPR RNA-guided adenine base editors. *Nat. Biotechnol.* **37**, 430–435 (2019).
5. Gehrke, J. M. et al. An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities. *Nat. Biotechnol.* **36**, 977–982 (2018).
6. Kim, D. et al. Genome-wide target specificities of CRISPR RNA-guided programmable deaminases. *Nat. Biotechnol.* **35**, 475–480 (2017).
7. Zuo, E. et al. Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science* **364**, 289–292 (2019).
8. Jin, S. et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292–295 (2019).
9. Poulsen, L. K., Larsen, N. W., Molin, S. & Andersson, P. Analysis of an *Escherichia coli* mutant strain resistant to the cell-killing function encoded by the gef gene family. *Mol. Microbiol.* **6**, 895–905 (1992).
10. Sowden, M., Hamm, J. K. & Smith, H. C. Overexpression of APOBEC-1 results in mooring sequence-dependent promiscuous RNA editing. *J. Biol. Chem.* **271**, 3011–3017 (1996).
11. Wolf, J., Gerber, A. P. & Keller, W. tadA, an essential tRNA-specific adenosine deaminase from *Escherichia coli*. *EMBO J.* **21**, 3841–3851 (2002).
12. Conticello, S. G. The AID/APOBEC family of nucleic acid mutators. *Genome Biol.* **9**, 229 (2008).
13. Blanc, V. & Davidson, N. O. APOBEC-1-mediated RNA editing. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2**, 594–602 (2010).
14. Villiger, L. et al. Treatment of a metabolic liver disease by *in vivo* genome base editing in adult mice. *Nat. Med.* **24**, 1519–1525 (2018).
15. Maeder, M. L. et al. Development of a gene-editing approach to restore vision loss in Leber congenital amaurosis type 10. *Nat. Med.* **25**, 229–233 (2019).
16. Rossidis, A. C. et al. In utero CRISPR-mediated therapeutic editing of metabolic genes. *Nat. Med.* **24**, 1513–1518 (2018).
17. Green, P., Ewing, B., Miller, W., Thomas, P. J. & Green, E. D. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33**, 514–517 (2003).
18. Mitchell, A. & Graur, D. Inferring the pattern of spontaneous mutation from the pattern of substitution in unitary pseudogenes of *Mycobacterium leprae* and a comparison of mutation patterns among distantly related organisms. *J. Mol. Evol.* **61**, 795–803 (2005).
19. Duret, L. Mutation patterns in the human genome: more variable than expected. *Plos Biol.* **7**, 217–219 (2009).
20. Chen, K. M. et al. Structure of the DNA deaminase domain of the HIV-1 restriction factor APOBEC3G. *Nature* **452**, 116–119 (2008).
21. Holden, L. G. et al. Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications. *Nature* **456**, 121–124 (2008).
22. 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).
23. Wang, X. et al. Efficient base editing in methylated regions with a human APOBEC3A-Cas9 fusion. *Nat. Biotechnol.* **36**, 946–949 (2018).
24. Stauch, B. et al. Model structure of APOBEC3C reveals a binding pocket modulating ribonucleic acid interaction required for encapsidation. *Proc. Natl Acad. Sci. USA* **106**, 12079–12084 (2009).
25. 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).
26. Xiang, S., Short, S. A., Wolfenden, R. & Carter, C. W. Jr The structure of the cytidine deaminase-product complex provides evidence for efficient proton transfer and ground-state destabilization. *Biochemistry* **36**, 4768–4774 (1997).
27. Kim, J. et al. Structural and kinetic characterization of *Escherichia coli* TadA, the wobble-specific tRNA deaminase. *Biochemistry* **45**, 6407–6416 (2006).
28. Grünewald, J. et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **569**, 433–437 (2019).
29. 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).
30. Grünewald, J. et al. CRISPR adenine and cytosine base editors with reduced RNA off-target activities. Preprint at https://doi.org/10.1101/631721 (2019).

Acknowledgements We thank M. Poo for discussions and comments on this manuscript; D. Li for discussions; and FACS facility H. Wu and L. Quan in the Institute of Neuroscience (ION) and M. Zhang in the Institute Pasteur of Shanghai (IPS), L. Yuan in Big Data Platform, Shanghai Institutes for Biological Sciences (SIBS) and Chinese Academy of Sciences (CAS). This work was supported by R&D Program of China (2017YFC1001302, 2018YFC2000100, 2018YFA0107701, and 2018YFC1003401), CAS Strategic Priority Research Program (XDB32060000), National Natural Science Foundation of China (31871502, 31522037, 31822035, 31822035, 31771590), Shanghai Municipal Science and Technology Major Project (2018SHZDZX05), Shanghai City Committee of Science and Technology Project (18411953700, 18JC1410100), National Science and Technology Major Project (2015ZX10004801-005) and National Key Research and Development Program of China (2017YFA0505500, 2016YFC0901704).

Author contributions C.Z. and H.Y. conceived the project. C.Z., Y. Liu, E.Z., L.H., Y.W., X.H. and H.Z. designed experiments, constructed plasmids and collected cells. Y.S., R.Z. and Y. Li performed bulk RNA-seq analysis. R.Y., C.G. and F.G. performed single-cell RNA-seq and analysis. H.Y. designed experiments and supervised the whole project. H.Z., Y.S. and H.Y. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1314-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1314-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Y. Li, H.Z., F.G. & H.Y.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

sgRNA and vector information. All gRNA and vector sequences are provided in Supplementary Tables 10, 11.

Transient transfection and sequencing. Plasmids were constructed using NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs) according to the standard protocol. HEK293T cells (Cell Bank of SIBCB, CAS) were authenticated by the supplier and free of mycoplasma contamination. Mycoplasma contamination was determined by PCR of the supernatant of HEK293T cells. HEK293T cells were seeded in 10-cm dishes and cultured in Dulbecco's modified Eagle's medium (DMEM, Thermo Fisher Scientific) supplemented with 10% FBS (Thermo Fisher Scientific) and penicillin–streptomycin at 37 °C with 5% CO_2. Cells were transfected with 30 μg plasmids using Lipofectamine 3000 (Thermo Fisher Scientific). Three days after transfection, cells were digested with 0.05% trypsin (Thermo Fisher Scientific) and prepared for FACS. GFP-positive cells were sorted and kept in DMEM or Trizol (Ambion) for determination of DNA base editing or RNA-seq. To determine the efficiency of DNA base editing, cells were lysed using One Step Mouse Genotyping Kit (Vazyme) and subsequently prepared for Sanger sequencing and quantified using EditR 1.0.8 (https://moriaritylab.shinyapps.io/editr_v10/). All experiments in Figs. 1b–e, 4a–d were performed simultaneously. Thus, data for GFP and BE3–site 3 at the site 3 locus were used in Figs. 1b, 4a, and data for GFP and ABE7.10–site 1 at the site 1 locus were used in Figs. 1d, 4c. For RNA-seq, ~500,000 cells (top 5% GFP signal) were collected and RNA was extracted according to the standard protocol. For library construction, mRNAs were fragmented and converted to cDNA using random hexamers or oligo(dT) primers. The 5′ and 3′ ends of cDNA were ligated with adaptors, and correctly ligated cDNA fragments were enriched and amplified by PCR. The concentration of the library was assessed using Bioanalyzer.

RNA-editing analysis by RNA-seq. We used fourteen groups of transfected cells: cells that expressed only GFP (36 h and 72 h), APOBEC1 or TadA–TadA*, cells that expressed BE3, BE3 with site 3 sgRNA (36 h and 72 h), BE3 with RNF2 sgRNA[2], BE3 (FNLS)[31] with site 3 sgRNA, ABE7.10, ABE7.10 with site 1 sgRNA (36 h and 72 h)[3], ABE7.10 with site 2 sgRNA and ABEmax[32] with site 1 sgRNA (Extended Data Figs. 1, 2).

High-throughput mRNA sequencing was carried out using Illumina Hiseq at mean coverages of 125×. FastQC (v.0.11.3) and Trimmomatic (v.0.36)[33] were used for quality control. Qualified reads were mapped to the reference genome (Ensemble GRCh38) using STAR (v.2.5.2b)[34] in two-pass mode with the parameters implemented by the ENCODE project. Picard tools (v.2.3.0) was then applied to sort and mark duplicates of the mapped BAM files. The refined BAM files were subject to split reads that spanned splice junctions, local realignment, base recalibration and variant calling with SplitNCigarReads, IndelRealigner, BaseRecalibrator and HaplotypeCaller tools from GATK (v.3.5)[35], respectively. To identify variants with high confidence, we filtered clusters of at least five SNVs that were within a window of 35 bases and retained variants with base-quality score >25, mapping quality score >20, Fisher strand values >30.0, qual by depth values <2.0 and sequencing depth >20. As the mRNAs were converted into cDNA before sequencing, both the nucleotide and its complementary base could be sequenced. For example, if there is a C in the mRNA, cDNA have both C and G at the specific site. When the reference genome was C, the sequence would be read as C, and if the reference was G at the site, G will be read oppositely. Therefore, we counted the sum of C to T + G to A mutations as the editing of BE3 and the sum of A to G + T to C for ABE7.10 editing.

Any confident variants found in wild-type HEK293T cells were considered to be SNPs and were filtered out from the GFP and base-editor-transfected groups for off-target analysis. The editing rate was calculated as the number of mutated reads divided by the sequencing depth for each site. To analyse the predicted variant effects of each off-target variant, we conducted variant annotation by Variant Effect Predicitor (VEP, v.94) with the GRCh38 database.

RSEM (v.1.2.21) was used to estimate the gene-expression levels on the alignment file with default parameters[36] and gene abundances were reported in TPM (transcripts per million kilobases). The off-target RNA SNVs identified in BE3- or ABE7.10-transfected cells were mapped to the gene level. We randomly selected the same number of genes from the transcriptome in each sample as that of the off-target SNVs, and then compared the expression levels between the two groups with $\log_2$-transformed TPM values.

The adjacent 3-bp sequences of the off-target RNA SNVs were extracted from the reference and subjected to motif prediction using WebLogo3 (http://weblogo.threeplusone.com/)[37].

All sequencing data have been deposited in the NCBI Sequence Read Archive (SRA) under project accession PRJNA528149.

Library construction for full-length RNA-seq from single cells. Individual human HEK293T cells were manually picked after FACS, lysed and subjected to cDNA synthesis using the Smart-seq2 protocol[38]. Single-cell cDNA was then amplified and fragmented as previously described[38,39]. The sequencing library was constructed (New England Biolabs), quality checked and sequenced with paired-end 150-bp reads on an Illumina HiSeq X-Ten platform (Novogene). We performed single-cell RNA-seq from 96 individual HEK293T cells, among which 16 were generated from single wild-type cells, 16 were generated from single GFP^+ cells, 32 were generated from single BE3–GFP^+ cells, and 32 were generated from single ABE–GFP^+ cells. After the quality check of all the libraries, 91 single-cell libraries passed our criteria and were subjected to deep sequencing. All the sequencing data were deposited in the SRA under PRJNA528561.

Processing of the single-cell RNA-seq data. Raw reads of single-cell RNA-seq data were first trimmed and aligned to the GRCh38 human transcriptome (STAR v2.5.2b)[34]. After de-duplication, RNA SNVs from individual cell were identified using GATK software (v3.5)[35]. Those SNVs detected in single cells with read depth ≥ 20.0, Fisher strand values ≤ 30.0 and qual by depth values ≥ 2.0 were retained for downstream analysis. Gene expression was quantified as $\log_2$(fragments per kilobase of transcript per million mapped reads (FPKM) + 1) using HTSeq (v0.10.0)[40]. On average, 10,932 RefSeq genes were detected in each single cell by about 6.07 million sequenced reads (Supplementary Table 12).

Statistical analysis. All values are shown as mean ± s.e.m. Unpaired Student's *t*-test (two-tailed) was used for comparisons and $P < 0.05$ was considered to be statistically significant. Details of statistical values are provided in Supplementary Tables. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

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

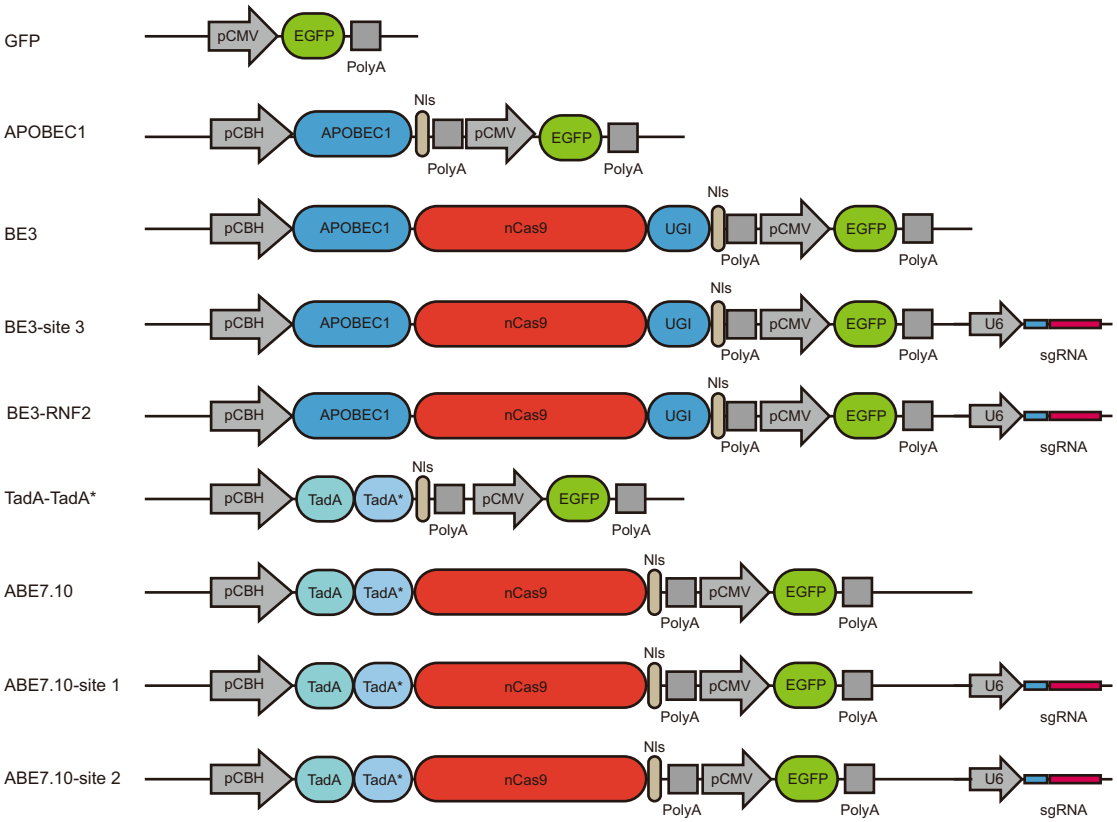
Data availability

All the sequencing data have been deposited in the NCBI SRA under project accession numbers PRJNA528149 and PRJNA528561 or at http://www.biosino.org/node/project/detail/OEP000277. All materials are available upon reasonable request.

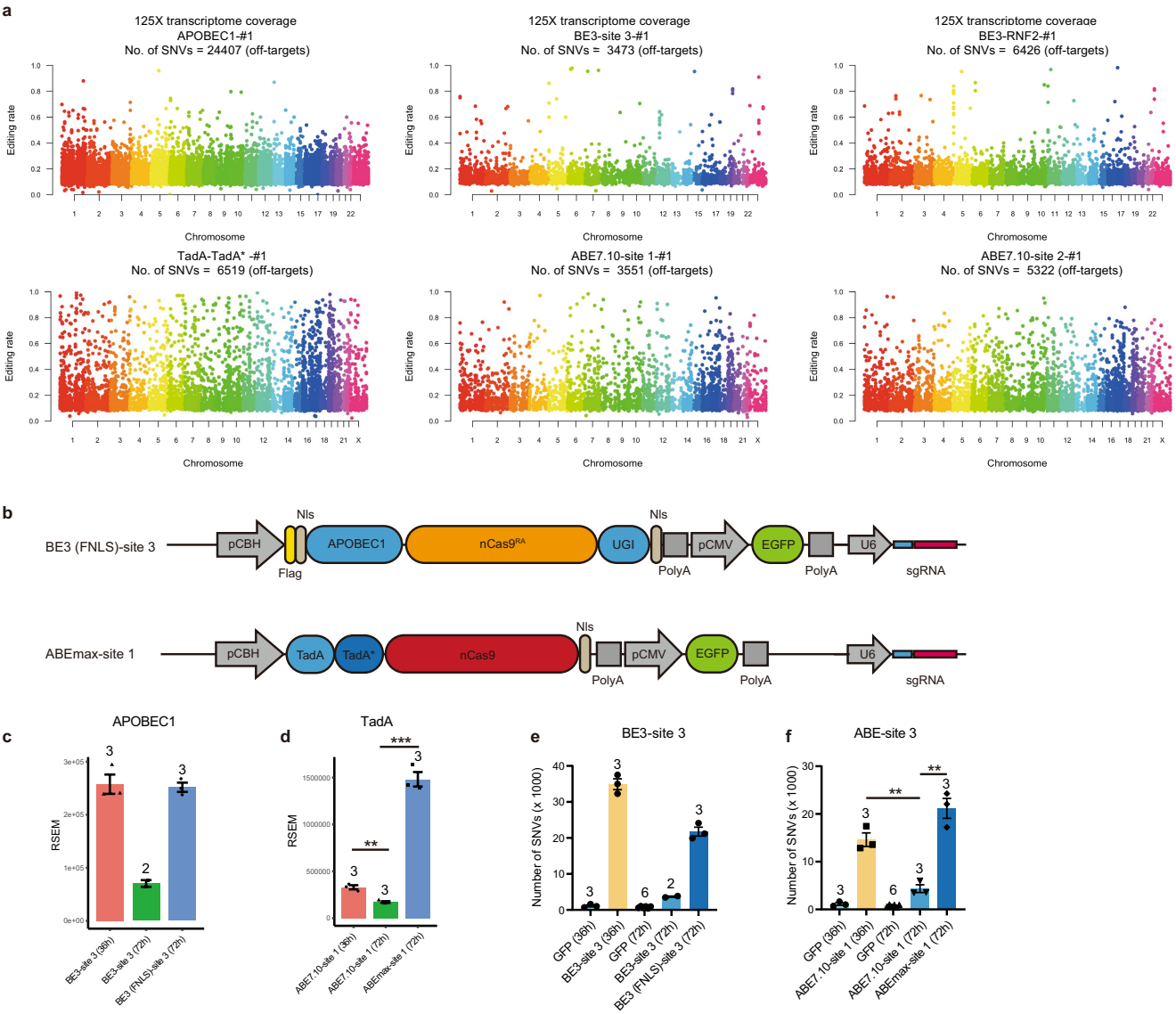
Code availability

The authors declare that all code used in this study are available within the article and its Extended Data or from the corresponding author upon reasonable request.

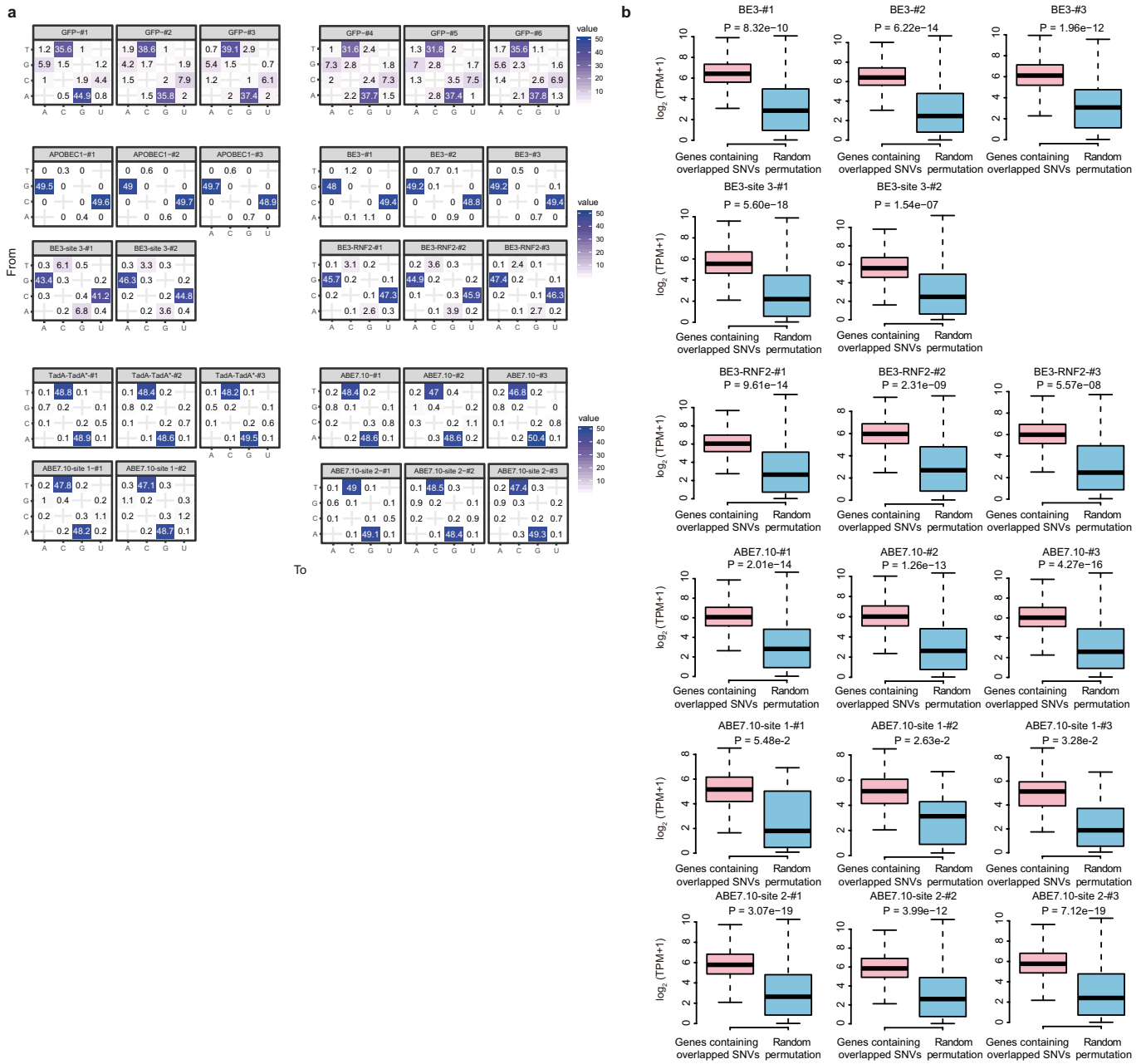
31. Koblan, L. W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).
32. Zafra, M. P. et al. Optimized base editors enable efficient editing in cells, organoids and mice. *Nat. Biotechnol.* **36**, 888–893 (2018).
33. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
34. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
35. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
36. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
37. Crooks, G. E. et al. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
38. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* **9**, 171–181 (2014).
39. Gu, C., Liu, S., Wu, Q., Zhang, L. & Guo, F. Integrative single-cell analysis of transcriptome, DNA methylome and chromatin accessibility in mouse oocytes. *Cell Res.* **29**, 110–123 (2019).
40. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).



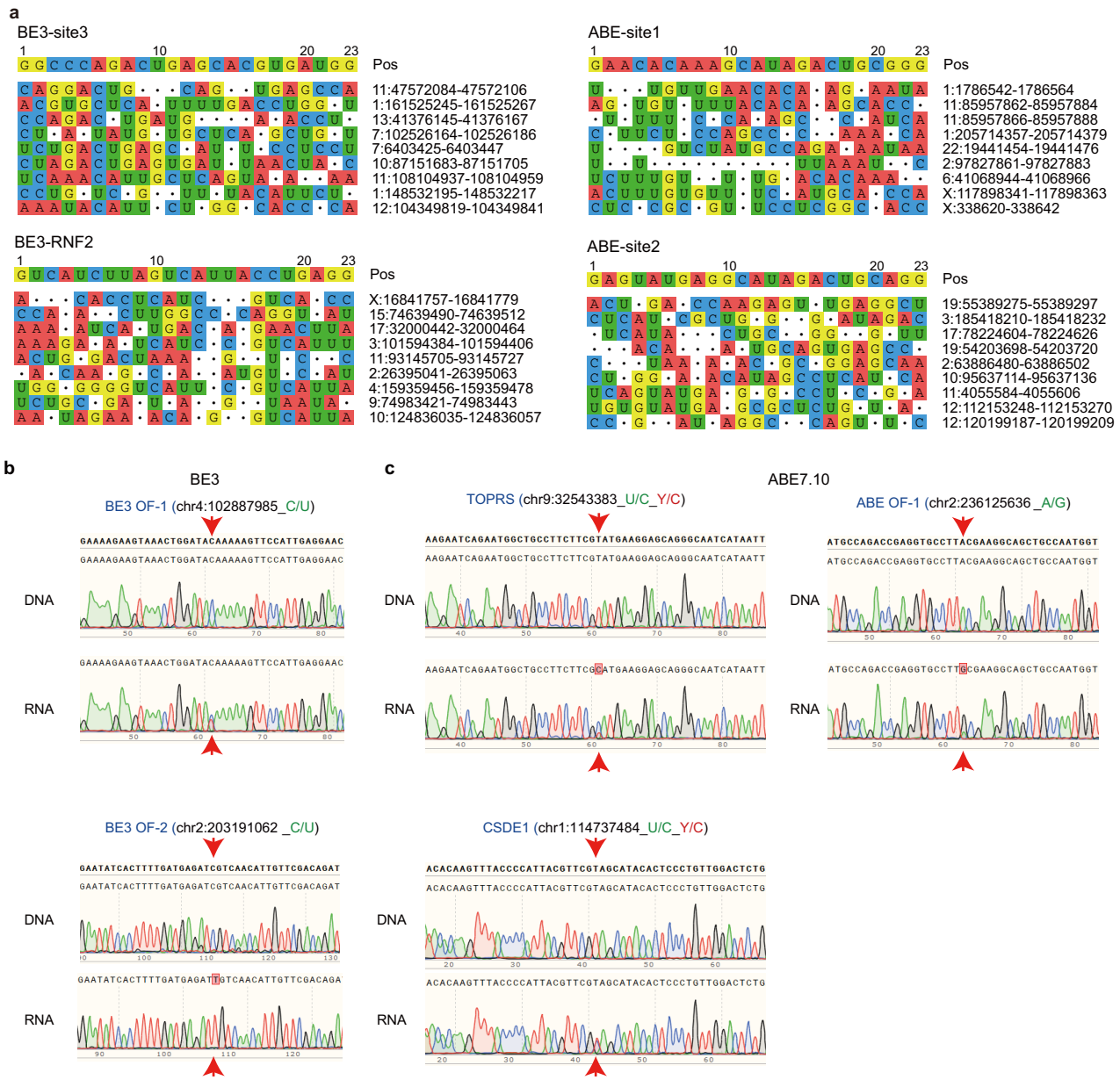
Extended Data Fig. 1 | Schematics of plasmids. The schematics show the plasmids used in this study.



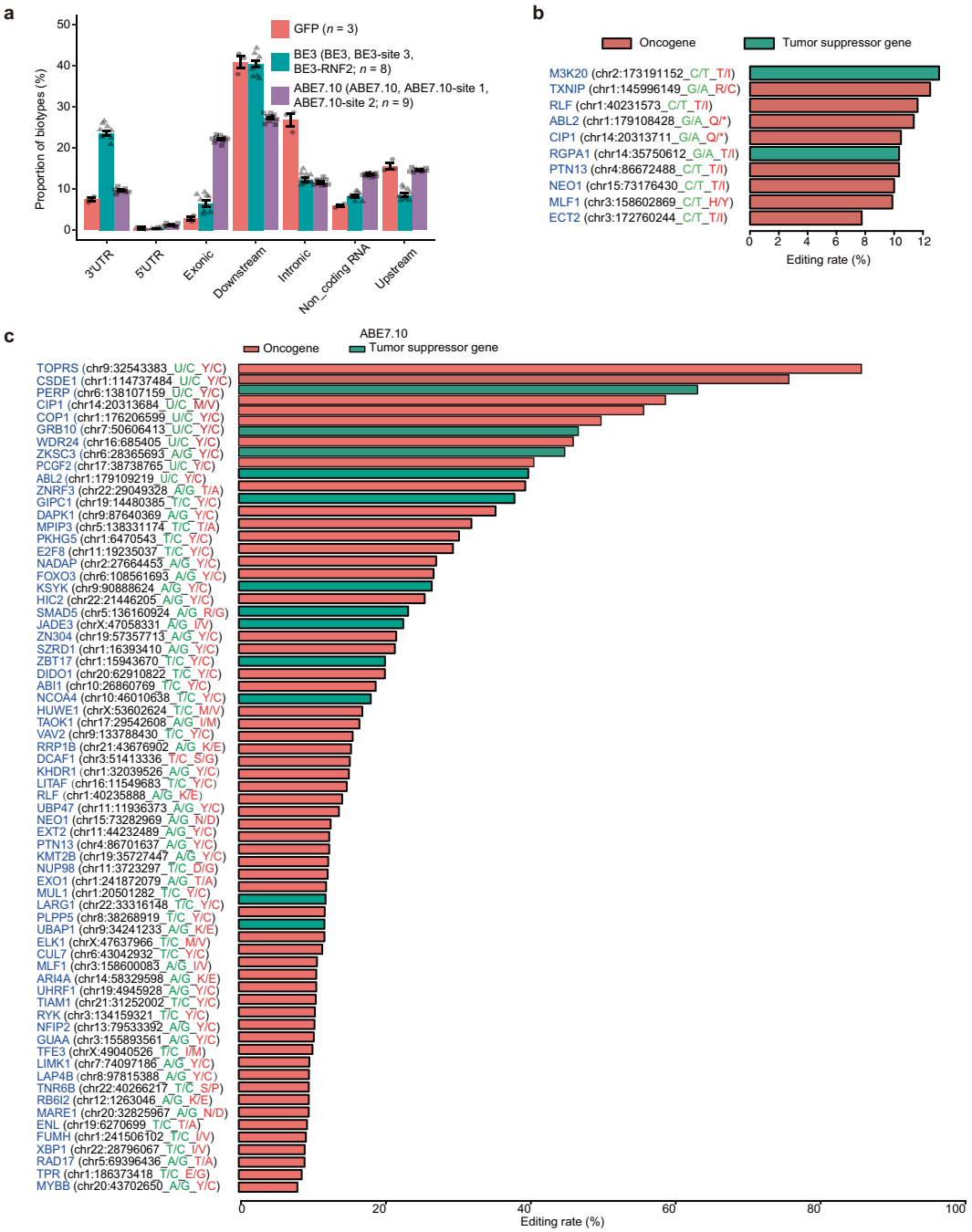
Extended Data Fig. 2 | Increased expression of deaminases induces an increase in off-target RNA SNVs. a, Representative distributions of off-target RNA SNVs on human chromosomes for APOBEC1, BE3–site 3, BE3–RNF2, TadA–TadA*, ABE7.10–site 1 and ABE7.10–site 2. **b**, Schematics of BE3(FNLS) and ABEmax. Note that BE3(FNLS)[31] and ABEmax[32] have previously been reported to greatly increase the expression of base editors. **c**, Expression of APOBEC1 in cells transfected with BE3–site 3 for 36 or 72 h, or with BE3(FNLS)–site 3 for 72 h. **d**, Expression level of in cells transfected with ABE7.10–site 1 for 36 and 72 h, or with ABEmax–site 1 for 72 h. **e**, The number of off-target RNA SNVs in cells transfected with BE3–site 3 for 36 or 72 h, or with BE3(FNLS)–site 3 for 72 h. **f**, The number of off-target RNA SNVs in cells transfected with ABE7.10–site 1 for 36 or 72 h, or with ABEmax–site 1 for 72 h. Transfections: GFP for 36 h; BE3–site 3 for 36 h; GFP for 72 h; BE3–site 3 for 72 h; BE3(FNLS)–site 3 for 72 h; ABE7.10–site 1 for 36 h; ABE7.10–site 1 for 72 h; ABEmax–site 1 for 72 h. RSEM, RNA-seq by expectation maximization. All values are presented as mean ± s.e.m. Number above the bar indicates the number of biologically independent samples. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, two-sided unpaired *t*-test. Exact *P* values are provided in Supplementary Table 15.



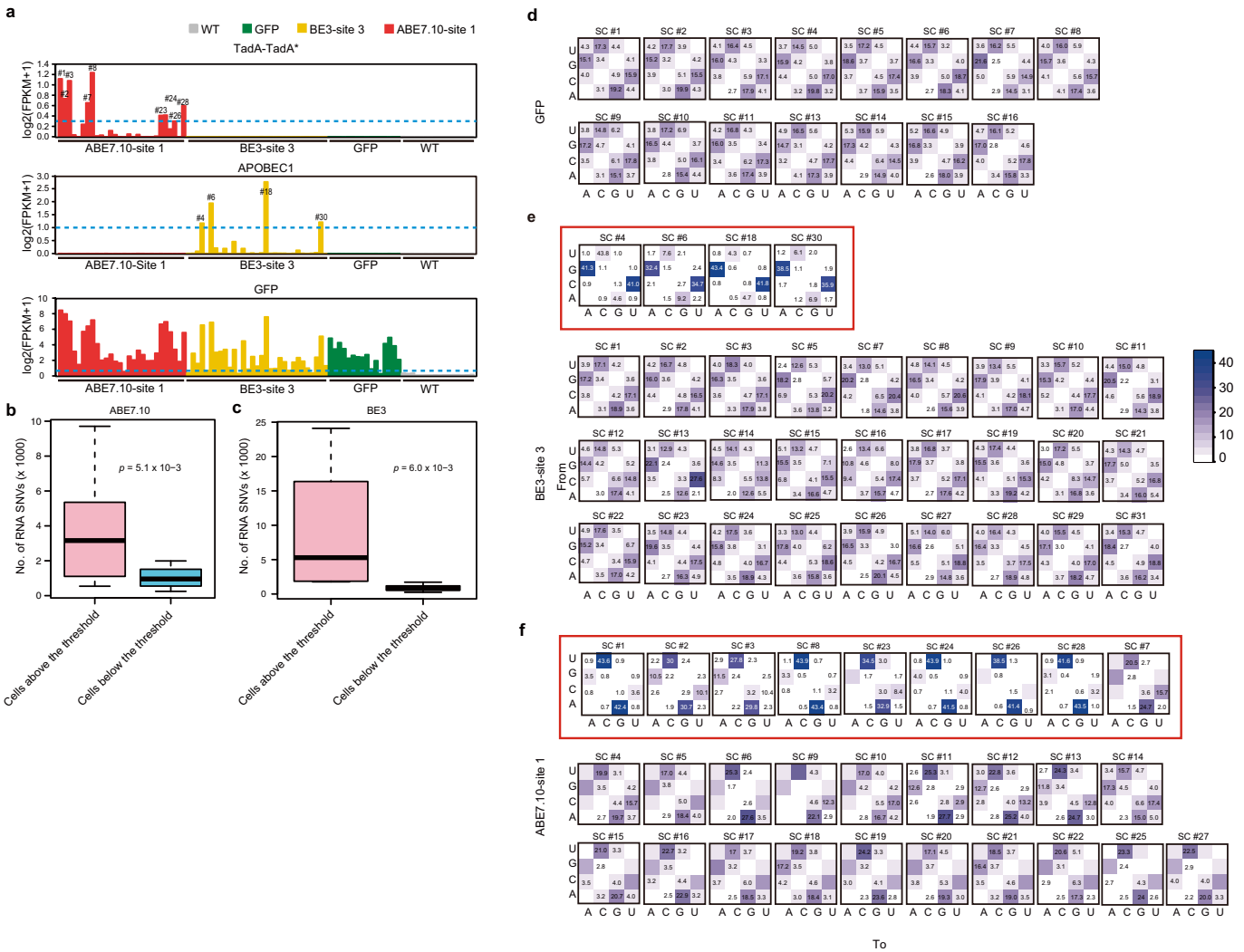
Extended Data Fig. 3 | Mutation types and gene expression of off-target RNA SNVs. a, Distribution of mutation types in each repeat of all groups. The number in each cell indicates the percentage of a certain type of mutation among all mutations. **b**, Expression of genes containing overlapping off-target RNA SNVs and random simulated genes in all groups transfected with BE3 ($n = 8$ biologically independent samples) or ABE7.10 ($n = 9$ biologically independent samples). Two-sided unpaired *t*-test. Box-and-whisker plots: centre line indicates median, box represents first and third quantiles, and whiskers indicate maximum and minimum values.



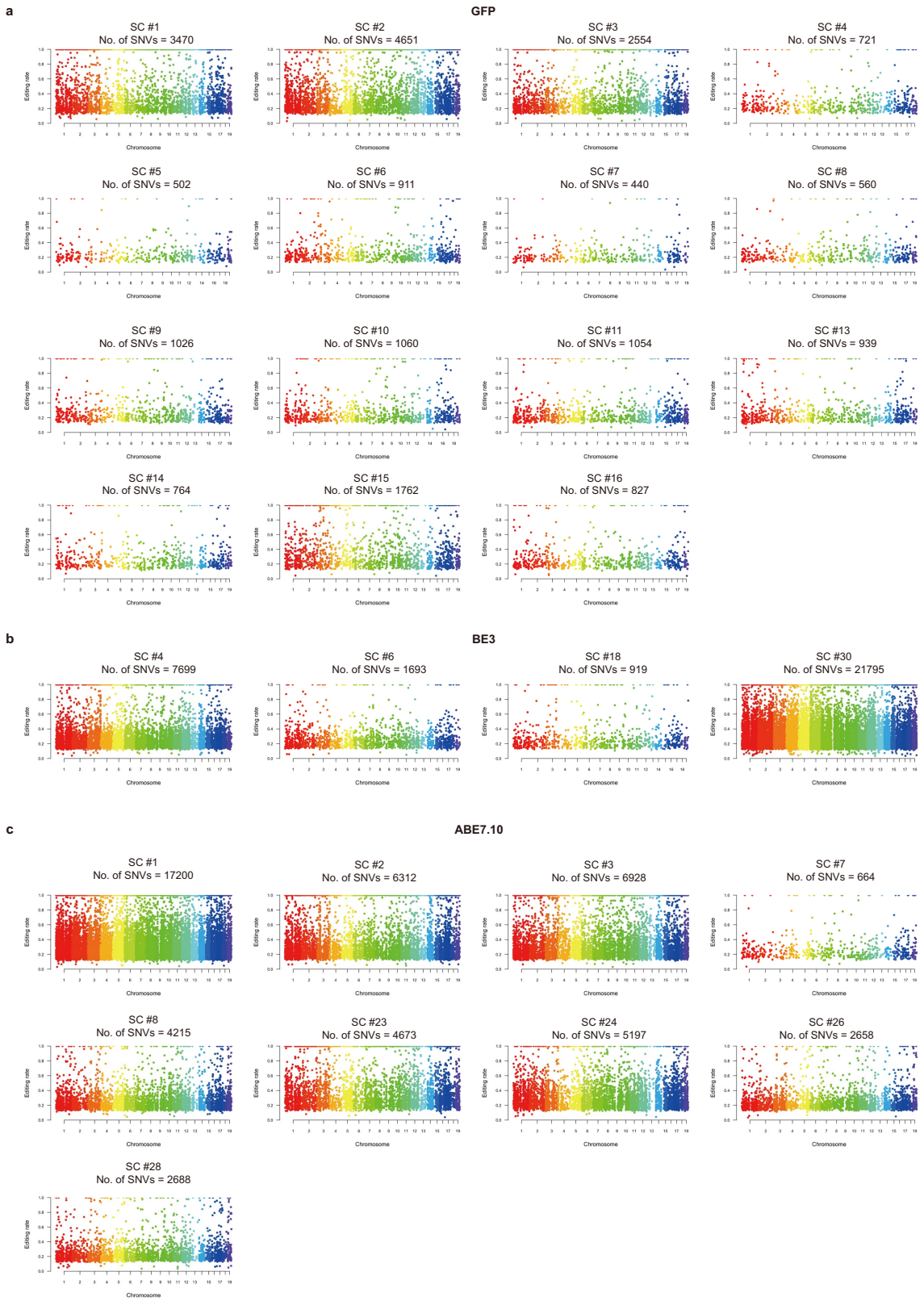
Extended Data Fig. 4 | Characteristics of off-target RNA SNVs. **a**, Similarity between adjacent sequences of off-target RNA SNVs and on-target sequences. $n = 2$ biologically independent samples for BE3–site 2, $n = 3$ for BE3–RNF2, $n = 3$ for ABE7.10–site 1, $n = 3$ for ABE7.10–site 2. **b**, Sanger sequencing chromatograms show that C to U mutation was observed only in RNA but not DNA for two BE3 off-target sites. OF, off-target. Gene names, amino acid mutations and single nucleotide conversions are indicated by blue, red and green, respectively. **c**, Sanger sequencing chromatograms show that U to C mutation was observed only in the RNA of three ABE7.10 off-target sites.



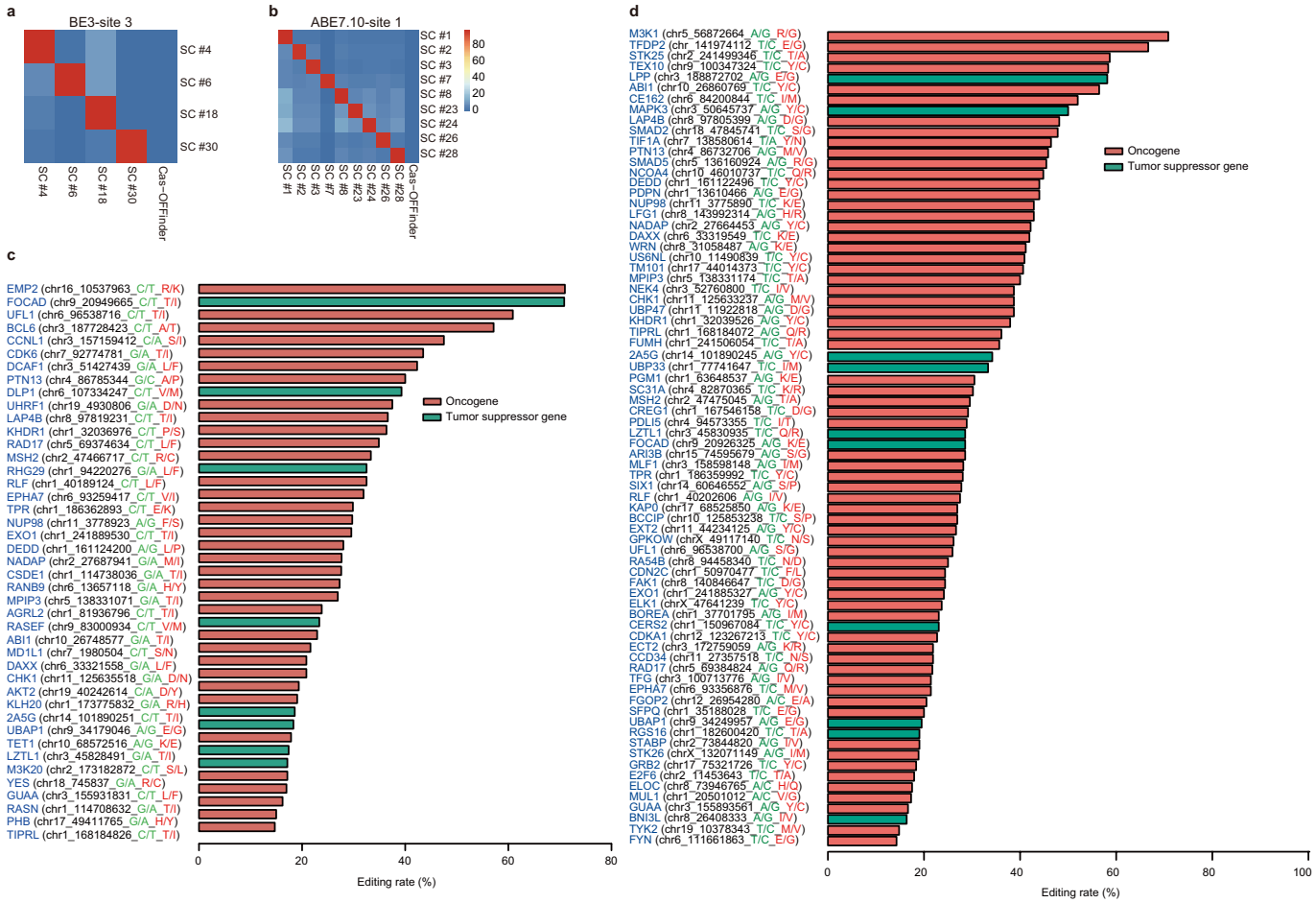
Extended Data Fig. 5 | Biotypes and tumour-associated genes of off-target RNA SNVs. a, Percentages of different locations of SNVs for GFP, BE3 (BE3, BE3–site 3 and BE3–RNF2) and ABE7.10 (ABE7.10, ABE7.10–site 1 and ABE7.10–site 2) groups. All values are presented as mean ± s.e.m. *n* denotes biologically independent samples. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, two-sided unpaired *t*-test. Exact *P* values are provided in Supplementary Table 16. **b**, Editing rate of BE3-induced non-synonymous mutations located on oncogenes and tumour suppressor genes. **c**, Editing rate of ABE7.10-induced non-synonymous mutations located on oncogenes and tumour suppressor genes. Gene names, amino acid mutations and single nucleotide conversions are indicated by blue, red and green, respectively.



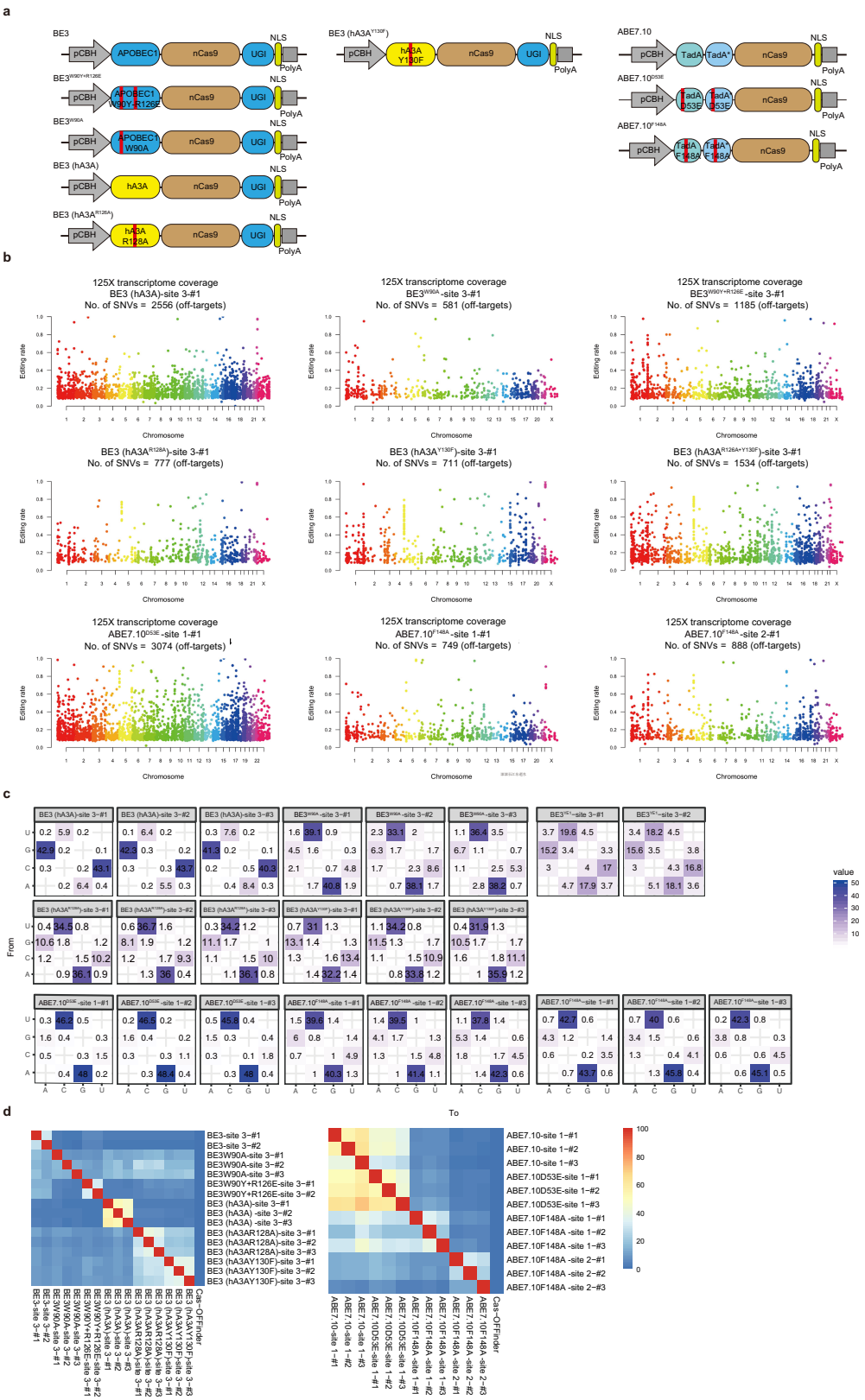
Extended Data Fig. 6 | Expression of transfected vectors and mutation types of off-target RNA SNVs in single cells. a, Expression of GFP, APOBEC1 and TadA–TadA* was quantified in all sequenced single cells. Thresholds are indicated by blue dashed lines. Thresholds of $\log_2$ (FPKM + 1) for GFP, BE3 and ABE7.10 are 0.3, 1 and 0.3, respectively. Cells with expression levels higher than the threshold were included for further analysis. **b**, **c**, Cells with high expression of TadA–TadA* or APOBEC1 showed greater numbers of RNA SNVs than those with low expressions in the ABE7.10 ($n = 9$ cells) or BE3 group ($n = 4$ cells), respectively. Box-and-whisker plots: centre line indicates median value, box represents first and third quantile, whisker indicates maximum and minimum values. **d**, Distribution of mutation types for GFP-transfected single cells ($n = 16$ cells). **e**, Distribution of mutation types for BE3–site 3-transfected single cells ($n = 31$ cells). Cells with expression of APOBEC1 higher than the threshold are included in the red square. **f**, Distribution of mutation types for ABE7.10–site 1-transfected single cells ($n = 28$ cells). Cells with expression of TadA–TadA* higher than the threshold are included in the red square. The number indicates the percentage of a certain type of mutation among all mutations. SC, single cell.



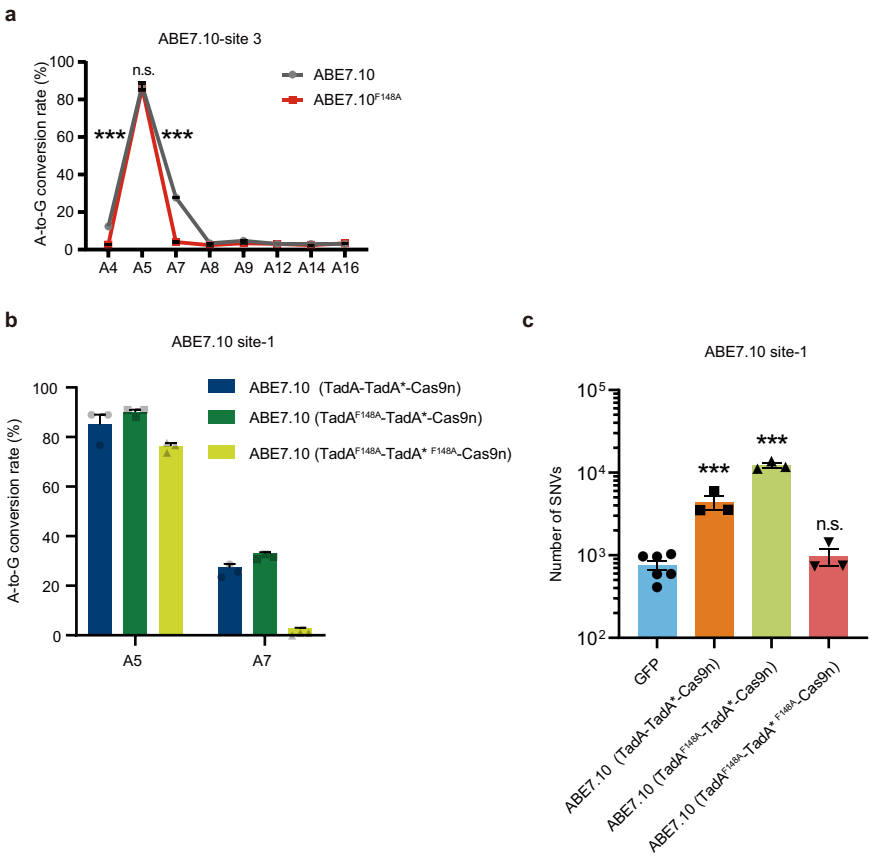
Extended Data Fig. 7 | Distribution of off-target RNA SNVs on human chromosomes for all single cells with expression above thresholds. **a**, Distribution of off-target RNA SNVs on human chromosomes for GFP-transfected single cells ($n = 15$ cells). **b**, Distribution of off-target RNA SNVs on human chromosomes for BE3–site 3-transfected single cells ($n = 4$ cells). **c**, Distribution of off-target RNA SNVs on human chromosomes for ABE7.10–site 1-transfected single cells ($n = 9$ cells).



Extended Data Fig. 8 | Characteristics of off-target RNA SNVs in single cells. a, **b**, The ratio of shared SNVs between any two samples in the same group or with predicted off-target sites by Cas-OFFinder. The proportion in each cell is calculated by the number of overlapping SNVs between two samples divided by the sample in the row. **c**, Editing rate of BE3-induced non-synonymous mutations on oncogenes and tumour suppressor genes in single cells. **d**, Editing rate of ABE7.10-induced non-synonymous mutations on oncogenes and tumour suppressor genes in single cells. Gene names, amino acid mutations and single nucleotide conversions are indicated by blue, red and green, respectively.



Extended Data Fig. 9 | Characteristics of off-target RNA SNVs for engineered BE3 and ABE7.10 variants. a, Schematic of BE3 and ABE7.10 variants. Point mutations are indicated by red lines. **b**, Representative distributions of off-target RNA SNVs on human chromosomes. **c**, Distribution of mutation types for each sample of the engineered variants of BE3 and ABE7.10. **d**, Ratio of shared RNA SNVs between any two samples in the engineered variants of BE3 and ABE7.10 or with off-target sites predicted by Cas-OFFinder. The proportion in each cell was calculated by the number of overlapping RNA SNVs between two samples divided by the number of RNA SNVs in the row.



Extended Data Fig. 10 | DNA on-target and RNA off-target activities of different engineered variants. a, Comparison of the width of editing windows between ABE7.10 and ABE7.10^{F148A}. $n = 3$ biologically independent samples for each group. **b**, DNA on-target efficiency on site 1 of TadA–TadA*–Cas9n (wild-type TadA–evolved TadA heterodimer), TadA^{F148A}–TadA*–Cas9n and TadA^{F148A}–TadA*^{F148A}–Cas9n. $n = 3$ biologically independent samples for each group. **c**, The number of RNA SNVs in the GFP, TadA–TadA*–Cas9n, TadA^{F148A}–TadA*–Cas9n and TadA^{F148A}–TadA*^{F148A}–Cas9n groups. $n = 3$ biologically independent samples for each group. All values are presented as mean ± s.e.m. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Two-sided unpaired *t*-test. Exact *P* values are provided in Supplementary Table 17.

nature research

Corresponding author(s): Hui Yang

Last updated by author(s): May 25, 2019

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	EditR 1.0.8 , FlowJo X, FastQC (v0.11.3), Trimmomatic (v0.36), Illumina HiSeq X-Ten
Data analysis	Microsoft Excel 2019, GraphPad Prism 8.0.2, STAR (v2.5.2b), Picard tools (v2.3.0), SplitNCigarReads, IndelRealigner, BaseRecalibrator and HaplotypeCaller tools from GATK (v3.5) ,Variant Effect Predicitor (VEP, v94) ,RSEM (v1.2.21)

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

All raw data are deposited in NCBI Sequence Read Archive (SRA) under project accession PRJNA528149 and PRJNA528561 or http://www.biosino.org/node/project/detail/OEP000277 and interpreted data is either available in the supplemental information or upon request.

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	Sample size was chosen based on the results of previous works in this field that used similar sample sizes to generate reproducible results.
Data exclusions	To avoid mistakes during sample preparation and sequencing, samples expressing the right transfected vectors were included in this study. One RNA-seq dataset was excluded due to expression of the wrong transfected vectors.
Replication	To ensure the robustness, we performed independent biological replicates for each experiments. Moreover, we replicated the results using different gRNAs.
Randomization	Randomization was not relevant to this study given that top 5% GFP positive cells were collected by FACS sorting strategy for all samples.
Blinding	Blinding was not relevant to our study. Based on previous studies in this field, these assays do not require blinding. Thus, blinding was not used.

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)	In this study, we only used one cell line HEK293T, which was from the Cell bank of Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences
Authentication	Cell lines were authenticated by the supplier.
Mycoplasma contamination	Cells used in this study were free of mycoplasma contamination. Mycoplasma contamination was determined by PCR the supernatant of HEK293T cells.
Commonly misidentified lines (See ICLAC register)	None of the cell lines used was listed in the database of ICLAC.

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	HEK293T cells were digested by trypsin (0.05%), contrifuged at 1000 rpm and filtered with a 35 μm nylon mesh.
Instrument	MoFlo XDP (Beckman)
Software	Summit Software version 5.2, FlowJo X.
Cell population abundance	Cell population abundance: Cell population abundance was influenced by the size of the plasmids. Normally, HEK293T cells transfected with plasmids were usually ~60% GFP+ and around 500000 cells were sorted for RNA-seq.
Gating strategy	Positive and negative boundaries were determined by control cells that were not transfected with any plasmids. Cells with top 5% of GFP signal were collected.

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