nature biotechnology

Efficient base editing in methylated regions with a human APOBEC3A-Cas9 fusion

Xiao Wang[1–3,6], Jianan Li[1–3,6], Ying Wang[4,6], Bei Yang[5,6], Jia Wei[4,6], Jing Wu[1], Ruixuan Wang[1], Xingxu Huang[1], Jia Chen[1] & Li Yang[4]

Base editors (BEs) enable the generation of targeted single-nucleotide mutations, but currently used rat APOBEC1-based BEs are relatively inefficient in editing cytosines in highly methylated regions or in GpC contexts. By screening a variety of APOBEC and AID deaminases, we show that human APOBEC3A-conjugated BEs and versions we engineered to have narrower editing windows can mediate efficient C-to-T base editing in regions with high methylation levels and GpC dinucleotide content.

BEs, which combine a cytidine deaminase with Cas9 or Cpf1, have been successfully applied to perform targeted base editing, including C-to-T edits[1–6]. Numerous human diseases have been reported to be driven by point mutations in genomic DNAs. With recently developed BEs, these disease-related point mutations can be potentially corrected, providing new therapeutic options. By analyzing disease-related T-to-C mutations that can be theoretically reverted to thymines by BEs, we found that ~43% of them are on cytosines in the context of CpG dinucleotides (**Fig. 1a**). It is well known that the cytosine of CpG is usually methylated in mammalian cells[7], and methylation of cytosine strongly suppresses cytidine deamination catalyzed by some APOBEC and AID deaminases[8]. Here we show that CpG dinucleotide methylation hinders the C-to-T base editing by current BEs and develop BEs for efficient C-to-T base editing in highly methylated regions.

We first examined the base editing efficiency of a commonly used BE, the rat APOBEC1 (rA1)-based BE3[1], in human cells having either increased or decreased levels of methylation. When DNA methylation was promoted by DNMT3 in regions with native low methylation levels, editing frequencies by BE3 decreased (**Supplementary Fig. 1a**). In addition, when DNA methylation was reduced by TET1 in regions with native high methylation levels, BE3-induced editing frequencies increased accordingly (**Supplementary Fig. 1b**). These results suggest that the canonical rA1-based BE3 is less efficient in editing cytosines embedded in highly methylated genomic regions. Notably, C-to-T editing was suppressed by DNA methylation at both CpG and flanking non-CpG sites (**Supplementary Fig. 1c**,**d**, median decrement ~28%, $P = 2 \times 10^{-8}$ for CpG sites and ~51%, $P = 7 \times 10^{-10}$ for flanking non-CpG sites). APOBECs deaminate cytidines on single-stranded DNA in a processive manner[9]. CpG methylation may affect the sliding of APOBEC and therefore impair its binding on the flanking non-CpG sites for deamination (**Supplementary Fig. 1d**).

To screen for efficient base editing in high-methylation background, we obtained a series of BEs by fusing Cas9 nickase with fifteen different APOBEC and AID deaminases (**Fig. 1b** and **Supplementary Fig. 2a–c**). We then tested these BEs in an *Escherichia coli*–derived vector system (**Fig. 1b**) that has been previously used to probe mutations[6,10]. In unmethylated vectors, these BEs showed varied levels of base editing. The BEs containing human APOBEC3A (hA3A-BE3, mean editing frequency ~39%), human APOBEC3B (hA3B-BE3, mean editing frequency ~33%) or human AID (hAID-BE3, mean editing frequency ~28%) mediated base editing at levels that are comparable to that of BE3 (mean editing frequency ~31%) (**Fig. 1c**). In methylated vectors only hA3A-BE3 induced efficient base editing (mean editing frequency ~35%), with relatively low editing efficiencies induced by BE3 (mean editing frequency ~12%) or other examined BEs (mean editing frequencies ~1%–20%) (**Fig. 1c**). Protein products of hA3A-BE3, BE3 and other examined BEs were comparable (**Fig. 1d** and **Supplementary Fig. 2d**).

Similarly to the observation in *E. coli*–derived vectors, hA3A-BE3 exhibited significantly higher base editing frequencies than rA1-based BE3 in all tested genomic regions, either those with a native high methylation background (median ~1.7-fold, $P = 2 \times 10^{-10}$, **Fig. 1e**,**f**) or those with an induced high-methylation condition (median ~1.8-fold, $P = 5 \times 10^{-4}$, **Supplementary Figs. 1a** and **3a**,**b**). Thus, using hA3A as the deaminase module in BE could generally achieve high base editing efficiency in genomic regions with high methylation levels. This finding is also in agreement with the fact that hA3A can catalyze efficient deamination of methylated cytidines[11,12].

The base editing on cytosines in a GpC context was observed to be generally inefficient by rA1-based BEs[1,6,13]. By contrast, we found that hA3A-BE3 could induce efficient base editing on most of the cytosines at GpC sites in both endogenously and induced high-methylation backgrounds (**Fig. 1e** and **Supplementary Fig. 3a**, shaded gray). We further compared their editing efficiencies under both endogenously and induced low-methylation backgrounds and observed a similar superiority of hA3A-BE3 over BE3 in editing cytosines in the GpC context (**Fig. 1g**,**h** and **Supplementary Fig. 4a**,**b**). Statistical analysis confirmed that the base editing efficiency induced by hA3A-BE3 was significantly higher than that induced by BE3 on cytosines in the GpC context in either high-methylation (median ~2.3-fold, $P = 1 \times 10^{-5}$, **Supplementary Fig. 3c**) or low-methylation (median ~1.8-fold, $P = 6 \times 10^{-9}$, **Supplementary Fig. 4c**) conditions. Notably,

[1]School of Life Science and Technology, ShanghaiTech University, Shanghai, China. [2]Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China. [3]University of Chinese Academy of Sciences, Beijing, China. [4]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]Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech University, Shanghai, China. [6]These authors contributed equally to this work. Correspondence should be addressed to J.C. (chenjia@shanghaitech.edu.cn), X.H. (huangxx@shanghaitech.edu.cn) or L.Y. (liyang@picb.ac.cn).

Received 30 March; accepted 3 July; published online 20 August 2018; doi:10.1038/nbt.4198

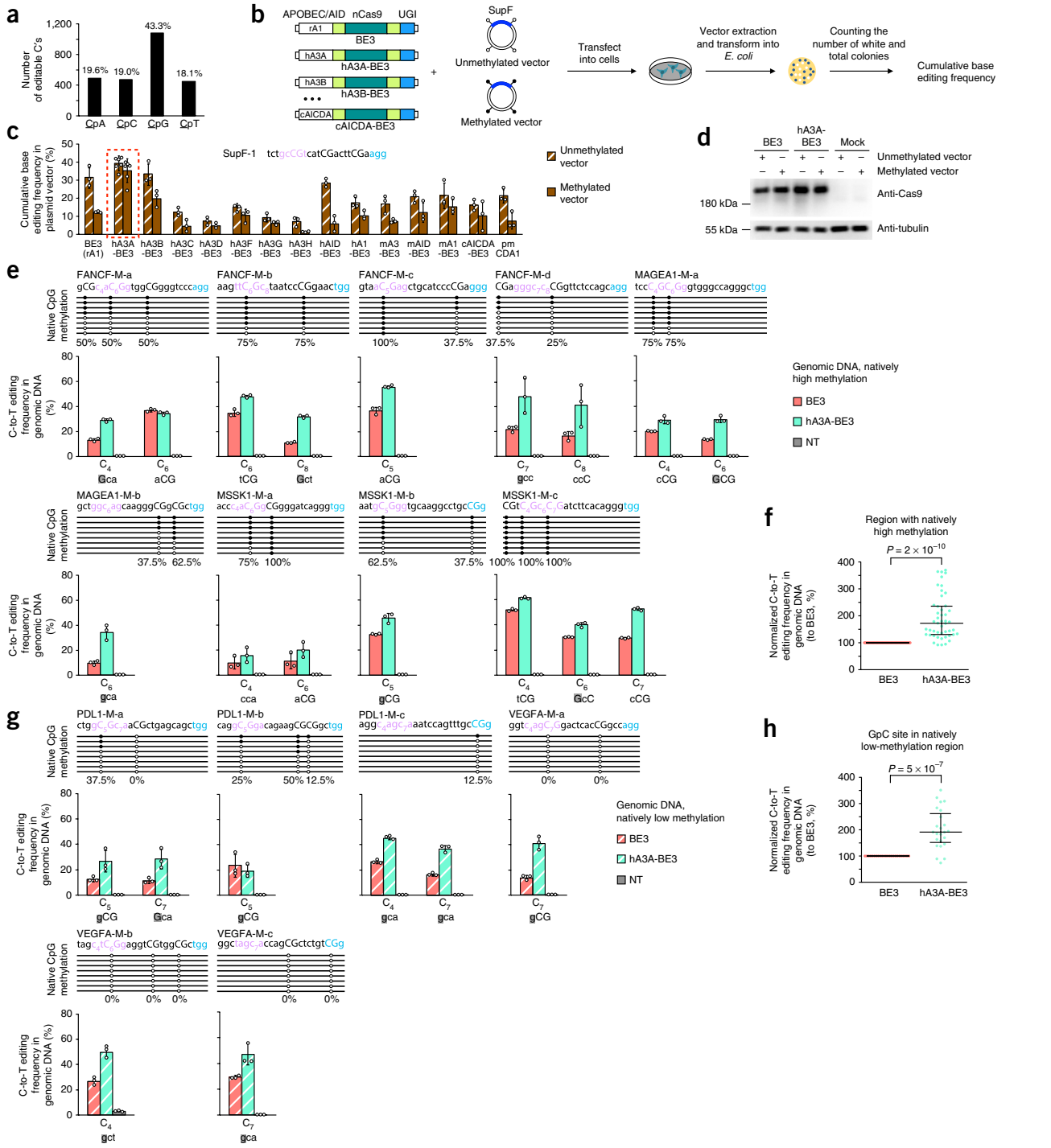


Figure 1 hA3A-BE3 induces efficient base editing in a methylated region and in a GpC context. (**a**) Distribution of BE-editable T-to-C (or A-to-G) variants. Potentially editable cytosines (underlined) are subclassified according to their 3′ adjacent bases. (**b**) Screening of BEs for efficient base editing in a high-methylation background. A series of new BEs were constructed by fusing different APOBEC and AID deaminases with Cas9 nickase (nCas9) and uracil DNA glycosylase inhibitor (UGI). (**c**) Cumulative base editing frequencies induced by different BEs in unmethylated and methylated vectors. A commonly used rA1-based BE3 was chosen for comparison. Means ± s.d. were from three (six for hA3A-BE3) independent experiments. (**d**) Immunoblots of BE3 and hA3A-BE3 cotransfected with unmethylated or methylated vectors. Tubulin was used as a loading control and immunoblot images are representative of three independent experiments. Uncropped blot images are shown in **Supplementary Figure 11**. (**e**) Comparison of base editing efficiencies induced by BE3 and hA3A-BE3 in genomic regions with natively high levels of DNA methylation. C-to-T editing frequencies of indicated cytosines were determined individually. Target site sequences are shown with the BE3 editing window (positions 4–8, setting the base distal to the protospacer-adjacent motif (PAM) as position 1) in lavender, PAM in cyan and CpG site in capitals. Guanines at the 5′ end of editable cytosines are shaded gray. NT, native HEK293T cells with no treatment. M, methylable regions. (**f**) Statistical analysis of normalized C-to-T editing frequencies in regions with natively high levels of DNA methylation shown in **e**, setting the ones induced by BE3 to 100%. $n = 48$ samples from three independent experiments. (**g**) Comparison of base editing efficiencies induced by BE3 and hA3A-BE3 at the cytosine of GpC in genomic regions with natively low levels of DNA methylation. (**h**) Statistical analysis of normalized C-to-T editing frequencies at GpC sites in regions with natively low levels of DNA methylation shown in **g**, setting the ones induced by BE3 to 100%. $n = 24$ samples from three independent experiments. (**e**,**g**) Means ± s.d. were from three independent experiments. (**f**,**h**) *P* value, one-tailed Student's *t* test. The median and interquartile range are shown.

hA3A-BE3-mediated base editing was as efficient as that of BE3 at cytosines in non-GpC contexts in all tested low-methylation regions (median ~1.1-fold, $P = 0.045$, **Supplementary Fig. 4d**). We also found that hA3A-BE3 yielded a higher C-to-T fraction (product purity) than BE3 in both high-methylation (**Supplementary Fig. 5a**,**b**, median ~97% by hA3A-BE3 compared to ~94% by BE3, $P = 3 \times 10^{-4}$) and low-methylation regions (**Supplementary Fig. 5c**,**d**, median ~92% by hA3A-BE3 compared to ~90% by BE3, $P = 4 \times 10^{-6}$). Both BE3 and hA3A-BE3 yielded a higher C-to-T fraction at CpG sites with high methylation status than at CpG sites with low methylation status (**Supplementary Fig. 5g**, median ~95% vs. ~90%, $P = 3 \times 10^{-5}$ for BE3 and median ~95% vs. ~92%, $P = 5 \times 10^{-4}$ for hA3A-BE3). We also found that hA3A-BE3 induced higher indel frequencies than BE3 (**Supplementary Fig. 5e**,**f**,**h**, median ~2 in both high- and low- methylation regions). Such an increase may be caused by the high deaminase activity of hA3A[12,14], which can trigger downstream DNA repair pathways to generate DNA double-strand breaks[15,16].

Our results suggest that hA3A-BE3 can efficiently induce base editing in a broader scope (**Fig. 1**). However, the editing window of hA3A-BE3 is wider (~12 nt, positions 2–13 in the sgRNA target site, **Supplementary Fig. 6a**) than that of BE3 (~5 nt, positions 4–8)[1]. As the wide editing window of hA3A-BE3 may result from the high deaminase activity of hA3A, mutations in hA3A that can reduce deaminase activity might correspondingly narrow

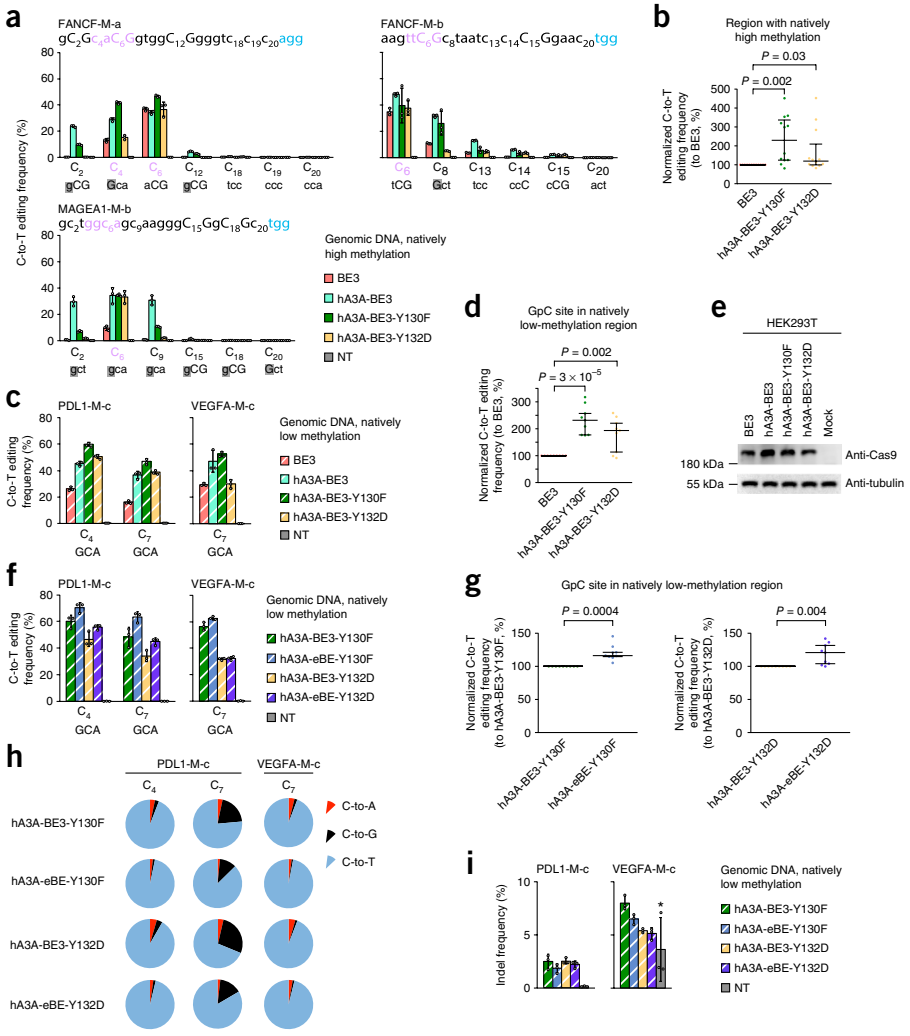


Figure 2 Improvements in hA3A-BE3. (**a**) Comparison of base-editing efficiencies induced by BE3, hA3A-BE3, hA3A-BE3-Y130F and hA3A-BE3-Y132D in genomic regions with natively high levels of DNA methylation. Target site sequences are shown with the overlapping editing window (positions 4–7) in lavender, PAM in cyan and CpG site in capitals. NT, native HEK293T cells with no treatment. (**b**) Statistical analysis of normalized C-to-T editing frequencies in the overlapping editing window shown in **a**, setting the ones induced by BE3 to 100%. $n = 12$ samples from three independent experiments. (**c**) Comparison of base editing efficiencies induced by BE3, hA3A-BE3, hA3A-BE3-Y130F and hA3A-BE3-Y132D at the cytosine of GpC in the overlapping editing window in genomic regions with natively low levels of DNA methylation. (**d**) Statistical analysis of normalized C-to-T editing frequencies shown in **c**, setting the ones induced by BE3 to 100%. $n = 9$ samples from three independent experiments. (**e**) Immunoblots of BEs transfected into HEK293T cells. Tubulin was used as a loading control and immunoblot images are representative of three independent experiments. Uncropped blot images are shown in **Supplementary Figure 11**. (**f**) Comparison of base editing efficiencies induced by hA3A-BE3-Y130F, hA3A-eBE-Y130F, hA3A-BE3-Y132D and hA3A-eBE-Y132D at the cytosine of GpC in the overlapping editing window in genomic regions with natively low levels of DNA methylation. (**g**) Statistical analysis of normalized C-to-T editing frequencies shown in (**f**), setting the ones induced by hA3A-BE3-Y130F (left) or hA3A-BE3-Y132D (right) as 100%. $n = 9$ samples from three independent experiments. (**h**,**i**) Comparison of product purity (**h**) and indels (**i**) yielded by hA3A-BE3-Y130F, hA3A-eBE-Y130F, hA3A-BE3-Y132D and hA3A-eBE-Y132D in genomic DNA regions with natively low levels of DNA methylation. Asterisk denotes an unusually high basal indel frequency (or amplification, sequencing or alignment artifact) at the examined VEGFA-M-c site in NT. (**a**,**c**,**f**,**i**) Means ± s.d. were from three independent experiments. (**b**,**d**,**g**) *P* value, one-tailed Student's *t* test. The median and interquartile range are shown.

the editing window of hA3A-BE3. Based on a previous finding that mutations of residue Y130, D131, or Y132 partially reduced the deamination activity of hA3A[17], we introduced similar mutations into hA3A-BE3. Designated mutations (Y130F, D131Y or Y132D) successfully narrowed the editing window with little effect on the base editing efficiency (**Supplementary Fig. 6**), whereas mutations in the zinc-coordination motif[18] (C101S and C106S) almost completely eliminated the deaminase activity (**Supplementary Fig. 7**).

We then focused on two engineered hA3A-BE3s, hA3A-BE3-Y130F and hA3A-BE3-Y132D, which have similar editing windows (positions 3–8 for hA3A-BE3-Y130F and positions 3–7 for hA3A-BE3-Y132D, **Supplementary Fig. 6b**,**d**) to that of BE3 (positions 4–8). In highly methylated regions, hA3A-BE3-Y130F and hA3A-BE3-Y132D induced higher editing efficiencies than BE3 at all editable sites in overlapping editing windows (positions 4–7) (**Fig. 2a**, cytosines in lavender, and **Fig. 2b**, median ~2.3-fold, $P = 0.002$ for hA3A-BE3-Y130F and median ~1.2-fold, $P = 0.03$ for hA3A-BE3-Y132D). For cytosines outside of overlapping editing windows, hA3A-BE3-Y132D induced C-to-T editing frequencies similar to those of BE3 while hA3A-BE3-Y130F induced higher editing frequencies (**Fig. 2a**, cytosines in black). Like the original hA3A-BE3, both engineered hA3A-BE3-Y130F and hA3A-BE3-Y132D edited cytosines in GpC contexts more efficiently than BE3 in overlapping editing windows (**Supplementary Fig. 8e** and **Fig. 2c**,**d**, median ~2.3-fold, $P = 3 \times 10^{-5}$ for hA3A-BE3-Y130F and median ~1.9-fold, $P = 0.002$ for hA3A-BE3-Y132D). Protein expression levels of hA3A-BE3-Y130F and hA3A-BE3-Y132D were very similar to that of BE3 (**Fig. 2e**), though the two engineered hA3A-BEs induced higher C-to-T editing efficiencies (**Fig. 2b**,**d**). In terms of product purity, we found that hA3A-BE3-Y130F yielded higher C-to-T fractions (**Supplementary Fig. 8a**,**b**, median ~96.3% by hA3A-BE3-Y130F compared to ~95.6% by BE3, $P = 0.03$ in high-methylation regions and **Supplementary Fig. 8f**,**g**, median ~92% by hA3A-BE3-Y130F compared to ~90% by BE3, $P = 0.002$ in low-methylation regions) but more indels (**Supplementary Fig. 8c**,**d**, median ~2.1-fold, $P = 0.0002$ in high-methylation regions and **Supplementary Fig. 8h**,**i**, median ~1.3-fold in low-methylation regions, $P = 0.12$) than BE3. The product purity induced by hA3A-BE3-Y132D was higher than BE3 in native low-methylation regions (**Supplementary Fig. 8f**,**g**, median ~93% by hA3A-BE3-Y132D compared to ~90% by BE3, $P = 0.001$), but lower in native high-methylation regions (**Supplementary Fig. 8a**,**b**, median ~94.9% by hA3A-BE3-Y132D compared to ~95.6% by BE3, $P = 0.03$). Nevertheless, indel frequencies induced by hA3A-BE3-Y132D were comparable to those induced by BE3 at all tested sites (**Supplementary Fig. 8c**,**d**,**h**,**i**, median ~1.2-fold in both high- and low- methylation regions).

To further enhance the C-to-T base editing system[13,19], we fused three copies of the 2A (self-cleaving peptide)-uracil DNA glycosylase inhibitor (UGI) sequence to the C terminus of hA3A-BE3-Y130F and hA3A-BE3-Y132D to develop hA3A-eBE-Y130F and hA3A-eBE-Y132D. In low-methylation regions, hA3A-eBE-Y130F and hA3A-eBE-Y132D induced significantly higher editing efficiencies (**Fig. 2f**,**g**, median ~1.2-fold, $P = 0.0004$ for hA3A-eBE-Y130F and median ~1.2-fold, $P = 0.004$ for hA3A-eBE-Y132D), higher product purity (**Fig. 2h** and **Supplementary Fig. 9a**, median ~96% by hA3A-eBE-Y130F compared to ~94% by hA3A-BE3-Y130F, $P = 0.006$ and median ~96% by hA3A-eBE-Y132D compared to ~92% by hA3A-BE3-Y132D, $P = 0.004$) and lower indel frequencies (**Fig. 2i** and **Supplementary Fig. 9b**, median decrement ~21%, $P = 4 \times 10^{-5}$ for hA3A-eBE-Y130F and median decrement ~9%, $P = 0.03$ for hA3A-eBE-Y132D) than hA3A-BE3-Y130F and hA3A-BE3-Y132D. In high-methylation regions, hA3A-eBE-Y130F and hA3A-eBE-Y132D yielded significantly higher product purity (**Supplementary Fig. 9c**,**d**, median ~97% by hA3A-eBE-Y130F compared to ~95% by hA3A-BE3-Y130F, $P = 0.003$ and median ~97% by hA3A-eBE-Y132D compared to ~95% by hA3A-BE3-Y132D, $P = 0.003$) and lower indel frequencies (**Supplementary Fig. 9e**,**f**, median decrement ~23%, $P = 2 \times 10^{-7}$ for hA3A-eBE-Y130F and median decrement ~21%, $P = 4 \times 10^{-5}$ for hA3A-eBE-Y132D) than hA3A-BE3-Y130F and hA3A-BE3-Y132D, respectively, though editing efficiencies remained the same (**Supplementary Fig. 9g**,**h**, median ~1-fold for hA3A-eBE-Y130F and hA3A-eBE-Y132D). Together, these results indicated that hA3A-BE3-Y130F, hA3A-BE3-Y132D, hA3A-eBE-Y130F and hA3A-eBE-Y132D can mediate highly efficient base editing in narrowed editing windows compared to the original hA3A-BE3 in all contexts examined.

Here we have demonstrated that hA3A-BE3 and its engineered forms can comprehensively induce efficient base editing in all contexts examined, including both methylated DNA regions and GpC dinucleotides (**Supplementary Fig. 10**). In the future, hA3A can also be conjugated with other Cas proteins to further expand the scope of base editing[5,20].

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to G.G. Carmichael and L.-L. Chen for critical reading of this paper, L. Lei and L. Wang for participating in the western blot and plasmid construction, and J. Liang for technical support. This work was supported by grants 31730111 (L.Y.), 91540115 (L.Y.), 31600654 (J.C.), 31600619 (B.Y.) and 31471241 (L.Y.) from the NSFC and grants 16PJ1407000 (J.C.) and 16PJ1407500 (B.Y.) from the Shanghai Pujiang Program.

AUTHOR CONTRIBUTIONS

J.C., L.Y. and X.H. conceived, designed and supervised the project. J.C. managed the project. X.W. and J.L. performed most experiments with the help of J. Wu and R.W. on cell culture and plasmid construction. J. Wei prepared libraries for deep sequencing and Y.W. performed bioinformatics analyses, supervised by L.Y. J.C., L.Y. and B.Y. wrote the paper with input from all authors.

COMPETING INTERESTS

The authors declare no competing interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A. & Liu, D.R. *Nature* **533**, 420–424 (2016).
2. Nishida, K. *et al. Science* **353**, 919–921 (2016).
3. Kim, K. *et al. Nat. Biotechnol.* **35**, 435–437 (2017).
4. Zong, Y. *et al. Nat. Biotechnol.* **35**, 438–440 (2017).
5. Hu, J.H. *et al. Nature* **556**, 57–63 (2018).
6. Li, X. *et al. Nat. Biotechnol.* **36**, 324–327 (2018).
7. Schübeler, D. *Nature* **517**, 321–326 (2015).
8. Nabel, C.S. *et al. Nat. Chem. Biol.* **8**, 751–758 (2012).
9. Chelico, L., Pham, P., Calabrese, P. & Goodman, M.F. *Nat. Struct. Mol. Biol.* **13**, 392–399 (2006).
10. Chen, J., Miller, B.F. & Furano, A.V. *Elife* **3**, e02001 (2014).
11. Carpenter, M.A. *et al. J. Biol. Chem.* **287**, 34801–34808 (2012).
12. Ito, F., Fu, Y., Kao, S.A., Yang, H. & Chen, X.S. *J. Mol. Biol.* **429**, 1787–1799 (2017).
13. Komor, A.C. *et al. Sci. Adv.* **3**, eaao4774 (2017).
14. St. Martin, A. *et al. Nucleic Acids Res.* https://dx.doi.org/10.1093/nar/gky332 (2018).
15. Lei, L. *et al. Nat. Struct. Mol. Biol.* **25**, 45–52 (2018).
16. Yang, B., Li, X., Lei, L. & Chen, J. *J. Genet. Genomics* **44**, 423–437 (2017).
17. Shi, K. *et al. Nat. Struct. Mol. Biol.* **24**, 131–139 (2017).
18. Mitra, M. *et al. Nucleic Acids Res.* **42**, 1095–1110 (2014).
19. Wang, L. *et al. Cell Res.* **27**, 1289–1292 (2017).
20. Kleinstiver, B.P. *et al. Nature* **523**, 481–485 (2015).

ONLINE METHODS

Plasmid construction. Primer sets (hA3A_PCR_F/hA3A_PCR_R) were used to amplify the fragment Human_APOBEC3A with template pUC57-Human_APOBEC3A (synthesized by Genscript). Then the fragment Human_APOBEC3A was cloned into the SacI and SmaI linearized pCMV-BE3 (Addgene, 73021) with plasmid recombination kit Clone Express (Vazyme, C112-02) to generate the hA3A-BE3 expression vector pCMV-hAPOBEC3A-XTEN-D10A-SGGS-UGI-SGGS-NLS. hA3B-BE3, hA3C-BE3, hA3D-BE3, hA3F-BE3, hA3G-BE3, hA3H-BE3, hAID-BE3, hA1-BE3, mA3-BE3, mAID-BE3, mA1-BE3 and cAICDA-BE3 expression vectors were constructed with the same strategy. The pmCDA1 expression vector pcDNA3.1_pCMV-nCas-PmCDA1-ugi pH1-gRNA(HPRT) was purchased from Addgene (79620).

Primer sets (SupF_PCR_F/SupF_PCR_R) were used to amplify the fragment SupF with template shuttle vector pSP189. Then the fragment SupF was cloned into pEASY-ZERO-BLUNT (TransGen Biotech, CB501) to generate the vector pEASY-SupF-ZERO-BLUNT.

Oligonucleotides SupF_sg1_FOR/SupF_sg1_REV and SupF_sg2_FOR/SupF_sg2_REV were annealed and ligated into BsaI-linearized pGL3-U6-sgRNA-PGK-puromycin (Addgene, 51133) to generate the sgRNA expression vectors psgSupF-1 and psgSupF-2 that target the *SupF* gene in pEASY-SupF-ZERO-BLUNT.

Two primer sets, (hA3A_PCR_F/hA3A_Y130F_PCR_R) and (hA3A_Y130F_PCR_F/hA3A_PCR_R), were used to amplify the Y130F-containing fragment hA3A-Y130F. Then the fragment was cloned into the ApaI- and SmaI-linearized hA3A-BE3 expression vector to generate the hA3A-BE3-Y130F expression vector pCMV-hAPOBEC3A_Y130F-XTEN-D10A-SGGS-UGI-SGGS-NLS. hA3A-BE3-D131Y, hA3A-BE3-Y132D, hA3A-BE3-C101S and hA3A-BE3-C106S expression vectors were constructed with the same strategy.

Primer sets (hA3A_PCR_F/hA3A_PCR_R) were used to amplify the fragment Human_APOBEC3A_Y130F with template hA3A-BE3-Y130F. Then the fragment Human_APOBEC3A_Y130F was cloned into the SacI- and SmaI-linearized pCMV-eBE-S3[19] to generate the hA3A-eBE-Y130F expression vector pCMV-hAPOBEC3A_Y130F -XTEN-D10A-SGGS-UGI-SGGS-NLS-T2A-UGI-NLS-P2A-UGI-NLS-T2A-UGI-NLS. The hA3A-eBE-Y132D expression vector was constructed in a similar way.

Oligonucleotides hEMX1_FOR/hEMX1_REV were annealed and ligated into BsaI-linearized pGL3-U6-sgRNA-PGK-puromycin to generate sgEMX1 expression vector psgEMX1. Other sgRNA expression vectors were constructed with the same strategy.

The sequences of the oligonucleotides used for plasmid construction were listed in **Supplementary Table 1** and the sequences of plasmids are listed in the **Supplementary Note**.

Antibodies. Antibodies were purchased from the following sources: against α-tubulin (T6199), Sigma; against Cas9 (ab204448), Abcam.

Immunoblotting analysis. Protein samples were incubated at 95 °C for 20 min and separated by SDS–PAGE in sample loading buffer, and proteins were transferred to nitrocellulose membranes (Thermo Fisher Scientific). After blocking with TBST (25 mM Tris, pH 8.0, 150 mM NaCl and 0.1% Tween 20) containing 5% (w/v) nonfat dry milk for 2 h, the membrane was reacted overnight with the indicated primary antibody. After extensive washing, the membranes were reacted with HRP-conjugated secondary antibodies for 1 h. Reactive bands were developed in ECL (Thermo Fisher Scientific) and detected with an Amersham Imager 600.

Cell culture and transfection. HEK293T cells from ATCC were maintained in DMEM (10566, Gibco/Thermo Fisher Scientific) + 10% FBS (16000-044, Gibco/Thermo Fisher Scientific) and regularly tested to exclude mycoplasma contamination.

The dCas9-Suntag-TetCD system[21] was used to induce targeted demethylation of genomic regions with natively high levels of methylation—for example, the *FANCF*, *MAGEA1* and *MSSK1* regions. The dCas9-DNMT3a-DNMT3l system[22] was used to induce targeted methylation of genomic regions with natively low levels of methylation—for example, the *VEGFA* and *PDL1* regions. HEK293T cells were transfected using LIPOFECTAMINE 2000 (Life, Invitrogen) with 3 μg pCAG-scFvGCN4sfGFPTET1CD (synthesized by Genscript) and 1 μg sgRNA expression vector or with 3 μg dCas9-DNMT3a-DNMT3l (synthesized by Genscript) and 1 μg sgRNA expression vector. Blasticidin (10 μg/ml, Sigma, 15205) and puromycin (1 μg/ml, Merck, 540411) were added 24 h after transfection. One week later, a portion of cells were collected to determine DNA methylation level and others were stored in liquid nitrogen for base editing. The sgRNAs used to induce genomic DNA methylation or demethylation are the ones used to induce base editing.

For base editing in genomic DNA, HEK293T cells were seeded in a 24-well plate at a density of 1.6×10^5 per well and transfected with 200 μl serum-free Opti-MEM that contained 5.04 μl Lipofectamine LTX (Life, Invitrogen), 1.68 μl Lipofectamine Plus (Life, Invitrogen), 1 μg BE3 expression vector (or hA3A-BE3, hA3A-BE3-Y130F, hA3A-BE3-D131Y, hA3A-BE3-Y132D, hA3A-BE3-C101S, hA3A-BE3-C106S, hA3A-eBE-Y130F or hA3A-eBE-Y132D expression vector) and 0.68 μg sgRNA expression vector. After 72 h, the genomic DNA was extracted from the cells with QuickExtract DNA Extraction Solution (QE09050, Epicentre) or the cells were lysed in 2× SDS loading buffer for western blotting.

For base editing in a plasmid vector, 293T cells were seeded in a six-well plate at a density of 3×10^5 per well and transfected with 500 μl serum-free Opti-MEM that contained 4 μl Lipofectamine LTX (Life, Invitrogen), 2 μl Lipofectamine Plus (Life, Invitrogen), 1 μg BE3 expression vector (or hA3A-BE3, hA3B-BE3, hA3C-BE3, hA3D-BE3, hA3F-BE3, hA3G-BE3, hA3H-BE3, hAID-BE3, hA1-BE3, mA3-BE3, mAID-BE3, mA1-BE3, cAICDA-BE3 or pmCDA1 expression vector) and 0.5 μg sgRNA expression vector. After 24 h, these cells were transfected with 500 μl serum-free Opti-MEM that contained 4 μl Lipofectamine LTX, 2 μl Lipofectamine Plus and 1.5 μg unmethylated (or methylated) pEASY-SupF-ZERO-BLUNT. After 48 h, the plasmids were extracted from the cells with a TIANprep Mini Plasmid Kit (DP103-A, TIANGEN) or the cells were lysed in 2× SDS loading buffer for western blot.

Bisulfite sequencing analysis. Genomic DNA was isolated and treated with bisulfite according to the instruction of the EZ DNA Methylation-Direct Kit (Zymo Research, D5021). The bisulfite-treated DNA was PCR-amplified with Taq Hot Start Version (Takara, R007B). The PCR products were ligated into T-Vector pMDTM19 (Takara, 3271). Eight clones were picked and sequenced by Sanger sequencing (Genewiz). The primers used for bisulfite PCR are listed in **Supplementary Table 2**.

Plasmid DNA methylation. For *in vitro* methylation, 1 μl CpG methyltransferase (M.SssI, Life, EM0821) was used to methylate 2 μl plasmid DNA in a 20-μl reaction. After *in vitro* methylation, pEASY-SupF-ZERO-BLUNT was restricted with BstUI (NEB, R0518S) to determine the methylation level.

Blue vs. white colony screening. The plasmids extracted from transfected cells were transformed into *E. coli* strain MBM7070 (*lacZ*^{uag_amber}), which were grown on LB plates containing 50 μg/ml kanamycin, 1 mM IPTG and 0.03% Bluo-gal (Life, Invitrogen) at 37 °C overnight and then at room temperature for another day for maximal color development. The cumulative base editing frequency is calculated by dividing the number of white colonies by the total number of colonies.

DNA library preparation and sequencing. Target genomic sites were PCR amplified by high-fidelity DNA polymerase Primestar HS (Clontech) with primers flanking each examined sgRNA target site. The PCR primers used to amplify target genomic sequences are listed in **Supplementary Table 2**. Indexed DNA libraries were prepared by using the TruSeq ChIP Sample Preparation Kit (Illumina) with some minor modifications. Briefly, the PCR products were fragmented by Covaris S220 and then amplified by using the TruSeq ChIP Sample Preparation Kit (Illumina). After being quantitated with a Qubit High-Sensitivity DNA kit (Life, Invitrogen), PCR products with different tags were pooled together for deep sequencing by using the Illumina NextSeq 500 (2 × 150) or Hiseq X Ten (2 × 150) at the CAS-MPG Partner Institute for Computational Biology Omics Core, Shanghai, China. Raw read qualities were evaluated by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). For paired-end sequencing, only R1 reads were used. Adaptor sequences and read sequences on both ends with Phred quality score lower than 28 were trimmed. Trimmed reads were then mapped with the

doi:10.1038/nbt.4198

BWA-MEM algorithm (BWA v0.7.9a) to target sequences. After pile-up with samtools (v0.1.18), indels and base substitutions were further calculated.

Indel frequency calculation. Indels were estimated in the aligned regions spanning from upstream 8 nt of the target site to downstream 19 nt of PAM sites (50 bp). Indel frequencies were subsequently calculated by dividing reads containing at least one inserted and/or deleted nucleotide by all the mapped reads at the same region. Counts of indel-containing reads and total mapped reads are listed in **Supplementary Table 3**.

Base substitution calculation. Base substitutions were selected at each position of the examined sgRNA target sites that mapped with at least 1,000 independent reads, and obvious base substitutions were only observed at the targeted base editing sites. Counts of reads for each base and total reads are listed in **Supplementary Table 4**. Base substitution frequencies were calculated by dividing base substitution reads by total reads.

Calculation of BE-targetable genetic variants. The single nucleotide variants (SNVs) from the NCBI ClinVar database were overlapped with the pathogenic human allele sequence from the NCBI dbSNP database to calculate pathogenic T-to-C and A-to-G mutations. In 3,089 pathogenic T-to-C or A-to-G mutations, 2,499 are potentially editable by SpCas9-BE3, SaCas9-BE3, dLbCpf1-BE or xCas9-BE3 with nearby PAM sequences. These 2,499 BE-targetable SNVs are further subclassified according to their 3′ adjacent base preferences: CpA, CpC, CpG and CpT (**Fig. 1a**).

Statistical analysis. *P* values were calculated from one-tailed Student's *t* tests in this study.

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

Data availability. The deep-sequencing data from this study have been deposited in the NCBI Gene Expression Omnibus (accession code GSE114999) and the National Omics Data Encyclopedia (accession code OEP000030). The datasets used in this study are provided in **Supplementary Tables 3** and **4**.

21. Morita, S. *et al. Nat. Biotechnol.* **34**, 1060–1065 (2016).
22. Stepper, P. *et al. Nucleic Acids Res.* **45**, 1703–1713 (2017).

nature research

Corresponding author(s): Jia Chen, Xingxu Huang, Li Yang

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The single nucleotide variants (SNV) information was downloaded from NCBI ClinVar database (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz);The pathogenic human allele sequences were downloaded from NCBI dbSNP database (https://www.ncbi.nlm.nih.gov/snp/?term=((%22pathogenic%22%5BClinical+Significance%5D+%20AND+%22snp%22%5BSNP+Class%5D+AND+homo+sapiens%5BOrganism%5D+)+AND(%22y%22%5BAllele%5D%20).
Data analysis	High-throught sequencing read quality was evaluated by FastQC (v0.11.4) and then aligned to reference sequences by BWA (v0.7.9a). Substitution calling was performed by SAMtools (v0.1.18) mpileup function and indel calling was based on the CIGAR values of the SAM/BAM files (Li et al, Nat Biotechnol, 2018). BE-targetable genetic variants were selected by judging whether a nearby PAM sequence exists with a Python (v.2.7.13) script (Kim et al, Nat Biotechnol, 2017).

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 upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Nature Biotechnology: doi:10.1038/nbt.4198

Data

Policy information about availability of data

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

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

The deep-sequencing data from this study have been deposited in the NCBI Gene Expression Omnibus (accession no. GSE114999) and the National Omics Data Encyclopedia (accession no. OEP000030). The dataset will be released prior to publication.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-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 three times independently unless indicated. In previous studies using related experiments, the sample size has been determined to be sufficient to ensure reproducibility.
Data exclusions	No data were excluded
Replication	The experimental findings in all figures were reproduced successfully
Randomization	Samples were not randomized
Blinding	The investigators were not blinded to group allocation

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used	Antibodies were purchased from the following sources: against alpha-tubulin (T6199) - Sigma; against Cas9 (ab204448) - Abcam.
Validation	The alpha-tubulin (T6199, Sigma) antibody has been validated by western blot in human HeLa cell lysate (https://www.sigmaaldrich.com/catalog/product/sigma/t6199?lang=en&region=US); the Cas9 (ab204448, Abcam) antibody has been validated by western blot in human 293T cells transfected with DDDDK tag-CRISPR-Cas9 lysate (http://www.abcam.com/crispr-cas9-antibody-ab204448.html#description_images_1).

Nature Biotechnology: doi:10.1038/nbt.4198

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T cell line was from ATCC
Authentication	No cell lines were authenticated
Mycoplasma contamination	HEK293T cell line has been tested negative for mycoplasma contamination by PCR methods
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used

Nature Biotechnology: doi:10.1038/nbt.4198