




Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction

Luke W Koblan¹⁻³, Jordan L Doman¹⁻³, Christopher Wilson¹⁻³ , Jonathan M Levy¹⁻³ , Tristan Tay¹⁻³, Gregory A Newby¹⁻³ , Juan Pablo Maianti¹⁻³, Aditya Raguram¹⁻³ & David R Liu¹⁻³

Base editors enable targeted single-nucleotide conversions in genomic DNA. Here we show that expression levels are a bottleneck in base-editing efficiency. We optimize cytidine (BE4) and adenine (ABE7.10) base editors by modification of nuclear localization signals (NLS) and codon usage, and ancestral reconstruction of the deaminase component. The resulting BE4max, AncBE4max, and ABE4max editors correct pathogenic SNPs with substantially increased efficiency in a variety of mammalian cell types.

Point mutations represent the majority of known pathogenic human genetic variants¹. Base editors enable the direct installation and correction of targeted point mutations in genomic DNA. These fusion proteins include a catalytically impaired Cas9, natural or laboratory-evolved nucleobase deaminases, and, in some cases, proteins that help preserve the resulting single-nucleotide change^{2,3}. Cytidine base editors (e.g., BE4)⁴ convert target C•G base pairs to T•A and adenine base editors (e.g., ABE7.10)³ convert A•T to G•C. Collectively, these editors enable the targeted installation of all four transition mutations (C-to-T, G-to-A, A-to-G, and T-to-C), which account for 61% of human pathogenic single-nucleotide polymorphisms (SNPs) in the ClinVar database (Supplementary Fig. 1a,b). Base editors have been successfully used in diverse systems including prokaryotes, plants, fish, amphibians, mammals, and human embryos⁴⁻⁸. However, diminished efficiency of base editors at certain target sites or in particular cell types limits their utility.

To test if base editing in cells is limited by transfection efficiency of the base-editor plasmid or by base-editor expression, we transfected HEK293T cells with three-plasmid mixtures in which one plasmid expresses mCherry (transfection marker), another expresses a targeting sgRNA, and a third expresses either (1) BE4 alone, (2) BE4 and GFP on separate promoters to follow transfection of this plasmid,

or (3) a BE4-P2A-GFP fusion to directly follow BE4 expression (Fig. 1a). P2A is a self-cleaving peptide⁹ that couples GFP production with full-length BE4 production.

Transfection with (1) and collecting mCherry-positive cells resulted in $45 \pm 7.1\%$ average C•G-to-T•A conversion within the base-editing activity window (positions 4–8, counting the PAM as positions 21–23) at five genomic loci (Fig. 1b,c). The average editing efficiency among mCherry and GFP double-positive cells did not improve following transfection with (2) (Fig. 1c), suggesting that transfection efficiency was not limiting editing efficiency. In contrast, double-positive cells following transfection with (3) exhibited $65 \pm 6.4\%$ editing, 1.9-fold higher than in sorted cells following (2) (Fig. 1c), indicating that cells expressing base editors and/or the amount of functional editor protein produced by each cell are major bottlenecks of editing efficiency.

To optimize nuclear localization, we tested all six combinations of BE4 N- and C-terminal fusions to the SV40 NLS used in BE4 or to a bipartite NLS (bpNLS)¹⁰ (Fig. 1d and Supplementary Fig. 2). A bpNLS at both the N and C termini (bis-bpNLS) performed best, resulting in a 1.3-fold average improvement in BE4-mediated C•G-to-T•A editing efficiency at five genomic loci (Fig. 1d and Supplementary Fig. 2a).

Next, we generated bis-bpNLS BE4 variants using eight codon usages: from IDT (used in BE4)⁴, GeneArt, Coller and co-workers¹¹, and GenScript. Every codon optimization method improved editing efficiency over IDT codon usage in HEK293T cells (Fig. 1e and Supplementary Fig. 2b). We also tested four chimeric codon-optimized BE4 variants that mixed different deaminase and Cas9 nickase codon usages (Supplementary Fig. 3a,b), but none outperformed the GenScript-only variant (BE4max), which induced 1.8-fold higher editing over bis-bpNLS BE4 with IDT codons (Fig. 1e and Supplementary Fig. 3b).

Chimeric editor experiments implicated expression of the APOBEC1 cytidine deaminase and Cas9 nickase as determinants of base-editing efficiency (Fig. 1e and Supplementary Fig. 3a). To further enhance APOBEC1 expression, we performed ancestral sequence reconstruction (ASR) using 468 APOBEC homologs (Supplementary Data 1 and Supplementary Sequences 1). ASR uses an alignment of protein sequences, an evolutionary model, and a resulting phylogenetic tree to infer ancestral sequences¹², and can improve protein expression while preserving activity^{13,14}. We created a maximum-likelihood APOBEC phylogeny and inferred the most likely sequences at ancestral nodes (Fig. 1f), then constructed five GenScript-coded bis-bpNLS-BE4 variants from five ancestral cytidine deaminases (Supplementary Fig. 4). Two ancestors, Anc689 and Anc687, containing 36 and 45 amino acid substitutions relative to rAPOBEC1, respectively, resulted in bis-bpNLS-BE4 variants that efficiently edited five test loci in HEK293T cells (Fig. 1g).

¹Merkin Institute of Transformative Technologies in Healthcare, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. ²Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts, USA. ³Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, USA. Correspondence should be addressed to D.R.L. (drliu@fas.harvard.edu).

Received 29 April; accepted 21 May; published online 29 May 2018; doi:10.1038/nbt.4172

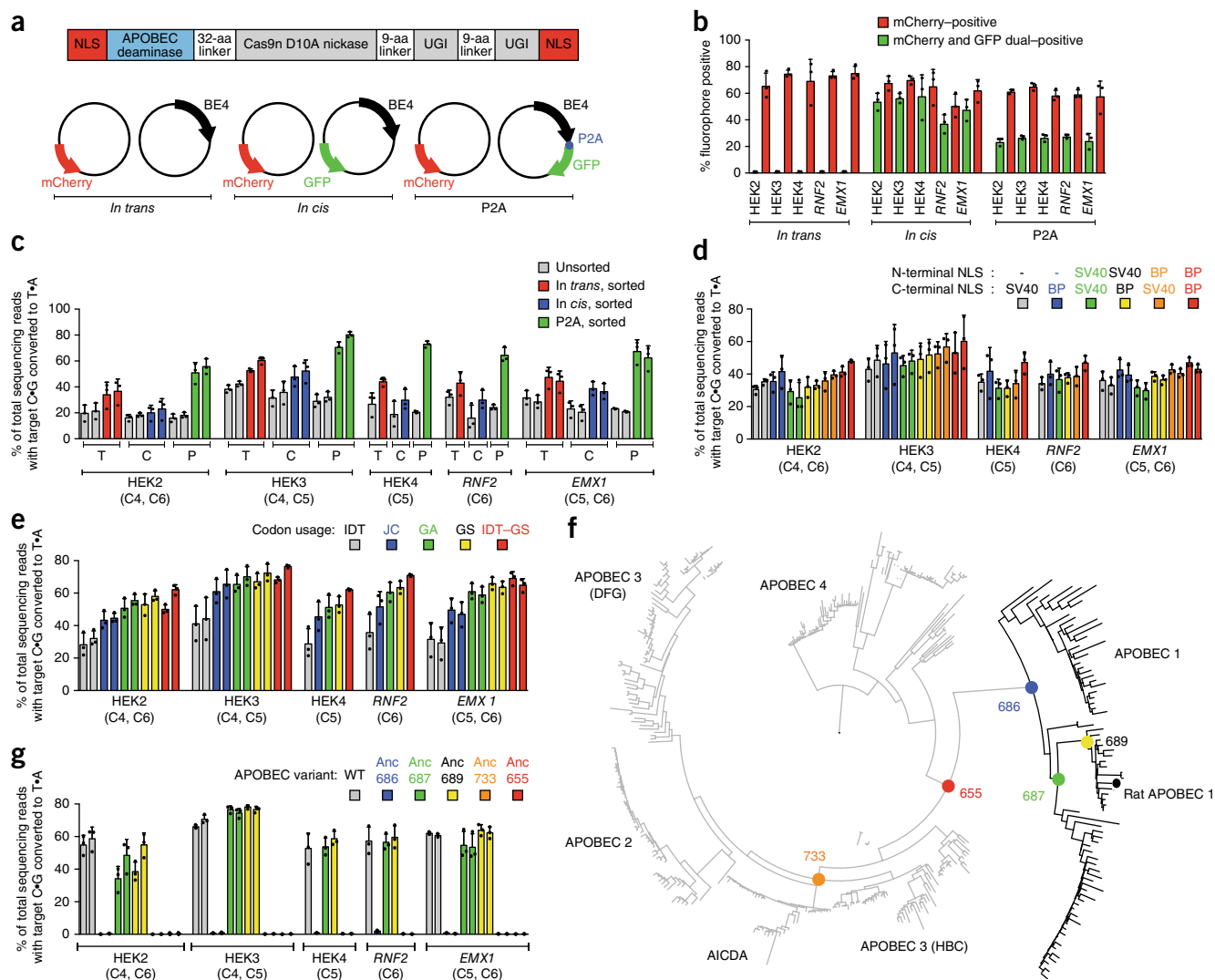


Figure 1 Identifying and addressing factors that limit base-editing efficiency in mammalian cells. **(a)** Plasmids used to elucidate the relationship between base-editor expression and editing efficiency in mammalian cells: mCherry (transfection control), and either BE4 (*in trans*), BE4 and GFP on separate promoters (*in cis*), or BE4–P2A–GFP (P2A). **(b)** Percent mCherry-positive or GFP-positive HEK293T cells 3 d after transfection of the constructs in **a**. **(c)** Target C>G-to-T>A editing in unsorted and sorted HEK293T cells. Sorted *in trans* cells were mCherry-positive, while sorted *in cis* and P2A cells were mCherry and GFP double-positive. **(d)** Effects of six NLS configurations on BE4 editing efficiency at five genomic loci in HEK293T cells. **(e)** Effects of five different codon usages on editing efficiency of bis-bpNLS-BE4 in HEK293T cells. IDT, Integrated DNA Technologies; JC, Jeff Collier; GA, GeneArt; GS, GenScript; IDT-GS, IDT APOBEC+GenScript Cas9 nickase. **(f)** Phylogenetic tree for ancestral APOBEC reconstruction. **(g)** Base editing of bis-bpNLS-BE4 variants with GenScript codons using the ancestral APOBEC domains in **f** in HEK293T cells. Values and error bars represent the mean and s.d. of $n = 3$ biologically independent experiments (dots) 3 d after transfection.

To characterize the base-editing activities of these optimized variants under suboptimal conditions, we compared eight different plasmid doses of BE4, BE4max, and AncBE4max (bis-bpNLS BE4 with the Anc689 APOBEC and GenScript codons) at three genomic loci in HEK293T cells (**Fig. 2a**). AncBE4max showed the highest activity across all tested sites over a range of plasmid doses, with BE4max performing slightly below, or similar to, AncBE4max (**Fig. 2a** and **Supplementary Fig. 5**). AncBE4max and BE4max improved over BE4 at rates ranging from 1.7-fold at higher plasmid doses to greater than ninefold at lower plasmid doses (**Fig. 2a**). Product purities of BE4max and AncBE4max—ratios of desired point mutations to indels and undesired mutations at the target nucleotide—were better than or comparable to those of BE4 (**Supplementary Figs. 6 and 7a**). The shape of the base-editing activity window for

BE4max and AncBE4max was unchanged compared to that of BE4 (**Supplementary Fig. 8a**).

BE4max and AncBE4max showed, respectively, greater than threefold and greater than fivefold higher mRNA (**Supplementary Fig. 9a**) and protein (**Supplementary Fig. 9b**) expression in HEK293T cells relative to BE4, improvements that correlated with editing efficiency. Among transfectable HEK293T cells expressing BE4max–P2A–GFP and AncBE4max–P2A–GFP (mCherry and GFP double-positive), base editing at three genomic loci averaged $89 \pm 0.9\%$ and $90 \pm 1.5\%$, respectively, while double-positive cells expressing BE4–P2A–GFP averaged $48 \pm 8.0\%$ (**Fig. 1e** and **Supplementary Fig. 9c,d**). Thus isolating cells expressing BE4max and AncBE4max results in much higher editing frequencies, which could facilitate creation of cell lines, agricultural strains, or animal models.

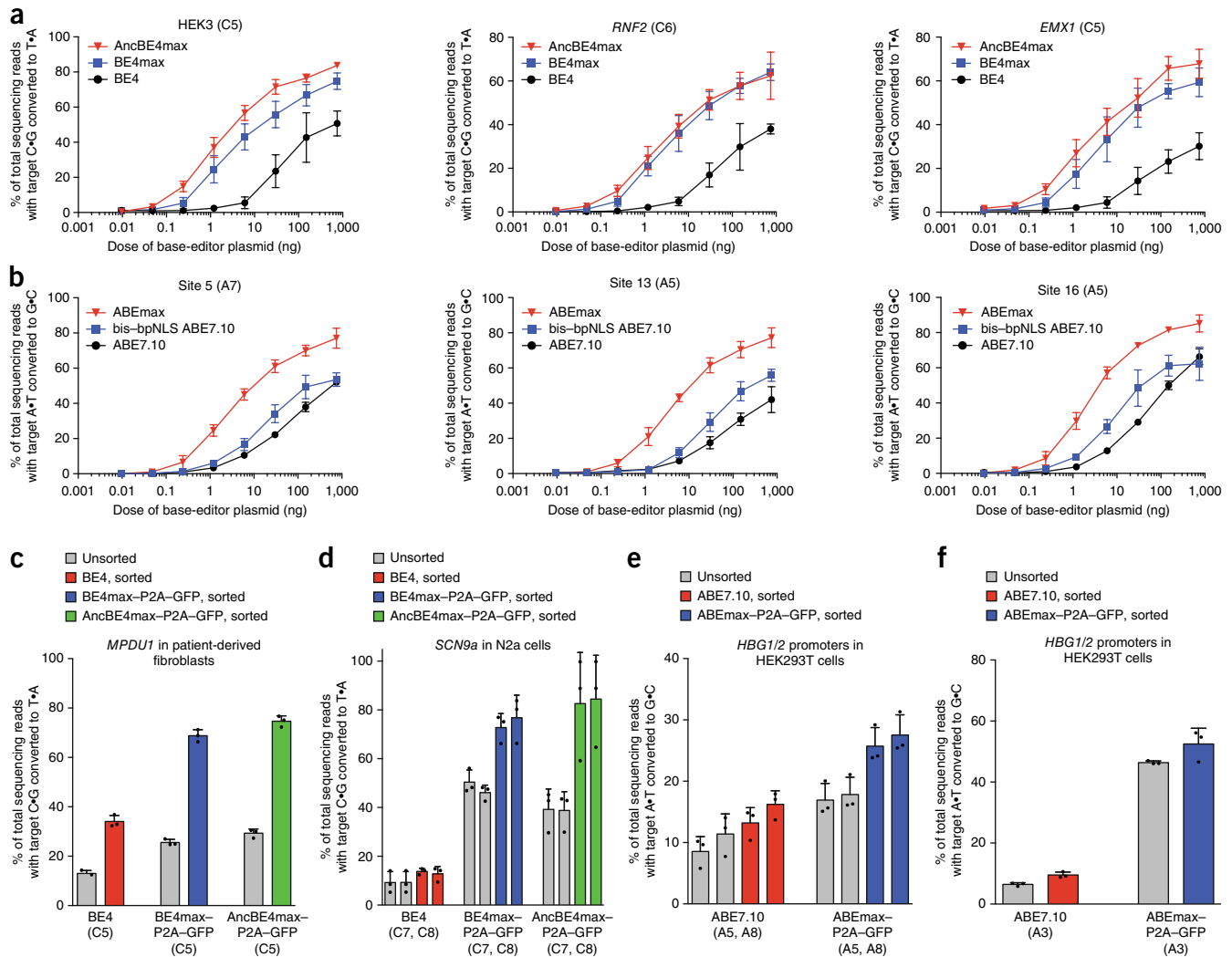


Figure 2 Properties of optimized AncBE4max, BE4max, and ABEmax compared to those of BE4 and ABE7.10. **(a)** Comparison at three genomic loci in HEK293T cells across eight plasmid doses. **(b)** Comparison of ABE7.10, bis-bpNLS-ABE with IDT codons, and ABEmax at three genomic loci in HEK293T cells across eight plasmid doses. **(c)** C•G-to-T•A editing for the correction of *MPDU1* Leu119Pro by BE4, BE4max, or AncBE4max in unsorted or sorted fibroblasts derived from congenital disorder of glycosylation (CDG) type 1f patients. BE4 samples were sorted for mCherry and BE4max–P2A–GFP and AncBE4max–P2A–GFP samples for GFP-positive cells. **(d)** C•G-to-T•A editing for the installation of the 3' splice acceptor in intron 6 of *SCN9a* in mouse N2a neuroblastoma cells, unsorted or sorted as described in **(c)**. **(e)** A•T-to-G•C editing for the installation of –116 A→G and –113 A→G in the *HBG* fetal hemoglobin promoters by ABE7.10 or ABEmax in HEK293T cells. ABE samples were sorted for mCherry-positive cells, and ABEmax–P2A–GFP samples were sorted for mCherry and GFP double-positive cells. **(f)** A•T-to-G•C editing to install *HBG* promoter mutation –175 T→C by ABE7.10 or ABEmax in HEK293T cells, unsorted or sorted as described in **(e)**. Values and error bars represent the mean and standard deviation of $n = 3$ biologically independent experiments (dots) 3 d after transfection.

Adenine base editors (ABEs) use a laboratory-evolved deoxyadenosine deaminase and a Cas9 nickase to mediate the conversion of target A•T to G•C base pairs³, reversing the most common class of point mutations in living systems¹⁵ (**Supplementary Fig. 1b**). We applied the above improvements to ABE7.10 (ref. 3). Replacing the SV40 NLS in ABE7.10 with the bis-bpNLS increased editing efficiencies ~1.5- to twofold at suboptimal ABE doses in HEK293T cells (**Fig. 2b**). GenScript codon optimization of bis-bpNLS ABE7.10 (ABEmax) resulted in 1.3- to 7.9-fold higher editing levels than IDT codon usage (in ABE7.10) at high and low plasmid doses, respectively (**Fig. 2b** and **Supplementary Figs. 10** and **11**). Although indels from ABEmax remained rare ($\leq 1.6\%$), they were elevated from the virtually undetectable indel levels of ABE7.10 (ref. 3 and **Supplementary Fig. 12**). ABEmax exhibited increases in mRNA and protein levels in HEK293T cells compared to those of ABE (**Supplementary**

Fig. 13), and product purity and the editing window remained unchanged (**Supplementary Figs. 7b** and **8b**). These findings establish that improvements in nuclear localization and codon usage that benefit BE4 also enhance ABE efficiency.

We evaluated the ability of BE4max, AncBE4max, and ABEmax (**Supplementary Sequences 2** and **3**) to edit disease-relevant loci in diverse cell types. Patient-derived fibroblasts harboring the *MPDU1* Leu119Pro T→C mutation, which drives congenital disorder of glycosylation type 1f (ref. 16), were nucleofected with plasmids expressing BE4, BE4max–P2A–GFP, or AncBE4max–P2A–GFP. The pathogenic SNP was corrected 2.0- and 2.2-fold more efficiently by BE4max ($26 \pm 1.3\%$ unsorted, $69 \pm 2.5\%$ sorted) and AncBE4max ($29 \pm 1.7\%$ unsorted, $75 \pm 2.2\%$ sorted) than by BE4 ($13 \pm 1.2\%$ unsorted, $34 \pm 2.4\%$ sorted) (**Fig. 2c** and **Supplementary Fig. 14a**). Second, we used BE4max and AncBE4max to mutate the splice acceptor of *SCN9a*

intron 6a in mouse N2a neuroblastoma cells in the chronic-pain-associated voltage-gated sodium channel $\text{NaV}_{1.7}$ (*SCN9a* gene)¹⁷. BE4, BE4max, and AncBE4max, respectively, resulted in $9.3 \pm 4.4\%$, $50 \pm 5.0\%$, and $39 \pm 7.7\%$ (unsorted) or $14 \pm 1.3\%$, $77 \pm 9.3\%$, and $84 \pm 18\%$ (sorted) editing (Fig. 2d and Supplementary Fig. 14a), 4.2- to 6.0-fold improvements favoring BE4max and AncBE4max. In one sorted sample, 99.8% of cells expressing AncBE4max contained mutations at both targeted *SCN9a* nucleotides (Supplementary Fig. 15). Third, we used ABEmax to install activating mutations in the promoters of *HBG1* or *HBG2* (γ -globin) that can rescue β -globin disorders with two sgRNAs: (1) $-116 \text{ A} \rightarrow \text{G}$ and $-113 \text{ A} \rightarrow \text{G}$; and (2) $-175 \text{ T} \rightarrow \text{C}$ ^{18,19}. For the first sgRNA, ABEmax resulted in approximately double the $-116 \text{ A} \rightarrow \text{G}$ and $-113 \text{ A} \rightarrow \text{G}$ conversion than ABE7.10 in both unsorted and sorted HEK293T cells (Fig. 2e). For the second sgRNA, ABE7.10 and ABEmax, respectively, induced $6.5 \pm 0.57\%$ and $46 \pm 0.55\%$ (unsorted) and $10 \pm 1.0\%$ and $52 \pm 5.2\%$ (sorted) editing in HEK293T cells (Fig. 2f and Supplementary Fig. 14c), representing 5.2- and 7.1-fold improvements favoring ABEmax.

BE4max, AncBE4max, and ABEmax thus offer increased editing in a variety of settings, especially under suboptimal conditions or at sites previously edited with modest efficiency. These improvements in expression and nuclear localization may also benefit other base-editor delivery methodologies, including viral, mRNA, and RNP delivery.

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

This work was supported by the Ono Pharma Foundation, DARPA HR0011-17-2-0049, US NIH RM1 HG009490, R01 EB022376, and R35 GM118062, and HHMI.

Flow cytometry was supported by NCI P30CCA14051. L.W.K. is an NSF Graduate Research Fellow and was supported by NIH Training Grant T32 GM095450. J.L.D. gratefully acknowledges graduate fellowship support from the NSF and Hertz Foundation. We thank J. Collier, G. Hansen, M. Weiss, and A. Sharma for helpful discussions.

AUTHOR CONTRIBUTIONS

L.W.K., J.L.D., C.W., J.M.L., T.T., G.A.N., and J.P.M. generated reagents and conducted experiments. C.W. and A.R. performed computational analyses. D.R.L. supervised the research. All authors contributed to writing the manuscript.

COMPETING INTERESTS

D.R.L. is a consultant and co-founder of Editas Medicine, Pairwise Plants, and Beam Therapeutics, companies that use genome editing. L.W.K., J.L.D., C.W., and D.R.L. have filed patent applications on aspects on this work. The authors declare no competing non-financial 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.

- Landrum, M.J. *et al. Nucleic Acids Res.* **44**, D862–D868 (2016).
- Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A. & Liu, D.R. *Nature* **533**, 420–424 (2016).
- Gaudelli, N.M. *et al. Nature* **551**, 464–471 (2017).
- Komor, A.C. *et al. Sci. Adv.* **3**, eaao4774 (2017).
- Li, G. *et al. Protein Cell* **8**, 776–779 (2017).
- Liang, P. *et al. Protein Cell* **8**, 811–822 (2017).
- Ryu, S.-M. *et al. Nat. Biotechnol.* <http://dx.doi.org/10.1038/nbt.4148> (2018).
- Hess, G.T., Tycko, J., Yao, D. & Bassik, M.C. *Mol. Cell* **68**, 26–43 (2017).
- Kim, J.H. *et al. PLoS One* **6**, e18556 (2011).
- Suzuki, K. *et al. Nature* **540**, 144–149 (2016).
- Hanson, G. & Collier, J. *Nat. Rev. Mol. Cell Biol.* **19**, 20–30 (2018).
- Harms, M.J. & Thornton, J.W. *Nat. Rev. Genet.* **14**, 559–571 (2013).
- Wheeler, L.C., Lim, S.A., Marqusee, S. & Harms, M.J. *Curr. Opin. Struct. Biol.* **38**, 37–43 (2016).
- Risso, V.A., Gavira, J.A., Mejia-Carmona, D.F., Gaucher, E.A. & Sanchez-Ruiz, J.M. *J. Am. Chem. Soc.* **135**, 2899–2902 (2013).
- Krokan, H.E., Drabløs, F. & Slupphaug, G. *Oncogene* **21**, 8935–8948 (2002).
- Schenk, B. *et al. J. Clin. Invest.* **108**, 1687–1695 (2001).
- Bennett, D.L. & Woods, C.G. *Lancet Neurol.* **13**, 587–599 (2014).
- Liu, N. *et al. Cell* **173**, 430–442 e417 (2018).
- Amato, A. *et al. Int. J. Lab. Hematol.* **36**, 13–19 (2014).

ONLINE METHODS

General methods. PCR was performed using either Phusion U Green Multiplex PCR Master Mix (ThermoFisher Scientific) or Q5 Host Start High-Fidelity 2× Master Mix (New England BioLabs) unless otherwise noted. All plasmids were assembled by either the USER cloning method as previously described²⁰ or by Gibson assembly²¹. Plasmids for mammalian cell transfections were prepared using an endotoxin removal plasmid purification system, ZymoPURE Plasmid Midiprep (Zymo Research Corporation).

Cell culture conditions. HEK293T cells (ATCC CRL-3216) were cultured in Dulbecco's Modified Eagle's Medium (DMEM, Corning) supplemented with 10% FBS (FBS), penicillin, and streptomycin (ThermoFisher Scientific). Fibroblast cell lines were maintained in DMEM supplemented with 15% FBS. N2a cells were maintained in DMEM supplemented with 10% FBS.

HEK293T transfection and genomic DNA preparation. HEK293T cells were seeded into 48-well Poly-D-Lysine-coated plates (Corning) in the absence of antibiotic. 12–15 h after plating, cells were transfected with 1 μ L of Lipofectamine 2000 (ThermoFisher Scientific) using 750 ng of base-editor plasmid, 250 ng of guide RNA plasmid, and 20 ng of fluorescent protein expression plasmid as a transfection control. Unless otherwise stated, cells were cultured for 3 d before they were washed with PBS (ThermoFisher Scientific). Genomic DNA was extracted by addition of 150 μ L of freshly prepared lysis buffer (10 mM Tris-HCl, pH 7.5, 0.05% SDS, 25 μ g/mL proteinase K (ThermoFisher Scientific)) directly into each transfected well. The resulting mixture was incubated for 1 h at 37 °C before a 30-min enzyme inactivation step at 80 °C. Guide RNA sequences for HEK2, HEK3, HEK4, RNF2, EMX1, site 2, site 5, site 13, and site 16 were previously reported^{2–4}. SCN, MPDU1, HBG Site 1, and HBG site 2 were cloned as described in **Supplementary Sequences 4**.

HEK293T base editing dose titrations. HEK293T cells were seeded as described above and transfected with a mixture of base-editor plasmid, guide RNA plasmid, pUC, and GFP. 250 ng of guide RNA plasmid and 20 ng of GFP transfection control plasmid were used for all samples. Base editor and pUC plasmids were combined in different amounts to maintain a constant amount of total DNA per transfection.

Fluorescence-activated cell sorting. Flow cytometry analysis was carried out using a FACSAria II (BD Biosciences). HEK293T cells were transfected with guide RNA expression plasmids, fluorophore expression plasmids, and editor expression plasmids. *In trans* samples were sorted for mCherry-positive cells. Both the *in cis* and P2a samples were sorted for both GFP and mCherry double-positive cells. A stringent mCherry-positive gate was used to avoid mCherry false positives. N2a cells and fibroblasts were sorted for mCherry-positive and GFP-positive cells. Genomic DNA for sorted and unsorted FACS samples was isolated using the Agencourt DNAdvance Genomic DNA Isolation Kit (Beckman Coulter) according to the manufacturer's instructions. Gating for all cell types can be found in **Supplementary Data 2**.

Nucleofection of fibroblasts and genomic DNA extraction. Cells were nucleofected using the Primary P2 Cell Line 4D-Nucleofector X Kit (Lonza) according to manufacturer's protocol. 1.25×10^5 cells were nucleofected in 20 μ L of P2 buffer supplemented with 750 ng of editor, 250 ng of guide RNA plasmid, and 20 ng of mCherry nucleofection marker. Cells were nucleofected in a 16-well nucleocuvette strip using the DT-130 program. Following a 3-d incubation, cells were flow-sorted and genomic DNA was extracted as described for HEK293T cells above.

High-throughput DNA sequencing (HTS) of genomic DNA. HTS of genomic DNA from HEK293T cells was performed as described previously^{2–4}. For fibroblasts, 34 cycles of amplification were used for PCR1. Primers for PCR 1 of HEK2, HEK3, HEK4, RNF2, EMX1, ABE Site 2, ABE Site 5, ABE Site 13, ABE Site 16, and HBG loci were used as previously described^{3,4,22}. PCR 1 primers for type 1F congenital glycosylation disorder, SCN9a, and all previously used loci are listed in **Supplementary Sequences 5**.

General HTS analysis. Sequencing reads were demultiplexed using the MiSeq Reporter (Illumina) and Fastq files were analyzed using open-source analysis tools. FASTQ files were aligned to the reference genome using the Burrows–Wheeler aligner (bwa-mem)²³. Statistics for each base were calculated using the pysamstats utility available at <https://github.com/alimanfoo/pysamstats>. All reads for a given base were aligned to the reference sequence. Total reads were the sum of all base calls, insertions, and deletions at any given nucleotide position. Percent representation of each base was calculated as reads of a given base divided by total reads. Indel frequencies were quantified with a custom Matlab script as previously described^{3,24}.

Quantitative RT-PCR and quantitative PCR. HEK293T cells were transfected with base editor–P2a–GFP plasmids and incubated 3 d before harvesting DNA and RNA from each sample. DNA samples were harvested using the genomic DNA preparation protocol described above. RNA was isolated and amplified using the Cells-to-Ct (ThermoFisher) kit according to the manufacturer's protocol except the DNase treatment step used 2× DNase for twice as long to ensure complete degradation of plasmid DNA. Levels of mRNA were calculated by normalizing base editor mRNA levels to β -actin levels by $\Delta\Delta$ Ct. Plasmid DNA levels, as determined by qPCR of the BGH poly-adenylation sequence present on the base-editor plasmid, were normalized to β -actin levels to ensure that mRNA abundance was not skewed by transfection efficiency.

Western blotting. HEK293T cells were transfected with 750 ng of base editor–3× HA tag plasmid and 250 ng of guide RNA plasmid. After 3 d, cells were lysed using RIPA buffer with PMSF and cOmplete Protease Inhibitor Cocktail (Roche). Samples were boiled and quantified using a (bicinchoninic acid) BCA assay. 10 μ g of protein was loaded per well into a 12-well 4–12% Tris gel (Novex), dry-transferred to nitrocellulose paper for 7 min at 20 V before blocking and incubation with anti-HA (Cell Signaling Technology) and anti-Actin antibodies (Cell Signaling Technology) and visualized using an Odyssey imager. Uncropped blots are shown in **Supplementary Figure 16**.

APOBEC sequence collection. APOBEC protein sequences used in phylogenetic analyses were identified through searches of the UniProt database²⁵ with the BLASTP algorithm²⁶ using selected query sequences. All sequences from these searches that returned BLASTP E-values $< 10^{-7}$ were downloaded from UniProt. To reduce phylogenetic complexity, sequences were curated based on character length and pairwise sequence identity within each data set. The data set used for the construction of the non-redundant phylogeny was generated using four query sequences: UniProt IDs P41238, H2P4E7, E1BTD6, and H2P4E9. Multiple sequences were necessary to generate full coverage due to the low sequence identity across the family, which is $< 25\%$ between some members. Limits were chosen to remove truncated and partial sequences and those featuring large insertions or terminal extensions. Sequences greater than 97% identical, determined by pairwise alignment within the data set, were also removed. This level of identity provides a high level of detail within the tree while accelerating computational time by removing redundant taxa. The final data set contains 468 taxa (**Supplementary Sequences 1**).

Phylogeny construction. A multiple sequence alignment of the data set was generated with the program MAFFT using the FFT-NS-I x1000 algorithm²⁷ (**Supplementary Data 1**). Model selection used the Bayesian information criteria (BIC) to determine the evolutionary model that best fit the input alignment²⁸. 228 models were tested. The Jones Taylor Thornton (JTT) substitution matrix with empirical frequencies (F) and free rates with five categories (R5) was the model that best fit the data. A maximum likelihood (ML) phylogenetic tree was inferred with IQ-TREE²⁹ using the best fit model (JTT+F+R5). The starting trees were generated by randomized maximum parsimony and searched by fast hill-climbing Nearest Neighbor Interchange (NNI). Tree topology, branch lengths, and rate parameters were optimized. Branch supports were estimated with Ultrafast bootstrapping, implemented in IQ-TREE³⁰ (**Supplementary Fig. 17 and Supplementary Data 3**).

Ancestral sequence reconstruction. Sequences at internal nodes in the phylogeny were inferred using the *codeml* program from the PAML software package³¹. Posterior amino acid probabilities at each site were calculated using the JTT substitution matrix, given the ML tree and estimated background frequencies generated by IQ-TREE. N and C termini of ancestral sequences were modified manually to match those of rat APOBEC1.

Statistics and reproducibility. All statistical analyses were performed on $n = 3$ biologically independent experiments using unpaired Student's *t*-test. Biologically independent experiments reported here are from independent splits of each cell type used. Degrees of freedom = 4.

ClinVar analysis. Custom code provided in **Supplementary Note 1** was used to determine base pair changes required to correct pathogenic SNPs in the ClinVar database¹.

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

Data availability. Plasmids encoding BE4max, AncBE4max, and ABEmax have been deposited to Addgene. High-throughput sequencing data are deposited in the NCBI Sequence Read Archive ([SRP145378](https://www.ncbi.nlm.nih.gov/sra/SRP145378)).

20. Badran, A.H. *et al.* *Nature* **533**, 58–63 (2016).
21. Gibson, D.G. *et al.* *Nat. Methods* **6**, 343–345 (2009).
22. Kim, Y.B. *et al.* *Nat. Biotechnol.* **35**, 371–376 (2017).
23. Li, H. & Durbin, R. *Bioinformatics* **25**, 1754–1760 (2009).
24. Hu, J.H. *et al.* *Nature* **556**, 57–63 (2018).
25. UniProt Consortium *Nucleic Acids Res.* **46**, 2699 (2018).
26. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. *J. Mol. Biol.* **215**, 403–410 (1990).
27. Katoh, K. & Standley, D.M. *Mol. Biol. Evol.* **30**, 772–780 (2013).
28. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermini, L.S. *Nat. Methods* **14**, 587–589 (2017).
29. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. *Mol. Biol. Evol.* **32**, 268–274 (2015).
30. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. & Vinh, L.S. *Mol. Biol. Evol.* **35**, 518–522 (2018).
31. Yang, Z. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

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

A custom python script, provided in the SI, was used to assess SNPs targetable by base editors on the ClinVar database.

Data analysis

The burrows-wheeler aligner (bwa-mem) and pysamstats are two publically available data analysis tools for HTS analysis. Ancestral reconstruction used: IQ-TREE v 1.6.1, PAML – v 1.8, and MAFFT – Ubuntu version 7.394. Prism 7 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/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.

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

Plasmids encoding BE4max, AncBE4max, and ABE4max have been deposited to Addgene. High-throughput sequencing data are deposited in the NCBI Sequence Read

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](https://www.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	Experiments were performed in biological triplicate n=3 unless otherwise noted. In previous studies using related experiments we determined this sample size to be sufficient to ensure reproducibility. No statistical tests were used to determine sample size.
Data exclusions	No data was excluded.
Replication	All attempts at replication were successful, and standard deviations were within expected ranges.
Randomization	Different cell passages were used for each biological replicate.
Blinding	Not applicable, as samples were processed identically through standard and in some cases automated procedures (DNA sequencing, transfection, DNA isolation) that should not bias outcomes.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	Rabbit anti-HA antibody from Cell Signaling Technology, product # 3724S, Lot 8, C2954 used at 1:1000 dilution; Rabbit anti-B-actin from Cell Signaling, product # 4970S, Lot 14, 13E5 used at 1:200 dilution.
Validation	Validation was performed by supplier.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	HEK293T (ATCC), N2A (ATCC), Fibroblast (Coriell, GM20958)
Authentication	Cells were authenticated by the supplier.
Mycoplasma contamination	HEK293T cells tested negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

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

Cells were grown in tissue culture, transfected or nucleofected as described in the text and grown for 3 days. Cells were then trypsinized and filtered to remove debris before sorting.

Instrument

FACSAria II

Software

BD FACS DIVA software was used for analysis.

Cell population abundance

The abundance of cells in the post-sort fraction was dependent upon cell type and sorting condition. HEK293T mCherry+ in trans cells were typically over 50% of the population, GFP/mCherry dual+ cells were also typically over 50% of the population, BE4-P2A-GFP GFP/mCherry dual+ cells were typically over 20% of the population, BE4max-P2A-GFP and ancBE4max-P2A-GFP GFP/mCherry dual+ cells were typically over 40% of the population. N2A In trans mCherry+ cells were typically over 60% of the population. N2A GFP+ cells for BE4max-P2A-GFP and ancBE4-P2A-GFP were typically over 30% of the population. Fibroblast in trans mCherry+ cells were typically over 20% of the population. Fibroblast GFP+ cells for BE4max-P2A-GFP and ancBE4max-P2A-GFP were typically over 10% of the population.

Gating strategy

Negative control (unstained) and fluorophore-positive cells were used to establish gates for each cell type. Gates were drawn to collect cells expressing either fluorophore. See the provided examples for gates used.

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