APPLIED SCIENCES AND ENGINEERING

Single C-to-T substitution using engineered APOBEC3G-nCas9 base editors with minimum genome- and transcriptome-wide off-target effects

Sangsin Lee[1]*, Ning Ding[2]*, Yidi Sun[3,4], Tanglong Yuan[5], Jing Li[1], Qichen Yuan[2], Lizhong Liu[6], Jie Yang[2], Qian Wang[7], Anatoly B. Kolomeisky[2,7,8], Isaac B. Hilton[1,6], Erwei Zuo[5]†, Xue Gao[1,2]†

Cytosine base editors (CBEs) enable efficient cytidine-to-thymidine (C-to-T) substitutions at targeted loci without double-stranded breaks. However, current CBEs edit all Cs within their activity windows, generating undesired bystander mutations. In the most challenging circumstance, when a bystander C is adjacent to the targeted C, existing base editors fail to discriminate them and edit both Cs. To improve the precision of CBE, we identified and engineered the human APOBEC3G (A3G) deaminase; when fused to the Cas9 nickase, the resulting A3G-BEs exhibit selective editing of the second C in the 5′-CC-3′ motif in human cells. Our A3G-BEs could install a single disease-associated C-to-T substitution with high precision. The percentage of perfectly modified alleles is more than 6000-fold for disease correction and more than 600-fold for disease modeling compared with BE4max. On the basis of the two-cell embryo injection method and RNA sequencing analysis, our A3G-BEs showed minimum genome- and transcriptome-wide off-target effects, achieving high targeting fidelity.

INTRODUCTION

Fusing a deaminase with the Cas9 nickase (nCas9) forms cytosine base editors (CBEs), which enable programmable conversion of cytidine-to-thymidine (C-to-T) mutations within a specific region of the genomic DNA without causing double-stranded breaks (*1*–*3*). CBEs have displayed substantially higher editing efficiency than the conventional Cas9 endonuclease-mediated homology-directed repair method for installing point mutations (*4*, *5*). In addition, recent protein engineering efforts have improved their product purities and efficiencies (*6*, *7*), greatly expanded the genome targeting scope (*8*), and minimized the undesirable RNA off-target effects (*9*–*11*). CBEs are important genetic tools and could potentially correct more than 5000 pathogenic single-nucleotide polymorphisms (SNPs) associated with human-inherited diseases caused by T-to-C (or G-to-A) mutations (*3*, *12*, *13*).

The presence of multiple targets within the CBEs' activity window [e.g., the editing window of BE4max is approximately from positions 4 to 8 of the protospacer, counting the protospacer adjacent motif (PAM) as positions 21 to 23] can introduce unwanted bystander editing, resulting in deleterious multi–C-to-T conversions (*14*). Earlier studies have shown that the activity window size can be narrowed using strategies such as modulating the catalytic activity of deaminase (*15*), using more rigid linkers between Cas9 and deaminase, or deleting nonessential deaminase sequences (*16*, *17*). These approaches can systematically enhance precision for position-dependent single-nucleotide editing irrespective of nearby sequence contexts, although the genome targeting scope might be compromised because of the requirement that the target nucleotide needs to be placed at a specific position relative to an available PAM. Alternatively, sequence context-specific CBE can avoid bystander editing without sacrificing the activity window size (*3*). The engineered APOBEC3A (A3A) enzyme preferentially deaminates in the TCR motif (target C underlined), which has been exploited for more precise base editing, and the resulting eA3A-BE3 base editor exhibited high on-target precision with minimized bystander editing (*18*). However, in the most challenging case, when editable Cs are located consecutively within the activity window, especially in the case of CC dinucleotides when a bystander C is located right upstream of the target C, the existing CBEs non-selectively edit both of the Cs. Nearly 38% of the human pathogenic SNPs that are caused by T-to-C disease point mutations lie in the context of CC, followed by AC (29%), GC (21%), and TC (13%) (see data file S1) (*1*, *12*), necessitating the development of new CBEs that can precisely discriminate between the target and bystander Cs.

Various APOBEC enzymes in vertebrates mediate defense against infections from retroviruses or retrotransposons by deaminating C to U in the viral complementary DNA (cDNA) (*19*, *20*), suggesting that these cytosine deaminases could have unique preferences for particular sequence motifs to distinguish DNA sequences from the native host (*21*–*23*). In this study, we identified human APOBEC3G (A3G) as a candidate for developing sequence-specific BEs in multiple C contexts. We characterized and engineered A3G-BE variants to efficiently edit a single C at various endogenous sites in human embryonic kidney–293T (HEK293T) cells. By introducing mutations that improve catalytic activity, solubility, and overall protein scaffold, we obtained and characterized three novel variants (A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14) that exhibit high editing efficiencies and precision in the context of the CC motif. A3G-BE variants have broader activity windows than BE4max that could expand the

[1]Department of Bioengineering, Rice University, Houston, TX 77030, USA. [2]Department of Chemical and Biomolecular Engineering, Rice University, Houston, TX 77005, USA. [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, University of Chinese Academy of Sciences, Shanghai 200031, 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 200031, China. [5]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China. [6]Department of Biosciences, Rice University, Houston, TX 77005, USA. [7]Center for Theoretical and Biological Physics, Rice University, Houston, TX 77005, USA. [8]Department of Chemistry, Rice University, Houston, TX 77005, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: xue.gao@rice.edu (X.G.); erweizuo@163.com (E.Z.)

targeting scope for precision base editing. We also demonstrated that these variants could efficiently and precisely generate or correct mutated alleles associated with the known pathogenic phenotypes, illuminating A3G-BEs' potential application in treating human genetic diseases. Last, we performed whole-genome sequencing (WGS) by using the most recently developed genome-wide off-target analysis by two-cell embryo injection (GOTI) method to detect DNA off-targets and used RNA sequencing (RNA-seq) to examine the RNA off-targets of cells treated with A3G-BE5.13. Our results showed that the most active A3G-BE5.13 induces baseline levels of the genome- and transcriptome-wide off-target mutations, suggesting high editing fidelity for future clinical applications.

RESULTS

Identification of a sequence-specific A3G deaminase

Previous studies have demonstrated that A3G predominantly deaminates the third C in the 5′-CCC-3′ motif of a single-stranded DNA (ssDNA) substrate (*24*). To test whether this motif preference could be preserved when A3G is fused to nCas9 as A3G-BE, we replaced the rAPOBEC1 deaminase domain of BE4max with the full-length, human codon-optimized A3G to construct A3G-BE2.1 (*6*). Since it has been reported that the N-terminal domain (NTD) could mediate aggregation of A3G monomers to impede A3G's mobility (*25*) and because the C-terminal domain (CTD) of A3G is sufficient for deamination activity in vitro (*26*, *27*), we therefore truncated the NTD of A3G to construct A3G-BE4.4, which only contains the CTD of A3G (Fig. 1A). HEK293T cells were then transfected with plasmids expressing BE4max, A3G-BE2.1, and A3G-BE4.4 with single-guide RNAs (sgRNAs) targeting *EMX1* #1 and *FANCF* #a3 sites, which contain dinucleotide Cs (C5 and C6 of *EMX1* #1 and C6 and C7 of *FANCF* #a3) within the canonical BE4max activity window. We extracted the genomic DNA after 72 hours and amplified the target regions for high-throughput sequencing (HTS). Analysis of the C-to-T editing efficiencies of the dinucleotide Cs showed that A3G-BE2.1 and A3G-BE4.4 edited 21 to 42% of the cognate Cs (C6 of *EMX1* #1 and C7 of *FANCF* #a3) but only 1 to 3% of the bystander Cs (C5 of *EMX1* #1 and C6 of *FANCF* #a3), while BE4max edited 47 to 62% of both the cognate and bystander Cs without obvious selectivity (Fig. 1B). No significant difference was observed between A3G-BE2.1 and A3G-BE4.4 for editing efficiencies of the cognate Cs, suggesting that the CTD itself adequately determines the enzymatic activity and sequence specificity of A3G.

Because the wild-type A3G in nature preferentially deaminates in the $C_{-2}C_{-1}C0A_{+1}$ sequences of the HIV-1 genome (*28*), we next examined whether nucleotides at positions −2 and +1 of the cognate C_0 also affect the base editing efficiency and specificity. We tested BE4max and A3G-BE4.4 at nine different loci containing the dinucleotide Cs motif with different combinations of nucleotides placed at the −2 and +1 positions ($N_{-2}C_{-1}C_0D_{+1}$, where D denotes A, T, and G) (Fig. 1C). HTS analysis confirmed that A3G-BE4.4 showed selective editing of the cognate Cs across all the sites. At six of the nine sites, A3G-BE4.4 reached at least 79% of the editing efficiencies of the cognate Cs of those of BE4max (Fig. 1D). Notably, at *DMD* #1 site, which contains the ACCA motif, similar to the native CCCA, and harbors the cognate C9 outside the canonical BE4max activity window, A3G-BE4.4 induced threefold higher editing of the cognate C9 compared to BE4max. However, although being selective, A3G-BE4.4 displayed very low cognate C editing efficiencies with only 13, 6, and 3% C-to-T conversion rates at the remaining three *PPP1R12C* #a3, *BCS1L* #1, and *EMX1* #a18 sites, respectively. These results may have occurred because the wild-type A3G disfavors deamination of certain motifs such as GCC, suggesting that the motif-dependent deamination activity of A3G could influence the efficiency of the selective base editing (*29*). We then quantified the specificity by dividing the editing efficiency of the cognate C by that of bystander C (cognate-to-bystander editing ratio). Across the nine sites, A3G-BE4.4 recorded the editing ratios ranging from 11 to 290, while BE4max achieved a maximum ratio of 6 at *EMX1* #a18 and less than 2 at all other sites (fig. S1A). Non-T by-products generated by A3G-BE4.4 averaged slightly higher than BE4max in most of the sites (fig. S1B), consistent with previous observations that generally lower product purity is generated by editing of a single C versus multiple Cs (*6*). A3G-BE4.4 also showed significantly fewer indels than BE4max at three of the nine sites (*HEK3* #1, *HEK4* #a1, and *EMX1* #a3), supporting an earlier study suggesting that single-nucleotide and multiple base editing have no significant correlation in terms of indel generation (fig. S1C) (*18*). Together, these results indicated that A3G-BE4.4 has sufficient editing efficiency to precisely edit the second C in the sequence context of 5′-CC-3′ dinucleotides.

Improvement of A3G-BE's editing efficiency

Given the relatively low base editing efficiencies of A3G-BE4.4 for cognate Cs observed from the *PPP1R12C* #a3, *BCS1L* #1, and *EMX1* #a18 sites, we envisioned that the wild-type A3G-CTD activity could be further improved. We devised three subsets of mutations that could be introduced into the A3G-CTD of A3G-BE4.4 based on different possible functional effects, including set A (P200A + N236A + P247K + Q318K + Q322K) to improve catalytic activity, set B (partial replacement of A3G's loop 3 with A3A's, that is H248N + K249L + H250L + G251C + F252G + L253F + E254Y) to increase ssDNA binding affinity, and set C (L234K + C243A + F310K + C321A + C356A) to enhance protein solubility (Fig. 2A and fig. S2A) (*27*, *30*, *31*). We first introduced set A to A3G-BE4.4 to construct A3G-BE5.1 and introduced sets B and C mutations to A3G-BE5.1 to construct A3G-BE5.3 and 5.4, respectively (fig. S2B and table S1). To further maximize A3G's potential deamination activity, two additional mutations, T311A + R320L, were introduced to A3G-BE5.3 to construct A3G-BE5.10 (fig. S2B and table S1) (*27*, *31*). We tested A3G-BE4.4, A3G-BE5.1, A3G-BE5.3, A3G-BE5.4, and A3G-BE5.10 at *EMX1* #1 and *FANCF* #a3; all of the further improved mutants showed substantially higher editing efficiency than A3G-BE4.4 did on both the cognate Cs and the bystander Cs (Fig. 2B and fig. S2C). Notably, when the loop 3 of A3G was partially replaced with A3A's by set B mutations, A3G-BE5.3 and A3G-BE5.10 exhibited substantial loss of the motif preference, and both Cs were efficiently edited. Structural alignment of the wild-type A3A, wild-type A3G, and the A3G containing the set A mutations, among which P247K lies in loop 3, showed that loop 3 of the wild-type A3A, as well as the A3G with set A mutations, exhibits greater proximity to the ssDNA substrate, suggesting that the observed increase in the editing efficiency and relaxation of the sequence specificity might be partly due to the stronger nonspecific binding to the ssDNA substrate (fig. S2D).

We hypothesized that modulating the nonspecific binding to DNA could restore the sequence specificity. Using structure-guided analysis, Tyr^{315} of A3G was identified as a key residue that interacts with both the DNA backbone and the target C (Fig. 2C). We speculated that changing Tyr^{315} to Phe, which lacks only the hydroxyl group

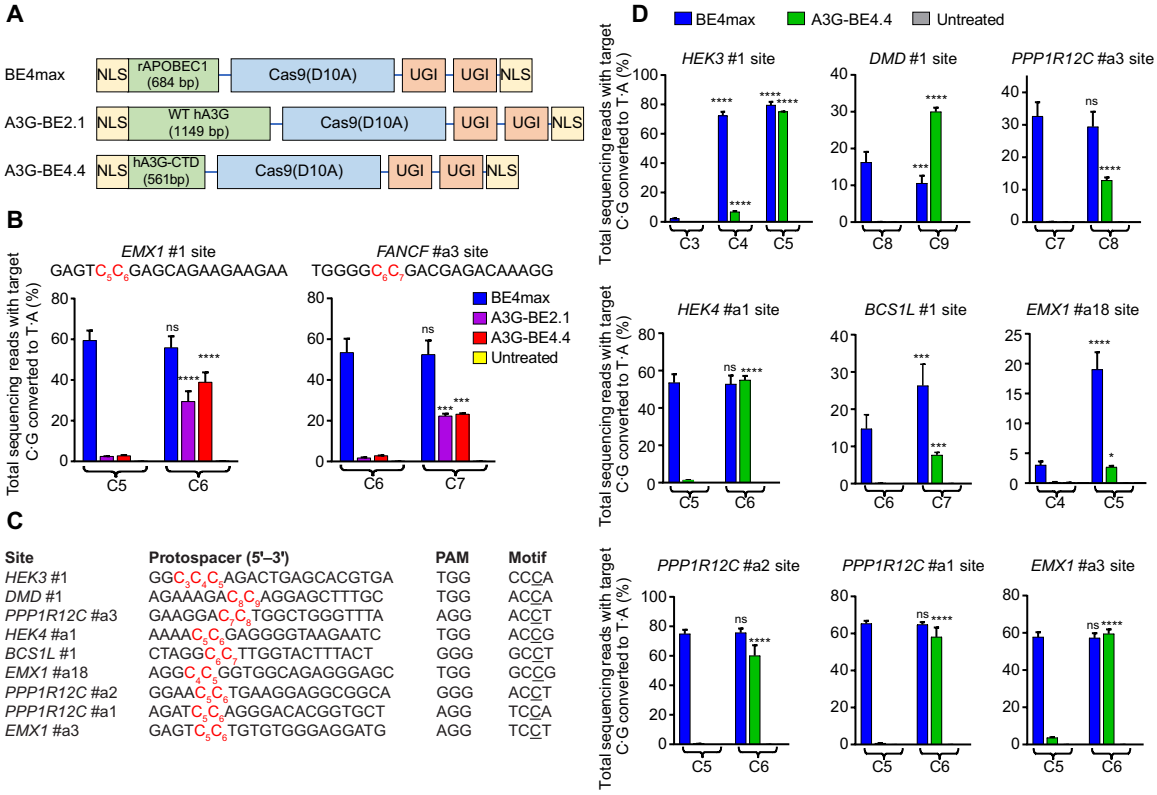

Fig. 1. Base editing specificity of the wild-type A3G-nCas9 fusions. (**A**) Schematic showing the protein architecture of base editors. BE4max is used to replace the rAPOBEC1 with either full-length (NTD + CTD) or CTD-only human A3G to construct A3G-BE2.1 or A3G-BE4.4, respectively. Linkers between functional domains are shown as horizontal blue lines. NLS, nuclear localization signal; UGI, uracil glycosylase inhibitor. (**B**) C-to-T editing efficiency and specificity of A3G-BE2.1 and A3G-BE4.4 at *EMX1* #1 and *FANCF* #a3 sites bearing the CC motif (red). (**C**) Nine endogenous sites of HEK293T bearing either CC or CCC motif (red) within the canonical BE4max activity window. Each PAM and the sequence motif identifying the nucleotides at +1 and −2 positions from the target C (underlined) are shown. (**D**) C-to-T editing efficiency and specificity of BE4max and A3G-BE4.4 at the endogenous sites listed in (C). Bar figures of (B) and (D) show means and error bars representing SD of $n = 2$ and $n = 3$ independent biological replicates performed on different days, respectively. Statistical significance shown on top of each bar using two-tailed Student's *t* test compares to editing efficiency of the preceding bystander C of the same BE. For example, *t* test was performed between the BE4max editing efficiencies of C8 and C9 at *DMD* #1 site. ns (not significant), $^{*}P < 0.05$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

from Tyr, could remove the hydrogen bond with the 5′ phosphate group of ssDNA while maintaining the π-π interaction with the target C. We introduced Y315F to A3G-BE5.1, A3G-BE5.3, A3G-BE5.4, and A3G-BE5.10 to construct A3G-BE5.12, A3G-BE5.13, A3G-BE5.14, and A3G-BE6.11, respectively (fig. S2B and table S1). Y315W (to provide steric hindrance) and Y315L (to remove both the hydrogen bond and the π-π interaction) were also introduced into A3G-BE5.10, resulting in A3G-BE6.16 and A3G-BE6.17, respectively. Additional mutations to further reduce the nonspecific binding, including N244Q, S286A, and R313A, were introduced into A3G-BE6.11 to construct A3G-BE6.18, A3G-BE6.19, and A3G-BE6.20, respectively. Last, we reverted the replacement of the A3G's loop 3 with A3A's from A3G-BE6.11 to construct A3G-BE6.21 (fig. S2B and table S1). Testing all the above variants at *EMX1* #1 and *FANCF* #a3 showed that A3G-BE6.11 induced higher selectivity than A3G-BE5.10 by moderately reducing editing of the bystander Cs. At the same time, A3G-BE6.16 and A3G-BE6.17 displayed markedly reduced editing efficiencies of the cognate Cs, even below those of A3G-BE4.4 (Fig. 2B and fig. S2C). Although all A3G-BE6.18, A3G-BE6.19, A3G-BE6.20, and A3G-BE6.21 showed improved editing ratios of the cognate to bystander Cs compared with A3G-BE6.11, their cognate C editing efficiencies did not outperform A3G-BE4.4. Nevertheless, A3G-BE5.13 and A3G-BE5.14, both of which contain Y315F, exhibited greater cognate C editing efficiency than A3G-BE4.4 did and demonstrated appreciable restoration of the sequence specificity (Fig. 2B and fig. S2C).

We further tested A3G-BE5.13 and A3G-BE5.14 at the *PPP1R12C* #a3, *BCS1L* #1, and *EMX1* #a18 sites at which the editing efficiencies of A3G-BE4.4 were previously low (Fig. 1D). HTS analysis showed that both A3G-BE5.13 and A3G-BE5.14 gained superior editing efficiency for the cognate Cs as compared to A3G-BE4.4 (Fig. 2D). Moreover, bystander editing of A3G-BE5.13 and A3G-BE5.14 remained substantially lower than that of BE4max, resulting in significant improvement of base editing efficiency while maintaining the specificity. Together, these results suggested that through rational engineering, A3G-BE5.13 and A3G-BE5.14 overcame the low editing drawbacks of A3G-BE4.4 on discrete sequence contexts.

Broad targeting scope of A3G-BEs

To comprehensively understand the capability of sequence-specific base editing of A3G-BE5.13 and A3G-BE5.14, we tested them at eight other endogenous sites with the dinucleotide Cs motif positioned across the whole protospacer. HTS analysis confirmed that all A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14 selectively edited the second

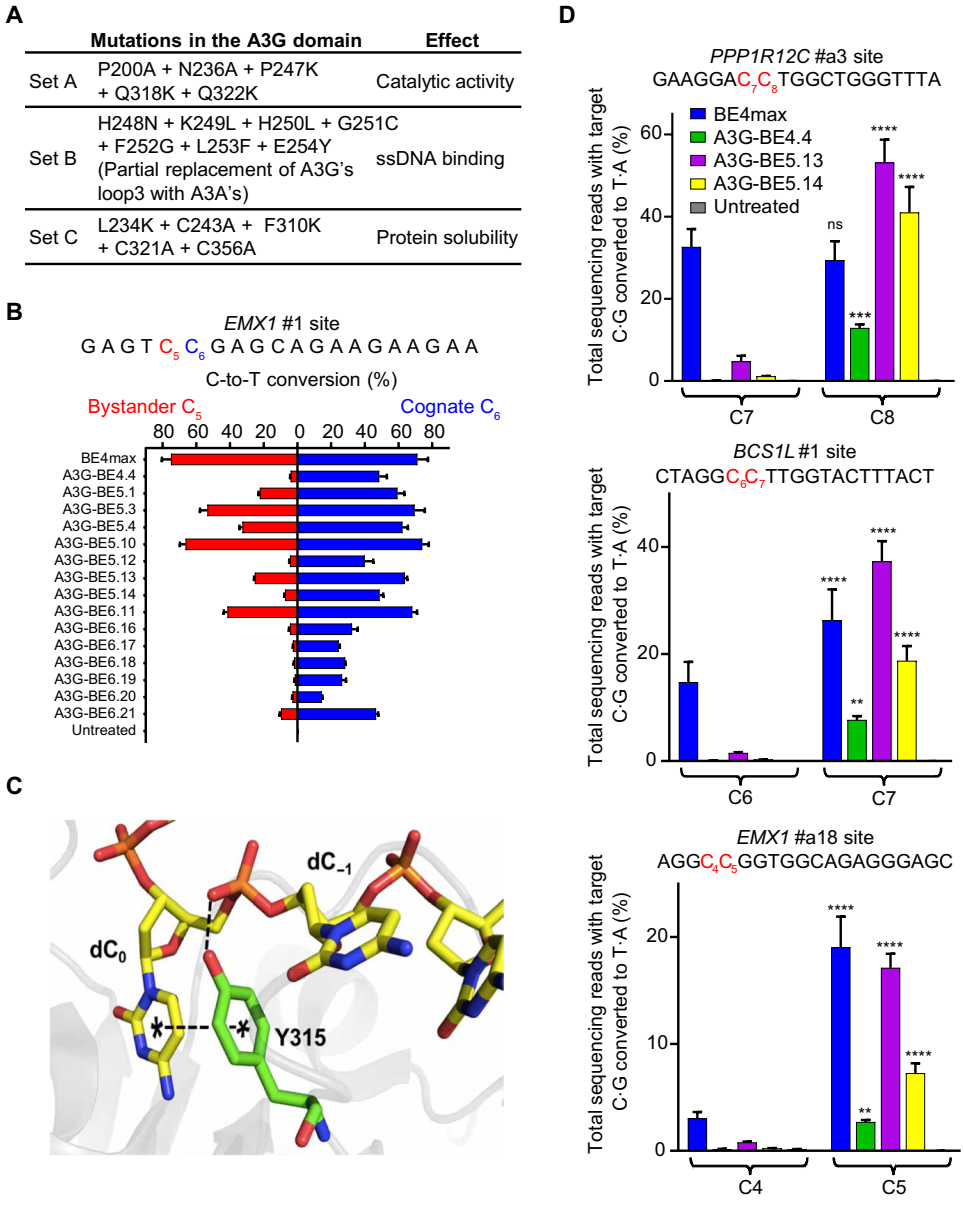

Fig. 2. Engineering A3G-BEs with enhanced base editing efficiency. (**A**) Set of residue mutations of A3G for improving catalytic activity (set A), ssDNA binding (set B), and protein solubility (set C) listed on each row. Counting of the residue number starts with the first residue of the original full-length A3G. (**B**) Screening of A3G-BE mutants at *EMX1* #1 site to determine variants with enhanced editing efficiency and retained sequence specificity. C-to-T editing efficiencies are represented as bidirectional bars with values for the cognate C6 (blue) on the right and the bystander C5 (red) on the left. (**C**) An enlarged view of the interactions of Tyr^{315} (green sticks) with the ssDNA substrate (yellow sticks). The hydrogen bond between the 5′ phosphate group of the DNA backbone and the hydroxyl group of Tyr^{315}, and the π–π interaction between the rings of the target cytidine (dC_0) and Tyr^{315} are represented as dashed lines. (**D**) C-to-T editing efficiency and specificity of A3G-BE5.13 and A3G-BE5.14 at three endogenous sites previously poorly edited by A3G-BE4.4. Panels (B) and (D) show means and error bars representing SD of $n = 3$ independent biological replicates performed on different days. For (D), statistical significance shown on top of each bar using two-tailed Student's *t* test compares to editing efficiency of the preceding bystander C of the same BE. ns (not significant), $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

C within the CC motifs across all the sites. The cognate-to-bystander editing ratios were calculated to be up to 186 (A3G-BE5.14 at *EMX1* #c16 site), while BE4max either nonselectively edited both Cs or failed to perform outside its canonical activity window (Fig. 3A and fig. S3A). At *BCS1L* #6 and *RNF2* #2 sites, which contained the cognate Cs at positions 12 and 15 of the protospacers, respectively, highly efficient and selective editing for the cognate Cs were only observed when using A3G-BE5.13 and A3G-BE5.14, while A3G-BE4.4 and BE4max did not yield efficient C-to-T editing (Fig. 3A). Notably, at both *BCS1L* #6 and *RNF2* #2 sites, the single C located at the fifth position was not efficiently edited by all A3G-BE variants, probably due to lack of the CC dinucleotide sequence context. Both A3G-BE5.13 and A3G-BE5.14 displayed efficient editing up to C15 of *RNF2* #2 but not C18 of *FANCF* #2 (Fig. 3A). For the two cognates Cs existing in *EMX1* #b1 (C7 and C15) and *FANCF* #2 (C6 and C10) sites, A3G-BE4.4 efficiently edited only the ones residing closer to the 5′ end

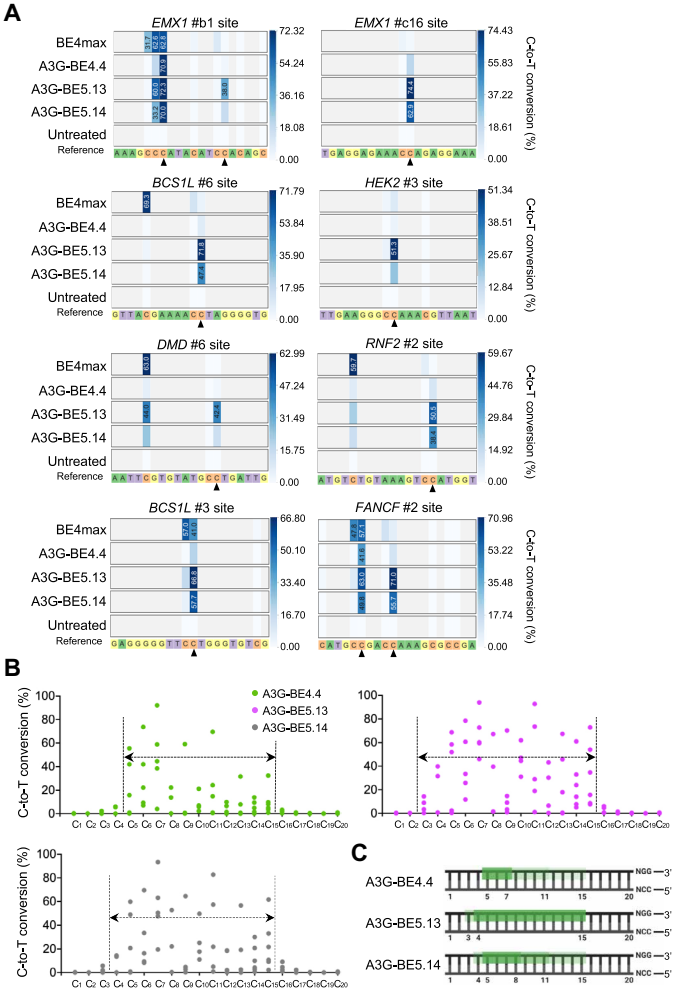

Fig. 3. Base editing specificity of A3G-BEs across the protospacer regions and the activity window size. (**A**) Heat maps are showing average C-to-T editing efficiencies of $n = 3$ independent biological replicates of BE4max, A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14 at eight endogenous sites containing the preferential C<u>C</u> or CC<u>C</u> motif across the whole region within the protospacers. The cognate Cs predicted to be preferentially editable by A3G-BEs are indicated by the black triangles. (**B**) Average C-to-T base editing frequencies at each protospacer position from the six poly-C endogenous sites shown in fig. S4. Bidirectional arrows in between vertical dashed lines show the base-editable ranges within the protospacer region by the indicated A3G-BEs (**C**) Schematic representation of the activity window sizes of A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14, with NGG PAM shown as positions 21 to 23. Standard, light, and near-transparent green represent the predicted relative base editing activity within the approximate regions of the protospacer.

(C7 of *EMX1* #b1 and C6 of *FANCF* #2), indicating a possible narrower window size compared with A3G-BE5.13 and A3G-BE5.14. The lowest cognate-to-bystander editing ratios for all three A3G-BEs occurred at *EMX1* #b1, which bears three consecutive Cs of the CC<u>C</u>A motif, suggesting that the requirement for single-nucleotide editing within more than two consecutive Cs might need to be more stringent. We did not find a consistent trend in the product purity following the treatment of all BEs, which might be due to the discrepancies among distinct properties of BEs that have different activity windows, deamination activities, and sequence specificities (fig. S3B) (*6*). We also observed indels being generated with varying frequencies across the sites without apparent correlation among BEs (fig. S3C).

To determine the sizes of the activity window of A3G-BEs, we tested A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14 at six endogenous genomic sites, which contain consecutive Cs within the protospacer, and analyzed their C-to-T editing efficiencies. For all the tested sites, A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14 revealed consistent and broad base editing activity window but differed mainly in their relative editing efficiencies, for which A3G-BE5.13 showed the highest followed by A3G-BE5.14 and A3G-BE4.4 (fig. S4). We observed that A3G-BE4.4 displayed comparatively lower editing efficiencies around positions 8 to 15 compared with those in positions 5 to 7 at four sites (*VEGF* #2, *EMX1* PolyC #1, *EMX1* PolyC #1, and *HEK4* PolyC #1), suggesting that editing toward the 3′ end of the protospacer, although targetable, could have lower editing efficiency. Next, we compared the average editing frequencies of Cs at each protospacer position from all the six sites. We found that the activity windows of A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14 span from positions 5 to 15, 3 to 15, and 4 to 15 of the protospacer, respectively (Fig. 3, B and C). Together, these data indicated that A3G-BEs enable sequence-specific editing with broadened targeting ranges.

Given that the preferential motif of A3G extends to three consecutive Cs, $C_{-2}C_{-1}\underline{C}_0$, we hypothesized to test whether the sequence specificity could be maintained when the middle C, the −1 position of the target, is altered to other nucleotides. To assess this possibility, we selected five endogenous sites that contained a T or A at the −1 position ($C_{-2}TC_0$ or $C_{-2}AC_0$ motifs) and, now, counting editing of the C at −2 position to be the bystander incidence (fig. S5A). We transfected HEK293T with BE4max, A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14 with sgRNAs targeted to the selected sites and performed HTS. After quantifying the C-to-T editing efficiencies, we found that, compared to BE4max, A3G-BEs indeed displayed significantly higher editing of the cognate Cs over bystander Cs within these altered sequence contexts (fig. S5B). A3G-BE5.14, among other A3G-BEs, exhibited the highest specificities (up to 89 cognate-to-bystander editing ratio) at four of the five sites (fig. S5B). While A3G-BE5.13 and A3G-BE5.14 have comparable or higher cognate C editing efficiency than BE4max, A3G-BE4.4 editing efficiencies of the cognate Cs were below 9% at four of the five sites, indicating that the absence of C at the −1 position might restrain A3G-BE4.4 from efficient editing. In addition, we observed relatively higher bystander C_{-2} editing from A3G-BE5.13 at *HEK*3 #b1 and *HEK*3 #b2 sites, which contained T immediately upstream of the bystander C_{-2}. Since C and T are structurally similar compared to the other two nucleotides, we speculated that this sequence context might be more prone to bystander editing. These findings indicated that A3G-BEs could selectively edit a target C in the CT<u>C</u> and CA<u>C</u> motifs and therefore can further expand the targeting scope for precision base editing in broader sequence contexts.

Disease modeling and correction

To test A3G-BEs in disease-relevant contexts, we sought to precisely generate SNPs of reported human pathogenic diseases (*32*). Three genetic variants caused by C-to-T (or G-to-A) substitution in which the wild-type sequences lie within the preferential 5′-CC-3′ motif of A3G-BEs were selected, including cystic fibrosis (model 1), hypertonic myopathy (model 2), and transthyretin amyloidosis (model 3) (Fig. 4A). Individual sgRNAs targeted to these disease-associated sites were constructed and cotransfected into HEK293T with BE4max,

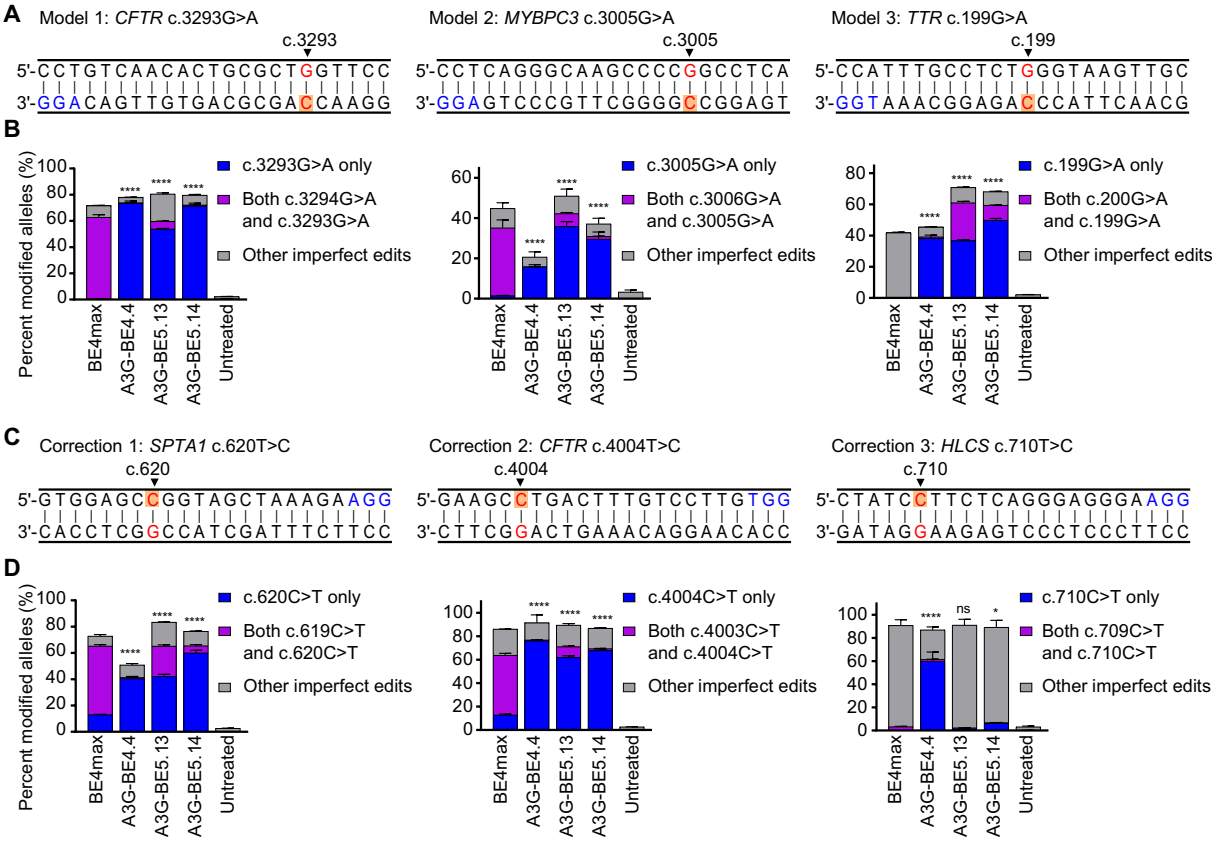

Fig. 4. Modeling and correcting human pathogenic SNPs in vitro using A3G-BEs. (**A**) Sequences of the protospacers and PAMs (blue) for model 1 (cystic fibrosis), model 2 (hypertonic myopathy), and model 3 (transthyretin amyloidosis). Position of the disease-relevant C>T (or G>A) point mutations are red and indicated by black triangles shown with the nucleotide numbers within the disease-associated genes. (**B**) Percent of alleles modified to the indicated genotypes following the treatment of BE4max and A3G-BEs for generating the three models presented in (A). (**C**) Sequences of the protospacers and PAMs (blue) for correction 1 (hereditary pyropoikilocytosis), correction 2 (cystic fibrosis), and correction 3 (holocarboxylase synthetase deficiency), bearing T>C (or A>G) point mutations for which the positions are indicated with black triangles showing the nucleotide numbers within the disease-associated genes. (**D**) Percent of alleles modified to the indicated genotypes following the treatment of BE4max and A3G-BEs for correcting the three disease-associated variants presented in (C). Panels (B) and (D) show means and error bars representing SD of $n = 3$ independent biological replicates performed on different days. Statistical significance shown on top of each bar using two-tailed Student's *t* test compares to the percentages of perfectly generated/corrected alleles by BE4max. ns (not significant), $^{*}P < 0.05$, $^{****}P < 0.0001$.

A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14. Genomic DNA was harvested after 72 hours and prepared for HTS to quantify the percentage of alleles perfectly modeled and of those that were imperfectly modified because of bystander editing. Direct comparison with BE4max of the modified allele frequencies demonstrated that A3G-BEs induced a substantially higher proportion of perfectly modified alleles for all three models (Fig. 4B). Despite the previous observations in which A3G-BE5.13 displayed more relaxed base-editing sequence specificity among other selected A3G-BEs, it achieved the highest percentage here of the perfectly modified alleles for hypertonic myopathy (model 2) (36%). For transthyretin amyloidosis (model 3), in which the target C lies at position 11 of the protospacer, all A3G-BEs produced the desired allele with high efficiencies (>35%), while BE4max failed to edit the target C (<0.1%) because of its inability to edit outside its activity window (fig. S6A). As a result, A3G-BE5.14 accomplished 613-fold higher correct modeling of transthyretin amyloidosis than BE4max did, highlighting the advantage of precise editing with an expanded activity window. Similarly, for cystic fibrosis (model 1), all A3G-BEs induced more than 50% of the perfectly modified alleles, while BE4max averaged 0.6%.

Next, to examine the therapeutic applicability of A3G-BEs, we selected three reported human pathogenic SNPs caused by T>C (or A>G) mutations, which can be preferentially targeted by A3G-BEs, including hereditary pyropoikilocytosis (correction 1), cystic fibrosis (correction 2), and holocarboxylase synthetase deficiency (correction 3) (Fig. 4C) (*32*). We generated three HEK293T lines containing 200 base pair (bp) of each disease-relevant sequence integrated into the genome (see Materials and Methods). Codelivery of the BEs and sgRNAs targeted to the disease-associated sites and analysis of the HTS data to quantify the perfectly corrected alleles verified that all A3G-BEs significantly outperformed BE4max by a minimum of threefold in corrections 1 and 2. In addition, A3G-BE4.4 exclusively induced more than 50% of perfectly corrected alleles among other BEs and accomplished 6496-fold higher correction than BE4max in correction 3 (Fig. 4D). Correction 3, in which the protospacer contained two motifs preferred by A3G-BEs, C<u>C</u> and CT<u>C</u>, interfered with the precise single C-to-T editing by A3G-BE5.13 and A3G-BE5.14 and resulted in substantial dual C editing due to their wide activity window sizes and high efficiencies (fig. S6A). Collectively, these comparisons indicated that A3G-BEs have higher

targeting precision than BE4max for reversing pathogenic SNPs within their preferred sequence contexts.

We further investigated the editing efficiency of A3G-BEs in therapeutically more relevant cell types, including the induced pluripotent stem cells (iPSCs) and human embryonic stem cells (hESCs). We nucleofected iPSC and ESI-017 hESC lines with BE4max, A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14 with sgRNA targeting the hypertonic myopathy (model 2)–associated site and performed clonal expansion of the successfully nucleofected cells for 10 to 14 days before analysis. In the iPSCs, analysis of the sequencing chromatograms revealed that A3G-BEs more efficiently edited the cognate C7 than the bystander C6, which were 10, 46, and 34% at C7 and 2, 15, and 5% at C6 by A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14, respectively. In contrast, BE4max nonselectively edited both Cs, 39 and 50% at C7 and C6, respectively (fig. S6B). The observed trend was consistent with the ESI-017 hESCs (fig. S6C), indicating the utility of A3G-BEs to serve as important tools to precisely model genetic variants in clinically relevant cell types.

DNA and RNA off-target determination of A3G-BEs

Several CBEs were reported to generate genome- and transcriptome-wide off-target editing, which became a major concern for their clinical uses (*9*, *10*, *33*, *34*). We then examined the propensity of A3G-BEs to cause deamination on off-target loci by performing orthogonal R-loop assay (*35*). Briefly, the nuclease-dead SaCas9 (dSaCas9) sgRNA complex creates an R-loop, recapitulation of a stochastic ssDNA exposure in the genome, at a DNA locus unassociated with the on-target site. Base editing mediated by cytosine deaminase in the off-target R-loop independently of SpCas9 nickase and its sgRNA is detected via targeted HTS (fig. S7A). We assessed six off-target loci (Sa #1 to #6 sites) by cotransfecting SpCas9-derived CBE (BE4max or A3G-BEs), on-target SpCas9 sgRNA, dSaCas9, and off-target dSaCas9 sgRNA into HEK293T (table S2). For the on-target editing at *EMX1* #1 site, specificities and efficiencies of all CBEs exhibited consistent results with our previous observations without the dSaCas9 system (fig. S7B). We then quantified the editing activities of 18 cytosines within those six off-target loci. We found that A3G-BEs show substantially reduced off-target editing compared with BE4max, except at those cytosines lying within the 5′-C<u>C</u>-3′ motif, e.g., C10 and C15 at Sa #2, C11 at Sa #5, and C8 at Sa #6 sites (fig. S7C). A3G-BE4.4 showed no significant off-target editing at 10 of the 18 cytosines. A3G-BE5.13 induced higher off-target mutations than both A3G-BE4.4 and A3G-BE5.14 at all cytosines but still significantly lower than BE4max at 11 of the 18 cytosines. Together, these results suggested that A3G-BEs generally exhibit lower propensities to cause Cas9/sgRNA-independent off-target mutations. We then selected A3G-BE5.13, the most active variant among the three selected ones, for further whole-genome off-target characterization.

To comprehensively understand the capability of A3G-BE5.13 to generate Cas9/sgRNA-independent DNA off-target mutations, we performed WGS using the most recently established GOTI method (*33*). A blastomere of two-cell embryos derived from Ai9 (CAG-LoxP-Stop-LoxP-tdTomato) mice was injected with Cre mRNA, A3G-BE5.13 mRNA, and sgRNA. At embryonic day 14.5 (E14.5), progeny cells were FACS (fluorescence-activated cell sorting)–sorted on the basis of tdTomato expression, and WGS was separately performed for the resulting two cell populations with (tdTomato$^+$) and without (tdTomato$^-$) the tdTomato expression (Fig. 5A) (*33*). Using the WGS data obtained from the tdTomato$^-$ sample as the reference, single nucleotide variants (SNVs) for the tdTomato$^+$ sample were called via three different algorithms, and the overlapping SNVs detected from all the three algorithms were counted as the true off-target variants. Notably, we detected only 17 and 24 SNVs per embryo in each replicate from those treated by A3G-BE5.13, similar to the spontaneous mutation rate found from embryos delivered with Cre alone, as compared to the average of 283 SNVs per embryo by BE3 as previously detected (Fig. 5B and fig. S8A) (*33*). The mutation patterns of A3G-BE5.13 only showed a slight bias toward C-to-T or G-to-A compared with BE3 (Fig. 5C). We also tested the on-target Tyr-C site used in the GOTI experiments, which harbors both $C_3\underline{C}_4$ and $C_4T\underline{C}_6$ motifs. The WGS results showed that the editing only happened at the $\underline{C}_6$ in the $C_4T\underline{C}_6$ motif, which is consistent with our previous data that A3G-BEs could selectively edit a target C in the CT<u>C</u> motif. (fig. S8B). Collectively, these data indicated that A3G-BE5.13 induces minimum DNA off-target SNVs across the genome while maintains highly efficient and selective editing at the on-target position.

Last, we characterized the transcriptome-wide off-target effect of A3G-BE5.13. We transfected HEK293T with sgRNA and nCas9, BE4max, or A3G-BE5.13 encoded in plasmid as cotranslational P2A fusion to green fluorescent protein (GFP). After 48 hours, we sorted cells with the top 5% GFP signal to isolate the high-expression population (Fig. 5D). We first confirmed the robust on-target efficiency of DNA editing by BE4max and A3G-BE5.13 in these cells using HTS (fig. S8C). We then performed RNA-seq and analyzed the sequencing data to call SNVs in each replicate sample according to the method described previously (*10*). Our results showed that the engineered A3G-BE5.13 did not induce significant RNA SNVs as compared to the control treated by the nCas9 (Fig. 5E). However, BE4max caused a substantial amount of off-target mutations, in line with the previous studies of the wild-type rAPOBEC1-based CBEs (*9*–*11*). Distribution of mutation types of the detected SNVs of A3G-BE5.13 was similar to that of the nCas9 control, indicating a minimum disturbance on the transcriptome despite the high expression of intracellular A3G-BE5.13 proteins (Fig. 5F). These results further demonstrate that the A3G-BEs developed in this study are with high precision and markedly reduced RNA editing activity (*9*, *10*) and indicate that A3G-BE5.13 could serve as a promising CBE variant with high fidelity and minimum risk of off-target effects.

DISCUSSION

Here, we developed and characterized three new base editors using the A3G deaminase that is capable of recognizing the unique natural motif of CC<u>C</u>A. A3G-BE4.4 displays considerable editing efficiency and selectivity when the target motif lies within around positions 5 to 11 of the protospacer. In most of the sites, A3G-BE4.4 exhibited remarkable sequence specificity by discriminating between two consecutive Cs. However, we also observed that A3G-BE4.4 editing efficiency was poor at certain sites, probably due to the presented motifs being disfavored by the wild-type A3G and its naturally moderate catalytic activity, which could be improved by our engineered A3G-BE5.13 and A3G-BE5.14 variants (*36*). Both A3G-BE5.13 and A3G-BE5.14 displayed high efficiency across broader activity windows, from positions 4 to 15, with slightly relaxed C<u>C</u> selectivity. An initial screening of these three A3G-BEs could be conducted to determine which one performs the best for the selective editing of a single desired C.

We estimated the scope of base-editable disease variants that could be corrected by using A3G-BEs. Among the total of 1515 pathogenic

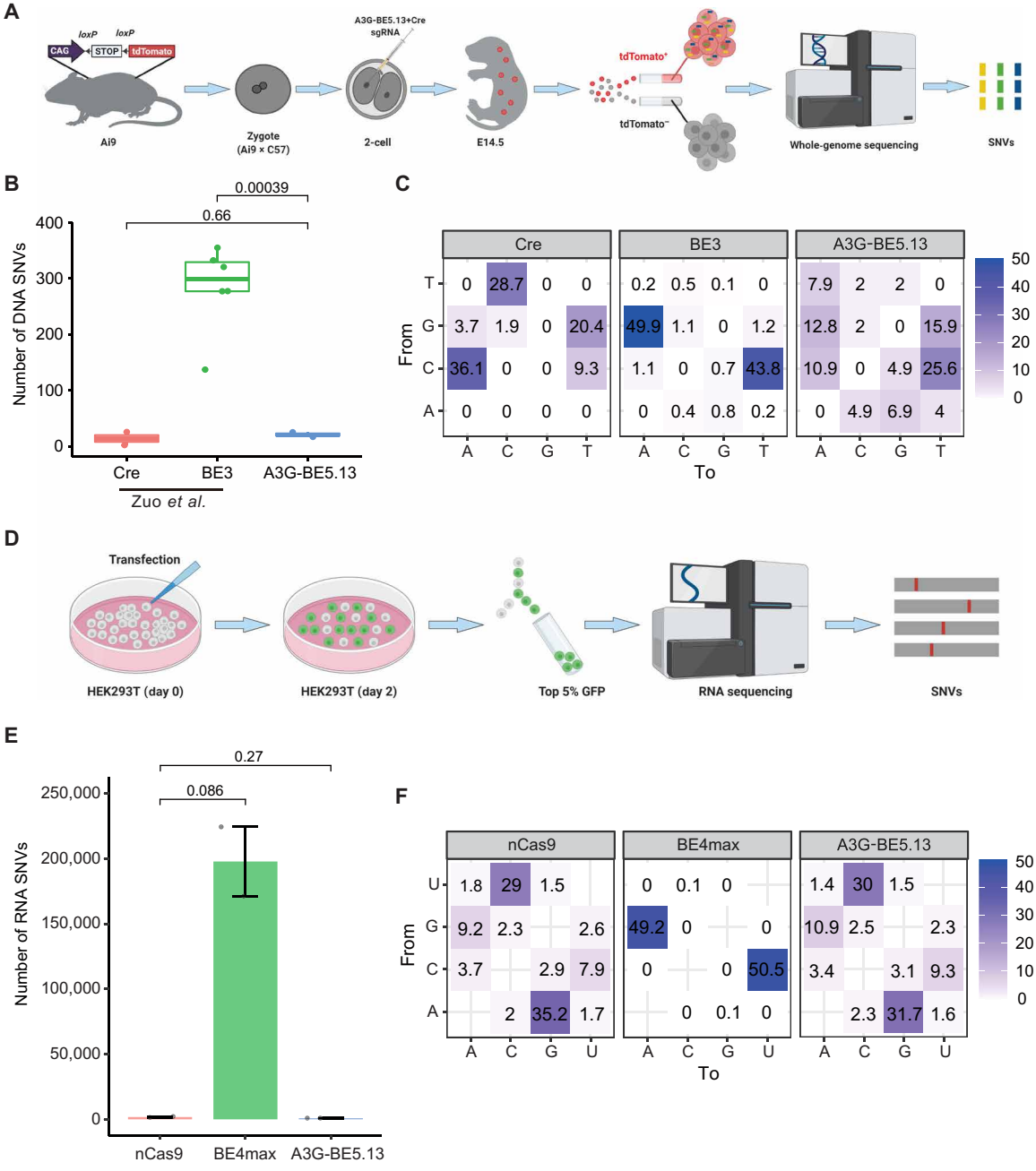

Fig. 5. Genome and transcriptome-wide off-target effects by A3G-BE5.13. (**A**) Scheme of the experimental workflow of GOTI. (**B**) Comparison of the total number of detected DNA off-target SNVs using the GOTI method. The number of SNVs identified in Cre-, BE3-, and A3G-BE5.13–treated embryos were 14 ± 12 (SD; $n = 2$), 283 ± 32 (SD; $n = 6$), and 20 ± 5 (SD; $n = 2$), respectively. (**C**) Distribution of DNA mutation types in each group. (**D**) Scheme of the experimental workflow of identifying transcriptome-wide off-target SNVs through RNA-seq. (**E**) Comparison of the total number of detected RNA off-target SNVs. The number of SNVs identified in nCas9-, BE4max-, A3G-BE5.13–treated cells were 2669 ± 712 (SD; $n = 2$), 198,688 ± 37,775 (SD; $n = 2$), and 1410 ± 39 (SD; $n = 2$), respectively. (**F**) Distribution of RNA mutation types in each group. For (C) and (F), the number in each cell indicates the percentage of a certain type of mutation among all mutations. For (B) and (E), each data point represents independent biological replicates performed on different days.

SNPs identified within the BEable-GPS (Base Editable prediction of Global Pathogenic-related SNVs) entries (*12*), 61% (929 of 1515) were found to lie within the CC or CNC sequence context preferred by A3G-BEs (*18*). We then identified 540 human pathogenic SNPs that could be precisely correctable by our A3G-BEs, occupying 36% of the total number (see data file S1). Manual filtering was conducted to ensure that neighboring bystander Cs within the activity window did not exist along with the target motif of A3G-BEs. This indicates that our engineered A3G-BEs greatly expand the number of precisely targetable genetic variants for potential therapeutic applications.

WGS and RNA-seq analysis suggested that our A3G-BEs variants induce minimum levels of both DNA and RNA off-target SNVs. A3G's

intrinsically high sequence specificity could reduce the probability of deaminating Cs other than its preferential motif. Our orthogonal R-loop assay showed that A3G-BEs exhibit a greater propensity to edit cytosines lying within the CC motif (fig. S7C). Apart from this reason, an earlier study indicated that mutations in the conserved zinc-coordinating, or catalytic, residues of either the NTD or CTD of the full-length A3G nearly abolished its capability to edit RNA and demonstrated that both domains are essential for optimal RNA editing (*37*). We speculate that the high fidelity of our engineered A3G-BEs could be due to the lack the NTD so that their ability to cause mutations in the transcriptome might be impaired (Fig. 5, D to F). These findings greatly mitigate the concerns about the off-target issues associated with A3G-BEs, showing great potential for their future therapeutic applications.

It is imperative that we develop genome editing tools that have the ability to produce anticipated results with the highest probability with minimum errors. Bystander editing is a major factor giving rise to imprecision, a limitation that should be improved for future clinical usage. Our engineered A3G-BEs here that recognize a specific CC motif could offer a toolkit to precisely edit a target C. These toolkits, if expanded, could allow versatile and precise editing of single nucleotides from various other distinct motifs. We envision that the continued development of novel base editing technology could facilitate the precise conversion of cytosines and treatment of human genetic diseases.

MATERIALS AND METHODS

Mammalian cell culture

HEK293T cells (American Type Culture Collection, CRL-3216) were cultured in the T-75 flask (Corning) using high-glucose Dulbecco's modified Eagle's medium (DMEM) with GlutaMAX and sodium pyruvate (Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (FBS) (Thermo Fisher Scientific) and 1× penicillin-streptomycin (Thermo Fisher Scientific) at 37°C with 5% CO_2. Upon reaching 80 to 90% confluency, cells were dissociated using TrypLE Express (Life Technologies) and passaged at a ratio of 1:3. Cells were verified mycoplasma-free using a mycoplasma detection kit (abm). ESI-017 hESCs (ESI BIO, CVCL_B854) and iPSCs (Coriell Institute, AICS-0058-067) were maintained in mTeSR1 (STEMCELL Technologies) in tissue culture dish coated with Matrigel (1:200; Corning). Dispase (STEMCELL Technologies) was used for routine passage. To perform nucleofection, a single-cell suspension was prepared using Accutase (Innovative Cell Technologies). The pluripotency of those cells was confirmed via staining of Oct4, Sox2, and Nanog. Both ESI-017 and iPSC lines were routinely tested for mycoplasma contamination and found negative.

Plasmid construction

A3G-BE2.1 was constructed by amplifying the BE4max plasmid (Addgene) outside the rAPOBEC1 region and In-Fusion cloning (Takara) with the synthesized human codon-optimized A3G fragment (Integrated DNA Technologies). Deletion of the NTD of A3G to construct A3G-BE4.4 was performed by polymerase chain reaction (PCR) amplification of A3G-BE2.1 outside the NTD region using Q5 High-Fidelity 2X Master Mix (New England Biolabs) and re-cloning the linearized fragment. Sets of mutations introduced into A3G-BE variants for enhancing editing efficiencies—including A3G-BE5.1, A3G-BE5.3, A3G-BE5.4, and A3G-BE5.10—were constructed using gBlocks (Integrated DNA Technologies) that contain the desired mutations and cloned with the remaining backbone of the A3G-BE4.4 plasmid. Other variants for introducing individual mutations, including Y315F, were constructed by site-directed mutagenesis using the general PCR method. Gibson assembly was used to attach P2A-GFP fragment to the C-terminal ends of nCas9, BE4max, and A3G-BE5.13 for the RNA-seq experiment that requires sorting of the transfected cells with the top 5% GFP signal. Similarly, the P2A-PuroR fragment was attached to the C-terminal ends of BE4max, A3G-BE4.4, A3G-BE5.13, and A3G-BE5.14 through Gibson assembly to select puromycin-resistant cells after nucleofection of iPSCs and hESCs. All assembled constructs were transformed into Stellar competent cells (Takara). Plasmids were extracted using either the QIAprep Spin Miniprep Kit (Qiagen) or the ZymoPURE II Plasmid Midiprep Kit (Zymo Research), and concentrations were measured using NanoDrop One (Thermo Fisher Scientific). sgRNAs were constructed by using the previous method (*38*). Briefly, a pair of primers for top and bottom strands encoding the 20-bp target sequence were 5′ phosphorylated using T4 polynucleotide kinase (New England Biolabs) and annealed by heating the oligos to 95°C and cooling down to room temperature at 5°C/min^{-1}. The mixture was diluted 1:25 using water and ligated into a sgRNA expression vector using T4 DNA ligase (New England Biolabs) and BsaI–HF v2 (New England Biolabs) following the manufacturer's instructions.

Creating stable cell line disease model

The HEK293T stable cell line was constructed by cloning a 200-bp fragment of disease-associated gene upstream of an EF1α promoter to drive the expression of the puromycin-resistant gene in a lentiviral vector. The single-base mutation of a disease-associated gene was inserted by PCR and In-Fusion cloning (Takara). The lentiviral vector was transfected into HEK293T cells in a 24-well plate (Olympus) at 80 to 90% confluency. For each well, 288 ng of the plasmid containing the vector of interest, 72 ng of pMD2.G, and 144 ng of psPAX2 were transfected using 1.0 μl of Lipofectamine 2000 and 25 μl of Opti-MEM I reduced serum medium (Life Technologies). Viral supernatant was harvested 48 hours after transfection, filtered with a 0.45-μm polyvinylidene difluoride filter (Millipore), and then serially diluted to add into a 24-well plate cultured with 5×10^4 HEK293T cells per well. After 24 hours, cells transduced with lentivirus were split into new plate wells supplemented with puromycin (3 μg/ml^{-1}). Seventy-two hours after the puromycin selection, cells were harvested from the well with the fewest surviving colonies to ensure single-copy integration and were then further cultured for expansion.

HEK293T transfection and genomic DNA extraction

Transfection and extraction of the genomic DNA were adopted from the previous method (*7*). Briefly, HEK293T cells were counted using Countess II FL (Thermo Fisher Scientific) and plated into a poly-D-lysine–coated 48-well plate (Corning) under 250 μl of the cell culture medium with a density of 4.5×10^4 cells per well. After ~16 hours, cells were transfected using 1.2 μl of Lipofectamine 2000 (Thermo Fisher Scientific) with 750 ng of base editor, plasmid and 250 ng of sgRNA plasmid per well following the manufacturer's protocol. For orthogonal R-loop assay, 300 ng of BE, 300 ng of dSaCas9, 200 ng of SpCas9 sgRNA, and 200 ng of SaCas9 sgRNA plasmids were cotransfected per well using 1.2 μl of Lipofectamine 2000. After incubation at 37°C for 72 hours, the medium was aspirated and incubated under 100 μl of lysis buffer [10 mM tris-HCl (pH 7.5), 0.05% SDS, and

proteinase K (25 μg/ml^{-1}) (Fisher BioReagents)] for 1 hour at 37°C. The lysed mixture was heat inactivated at 80°C for 30 min and stored at 4°C until use. For preparing RNA-seq samples, 7.5×10^6 cells were seeded in 10-cm culture dish and transfected after 20 hours with 22.5 μg of base editor P2A-GFP expression plasmid and 7.5 μg of *EMX1* #1–targeting sgRNA plasmid mixed with 90 μg of PEI MAX (Polysciences) in 1.0 ml of Opti-MEM I. The mixture was incubated for 30 min in room temperature and applied to the cells dropwise before cell sorting after 48 hours.

HTS library preparation

The HTS library was prepared using two rounds of PCR. For the first round, a 200-bp DNA fragment of the target region was amplified in a total volume of 25 μl mixed with 12.5 μl of the Q5 High-Fidelity 2X Master Mix, 1 μl of the extracted genomic DNA, and a pair of primers (see the Supplementary Materials). Successful amplification of individual samples was checked using 1% agarose gel. For the second round, combinations of different Illumina indexes were attached at each 5′ and 3′ end of the first PCR products using the same total PCR volume. The PCR products were combined and column purified using a QIAquick PCR Purification kit (Qiagen) and further gel extracted to remove nonspecific amplifications. The final mixture of the library was quantified using the Qubit dsDNA HS Assay Kit (Life Technologies) and prepared for loading into a 150-cycle MiSeq reagent kit v3 (Illumina) according to the manufacturer's protocol.

General HTS analysis

FASTQ files were generated by demultiplexing total sequencing reads using the MiSeq Reporter or Illumina's bcl2fastq 2.17 software. CRISPResso2 (available in GitHub; https://github.com/pinellolab/CRISPResso2) was used with the batch mode function to quantify the base editing conversion rates, indel frequencies, and product purities of the aligned reads (*39*). Heat maps displaying average base editing frequencies at each nucleotide position of three independent biological replicates were generated by running the CRISPResso2 analysis.

GOTI experiments using mouse embryo

The use and care of animals followed the guidelines of the Biomedical Research Ethics Committee of Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. GOTI experiments were performed according to the previous method (*33*). Briefly, mRNA of A3G-BE5.13 or Cre was generated by attaching the T7 promoter to the coding region through PCR amplification and using its purified PCR product as the template for in vitro transcription (IVT) using the mMESSAGE mMACHINE T7 ULTRA Kit (Invitrogen). Similarly, for sgRNA, the T7 promoter was attached, and the MEGAshortscript T7 Transcription Kit (Invitrogen) was used for IVT. mRNA and sgRNA products were purified using the MEGAclear Transcription Clean-Up Kit (Invitrogen). Fertilized embryos were obtained from C57BL/6 females (4 weeks old) mated to heterozygous Ai9 males (JAX strain 007909). A3G-BE5.13 mRNA (50 ng/μl), Cre mRNA (2 ng/μl), and sgRNA (50 ng/μl) were mixed and injected using a FemtoJet microinjector (Eppendorf) into the cytoplasm of one blastomere of the two-cell embryo in a droplet of Hepes-CZB (Chatot-Ziomek-Bavister) medium containing cytochalasin B (5 μg/ml). The embryos were incubated at 37°C with 5% CO_2 under KSOM (Potassium simplex optimized medium) medium for 2 hours and transferred into oviducts of ICR (Institute for Cancer Research) females at 0.5 days post coitum.

WGS and data analysis

WGS and data analysis were performed according to the previous method (*33*). Briefly, at E14.5, prepared fetal tissues were dissociated using trypsin-EDTA (0.05%) and homogenized by passing through pipette tips multiple times. Cells were centrifuged, and the resulting pellet was resuspended in DMEM supplemented with 10% FBS before filtering through a 40-μm cell strainer. tdTomato$^-$ and tdTomato$^+$ cells were isolated through FACS, and their genomic DNA were each extracted using the DNeasy Blood and Tissue Kit (Qiagen). WGS was performed at mean coverages of 50× by Illumina HiSeq X Ten. Burrows-Wheeler Aligner (version 0.7.12) was used to map qualified sequencing reads to the reference genome (mm10), and then the mapped BAM files were sorted and marked using Picard tools (version 2.3.0). SNVs were called from three algorithms, Mutect2 (version 3.5), LoFreq (version 2.1.2), and Strelka (version 2.7.1) with default parameters, separately (*40*–*42*). Using the tdTomato$^-$ sample from the same embryo as the reference, only variants shown to be mutated in the tdTomato$^+$ at the same coordinate were counted within the mapped BAM file. SNVs overlapping from all the three algorithms were considered as the true variants.

RNA-seq experiments

Forty-eight hours after transfection, HEK293T cells cultured in 10-cm dish were washed with phosphate-buffered saline (Thermo Fisher Scientific) and dissociated by TrypLE Express. Cells were centrifuged, and the resulting pellet was resuspended in 5 ml of normal culture medium. Cells (0.5 to 0.7×10^6) with the top 5% GFP signal were sorted using SH800S cell sorter (Sony). Approximately a quarter of the sorted cells were collected in separate tubes for genomic DNA extraction and HTS analysis of the on-target base editing. For the remaining cells, the RNeasy Plus Mini Kit (Qiagen) was used to purify the total RNA. RNA library preparations and sequencing reactions were conducted at GENEWIZ LLC. (South Plainfield, NJ, USA). RNA samples were quantified using Qubit 2.0 fluorometer (Life Technologies), and RNA integrity was checked using Agilent TapeStation 4200 (Agilent Technologies). Sequencing libraries were prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina following the manufacturer's instructions (New England Biolabs). Briefly, mRNAs were enriched with Oligo(dT) beads and were fragmented for 15 min at 94°C. First- and second-strand cDNAs were subsequently synthesized. cDNA fragments were end-repaired and adenylated at 3′ ends, and universal adapters were ligated to cDNA fragments, followed by index addition and library enrichment by limited-cycle PCR. The sequencing libraries were validated on the Agilent TapeStation (Agilent Technologies) and quantified by using Qubit 2.0 fluorometer and by quantitative PCR (Kapa Biosystems). The sequencing libraries were clustered on one lane of a flowcell and loaded on the Illumina HiSeq 4000 to be sequenced using a 2× 150-bp paired-end configuration.

RNA-seq data analysis

RNA-seq data analysis was carried out using the previous method (*10*). Qualified reads obtained from FastQC (version 0.11.3) and Trimmomatic (version 0.36) were aligned to the reference genome (Ensembl GRCh38) using STAR (version 2.5.2b) in two-pass mode with default parameters (*43*). Picard tools (version 2.3.0) were 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 (version 3.5), respectively (*44*). Clusters of more than four SNVs identified within a 35-bp window were filtered to maintain high-confidence variants, and found variants with base quality of >25, mapping quality score of >20, Fisher strand values of >30.0, qual by depth values of <2.0, and sequencing depth of >20 were counted.

Base editing in iPSC and hESC through nucleofection

For nucleofection of iPSCs and hESCs, cells were detached by using Accutase. For each reaction, 1.0×10^6 cells were resuspended in 82 μl of P3 Primary Cell Nucleofector Solution and 18 μl of supplement 1 using the P3 Primary Cell 4D-Nucleofector X Kit L (Lonza). Three micrograms of base editor P2A-PuroR expression plasmid and 1 μg of sgRNA plasmid were added in the single-cell suspension and mixed well. The single-cell suspension was then transferred into a Nucleocuvette. Nucleofection was carried out in 4D-Nucleofector X Unit (Lonza) using code CB200, and cells were immediately plated on a Matrigel-coated 35-mm dish in mTeSR supplemented with 1× CloneR (STEMCELL Technologies). After 24 hours, puromycin ($1.0\ \mu g/ml^{-1}$) was supplemented into the medium for 1 day selection, and the surviving colonies were expanded for 10 to 14 days until extraction of the genome using the DNeasy Blood and Tissue Kit (Qiagen). The target region was PCR amplified using 30 cycles and sent for Sanger sequencing. EditR (baseeditr.com) was used to quantify the mutation peaks of Sanger chromatograms for analyzing the base conversion.

Analysis of potential genetic diseases correctable by A3G-BEs

Bioinformatic analysis of pathogenic SNPs obtained from the BEable-GPS database (https://picb.ac.cn/rnomics/BEable-GPS/) was performed by finding correctable pathogenic SNPs that contain the target C located within the activity window of positions 4 to 8 of the protospacer, with NGG PAM positioned 21 to 23 (*12*). We then manually filtered the list on the basis of the sequence contexts containing the C<u>C</u> and/or CN<u>C</u> motif preferred by A3G-BEs. We counted precisely correctable pathogenic SNPs by manually filtering each disease on the basis of whether another base-editable bystander C was present within the activity window. For example, variant NM_012203.1(GRHPR): c.84-2A>G (protospacer; 5′-TCACAGC<u>C</u>GCGGGGAAAGGG-3′), in which the target C lies in the C<u>C</u> context but has a nearby bystander C lying in a CA<u>C</u> context potentially editable by A3G-BEs was removed from counting. The summarized list of SNPs can be found in data file S1.

Statistical analysis

Three biologically independent replicates performed on different days were used to calculate means and SD unless stated otherwise. All bar plots and figures except for heat maps were generated using Prism 8 (GraphPad). *P* values were calculated using Prism 8 by performing two-tailed Student's *t* test, with a statistical significance level represented on each figure as ns (not significant), $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, and $****P < 0.0001$.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/29/eaba1773/DC1

REFERENCES AND NOTES

1. A. C. Komor, Y. B. Kim, M. S. Packer, J. A. Zuris, D. R. Liu, Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
2. K. Nishida, T. Arazoe, N. Yachie, S. Banno, M. Kakimoto, M. Tabata, M. Mochizuki, A. Miyabe, M. Araki, K. Y. Hara, Z. Zhimatani, A. Kondo, Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353**, aaf8729 (2016).
3. H. A. Rees, D. R. Liu, Base editing: Precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **19**, 770–788 (2018).
4. A. C. Komor, A. H. Badran, D. R. Liu, CRISPR-based technologies for the manipulation of eukaryotic genomes. *Cell* **168**, 20–36 (2017).
5. A. V. Anzalone, P. B. Randolph, J. R. Davis, A. A. Sousa, L. W. Koblan, J. M. Levy, P. J. Chen, C. Wilson, G. A. Newby, A. Raguram, Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
6. A. C. Komor, K. T. Zhao, M. S. Packer, N. M. Gaudelli, A. L. Waterbury, L. W. Koblan, Y. B. Kim, A. H. Badran, D. R. Liu, Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Sci. Adv.* **3**, eaao4774 (2017).
7. L. W. Koblan, J. L. Doman, C. Wilson, J. M. Levy, T. Tay, G. A. Newby, J. P. Maianti, A. Raguram, D. R. Liu, Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).
8. R. T. Walton, K. A. Christie, M. N. Whittaker, B. P. Kleinstiver, Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**, 290–296 (2020).
9. J. Grünewald, R. Zhou, S. P. Garcia, S. Iyer, C. A. Lareau, M. J. Aryee, J. K. Joung, Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **569**, 433–437 (2019).
10. C. Zhou, Y. Sun, R. Yan, Y. Liu, E. Zuo, C. Gu, L. Han, Y. Wei, X. Hu, R. Zeng, Y. Li, H. Zhou, F. Guo, H. Yang, Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis. *Nature* **571**, 275–278 (2019).
11. E. Zuo, Y. Sun, T. Yuan, B. He, C. Zhou, W. Ying, J. Liu, W. Wei, R. Zeng, Y. Li, H. Yang, High-fidelity base editor with no detectable genome-wide off-target effects. *bioRxiv* 10.1101/2020.02.07.939074 (2020).
12. Y. Wang, R. Gao, J. Wu, Y.-C. Xiong, J. Wei, S. Zhang, B. Yang, J. Chen, L. Yang, Comparison of cytosine base editors and development of the BEable-GPS database for targeting pathogenic SNVs. *Genome Biol.* **20**, 1–7 (2019).
13. R. Hunt, Z. E. Sauna, S. V. Ambudkar, M. M. Gottesman, C. Kimchi-Sarfaty, Silent (synonymous) SNPs: Should we care about them? in *Single Nucleotide Polymorphisms* (Humana Press, Totowa, NJ, 2009), vol. 578, pp. 23–39.
14. H. K. Lee, M. Willi, S. M. Miller, S. Kim, C. Liu, D. R. Liu, L. Hennighausen, Targeting fidelity of adenine and cytosine base editors in mouse embryos. *Nat. Commun.* **9**, 4804 (2018).
15. Y. B. Kim, A. C. Komor, J. M. Levy, M. S. Packer, K. T. Zhao, D. R. Liu, Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat. Biotechnol.* **35**, 371–376 (2017).
16. J. Tan, F. Zhang, D. Karcher, R. Bock, Engineering of high-precision base editors for site-specific single nucleotide replacement. *Nat. Commun.* **10**, 439 (2019).
17. J. Tan, F. Zhang, D. Karcher, R. Bock, Expanding the genome-targeting scope and the site selectivity of high-precision base editors. *Nat. Commun.* **11**, 629 (2020).
18. J. M. Gehrke, O. Cervantes, M. K. Clement, Y. Wu, J. Zeng, D. E. Bauer, L. Pinello, J. K. Joung, An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities. *Nat. Biotechnol.* **36**, 977–982 (2018).
19. R. S. Harris, K. N. Bishop, A. M. Sheehy, H. M. Craig, S. K. Petersen-Mahrt, I. N. Watt, M. S. Neuberger, M. H. Malim, DNA deamination mediates innate immunity to retroviral infection. *Cell* **113**, 803–809 (2003).
20. S. G. Conticello, The AID/APOBEC family of nucleic acid mutators. *Genome Biol.* **9**, 229 (2008).
21. R. C. Beale, S. K. Petersen-Mahrt, I. N. Watt, R. S. Harris, C. Rada, M. S. Neuberger, Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: Correlation with mutation spectra in vivo. *J. Mol. Biol.* **337**, 585–596 (2004).
22. M.-A. Langlois, R. C. Beale, S. G. Conticello, M. S. Neuberger, Mutational comparison of the single-domained APOBEC3C and double-domained APOBEC3F/G anti-retroviral cytidine deaminases provides insight into their DNA target site specificities. *Nucleic Acids Res.* **33**, 1913–1923 (2005).
23. T.-L. Cheng, S. Li, B. Yuan, X. Wang, W. Zhou, Z. Qiu, Expanding C–T base editing toolkit with diversified cytidine deaminases. *Nat. Commun.* **10**, 3612 (2019).
24. R. Nowarski, P. Prabhu, E. Kenig, Y. Smith, E. Britan-Rosich, M. Kotler, APOBEC3G inhibits HIV-1 RNA elongation by inactivating the viral trans-activation response element. *J. Mol. Biol.* **426**, 2840–2853 (2014).
25. M. Morse, R. Huo, Y. Feng, I. Rouzina, L. Chelico, M. C. Williams, Dimerization regulates both deaminase-dependent and deaminase-independent HIV-1 restriction by APOBEC3G. *Nat. Commun.* **8**, 597 (2017).

26. L. G. Holden, C. Prochnow, Y. P. Chang, R. Bransteitter, L. Chelico, U. Sen, R. C. Stevens, M. F. Goodman, X. S. Chen, Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications. *Nature* **456**, 121–124 (2008).
27. K.-M. Chen, E. Harjes, P. J. Gross, A. Fahmy, Y. Lu, K. Shindo, R. S. Harris, H. Matsuo, Structure of the DNA deaminase domain of the HIV-1 restriction factor APOBEC3G. *Nature* **452**, 116–119 (2008).
28. Q. Yu, R. König, S. Pillai, K. Chiles, M. Kearney, S. Palmer, D. Richman, J. M. Coffin, N. R. Landau, Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. *Nat. Struct. Mol. Biol.* **11**, 435–442 (2004).
29. S. J. Ziegler, C. Liu, M. Landau, O. Buzovetsky, B. A. Desimmie, Q. Zhao, T. Sasaki, R. C. Burdick, V. K. Pathak, K. S. Anderson, Y. Xiong, Insights into DNA substrate selection by APOBEC3G from structural, biochemical, and functional studies. *PLOS ONE* **13**, e0195048 (2018).
30. A. Maiti, W. Myint, T. Kanai, K. Delviks-Frankenberry, C. S. Rodriguez, V. K. Pathak, C. A. Schiffer, H. Matsuo, Crystal structure of the catalytic domain of HIV-1 restriction factor APOBEC3G in complex with ssDNA. *Nat. Commun.* **9**, 2460 (2018).
31. A. Rathore, M. A. Carpenter, Ö. Demir, T. Ikeda, M. Li, N. M. Shaban, E. K. Law, D. Anokhin, W. L. Brown, R. E. Amaro, R. S. Harris, The local dinucleotide preference of APOBEC3G can be altered from 5′-CC to 5′-TC by a single amino acid substitution. *J. Mol. Biol.* **425**, 4442–4454 (2013).
32. M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, D. R. Maglott, ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
33. E. Zuo, Y. Sun, W. Wei, T. Yuan, W. Ying, H. Sun, L. Yuan, L. M. Steinmetz, Y. Li, H. Yang, Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science* **364**, 289–292 (2019).
34. S. Jin, Y. Zong, Q. Gao, Z. Zhu, Y. Wang, P. Qin, C. Liang, D. Wang, J.-L. Qiu, F. Zhang, C. Gao, Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292–295 (2019).
35. J. L. Doman, A. Raguram, G. A. Newby, D. R. Liu, Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors. *Nat. Biotechnol.* **38**, 620–628 (2020).
36. M. A. Carpenter, M. Li, A. Rathore, L. Lackey, E. K. Law, A. M. Land, B. Leonard, S. M. Shandilya, M.-F. Bohn, C. A. Schiffer, W. L. Brown, R. S. Harris, Methylcytosine and normal cytosine deamination by the foreign DNA restriction enzyme APOBEC3A. *J. Biol. Chem.* **287**, 34801–34808 (2012).
37. S. Sharma, S. K. Patnaik, R. T. Taggart, B. E. Baysal, The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. *Sci. Rep.* **6**, 39100 (2016).
38. F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, F. Zhang, Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
39. K. Clement, H. Rees, M. C. Canver, J. M. Gehrke, R. Farouni, J. Y. Hsu, M. A. Cole, D. R. Liu, J. K. Joung, D. E. Bauer, L. Pinello, CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).
40. K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, G. Getz, Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
41. C. T. Saunders, W. S. Wong, S. Swamy, J. Becq, L. J. Murray, R. K. Cheetham, Strelka: Accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
42. A. Wilm, P. P. K. Aw, D. Bertrand, G. H. T. Yeo, S. H. Ong, C. H. Wong, C. C. Khor, R. Petric, M. L. Hibberd, N. Nagarajan, LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).
43. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
44. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

Acknowledgments: We thank D. Zhang's NABLab (Rice University) and G. Bao's laboratory (Rice University) for providing the usage of the MiSeq Sequencing System**. Funding:** This work was supported by the Robert A. Welch Foundation (C-1952 to X.G. and C-1559 to A.B.K.), the NIH grant (HL151545 to X.G.), the Rice University Creative Ventures Fund (to X.G. and A.B.K.), the NSF grants (CHE-1664218 to A.B.K. and PHY-1427654 to the Center for Theoretical Biological Physics), the National Natural Science Foundation of China (31922048 to E.Z.), and the Agricultural Science and Technology Innovation Program (to E.Z.). **Author contributions:** S.L., N.D., and X.G. designed the study. S.L. and N.D. constructed plasmids, performed FACS, and prepared the HTS library. S.L. performed transfection, HTS, and HTS data analysis. S.L. and Q.Y. maintained HEK293T cells and created disease-associated stable cell lines. Y.S., T.Y., and E.Z. performed GOTI, WGS, and software analysis of the off-target SNVs. J.L. and I.B.H. helped with RNA-seq sample preparation. S.L. and L.L. performed nucleofection and clonal expansion of iPSCs and ESI-017 hESCs. N.D. performed the analysis of pathogenic SNPs statistics. S.L. and J.Y. performed statistical analysis. S.L., N.D., Q.W., and A.B.K. provided structural insights into A3G. All authors wrote and edited the manuscript. **Competing interests:** S.L., N.D., and X.G. are inventors on a pending provisional patent application submitted by the William Marsh Rice University related to this work. The authors declare that they have no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. FASTQ files containing HTS reads have been deposited in the National Center for Biotechnology Information, NIH Sequencing Read Archive and are available with accession number PRJNA623461. Additional data related to this paper may be requested from the authors.

Submitted 14 November 2019
Accepted 2 June 2020
Published 15 July 2020
10.1126/sciadv.aba1773

Citation: S. Lee, N. Ding, Y. Sun, T. Yuan, J. Li, Q. Yuan, L. Liu, J. Yang, Q. Wang, A. B. Kolomeisky, I. B. Hilton, E. Zuo, X. Gao, Single C-to-T substitution using engineered APOBEC3G-nCas9 base editors with minimum genome- and transcriptome-wide off-target effects. *Sci. Adv.* **6**, eaba1773 (2020).

Single C-to-T substitution using engineered APOBEC3G-nCas9 base editors with minimum genome- and transcriptome-wide off-target effects

Sangsin Lee, Ning Ding, Yidi Sun, Tanglong Yuan, Jing Li, Qichen Yuan, Lizhong Liu, Jie Yang, Qian Wang, Anatoly B. Kolomeisky, Isaac B. Hilton, Erwei Zuo and Xue Gao

Sci Adv **6** (29), eaba1773.
DOI: 10.1126/sciadv.aba1773