Title: Precise and programmable C:G to G:C base editing in genomic DNA

Authors: Liwei Chen, Jung Eun Park, Peter Paa, Priscilla D. Rajakumar, Yi Ting Chew, Swathi N. Manivannan, Wei Leong Chew*

Affiliation: Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore.

*Correspondence to: chewwl@gis.a-star.edu.sg

Abstract: Many genetic diseases are caused by single-nucleotide polymorphisms (SNPs). Base editors can correct SNPs at single-nucleotide resolution, but until recently, only allowed for C:G to T:A and A:T to G:C transition edits, addressing four out of twelve possible DNA base substitutions. Here we developed a novel class of C:G to G:C Base Editors (CGBEs) to create single-base genomic transversions in human cells. Our CGBEs consist of a nickase CRISPR-Cas9 (nCas9) fused to a cytosine deaminase and base excision repair (BER) proteins. Characterization of >30 CGBE candidates and 27 guide RNAs (gRNAs) revealed that CGBEs predominantly perform C:G to G:C editing (up to 90% purity), with rAPOBEC-nCas9-rXRCC1 being the most efficient (mean C:G to G:C edits at 15% and up to 37%). CGBEs target cytosine in WCW, ACC or GCT sequence contexts and within a precise two-nucleotide window of the target protospacer. We further targeted genes linked to dyslipidemia, hypertrophic cardiomyopathy, and deafness, showing the therapeutic potential of CGBE in interrogating and correcting human genetic diseases.

Main text:

Many human genetic diseases are caused by single-nucleotide polymorphism (SNPs), in which the disease and healthy alleles differ by a single DNA base. Base editors can correct these SNPs by converting the targeted DNA bases into another base in a controllable and efficient fashion. Current technology enables the conversion of C:G base pairs to T:A base pairs using cytosine base editors (CBEs) and A:T base pairs to G:C base pairs using adenine base editors (ABEs)¹⁻³, which together represent half of all known disease-associated SNPs. CBEs and ABEs are also known to effect some C:G to G:C edits as byproducts^{4,5}, but they cannot effect such transversions at efficiencies or purities necessary for therapeutic use, and hence the remaining half of these SNPs are not addressable by current base-editors^{1,3}. Here, we developed a base editor capable of editing C:G to G:C (Fig. 1a), which opens up treatment avenues to half of all known SNP-associated human genetic diseases (Extended Data Table 1).

To effect C:G to G:C editing, we looked to leverage the cell's innate base excision repair (BER) pathway, in which DNA polymerase β , DNA ligase III and XRCC1 are major players⁶. Previous work suggests that removal of uracil glycosylase inhibitor (UGI) from a CBE might lead to an increase in C:G to G:C levels². UGI inhibits uracil DNA glycosylase (UDG), thereby inhibiting the downstream BER^{1,2,7}. We speculate that the endogenous BER pathway – in the presence of a bound nCas9 – is linked to the observed C:G to G:C editing and that replacing UGI with BER proteins might increase such editing events. Using a version of CBE – BE3 – as a starting point, we removed UGI, instead fusing rAPOBEC-nCas9 with DNA ligase 3, DNA repair protein XRCC1, or the DNA binding and lyase domain of DNA polymerase β (PB)⁸.

Because the relative orientation of Cas9 fusions may affect the activity of the fusion, we built several versions of our CGBE candidates with the fused rAPOBEC and BER proteins at different orientations with respect to nCas9 (Fig. 1b). We then separately treated HEK293AAV cells with each candidate along with gRNAs designed to target genomic sites HEK2 and HEK3, both of which were previously used to characterize base editors¹. As controls, we also treated cells with BE3 and BE4. Since no effective means of C:G to G:C base editing is currently known and BE3 has been observed to effect this reaction as a byproduct of C:G to T:A editing⁴, we used BE3 as a benchmark for our CGBEs.

High-throughput sequencing of the HEK2 site revealed that our CGBE candidates were able to edit C:G to both G:C and T:A. Out of the 31 candidates tested, 20 candidates showed an increased level of C:G to G:C editing relative to BE3 at position 6 within the protospacer (Extended Data Fig. 1a). The next closest available C was at position 4, where editing levels were lower (< 10%, Extended Data Fig. 1b), suggesting a narrower editing window than BE3¹. In the best performing candidates, up to 29% C:G to G:C editing and 6% C:G to T:A editing were observed for the CGBE candidates, compared with 14% and 36% respectively for BE4, and 16% and 18% respectively for BE3.

Similarly, sequencing at the HEK3 site revealed editing of C:G to both G:C and T:A. 12 out of 31 candidates gave C:G to G:C editing levels at position 5 that were higher than that achieved by BE3, with the best performing candidates effecting up to 13% C:G to G:C editing compared with 4% for BE3 (Extended Data Fig. 2a). Despite an increase in C:G to T:A editing with CGBEs compared to BE3, a more than twofold increase in C:G to G:C editing with CGBEs led to an increase in the editing ratio of C:G to G:C compared to C:G to T:A. Compared to position 5, C:G to G:C editing at position 3 and position 4 are significantly lower (Extended Data Fig. 2b and 2c). From this pilot screen, we shortlisted seven of the best performing candidates for a secondary screen (Fig. 1c).

We further tested the CGBEs for C:G to G:C editing at four genomic sites known to be amenable to BE3-mediated editing – EMX1, HEK4, RNF2, and FANCF (Extended Data Fig. 3a). With CGBEs, C:G to G:C edits were efficiently induced (17-24%, compared to 8-10% with BE3) as the predominant product (up to 69% purity) at HEK4 and RNF2. At sites EMX1 and FANCF, up to 9% C:G to G:C editing was observed despite this not being the predominant edit, while BE3 effected up to 3% C:G to G:C editing. From this secondary screen, we chose ACX, rXRCC1 and ACX, rPB(8kD) for further characterization.

We next determined if target sequence context impacts editing efficiency. In characterizing the two CGBEs with eight other gRNAs – including dyslipidemia-associated gene ADRB2⁹, hearing loss-associated gene GJB2¹⁰, and hypertrophic cardiomyopathy-associated gene MYBPC3¹¹ – we noticed that not only did CGBEs efficiently interrogate disease-associated genes, but they also gave higher levels of C:G to G:C editing at C's immediately following an A/T (Extended Data Fig. 3b, Extended Data Fig. 4, and Extended Data Table 2). Hence, we suspected that the difference in editing efficiency between gRNAs might be due to the different motifs within which the targeted C is located (e.g. ACA at HEK2, CCA at HEK3; targeted C is underlined). To address such potential differences, we designed 16 guides targeting HEK2. These guides collectively contained all possible NCN motifs with targeted C's at position 6 (Fig. 2a, Extended Data Fig. 3c). Sequencing revealed that the most readily edited motif was ACA, while the least edited motif was GCC. Up to 10% C:G to G:C editing

was observed for ACA, ACC, ACT, GCT, TCA, and TCT. This observation applied to both ACX, rPB(8kD) and ACX, rXRCC1. From these results, we propose that the favored DNA motifs are WCW, ACC, and GCT (W is either A or T, Fig. 2b). This understanding is consistent with our observation that C:G to G:C editing is more efficient at HEK2, HEK4, and RNF2, which all had favored DNA motifs of ACA, ACT, and TCT, whereas EMX1, FANCF, and HEK3, with suboptimal DNA motifs of TCC and CCA, are less amenable to C:G to G:C editing. With a preference for WCW, ACC, and GCT, CGBEs most efficiently target 6 out of 16 possible motifs.

We previously observed a reduction in editing levels at position 4 relative to position 6 of HEK2 (Extended Data Fig. 1). Similarly, editing levels at position 3 and position 4 of HEK3 are lower than that at position 5 (Extended Data Fig. 2). Both observations suggested that our candidates might not inherit the exact base editing window of BE3¹. To elucidate the editing window of the CGBEs, we chose a genomic site that has an alternating 5'-W-C-W-C-W-C-3' sequence such that gRNAs can be designed with C's located either at every odd position or at every even position (Fig. 2c). We detected BE3 C:G to T:A editing between positions 1 and 9, with higher editing efficiency between positions 4 and 8. For both ACX, rPB(8kD) and ACX, rXRCC1, C:G to G:C editing was observed in a nine nucleotide window from positions 2 to 10. However, only at positions 5 and 6 was C:G to G:C editing the predominant and appreciable outcome.

Recognizing that simply removing UGI can potentially increase C:G to G:C editing at the expense of C:G to T:A editing², we next sought to quantify the effect of fusing rXRCC1 or rPB(8kD) to rAPOBEC-nCas9. Across 28 independent treatments using a variety of gRNAs (Fig. 3, Extended Data Fig. 5), BE3 was able to effect on average 4.4% C:G to G:C editing. Removing UGI from BE3 raised mean C:G to G:C editing to 11.5%. Fusing rPB(8kD) did not provide additional significant effect ($p > 0.05$). However, fusing rXRCC1 further raised mean C:G to G:C editing levels to 15.3% ($p = 0.01$). Meanwhile, the major byproduct C:G to T:A editing stayed between 6% and 9%. Our results suggest that UGI removal from BE3 promotes C:G to G:C editing; rXRCC1 fusion further increases C:G to G:C editing. Therefore, we narrowed down to rAPOBEC-nCas9-rXRCC1 as the preferred CGBE embodiment that effects C:G to G:C editing at $15 \pm 7\%$ efficiency in human cells, at a $68 \pm 14\%$ purity, within a two-nucleotide target window and WCW, ACC, and GCT target sequence contexts.

Here, we developed CGBEs that target cytosine within a specified window and convert it into guanine as a predominant editing product. In separate work, Liu and Koblan designed CGBE candidates by fusing UDG, UdgX, and polymerases with rAPOBEC-nCas9¹² (Extended Data Fig. 6). Both studies induce a C to U change via rAPOBEC and envision this U to be further converted to an abasic site; but the downstream resolution of this abasic site is distinctly different. As a direct opposite of our study, which seeks to induce BER and hence repair the targeted abasic site (Extended Data Fig. 7), Liu and Koblan designed candidates to maintain the abasic site while polymerases perform translesion synthesis on the opposite strand. A similar translesion polymerization hypothesis was also put forth by Gajula¹³. This proposed mechanism is unlikely to be the case for our CGBEs. Instead, we envisage that upon creation of an abasic site in the Cas9-induced R-loop, cellular UDG is displaced by APE1, after which XRCC1 recruits various BER components^{6,14} to repair the abasic site independently of the unedited opposite strand, giving rise to guanine as the major product. Subsequent DNA repair

converts the G:G mismatch to G:C. Such a hypothesis would be consistent with the tight binding of Cas9 on its target strand¹⁵, which renders that strand less accessible to other enzymes; the accessibility of the deaminated strand as a single-strand of the R-loop^{16,17}; the detrimental effect of UDG and UdgX on C:G to G:C editing¹² (Extended Data Fig. 6), which suggests that the persistence of an UDG-bound site or abasic site might act against the C:G to G:C reaction; and the C:G to G:C editing effected by our CGBEs with domains that do not have intrinsic polymerase activity but are key drivers of abasic site repair. While further mechanistic studies and development continue, CGBEs expand the growing suite of precise genome-editing tools that include CBEs^{1,2}, ABEs³, and prime editors¹⁶, together enabling the precise and efficient engineering of DNA for biological interrogation and disease correction.

Main References:

1. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420-424 (2016).
2. Nishida, K. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353**, aaf8729 (2016).
3. Gaudelli, N. M. et al. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* **551**, 464-471 (2017).
4. Komor, A. C. et al. 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).
5. Kim, H. S., Jeong, Y. K., Hur, J. K., Kim, J.-S., Bae, S. Adenine base editors catalyze cytosine conversions in human cells. *Nat. Biotechnol.* **37**, 1145-1148 (2019).
6. Wallace, S. S. Base excision repair: A critical player in many games. *DNA Repair* **19**, 14-26 (2014).
7. Mol, C. D. et al. Crystal Structure of Human Uracil-DNA Glycosylase in Complex with a Protein Inhibitor: Protein Mimicry of DNA. *Cell* **82**, 701-708 (1995).
8. Husain, I. et al. Specific inhibition of DNA polymerase β by its 14 kDa domain: role of single- and double-stranded DNA binding and 5'-phosphate recognition. *Nucleic Acid Res.* **23**, 1597-1603 (1995).
9. Litonjua, A. A. et al. Very important pharmacogene summary ADRB2. *Pharmacogenet. Genomics* **20**, 64-69 (2010).
10. Snoeckx, R. L. et al. GJB2 mutations and degree of hearing loss: a multicenter study. *Am. J. Hum. Genet.* **77**, 945-957 (2005).
11. Carrier, L., Mearini, G., Stathopoulou, K., Cuello, F. Cardiac myosin-binding protein C (MYBPC3) in cardiac pathophysiology. *Gene* **573**, 188-197 (2015).
12. Liu, D. R., Koblan, L. W. Cytosine to Guanine Base Editor. *World Intellectual Property Organization* (2018).
13. Gajula, K. S. Designing an Elusive C•G→G•C CRISPR Base Editor. *Trends Biochem Sci.* **44**, 91-94 (2018).
14. London, R. E. The structural basis of XRCC1-mediated DNA repair. *DNA Repair* **30**, 90-103 (2015).
15. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62-67 (2014).

16. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149-157 (2019).
17. Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L., Corn, J. E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat. Biotechnol.* **34**, 339-344 (2016).

Figures and legends:

Fig. 1. Initial screen of CGBE candidates for C:G to G:C editing. (a) CBEs like BE3 and BE4 predominantly convert C:G to T:A while CGBE aims to predominantly convert C:G to G:C. **(b)** CGBE candidates were designed in three orientations – ACX, AXC, and XAC, where X denotes the fused BER protein. **(c)** Seven candidates were selected for their high C:G to G:C editing at both HEK2 and HEK3. Targeted C's are in red. PAMs are underlined. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (one-way ANOVA against 'Untreated,' Dunn-Šidák; $n=3$, mean $\pm$ standard error).

Fig. 3. CGBE induces efficient C:G to G:C editing as the predominant product. (a) Removal of UGI from BE3 increases C:G to G:C editing (blue triangles); fusion of rXRCC1 further increases C:G to G:C editing. The major byproduct is C:G to T:A editing (orange circles). **(b)** Mean C:G to G:C editing/C:G to T:A editing ratio. Data includes all biological replicates across 15 genomic loci that reside within a WCW, ACC or GCT motif. * $p < 0.05$; *** $p < 0.001$ (two-tailed Student's T test; $n=28-45$ biological replicates).

Methods:

Constructs and cell culture

DNA amplification was done via PCR using Q5 Hot Start HiFi 2X Master Mix (NEB, M0494). All plasmids were constructed using either blunt end ligation or Gibson Assembly. For blunt end ligation, PCR products were first treated with DpnI (NEB, R0176) and T4 Polynucleotide Kinase (NEB, M0201) before being ligated to plasmid vectors using T4 DNA Ligase (NEB, M0202). NEBuilder HiFi DNA Assembly Master Mix (NEB, E2621) was used for Gibson Assemblies. hXRCC1 (pTXG-hXRCC1) and hLIG3 (pGEX4T-hLIG3) were gifts from Primo Schaer (Addgene plasmid # 52283 and # 81055 respectively). BE3 (pCMV-BE3) was a gift from David R. Liu (Addgene plasmid # 73021). The mutation R400Q is introduced to hXRCC1; N628K is introduced to hLIG3. rXRCC1, rLIG3, hPBs, and rPBs were obtained as human codon-optimized *de novo* synthesized gene fragments (Twist Biosciences). All other oligonucleotides used in the study were *de novo* synthesized (IDTDNA).

Following transformation of constructs and bacterial outgrowth, gRNA plasmids and base-editor plasmids were purified using the PureYield Plasmid Miniprep System (Promega, A1223) and Plasmid Plus Maxi Kit (Qiagen, 12965) respectively. Both plasmid kits' protocols include endotoxin removal steps. Plasmids were Sanger sequenced for sequence verification.

Cell culture, transfection, and genomic DNA harvest

HEK293AAV cells (Agilent, 240073) were maintained in DMEM (Thermo Fisher, 10569-010) supplemented with 10% HI FBS (Thermo Fisher). 30,000 HEK293AAV cells were added to each well of a 48 well plate one day before transfection. For each well, 750 ng of base-editor plasmid, 250 ng of gRNA plasmid, and 20 ng of GFP plasmid were transfected into these cells using Lipofectamine 3000 (Invitrogen, L3000015) according to the manufacturer's protocol. The media were replaced with fresh media one day after transfection. Three days after transfection, media were removed, cells were washed with 50 μ L PBS, pH 7.2 (Thermo Fisher, 20012-027), and genomic DNA was extracted using 50 μ L of Quick Extract DNA Extract Solution (Lucigen, QE09050) per well according to manufacturer's protocol. All sample sizes indicate biological replicates.

Sequencing of genomic DNA

Sites of interest were prepared for high-throughput sequencing via two PCR amplifications – the first PCR amplifies the region of interest while the second PCR adds appropriate sequencing barcodes. Primers for the first PCR are included in Extended Data Table 3. Primers for the second PCR are based off Illumina adaptors. Amplicons from the second PCR were then pooled and gel extracted (Promega, A9282) to make the final library, which was quantified via Qubit fluorometer (Thermo Fisher) and sequenced on an Illumina iSeq 100 according to the manufacturer's protocol. The resultant FASTQ files were analyzed using CRISPResso2¹⁸. All sample sizes indicate biological replicates.

Statistical analyses were performed on Matlab. Weblogos were created using Weblogo 3¹⁹.

Data availability: High-throughput sequencing data will be deposited to the NCBI Sequence Read Archive database. Plasmid encoding ACX, rXRCC1 will be available on Addgene.

Method references:

18. Clement, K. et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224-226 (2019).
19. Crooks, G. E., Hon, G., Chandonia, J.-M., Brenner, S. E., Weblogo: A sequence logo generator. *Genome Res.* **14**, 1188-1190 (2004). Website: <http://weblogo.threeplusone.com/>

Acknowledgements:

We thank Anna Traczyk for helpful discussions; Eddie Keng for help with HTS; Hannah Nicholas, Nicholas Ong, and Hong-Ting Prekop for editing; and Ke Guo and Daryl Lim for general administrative support. This work is supported by Agency for Science, Technology and Research (A*STAR) and Industrial Alignment Fund Pre-Positioning (IAF-PP) grant H17/01/a0/012. Work by P.D.R. and W.L.C. is further supported by Advanced Manufacturing and Engineering (AME) Programmatic grant A18A9b0060.

Author contributions:

L.C. made the initial discovery. L.C., J.E.P., and W.L.C. designed the research. L.C., J.E.P., and P.P. performed the experiments. P.D.R. performed biological replicates. Y.T.C. and S.N.M. cloned plasmids. L.C., J.E.P., P.P., and W.L.C. analyzed data. L.C. and W.L.C. wrote the manuscript with input from other authors. W.L.C. supervised the research.

Competing interest declaration: L.C. and W.L.C. have filed a patent application based on this work.

Additional information:

Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to W. L. C.

284

285 **Extended Data Fig. 2. Initial screen of CGBE candidates on HEK3 at (a) position 5, (b)**
 286 **position 4, and (c) position 3.** The seven candidates selected for further studies are marked
 287 with †. Targeted C is in red. PAM is underlined. *p < 0.05; **p < 0.01; ***p < 0.001 (one-
 288 way ANOVA against 'Untreated,' Dunn-Šidák; n=3, mean ± standard error).

Extended Data Fig. 3. Further characterization of shortlisted CGBE candidates. (a) CGBE candidates effect C:G to G:C mutations at EMX1, HEK4, RNF2, and FANCF. C:G to G:C editing is the main edit at HEK4 and RNF2; C:G to T:A editing is the main edit at FANCF and EMX1. (one-way ANOVA against 'BE3,' Dunn-Šidák; $n=3$, mean $\pm$ standard error) **(b)** CGBE editing at disease-associated genes ADRB2, GJB2, MYBPC3, and GAL 292. Note that ADRB2 contains naturally occurring polymorphism in HEK293AAV cells, and hence this data is not included in Fig 3. (one-way ANOVA against 'Untreated,' Dunn-Šidák; $n=5$ for ADRB2 and MYBPC3; $n=4$ for GJB2; $n=1$ for GAL 292, mean $\pm$ standard error) **(c)** Mean C:G to G:C editing as a percent of all reads across 16 NCN sites. CGBEs increase C:G to G:C editing by three to four fold compared to BE3, across all possible NCN sequences. gRNA sequences are included in Extended Data Table 2. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (one-way ANOVA against 'BE3,' Dunn-Šidák; $n=32$, mean $\pm$ standard error).

a ADRB2

WT	C ₁	C ₂	C ₃	T ₄	T ₅	T ₆	C ₇	C ₈	T ₉	G ₁₀	C ₁₁	G ₁₂	T ₁₃	G ₁₄	A ₁₅	C ₁₆	G ₁₇	T ₁₈	C ₁₉	G ₂₀
A	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.1	0.1	99.8	0.0	0.3	0.0	0.0	0.2
C	99.9	99.9	99.8	0.2	0.2	0.1	54.5	99.4	0.1	0.1	99.8	0.4	0.1	0.1	0.0	99.9	0.1	0.4	99.9	0.1
G	0.0	0.0	0.0	0.1	0.1	0.1	44.9	0.0	0.1	99.7	0.0	99.5	0.2	99.7	0.2	0.0	99.6	0.3	0.0	99.5
T	0.1	0.1	0.1	99.6	99.7	99.9	0.6	0.5	99.7	0.1	0.2	0.0	99.5	0.0	0.1	0.1	0.1	99.3	0.1	0.2

XRCC	C ₁	C ₂	C ₃	T ₄	T ₅	T ₆	C ₇	C ₈	T ₉	G ₁₀	C ₁₁	G ₁₂	T ₁₃	G ₁₄	A ₁₅	C ₁₆	G ₁₇	T ₁₈	C ₁₉	G ₂₀
A	0.1	0.0	0.1	0.0	0.1	0.0	0.7	0.1	0.0	0.1	0.0	0.3	0.0	0.1	99.8	0.0	0.2	0.0	0.0	0.3
C	99.6	99.9	99.6	0.2	0.1	0.1	10.6	99.2	0.1	0.1	99.8	0.2	0.1	0.1	0.0	99.9	0.0	0.3	99.9	0.1
G	0.0	0.0	0.0	0.2	0.1	0.1	87.6	0.1	0.1	99.8	0.0	99.4	0.3	99.8	0.1	0.0	99.6	0.2	0.0	99.5
T	0.3	0.1	0.3	99.6	99.7	99.8	1.1	0.6	99.8	0.1	0.2	0.0	99.6	0.0	0.1	0.1	0.1	99.4	0.1	0.2

rPB	C ₁	C ₂	C ₃	T ₄	T ₅	T ₆	C ₇	C ₈	T ₉	G ₁₀	C ₁₁	G ₁₂	T ₁₃	G ₁₄	A ₁₅	C ₁₆	G ₁₇	T ₁₈	C ₁₉	G ₂₀
A	0.1	0.0	0.1	0.0	0.1	0.0	1.0	0.3	0.1	0.1	0.0	0.2	0.1	0.1	99.7	0.0	0.3	0.0	0.0	0.2
C	99.5	99.7	99.1	0.2	0.1	0.1	25.4	98.5	0.1	0.1	99.8	0.3	0.1	0.1	0.0	99.9	0.1	0.4	99.9	0.1
G	0.0	0.0	0.0	0.1	0.1	0.1	71.4	0.1	0.1	99.8	0.0	99.5	0.2	99.8	0.2	0.0	99.5	0.3	0.0	99.6
T	0.4	0.3	0.7	99.6	99.7	99.8	2.2	1.1	99.7	0.1	0.2	0.0	99.6	0.1	0.1	0.1	0.1	99.3	0.1	0.1

b MYBPC3

WT	G ₁	A ₂	T ₃	A ₄	G ₅	A ₆	C ₇	C ₈	T ₉	G ₁₀	T ₁₁	G ₁₂	T ₁₃	G ₁₄	C ₁₅	A ₁₆	T ₁₇	G ₁₈	G ₁₉	A ₂₀
A	0.1	99.8	0.7	99.8	0.1	99.9	0.1	0.0	2.3	1.0	0.7	7.2	0.1	0.1	0.0	99.9	1.4	3.0	0.2	99.9
C	0.0	0.0	0.0	0.0	0.1	0.0	99.7	99.6	0.1	0.0	0.0	0.0	0.0	0.0	99.2	0.0	0.0	0.0	0.0	0.0
G	99.8	0.1	0.0	0.1	99.8	0.1	0.1	0.3	0.0	99.0	0.0	92.8	0.0	99.9	0.4	0.0	0.0	97.0	99.8	0.1
T	0.0	0.0	99.3	0.1	0.0	0.0	0.1	0.1	97.6	0.0	99.3	0.0	99.8	0.0	0.3	0.1	98.5	0.0	0.0	0.0

XRCC	G ₁	A ₂	T ₃	A ₄	G ₅	A ₆	C ₇	C ₈	T ₉	G ₁₀	T ₁₁	G ₁₂	T ₁₃	G ₁₄	C ₁₅	A ₁₆	T ₁₇	G ₁₈	G ₁₉	A ₂₀
A	0.2	99.7	0.9	99.8	0.2	99.8	5.3	2.0	2.1	1.1	0.6	7.2	0.1	0.1	0.0	99.9	1.1	3.0	0.2	99.9
C	0.2	0.1	0.1	0.0	0.1	0.0	69.4	94.0	0.1	0.1	0.0	0.0	0.1	0.0	99.3	0.0	0.0	0.0	0.0	0.0
G	99.6	0.1	0.0	0.1	99.7	0.1	13.9	1.3	0.0	98.8	0.0	92.7	0.0	99.9	0.4	0.0	0.0	97.0	99.8	0.1
T	0.0	0.0	99.0	0.1	0.0	0.1	11.3	2.7	97.8	0.0	99.3	0.1	99.8	0.0	0.3	0.1	98.8	0.0	0.0	0.0

rPB	G ₁	A ₂	T ₃	A ₄	G ₅	A ₆	C ₇	C ₈	T ₉	G ₁₀	T ₁₁	G ₁₂	T ₁₃	G ₁₄	C ₁₅	A ₁₆	T ₁₇	G ₁₈	G ₁₉	A ₂₀
A	0.3	99.7	0.9	99.7	0.2	99.8	5.2	2.4	2.2	1.2	0.6	7.6	0.1	0.1	0.1	99.9	1.2	2.9	0.2	99.9
C	0.1	0.1	0.1	0.0	0.1	0.0	71.1	92.9	0.1	0.1	0.1	0.0	0.0	0.0	99.2	0.0	0.0	0.0	0.0	0.0
G	99.6	0.1	0.0	0.1	99.7	0.1	12.3	1.4	0.0	98.7	0.0	92.3	0.1	99.9	0.4	0.0	0.0	97.1	99.8	0.1
T	0.0	0.1	99.0	0.2	0.0	0.1	11.3	3.3	97.7	0.0	99.3	0.1	99.8	0.0	0.3	0.1	98.8	0.0	0.0	0.0

Extended Data Fig. 4. Representative data of CGBE editing at (a) ADRB2 and (b) MYBPC3. WT denotes wild-type untreated cells; XRCC denotes ACX, XRCC1; rPB denotes ACX, rPB(8kD). Note that ADRB2 contains naturally occurring polymorphism in HEK293AAV cells.

Extended Data Fig. 5. ACX, rXRCC1 is the best performer out of shortlisted CGBE candidates. (a) C:G to G:C editing (blue triangles) vs. C:G to T:A editing (orange circles) as percent of all reads across gRNAs used in this study. All biological replicates are included except those targeting the 10 suboptimal C:G to G:C base editing motifs (Fig. 2a and Fig. 2b) and ADRB2 due to naturally occurring polymorphism (Extended Data Fig. 3b). (b) Ratio of C:G to G:C editing to C:G to T:A editing across gRNAs used in this study. Only BE3 (no UGI) and ACX, rXRCC1 give a significantly higher ratio of C:G to G:C editing/C:G to T:A editing. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (one-way ANOVA against 'BE3,' Dunn-Šidák, $n=20-45$ biological replicates).

Extended Data Fig. 6. Comparison of CGBEs at (a) HEK2, (b) FANCF, and (c) RNF2. Since datasets were generated independently and in different cell types, comparisons should be made only against BE3 common to the two studies. Fusion of base excision enzymes such as UDG and UdgX decreases C:G to G:C editing beyond BE3 at 2 of 3 sites ($n=1$; Liu and Koblan, 2018). Fusion of base excision repair enzymes such as rXRCC1 increases C:G to G:C editing beyond BE3. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (one-way ANOVA against 'BE3,' Dunn-Šidák; $n=3$, mean $\pm$ standard error; this study).

Extended Data Fig. 7. Distinct strategies for CGBE design. This study employs a CGBE design strategy (left) where Cas9 is fused to protein(s) involved in repairing uracil-containing or abasic sites (AP). The activities of these BER proteins are expected to convert the AP to G before the nucleotide on the opposite strand is converted from G to C. In contrast, the strategy employed by Liu and Koblan (right) seeks to maintain the abasic site (AP) throughout a translesion synthesis envisioned to occur on the opposite strand. The AP is repaired after the nucleotide on the opposite strand is converted from G to C. In other words, this study employs proteins that repair and not maintain abasic sites whereas Liu and Koblan employ proteins that maintain and not repair abasic sites.

From	To	% of SNPs in ClinVar	Equivalent to		Base-editing route for correction
A	T	7	T	A	A to G (ABE), G to C (CGBE), C to T (CBE)
	G	48		C	ABE
	C	15		G	C to G (CGBE), G to A (CBE)
G	A	6	C	T	CBE
	T	14		A	A to G (ABE), G to C (CGBE)
	C	11		G	CGBE

Extended Data Table 1. CGBEs enable treatment avenues to previously unaddressable SNPs that represent half of all known human disease-associated SNPs. CBE enables treatment to 48% of all known disease-associated SNPs, while ABE enables treatment to 6%. CGBEs effect primarily C:G to G:C and G:C to C:G changes (*yellow highlight*) that can correct 11% of disease-associated SNPs. In combination with CBEs or ABEs, CGBEs effect secondarily G to T, C to A, A to C, and T to G edits (*green highlight*). With CBEs, ABEs, and CGBEs, the remaining 7% of SNPs (A to T and T to A) can also be corrected (*orange highlight*).

Name	Target Sequences and PAMs
EMX1	GAGT <u>CC</u> GAGCAGAAGAAG AGG
FANCF	GGAAT <u>CC</u> CTTCTGCAGCACCT TGG
HEK2-1	GAACACAAAGCATAGACTGC GGG
HEK2-2	GAGCA <u>CC</u> ACACCCTAAACTAT TGG
HEK2-3	GGAAACGGATAGTTCTGAA AGG
HEK2-4	CTTA <u>ACT</u> ATTTGTATTCCACT TGG
HEK2-5	CTTCC <u>CA</u> AGTGAGAAGCCAG TGG
HEK2-6	CCAGCCCGCTGGCCCTGTAA AGG
HEK2-7	CATTCCGTTATTTTACATAT TGG
HEK2-8	GTTTC <u>CT</u> TTACAGGGCCAGC GGG
HEK2-9	ATACGCACAGTTTGACAGAT GGG
HEK2-10	GCTGGCCCTGTAAAGGAACT TGG
HEK2-11	GCATGC <u>CG</u> TGTGTGTTAAGCT TGG
HEK2-12	TTGGGCTGCAGTAACTTGA AGG
HEK2-13	TCTTTCAAGCAGGTGATTAC AGG
HEK2-14	AGTTTCCTTTACAGGGCCAG CGG
HEK2-15	GAGGT <u>CG</u> TGGCTGAGCACA AGG
HEK2-16	GGCCTC <u>T</u> ATTGTTGGTAGAAT TGG
HEK3	GGCC <u>CC</u> AGACTGAGCACGTGAT TGG
HEK4	GGCA <u>CT</u> GC GGCTGGAGGTGG GGG
RNF2-1	GTCATC <u>CT</u> TAGTCATTACCTG AGG
RNF2-2	<u>CACACAC</u> ACTTAGAATCTGT GGG
RNF2-3	<u>ACACACAC</u> ACTTAGAATCTGT TGG
GJB2	GGACACGAAGATCAGCTGCAG GGG
ADRB2	CCCTTT <u>C</u> CTGCGTGACGTCG TGG
MYBPC3	CCCTTT <u>C</u> CTGCGTGACGTCG TGG
GAL 292	GAAGTCGTTGTCAAACAGGA AGG

Extended Data Table 2. Target protospacer sequences used in this study. Targeted C's are underlined. PAMs are in bold.