Glycosylase base editors enable C-to-A and C-to-G base changes

Editorial summary

New base editors change C to A in bacteria and C to G in mammalian cells.

Glycosylase base editors enable C-to-A and C-to-G base changes

Dongdong Zhao[1,3,#], Ju Li[2,#], Siwei Li[1,3,#], Xiuqing Xin[1,3,4], Muzi Hu[1,3,5], Marcus A. Price[6], Susan J. Rosser[6], Changhao Bi[1,3*]& Xueli Zhang[1,3*]

[1]Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China

[2]College of Life Science, Tianjin Normal University, Tianjin, China

[3]Key Laboratory of Systems Microbial Biotechnology, Chinese Academy of Sciences, Tianjin, China

[4]College of Biotechnology, Tianjin University of Science and Technology, Tianjin, China

[5]School of Biological Engineering, Dalian Polytechnic University, Dalian, China

[6]Centre for Synthetic and Systems Biology and UK Centre for Mammalian Synthetic Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

[#]These authors contributed equally to this work

*Corresponding authors

Email: bi_ch@tib.cas.cn; zhang_xl@tib.cas.cn

Current base editors (BEs) catalyze only base transitions (C-to-T and A-to-G), but cannot produce base transversions. Here we present BEs that cause C-to-A transversions in E. coli and C-to-G transversions in mammalian cells. Our glycosylase base editors (GBEs) consist of a Cas9 nickase, a cytidine deaminase and a Uracil-DNA glycosylase (Ung). Ung excises the U base created by the deaminase, creating an apurinic/apyrimidinic (AP) site that initiates the DNA repair process. [AU: unclear how this results in a transversion. Can this be

clarified?]In *E. coli*, we used the activation-induced cytidine deaminase (AID) to construct AID-nCas9-Ung and found that it converts C to A with an average specificity of 93.8±4.8% [AU: s. question below in the main text]. For use in mammalian cells, we replaced AID with rat APOBEC1 (APOBEC-nCas9-Ung). Testing APOBEC-nCas9-Ung at 30 endogenous sites, we observed C-to-G conversions with a high specificity at the 6th position of the protospacer [AU: OK?] with efficiencies between 92.15% and 29.65%[AU: give range of observed C-to-G conversion efficiencies] and a product purity of 50-90%. APOBEC-nCas9-Ung supplements the current A and C base editors and could be used to target G/C disease-causing mutations.

[Edited Abstract ok? Word count limit 160]

The CRISPR/Cas9 system is able to recognize target sequences based purely on nucleotide sequences, and it has been adapted for genome editing in both eukaryotes and prokaryotes[1-5]. Recently, a new family of CRISPR-based genome editing methods, named base editor techniques, was developed. Initially, for precise cytosine (C) to thymine (T) editing (CBE) without the use of an editing template, Komor *et al.*[6] and Nishida *et al.*[7] fused rat cytidine deaminase APOBEC1 to the N-terminus of *Streptococcus pyogenes* Cas9 (SpCas9) variants or the PmCDA1 activation-induced cytidine deaminase (AID) ortholog from the sea lamprey to the C-terminus of SpCas9 variants (termed Target-AID). More recently, the base editing capacity was expanded to adenine (A) to guanine (G) adenine base editing (ABE) by nCas9 (D10A)-fused adenosine deaminase[8]. These techniques have been a breakthrough for precise base editing in the chromosomes of various species[9-15].

Since current base editing techniques can only facilitate C-to-T editing (CBE) and A-to-G editing (ABE), a complete base editing technique for converting any base to any other base is highly desirable. There have been several reports observing C mutating to any other base during error-prone DNA repair[7, 9, 15-18]. If the random mutation events could be enhanced during DNA replication, then the genome editing outcome would be significantly changed. If correctly harnessed and controlled, C

might be specifically converted to any other base, further enhancing the base editing toolbox.

To examine the frequency of different point mutations induced by BE, the effect of nCas9-AID was determined in *Escherichia coli* wild-type strains K12 MG1655 and ATCC 8739. The genome locus *lacZ* (1444-1463) in MG1655 was edited by nCas9-AID, and cells were spread on plates containing IPTG and X-gal. A total of 83.0% ± 2.2% of the obtained colonies were white, indicating that DNA was edited at the target locus. When additional loci in MG1655 and ATCC 8739 were edited, C-to-T and C-to-A mutations were observed in the target editing region. While no C-to-G mutations were detected, all 7 tested target loci in MG1655 had similar C-to-T and C-to-A conversion specificity; both were approximately 50% (Fig. 1A). The frequency of C-to-A mutations was slightly higher than that of C-to-T mutations in the 8739 strain (Fig. 1B). However, the results were generally similar to BE rates reported in mammalian cells, where the error-prone repair mechanism had a significant effect[7].

Uracil-DNA glycosylase (Ung) removes U from the DNA double strand and initiates base-excision repair (BER)[19]. In *E. coli*, Banno *et al.*[11] and Zheng *et al.*[20] utilized uracil DNA glycosylase inhibitor protein (UGI) to repress Ung and only observed C-to-T mutations. To examine the cause of the C-to-A editing events, the *ung* gene was deleted in *E. coli* MG1655. In the resulting strain, the U-G base pair could not be processed by BER following deamination of C. Five loci in *lacZ* (984-1003, 1431-1450, 1444-1463, 1608-1627 and 2293-2312) in MG1655Δ*ung* were targeted by nCas9-AID, and all five loci exhibited C-to-T mutations (Extended Data Figure 1A). Thus, the C-to-A editing events were most likely caused by DNA repair initiated by Ung in *E. coli*.

While dCas9-AID only eliminates the amine group of Cs within the target editing region on one DNA strand, nCas9-AID also cleaves the other strand at a nearby location. To determine whether the pattern of DNA damage is the key factor for C-to-A mutations, nCas9-AID was replaced with dCas9-AID to perform base editing

under the same conditions. Four loci within *lacZ* (984-1003, 1431-1450, 1444-1463 and 1608-1627) were tested, and all five loci exhibited C-to-T mutations. No C-to-A mutations were observed (Extended Data Figure 1B). This result supports our hypothesis that the specific DNA damage by nCas9-AID was the cause of Ung-initiated C-to-A conversions. This repair mechanism could provide a profound and rewarding goal to research, which, once illustrated, might be applied to plant or animal cells for specific C-to-A base editing.

Since C-to-A conversion was caused by Ung-initiated DNA repair of damage caused by nCas9-AID, we constructed enzyme complexes of fused nCas9, AID, and Ung. These fusions perform a series of functions, including specific DNA binding, cleaving the amine group from C, and excising U to create AP sites (Fig. 2A). To obtain an active fusion complex, three patterns for fusing these enzymes together were attempted: Ung-nCas9-AID, nCas9-Ung-AID and nCas9-AID-Ung. However, only Ung-nCas9-AID was successfully constructed. The other patterns of fusion were unable to be obtained due to random mutations appearing with these fusions, indicating that these constructs were toxic.

The subsequent proximity of Ung to the U base created by AID allows efficient creation of an AP site and prevents C-to-T editing events facilitated by DNA replication with a U base. Four loci within *lacZ* (984-1003, 1431-1450, 1444-1463 and 1608-1627) were targeted for editing. Ung-nCas9-AID was able to yield C-to-A conversions with a high specificity at all four loci (Fig. 2B), 97.6%95.4±4.0% at *lacZ* (984-1003), 97.2%90.3±10.0% at *lacZ* (1608-1627), 93.2%80.6±6.9% at *lacZ* (1444-1463) and 87.3%82.5±7.5% at *lacZ* (1431-1450) in the tested wild-type *E. coli* MG1655 strains, with an average specificity of 93.8±4.8% **[AU: Does this mean that 93.8% of all sites showed a C-to-A conversion or that of the edited sites 93.8% had a C-to-A conversion? Please clarify. If the latter what was the overall rate of C to A conversion?]** . To test whether Ung-nCas9-AID had off-target effects in *E. coli*, ten loci containing the most similar sequences of either *lacZ* (984-1003) or *lacZ* (1444-1463) in the edited MG1655 cells were amplified by PCR for sequence analysis. The sequencing results matched that of the original *E. coli* MG1655, suggesting that

no significant off-target events occurred at any of these loci (Supplementary Table S1).

During the development process of CBE editing in mammalian cells, C-to-G base conversion was reported to be a byproduct of C-to-T, most likely resulting from DNA repair of AP sites formed by direct elimination of Us by Ung. To decrease the frequency of C-to-G conversion, the protein UGI was employed to repress the function of Ung and protect the formation of AP sites from the resulting Us in DNA[6, 7]. In this work, to achieve specific conversion of C-to-G in mammalian cells, we proposed a route that was opposite to current C-to-T editors, which specifically create AP sites from Cs instead of from Us. To carry out the strategy, we constructed a glycosylase base editor, APOBEC-nCas9-Ung, with the proven fusion strategy from *E. coli* experiments. Eighteen positions of protospacers (C1, C3, C4, C5, C6, C7, C9, C10, C11, C12, C13, C14, C15, C16, C17, C18, C19, and C20) from 10 different loci (EMX1-site2，EMX1-site3, FANCF-site1, FANCF-site2, FANCF-site3, RNF2, RNF2-site1, PSMB2-1, VEGFA-site2, and PDL1) on HEK293T chromosome were selected for base editing using APOBEC-nCas9-Ung and the control, APOBEC-nCas9-UGI. The results are illustrated in Fig. 3. While APOBEC-nCas9-UGI generally converted C to T at the C3 to C7 positions of the target loci, APOBEC-nCas9-Ung had a much narrower target editing region; it acted mainly on the C6 position and specifically converted the 6th position from C to G. To further determine the editing efficiency and specificity of APOBEC-nCas9-Ung, 27 more loci containing C6 were selected for GBE editing. As shown in Fig. 4, the control APOBEC-nCas9-UGI generally edited by conversion of the Cs in the target loci to Ts. However, the GBE editor was able to obtain C-to-G conversions with a high specificity: more than 90% at two tested loci (VISTA hs267, RP11-177B4-2), 80% to 90% at seven loci (FANCF-site4, RP1-97D16, HIRA-1, EMX1-site5, NIBAN1, RP11-177B4-1, PSMB2-1), 70% to 80% at eight loci (HEK4 OT2, HEK2 OT1, VEGFA-site3, EMX1-site4, EMX1-site6, EMX1-site7, RP11-177B4-3, RP11-177B4-4), 50% to 70% at six loci. Seven loci resulted in C-to-G conversion specificities lower than 50%. The results indicated that the GBE editor,

APOBEC-nCas9-Ung, could specifically convert C to G in mammalian cells and mainly at the 6th position of target loci, which is unlike other techniques.

Indel (insertions and deletions) data was also obtained by the deep sequencing. Thirty loci were tested and the observed indel frequencies of GBE ranged from 0.03 ±0.03% to 12.18±3.32% with an average of 2.71±0.57% (Fig. 4), which was similar to that of APOBEC-nCas9-UGI (Supplementary Table S2), which ranged from 0.17±0.10% to 16.58±0.85% with an average of 2.59±0.28%. To analyze the off-target frequency of APOBEC-nCas9-Ung, ten loci containing the most similar sequences of RP11-177B4-3, PSMB2-1 and EMX1-site5 in edited HEK293T cells were amplified by PCR for deep sequence analysis. The results showed that the mutation rates in off-target sites by APOBEC-nCas9-Ung were in the range of 2% to 3%, and the rates for APOBEC-nCas9-UGI were also in the same range (Supplementary Table S3). Thus, the off-target effect of APOBEC-nCas9-Ung was similar to that of the BE editor APOBEC-nCas9-UGI.

It was revealed that there is a correlation of C-to-G conversion efficiencies between the two editors, APOBEC-nCas9-UGI and APOBEC-nCas9-Ung. At the loci where APOBEC-nCas9-UGI had relatively low C-to-G conversion, that of APOBEC-nCas9-Ung was also relatively low, though the C-to-G conversion efficiency of APOBEC-nCas9-Ung was higher than APOBEC-nCas9-UGI. One plausible explanation might be that the Ung enzyme could not contact the Us in these loci probably due to space limitation, so that Us of these loci could only go through DNA replication and be converted to Ts. The deep sequencing data also revealed that the average editing efficiency of C-to-A conversions by GBE was 2.67±0.31%, which was low, and not a major outcome of GBE editing (Supplementary Table S4).

With the high specificity to act mainly on one position, GBE could be applied in scenarios in which one C needs to be edited in the context of other Cs. Although the target editing region is narrowed, using Cas9 homologs[21] or variants with different PAM sequence requirements[22, 23], GBE could be applied to more loci for position-specific editing.

With the establishment of specific C-to-A editing, it is possible to fill the gaps of

the current BE techniques and design a method to achieve any base editing (NBE) with a one-, two- or three-step process. Although multiple steps are required for some types of base editing, we think this editing strategy is still useful with DNA repair templates for microbes with low recombination efficiency. As illustrated in Fig. 5, C could be converted to A or T specifically with nCas9-AID, and subsequently, A could be converted to G with nCas9-TadA, which reveals a complete set of BE methods where C could be converted to all three other bases. Since G is paired to C, G could also be converted to any base with the same method (Fig. 5). The only difference is the need to design and facilitate the base editing process for the C on the opposite strand that is paired with the target G.

Starting from A, nCas9-TadA could be used for conversion of A-to-G, from which Ung-nCas9-AID editing could be designed based on the opposite strand to convert its paired base C-to-A, which could cause G-to-T conversion on the positive strand. Subsequently, the A on the antisense strand could be converted to G with nCas9-TadA, converting the complimentary T to C on the sense strand. Thus, the A-T pair could be converted to any other bases with up to three steps (Fig. 5).

The NBE strategy was utilized for modification of bases in *lacZ* (2293-2312) and *lacZ* (2635-2654) in *E. coli* MG1655 (Fig. 5). At target *lacZ* (2293-2312), C was mutated to T using CBE with an efficiency of 44.1±10.6%. On the opposite strand, C was mutated to A using GBE at an efficiency of 92.5±6.6%; subsequently, A was mutated to G by ABE with an efficiency of 35.8±5.2%. At target *lacZ* (2635-2654), T was converted to C by ABE on the complementary strand with an efficiency of 28.5±9.9%, and C was mutated to A by a subsequent GBE with an efficiency of 89.3±11.1%. Last, A was mutated to G by ABE with an efficiency of 38.3±7.6%. The results indicated that all four DNA bases could be converted, *in situ* on the *E. coli* chromosome, to any other base by the NBE method. With the same strategy, we attempted to combine the C-to-G specific GBE base editor with C-to-T and A-to-G conversion techniques to carry out a multistep NBE in mammalian cells. However, low editing efficiencies were observed from multiple consecutive rounds of base editing in HEK293T cells, suggesting that the NBE strategy might only be feasible for

microbes (Supplementary Figure 1).

To test whether the *E. coli*-based Ung-nCas9-AID could function in mammalian cells, mammalian-based APOBEC-nCas9-Ung could work in prokaryotic cells. We constructed an expression cassette of Ung-nCas9-AID with mammalian codons and tested it in HEK293T cells, as well as an expression cassette of APOBEC-nCas9-Ung with *E. coli* codons and tested it in *E. coli.* However, both enzyme complexes failed to edit the genomes of their hosts (Supplementary Figure 2). These results indicated that it was difficult to keep the activity of such a large complex by a direct transition of preference of codon usage. A subtler optimization procedure might be needed to complete such a task.

In this work, a novel GBE technique was established to specifically convert C to G. First, the editing mechanism of GBE is different from previously reported base editors. Current cytosine base editing techniques employ UGI to repress the function of Ung and increase editing events via the DNA replication route. However, with Ung-nCas9-AID and APOBEC-nCas9-Ung, the fused Ung was able to efficiently create an AP site, thus facilitating C-to-G base editing through DNA repair. Second, the GBE technique is the first base editing technique that allows specific C-to-G conversion, a base transversion that was not performed by previous base editors. Third, APOBEC-nCas9-Ung demonstrated a high position specificity, which mainly edited the 6th C of an N20 sequence. This is quite different from previous editors, which normally have a wide target editing region of an N20 sequence. With these novel functions, the GBE technique could expand the BE toolbox and directly impact biological studies in which C/G could be converted back and forth with high specificity in mammalian cells. In terms of applications, there are approximately three thousand G/C SNPs that cause human diseases, which is approximately 11% of the total[8]. With the high G/C conversion specificity and position specificity, APOBEC-nCas9-Ung holds great potential to be developed as a medical treatment for some G/C SNP diseases.

A novel genome editing technique referred to as the prime editor (PE)[24] was developed during the revision process of this paper. PEs employ a prime editing guide

RNA (pegRNA) as a reverse transcription template, and the resulting DNA sequence is incorporated into the genome after the editing process. We think that while PRIME exploits a brand-new method of genome editing, there are still some aspects that need to be improved for a mature method. A major problem for PRIME is that the editing efficiency was low; the efficiency for most tested loci was below 10% when one nick was formed by nCas9, as was the case for PE1 and PE2 editing strategies. The efficiency was improved to some extent when both DNA strands were nicked, which was the case for PE3, which introduced a substantial amount of indels. Even worse, the nickase can remain active on one or both of the DNA strands in some cases when any of the two N20s used was not covered by the editing, which could induce various types of DNA repair, recombination or lethal effects in the edited cell. Compared with PRIME editing, base editing techniques including GBE are more mature, have a higher editing efficiency, are easier to perform and do not make double strand breaks in genomic DNA. Currently, BE is much closer as a potential medical treatment for genomic diseases caused by G/C SNPs.

References

1. Jinek, M. et al. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).
2. Fujii, W., Kawasaki, K., Sugiura, K. & Naito, K. Efficient generation of large-scale genome-modified mice using gRNA and CAS9 endonuclease. *Nucleic acids research* **41**, e187-e187 (2013).
3. Hwang, W.Y. et al. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature Biotechnology* **31**, 227-229 (2013).
4. Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014).
5. Jiang, Y. et al. Multigene Editing in the *Escherichia coli* Genome via the CRISPR-Cas9 System. *Applied and environmental microbiology* **81**, 2506-2514 (2015).
6. 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).
7. K, N. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353** (2016).
8. Gaudelli, N.M. et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464 (2017).
9. Ma, Y. et al. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic

diversification in mammalian cells. *Nature Methods* **13**, 1029 (2016).
10. Zhang, Y. et al. Programmable base editing of zebrafish genome using a modified CRISPR-Cas9 system. *Nature Communications* **8**, 118 (2017).
11. Banno, S., Nishida, K., Arazoe, T., Mitsunobu, H. & Kondo, A. Deaminase-mediated multiplex genome editing in *Escherichia coli*. *Nature Microbiology* **3**, 423-429 (2018).
12. Li, C. et al. Expanded base editing in rice and wheat using a Cas9-adenosine deaminase fusion. *Genome Biology* **19**, 59 (2018).
13. Ryu, S.-M. et al. Adenine base editing in mouse embryos and an adult mouse model of Duchenne muscular dystrophy. *Nature Biotechnology* **36**, 536 (2018).
14. Shimatani, Z. et al. Inheritance of co-edited genes by CRISPR-based targeted nucleotide substitutions in rice. *Plant Physiology & Biochemistry* (2018).
15. Wang, Y. et al. MACBETH: multiplex automated *Corynebacterium glutamicum* base editing method. *Metabolic Engineering* **47**, 200-210 (2018).
16. Hess, G.T. et al. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nature Methods* **13**, 1036 (2016).
17. Komor, A.C., Badran, A.H. & Liu, D.R. CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. *Cell* **168**, 20 (2016).
18. Kim, K. et al. Highly efficient RNA-guided base editing in mouse embryos. *Nature Biotechnology* **35**, 435 (2017).
19. Lindahl, T. An N-Glycosidase from *Escherichia coli* That Releases Free Uracil from DNA Containing Deaminated Cytosine Residues. *Proceedings of the National Academy of Sciences of the United States of America* **71**, 3649-3653 (1974).
20. Zheng, K. et al. Highly efficient base editing in bacteria using a Cas9-cytidine deaminase fusion. *Communications Biology* **1**, 32 (2018).
21. Ran, F.A. et al. In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186 (2015).
22. Kleinstiver, B.P. et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481-485 (2015).
23. Hu, J.H. et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57 (2018).
24. Anzalone, A.V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149-157 (2019).

Methods

Strains and culture conditions

E. coli DH5α was used as a cloning host. Wild-type *E. coli* MG1655, 8739 and MG1655Δ*ung*, where *ung* was deleted using our previously described method, were used in the genome editing experiments. Strains were grown at 30 °C in lysogeny broth (LB, 1% (w/v) tryptone, 0.5% (w/v) yeast extract, and 1% (w/v) NaCl). Kanamycin (50 mg/L), chloramphenicol (30 mg/L), and apramycin (50 mg/L) were added to the medium when appropriate. One percent (w/v) glucose and 2 g/L

L-arabinose were added to the culture for Cas9 expression repression and induction. Isopropyl β-D-1-thiogalactopyranoside (IPTG, 0.4 mM) was used for the nCas9-AID system, the APOBEC-nCas9 system or for nCas9-TadA induction in *E. coli* when necessary. IPTG and X-gal were added at concentrations of 0.4 mM and 40 μg/mL, respectively, for blue/white selection.

Plasmid construction

Plasmids pnCas9-AID, pdCas9-AID, pnCas9-TadA, pUng-nCas9-AID and pAPOBEC-nCas9-Ung with expression for nCas9-AID, dCas9-AID, nCas9-TadA, Ung-nCas9-AID and APOBEC-nCas9-Ung fusion proteins, respectively, were assembled with the Golden Gate method[25]. PCR primers for Golden Gate and Gibson assembly were designed with the J5 Device Editor[26]. The backbones of pnCas9-AID, pdCas9-AID, pnCas9-TadA and pUng-nCas9-AID, containing the P_{trc} promoter, were PCR amplified from pACYC184-M[27]. Nuclease-deficient Cas9 (dCas9) for *E. coli* was PCR amplified from plasmid pdCas9-bacteria[28]. nCas9 (D10A) was also generated from plasmid pdCas9-bacteria; however, the catalytic H840 residue was reintroduced via PCR during its generation. Ung used in *E. coli* was PCR amplified from the MG1655 genome. The *E. coli* codon optimized DNA sequence of *aid*[15], an evolved TadA variant (ABE7.10)[8] and human expression of Ung, with the Cas9 variant linking regions (C-terminal and N-terminal), were synthesized (GenScript, China). The gRNA expression plasmids used in *E. coli* were constructed using a modular construction method, as described in our previous study[29]. Plasmid pAgRNA was used as the template for subsequent construction of *E. coli* gRNA expression plasmids. The backbone of pAgRNA was PCR amplified from pTrac-99A-apr-93, and the gRNA expression cassette was PCR amplified from pgRNA[23]. The two parts were assembled to yield plasmid pAgRNA. Plasmid pBadCas9, with inducible expression of Cas9, was also assembled with the Golden Gate method, and the assembly fragments were PCR amplified from pRed-Cas9[29] and pCAGO[30]. The backbones of pAPOBEC-nCas9-Ung, containing APOBEC-nCas9, were PCR amplified from pAPOBEC-nCas9-UGI. The shuttle vector pAPOBEC-nCas9-UGI was a gift from Dr. Feng Gu's laboratory. HEK293T gRNA expression plasmids were assembled with the

Golden Gate method with the N20 sequence embedded in the primers, and *RNF2* sgRNA expression plasmids[6] were used as the template.

E. coli vectors expressing Ung and AID with mammalian codons, and human vectors expressing Ung and APOBEC with *E. coli* codons were synthetized and assembled to produce pAPOBEC-nCas9-Ung and pUng-nCas9-AID, respectively, by GenScript (China); thus, pUng-nCas9-AID-homo and pAPOBEC-nCas9-Ung-E. coli plasmids were obtained.

All the DNA templates were PCR amplified with Phusion DNA polymerase (NEB, USA). PCR products were gel purified, digested with DpnI restriction enzyme (NEB, USA), and assembled with the Golden Gate assembly method. The main primers, plasmids and protospacer sequences used in HEK293T cells are listed in Supplementary Table S5, Supplementary Table S6 and Supplementary Table S7, respectively.

***E. coli* base editing procedure**

Fifty microliters of competent cells were mixed with 100 ng of pnCas9-AID, pdCas9-AID, pnCas9-TadA, pUng-nCas9-AID or pAPOBEC-nCas9-Ung-E.coli and 100 ng of pgRNA in a 2 mm Gene Pulser cuvette (Bio-Rad). After electroporation at 2.5 kV, cells were immediately suspended in 1 ml of ice-cold medium and then were incubated for 1 h at 30 °C and plated on solid LB medium with chloramphenicol (30 mg/L) and apramycin (100 mg/L). A single colony was selected and used to inoculate LB medium with chloramphenicol and apramycin, which then was incubated for 3 h at 30 ℃ and 250 rpm before the addition of 0.4 mM IPTG for initiating the genome editing process. The culture was then plated on agar plates containing LB with chloramphenicol and apramycin. Colonies obtained from the plates were analyzed by colony PCR with primers that targeted regions upstream and downstream of the 20 nt sgRNA target site. Successful editing was verified by DNA sequencing. All experiments were repeated three times to obtain the means and standard variations. When the *lacz* locus was targeted, white colonies in blue/white selection plates were chosen for further analysis.

HEK293T cell culture and transfection

HEK293T cells (from ATCC) were cultivated in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% (v/v) fetal bovine serum (FBS) at 37 °C under 5% CO_2. Cells were seeded in 24-well plates (Corning, USA). Approximately 24 h after seeding, cells were transfected at approximately 40% confluency with Lipofectamine 2000 (Life, Invitrogen, USA) according to the manufacturer's protocols. Then, 600 ng of Cas9 plasmid and 300 ng of sgRNA-expressing plasmid were transfected with 50 μl of DMEM containing 1.8 μl of Lipofectamine 2000. Twenty-four hours after transfection, 5 μg/ml puromycin (Merck, USA) was added to the media. In addition, 144 h after transfection, genomic DNA was extracted from the cells using a QuickExtract DNA Extraction Solution (Epicentre, USA). On-target and off-target genomic regions (200 bp~300 bp) of interest were amplified by PCR for high-throughput DNA sequencing.

High-throughput DNA sequencing of genomic DNA samples and data analysis

Next-generation sequencing library preparations were constructed following the manufacturer's protocol (VAHTS Universal DNA Library Prep Kit for Illumina). For each sample, >50 ng of purified PCR fragments were used for direct library preparation. The fragments were treated in one reaction with End Prep Enzyme Mix for end repair, 5' phosphorylation and dA tailing, which was followed by T-A ligation to add adaptors to both ends. Size selection of adaptor-ligated DNA was then performed using VAHTSTM DNA Clean Beads. Each sample was then amplified with 4 cycles of PCR using P5 and P7 primers; both primers carried sequences that could anneal with flow cells to perform bridge PCR, and P7 primers carried a six-base index allowing for multiplexing. The PCR products were purified using beads, validated using a Qsep100 (BiOptic, Taiwan, China), and quantified by a Qubit3.0 Fluorometer (Invitrogen, Carlsbad, CA, USA).

Then, libraries with different indexes were multiplexed and loaded on an Illumina HiSeq instrument according to the manufacturer's instructions (Illumina, San Diego, CA, USA). Sequencing was carried out using a 2x150 paired-end configuration; image analysis and base calling were conducted by HiSeq Control Software (HCS) + RTA 2.7 (Illumina) on a HiSeq instrument. For pair-end sequencing

results, read 1 and read 2 were merged to generate a complete sequence according to their overlapping regions, and a file in FASTA (fa) format was generated. Data were split according to their barcodes. The merged sequences were aligned to the reference sequence by using BWA (version 0.7.12) software. Examined target sites that mapped with around 100,000~1,000,000 independent reads were selected, and obvious base substitutions were observed at only the targeted base-editing sites. Base substitution frequencies were calculated by dividing base-substitution reads by total reads. Indel frequencies were calculated as previously described[6].

Selection of off-target sites

Off-target sites of selected target loci were analyzed by Cas-OFFinder[31], and ten loci containing the most similar sequences of selected target loci were chosen as predicted off-target sites. These off-target sites and the associated primers are listed in the supplementary information.

Data availability

There is no restriction on data associated with this study. Sequencing data supporting Figures 1, 2, 3, 4 and 5 are available upon request.

Acknowledgments

This research was financially supported by the National Key Research and Development Program of China (2019YFA0904900), the Key Research Program of the Chinese Academy of Science (KFZD-SW-215), National Natural Science Foundation of China (31522002, 31770105), and a Newton Fund PhD placement program grant, ID 352639434, under the UK-China Joint Research and Innovation Partnership Fund. We gratefully thank Professor Feng Gu (Wenzhou Medical University, China) for endowing the plasmid pAPOBEC-nCas9-UGI, and Dr. Yaqiu Li (Tianjin Institute of Industrial Biotechnology CAS, China) for assiting with data analysis.

Author Contributions

X.Z., C.B. and J.L. designed the research, analyzed data, and wrote the manuscript. D.Z. designed the research, performed experiments, analysed data, and

wrote the manuscript. S.L., performed experiments, analysed data. M.A.P. performed experiments and wrote the manuscript. S.J.R. wrote the manuscript. X.X., and M.H. performed experiments.

Competing Interests

A provisional patent has been submitted in part entailing the reported approach.

Materials & Correspondence

Correspondence and requests for materials should be addressed to C.B. and X.Z. (bi_ch@tib.cas.cn; zhang_xl@tib.cas.cn).

References

25. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PloS one* **3**, e3647 (2008).
26. Hillson, N.J., Rosengarten, R.D. & Keasling, J.D. j5 DNA assembly design automation software. *ACS Synthetic Biology* **1**, 14-21 (2011).
27. Zhao, J. et al. Engineering central metabolic modules of *Escherichia coli* for improving β-carotene production. *Metabolic engineering* **17**, 42-50 (2013).
28. Qi, L.S. et al. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152**, 1173 (2013).
29. Feng, X., Zhao, D., Zhang, X., Ding, X. & Bi, C. CRISPR/Cas9 assisted Multiplex Genome Editing Technique in *Escherichia coli*. *Biotechnology Journal*, e1700604 (2018).
30. Zhao, D. et al. CRISPR/Cas9-assisted gRNA-free one-step genome editing with no sequence limitations and improved targeting efficiency. *Scientific Reports* **7**, 16624 (2017).
31. Bae, S., Park, J. & Kim, J.S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473-1475 (2014).

Figure legends

Figure 1. The specificity efficiencies of base editing in *E. coli* ATCC 8739 and MG1655 strains by nCas9-AID. Eight loci were tested in both MG1655 cells and seven loci were tested in ATCC 8739 cells. Base editing experiments were performed in triplicate for each locus. The average

specificity efficiencies **[AU: Not quite sure what this means please rephrase here and in the figure]**and standard deviations were calculated based on sequencing results.

Figure 2. Base editing in *E. coli* using Ung-nCas9-AID fusion. **A)** Illustration of the functional mechanism of Ung-nCas9-AID. C is deaminated to U by AID, which is excised by Ung to form an AP site. nCas9 also cleaves the other strand at a nearby location. **B)** Base editing of *E. coli* MG1655 by Ung-nCas9-AID. Four loci were tested in MG1655, and base editing experiments were performed in triplicate for each locus. The average specificity efficiencies **[AU: Not quite sure what this means please rephrase here and in the figure]** and standard deviations were calculated based on sequencing the results.

Figure 3. Base editing in HEK293T cells using GBE and the APOBEC-nCas9-Ung system, with APOBEC-nCas9-UGI as a control. **A)** The base editing efficiency of the APOBEC-nCas9-UGI system by deep sequencing. The columns represent an average efficiency of C-to-G, C-to-T and C-to-A at 18 positions of protospacers (C1, C3, C4, C5, C6, C7, C9, C10, C11, C12, C13, C14, C15, C16, C17, C18, C19, and C20) from 10 loci (*EMX1-site2, EMX1-site3, FANCF-site1, FANCF-site2, FANCF-site3, RNF2, RNF2-site1, PSMB2-1, VEGFA-site2,* and *PDL1*) **B)** The base editing efficiency of the APOBEC-nCas9-Ung system in the same 18 positions listed above. Base editing experiments were performed in triplicate for each locus, and the average efficiencies and standard deviations were calculated based on sequencing the results.

Figure 4. The base editing efficiency and indels of the APOBEC-nCas9-Ung system at the C6 position from 30 loci, and fraction of cytosine substitutions at the C6 from APOBEC-nCas9-Ung and APOBEC-nCas9-UGI systems. Base editing experiments were performed in triplicate for each locus, and the average efficiencies and standard deviations were calculated based on sequencing the results.

Figure 5. Any base editing (NBE) in *E. coli*. **A)** Diagram showing the steps involved in NBE with the bases targeted for initial editing in red. C and A conversion acts as an initial target when editing of their complimentary G and T bases is desired. **B)** DNA sequences obtained following sequential NBE steps. Horizontal bar graphs indicate the editing efficiency observed at each step following sequencing analysis. Base editing experiments were performed in triplicate for each locus, and the average efficiencies and standard deviations were calculated based on sequencing the results.

Extended Data Figure 1. Sequencing data results following CBE. **A)** Results obtained following treatment with nCas9-AID in *E. coli* MG1655 Δung. **B)** Results obtained following treatment with dCas9-AID in wild type *E. coli* MG1655. Base editing experiments were performed in triplicate for each locus, and the average efficiencies and standard deviations were calculated based on sequencing the results.