



Published in final edited form as:

Nat Biotechnol. 2021 January ; 39(1): 41–46. doi:10.1038/s41587-020-0609-x.

CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells

Ibrahim C. Kurt^{1,2,3,5}, Ronghao Zhou^{1,2,5}, Sowmya Iyer¹, Sara P. Garcia¹, Bret R. Miller^{1,2}, Lukas M. Langner^{1,2}, Julian Grünewald^{1,2,4,6}, J. Keith Joung^{1,2,4,6}

¹Molecular Pathology Unit, Massachusetts General Hospital, Charlestown, MA, USA

²Center for Cancer Research and Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA, USA

³Biological Sciences in Public Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA

⁴Department of Pathology, Harvard Medical School, Boston, MA, USA

⁵These authors contributed equally to this work

⁶These authors jointly supervised this work

Abstract

CRISPR-guided DNA cytosine and adenine base editors (CBEs and ABEs) are widely used for many applications^{1–4} but primarily create DNA base transitions (i.e., pyrimidine-to-pyrimidine, or purine-to-purine). Here we describe the engineering of two base editor architectures that can efficiently induce targeted C-to-G base transversions, with reduced levels of unwanted C-to-W (W = A or T) and indel mutations. One of these C-to-G base editors (CGBE1), consists of an RNA-guided Cas9 nickase, an *E. coli*-derived uracil DNA N-glycosylase (eUNG), and a rat APOBEC1 cytidine deaminase variant (R33A) previously shown to have reduced off-target RNA and DNA editing activities^{5, 6}. We show that CGBE1 can efficiently induce C-to-G edits, particularly in AT-rich sequence contexts in human cells. We also removed the eUNG domain to yield miniCGBE1, which reduced indel frequencies but only modestly decreased editing efficiency. CGBE1 and miniCGBE1 enable C-to-G edits and will serve as a basis for optimizing C-to-G base editors for research and therapeutic applications.

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

*Correspondence should be addressed to: jjoung@mgh.harvard.edu or jgrunewald@mgh.harvard.edu.

Author Contributions

S.I. and S.P.G. contributed equally to this work and are co-second authors. Wet lab experiments were performed by I.C.K., R.Z., B.R.M., and L.M.L., while S.P.G. and S.I. performed computational analysis of the data. I.C.K., R.Z., J.G., and J.K.J. conceived of and designed the study. J.G. and J.K.J. supervised the work. I.C.K., R.Z., J.G., and J.K.J. wrote the initial manuscript draft and all authors contributed to the editing of the manuscript.

Competing Financial Interests Statement

J.K.J. has financial interests in Beam Therapeutics, Editas Medicine, Excelsior Genomics, Pairwise Plants, Poseida Therapeutics, Transposagen Biopharmaceuticals, and Verve Therapeutics (*f/k/a* Endcadia). J.K.J.'s interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies. J.K.J. is a member of the Board of Directors of the American Society of Gene and Cell Therapy. J.G., I.C.K., R.Z., and J.K.J. are co-inventors on patent applications that have been filed by Partners Healthcare/Massachusetts General Hospital on engineered C-to-G transversion base editors and other related base editor technologies.

Editorial summary

A new base editor enables the creation of C-to-G base changes in human cells.

Our efforts to develop a C-to-G base editor were motivated by our previous observation that the A-to-G editor ABE_{max} could induce unexpected C-to-G edits at sites in which a C was present at position 6 of the protospacer⁷ (numbering starting from the position most distal to the protospacer adjacent motif [PAM]). (These results were recently confirmed by another group who reported C-to-G edits for Cs at positions 5, 6, and 7 of the protospacer by the same ABE⁸.) We hypothesized that the heterodimeric TadA-based adenosine deaminase domain present in ABE_{max} might deaminate the C at position 6 to a U with subsequent creation of an abasic site by cellular UNG (base excision repair) followed by the preferential introduction of a G at that position by an as-yet undefined mechanism (Fig. 1a). Consistent with this, we found that the addition of two uracil glycosylase inhibitor (**UGI**) domains to ABE_{max} resulted in a reduction of C-to-G edits and indels, and an increase in C-to-T edits (Extended Data Figs. 1 and 2a, Supplementary Tables 1 and 2), presumably due to reduced efficiency of uracil excision.

Given this observation about ABE-mediated C-to-G alterations, we wondered whether we could induce these edits more efficiently by modifying the BE4_{max} CBE^{9, 10}, which harbors an enzyme actually intended to deaminate cytosines (the rat APOBEC1 cytidine deaminase). Removal of the two UGIs from BE4_{max} to create BE4_{max} UGI resulted in an increase in C-to-G (and to a lesser degree C-to-A) edits relative to wild-type BE4_{max} when tested with seven different gRNAs targeted to sites with Cs at protospacer positions 5, 6, and 7 in HEK293T cells (Fig. 1b and Supplementary Table 2). In general, C-to-G editing was observed with BE4_{max} UGI at Cs that were preceded by A, C, or T, with the most efficient editing generally observed with Cs at protospacer position 6 (Fig. 1b). We also observed a substantially higher frequency of indels with BE4_{max} UGI relative to BE4_{max} (Fig. 1b, Extended Data Fig. 2b), consistent with the idea that this fusion is likely more efficient at creating abasic sites^{2, 9}. Reasoning that generation of an abasic site is important for increased C-to-G editing, we further hypothesized that adding human UNG (**hUNG**) enzyme to BE4_{max} UGI might enhance the frequency of desired edits. However, a BE4_{max} UGI-hUNG fusion possessed somewhat decreased C-to-G editing activity and did not induce appreciably changed frequencies of indels with the seven gRNAs tested (although it did show decreased C-to-T editing activity) (Fig. 1b, Extended Data Fig. 2b, and Supplementary Table 2). Similar results were obtained when hUNG was fused at the N-terminus of BE4_{max} UGI (Extended Data Fig. 1 and Supplementary Table 2). Fusion of UNG to ABE_{max} did not yield enhanced C-to-G editing compared to ABE_{max} (Extended Data Fig. 1 and Supplementary Table 2). We also tested a variety of CBEs that are based on non-APOBEC1 deaminase architectures, such as human APOBEC3A (**A3A**), engineered A3A-BE3¹¹, human AID-BE3⁹, and the *Petromyzon marinus* CDA1 (**pmCDA1**)-based Target-AID³, as well as variants thereof lacking UGIs and having added UNGs. Among this larger ensemble of variants, none consistently showed higher C-to-G editing activity than the BE4_{max} UGI-hUNG editor (Supplementary Fig. 1 and Supplementary Table 2).

We also investigated whether introducing mutations into the APOBEC1 part of BE4max UGI-hUNG might further increase the frequency of C-to-G editing. Although we do not have a mechanistic understanding of how C-to-G edits are induced, we reasoned that altering the deamination dynamics of APOBEC1 might also influence the editing outcome. We focused on the APOBEC1 R33A mutation, a substitution we previously showed can decrease off-target RNA editing while substantially preserving the efficiency and increasing the precision of on-target DNA editing by CBEs⁵. We found that introduction of R33A into BE4max UGI-hUNG increased C-to-G editing frequencies with three of the seven gRNAs tested in HEK293T cells while leaving editing frequencies essentially unaltered with the other four (Fig. 1b and Supplementary Table 2). The effect of the R33A variant was most striking with the *FANCF* site 1 gRNA, which had shown virtually no C-to-G editing with any of the other editors we tested but now showed a mean editing frequency of 14.0% (Fig. 1b). Interestingly, BE4max(R33A) UGI-hUNG on average showed lower indel byproducts with 6 out of 7 gRNAs compared to BE4max UGI-hUNG (Extended Data Fig. 2b).

We additionally explored whether replacing the hUNG present in the BE4max(R33A) UGI-hUNG editor with an orthologous UNG from *Escherichia coli* (eUNG) might increase the efficiency of C-to-G edits. We created two additional editors: BE4max(R33A) UGI-eUNG and eUNG-BE4max(R33A) UGI with an eUNG added to the carboxy- or amino-terminal ends, respectively. Testing of these fusions in HEK293T cells revealed that both induced C-to-G edits with higher frequencies than BE4max(R33A) UGI-hUNG for six out of seven gRNAs tested (mean editing frequencies ranging from 3.3–57.0% and 8.5–62.6% for BE4max(R33A) UGI-eUNG and eUNG-BE4max(R33A) UGI, respectively) (Fig. 1b and Supplementary Table 2). Indel frequencies with both fusions were generally comparable to those observed with BE4max(R33A) UGI-hUNG (Extended Data Fig. 2b). Given its higher C-to-G editing activity, we chose the eUNG-BE4max(R33A) UGI fusion (hereafter referred to as C-to-G Base Editor 1 (CGBE1)) for additional characterization.

To more comprehensively characterize CGBE1, we tested its activity with 18 additional gRNAs in human HEK293T cells. 12 of the sites targeted by these 18 gRNAs have a C at position 6 (“C6-sites”) (Fig. 2a and Extended Data Fig. 3) and 6 have a C at positions 4, 5, 7, or 8 (“non-C6-sites”) (Fig. 2b and Extended Data Fig. 4a). For 16 of the 18 sites, CGBE1 induced C-to-G edits with substantially higher frequencies than what was observed with its parental CBE control (BE4max(R33A)) (Fig. 2a and b, Supplementary Table 3). Highly efficient C-to-G edits were observed for 4 of the 18 sites (ABE site 7, ABE site 8, HEK site 2, and *PPP1R12C* site 6), with mean editing frequencies ranging from 41.7 to 71.5% (Fig. 2a and b). C-to-G edits were by far the most efficiently induced edits at these 4 sites with only very low levels of C-to-T or C-to-A byproducts observed (Fig. 2a and b). C-to-G was also the most efficiently induced edit for 6 additional sites albeit at lower frequencies (three C6-sites and three non-C6-sites) (Fig. 2a and b). In total, when combined with the results obtained with the initial seven gRNAs described above (Fig. 1b), CGBE1 induced C-to-G editing with mean frequencies of 20% or higher at 14 of the 25 sites tested (Figs. 1b, 2a and b). Notably, C-to-G editing was most efficient for Cs embedded in an AT-rich sequence context (Figs. 1b, 2a and b). Analysis of the spatial distribution of editing across all 25 sites tested shows that the mean frequency of C-to-G editing was highest at position 6 and that

indels (Extended Data Figs. 2b and 4b) were distributed throughout the protospacer (Fig. 2c).

We next explored the impact of deleting the eUNG domain from the CGBE1 editor on its activity. This particular editor architecture, which we named miniCGBE1 (Fig. 3), had not been made or tested over the course of the stepwise progression from BE4max to CGBE1 and also had the added advantage of being smaller in size. Side-by-side comparisons of miniCGBE1 with CGBE1 at the same 25 sites we had previously tested showed that the frequencies of editing observed with miniCGBE1 were comparable but moderately lower at 6 out of 25 sites tested (mean editing frequencies across all 25 sites of 14.4% and 13% with CGBE1 and miniCGBE1, respectively), whereas the indel frequencies induced by miniCGBE1 were lower at 15 out of 25 sites (mean indel frequencies of 10.4% and 8.5% for CGBE1 and miniCGBE1, respectively; Fig. 3, Supplementary Fig. 2 and Supplementary Table 4).

To more fully characterize the positional preferences within the editing windows of CGBE1 and miniCBGE1, we tested these two editors side-by-side with BE4max and BE4max(R33A) using 23 additional gRNAs that target sites with cytosines at protospacer positions 4, 5, 7, and 8 (Supplementary Fig. 3). The targets of these 23 gRNAs included six sites with a C5, five with a C7, four with a C8, and eight with two Cs at various positions (C4 and C7, C4 and C8, C5 and C7, C5 and C8, and C7 and C8). Mean editing frequencies induced by CGBE1 were comparable to those of miniCGBE1: 1.7% and 1.5% at C4, 7.3% and 6.7% at C5, 16.0% and 13.5% at C7 and 3.4% and 2.9% at C8 for CGBE1 and miniCGBE1, respectively (Supplementary Fig. 3a and Supplementary Table 5). In addition, indel frequencies induced by CGBE1 and miniCGBE1 were comparable at 10 sites, lower with CGBE1 at five sites, and lower with miniCGBE1 at eight sites (Supplementary Fig. 3b). Collectively, our testing of CGBE1 and miniCGBE1 with 48 different gRNAs demonstrates that both have a favorable editing window for cytosines at positions 5–7 in the protospacer with those at position 6 being edited most efficiently (Extended Data Fig. 5). This finding is consistent with our previously published studies showing that a CBE with the APOBEC1-R33A variant edits optimally on positions 5–7 of the protospacer and more weakly on positions 4 and 8⁵.

To characterize gRNA-dependent DNA off-target profiles of CGBE1 and miniCGBE1, we assessed their editing activities side-by-side with BE4max and BE4max(R33A) at 23 known SpCas9 off-target sites of five different gRNAs (previously identified by GUIDE-seq¹²) (Supplementary Fig. 4). BE4max induced C-to-D (D = A, G, or T) edits at 15 of the 23 off-target sites, while BE4max-R33A induced edits at lower frequencies across all of these 15 sites (Supplementary Fig. 4a and Supplementary Table 6). Similarly, both CGBE1 and miniCGBE1 showed lower C-to-D off-target editing at 14 out of the 15 off-target sites that were edited by BE4max (Supplementary Fig. 4a and Supplementary Table 6). As expected, off-target indel frequencies were higher with CGBE1 and miniCGBE1 relative to BE4max at 18 out of 23 sites, although miniCGBE1 again showed reduced activity compared with CGBE1 at 14 out of these 18 sites (Supplementary Fig. 4b). Overall, this assessment of gRNA-dependent DNA off-target editing shows that CGBE1 and miniCGBE1 induce fewer

off-target DNA base edits than BE4max, that CGBE-induced indels can occur at off-target sites, and that indels are reduced with miniCGBE1 relative to CGBE1.

We additionally tested whether we could improve the somewhat more restricted targeting range of CGBEs by using previously described SpCas9-NG and SpCas9-VRQR variants that recognize shorter NG¹³ and alternative NGA¹⁴ PAMs, respectively. We targeted six sites with NGT PAMs using modified CGBE1-NG and miniCGBE1-NG variants and six sites with NGAG PAMs using CGBE1-VRQR and miniCGBE1-VRQR variants. Each of these 12 sites have a cytosine at position 6 embedded within an AT-rich sequence context to provide an optimal target for C-to-G editing (Extended Data Fig. 6). At these target sites, CGBE1-NG and miniCGBE1-NG induced C-to-G edits with frequencies as high as 27% and 26%, respectively, and CGBE1-VRQR and miniCGBE1-VRQR induced C-to-G edits with frequencies of up to 31% (Extended Data Fig. 6 and Supplementary Table 7). These results show that the targeting range of CGBE constructs can be expanded by using Cas9 variants with altered or relaxed PAM recognition specificities.

Lastly, we compared our CGBEs with new Prime Editing (**PE**) methods that can introduce a diverse range of different edits and that were published¹⁵ while we were completing this project. The PE2 system uses two components: (1) a Prime Editor fusion protein and (2) a prime editing gRNA (**pegRNA**) (Extended Data Fig. 7a)¹⁵. The more efficient PE3 system adds a secondary “nicking gRNA” (**ngRNA**) that directs a nick to the non-edited DNA strand, thereby increasing editing efficiency (Extended Data Fig. 7a)¹⁵. We performed side-by-side comparisons of our CGBEs with PE2 and PE3 systems for installing four different C-to-G edits, assessing frequencies of these alterations across four different human cell lines (HEK293T, K562, U2OS, and HeLa cells) (Online Methods). Positive control experiments we performed in all four cell lines re-confirmed that two other previously described pegRNAs could induce a G-to-T transversion in *FANCF* site 1 and a CTT insertion in HEK site 3, that PE3 outperforms PE2, and that the highest prime edit frequencies are observed in HEK293T cells (Extended Data Fig. 7b-d and Supplementary Table 8). For all four C-to-G edits (which we had already established could be efficiently induced by CGBEs in HEK293T cells), we found that both PE2 and PE3 were substantially less efficient than CGBE and miniCGBE1 across all four cell lines (Extended Data Fig. 7e, Supplementary Fig. 5a and Supplementary Table 8). Importantly, these data also show that our CGBEs can function robustly and efficiently across multiple human cancer cell lines. In addition, we found that the frequencies of unwanted indels were lower with prime editors compared to the CGBEs in all four cell lines, an unsurprising finding given that on-target editing of PEs was also much less efficient (Supplementary Fig. 5b). To rule out that the pegRNAs and ngRNAs we designed were inactive or unable to interact with Cas9, we tested their abilities to induce Cas9-mediated indels at their target sites in HEK293T cells (note that we could not assess the activity of the HEK site 3 ngRNA due to its overlap with a required PCR primer). The indel frequencies induced by these pegRNAs and ngRNAs were comparable to those observed with the two positive controls (Extended Data Fig. 7f).

Our development and characterization of CGBE1 and miniCGBE1 provide an important proof-of-concept for a modified base editor architecture that can be used to induce programmable C-to-G transversion mutations. Additional optimization and engineering will

be required to enable a more robust and efficient CGBE platform. Although highly efficient C-to-G editing with high product purity could be observed at some sites with a C at position 6 in protospacer, not all C6 sites showed equally high activities. At least some of this variability may be influenced by the identities of bases flanking C6. For example, C6 sites with the most efficient editing and optimal product purity had As and/or Ts at the flanking 5 and 7 positions. In addition, C6 sites with a preceding G showed consistently absent or very low C-to-G editing, a known limitation of the rat APOBEC1 cytidine deaminase². In addition, our combined data from 48 different sites demonstrated that C-to-G editing by CGBE constructs was favored when the target C was in position 6 and less favored when in positions 5, 7 and 8. Collectively, these observations show that the targeting range of the CBGE constructs is constrained by both the location of the target C and its sequence context, although additional work could further refine these parameters. The use of CGBEs harboring Cas9 PAM recognition variants with altered^{14, 16} or broadened^{13, 17, 18} PAM specificities may help to further alleviate some of these targeting range limitations.

Among four target sites we tested, we also found that our CGBE variants generally induced higher frequencies of desired C-to-G edits than PE2 or PE3 prime editors directed to induce the same alterations. This was consistently observed for these four C-to-G edits across four different human cancer cell lines. Although we designed the pegRNAs and ngRNAs following parameters previously described by the Liu group¹⁵, it is possible that additional optimization (e.g., of the pegRNA target site choice, primer binding site (PBS) or reverse transcription template (RTT) lengths, distance between the pegRNA and ngRNA nicking sites) might lead to more efficient prime editing outcomes. For the four sites we tested, prime editors appear to induce lower frequencies of unwanted indels relative to our CGBEs. All of these findings are consistent with the conclusions of Liu and colleagues that base editors induce a more restricted range of edits than prime editors but also generally show more efficient editing if the target base is positioned optimally¹⁵.

Although we do not yet understand why the use of the rat APOBEC1 R33A variant increases the efficiency of C-to-G editing, this observation strongly suggests that altering the activities of APOBEC1 with other mutations might yield further desirable increases in editing activities. We also found that C-to-G editing frequencies showed cell-type-dependence across the four cell lines tested, a finding consistent with previous reports, which described that even standard CBEs can induce higher efficiencies of unwanted C-to-G edits in certain cell types^{9, 11}. Finally, a better understanding of the mechanistic parameters that govern both the frequencies and the product purity of C-to-G edits may suggest additional strategies to further increase the targeting range and efficiency of CGBEs.

The availability of efficient and programmable C-to-G base editing should further expand both research and therapeutic applications of base editing. C-to-G editing permits the introduction of new codon and sequence changes not possible with CBEs and ABEs. C-to-T base edits performed by CBE at the third nucleotide (wobble) position of codons can induce only two amino acid alterations Met→Ile (ATG→ATA) and Trp→Stop (TGG→TGA). Similarly, ABE-induced A-to-G edits can induce only Ile→Met (ATA→ATG) and Stop→Trp (TGA→TGG) alterations. By contrast, C-to-G editing by CGBE at the third nucleotide position of codons enables 16 different amino acid alterations not possible with

CBEs or ABEs. C-to-G alterations at the first and second nucleotide positions of codons also enable additional amino acid substitutions. Furthermore, transversion mutations introduced into transcription factor binding sites would be expected to have more pronounced effects on binding and gene expression as compared to transition mutations¹⁹. Finally, and most importantly, the ability to install C-to-G edits will permit the correction of additional disease-causing mutations in both coding and non-coding regions that cannot be accessed with existing technologies. As was the case with both CBEs and ABEs^{9, 10, 20, 21}, we envision that our initial description of CGBE1 and miniCGBE1 will spur further optimization and development of next-generation C-to-G editors with more efficient and robust activities.

Methods

Molecular Cloning

All base editor (BE) and prime editor (PE) constructs (PE2, Addgene #132775) used in this study (Supplementary Table 1) were cloned into a mammalian expression plasmid backbone under the control of a pCMV promoter (AgeI and NotI restriction digest of parental plasmid Addgene #112101). The wild-type SpCas9 construct (SQT 817; Addgene #53373) used in Extended Data Fig. 7f is expressed under the control of a CAG promoter. All BE and PE constructs were encoded as P2A-eGFP fusions for co-translational expression of the base/prime editors and eGFP. Gibson fragments with matching overlaps were PCR-amplified using Phusion High-fidelity polymerase (NEB). Fragments were gel-purified and assembled for 1 hour at 50°C and transformed into chemically competent *E. coli* (XL1-Blue, Agilent). The UNGs used in our experiments originated either from *E. coli* (eUNG; UniProtKB-P12295) or *Homo sapiens* (hUNG; UniProtKB-P13051), were codon-optimized for expression in human cells and synthesized as gblocks (IDT). All guide RNA (gRNA) constructs (Supplementary Table 9) used in this study were cloned into a BsmBI-digested pUC19-based entry vector (BPK1520, Addgene #65777) with a U6 promoter driving gRNA expression. We designed the pegRNAs (Supplementary Table 9) to implement the same C-to-G changes that the CGBE constructs would install and followed previously described default design rules for designing pegRNAs and ngRNAs¹⁵. PegRNAs were cloned into the BsaI-digested pU6-pegRNA-GG-acceptor entry vector (Addgene #132777) and ngRNAs were cloned into the above-mentioned BsmBI-digested entry vector BPK1520. For pegRNA cloning, oligos containing the spacer, the 5' phosphorylated pegRNA scaffold, and the 3' extension sequences were annealed to form dsDNA fragments with compatible overhangs and ligated using T4 ligase (NEB). All plasmids used for transfection experiments were prepared using Qiagen Midi or Maxi Plus kits.

Cell Culture

STR-authenticated HEK293T (CRL-3216), K562 (CCL-243), HeLa (CCL-2), and U2OS cells (similar match to HTB-96; gain of #8 allele at the D5S818 locus) were used in this study. HEK293T and HeLa cells were grown in Dulbecco's Modified Eagle Medium (DMEM, Gibco) with 10% heat-inactivated fetal bovine serum (FBS, Gibco) supplemented with 1% penicillin-streptomycin (Gibco) antibiotic mix. K562 cells were grown in Roswell Park Memorial Institute (RPMI) 1640 Medium (Gibco) with 10% FBS supplemented with

1% Pen-Strep and 1% GlutaMAX (Gibco). U2OS cells were grown in DMEM with 10% FBS supplemented with 1% Pen-Strep and 1% GlutaMAX. Cells were grown at 37°C in 5% CO₂ incubators and periodically passaged upon reaching around 80% confluency. Cell culture media supernatant was tested for mycoplasma contamination every 4 weeks using the MycoAlert PLUS mycoplasma detection kit (Lonza) and all tests were negative throughout the experiments.

Transfection & Electroporation Experiments

HEK293T cells were seeded at 1.25×10^4 cells per well into 96-well flat bottom cell culture plates (Corning) for DNA on-target experiments or at 6.25×10^4 cells per well into 24-well cell culture plates (Corning) for DNA off-target experiments. 24 hours post-seeding, cells were transfected with 30 ng of control or base/prime editor plasmid and 10 ng of peg or gRNA plasmid (and 3.3 ng nicking gRNA plasmid for PE3) using 0.3 μ L of TransIT-X2 (Mirus) lipofection reagent for experiments in 96-well plates, or 150 ng control or base editor plasmid and 50 ng gRNA, and 1.5 μ L TransIT-X2 for experiments in 24-well plates. K562 cells were electroporated using the SF Cell Line Nucleofector X Kit (Lonza), according to the manufacturer's protocol with 2×10^5 cells per nucleofection and 800 ng control or base/prime editor plasmid, 200 ng gRNA or pegRNA plasmid, and 83 ng nicking gRNA plasmid (for PE3). U2OS cells were electroporated using the SE Cell Line Nucleofector X Kit (Lonza) with 2×10^5 cells and 800 ng control or base/prime editor plasmid, 200 ng gRNA or pegRNA, and 83 ng nicking gRNA (for PE3). HeLa cells were electroporated using the SE Cell Line 4D-Nucleofector X Kit (Lonza) with 5×10^5 cells and 800 ng control or base/prime editor, 200 ng gRNA or pegRNA, and 83 ng nicking gRNA (for PE3). 72 hours post-transfection, cells were lysed for extraction of genomic DNA (gDNA).

DNA Extraction

HEK293T cells were washed with 1X PBS (Corning) and lysed overnight by shaking at 55°C with 43.5 μ L of gDNA lysis buffer (100 mM Tris-HCl at pH 8, 200 mM NaCl, 5 mM EDTA, 0.05% SDS) supplemented with 5.25 μ L of 20 mg/ml Proteinase K (NEB) and 1.25 μ L of 1M DTT (Sigma) per well for experiments in 96-well plates, or with 174 μ L DNA lysis buffer, 21 μ L Proteinase K, and 5 μ L 1M DTT per well for experiments in 24-well plates. K562 cells were centrifuged for 5 min, media removed, and lysed overnight by shaking at 55°C with 174 μ L DNA lysis buffer, 21 μ L Proteinase K, and 5 μ L 1M DTT per well in 24-well plates. U2OS cells and HeLa cells were washed with 1X PBS and lysed overnight shaking at 55°C with 174 μ L DNA lysis buffer, 21 μ L Proteinase K, and 5 μ L 1M DTT per well in 24-well plates. Subsequently, gDNA was extracted from lysates using 1–2X paramagnetic beads as previously described⁵ and eluted in 45 μ L of 0.1X EB buffer. DNA extraction was performed using a Biomek FX^P Laboratory Automation Workstation (Beckman Coulter).

Targeted Amplicon Sequencing

DNA targeted amplicon sequencing was performed as previously described⁵. Briefly, extracted gDNA was quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher). Amplicons were constructed in 2 PCR steps. In the first PCR, regions of interest (170–250

bp) were amplified from 5–20 ng of gDNA with primers containing Illumina forward and reverse adapters on both ends (Supplementary Table 9). PCR products were quantified on a Synergy HT microplate reader (BioTek) at 485/528 nm using a Quantifluor dsDNA quantification system (Promega), pooled and cleaned with 0.7X paramagnetic beads, as previously described⁵. In a second PCR step (barcoding), unique pairs of Illumina-compatible indexes (equivalent to TruSeq CD indexes, formerly known as TruSeq HT) were added to the amplicons. The amplified products were cleaned up with 0.7X paramagnetic beads, quantified with the Quantifluor or Qubit systems, and pooled before sequencing. The final library was sequenced on an Illumina MiSeq machine using the Miseq Reagent Kit v2 (300 cycles, 2×150bp, paired-end). Demultiplexed FASTQ files were downloaded from BaseSpace (Illumina).

Analysis and Plotting of NGS data

Amplicon sequencing data were analyzed with CRISPResso2 version 2.0.31 in batch mode²². Downstream analysis was performed using R version 3.5.1 with data sourced from ‘Nucleotide_percentage_summary.txt’ for nucleotide distribution barplots and ‘MODIFICATION_PERCENTAGE_SUMMARY.txt’ tables for indels. Each point in position-wise indel plots (Fig. 2c indel frequency dot plots) is the sum of “Insertions_Left” and “Deletions” columns in the text file for each position in the protospacer. CRISPResso2 output table “CRISPRessoBatch_quantification_of_editing_frequency.txt”, was used to quantify percentage of alleles that contain an insertion or deletion across the protospacer sequence for base/prime editor experiments, with window parameters set to -wc 10 -w 10 (Extended Data Figs. 2, 4b, 6b, and 7c, d and Supplementary Figs. 1b, 2c, 3b, 4b, 5b). For WT-Cas9 experiments (Extended Data Fig. 7f) default window parameters were used (-wc 3 -w 1) to quantify indels around the expected cut site. Each point on the plot is calculated from columns in the above table as follows: $\frac{\text{Insertions} + \text{Deletions}}{\text{Reads_aligned}}$. Extended Data Fig. 7d: From tables “Alleles_frequency_table_around_sgRNA_*.txt” that are part of CRISPResso2 output, percentages of alleles that contained a CTT insertion in positions 18, 19 and 20 of the protospacer were summed and reported in Extended Data Fig. 7d (% alleles bar plot). Sums of percentages of other indels besides the above insertions were plotted separately in Extended Data Fig. 7d (indel frequency dot plot).

Statistics & Data Reporting

Boxes span the interquartile range (IQR; 25th to 75th percentile), horizontal lines indicate the median (50th percentile), and whiskers extend to $\pm 1.5 \times \text{IQR}$. Data points in box and dot plots represent full range of values plotted (Fig. 2c and Extended Data Fig. 5). A two-tailed Student’s t-test with *p*-values adjusted for multiple testing was used to calculate the *p*-values in Extended Data Fig. 7e. The error bars in all dot and bar plots show the standard deviation (s.d.) and were plotted with ggplot (R-3.5.1). The measure of center for the error bars is the mean. We did not predetermine sample sizes based on statistical methods. Investigators were not blinded to experimental conditions or assessment of experimental outcomes.

Data availability

Plasmids encoding CGBE1 (Addgene #140252) and miniCGBE1 (Addgene #140253), as well as other constructs used in this work are available on Addgene via https://www.addgene.org/Keith_Joung/. Targeted amplicon sequencing data (obtained from Illumina Basespace) have been deposited at the Sequence Read Archive (SRA): <https://www.ncbi.nlm.nih.gov/sra/PRJNA622835>. All other relevant data are available from the corresponding authors upon request.

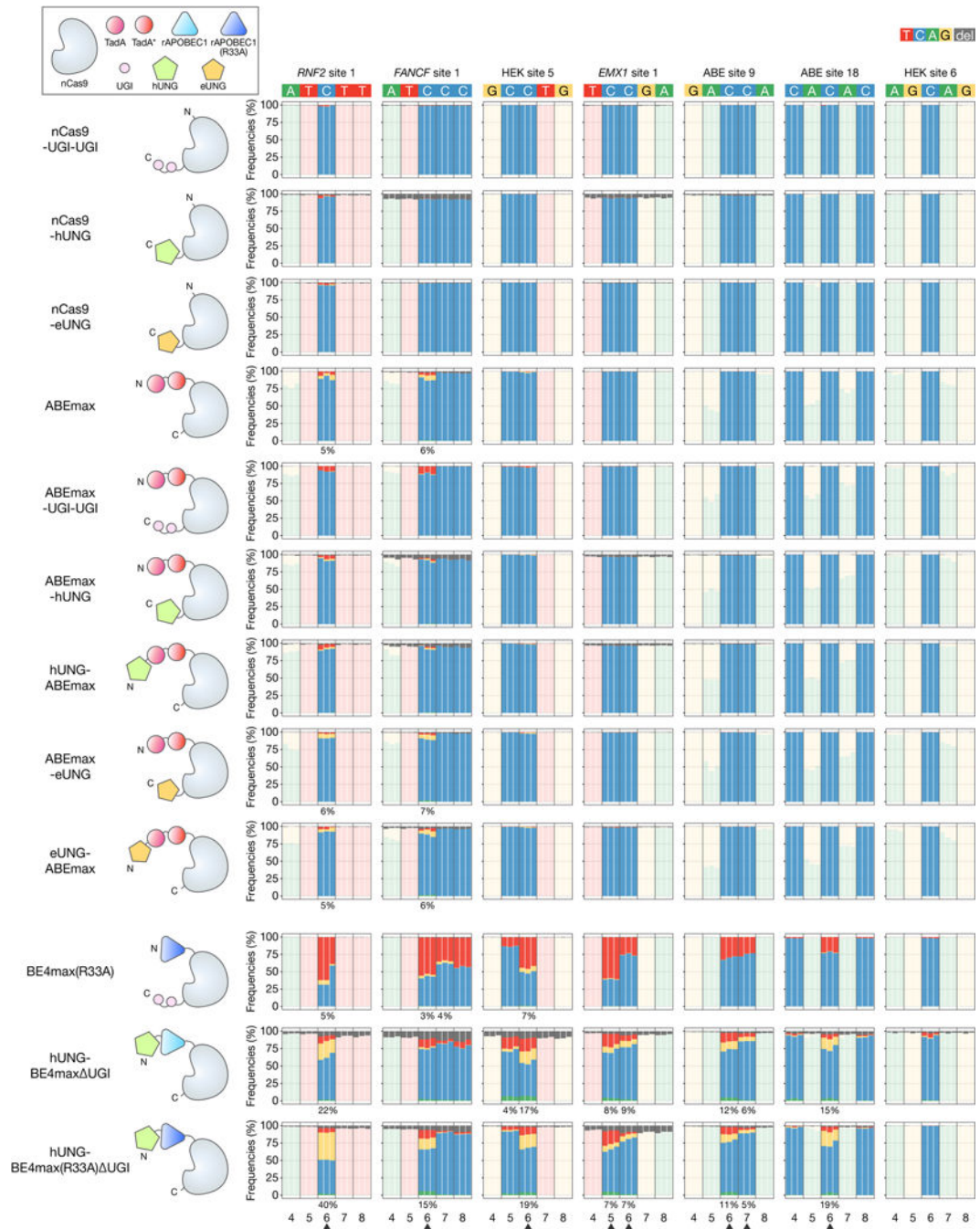
Code availability

No custom code was used in this study that was central to its conclusions. Code to generate plots from CRISPResso output tables will be provided upon request.

Life Sciences Reporting Summary.

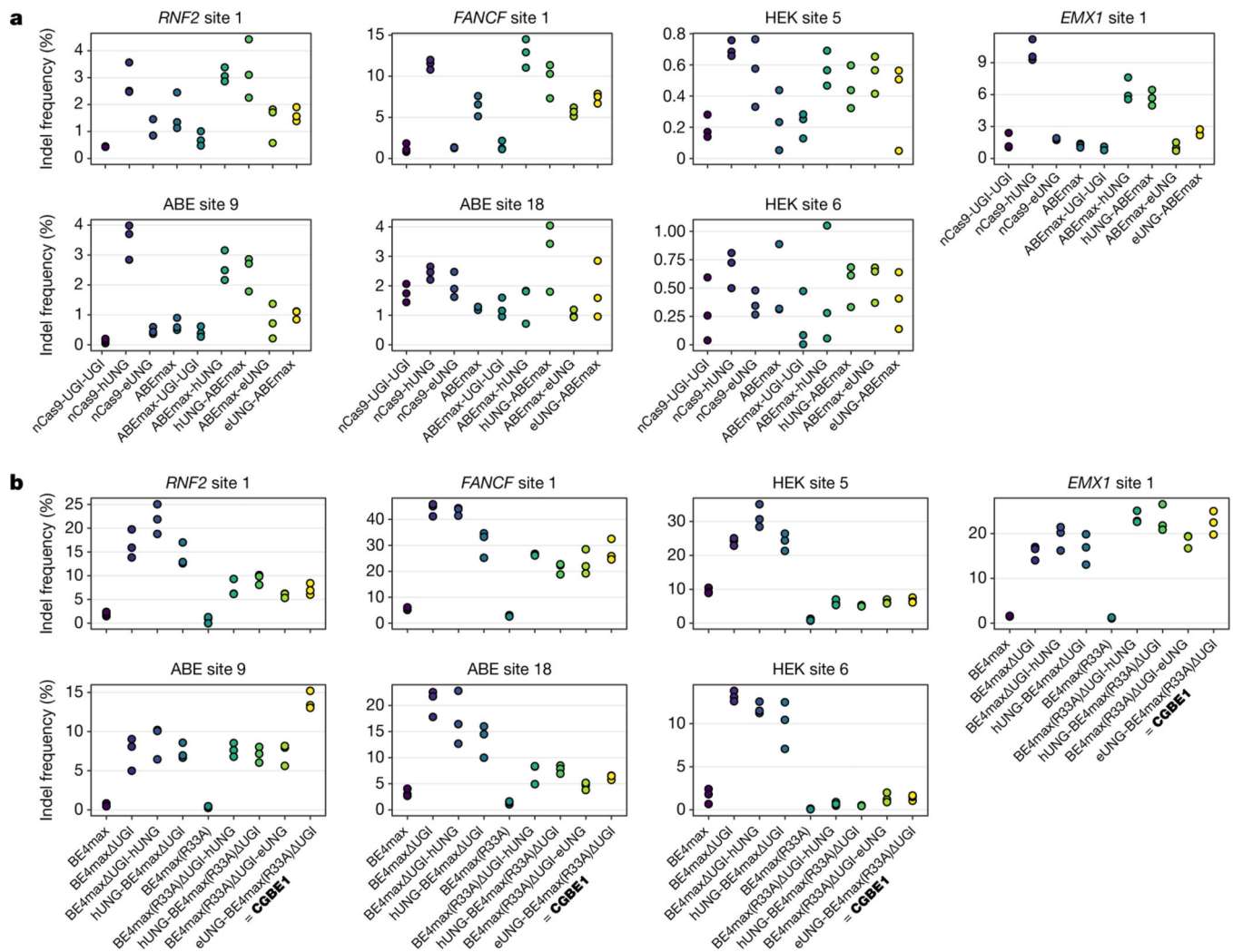
More information concerning statistical tests and experimental design are reported in the Nature Research Reporting Summary that is attached to this article.

Extended Data



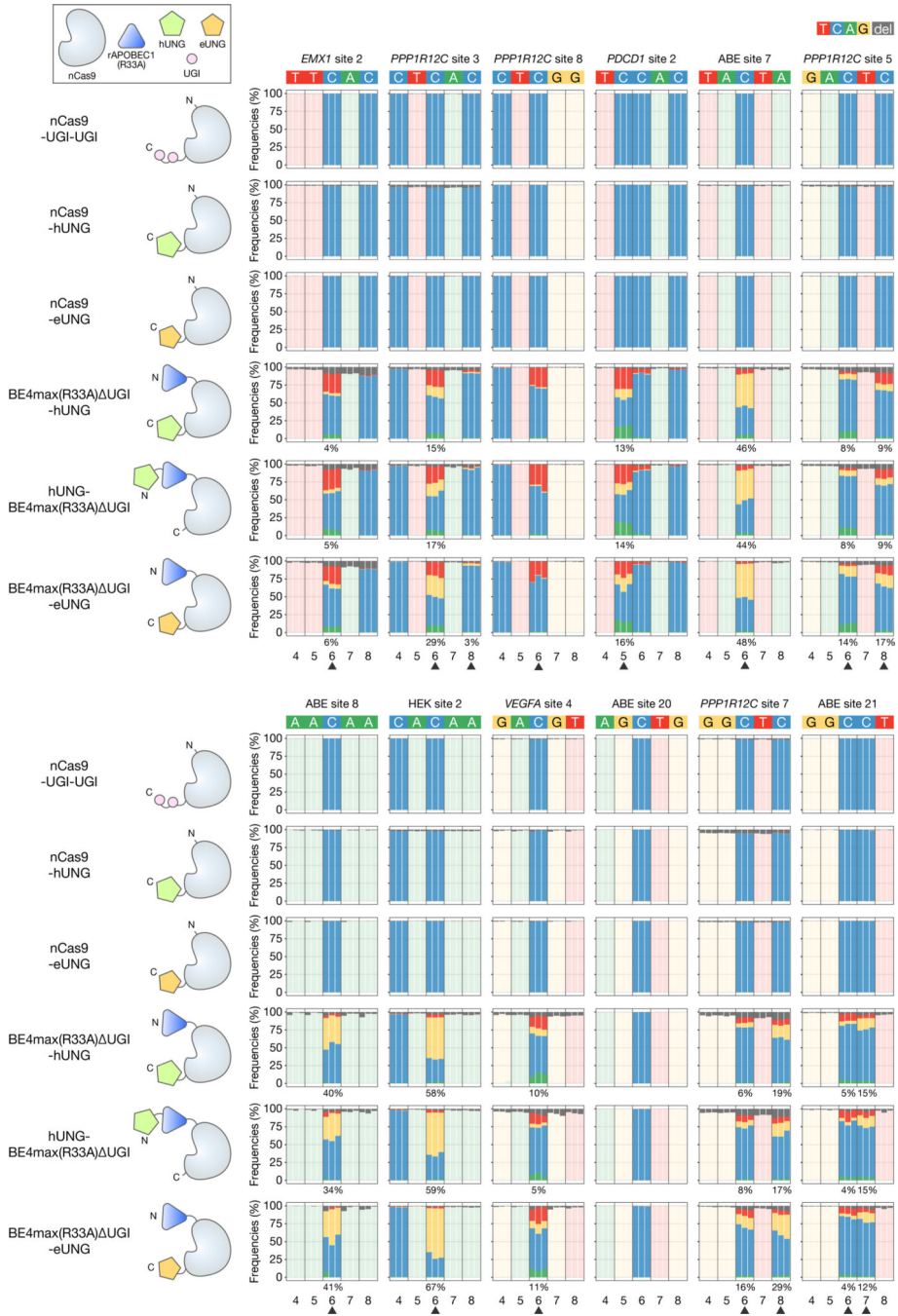
Extended Data Fig. 1. On-target activities of nCas9 controls, ABE variants and more CBE variants tested for C-to-G editing in HEK293T cells

Bar plots showing the on-target DNA base editing frequencies induced by nCas9 negative controls, ABE and ABE variants and other CBE variants with seven gRNAs in HEK293T cells. Editing frequencies of three independent replicates ($n = 3$) at each base are displayed side-by-side. Arrowheads indicate cytosines showing C-to-G edits by CGBE1.



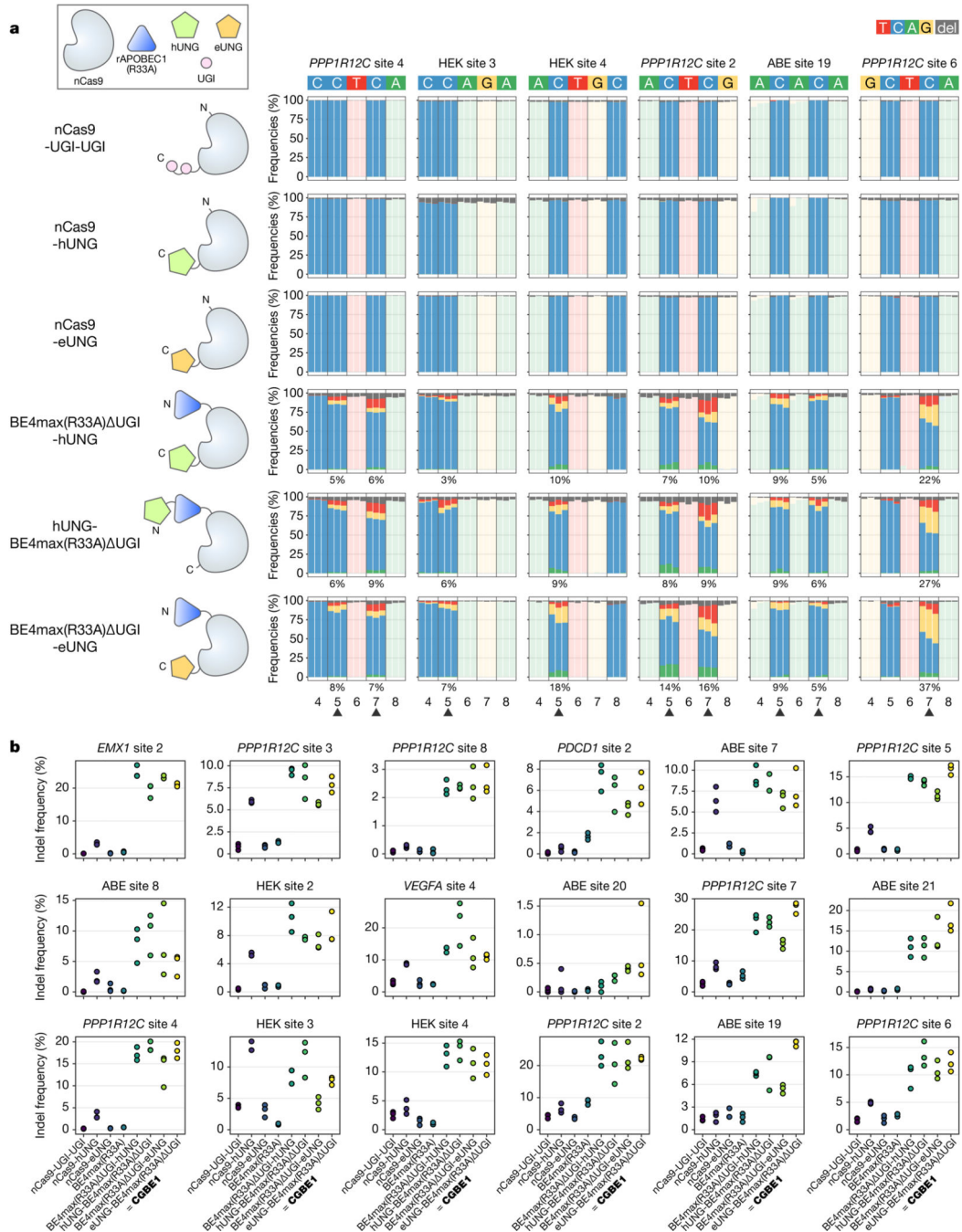
Extended Data Fig. 2. Indel frequencies of nCas9 controls, ABE variants and CBE variants tested for C-to-G editing in HEK293T cells

a,b, Dot plots representing percentage of alleles that contain an insertion or deletion across the entire protospacer from experiments with various base editor architectures reported in (a) Extended Data Fig. 1 or (b) Fig. 1b and Extended Data Fig. 1. Single dots represent individual replicates ($n = 3$ independent replicates).



Extended Data Fig. 3. On-target activities of nCas9 controls and CGBE1-related variants with 12 C6 gRNAs in HEK293T cells

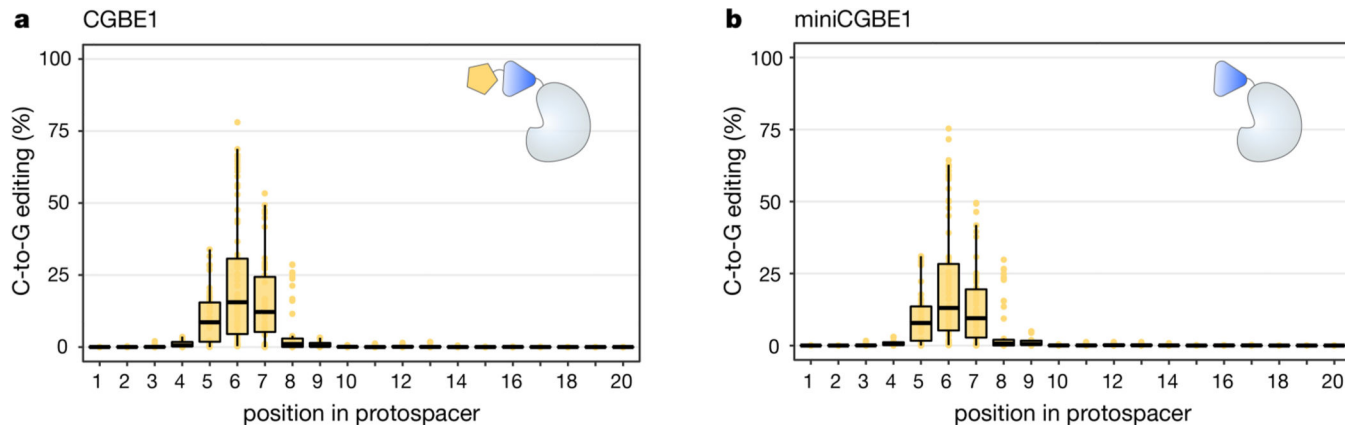
Bar plots showing the on-target DNA base editing frequencies of nCas9 controls and CGBE1-related variants using 12 gRNAs for sites with a C at position 6 (C6-sites) in HEK293T cells. Editing frequencies of three independent replicates ($n = 3$) at each base are displayed side-by-side.



Extended Data Fig. 4. On-target activities of nCas9 controls and CGBE1-related variants with 6 non-C6 gRNAs in HEK293T cells and indel frequencies across 18 targeted sites

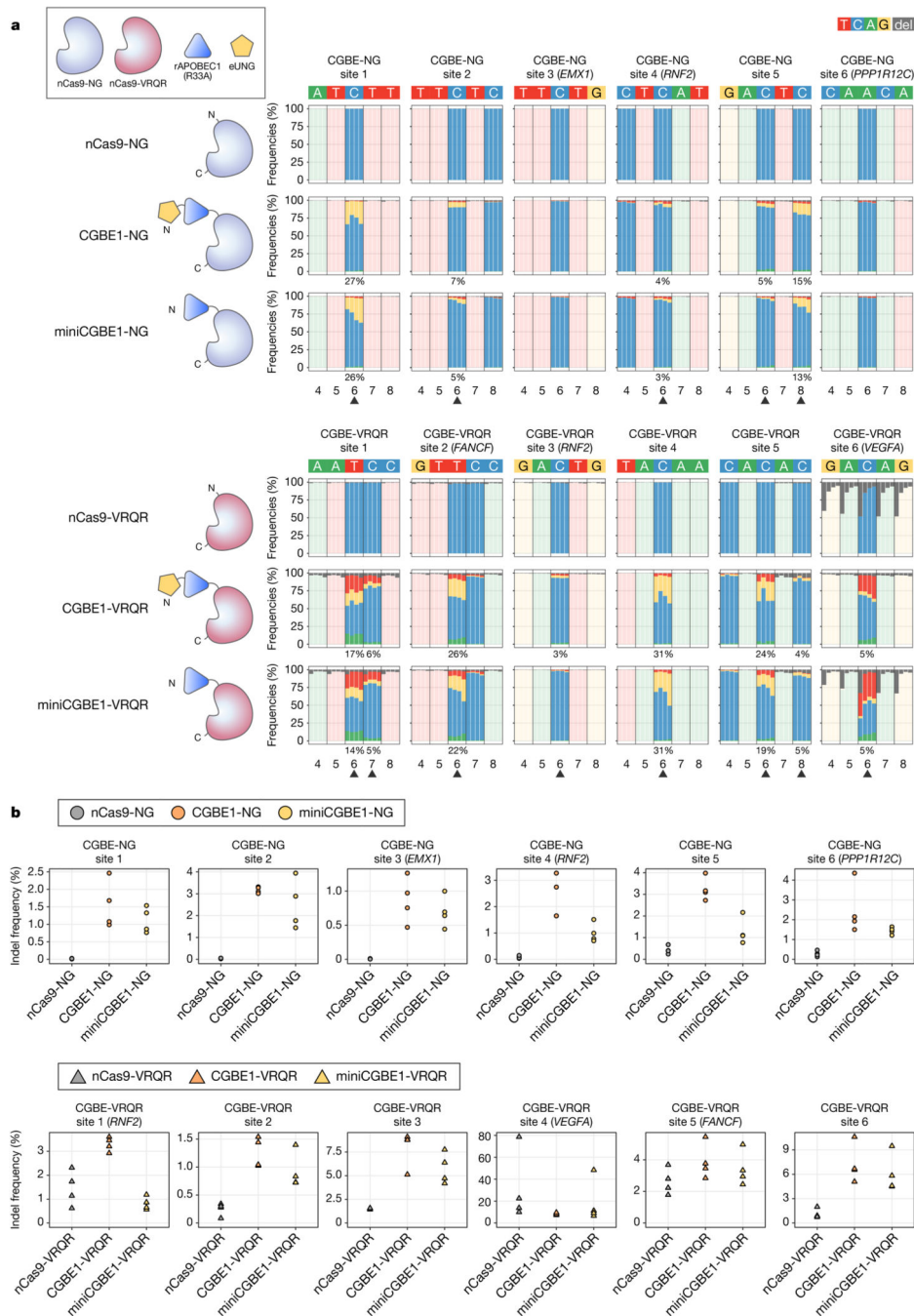
a. Bar plots showing the on-target DNA base editing frequencies of nCas9 controls and CGBE1-related variants using 6 gRNAs for sites with a C at position 4, 5, 7, or 8 (non-C6-sites) in HEK293T cells. Editing frequencies of three independent replicates ($n = 3$) at each base are displayed side-by-side. **b.** Dot plots representing percentage of alleles that contain an insertion or deletion across the entire protospacer from experiments with CGBE1-related

variants reported in Fig. 2a and b and Extended Data Figs. 3 and 4a. Single dots represent individual replicates ($n = 3$ independent replicates).



Extended Data Fig. 5. Aggregated distribution of C-to-G editing frequencies across protospacer with CGBE1 and miniCGBE1 in HEK293T cells

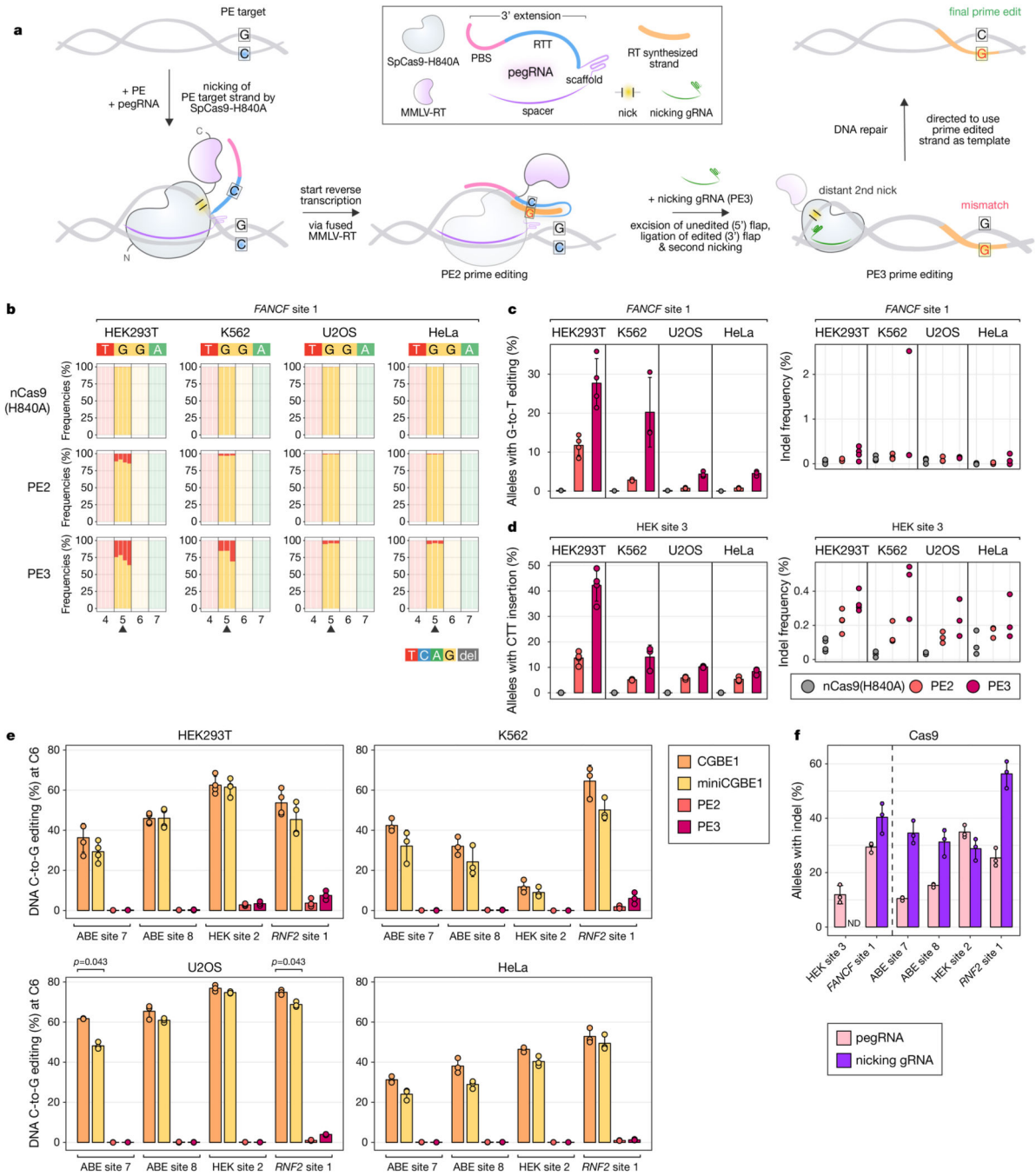
a,b, Dot and box plots representing the aggregate distribution of C-to-G (yellow) editing frequencies per nucleotide across the entire protospacer from experiments performed with CGBE1 (**a**) and miniCGBE1 (**b**) with all 48 tested gRNAs. Boxes span the interquartile range (IQR; 25th to 75th percentile), horizontal lines indicate the median (50th percentile), and whiskers extend to $\pm 1.5 \times$ IQR. Data points in plots represent full range of values plotted. Single dots represent individual replicates. The graphs were derived from the data shown in Fig. 3a,b ($n = 4$ independent replicates per site), and Supplementary Fig. 3a ($n = 3$ independent replicates per site).



Extended Data Fig. 6. On-target DNA editing activities of NG and VRQR variants of CGBE1 and miniCGBE1 in HEK293T cells

a, Bar plots showing the on-target DNA base editing frequencies induced by NG and VRQR variants of nCas9, CGBE1, and miniCGBE1 using 6 gRNAs that target AT-rich genomic loci with PAMs that are compatible with SpCas9-NG (NGT) and SpCas9-VRQR (NGAG) variants in HEK293T cells. Editing frequencies of four independent replicates ($n = 4$) at each base are displayed side-by-side. **b**, Dot plots representing percentage of alleles that contain an insertion or deletion across the entire protospacer from experiments with NG and

VRQR variants of CGBE1 and miniCGBE1 reported in **a**. Single dots or triangles represent individual replicates ($n = 4$ independent replicates).



Extended Data Fig. 7. Comparing the editing activities of CGBEs and PEs in multiple human cell lines

a, Schematic of prime editing (PE) used to install a C-to-G substitution. PE fusion protein consists of an SpCas9-H840A nickase fused to an engineered Moloney murine leukemia virus reverse transcriptase (MMLV-RT). The prime editing guide RNA (pegRNA) consists of a standard targetable SpCas9 gRNA that also harbors a 3' extension containing a primer

binding site (PBS) and a reverse transcription template (RTT) that encodes the desired edit. The PE2 system encompasses the prime editor fusion protein and a pegRNA. The PE3 system additionally includes a nicking gRNA (ngRNA). **b**, Bar plots showing the on-target DNA prime editing frequencies induced by nCas9(H840A), PE2 and PE3 using a pegRNA that targets *FANCF* site 1 across four human cancer cell lines. Gray overlay bars at top represent deletions at each nucleotide. Editing frequencies of four independent replicates ($n = 4$) for HEK293T cells or three independent replicates ($n = 3$) for K562, U2OS, and HeLa cells at each base are displayed side-by-side. Numbering on the bottom indicates the position of the base with 1 being the first nucleotide 3' of the pegRNA/Cas9-induced nick. Arrowheads indicate guanines that exhibit desired G-to-T prime edits. **c,d**, Bar and dot plots representing the average on-target DNA prime editing and indel frequencies of PE2 and PE3 targeting *FANCF* site 1 for G-to-T prime editing (**c**; data from the same experiment as **b**) and HEK site 3 for PE-induced CTT insertion (**d**) in 4 cell lines. Single dots represent individual replicates ($n = 4$ for HEK293T and $n = 3$ for K562, U2OS, and HeLa cells). Error bars represent standard deviation (s.d.). Measure of center for the error bars = mean. **e**, Bar and dot plots showing the average on-target DNA C-to-G base or prime editing frequencies induced by CGBE1, miniCGBE1, PE2 or PE3 on four genomic target loci. Single dots represent individual replicates ($n = 4$ for HEK293T and $n = 3$ for K562, U2OS, and HeLa cells). A two-tailed Student's t-test with p -values adjusted for multiple testing was used to calculate the shown p -values ($p = 0.043$ for both). Error bars represent (s.d.). Measure of center for the error bars = mean. **f**, Bar and dot plots representing the average frequency of alleles with indels (%) induced by pegRNAs and nicking gRNAs used in the experiments shown above (and *FANCF* site 1 +21 ngRNA control, Supplementary Table 9) with wild-type SpCas9 in HEK293T. pegRNAs/ngRNAs designed by Anzalone *et al.* (left) and by us (right) are separated by the dashed line. Single dots represent individual replicates ($n = 3$ independent replicates). Error bars represent (s.d.). Measure of center for the error bars = mean. ND, not done.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the National Institutes of Health (RM1 HG009490 and R35 GM118158) and by the Defense Advanced Research Projects Agency Safe Genes Program (HR0011-17-2-0042). J.G. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 416375182. L.M.L. was supported by a Boehringer Ingelheim Fonds MD fellowship. J.K.J. is additionally supported by the Robert B. Colvin, M.D. Endowed Chair in Pathology and the Desmond and Ann Heathwood MGH Research Scholar Award. We thank B. Kleinstiver and R. Walton for providing a codon-optimized version of VRQR-SpCas9 (D10A). We thank M.K. Clement for technical advice, K. Petri, P. Cabeceiras, J. Angstman, J. Hsu, C. Lareau, and M. Aryee for discussions and technical advice, and L. Paul-Pottenplackel for help with editing the manuscript. J.K.J. dedicates this work to the memory of Ki Sub Joung with gratitude for his lifelong advice, support, and inspiration.

References:

1. Rees HA & Liu DR Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat Rev Genet* 19, 770–788 (2018). [PubMed: 30323312]

2. Komor AC, Kim YB, Packer MS, Zuris JA & Liu DR Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533, 420–424 (2016). [PubMed: 27096365]
3. Nishida K. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* 353 (2016).
4. Gaudelli NM et al. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* 551, 464–471 (2017). [PubMed: 29160308]
5. Grünewald J. et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* 569, 433–437 (2019). [PubMed: 30995674]
6. Doman JL, Raguram A, Newby GA & Liu DR Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors. *Nat Biotechnol* (2020).
7. Grünewald J. et al. CRISPR DNA base editors with reduced RNA off-target and self-editing activities. *Nat Biotechnol* 37, 1041–1048 (2019). [PubMed: 31477922]
8. Kim HS, Jeong YK, Hur JK, Kim JS & Bae S. Adenine base editors catalyze cytosine conversions in human cells. *Nat Biotechnol* 37, 1145–1148 (2019). [PubMed: 31548727]
9. Komor AC 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).
10. Koblan LW et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat Biotechnol* 36, 843–846 (2018). [PubMed: 29813047]
11. Gehrke JM et al. An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities. *Nat Biotechnol* 36, 977–982 (2018). [PubMed: 30059493]
12. Tsai SQ et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* 33, 187–197 (2015). [PubMed: 25513782]
13. Nishimasu H. et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* 361, 1259–1262 (2018). [PubMed: 30166441]
14. Kleinstiver BP et al. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* 529, 490–495 (2016). [PubMed: 26735016]
15. Anzalone AV et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576, 149–157 (2019). [PubMed: 31634902]
16. Kleinstiver BP et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* 523, 481–485 (2015). [PubMed: 26098369]
17. Hu JH et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* 556, 57–63 (2018). [PubMed: 29512652]
18. Walton RT, Christie KA, Whittaker MN & Kleinstiver BP Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* 368, 290–296 (2020). [PubMed: 32217751]
19. Guo C. et al. Transversions have larger regulatory effects than transitions. *BMC Genomics* 18, 394 (2017). [PubMed: 28525990]
20. Kim YB et al. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat Biotechnol* 35, 371–376 (2017). [PubMed: 28191901]
21. Thuronyi BW et al. Continuous evolution of base editors with expanded target compatibility and improved activity. *Nat Biotechnol* 37, 1070–1079 (2019). [PubMed: 31332326]

Methods-only References

22. Clement K. et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* 37, 224–226 (2019). [PubMed: 30809026]

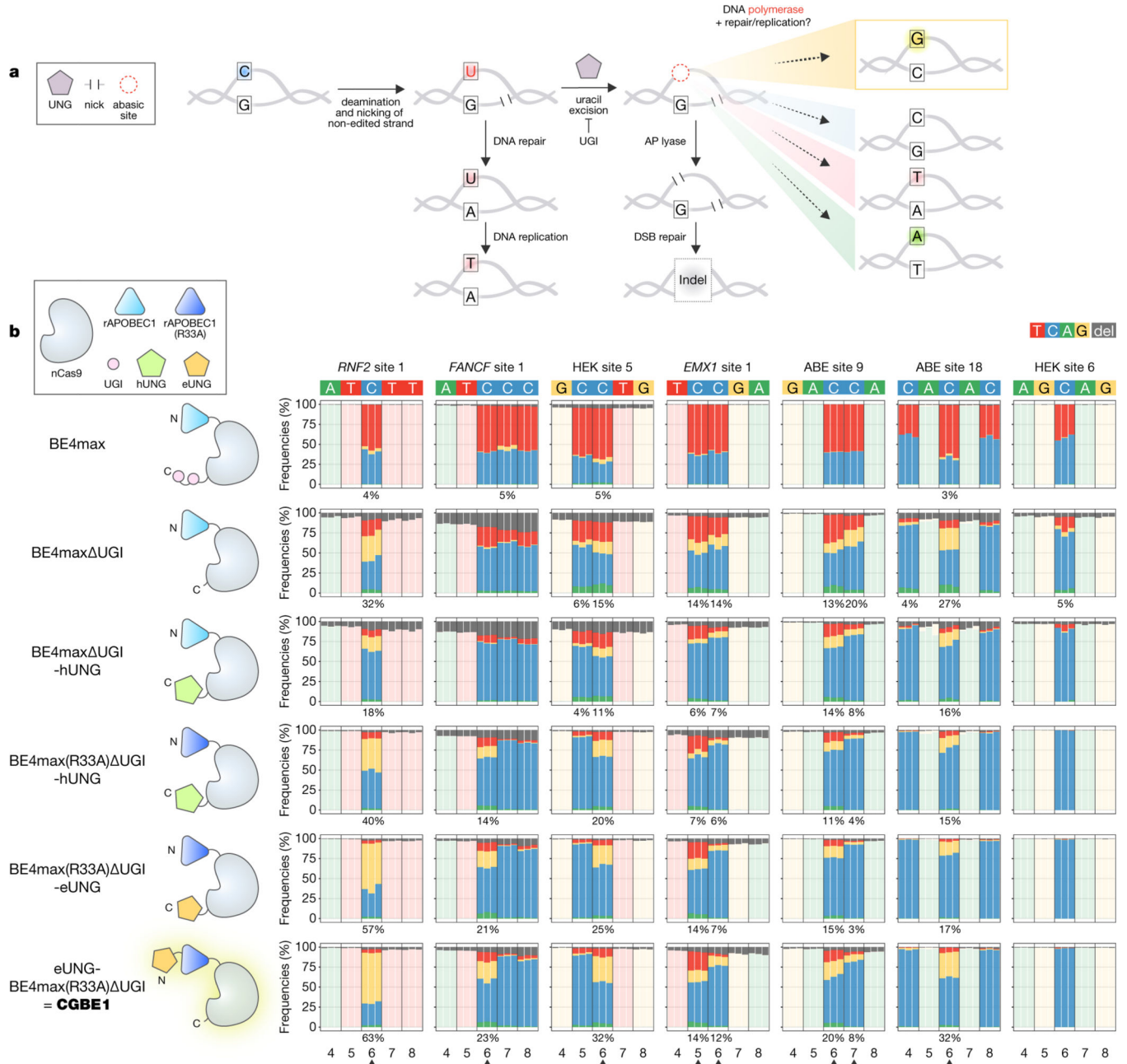


Figure 1: Engineering of a C-to-G base editor

a. Schematic of potential cellular mechanisms and outcomes downstream of cytosine deamination by base editors. Uracil excision by endogenous uracil *N*-glycosylase (UNG, purple pentagon), nicking on the non-edited strand by Cas9 nickase, followed by DNA repair and replication can lead to diverse editing outcomes. C, Cytosine; G, Guanine; U, Uracil; T, Thymine; A, Adenine; UGI, uracil glycosylase inhibitor; AP lyase, apurinic/apyrimidinic site lyase; DSB, double strand break.

b. Bar plots showing on-target DNA base editing frequencies with various base editor architectures using seven gRNAs targeting genomic sites in HEK293T cells. N and C in the base editor illustrations indicate amino-terminal and carboxy-terminal ends, respectively. Gray overlay bars at top represent

deletions at each nucleotide. Target cytosines are highlighted. Editing frequencies of three independent replicates ($n = 3$) at each base are displayed side-by-side. Percentage values below specific cytosine bases indicate the average C-to-G editing observed (values below 3% not reported). Numbering on the bottom indicates position of the base in the protospacer with 1 being the most PAM-distal base. Arrowheads indicate cytosines with C-to-G edits.

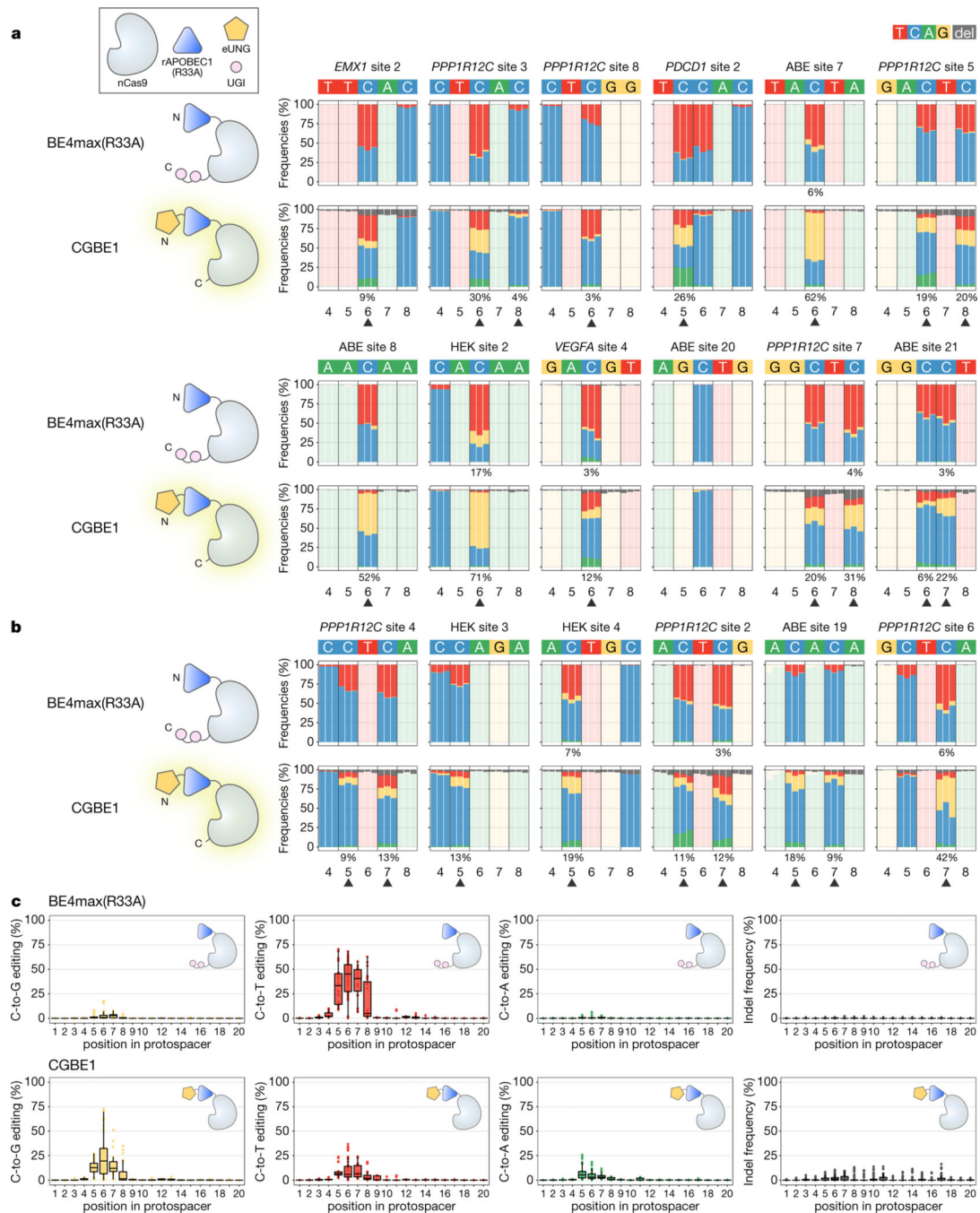


Figure 2: Additional characterization of CGBE1 on-target editing activities in HEK293T cells
a,b, Bar plots showing the on-target DNA base editing frequencies induced by BE4max(R33A) and CGBE1 using 12 gRNAs for sites with a C at position 6 (C6-sites; **a**) and 6 gRNAs for sites with a C at position 4, 5, 7, or 8 (non-C6-sites; **b**) in HEK293T cells. Editing frequencies of three independent replicates ($n = 3$) at each base are displayed side-by-side. **c**, Dot and box plots representing the combined distribution of C-to-G (yellow), C-to-T (red), C-to-A (green), and indel (gray) frequencies per nucleotide across the entire protospacer from experiments performed with BE4max(R33A) and CGBE1 using 25 guides.

Boxes span the interquartile range (IQR; 25th to 75th percentile), horizontal lines indicate the median (50th percentile), and whiskers extend to $\pm 1.5 \times \text{IQR}$. Data points in plots represent full range of values plotted. Single dots represent individual replicates ($n = 3$ independent replicates per site). The graphs were derived from the data shown in Figs. 1b, 2a and b, and Extended Data Fig. 1.

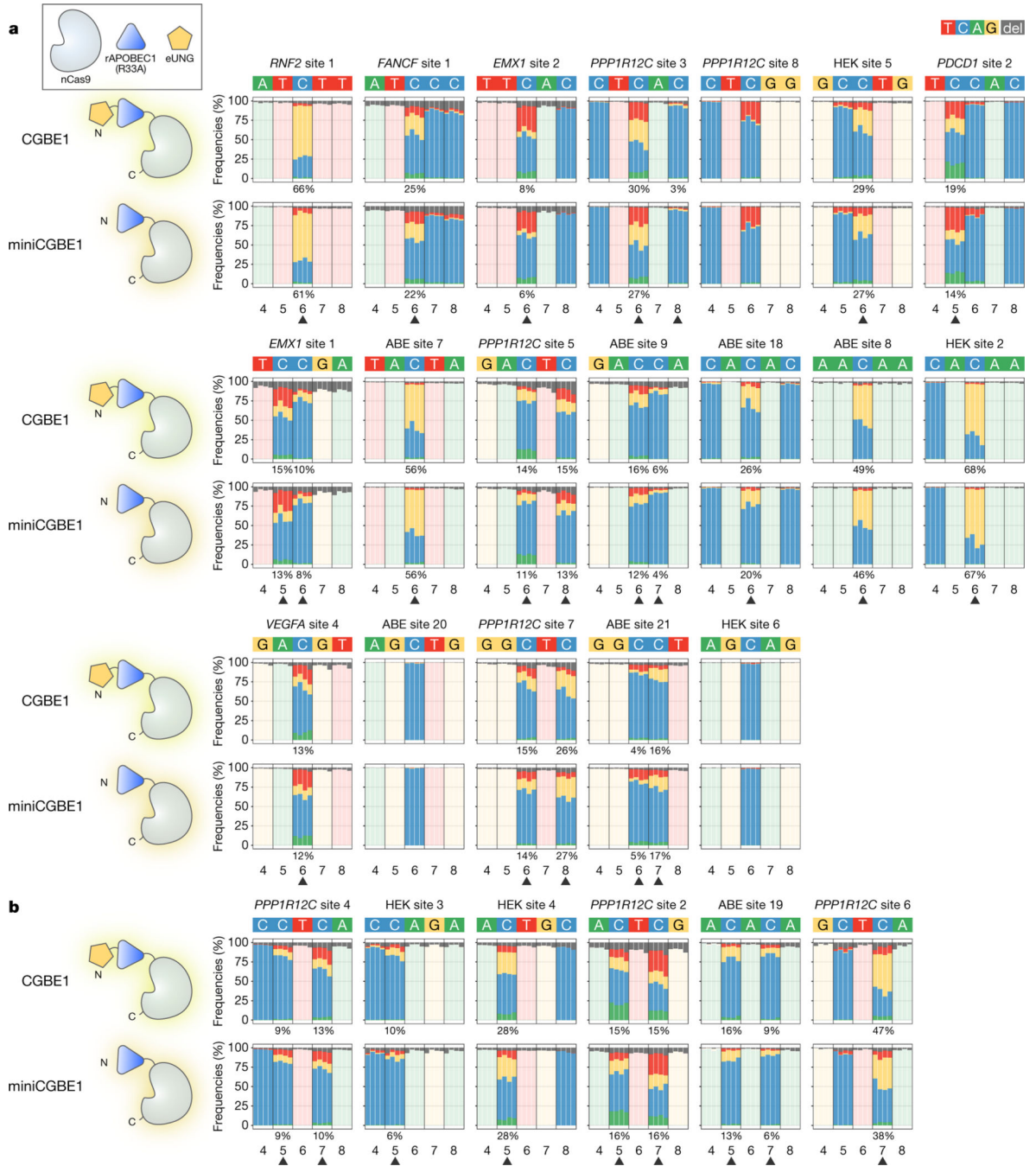


Figure 3: Comparison of CGBE1 and miniCGBE1 on-target editing activities in HEK293T cells
a,b, Bar plots showing the on-target DNA base editing frequencies of CGBE1 and miniCGBE1 using 19 gRNAs for sites with a C at position 6 (C6-sites; **a**) and 6 gRNAs for sites with a C at position 4, 5, 7, or 8 (non-C6-sites; **b**) in HEK293T cells. Editing frequencies of four independent replicates ($n = 4$) at each base are displayed side-by-side.