Title

The Wild-Type tRNA Adenosine Deaminase Enzyme TadA Is Capable of Sequence-Specific DNA Base Editing

Journal

ChemBioChem, 24(16)

ISSN

1439-4227

Authors

Ranzau, Brodie L
Rallapalli, Kartik L
Evanoff, Mallory
et al.

Publication Date

2023-03-22

DOI

10.1002/cbic.202200788

The wild-type tRNA adenosine deaminase enzyme TadA is capable of sequence-specific DNA base editing

Authors: Brodie L. Ranzau[1], Dr. Kartik L. Rallapalli[1], Mallory Evanoff[1], Prof. Francesco Paesani[1-4], and Prof. Alexis C. Komor[1*]

Affiliations: [1]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA 92093, USA

[2] Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, California 92093, USA

[3] Materials Science and Engineering, University of California San Diego, La Jolla, California 92093, USA

[4]San Diego Supercomputer Center, University of California San Diego, La Jolla, California 92093, USA

*Correspondence: akomor@ucsd.edu

Abstract

Base editors are genome editing tools that enable site-specific base conversions via the chemical modification of nucleobases in DNA. Adenine base editors (ABEs) convert A•T to G•C base pairs in DNA by utilizing an adenosine deaminase enzyme to modify target adenosines to inosine intermediates. Due to the lack of a naturally occurring adenosine deaminase that can modify DNA, ABEs were evolved from a tRNA-deaminating enzyme, TadA. Previous experiments utilizing an ABE comprised of a wild-type (wt) TadA showed no detectable activity on DNA, and directed evolution was therefore required to enable this enzyme to accept DNA as a substrate. Here we show that wtTadA can perform base editing in DNA in both bacterial and mammalian cells, with a strict sequence motif requirement of TAC. We leverage this discovery to optimize a reporter assay to detect base editing levels as low as 0.01%. Finally, we use this assay along with molecular dynamics simulations of full ABE:DNA complexes to better understand how the sequence recognition of mutant TadA variants change as they accumulate mutations to better edit DNA substrates.

Introduction

Base editing is a genome editing technique that enables the targeted introduction of single nucleotide variants (SNVs) without using double strand breaks (DSBs)[1,2]. These mutations are introduced using base editors, which consist of a single-stranded DNA (ssDNA)-modifying enzyme linked to a catalytically impaired or inactivated Cas9 (nCas9 or dCas9, **Figure 1A**)[3]. During the process of base editing, Cas9 uses a guide RNA (gRNA) to bind to a target location in the genome, driven by sequence complementarity between the DNA (called the protospacer) and the 5′ end of the gRNA (called the spacer). The protospacer must also be next to a protospacer adjacent motif (PAM) for Cas9 binding. Upon DNA binding, Cas9 forms an R-loop, exposing a small window of ssDNA to the ssDNA-modifying enzyme[4], which then modifies the target nucleotide(s) within the protospacer (**Figure 1A**). DNA repair or replication predictably resolves the modified nucleotide to a canonical base pair to complete the process of base editing. Thus far, two classes of ssDNA-modifying enzymes (cytidine deaminases and adenosine deaminases) have been utilized for base editing. Cytosine base editors (CBEs) facilitate C•G to T•A base pair conversions through the deamination of cytosine to uracil, and adenine base editors (ABEs) facilitate A•T to G•C base pair conversions through the deamination of adenosine to inosine (**Figure 1A**).

While many CBEs have been developed by linking naturally occurring ssDNA-modifying enzymes to dCas9, there are no natural enzymes that deaminate adenosine within DNA[1,5–7]. ABEs were therefore developed by evolving a natural tRNA-modifying enzyme (*E. coli* TadA) to accept DNA as a substrate (**Figure 1B-D**)[2]. Initial experiments tested the ability of wild-type (wt) TadA to facilitate A•T to G•C base editing when fused to nCas9 (the resulting construct is called ABE0.1) and observed no detectable base editing with this construct at six genomic targets in HEK293T cells. However, after accumulating 15 mutations over seven rounds of directed evolution, the resulting ABE7.10 construct could introduce A•T to G•C mutations with high efficiencies across diverse sequence contexts (**Figure 1C**). Additional rounds of directed evolution have since been undertaken to engineer ABE8 (**Figure 1C-E**) and ABE9 editors, which perform base editing with even higher efficiencies and faster kinetics[8–10].

Expansion of the base editor toolbox will require the development of additional ssDNA-modifying enzymes. However, despite the successful development of ABEs, thus far no other RNA-modifying enzymes have been repurposed as base editors by engineering or evolving them to accept DNA as a substrate. To better inform the creation of new base editors, we have sought to better understand the evolution of ABEs, particularly the first-round mutation, D108N (**Figure 1C-D**), which is sufficient and imperative for imparting TadA with detectable DNA editing activity[11,12].

In nature, tRNA editing enzymes have evolved to recognize specific RNA structures or sequences[13]. As a result, they modify only a specific nucleotide on one or multiple tRNAs within the cell. Specifically, *E. coli* TadA recognizes the UACG motif within the anticodon loop of the $tRNA^{Arg}_{ACG}$ and deaminates the indicated adenosine, which is in the wobble position **(Figure 1B)**[14]. Several studies found that the mutated TadA in ABE7.10 (which we will refer to as TadA7.10) has retained its natural RNA-editing ability, with the majority of off-target RNA edits by TadA7.10 occurring at UACG motifs throughout the transcriptome[12,15–17]. We subsequently observed that the wtTadA enzyme also efficiently deaminates adenosines in these motifs transcriptome-wide, which was quite surprising given wtTadA was previously thought to discriminately target the $tRNA^{Arg}_{ACG}$ from other RNAs in the cell[12].

While initial experiments concluded that ABE0.1 was incapable of editing DNA, none of the tested protospacers contained a TACG target (**Figure 2A**)[2]. Given wtTadA's natural affinity for the UACG sequence, even outside of its tRNA anticodon loop structure, we sought to evaluate if ABE0.1 could edit DNA in these sequence contexts. Here we show that ABE0.1 is capable of A•T to G•C editing in TACG motifs, but with efficiencies of less than 0.2% as evaluated with next generation sequencing (NGS). We then describe the development of a fluorescence-based reporter of ABE0.1 activity that incorporates the TACG motif at the target site. This reporter is able to detect A•T to G•C editing efficiencies as low as 0.01% (below the detection of NGS) and is easily modified to enable evaluation of editing at additional sequence contexts. The enhanced sensitivity of these reporters allows facile characterization of the sequence specificity of several ABE variants. Additionally, we revisit the activity of ABE0.1 in an *E. coli*-based antibiotic survival assay under optimal target base editing conditions with respect to editing window and sequence motif and detect adenosine deamination activity by ABE0.1 in this optimized system. These collective experiments explain that the prior conclusion that wtTadA is incapable of editing DNA was due to the choice of sequence targets. Finally, using both experimental and computational methods, we elucidate the molecular factors governing the sequence specificity of these ABE variants. These results provide a general framework for the identification of other wild-type RNA-modifying enzymes with inherent DNA editing activities, with the potential to aid the development of novel base editors.

Results and Discussion

ABE0.1 can edit DNA at TACG motifs

We previously reported that ABE0.1 showed high (>50% A to I conversion efficiencies) gRNA-independent editing at RNA sites that contain the UACG motif from the natural $\text{tRNA}^{\text{Arg}}_{\text{ACG}}$ target of wtTadA. In fact, at all six mRNA sites that were analyzed, ABE0.1 displayed equivalent or higher editing than ABE7.10[12]. These observations suggested to us that perhaps ABE0.1 could also deaminate adenosines in DNA that are embedded in this motif. We found that A•T to G•C DNA editing activity by ABE0.1 was previously evaluated at only non-TACG motifs (**Figure 2A**)[2]. We therefore identified three genomic loci containing this motif and designed protospacers that place the target A in position 5 or 7 (**Figure 2B**).

We first evaluated the quality of the protospacers by transfecting HEK293T cells with plasmids encoding wtCas9 and a non-targeting gRNA or a gRNA targeting one of the three protospacers. Cells were then lysed after 72 hours, genomic loci of interest were amplified, and indel introduction efficiencies were quantified by NGS. All gRNAs facilitated indel introduction efficiencies ranging from 30-89%, indicating efficient targeting and binding by Cas9 (**Figure 2B**). We then repeated the experiment with ABE0.1 and quantified A•T to G•C editing efficiencies. When ABE0.1 was originally tested for genomic DNA (gDNA) editing activity, a single monomer of TadA was fused to Cas9n (SI Figure 1B)[2]. However, later generations of ABEs utilized dimeric TadA fused to Cas9n, as these demonstrated increased editing efficiency[2]. Here, we tested both monomeric and dimeric ABE0.1 constructs (ABE0.1m and ABE0.1d, respectively), as wtTadA is known to dimerize when editing its native $\text{tRNA}^{\text{Arg}}_{\text{ACG}}$ substrate, but the mechanism of DNA editing by evolved TadA enzymes is not currently known[14,18–20]. Specifically, monomeric ABE7.10 and ABE8 constructs do not show any apparent decrease in editing efficiency compared to their dimeric counterparts, but these constructs may be dimerizing in trans to perform deamination (**Figure 1A**)[11,15,17,21]. These initial experiments did not show A•T to G•C editing levels above background (SI Figure 1C), but we hypothesized that this may be due to inefficient transfection efficiency and/or ABE0.1 expression.

To enrich for transfected cells with high ABE0.1 expression, we next utilized an ABE0.1-P2A-EGFP construct (**Figure 2C**), in which *ABE0.1* and *EGFP* are transcribed on the same mRNA transcript but translated into separate proteins due to self-cleavage by the P2A linker during translation[22]. We repeated the prior experiment with these EGFP-containing constructs, and used fluorescence activated cell sorting (FACS) to sort for cells with the top 20-30% EGFP fluorescence (SI Figure 2). We additionally included ABE8e targeting and non-target controls. NGS analysis of the cells enriched for EGFP fluorescence revealed robust levels of A•T to G•C editing (greater than 56 ± 6.4% at all three sites) with the ABE8e construct, demonstrating that these protospacers are capable of efficient targeting and editing by ABEs (SI Figure 3B). Importantly, at two of the three sites, we observed A•T to G•C editing above non-targeting control levels with the dimeric ABE0.1d construct (**Figure 2D**): 0.11 ± 0.01% A•T to G•C editing was observed at the *PSMB2* site, and 0.23 ± 0.03% A•T to G•C editing was observed at the *SCAP* site (see Methods for statistical analysis details). We also observed A•T to G•C editing above non-targeting control levels by the ABE0.1m monomeric construct at the *SCAP* site (0.14 ± 0.03% A•T to G•C editing), but not at the *PSMB2* site (**Figure 2D**). Editing above background levels was not observed at the *FANCF* site for any of the editors (**Figure 2D**), which may be due to the location of the target A within this protospacer (position 7 in the *FANCF* protospacer, and position 5 in the *PSMB2* and *SCAP* protospacers). These data are the first to demonstrate that wtTadA possesses DNA editing activity when the target A is optimally positioned and embedded within the TACG sequence motif.

A GFP reporter assay enhances editing detection

The observation of DNA editing by wtTadA is especially noteworthy in the context of using the development of ABE as a model for the generation of new base editors; directed evolution efforts that aim to enhance an existing activity, rather than evolve a completely new activity, are much more successful[23,24]. The ability to detect low levels of DNA editing by additional enzymes would greatly enhance future efforts to develop novel base editors. However, detecting such low levels of editing using HTS is laborious and is limited by the error rate of the instrument, which is typically 0.1% when performing targeting amplicon sequencing[25,26]. We therefore sought to more robustly detect such low levels of DNA editing through the development of a plasmid-based colorimetric reporter for A•T to G•C editing. We reasoned that the combination of a plasmid substrate (which would provide the editor with more substrate targets and therefore additional editing opportunities within each cell) and a fluorescence-based turn-on signal (which would create a more drastic and easily detected phenotypic change upon base editing, therefore enhancing the signal) would collectively increase the limit of detection of editing. We adapted a recently described EGFP reporter for cytosine base editing which utilizes an mCherry-P2A-EGFP construct (**Figure 3A**)[7]. mCherry fluorescence is used to monitor transfection efficiency of the reporter, while the *EGFP* gene is mutated to express an inactive variant of EGFP (which we call dGFP). The mutation is targeted by a base editor to correct the loss-of-function mutation, leading to the expression of active EGFP (**Figure 3A**), which can be detected with fluorescence microscopy and/or quantified with flow cytometry. As mentioned previously, we have shown that wtTadA can deaminate UACG motifs in mRNA, independent of Cas9 targeting[12]. Therefore, to avoid EGFP fluorescence due to editing of the *EGFP* mRNA, we incorporated the target adenosine into the non-coding strand of the reporter. With this in mind, we developed two reporters in which a G•C to A•T mutation within a TGCG motif resulted in a non-functional EGFP protein. One uses the A111V EGFP mutant (in which the target A is in the second position of the Val codon, and position 5 within the protospacer, **Figure 3A**), and one uses the H182Y EGFP mutant (in which the target A is in the first position of the Tyr codon, and position 6 within the protospacer, SI Figure 4A)[27,28]. To incorporate the TACG motif in this second reporter, we additionally installed a D181S mutation next to the H182Y mutation, which maintains the electrostatic properties of this outward facing residue and therefore does not abolish fluorescence itself. Between these two reporters, all three neighboring bases within the TACG motif could be mutated silently, allowing for facile evaluation of the sequence context requirements of ABE variants.

We then transfected HEK293T cells with plasmids encoding ABE8e, one of the mCherry-P2A-dGFP reporter constructs, and either a non-targeting gRNA or a gRNA targeting the dGFP loss-of-function mutation. After 72 hours, cells were imaged by fluorescence microscopy (**Figure 3B**) and analyzed by flow cytometry (**Figures 3C and 3E-F,** and SI Figure 4B-D). In non-targeting controls, the A111V reporter showed a complete knock-out of EGFP fluorescence, but the H182Y reporter showed low levels of background fluorescence (SI Figure 4B). We found that reducing the voltage of the photomultiplier tube used for EGFP detection in combination with appropriate gating allowed us to discriminate between background fluorescence and cells with edited plasmid (SI Figure 4B). With these modifications, we observed efficient and gRNA-dependent EGFP turn-on for both reporters with ABE8e (65.5 ± 5.2% of transfected cells displayed EGFP fluorescence with the A111V reporter, and 60.2 ± 4.6% of transfected cells displayed EGFP fluorescence with the H182Y reporter, **Figure 3E** and SI Figure 4C), demonstrating their utility for the detection of adenine base editing.

We then repeated the experiment with the A111V reporter using both ABE0.1m and ABE0.1d, and again observed gRNA-dependent EGFP fluorescence with both ABE constructs (for the ABE0.1m editor, 0.74 ± 0.08% of transfected cells displayed EGFP fluorescence**, Figure 3C-E**). Both ABE0.1m and ABE0.1d displayed EGFP fluorescence above background, and we observed no statistically significant difference in editing activity between the two constructs, as measured by percent of transfected cells with EGFP fluorescence (**Figure 3D**). We then repeated this experiment with ABE1.1m and ABE1.1d, and again observed no statistically significant difference in editing activity between the monomeric and dimeric constructs (**Figure 3D**). To further confirm this, we measured editing efficiency of ABE1.1m and ABE1.1d by NGS at the three TACG-containing genomic sites from our previous experiments, and found no consistent difference in A•T to G•C editing efficiencies between these two constructs (SI Figure 3D). This indicates that ABE0.1 and ABE1.1 either perform DNA editing as monomers, or the intracellular concentration of ABE protein is high enough for dimerization in trans. We therefore performed all future experiments with monomeric constructs. Finally, we will note that in addition to a subsequent increase in the percent of transfected cells displaying EGFP fluorescence when comparing ABE0.1m to ABE1.1m to

ABE8e, we also observed a subsequent increase in the median fluorescence intensity (MFI) of the EGFP-positive cells when comparing ABE0.1m to ABE1.1m to ABE8e for both reporters (16 ± 3 AU for ABE0.1m, 143 ± 46 AU for ABE1.1m, and 376 ± 109 AU for ABE8e with the A111V reporter, and 16 ± 3 AU for ABE0.1m, 81 ± 17 AU for ABE1.1m, and 110 ± 29 AU for ABE8e with the H182Y reporter, **Figure 3F** and SI Figure 4D).

To test the sensitivity of both reporters, we performed an experiment in which the A111V or H182Y dGFP-based reporter plasmid was mixed with decreasing amounts of the equivalent wild-type EGFP-based reporter plasmid, ranging from 100% wild-type to 0% wild-type. We transfected the resulting plasmid mixtures into HEK293T cells, waited 72 hours, and analyzed the cells by flow cytometry. With both reporters, we found that cells with EGFP fluorescence above background levels can be reliably detected when 0.01% of the total plasmid had the wild-type EGFP sequence (SI Figure 5A). At this minimum detection level, EGFP fluorescence was observed in 0.79 ± 0.13 % of cells transfected with the A111V reporter, and in 0.29 ± 0.03 % of cells transfected with the H182Y reporter (SI Figure 5A). We additionally analyzed the MFI of EGFP-positive cells of all samples for both reporters. We found that using this analysis method, the limit of detection for the A111V reporter was still around 0.01% (in which case the MFI of EGFP-positive cell was 13.6 ± 1.1 AU, versus 8.1 ± 0.5 AU for the negative control, SI Figure 5B). The limit of detection for the H182Y reporter, however, was 1% (in which case the MFI of EGFP-positive cell was 28.2 ± 1.7 AU, versus 19.3 ± 2.2 AU for the negative control, SI Figure 5B) using this analysis method, potentially due to a combination of the higher background signal and reduced voltage used to observe this reporter (SI Figure 5B). These data therefore demonstrate that these fluorescent reporters boast limits of detection an order of magnitude lower than NGS-based strategies. Further, a combination of these two analysis methods can be used to draw conclusions regarding relative comparisons of editing efficiencies among ABE variants.

ABE0.1 shows strict sequence requirements for editing

We next sought to analyze the sequence requirement of ABE0.1m by installing all possible single mutations at each of the three positions surrounding the target A in the $T^{-1}\underline{A}^{0}C^{+1}G^{+2}$ motif and assessing editing by ABE0.1m with each of these mutated targets. The A111V reporter enables both the T^{-1} and G^{+2} positions to be mutated, as these reside at the wobble positions of the Arg and Val codons (**Figure 3A**), and the H182Y reporter enables mutation of the C^{+1} as this resides at the wobble position of a Ser codon (SI Figure 4A).

We transfected HEK293T cells with plasmids encoding ABE0.1m, each of the mCherry-P2A-dGFP(A111V) or mCherry-P2A-dGFP(H182Y) reporter constructs, and either a non-targeting gRNA or a gRNA targeting the dGFP loss-of-function mutation. After 72 hours, cells were analyzed by flow cytometry (**Figure 4A-C** and SI Figure 6A-C). Mutation of the T^{-1} position had a drastic effect on editing activity by ABE0.1m, as measured by percent of transfected cells with EGFP fluorescence. Specifically, mutation of T^{-1} to A caused a 13.1 ± 4.7-fold decrease, mutation of T^{-1} to C caused a 6.1 ± 2.8-fold decrease, and mutation of T^{-1} to G caused a 6.8 ± 4.2-fold decrease (**Figure 4A**). However, we will note that ABE0.1m editing activity with these reporters was still above levels of non-targeting controls, demonstrating that wtTadA can deaminate at sequence motifs beyond T$\underline{A}$CG, but likely with efficiencies below what can be detected with NGS. When analyzing the MFI of EGFP-positive cells of these samples, only the "wild-type" TACG sample was statistically significantly higher than that of its non-targeting gRNA control (SI Figure 6A), suggesting that editing levels of these samples are very close to the limit of detection of the reporter.

Mutation of the C^{+1} position appeared to have a smaller effect on the editing activity by ABE0.1m, but this may be explained by the use of the H182Y reporter, which results in slightly lower overall editing efficiencies as well as higher background signal (when comparing ABE0.1m editing at the two T$\underline{A}$CG reporters, there is a 3.0 ± 0.6-fold overall reduction in editing observed with the H182Y reporter compared to the A111V reporter, as measured by percent of transfected cells with EGFP fluorescence). Mutation of C^{+1} to T (a 2.1 ± 0.6-fold decrease) and C^{+1} to G (a 2.5 ± 0.7-fold decrease) caused the largest decreases, as measured by percent of transfected cells with EGFP fluorescence, with both of these sequences showing EGFP fluorescence levels near those of the non-targeting controls (**Figure 4B**). Interestingly, the C^{+1} to A mutation was not statistically significantly different from the WT C^{+1} reporter (1.2 ± 0.3-fold decrease in editing), showing that wtTadA can potentially edit at TACG and TAAG motifs. Again, when analyzing the MFI of EGFP-positive cells of these samples, only the "wild-type" TACG sample was statistically significantly higher than that of its non-targeting gRNA control (SI Figure 6B), consistent with the lower sensitivity of this reporter when using this analysis method.

Mutation of the G^{+2} position to C (which was assayed using the A111V reporter) caused a less drastic decrease in EGFP fluorescence levels than mutation of the T^{-1} position (which was evaluated using the same reporter), causing a 3.2 ± 0.4-fold decrease in the percent of transfected cells with EGFP fluorescence (**Figure 4C**). Mutation of G^{+2} to T (a 1.8 ± 0.2-fold decrease) and G^{+2} to A (a 1.4 ± 0.2-fold decrease) showed smaller effects on editing efficiency. The G^{+2} base appears to be favored, but the other bases can be better tolerated in this position than at the T^{-1} and C^{+1} positions. These results collectively indicate that the identity of the T^{-1} and C^{+1} bases are important for the editing of DNA by wtTadA, as mutations at these positions reduce activity to near background levels (with the exception of the A^{+1} mutation, which seems to be well-tolerated).

Later generation ABEs show relaxed sequence requirements for editing

We recently reported that the first-round mutation D108N is crucial for enabling efficient A•T to G•C editing at genomic targets by higher-generation ABEs[11]. When incorporated into the ABE0.1 construct (which generates the ABE1.1 variant), this single mutation facilitated A•T to G•C editing levels above background (the highest at a CACA sequence motif), as quantified by NGS[2]. Furthermore, reversion of this mutation back to wild-type caused the ABE7.10 construct to lose nearly all activity[11]. To better understand the impact of this mutation on the sequence specificity of TadA, we repeated the previous experiments using ABE1.1m. At the "wild-type" TACG target in the A111V reporter, editing by ABE1.1m activated EGFP fluorescence in 34.6 ± 3.8% of transfected cells, which represented a 9.3 ± 1.8-fold increase compared to ABE0.1m (**Figure 4A** and **4D**). Further, the MFI of EGFP-positive cells treated with ABE1.1m were 8.6 ± 3.2-fold higher than those treated by ABE0.1m (SI Figure 6A and 6C). Mutation of the T^{-1} position had a much less drastic impact on editing activity by ABE1.1m compared to ABE0.1m, as measured by percent of transfected cells with EGFP fluorescence. Specifically, mutation of T^{-1} to A caused a 1.9 ± 0.5-fold decrease, mutation of T^{-1} to C caused a 1.7 ± 0.4-fold decrease, and mutation of T^{-1} to G caused a 1.4 ± 0.4-fold decrease (**Figure 4D**). Analysis of the MFI of EGFP-positive cells of these samples revealed the same trend (SI Figure 6C). Interestingly, mutations at the G^{+2} position did not cause statistically significant changes in editing activity by ABE1.1m (**Figure 4F**).

At the "wild-type" TACG target in the H182Y reporter, editing by ABE1.1m activated EGFP fluorescence in 32.4 ± 7.3% of transfected cells (**Figure 4E**). Notably, this represents a 26.6 ± 6.7-fold increase when comparing ABE1.1m to ABE0.1m (**Figure 4B** and **4E**). Further, the MFI of EGFP-positive cells treated with ABE1.1m was 5.0 ± 1.4-fold higher than those treated by ABE0.1m with this reporter (SI Figure 6B and 6D). Interestingly, mutation of the C^{+1} position had a drastic impact on editing activity by ABE1.1m, as assessed by both percent of transfected cells with EGFP fluorescence and the MFI of EGFP-positive cells. Mutation of the C^{+1} to all three other bases resulted in EGFP fluorescence levels that were barely above those of the non-targeting controls (**Figure 4E** and SI Figure 6D). Specifically, mutation of C^{+1} to A caused a 8.4 ± 3.3-fold decrease, mutation of C^{+1} to T caused a 6.6 ± 3.1-fold decrease, and mutation of C^{+1} to G caused a 13.4 ± 6.9-fold decrease, as measured by percent of transfected cells with EGFP fluorescence (**Figure 4E**).

These results collectively demonstrate that ABE1.1m prefers the TAC motif, with more flexibility at the T^{-1} position and less flexibility at the C^{+1} position compared to ABE0.1. The importance of the C^{+1} base is consistent with data from the original development of this editor, as ABE1.1 displayed the highest levels of A•T to G•C editing at site 1, which targeted an A in position 5, with a CACA motif (**Figure 2A**)[2]. We have shown computationally that the increased editing activity facilitated by the D108N mutation is due to the elimination of unfavorable electrostatic interactions between the negatively charged Asp108 residue and the phosphate backbone of the DNA at T^{-1}, which may suggest that this drastic protein-DNA interaction change is more important than any base-specific contacts between TadA1.1 and the DNA at position T^{-1}[11].

We next evaluated the sequence preferences of ABE7.10 and ABE8e using our fluorescent reporters, as these editors were explicitly evolved for broad sequence tolerance[2,9]. At the "wild-type" TACG target in the A111V reporter, editing by ABE7.10 activated EGFP fluorescence in 46.0 ± 2.5% of transfected cells. Mutation of the T^{-1} position to A had the largest impact on editing by ABE7.10, with a 1.79 ± 0.35-fold decrease, while the C and G mutations modestly reduced the editing efficiency by only 1.26 ± 0.10-fold and 1.13 ± 0.22-fold, respectively

(**Figure 4D**). Similar to ABE1.1, mutations to the G^{+2} position did not significantly affect editing efficiencies (**Figure 4F**). At the “wild-type” TACG target in the H182Y reporter, editing by ABE7.10 activated GFP fluorescence in 55.0 ± 6.2% of transfected cells. Interestingly, mutations to the C^{+1} position had no impact on editing by ABE7.10, as measured by percent of transfected cells with EGFP fluorescence (**Figure 4E**). These same trends were also observed when analyzing the MFI of EGFP-positive cells (SI Figure 6C-D).

Editing by ABE8e was higher than that of ABE7.10 at the “wild-type” TACG target in the A111V reporter, with 65.5 ± 5.2% of transfected cells displaying EGFP fluorescence (**Figure 4D**). Mutation of the T^{-1} position did not cause any statistically significant changes in editing, except for the T^{-1} to A mutation, which caused a modest 1.35 ± 0.23-fold decrease in percent of transfected cells with EGFP fluorescence. Mutations at the C^{+1} and G^{+2} positions also had no effect on ABE8e activity. Again, these same trends were also observed when analyzing the MFI of EGFP-positive cells (SI Figure 6C-D). Collectively, these data demonstrate a gradual loss of sequence-specificity by higher-generation ABEs, which is consistent with the selection strategies of these later rounds of directed evolution (SI Figure 7)[2].

To better understand how mutations to TadA may affect the stability of the base editor, we created mCherry-P2A-ABE-EGFP plasmids for ABE0.1, ABE1.1, ABE7.10, and ABE8e. With these constructs, mCherry fluorescence reflects transfection efficiency, while EGFP fluorescence reflects intracellular ABE concentration (as the EGFP protein is covalently attached to the ABE protein via a 10-amino acid linker). We transfected these plasmids into HEK293T cells and quantified EGFP intensities normalized to mCherry intensities to compensate for changes in transfection efficiencies. To our surprise, we observed the lowest normalized EGFP intensity with the ABE8e construct, which showed a 2.3 ± 0.3-fold lower intensity than the ABE0.1 construct (SI Figure 8A). In light of this observation, we analyzed the EGFP MFI’s from our earlier gDNA editing experiment (**Figure 2**), where cells were transfected with ABE-P2A-EGFP constructs and sorted for cells with the top 20-30% EGFP fluorescence. We again observed that the EGFP MFI of the ABE8e-P2A-EGFP sample was lower than that of the identical ABE0.1 and ABE1.1 constructs (SI Figure 8B-C). These data suggest that despite the significant improvement in editing efficiency with ABE8e, the accumulated mutations may have reduced its stability or expression.

Optimization of bacterial directed evolution selection target results in measurable editing activity

The original development of the ABEs utilized a bacterial directed evolution strategy that required the installation of an A•T to G•C point mutation in an antibiotic resistance gene by active library members (ABE mutants) to confer survival advantage[2]. When building the selection system for the first round of evolution, a wtTadA-dCas9 (ABE0.1) construct was tested for baseline activity by directing it to a TAGT motif in the inactive chloramphenicol resistance gene CmR H193Y (**Figure 5A**). Additionally, the target A was placed at position 9 within the protospacer. Consequently, ABE0.1 was found to be completely inactive in this bacterial selection system[2]. This, combined with the mammalian cell editing data of ABE0.1 mentioned earlier, resulted in the conclusion that wtTadA was incapable of DNA editing. We re-examined the activity of early generation ABEs in bacterial selection systems in light of our new understanding of the sequence context and editing window preferences of these ABE variants. We first tested the editing activity of ABE0.1 and ABE1.1 on the original round 1 selection system (TAGT sequence motif with the target A in position 9). S1030 *E. coli* cells harboring a plasmid encoding the inactive CmR H193Y gene were transformed with a plasmid encoding ABE0.1, ABE1.1, ABE7.10, or Cas9n under the control of a theophylline-responsive riboswitch (**Figure 5A**)[29]. After recovery, ABE expression was induced for 18 hours, and cultures were plated on both 0 mg/mL and 25 mg/mL chloramphenicol plates. Survival rate was calculated by taking the fraction of surviving colonies at 25 ng/uL chloramphenicol compared to those plated at 0 ng/uL chloramphenicol. We found that cells transformed with ABE0.1 or Cas9n were unable to rescue chloramphenicol resistance at detectable levels, while cells transformed with ABE1.1 had a $\log_{10}$ survival rate of -6.5 ± 0.8, and those transformed with ABE7.10 had a $\log_{10}$ survival rate of -4.3 ± 0.8 (**Figure 5B**), consistent with what was observed during the original development of the ABEs[2]. We then used the PAM-relaxed Cas9n-NG variant to shift the protospacer by three bases and placed the target A at position 6[30]. We repeated the experiment and observed detectable activity by ABE0.1 ($\log_{10}$ survival rate of -5.5 ± 0.9), ABE1.1 ($\log_{10}$ survival rate of -4.1 ± 0.8), and ABE7.10 ($\log_{10}$ survival rate of -1.9 ± 0.8), but not the Cas9n negative control (**Figure 5B**). While it was previously thought that the D108N mutation identified from the first round of directed evolution imparted a completely new activity (DNA editing) on TadA, these data suggest that the directed evolution instead enhanced the pre-existing but nearly undetectable DNA editing activity of TadA.

Intrigued by these observations, we redesigned the selection system to have TACG or TAGG target motifs (target A in position 6, **Figure 5A**), and repeated the experiment with Cas9n, ABE0.1, ABE1.1, ABE7.10, and ABE8e (**Figure 5C**). We noted similar patterns as those observed in our mammalian cell-based fluorescence assay; ABE0.1 and ABE1.1 both demonstrated strict sequence preferences at the +1 position (ABE0.1 had a 56 ± 22-fold higher survival rate on the TACG motif, and ABE1.1 had a 16 ± 5-fold higher survival rate on the TACG motif), while ABE7.10 and ABE8e demonstrated no statistically significant differences between these two sequence motifs (**Figure 5C**). These findings collectively demonstrate that judicious design of the selection motif and target position can vastly improve the chances of success when designing a directed evolution strategy for the development of novel base editors.

Computational simulations reveal molecular basis of editing activity and sequence preference

Next, we sought to investigate the molecular basis for the strict sequence-context requirement observed for ABE0.1m and ABE1.1m, particularly when compared to their evolved successors ABE7.10, and ABE8e. We performed all-atom unbiased MD simulations of these full-length ABE variants in their substrate-bound forms and varied the sequence-context surrounding the target adenine base (SI Table 1). In our previous simulation studies of ABEs, we used minimalistic models (specifically, we simulated only the TadA variant bound to either its substrate ssDNA or tRNA)[11,12]. However, in this current study, we were able to expand upon these minimal models by using the recently published cryo-EM structure of the ABE8e R-loop complex (PDB ID: 6VPC)[21]. We will note that in this structure, ABE8e consists of two TadA subunits, only one of which is complexed with the DNA substrate 5′-GTTCCACTTT-3′. We therefore started with this sequence context in our simulations after removing the *in trans* TadA subunit (**Figure 1A**). Moreover, we generated experimental data using our reporter assay with this sequence context to complement these simulations. We installed two silent mutations into our A111V reporter (T^{-1} to C, and G^{+2} to T), and analyzed activity by ABE0.1m, ABE1.1m, ABE7.10, and ABE8e using this CACT reporter. We observed that this CACT-based reporter is highly edited by ABE8e (68.9 ± 4.5%), ABE7.10 (35.2 ± 4.3%), and ABE1.1m (22.0 ± 4.1%), but not by ABE0.1m (0.4 ± 0.2%) (**Figure 6A**) as measured by percent of transfected cells with EGFP fluorescence, consistent with the general trends that we observed at other targets with T^{-1} mutations (**Figure 4A** and **4D**).

We initially focused on ABE8e and ABE0.1m, as these represent the two extremes in the evolutionary lineage of the ABEs. Starting from the structure of ABE8e, we launched microsecond-timescale simulations, focusing on the dynamics of TadA8e and the exposed ssDNA strand, particularly the CAC target motif (**Figure 1D**)[21]. We then reverted the mutations in TadA8e back to wild-type and repeated the simulation (see Methods). To identify the molecular interactions that differ between ABE8e and ABE0.1, we defined a 4 Å "shell" surrounding the CAC nucleotides and analyzed the molecular interactions between the target DNA and the amino acids that lie within this first interaction shell during the production phase (last 1800 ns) of the simulations (**Figure 6B-C,** SI Figures 9-10). In these "interaction maps" (**Figure 6B-C**), the amino acids that lie within the first interaction shell of the CAC nucleotides are listed. Those that make direct interactions, such as hydrogen-bonds (H-bond) or hydrophobic contacts, with the nucleotides are shown with color-coded arrows pointing between their residue label and the appropriate location on the DNA, with the thickness of the arrows being proportional to the stability of the interaction (defined as the frequency of appearance of the interaction during the simulation) (**Figure 6B-C** and SI Figures 9-10).

A comparison between the interactions maps of ABE0.1m (**Figure 6B**) and ABE8e (**Figure 6C**) bound to the CAC substrate reveals that the mutations in ABE8e lead to the formation of several new H-bonds between TadA residues and the phosphate backbone of the DNA, particularly residues within the β4-β5 active site loop (residues 104 to 129). Specifically, two mutations in the β4-β5 loop of the enzyme, D108N (discovered in ABE1.1m) and T111R (discovered in ABE8e), interact strongly with the phosphate backbone (i.e., nonspecifically) of the target A in the ABE8e complex, but not in the ABE0.1m complex. Further, the A109S mutation (also in the β4-β5 loop, discovered in ABE8e) forms an H-bond directly with the N3 atom of the C^{-1} nucleobase. Additionally, there is a nonspecific H-bond interaction between Lys110 (also in the β4-β5 loop) and the phosphate backbone of the -1 nucleobase in both the ABE0.1 and ABE8e complexes (**Figure 6B-C**), which has been strengthened by the collective mutations in the ABE8e complex. We also observed a nucleobase-specific H-bond between the exocyclic amino group of the C^{+1} nucleobase and Glu27 in the ABE0.1 complex, but not ABE8e (**Figure 6B-C**). This H-bond formed spontaneously during the course of the ABE0.1 simulation, and once formed remained stable throughout the trajectory (**Figure 6D**). A closer inspection of the orientation of the central CAC nucleotides in the ABE0.1 simulation revealed that the C^{+1} nucleotide undergoes a conformational change, moving away from the α5 helix and towards the Glu27 residue in the α1-β1 loop to form this stable hydrogen bond (**Figure 6E**). This new conformation adopted by the nucleotides in ABE0.1 closely resembles the conformation of wtTadA bound to its native tRNA substrate, in which the nucleotides are splayed across the active site groove of the wtTadA enzyme (**Figure 1B**)[19]. In contrast, in the ABE8e complex, this C^{+1} base remains stable in its initial conformation and forms no notable interactions with any of the TadA residues (**Figures 6F** and **1D**). This conformation change is likely driven by a "kink" in the α5 helix of TadA8e, caused mainly by the seventh round R152P mutation, as there are no mutations in the α1-β1 loop, and the +1 base resides between these two secondary structural elements in our ABE0.1 simulations (**Figure 6B, E**) as well as in the wtTadA-RNA structure (**Figure 1B**)[2,19].

We next expanded the set of ABE variants and the sequence context surrounding the target A in our simulations to better understand the molecular details driving the sequence specificity we observed experimentally (**Figure 4**). Informed by the dynamics of the CAC systems which underwent conformational transitions within 200 ns (**Figure 6D**), we modelled ABE0.1m, ABE1.1m, ABE7.10, and ABE8e bound to TAC, AAC, and TAG target motifs, and conducted 400 ns-long simulations for each complex (**Figure 6I-J**). An analysis of the interactions between the target nucleotides and residues in the β4-β5 loop of TadA revealed a similar trend as observed with the CAC systems; specifically, the interaction network between the residues in the β4-β5 loop of TadA and the substrate adenine and the -1 nucleotide progressively strengthens, first with the introduction of the D108N mutation in ABE1.1m, and is subsequently reinforced by the T111R mutation in ABE8e, regardless of the identity of the flanking nucleotides (**Figure 6I** and SI Figures 11-14). It should be noted that given the non-specific nature of the interactions between the β4-β5 loop residues and the -1 nucleotide, these simulations do not explain the strict T^{-1} requirement of ABE0.1 observed in the reporter assays (**Figure 4A**), as Lys110 forms a H-bond with the phosphate backbone of the target motifs in all ABE0.1 simulations. Losey et al. made a similar observation in their TadA-RNA crystal structure (PDB ID: 2B3J) and noted that the U^{-1} base is not recognized by any significant nucleobase-specific interactions with the enzyme. Furthermore, analogous to the CAC simulations (**Figure 6B-F**), in the TAC and AAC simulations, the C^{+1} nucleotide adopted a conformation closer to the α1-β1 loop residues, where it forms a sequence-specific H-bond through its exocyclic amino group with either the peptide backbone of Arg26 or directly with the side chain of Glu27, only in the ABE0.1m and ABE1.1m systems (**Figure 6G-H, J** and SI Figures 11-14). However, consistent with the experimental observations (**Figure 4**), this critical H-bond with the α1-β1 loop residues is not formed in any of the TAG simulations, as the G^{+1} base did not undergo the conformational transition seen with the C^{+1} systems. Instead, the G^{+1} severely clashes with the α5 helix in the ABE0.1m and ABE1.1m variants (SI Figure 15).

On the basis of these findings, we hypothesize that there are two molecular mechanisms at play in higher generation ABEs (ABE7.10 and ABE8e) that are responsible for the loss of sequence specificity observed in ABE0.1 and ABE1.1. First, the mutations accumulated in the β4-β5 loop (particularly D108N and T111R) non-specifically increase TadA's affinity to the target DNA (**Figure 6**). Overall, these mutations enhance editing activity while simultaneously relaxing the T^{-1} requirement as these residues bind the target through its phosphate backbone (**Figure 4A** and **D**). Second, mutations in the higher generation ABE variants reduce the nucleobase-specific interactions between the C^{+1} base and the residues in the α1-β1 loop, thereby relaxing the C^{+1} requirement. Notably, the α1-β1 loop itself is not mutated in any ABE, and in fact these residues are highly conserved among its homologs as well (SI Figure 16). We therefore attribute this second molecular mechanism to secondary effects caused by the mutations in the α5 helix region of TadA, which introduce a kink in the helix (**Figures 1C-D** and **6**). The kink in the α5 helix caused by these mutations, particularly R152P, abolishes the contact between the C^{+1} base and α1-β1 loop, hence relaxing the sequence preference of the highly evolved ABE variants at this position (that is ABE7.10 and ABE8e, but not ABE1.1m) (**Figure 4E**).

Mutations to the α5-helix affect C^{+1} requirements

The α5 helix is comprised of twenty-one amino acids, spanning from residues 137-167, nearly half of which have been mutated in ABE8e. Six mutations were evolved in the ABE7.10 variant (S146C, D147Y, R152P, E155V, and K157N) and an additional four during the ABE8e evolution (F149Y, Q154R, T166I, and D167N, **Figure 1E**). Given the apparent importance of the α5 helix on defining the sequence specificity of ABEs at the +1 position (the C^{+1} position in TACG), we performed several "α5 helix swapping" experiments on ABE0.1m, ABE1.1m, and ABE8e. Specifically, we introduced the α5-helix mutations from ABE8e into ABE0.1m and ABE1.1m (which we call ABE0.1m(8e α5) and ABE1.1m(8e α5), respectively), and reverted the entire α5-helix in ABE8e back to wild-type (which we call ABE8e(WT α5), **Figure 7A**). We first compared editing efficiencies of these new variants at the "wild-type" TACG motif using both the A111V and H182Y reporters. With both reporters, ABE0.1m(8e α5) exhibited a complete abolishment of activity as measured by percent of transfected cells with EGFP fluorescence, with levels within error of non-targeting gRNA controls (compared to 2.21 ± 0.80% of cells transfected with ABE0.1m displaying EGFP fluorescence with the A111V reporter and 1.01 ± 0.23% of cells transfected with ABE0.1m displaying EGFP fluorescence with the H182Y reporter, **Figure 7B**). We then measured editing efficiencies with the C^{+1} to G reporter to assess the effects of these helix swaps on the sequence specificity at the +1 position. We found that editing efficiency by ABE0.1m(8e α5) was within error of non-targeting gRNA controls with the C^{+1} to G mutation (**Figure 7B**). These data suggest that the sequence-specific interactions between the α1-β1 loop and the exocyclic amino group of the C^{+1} base (driven by the orientation of the α5 helix in the lower-generation ABEs) may be crucial for editing by ABE0.1m (**Figure 6**)

ABE1.1m(8e α5) also exhibited decreases in editing compared to ABE1.1m at the TACG motif in both reporters, as measured by percent of transfected cells with EGFP fluorescence. However, this decrease was much less drastic, and editing was above the levels of the non-targeting gRNA controls (we observed a 1.82 ± 0.23-fold decrease with the A111V reporter and a 1.35 ± 0.18-fold decrease with the H182Y reporter, **Figure 7C**). Interestingly, ABE1.1m(8e α5) displayed much higher editing than ABE1.1m when the C^{+1} base was mutated to G (**Figure 7C**). Specifically, ABE1.1m activated EGFP fluorescence in only 2.54 ± 0.52% of cells transfected with the C^{+1} to G reporter, while ABE1.1m(8e α5) activated EGFP fluorescence in 22.64 ± 2.64% of transfected cells (an 8.90 ± 2.11-fold increase). These general trends were also observed when cells were analyzed for MFI of EGFP-positive cells (**Figure 7D**). These data suggest that the additional non-specific interactions between TadA and the -1 nucleotide due to the D108N mutation in ABE1.1 may be sufficient for this evolved TadA variant to edit DNA sequences even when the nucleobase-specific interaction with the C^{+1} base is eliminated. This evolutionary path was likely taken due to all seven rounds of directed evolution being undertaken with non-C^{+1} selection targets (SI Figure 7).

Interestingly, we found that removal of the 8e α5 helix mutations in ABE8e did not significantly impact the relative editing activity of the C^{+1} to G reporter (compared to the TACG H182Y reporter), as assessed by percent of transfected cells with EGFP fluorescence and MFI of EGFP-positive cells (**Figure 7C-D**). However, overall editing activity by ABE8e(WT α5) was slightly reduced compared to ABE8e at all three reporters tested (**Figure 7C-D**). This suggests that either the collective additional interactions between ABE8e and the target DNA (as compared to ABE0.1) are sufficient to overcome any sequence-specific interactions between the α1-β1 loop and the C^{+1} base that may be re-introduced by removal of the kink in the α5 helix, or the additional mutations present in ABE8e cause additional rearrangements throughout the protein that prevent the α1-β1 loop from interacting with the C^{+1} base.

Reversion mutations in the β4-β5 loop reduce editing efficiencies by evolved ABEs

To better understand the benefits of the D108N and T111R mutations in TadA, we reverted each of these mutations back to the wild-type residue in ABE7.10 and ABE8e, producing ABE7.10-N108D, ABE8e-N108D, and ABE8e-R111T. Editing efficiencies of these three ABE variants were evaluated with the two "wild-type" reporters, as well as one mutant reporter per each of the three positions surrounding the target A in the T^{-1}A^{0}C^{+1}G^{+2} motif (the T^{-1} to A reporter, the C^{+1} to G reporter, and the G^{+2} to C reporter; these were chosen as we previously observed the lowest editing efficiencies with these reporters). At the A111V "wild-type" reporter, the ABE7.10-N108D variant showed a 249 ± 17-fold decrease in editing compared to ABE7.10 (SI Figure 17). While this was near the levels of the non-targeting controls, editing was still statistically significantly above those of the controls, allowing us to quantify trends in editing efficiencies among the five tested reporters. With the T^{-1} to A reporter we observed a complete loss of editing, similar to the trends in editing efficiencies observed with the other ABEs. Editing efficiencies at the C^{+1} to G and G^{+2} to C reporters were within error of their respective TACG reporters, as is seen with other ABEs that contain a mutated α5 helix. Overall, the near loss of all editing by the ABE7.10-N108D variant further shows the importance of the D108N mutation which was first observed in our previous work evaluating the editing efficiency of this variant on gDNA targets[11].

The ABE8e-N108D variant also showed a drastic decrease in editing efficiency compared to its parental construct at all reporters, as measured by percent of transfected cells with EGFP fluorescence (we observed a 9.3 ± 1.8-fold decrease with the A111V "wild-type" reporter, and a 5.2 ± 0.4-fold decrease with the H182Y "wild-type" reporter, SI Figure 17). The sequence specificity of this variant followed the same trends as observed in the ABE7.10-N108D variant; editing efficiencies of ABE8e-N108D were within error of the "wild-type" reporters with the C^{+1} to G and G^{+2} to C reporters (SI Figure 17) but were greatly decreased with the T^{-1} to A reporter (we observed a 5.6 ± 1.9-fold decrease in editing efficiency compared to the "wild-type" reporter, SI Figure 17). The ABE8e-R111T variant displayed a slight decrease in editing efficiency at the A111V "wild-type" reporter compared to the ABE8e construct (a 1.27 ± 0.04-fold decrease, SI Figure 17). Again, the editing efficiency of ABE8e-R111T with the C^{+1} to G reporter was within error of the "wild-type" TACG reporter, and the T^{-1} to A mutation was less tolerated than the parental ABE8e construct. Specifically, we observed a 2.1 ± 0.2-fold decrease with the T^{-1} to A mutation compared to the A111V "wild-type" reporter for ABE8e-R111T, while that for ABE8e was only a 1.27 ± 0.04-fold decrease (SI Figure 17). Editing efficiency by ABE8e-R111T at the H182Y "wild-type" was within error of that of ABE8e, and editing efficiency increased slightly with the G^{+2} to C reporter (SI Figure 17). At all tested reporters, editing by the ABE8e-R111T variant was more similar to the editing profile of ABE7.10 than the parental ABE8e construct, which has been shown previously[21]. These data, along with the α5 helix swapping experiments, are supportive of the conclusions from our simulations; mutations in the β4-β5 loop (D108N and T111R) non-specifically increase TadA's affinity to the target DNA (and in the process relax the T^{-1} requirement), while mutations in the α5 helix relax the sequence preference at the C^{+1} base.

Conclusion

The lack of naturally-occurring DNA-modifying enzymes has proven challenging in the quest to develop base editors capable of introducing additional types of mutations. The evolution of TadA into an efficient DNA base editor showed that RNA-modifying enzymes can be used to expand the base editing tool kit, but no other RNA-modifying enzyme has been successfully evolved into a DNA base editor[2]. Here we show that wtTadA is not a strict RNA-modifying enzyme, but can also modify DNA with low efficiency at specific sequence contexts. This finding may ease the search for new enzymes for base editing by identifying enzymes that already show low levels of DNA editing.

To facilitate the detection of enzymes with low levels of DNA editing, we developed a fluorescence-based screening assay that takes advantage of the natural sequence recognition properties of TadA. Specifically, the incorporation of the preferred sequence motif of TadA (T$\underline{\text{A}}$CG) allowed for the reliable detection of DNA editing by wtTadA. This assay can detect down to 0.01% of corrected plasmid and can in theory be repurposed with any mutation that knocks out EGFP fluorescence. While GFP reporters have been used before to compare editing efficiencies between different deaminases and also to enrich for cells with higher levels of editing, here we have developed an additional use for these reporters for the detection of low levels of editing[7,31,32].

This assay also allowed us to probe the sequence context recognition requirements of multiple ABE variants by making silent mutations at the target site. Through this, we were able to observe strict requirements for the T^{-1} base, and for a C^{+1} or A^{+1} base, by ABE0.1. These sequence preferences change with the introduction of the D108N mutation in ABE1.1; the T^{-1} requirement is relaxed, and there is a strict C^{+1} requirement for this ABE variant. The higher generation ABE7.10 and ABE8e constructs display a slight aversion to A at position -1, with no apparent sequence preferences at the +1 position. This is particularly interesting given the sequences of the targets used for directed evolution, as seven out of nine of the selection targets contained a T^{-1}, (SI Figure 7)[2].

We used molecular dynamics simulations to explain the sequence specificity at the +1 position of ABE0.1 and ABE1.1 due to the mutations that occur in the α5 helix. Crystal structures of TadA have shown that this helix is flexible and hampers crystal formation but takes on a helical structure when bound to the target tRNA and can be resolved[19,20]. Our ABE0.1 simulations show that the DNA binding process causes the α5 helix to flip out the +1 base, where it is held in place by sequence-specific contacts through a H-bond between its exocyclic amine and the α1-β1 loop. In the process, this structural rearrangement positions the target A to better fit into the active site. If a G^{+1} is present though, the base cannot be properly positioned, and TadA activity is hampered. The relaxation of this requirement can be attributed in part to the mutations in the α5 helix that cause a sharp "kink" at the R152P, which is not present in ABE0.1 or ABE1.1. This kink was observed in the cryo-EM structure of ABE8e, and likely allows the target A to better access the active site independently of any sequence-specific interactions with the +1 base[21].

The α5 helix swapping experiments further confirmed the importance of these molecular interactions between the protein and the +1 position. Specifically, adding the α5 helix mutations to ABE0.1 (which would prevent the +1 base from interacting with the α1-β1 loop) abolished its activity at TACG and TAGG sites, while adding these mutations to ABE1.1 decreased its previously strict requirement for a C at position +1. Collectively, these data demonstrate that the base-specific interactions between the α1-β1 loop and the DNA substrate, caused by the overall orientation of the α5 helix, are crucial for DNA editing by wtTadA, but can be removed to expand sequence recognition once the enzyme has evolved a higher overall editing efficiency. Finally, reversion of the D108N and T111R mutations from the β4-β5 loop caused significant decreases in overall editing efficiency. This is consistent with the results from our simulations, which revealed that the residues in the β4-β5 loop interact with the phosphate backbone of the target DNA at the -1 position and thus non-specifically increase TadA's affinity to the target DNA.

Methods

Cloning

All primers in this study were ordered through Integrated DNA technologies (IDT). All PCR reactions were performed with Phusion DNA Green High-Fidelity Polymerase (F534L, Thermo Fisher) or Phusion U (F556L, Thermo Fisher) where appropriate. The mCherry-P2A-EGFP reporter plasmid was cloned via USER cloning following New England Biolabs (NEB) protocols (reference), by replacing the *ABE* gene in the ABE-P2A-EGFP plasmid (Addgene plasmid #112101) with the *mCherry* gene from the pBAD-mCherry plasmid (Addgene plasmid #54630). All variations (i.e. point mutations) on this plasmid were cloned by site directed mutagenesis[33]. To facilitate the cloning of mammalian cell base editor plasmids a "Golden Gate Destination Plasmid" was cloned using the NG-ABEmax plasmid (Addgene plasmid #124163). Briefly, USER cloning was used to delete the TadA dimer and insert a sequence containing two BsaI (a type IIS restriction enzyme) recognition sites. To clone new base editor variants, point mutations or helix swaps were performed in "reservoir" plasmids (containing only the *TadA* gene) using site directed mutagenesis or USER cloning, respectively. The TadA reservoir plasmids also contained BsaI recognition cut sites with overhangs matching the overhangs in the destination plasmid. New ABE plasmids were therefore cloned by following the BsaI-HFv2 Golden Gate Assembly protocol from NEB. Destination base editor plasmids for mammalian reporter and bacterial selection experiments were similarly cloned using USER cloning. A similar destination plasmid was used to clone mammalian cell gRNA plasmids as previously described[33]

Cell culture

HEK293T cells (ATCC CRL-3216) were cultured in high glucose Dulbecco's Modified Eagle's Medium (DMEM) supplemented with GlutaMAX (ThermoFisher Scientific #10566-016) and 10% (v/v) fetal bovine serum (ThermoFisher Scientific #10437-028) at 37°C with 5% CO_2. Cells were passaged every 2 days using TrypLE (ThermoFisher Scientific # 12605028).

Transfections

12-16 hours before transfection, 50,000 HEK293T cells in 250 µl DMEM media were plated per well into a 48-well cell culture plate (VWR # 10062-898). For fluorescent reporter assays, DNA mixtures were prepared with: 750 ng of base editor plasmid, 500 ng of mCherry-P2A-EGFP reporter plasmid, and 250 ng of gRNA plasmid. For reporter sensitivity experiments, DNA mixtures contained a combination of the active mCherry-P2A-EGFP plasmid and inactive mCherry-P2A-dGFP plasmid at the indicated percentages for a total of 500 ng plasmid. For gDNA editing experiments, the DNA mixtures were prepared with: 750ng of base editor plasmid and 250ng of gRNA plasmid. For all transfections, DNA mixtures were brought to a total volume of 12.5 µl using Opti-MEM (Gibco #31985-070) and then combined with a 12.5 µl solution comprised of 1.5 µl of Lipofectamine 2000 (Invitrogen #11668-019) and 11 µl Opti-MEM (Gibco #31985-070). The resulting 25 µl DNA/Lipofectamine mixture was then added to the cells. 24 hours after transfection, 250 µl of DMEM media was added to each transfected well. Cells were then incubated for 48 additional hours before harvesting for NGS or flow cytometry/FACS.

Flow cytometry and Fluorescence Activated Cell Sorting (FACS)

The media was removed from each well, and each well was washed with 150 µl of phosphate buffered saline (PBS, Gibco #10010-023). To detach cells, 40 µl of Accumax (Innovative Cell Technologies #AM-105) was added to each well. Cells were counted and diluted to a concentration of 1 x 10^6 cells/ml using PBS, then pipetted into a Falcon 5 ml test through the cell strainer cap (Corning #352235) and kept on ice. Flow cytometry data was collected using a Bio-Rad S3e cell sorter equipped with 488nm, 561nm and 640nm lasers, and analyzed using FlowJo v10.8.1 Software (BD Life Sciences)[34]. Scatter gates were applied to remove non-viable cells and doublets. For reporter experiments, gates were applied based on cells transfected with only mCherry or only EGFP plasmids. mCherry fluorescence was detected using FL3 (602-627 nm) and a PMT voltage of 360. EGFP fluorescence was detected using FL1 (510-540 nm), with a PMT voltage of 420 when detecting the A111V reporter and a PMT voltage of 330 when detecting the H182Y reporter. ~100,000 cells (after scatter gating) were collected for each sample. For FACS, the same protocols for gating and fluorescence detection were used, with an additional sort gate applied based on the EGFP fluorescence of non-transfected cells as a negative control. Cells were collected into 300 µl of DMEM media and kept on ice. Sorted cells were centrifuged at 300 rcf for 10 minutes. The supernatant was decanted, and 300 µl of PBS was added to wash the cells. After centrifuging at 300 rcf for 10 minutes, the supernatant was removed, and the cells were further processed for HTS (next section).

High-throughput amplicon sequencing of genomic DNA

For unsorted cells, the media was removed from each well and cells were washed with 150 µl of PBS. 100 µl of lysis buffer (10 mM Tris (pH 7.5), 0.1% SDS, and 25 µg/ml Proteinase K) was added to each well, then pipetted up and down several times to break up cell clumps. Cells were lysed by incubating at 37°C for 1 hour, followed by 80°C for 20 minutes.

For sorted cells, following sorting and washing, a volume of lysis buffer (10 mM Tris (pH 7.5), 0.1% SDS, and 25 µg/ml Proteinase K) was added to bring the cell concentration to ~2000 cells/µl. Cells were then lysed by incubating at 37°C for 1 hour, followed by 80°C for 20 minutes.

For both unsorted and sorted cells, genomic loci of interest were PCR amplified from the lysed cells using Phusion High-Fidelity DNA Polymerase, and primers that bind to the loci of interest (see Supporting Sequences) PCRs followed the manufacturer's protocol, using 1 µl of genomic DNA for template and 24 or fewer rounds of amplification. Unique combinations of forward and reverse Illumina adapter sequences were then appended with an additional round of PCR using Phusion High-Fidelity DNA Polymerase. Round two PCRs followed the manufacturer's protocol, using 1 µl of the previous PCR product as a template and 10 or fewer rounds or amplification. PCR products were gel purified from 2% agarose gel with QIAquick Gel Extraction Kit (Qiagen #28704) and quantified using NEBNext Ultra II DNA Library Prep Kit (NEB #E7805L) on a Bio-Rad CFX96 system. Samples were then sequenced on an Illumina MiniSeq according to the manufacturer's protocol.

Analysis of Illumina HTS was performed with CRISPResso2[35]. Specifically, fastq files were analyzed *via* Docker scripts that analyzed reads against the entire amplicons, with outputs for the gRNA and base editor (--guide_seq and –base_editor_output). A•T to G•C edits were calculated using the nucleotide frequency at the target site, by dividing the number of A reads by the total reads. Indel counts were calculated by subtracting reads with only substitutions from the total modified reads, then indel percentages were calculated by dividing by the total reads.

Data analysis and Statistics

Plots were made in R studio using the "ggplot2" package or with GraphPad Prism[36,37]. Statistics tests were performed in R Studio using the "rstatix" package or with GraphPad Prism[38]. One-tailed T-tests were used when comparing targeting samples with non-targeting gRNA negative controls. Two-tailed T-tests were used when comparing two targeting samples to each other.

Bacterial Survival Assay

10ng of antibiotic target plasmid was transformed into S1030 *E. coli*, allowed to recover for 1 hour at 37°C while shaking in super optimal broth with catabolite repression (SOC) media (NEB #B9020S), and plated on 2xYT agar plates supplemented with 50 ng/μL Kanamycin maintenance antibiotic[39]. Single colonies were inoculated into Kanamycin containing culture, from which chemically competent target plasmid-containing S1030 stocks were developed[40]. 10ng base editor plasmid was then chemically transformed into respective target plasmid-containing *E. coli* and allowed to recover for 1 hour 37°C in EZ Rich media (Teknova #M2105). Transformation efficiencies were monitored by plating 5 μL of 1 through 1000-fold dilutions of the transformation cultures on 2xYT agar plates supplemented with 50 ng/μL kanamycin and 50 ng/μL carbenicillin (BE plasmid maintenance antibiotic). Plates were then incubated for 16 hours at 37°C. The transformation cultures were also diluted 1:100 into two separate solutions of 5 mL of EZ Rich media supplemented with 50 ng/μL kanamycin, 50 ng/μL carbenicillin, and 0 or 1 mM theophylline (to induce ABE expression, which is controlled by a theophylline riboswitch). Cultures were incubated at 37°C while shaking for 16 hours. Saturated cultures were then diluted 1 to $1x10^7$-fold in PBS, and 5 μL of each dilution factor was plated on 2xYT agar plates supplemented with 50 ng/μL kanamycin, 50 ng/μL carbenicillin, and 0 or 25 ng/μL chloramphenicol. Plates were incubated for 18 hours at 37°C, and colonies were counted at a dilution factor where single colonies were visible. Survival rate was calculated by dividing the number of colonies that survived on the 25 ng/μL chloramphenicol plates by the number of colonies that survived on the 0ng/μL chloramphenicol plates.

Molecular Dynamics (MD) simulations

System Preparation

MD simulations for all ABE variants were performed starting from the full-length cryo-EM structure of ABE8e (PDB ID: 6VPC)[21]. The missing amino acid residues in TadA8e (residues 1-4 and 160-167), the XTEN linker (residues 168-200), and Cas9 (residues 910-915, 967-972, 1104-1120, and 1562-1565), as well as the missing bases in the exposed ssDNA (nucleotides 31-38) were modeled using Modeller 10.1[41]. The cryo-EM coordinates were kept fixed, and 100 independent models were generated for the ABE-R-loop structure. The top 10 models were selected based on the lowest DOPE score and Z-score value, and the final model was selected after thorough visual inspection of these ten models to ensure that no loops were entangled or knotted in a physiologically irrelevant conformation, and to ensure there were no clashes between the modeled portions and the rest of the resolved structure. Catalytic Mg^{+2} ion was added to the HNH domain, and the Ala840 residue of Cas9 was mutated back to His. Waters of crystallization were added in from PDB ID: 4UN3[42]. All titratable residues were protonated using the H++ server employing the default settings[43,44].

To prevent simulation artifacts, especially due to unwanted interactions between the flexible XTEN linker and the PAM-distal end of the DNA, an additional 10 base pairs of DNA was built using Chimera, based on the missing DNA sequence density in the original cryo-EM structure (non-target strand 5′-CGATCGGTGG-3′)[45]. Initially, this DNA decamer was constructed in isolation in Chimera, followed by manual adjustments to place it close to the existing PAM-distal end of the DNA sequence. The phosphate bonds were created manually between these DNA sequences to covalently link the missing decamer to the resolved DNA base pairs, and the atom names as well as numbering was fixed to reflect the connectivity between the new DNA bases and the pre-existing ones.

The complexes containing ABE0.1, ABE1.1, and ABE7.10 were modeled using a strategy similar to the one listed above, except the TadA8e was replaced with TadA0.1, TadA1.1, or TadA7.10, respectively. The structures of these TadA variants were predicted using Alphafold2[46]. This was done primarily because the X-ray structure of wtTadA (PDB ID: 1Z3A) lacks density for the terminal 10 amino acids of the α5-helix of the protein, which are critical for the hypotheses that we tested in this study[18]. The different DNA sequences were generated using the swapna command in Chimera. The list of all the combinations modelled can be found in SI Table 1.

All ABE systems were solvated in rectangular TIP3P water boxes with a buffer length of 13.5Å[47]. Na+ ions were added to the system to maintain electroneutrality. The protein was represented using the Amber ff14SB force field, the RNA was represented using the RNA.OL3 force field, and the DNA was represented using bsc1 parameters[48–53]. The Zinc-containing active site of TadA was represented with custom force field parameters obtained using the MCPB.py approach, at B3LYP/6-31G* level of theory previously shown to be effective at capturing its semi-bonded characteristics[12,54].

Simulation Protocol

All MD simulations were performed under periodic boundary conditions using the CUDA accelerated version of PMEMD implemented in the Amber20 suite of programs[53,55,56]. The structures were relaxed using a combination of steepest descent and conjugate gradient minimization. During the first minimization phase, all atoms except the waters were restrained with a 300 kcal/molÅ^2 force constant. During the second and final minimization phase, all the restraints were removed, and the system was allowed to freely minimize. The long-range electrostatics were cut-off at 12 Å.

This was followed by multi-step heating. During the first phase, the system was heated from 0-100 K, with the non-water atoms held with a 100 kcal/molÅ^2 force constant. This was followed by a 100-200K heating ramp where only the backbone atoms of the non-water atoms were restrained with a 100 kcal/molÅ^2 force constant. Finally, all restraints were removed, and the system was heated to the final temperature of 300K. The Langevin thermostat was employed in these NVT simulations with a collision frequency of 1 ps^{-1}. Finally, NPT equilibrations were performed for 2000 ns for all systems using a Brendsen barostat to maintain a 1 bar pressure. The hydrogen atom bond length was constrained by implementing the SHAKE algorithm. All MD simulations were propagated in time using the velocity Verlet with a time step of 2 fs. The initial 200 ns were discarded to compare the equilibrate dynamics of the various ABE systems.

The CPPTRAJ module implemented within Amber21 was used to analyze all the MD trajectories[57,58]. The visualization of the MD trajectories was rendered using ChimeraX, and data were plotted using Matplotlib[59–61].

Data and material availability

High-throughput sequencing data is deposited in the NCBI Sequencing Read Archive database under accession code PRJNA943422.

Acknowledgements

This research was supported by the University of California, San Diego and the NIH through grant no. 1R35GM138317 (to A.C.K). All computer modelling and simulations used resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF through grant no. ACI-1548562. B.L.R. was supported by the Chemistry-Biology Interface Training Program, NIH Grant T32 GM112584. M.E. was supported by the Cell and Molecular Genetics Training Program, NIH Grant T32 GM007240-41.

Author contributions

B.L.R. contributed to conceptualization of the research project, experimental design of all fluorescent reporter and gDNA editing experiments, experimental data curation/acquisition, data analysis, and writing of the manuscript. K.L.R. contributed to conceptualization of the computational aspects of the research project, design of all simulations, data curation/acquisition of all simulation work, data analysis, and writing of the manuscript. M.E. contributed to conceptualization of the research project, experimental design of bacterial editing experiments, experimental data curation/acquisition, data analysis, and writing of the manuscript. F.P. contributed to conceptualization of the research project, computational design, computational data analysis, and supervision of the work. A.C.K. contributed to conceptualization of the research project, experimental design, data analysis, supervision of the work, writing of the manuscript, and acquisition of the funding. All authors contributed to editing of the manuscript.

Competing interests

A.C.K. is a member of the SAB of Pairwise Plants, is an equity holder for Pairwise Plants and Beam Therapeutics, and receives royalties from Pairwise Plants, Beam Therapeutics, and Editas Medicine via patents licensed from Harvard University. A.C.K.'s interests have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. All other authors declare no competing financial interests.

Keywords

Base editing; Directed evolution; Genome editing; Nucleic acids; Adenosine deamination chemistry

Figures:

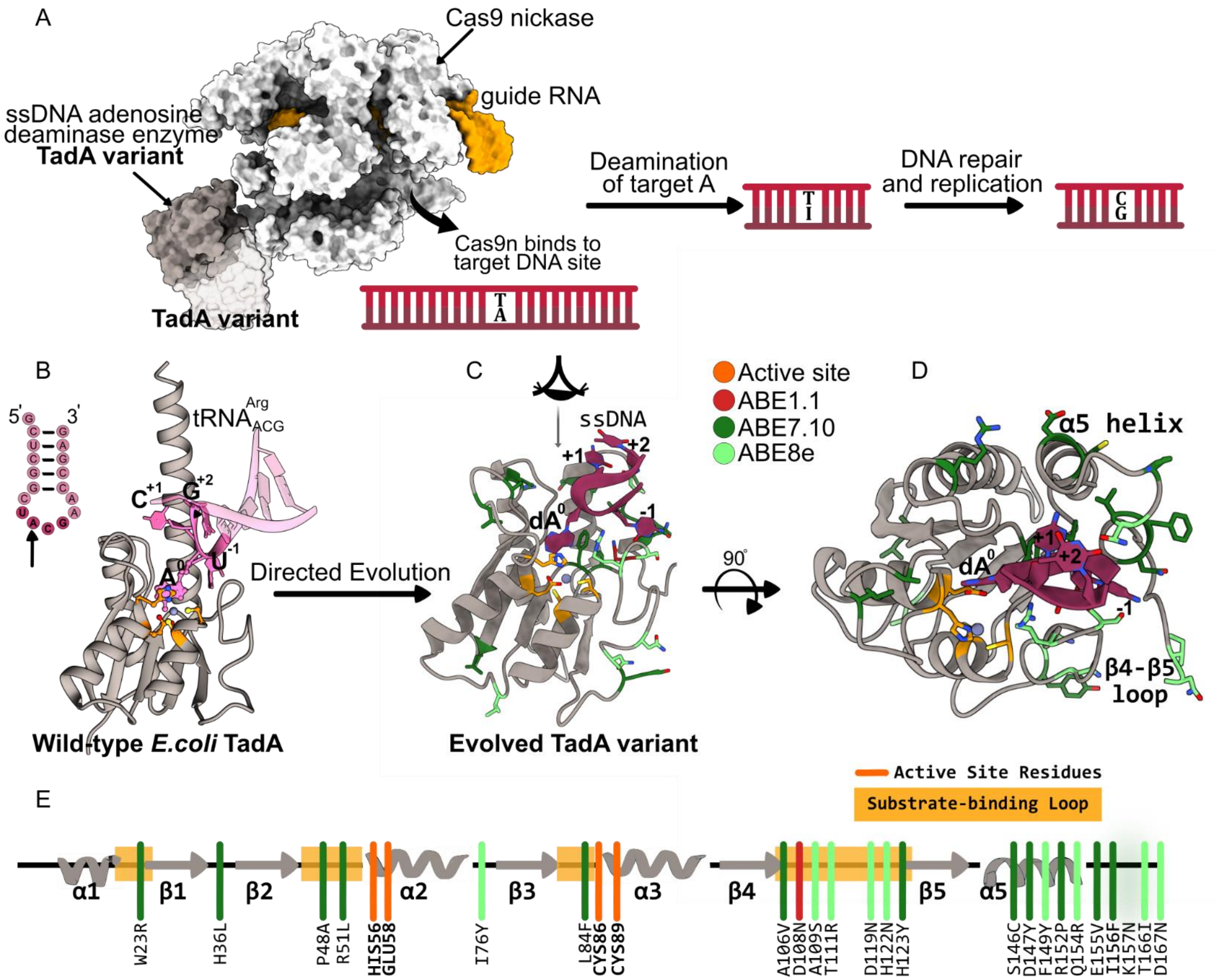
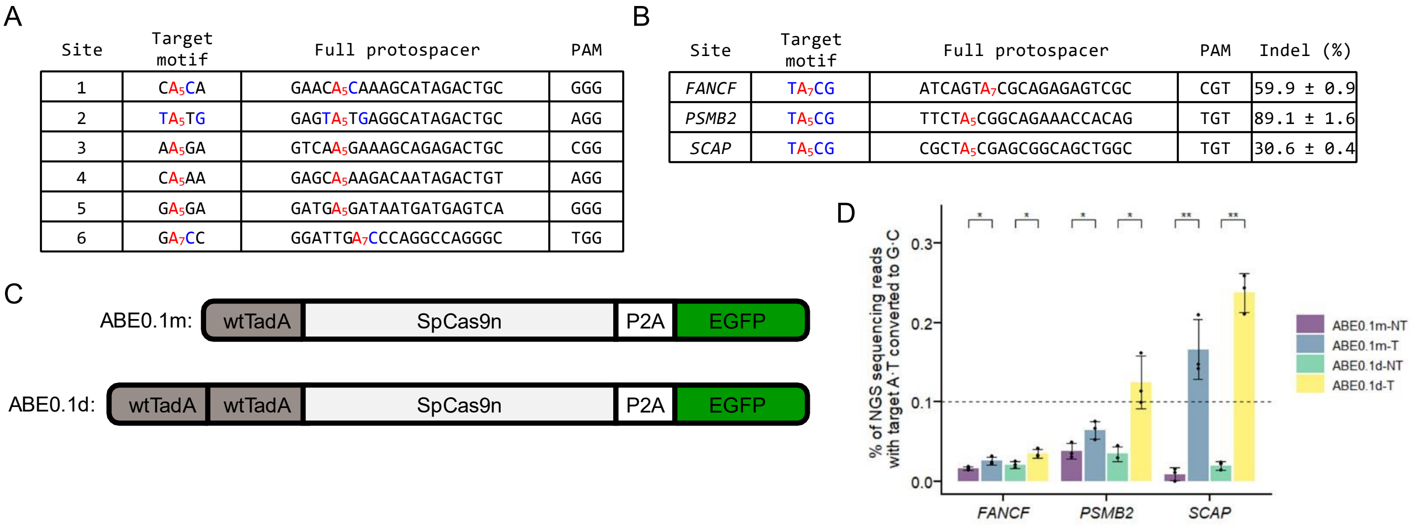


Figure 1. Overview of Adenine Base Editor (ABE) mechanism and evolution.

(A) ABEs (PDB ID: 6VPC) bind to the target genomic site via sequence complementary with the guide RNA (gRNA, orange). This base pairing results in the formation of an R-loop structure, in which one of the DNA strands protrudes out from the Cas9n:gRNA complex, and is presented to the Cas9n-fused TadA enzyme for deamination. TadA deaminates adenosine nucleobases within this exposed single-stranded DNA (ssDNA) “bubble” into inosines, which are subsequently modified into guanines by the DNA repair and/or replication machinery of the cell. Overall, ABEs introduce A•T to G•C base pair conversions. **(B)** Structure of the wild-type *E. coli* TadA bound to its substrate $tRNA^{Arg}{}_{ACG}$ (PDB ID: 1Z3A and 2B3J). The target U$\underline{A}$CG motif within the anticodon loop is splayed in the active site groove. **(C)** and **(D)** Structure of the evolved TadA8e bound to its ssDNA substrate, with the mutations identified through directed evolution in ABE1.1, ABE7.10, and ABE8e color-coded according to the legend (red, dark green, and light green, respectively). The active site residues are shown in orange. (**E**) Secondary structure of TadA, highlighting the position of critical active site elements and various ABE mutations. The color-coding matches that in **(C)** and **(D)**.



A

Site	Target motif	Full protospacer	PAM
1	CA$_5$CA	GAACA$_5$CAAAGCATAGACTGC	GGG
2	TA$_5$TG	GAGTA$_5$TGAGGCATAGACTGC	AGG
3	AA$_5$GA	GTCAA$_5$GAAAGCAGAGACTGC	CGG
4	CA$_5$AA	GAGCA$_5$AAGACAATAGACTGT	AGG
5	GA$_5$GA	GATGA$_5$GATAATGATGAGTCA	GGG
6	GA$_7$CC	GGATTGA$_7$CCCAGGCCAGGGC	TGG

B

Site	Target motif	Full protospacer	PAM	Indel (%)
FANCF	TA$_7$CG	ATCAGTA$_7$CGCAGAGAGTCGC	CGT	59.9 ± 0.9
PSMB2	TA$_5$CG	TTCTA$_5$CGGCAGAAACCACAG	TGT	89.1 ± 1.6
SCAP	TA$_5$CG	CGCTA$_5$CGAGCGGCAGCTGGC	TGT	30.6 ± 0.4

Figure 2. ABE0.1 can edit DNA at TACG motifs.

(A) Protospacer and PAM sequences of genomic loci at which ABE0.1 base editing activity was evaluated in the original evolution of the ABEs, with target motifs indicated. Nucleotides in blue are those that match the UACG motif that wtTadA targets in its native tRNA$^{Arg}{}_{ACG}$ target[2,14]. **(B)** Protospacer and PAM sequences of genomic loci at which ABE0.1 base editing activity was evaluated in this work, with target motifs indicated. Nucleotides in blue are those that match the UACG motif that wtTadA targets in its native tRNA$^{Arg}{}_{ACG}$ target. Average indel introduction efficiencies are also shown. HEK293T cells were transfected with plasmids encoding Cas9 and gRNA, and cells were lysed at 72 hours. The genomic DNA was extracted, and target loci were amplified via PCR and subjected to high-throughput sequencing (HTS). Indel introduction efficiencies were quantified with CRISPResso2. **(C)** Architectures of ABE0.1 constructs used in **(D)**. **(D)** A•T to G•C base editing efficiencies by ABE0.1m and ABE0.1d at the three genomic sites from **(B)**. HEK293T cells were transfected with plasmids encoding ABE0.1m or ABE0.1d and a NT or targeting gRNA. After 72 hours, fluorescence activated cell sorting (FACS) was used to sort for cells with the top 20-30% EGFP fluorescence (see SI Figure 2 for gating information). The cells were lysed, genomic DNA was extracted, and target loci were amplified via PCR and subjected to high-throughput sequencing (HTS). A•T to G•C base editing efficiencies were quantified with CRISPResso2. Values and error bars in **(B)** and **(D)** represent the mean and standard deviation for n = 2 biological replicates. Each replicate is marked individually in **(D)**. Data were analyzed with one-tailed T-tests and p-values are marked as following: p ≥ 0.05 not significant (ns), p = 0.01–0.05 significant (*), p = 0.001 – 0.01 very significant (**).

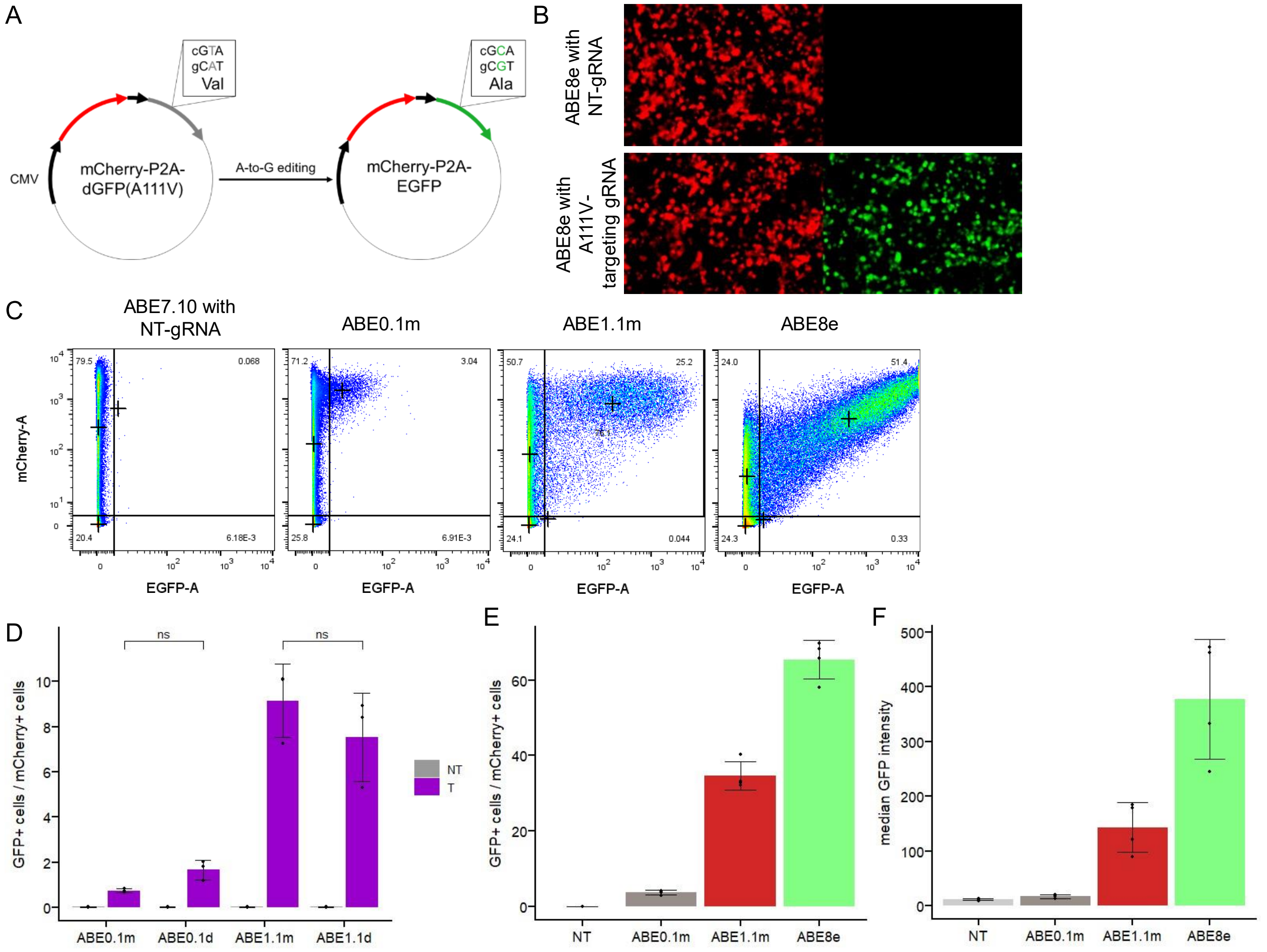


Figure 3. An EGFP reporter assay enhances editing detection by low-efficiency ABEs

(A) Schematic overview of the A111V EGFP turn-on reporter. **(B)** Confocal microscopy images of HEK239T cells treated with ABE8e, the A111V reporter, a non-targeting (NT) gRNA (top) or a targeting gRNA (bottom). HEK293T cells were transfected with plasmids encoding ABE8e, the A111V reporter, and a NT or targeting gRNA. After 72 hours, cells were imaged for mCherry (left) and EGFP (right) fluorescence on a confocal microscope. **(C)** Cells were treated as in **(B)**, but transfected with the ABE variants indicated, and analyzed by flow cytometry after 72 hours. Shown are mCherry fluorescence intensity (y-axis) and EGFP fluorescence intensity (x-axis) for the representative samples of the four conditions indicated. "+"'s in the upper right quadrants indicate the median EGFP fluorescence intensity of EGFP-positive cells. **(D)** Cells were treated as in **(C)**, but transfected with plasmids encoding ABE0.1m, ABE0.1d, ABE1.1m, or ABE1.1d, the A111V reporter, and a NT or targeting gRNA. Shown are the percent of transfected cells (determined based on mCherry fluorescence) with EGFP fluorescence (using gating strategies as shown in **(C)**) for all samples indicated. **(E)** Cells were treated as in **(C)**, but transfected with plasmids encoding ABE0.1m, ABE1.1m, or ABE8e, the A111V reporter, and a NT or targeting gRNA. Shown are the percent of transfected cells (determined based on mCherry fluorescence) with EGFP fluorescence (using gating strategies as shown in **(C)**) for all samples. **(F)** Cells were treated as in **(E)**. Shown are the median EGFP fluorescence intensity of mCherry, EGFP double positive cells for all samples. Values and error bars in **(D-F)** represent the mean and standard deviation for n = 3-4 biological replicates. Each replicate is marked individually. Data were analyzed with two-tailed T-tests and p-values are marked as following: p ≥0.05 not significant (ns).

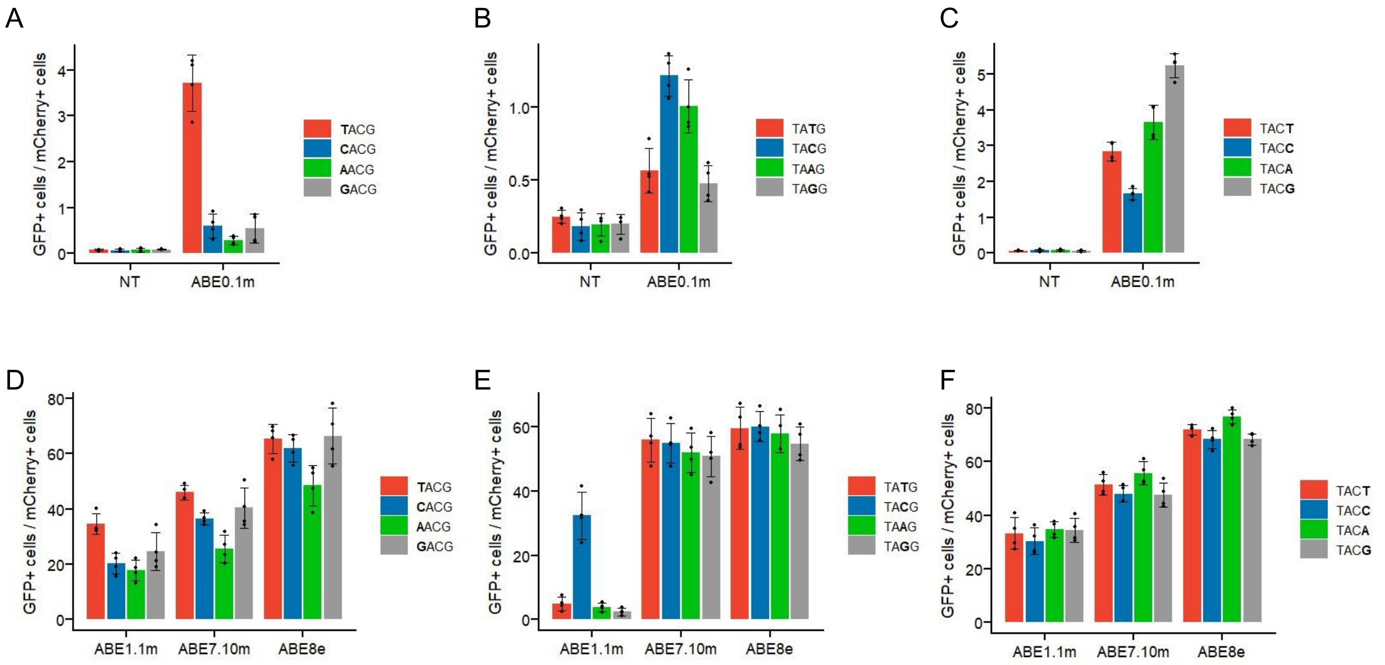


Figure 4. ABE0.1 shows strict sequence requirements for DNA base editing

(A-C) HEK293T cells were transfected with plasmids encoding ABE0.1m, the fluorescent reporter indicated, and a NT or targeting gRNA. After 72 hours, cells were analyzed by flow cytometry. Shown are the percent of transfected cells (determined based on mCherry fluorescence) with EGFP fluorescence (using gating strategies as shown in **Figure 3C**) for all samples indicated. Samples in **(A)** were transfected with A111V-derived reporters, with mutations at the -1 position (wild-type nucleotide is T^{-1}). Samples in **(B)** were transfected with H182Y-derived reporters, with mutations at the +1 position (wild-type nucleotide is C^{+1}). Samples in **(C)** were transfected with A111V-derived reporters, with mutations at the +2 position. **(D-F)** Samples were treated as in **(A-C)**, but transfected with ABE1.1m, ABE7.10m, or ABE8e constructs. Shown are the percent of transfected cells (determined based on mCherry fluorescence) with EGFP fluorescence (using gating strategies as shown in **Figure 3C**) for all samples indicated. Samples in **(D)** were transfected with A111V-derived reporters, with mutations at the -1 position (wild-type nucleotide is T^{-1}). Samples in **(E)** were transfected with H182Y-derived reporters, with mutations at the +1 position (wild-type nucleotide is C^{+1}). Samples in **(F)** were transfected with A111V-derived reporters, with mutations at the +2 position. Values and error bars represent the mean and standard deviation for n = 4 biological replicates. Each replicate is marked individually.

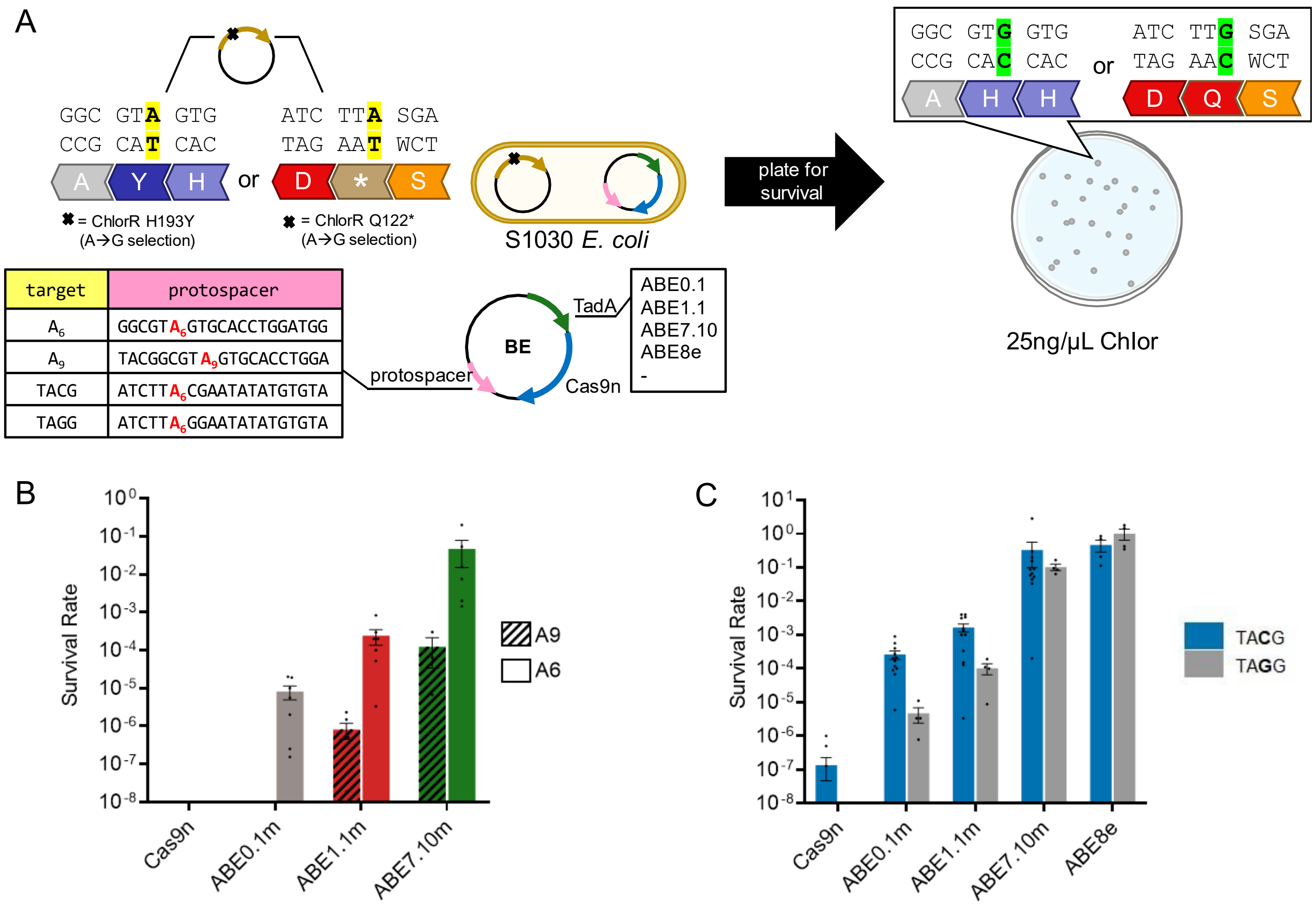
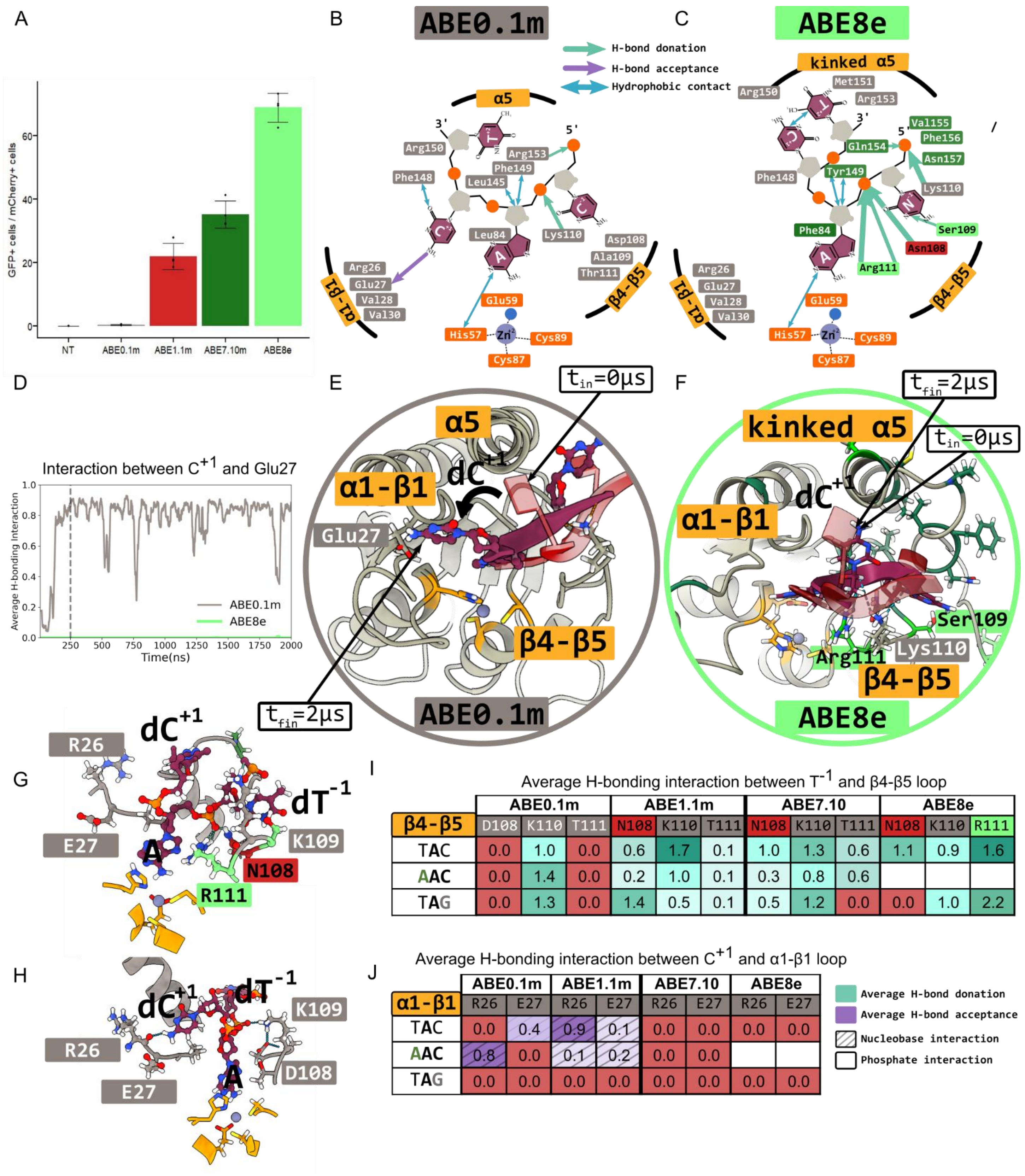


Figure 5. Optimization of bacterial directed evolution selection target results in measurable editing activity by ABE0.1

(A) Schematic of bacterial selection scheme for evaluating activity of ABEs in *E. coli*. S1030 *E. coli* harboring an inactivated chloramphenicol resistance gene (either via an H193Y mutation, as in the original directed evolution system[2], or via a sequence-optimized system using a Q122* mutation) were transformed with a plasmid encoding a theophylline-inducible ABE variant (ABE0.1, ABE1.1, ABE7.10, ABE8e, or Cas9n) and a gRNA targeting the appropriate ChlorR mutation. ABE expression was induced for 18 hours, and cultures were plated on both 0 mg/mL and 25 mg/mL chloramphenicol plates. Survival rate was calculated by taking the fraction of surviving colonies at 25 ng/uL chloramphenicol compared to those plated at 0 ng/uL chloramphenicol. **(B)** Survival rates for Cas9n, ABE0.1, ABE1.1, and ABE7.10 are shown for the H193Y ChlorR mutation-based selection systems, which use a TAGT motif. In one system, the target A is in position 9 within the protospacer (labeled as A9, which matches the original direction evolution selection system[2]), and in another, we used a PAM-relaxed Cas9n variant to move the target A to position 6 (labeled as A6). **(C)** Survival rates for Cas9n, ABE0.1, ABE1.1, ABE7.10, and ABE8e are shown for the Q122* ChlorR mutation-based selection systems, which use TACG (labeled as C^{+1} WT) and TAGG (labeled as G^{+1}) motifs. Values and error bars represent the mean and standard error of mean for n = 4-12 biological replicates. Each replicate is marked individually.



Average H-bonding interaction between T^{-1} and β4-β5 loop

	ABE0.1m			ABE1.1m			ABE7.10			ABE8e		
β4-β5	D108	K110	T111	N108	K110	T111	N108	K110	T111	N108	K110	R111
TAC	0.0	1.0	0.0	0.6	1.7	0.1	1.0	1.3	0.6	1.1	0.9	1.6
AAC	0.0	1.4	0.0	0.2	1.0	0.1	0.3	0.8	0.6			
TAG	0.0	1.3	0.0	1.4	0.5	0.1	0.5	1.2	0.0	0.0	1.0	2.2

Average H-bonding interaction between C^{+1} and α1-β1 loop

	ABE0.1m		ABE1.1m		ABE7.10		ABE8e	
α1-β1	R26	E27	R26	E27	R26	E27	R26	E27
TAC	0.0	0.4	0.9	0.1	0.0	0.0	0.0	0.0
AAC	0.8	0.0	0.1	0.2	0.0	0.0		
TAG	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

🟩 Average H-bond donation
🟪 Average H-bond acceptance
▨ Nucleobase interaction
□ Phosphate interaction

Figure 6. Conformational basis for sequence specificity and activity for ABE variants

(A) HEK293T cells were transfected with plasmids encoding ABE0.1m, ABE1.1m, ABE7.10m, or ABE8e, a A111V-based fluorescent reporter with a CACT target motif, and a NT or targeting gRNA. After 72 hours, cells were analyzed by flow cytometry. Shown are the percent of transfected cells (determined based on mCherry fluorescence) with EGFP fluorescence (using gating strategies as shown in **Figure 3C**) for all samples indicated. Values and error bars represent the mean and standard deviation for n = 4 biological replicates. Each replicate is marked individually. **(B-C)** All-atom molecular simulations of full-length ABE variants bound to gRNA and target DNA were initiated from PDB ID: 6VPC. The first 200 ns of simulation was excluded from analysis[21]. The molecular interactions between the target ssDNA (CAC motif) and TadA residues in the **(B)** ABE0.1m and **(C)** ABE8e complexes are summarized as interaction maps. The TadA residues that are within the first interaction shell of the CAC trinucleotides during the course of the simulation are shown. Hydrogen bonds (H-bonds) between these residues and the DNA bases are depicted as green and purple arrows, whose thickness is proportional to the stability of the H-bond itself (defined as the frequency of appearance of that H-bond during the simulation). Key hydrophobic contacts are also indicated with double-sided blue arrows. **(D)** Rolling averages (every 20 ns) of the H-bonding interaction between the exocyclic amino group of C^{+1} and the α1-β1 loop residues in the ABE0.1m (grey) and ABE8e (green) complexes. **(E-F)** Simulation snapshots of the initial and final states of the **(D)** ABE0.1m and **(E)** ABE8e complexes, highlighting the conformational difference in the position of the C^{+1} nucleobase during these simulations. **(G-H)** Simulation snapshots of the TAC "wild-type" motif target bound to **(G)** ABE8e and **(H)** ABE0.1highlighting the drastically different conformations adopted by the C^{+1} base in the two simulations. **(I-J)** Average H-bonding interactions between the target DNA and the **(I)** β4-β5 loop residues and **(J)** α1-β1 loop residues for the target motifs and ABE variants indicated.

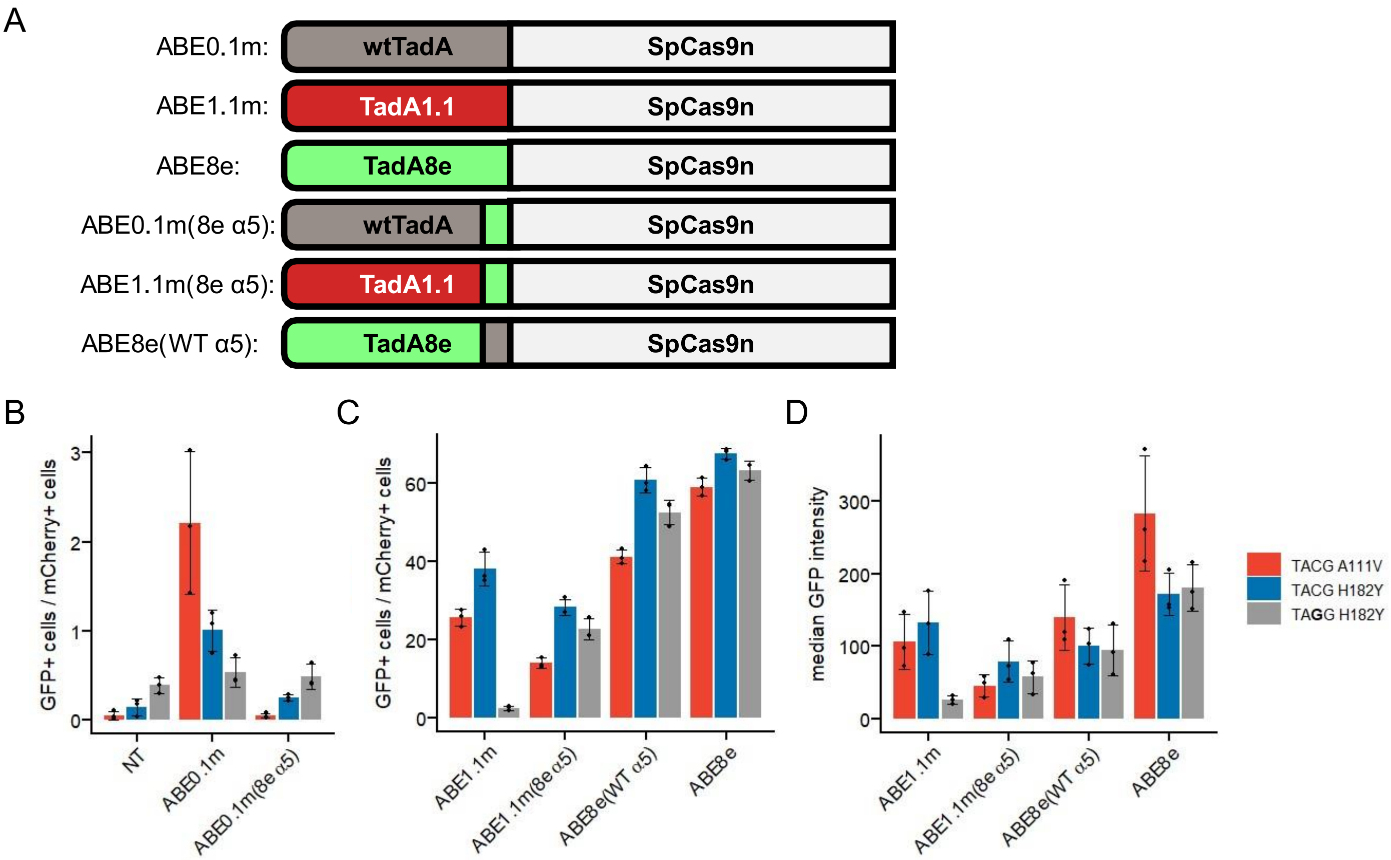


Figure 7. Mutations to the α5-helix affect C^{+1} requirements

(A) Architectures of "helix-swapping" ABE constructs used in **(B-D)**. HEK293T cells were transfected with plasmids encoding the ABE variants indicated in **(A)**, the "wild-type" A111V TACG reporter (labelled as TACG A111V), the "wild-type" H182Y TACG reporter (labelled as TACG H182Y), or the H182Y TAGG reporter (labelled as TAGG H182Y), and a NT or targeting gRNA. After 72 hours, cells were analyzed by flow cytometry. **(B-C)** Shown are the percent of transfected cells (determined based on mCherry fluorescence) with EGFP fluorescence (using gating strategies as shown in **Figure 3C**) for all samples indicated. Samples in **(B)** were transfected with ABE0.1m-derived variants. Samples in **(C)** were transfected with ABE1.1m- and ABE8e-derived variants. **(D)** Shown are the median EGFP fluorescence intensity of mCherry, EGFP double positive cells for all samples indicated. Values and error bars represent the mean and standard deviation for n = 3 biological replicates. Each replicate is marked individually.

References

[1] A. C. Komor, Y. B. Kim, M. S. Packer, J. A. Zuris, D. R. Liu, *Nature* **2016**, *533*, 420–424.

[2] N. M. Gaudelli, A. C. Komor, H. A. Rees, M. S. Packer, A. H. Badran, D. I. Bryson, D. R. Liu, *Nature* **2017**, *551*, 464–471.

[3] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, E. Charpentier, *Science* **2012**, *337*, 816–821.

[4] F. Jiang, D. W. Taylor, J. S. Chen, J. E. Kornfeld, K. Zhou, A. J. Thompson, E. Nogales, J. A. Doudna, *Science* **2016**, *351*, 867–871.

[5] X. Wang, J. Li, Y. Wang, B. Yang, J. Wei, J. Wu, R. Wang, X. Huang, J. Chen, L. Yang, *Nat Biotechnol* **2018**, *36*, 946–949.

[6] A. C. Komor, K. T. Zhao, M. S. Packer, N. M. Gaudelli, A. L. Waterbury, L. W. Koblan, Y. B. Kim, A. H. Badran, D. R. Liu, *Science Advances* **2017**, *3*, eaao4774.

[7] A. S. Martin, D. J. Salamango, A. A. Serebrenik, N. M. Shaban, W. L. Brown, R. S. Harris, *Sci Rep* **2019**, *9*, 497.

[8] N. M. Gaudelli, D. K. Lam, H. A. Rees, N. M. Solá-Esteves, L. A. Barrera, D. A. Born, A. Edwards, J. M. Gehrke, S.-J. Lee, A. J. Liquori, R. Murray, M. S. Packer, C. Rinaldi, I. M. Slaymaker, J. Yen, L. E. Young, G. Ciaramella, *Nat Biotechnol* **2020**, 1–9.

[9] M. F. Richter, K. T. Zhao, E. Eton, A. Lapinaite, G. A. Newby, B. W. Thuronyi, C. Wilson, L. W. Koblan, J. Zeng, D. E. Bauer, J. A. Doudna, D. R. Liu, *Nat Biotechnol* **2020**, DOI 10.1038/s41587-020-0453-z.

[10] L. Chen, S. Zhang, N. Xue, M. Hong, X. Zhang, D. Zhang, J. Yang, S. Bai, Y. Huang, H. Meng, H. Wu, C. Luan, B. Zhu, G. Ru, H. Gao, L. Zhong, M. Liu, M. Liu, Y. Cheng, C. Yi, L. Wang, Y. Zhao, G. Song, D. Li, *Nat Chem Biol* **2022**, 1–10.

[11] K. L. Rallapalli, A. C. Komor, F. Paesani, *Sci. Adv.* **2020**, *6*, eaaz2309.

[12] K. L. Rallapalli, B. L. Ranzau, K. R. Ganapathy, F. Paesani, A. C. Komor, *The CRISPR Journal* **2022**, *5*, 294–310.

[13] K. M. McKenney, M. A. T. Rubio, J. D. Alfonzo, in *The Enzymes* (Ed.: G.F. Chanfreau), Academic Press, **2017**, pp. 51–88.

[14] J. Wolf, A. P. Gerber, W. Keller, *The EMBO Journal* **2002**, *21*, 3841–3851.

[15] J. Grünewald, R. Zhou, S. Iyer, C. A. Lareau, S. P. Garcia, M. J. Aryee, J. K. Joung, *Nat Biotechnol* **2019**, *37*, 1041–1048.

[16] C. Zhou, Y. Sun, R. Yan, Y. Liu, E. Zuo, C. Gu, L. Han, Y. Wei, X. Hu, R. Zeng, Y. Li, H. Zhou, F. Guo, H. Yang, *Nature* **2019**, *571*, 275–278.

[17] H. A. Rees, C. Wilson, J. L. Doman, D. R. Liu, *Sci Adv* **2019**, *5*, DOI 10.1126/sciadv.aax5717.

[18] J. Kim, V. Malashkevich, S. Roday, M. Lisbin, V. L. Schramm, S. C. Almo, *Biochemistry* **2006**, *45*, 6407–6416.

[19] H. C. Losey, A. J. Ruthenburg, G. L. Verdine, *Nat Struct Mol Biol* **2006**, *13*, 153–159.

[20] Y. Elias, R. H. Huang, *Biochemistry* **2005**, *44*, 12057–12065.

[21] A. Lapinaite, G. J. Knott, C. M. Palumbo, E. Lin-Shiao, M. F. Richter, K. T. Zhao, P. A. Beal, D. R. Liu, J. A. Doudna, *Science* **2020**, *369*, 566–571.

[22] Y. Wang, F. Wang, R. Wang, P. Zhao, Q. Xia, *Sci Rep* **2015**, *5*, 16273.

[23] H. Zhao, *Biotechnology and Bioengineering* **2007**, *98*, 313–317.

[24] M. D. Lane, B. Seelig, *Current Opinion in Chemical Biology* **2014**, *22*, 129–136.

[25] N. J. Loman, R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, M. J. Pallen, *Nat Biotechnol* **2012**, *30*, 434–439.

[26] F. Meacham, D. Boffelli, J. Dhahbi, D. I. Martin, M. Singer, L. Pachter, *BMC Bioinformatics* **2011**, *12*, 451.

[27] J. L. Fu, T. Kanno, S.-C. Liang, A. J. M. Matzke, M. Matzke, *G3 (Bethesda)* **2015**, *5*, 1849–1855.

[28] Y. Sun, J. H. Ambrose, B. S. Haughey, T. D. Webster, S. N. Pierrie, D. F. Muñoz, E. C. Wellman, S. Cherian, S. M. Lewis, L. E. Berchowitz, G. P. Copenhaver, *PLoS Genet* **2012**, *8*, e1002968.

[29] S. A. Lynch, J. P. Gallivan, *Nucleic Acids Research* **2009**, *37*, 184–192.

[30] H. Nishimasu, X. Shi, S. Ishiguro, L. Gao, S. Hirano, S. Okazaki, T. Noda, O. O. Abudayyeh, J. S. Gootenberg, H. Mori, S. Oura, B. Holmes, M. Tanaka, M. Seki, H. Hirano, H. Aburatani, R. Ishitani, M. Ikawa, N. Yachie, F. Zhang, O. Nureki, *Science* **2018**, *361*, 1259–1262.

[31] L. W. Koblan, J. L. Doman, C. Wilson, J. M. Levy, T. Tay, G. A. Newby, J. P. Maianti, A. Raguram, D. R. Liu, *Nat Biotechnol* **2018**, *36*, 843–846.

[32] P. Wang, L. Xu, Y. Gao, R. Han, *Molecular Therapy* **2020**, *28*, 1696–1705.

[33] Z. Bodai, A. L. Bishop, V. M. Gantz, A. C. Komor, *Nat Commun* **2022**, *13*, 2351.

[34] FlowJo^{TM} Software for Windows Version 10.8.1. Ashland, OR: Becton, Dickinson and Company; 2021, **2022**.

[35] K. Clement, H. Rees, M. C. Canver, J. M. Gehrke, R. Farouni, J. Y. Hsu, M. A. Cole, D. R. Liu, J. K. Joung, D. E. Bauer, L. Pinello, *Nat Biotechnol* **2019**, *37*, 224–226.

[36] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical, Computing, Vienna, Austria. URL https://www.R-project.org/, **2022**.

[37] H. Wickham, *Ggplot2*, Springer Cham, **2016**.

[38] A. Kassambara, **2022**.

[39] J. C. Carlson, A. H. Badran, D. A. Guggiana-Nilo, D. R. Liu, *Nat Chem Biol* **2014**, *10*, 216–222.

[40] C. T. Chung, S. L. Niemela, R. H. Miller, *Proc Natl Acad Sci U S A* **1989**, *86*, 2172–2175.

[41] B. Webb, A. Sali, *Methods Mol Biol* **2014**, *1137*, 1–15.

[42] C. Anders, O. Niewoehner, A. Duerst, M. Jinek, *Nature* **2014**, *513*, 569–573.

[43] J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath, A. Onufriev, *Nucleic Acids Res* **2005**, *33*, W368-371.

[44] R. Anandakrishnan, B. Aguilar, A. V. Onufriev, *Nucleic Acids Res* **2012**, *40*, W537-541.

[45] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J Comput Chem* **2004**, *25*, 1605–1612.

[46] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583–589.

[47] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926–935.

[48] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J Chem Theory Comput* **2015**, *11*, 3696–3713.

[49] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, M. Orozco, *Biophysical Journal* **2007**, *92*, 3817–3829.

[50] M. Zgarbová, M. Otyepka, J. Sponer, A. Mládek, P. Banáš, T. E. Cheatham, P. Jurečka, *J Chem Theory Comput* **2011**, *7*, 2886–2902.

[51] P. Banáš, D. Hollas, M. Zgarbová, P. Jurečka, M. Orozco, T. E. Cheatham, J. Šponer, M. Otyepka, *J Chem Theory Comput* **2010**, *6*, 3836–3849.

[52] I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, F. Battistini, J. L. Gelpí, C. González, M. Vendruscolo, C. A. Laughton, S. A. Harris, D. A. Case, M. Orozco, *Nat Methods* **2016**, *13*, 55–58.

[53] R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, R. C. Walker, *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.

[54] P. Li, K. M. Merz, *J Chem Inf Model* **2016**, *56*, 599–604.

[55] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, R. J. Woods, *Journal of Computational Chemistry* **2005**, *26*, 1668–1688.

[56] Case, David & Ben-Shalom, Ido & Brozell, S.R. & Cerutti, D.S. & Cheatham, Thomas & Cruzeiro, V.W.D. & Darden, Thomas & Duke, Robert & Ghoreishi, Delaram & Gilson, Michael & Gohlke, H. & Götz, Andreas & Greene, D. & Harris, Robert & Homeyer, N. & Huang, Yandong & Izadi, Saeed & Kovalenko, Andriy & Kurtzman, Tom & Kollman, P.A. Amber 2018., **2018**.

[57] D. R. Roe, T. E. Cheatham, *J Chem Theory Comput* **2013**, *9*, 3084–3095.

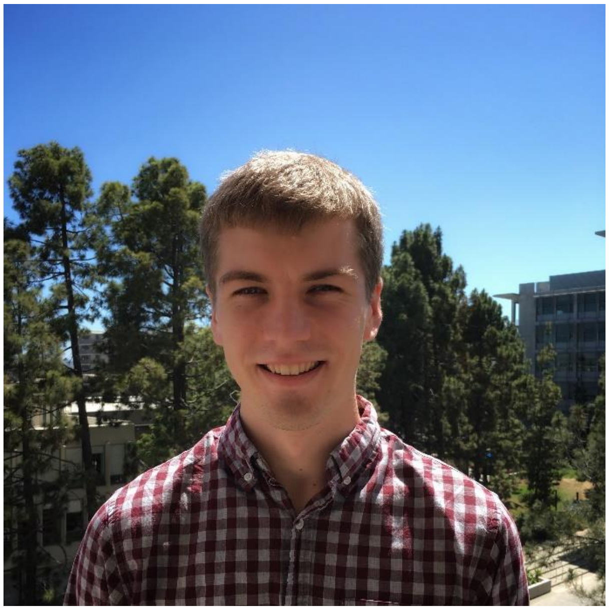
[58] D. R. Roe, T. E. Cheatham, *J Comput Chem* **2018**, *39*, 2110–2117.

[59] T. D. Goddard, C. C. Huang, E. C. Meng, E. F. Pettersen, G. S. Couch, J. H. Morris, T. E. Ferrin, *Protein Sci* **2018**, *27*, 14–25.

[60] E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, *Protein Sci* **2021**, *30*, 70–82.

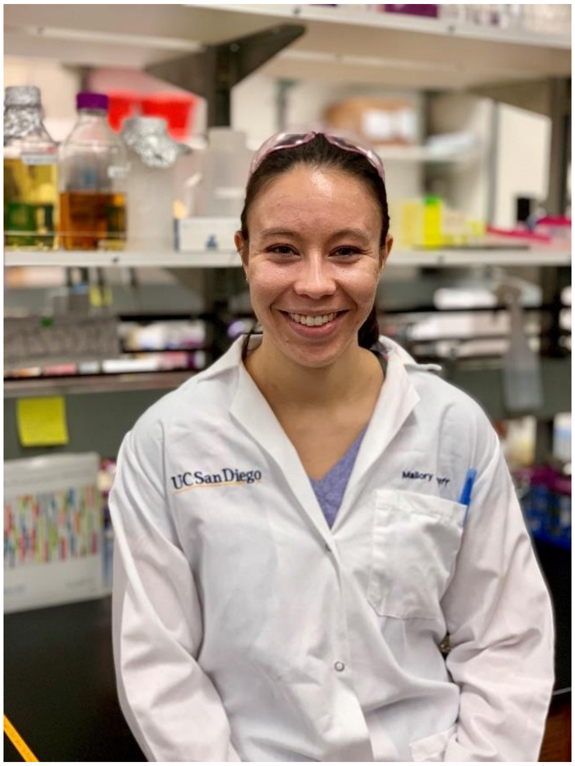
[61] J. D. Hunter, *Computing in Science & Engineering* **2007**, *9*, 90–95.

Biographical Information



Brodie Ranzau received his B. S. degree in Biochemistry and Molecular Biology from Otterbein University in 2017. While there he studied the alternative splicing of lipid droplet proteins in the lab of Prof. John Tansey. He is currently a PhD student in the lab of Prof. Alexis Komor at the University of California, San Diego where he studies the development of TadA from an RNA-editing protein to the DNA-editing protein used in the Adenine Base Editor.

Kartik Rallapalli is currently a PhD student in the labs of Prof. Francesco Paesani and Prof. Alexis Komor in the Department of Chemistry and Biochemistry, University of California San Diego. Her thesis research focuses on developing a molecular understanding of the activity of DNA-editing enzymes, specifically the Adenine Base Editors. She received her Master's degree at Indian Institute of Technology, Delhi and her Bachelor’s degree at Miranda House, University of Delhi.



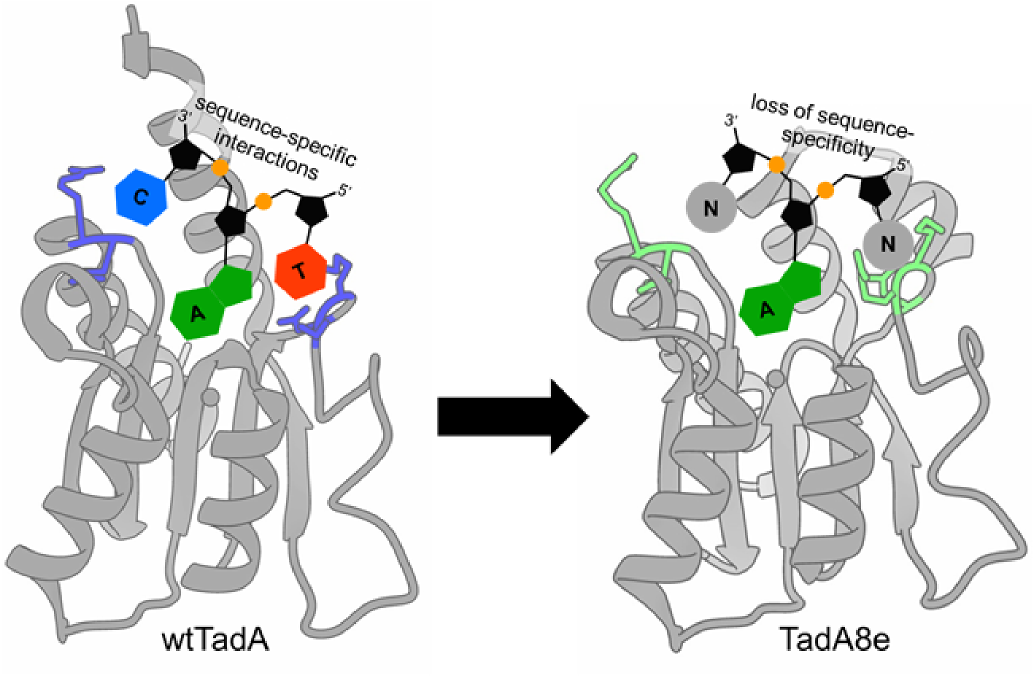
Mallory Evanoff grew up in Albuquerque, NM and attended Carnegie Mellon University for a Bachelor's of Science in Chemistry. While there, she studied nucleic acid conjugated fluorophore tools in the laboratory of Bruce Armitage. She is currently a PhD student under the mentorship of Alexis Komor at the University of California, San Diego where she develops new genome editing tools and studies the structure-function relationship of current base editors. In addition to her research, Mallory is a lecturer at UCSD, and a co-creator of the Komor Lab's Base Editing Research Practice Partnership with local high schools.

Francesco Paesani received his Ph.D. in Theoretical Physical Chemistry from the University of Rome "La Sapienza" in 2000. He was a postdoctoral fellow at the University of California, Berkeley, and the University of

Utah. In 2009, he joined the faculty of the University of California, San Diego, where he was promoted to Associate Professor in 2015 and Professor in 2017. He is currently the Kurt Shuler Faculty Scholar in the Department of Chemistry and Biochemistry. He was awarded the ACS OpenEye Outstanding Junior Faculty Award in Computational Chemistry in 2014, the NSF CAREER Award in 2015, the ACS Early Career Award in Theoretical Chemistry in 2016, and the Cozzarelli Prize (National Academy of Sciences U.S.A.).

Alexis C. Komor received her B.S. degree in chemistry from the University of California, Berkeley in 2008 and her Ph.D. in chemistry from the California Institute of Technology in 2014. She pursued postdoctoral work as a Ruth L. Kirschstein NIH Postdoctoral Fellow in the laboratory of David R. Liu, where she developed base editing, a new genome editing method that enables the introduction of point mutations in genomic DNA via the chemical modification of nucleobases. Alexis joined the Department of Chemistry and Biochemistry at the University of California, San Diego in 2017, where her lab develops and applies new precision genome editing techniques to the functional genomics field. Alexis's contributions in teaching, mentoring, and research have been recognized through the Cottrell Scholar Award, the "Talented 12" recognition by C&EN News, an NSF Faculty Early Career Development award, an NIH early stage investigator MIRA, and a "40 under 40" recognition in healthcare by Fortune Magazine.

TOC



Adenine base editors (ABEs) are genome editing tools that convert A•T base pairs to G•C using an extensively evolved tRNA- adenosine deaminase enzyme, TadA. Here we show that wtTadA can perform DNA base editing, with a strict sequence motif requirement of TAC. We additionally investigate how the sequence recognition of TadA variants change as they accumulate mutations to better edit DNA substrates.