Title

Studying the Sequence Recognition Properties of Base Editing Enzymes

Author

Ranzau, Brodie Linck

Publication Date

2023

UNIVERSITY OF CALIFORNIA SAN DIEGO

Studying the Sequence Recognition Properties of Base Editing Enzymes

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Biochemistry and Molecular Biophysics

by

Brodie L. Ranzau

Committee in charge:

Professor Alexis Komor, chair
Professor J. Andrew McCammon
Professor Bradley Moore
Professor Wei Wang
Professor Brian Zid

2023

The Dissertation of Brodie L. Ranzau is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

For Hannah,

> “We're in this forever thankfully
> I couldn't do this with anybody else
> And I don't think I could do it by myself”
>
> \- Daði Freyr

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

To Hannah, who has always believed in me and encouraged me through the difficulties of this season. You challenge me to grow beyond my boundaries as we learn new skills and experience new adventures together. Thank you for always listening to my ideas, and for helping me to turn the best ones into a reality. I look forward to returning all of your love and support as you finish your own PhD and during our future adventures.

To Alexis, thank you for your mentorship and providing all of the resources I needed to start all of the different projects reported here. Though I came to UCSD with other plans for research, thank you for believing in my abilities and bringing me into the field of genome editing. Thank you for creating a space where creativity and hard work can flourish.

To everyone I worked alongside over these last six years, thank you for being a wonderful group of people who lift each other up after failures and celebrate each others' success. Special thanks to my undergraduate mentee, Helen. Many of these experiments would not have seen the light of day without your hard work and your ability to learn new skills quickly. You helped me to be a better teacher and kept our experiments running at a quick pace. You made a great lab environment even better with your creativity and excitement. Thank you to Zsolt, whose mentorship and friendship has made me a better scientist and person. Thank you to Mallory, Sifeng, and Elizabeth for walking alongside me as we made a lot of mistakes, developed protocols, and tackled a seemingly endless number of difficulties in our projects. Thank you to Carlos, for your passion for teaching and lifting up others. Your dedication developing an outreach program kept the project moving during hard times, and I'm grateful to have been a part of it. Thank you to Kartik for being a great computational partner and for working on so many projects with me. Together, we built better stories than we could have alone.

Thank you to Randall, Joe, Meghan, Billy, Mary, Ethan, Brooke, Ricardo, and the rest of my friends at Grace City SD. Your friendship has meant so much to me throughout these years and made this new city feel like a home. You model grace and care in all that you do and helped me build confidence on a stage and as a musician.

Thank you to Nick, Geoff, Alex, Nik, and Danny for opening their arms to a stranger and making my evenings brighter after long days in the lab.

Thank you to my parents and family, who have always loved and supported me, even as I've moved further away every few years. I look forward to being closer and seeing you more often.

Thank you to Grandpa Roby, who made the best better, in all that he did. He showed his support for me in every season of my life and is greatly missed as I write the final words of this chapter. I hope to build things and love others as well as you did.

Chapter 1 has been reproduced, in full, with permission, from: **Ranzau, B. L.** & Komor, A. C. Genome, Epigenome, and Transcriptome Editing via Chemical Modification of Nucleobases in Living Cells. *Biochemistry* **58**, 330–335 (2019). The dissertation author was the primary author on all reprinted materials.

Chapter 2 has been adapted, in part, with permission, from Rallapalli, K. L., **Ranzau, B. L.**, Ganapathy, K. R., Paesani, F. & Komor, A. C. Combined Theoretical, Bioinformatic, and Biochemical Analyses of RNA Editing by Adenine Base Editors. *The CRISPR Journal* **5**, 294–310 (2022). The dissertation author was the primary author on all reprinted materials.

Chapter 3 has been reproduced, in full, with permission, from **Ranzau, B. L.**, Rallapalli, K. L., Evanoff, M., Paesani, F. & Komor, A. C. The Wild-Type tRNA Adenosine Deaminase Enzyme TadA Is Capable of Sequence-Specific DNA Base Editing. *ChemBioChem*, e202200788 (2023). The dissertation author was the primary author of this paper.

Chapter 7 has been adapted, in part, with permission, from Vasquez, C. A. Evanoff, M., **Ranzau, B. L.**, Gu, S., Deters, E., Komor, A. C. Curing “GFP-itis” in Bacteria with Base Editors: Development of a Genome Editing Science Program Implemented with High School Biology Students. *The CRISPR Journal* **6**, 186–195 (2023).

VITA

2017 Bachelor of Science in Biochemistry and Molecular Biology, Otterbein University, Westerville, OH

2019 Master of Science in Chemistry, University of California San Diego, La Jolla, CA

2023 Doctor of Philosophy in Biochemistry and Molecular Biophysics, University of California San Diego, La Jolla, CA

PUBLICATIONS

Ranzau, B. L.; Rallapalli, K. L.; Evanoff, M.; Paesani, F.; Komor, A. C. "The wild-type tRNA adenosine deaminase enzyme TadA is capable of sequence-specific DNA base editing," *ChemBioChem*, **2023**, e202200788.
C*over art*: e202300396

Vasquez, C. A.; Evanoff, M.; **Ranzau, B. L.**; Gu, S.; Deters, E.; Komor, A. C. "Curing "GFP-itis" in Bacteria with Base Editors: Development of a Genome Editing Science Program Implemented with High School Biology Students," *The CRISPR Journal*, **2023**, 6, 186-195.

Rallapalli, K. L.; **Ranzau, B. L.**; Ganapathy, K. R.; Paesani, F.; Komor, A. C. "Combined Theoretical, Bioinformatic, and Biochemical Analyses of RNA Editing by Adenine Base Editors," *The CRISPR Journal*, **2022**, 5, 294-310.

Ranzau, B. L.; Komor, A. C. "Genome, Epigenome, and Transcriptome Editing via the Chemical Modification of Nucleobases in Living Cells," *Biochemistry*, **2019**, 58, 330-335.

ABSTRACT OF THE DISSERTATION

Studying the Sequence Recognition Properties of Base Editing Enzymes

by

Brodie L. Ranzau

Doctor of Philosophy in Biochemistry and Molecular Biophysics

University of California San Diego, 2023

Professor Alexis Komor, Chair

Base editors are tools that chemically modify the nucleobases of DNA and RNA in a programmable manner, allowing for genome, epigenome, and transcriptome editing in live cells. These tools can be used to introduce specific base transitions in DNA or RNA, manipulate methylation patterns in the epigenome, and create genetically encoded libraries in

target genes. These various functions can be used to modulate every aspect of the central dogma. The efficiency and precision of base editors makes them useful in both basic research and in the development of new therapies.

The adenosine base editor (ABE) was developed by evolving the RNA-modifying enzyme TadA to also accept DNA as substrate. This suggests that the base editing tool kit can be expanded to use many of the naturally occurring RNA-modifications as intermediates. However, current base editors remain limited to deaminases, suggesting that more complex chemical reactions face additional barriers. To better inform the creation of new base editors, we set out to better understand the molecular mechanisms of the wild-type TadA enzyme (used in ABE0.1) and the DNA-modifying variants developed for base editing (ABE7.10, ABE8e, ABE8.20, etc.).

First, we observe that ABE0.1 shows high Cas9-independent, off-target RNA-editing at UACG motifs in accessible loops. In chapter 3, we ask if this discovery also extends to DNA and find that ABE0.1 can perform targeted DNA-editing at TACG motifs. This editing is inefficient though and may be easily missed, so we leveraged this discovery to develop a fluorescent reporter of base editing that greatly increases our ability to detect low-levels of base editing. We finish out the chapter by using this fluorescent reporter to better understand the sequence-dependent editing of ABE0.1 and evolved ABEs. Finally, we apply this new understanding of an early base editor to develop more sequence-optimized fluorescent reporters to: detect DNA editing by RNA-modifying enzymes that could be used for base editing in chapters 4 and 5, develop a protocol for evolving new base editors in mammalian cells in chapter 6, and developing an outreach experiment to introduce base editing to high school students in chapter 7. Through the experiments in each of these sections, we gain a better understanding of how the sequence recognition properties of RNA-modifying enzymes can be leveraged to develop new base editors.

Chapter 1 Introduction

Introduction

The development of tools that directly chemically modify the nucleobases of DNA and RNA (via enzymatic methylation, deamination, or demethylation) in living cells has opened the door for studying and manipulating the components of the central dogma. These tools have been enabled by an increasing understanding of CRISPR systems and their ability to recognize and bind to specific dsDNA (for Cas9 and Cas12 enzymes)[1] and ssRNA (for Cas13 enzymes)[2] sequences inside living cells. In these systems, a guide RNA (gRNA) is bound by a Cas enzyme to form a ribonucleoprotein (RNP) complex. The RNP recognizes a target locus (the protospacer) through base pairing between the gRNA and the target nucleic acid sequence. When the target is DNA, this process separates the two DNA strands, creating a DNA-RNA heteroduplex with a small single-stranded DNA (ssDNA) region at the target locus (known as an R-loop).[3] Once bound to its target dsDNA sequence, the Cas enzyme cleaves both phosphodiester backbones of the target nucleic acid using either a single (in the case of Cas12 enzymes)[4–6] or two distinct (in the case of Cas9)[1] endonuclease domains. If the target is ssRNA, the Cas13 enzyme becomes activated upon target binding, and cleaves the phosphodiester backbone of both the target sequence and any neighboring ssRNA molecules in a promiscuous manner.[7–9] This RNA-guided endonuclease activity has been utilized extensively for genome editing and RNA degradation purposes.[10]

The catalytic activity of Cas enzymes can be inactivated by introducing specific mutations into the endonuclease domains to create catalytically dead Cas variants (dCas) that use a gRNA to bind a target DNA or RNA sequence without cleaving the phosphodiester backbone.[11] A diverse set of genome, epigenome, and transcriptome editing tools (collectively referred to here as base editors) have been developed from these dCas proteins by physically attaching them to various nucleobase modifying enzymes.[12] In this way, the chemical activity of these enzymes can be confined to specific genomic and transcriptomic loci where they chemically modify a canonical nucleobase into either a noncanonical DNA base or a naturally occurring modified base. As such, different base editing tools allow researchers to modify and study the effects of nucleic acid primary sequences, chromatin organization, RNA activity and stability, and DNA damage. In this perspective, we discuss the various base editor tools available for directly chemically modifying target nucleobases in the genome and transcriptome, and their potential for manipulating the composition and expression of target genes.

Cytosine deamination

Base editing tools were first created by fusing cytidine deaminase enzymes to dCas9.[13,14] These deaminases recognize cytosine in a ssDNA context and deaminate the nucleobase to form uracil, which has the base pairing properties of thymine (**Figure 1.1**). Due to the substrate requirement of ssDNA, the deaminase activity of the tethered enzyme is confined to only the single-stranded portion of the dCas9:gRNA:DNA R-loop. This strategy results in an editing "window" of only five to nine nucleotides (depending on the Cas9 variant used) within the protospacer.[13,15,16]

Several tools have since been created that link a cytidine deaminase to a DNA targeting enzyme such as dCas9[13,14,17,18] or dCas12a.[15] Though each tool may have slight variations, they each use the Cas protein to bring the cytidine deaminase to a target locus and convert cytosines to uracils. The final outcome following uracil incorporation, however, depends on the tool being used. Mutagenesis tools such as Targeted AID-mediated mutagenesis (TAM)[19] and CRISPR-X[18] rely on the cell's efficient excision of uracil from genomic DNA by the base excision repair pathway (BER). Excision of uracil by uracil-DNA glycosylase (UDG) creates an abasic site which can be permanently converted to any of the four canonical DNA bases through an as-yet-unknown mechanism that potentially involves error-prone DNA polymerases (**Figure 1.1**).[20]

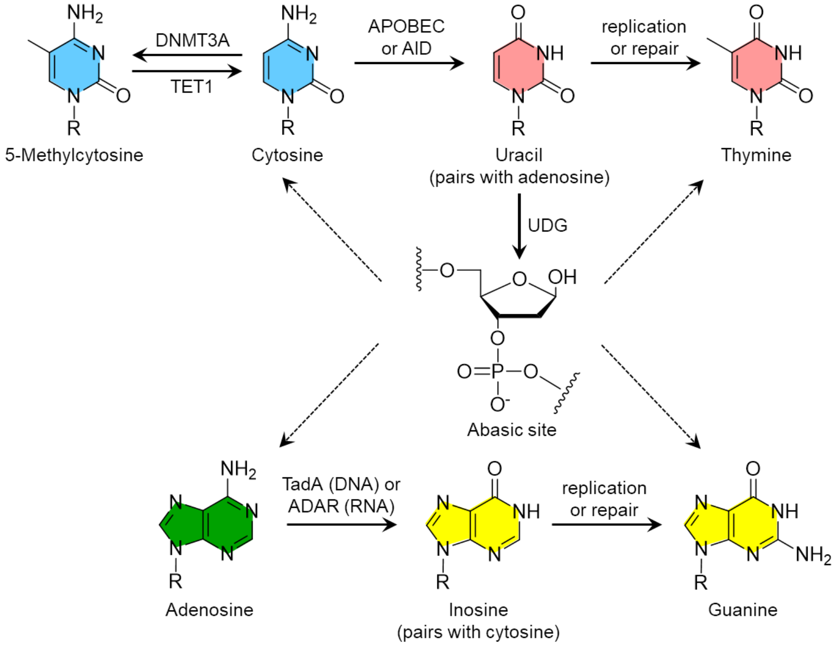


Figure 1.1: Nucleobase chemistries facilitated by current base editors

Cytosines can be methylated to 5-methylcytosines by DNMT3A-derived base editors. Likewise, 5-methylcytosines can be demethylated by TET1-containing base editors. These two types of tools can be used to control epigenetic patterns. Cytosine and adenine can be deaminated to uracil and inosine, respectively. These chemical transformations are catalyzed by base editors comprised of APOBEC or AID in the case of uracil, and TadA or ADAR in the case of inosine in DNA or RNA, respectively. Uracil is further converted to thymine and inosine is further converted to guanine by cellular replication or repair processes due to the new hydrogen-bonding properties of these base editing intermediates. Uracil is efficiently excised by the base excision repair protein UDG to form an abasic site, which is mutagenized to all four canonical DNA bases (dashed lines). The details of this process are still relatively unknown and the outcome is largely uncontrollable and unpredictable.

A particularly intriguing feature of these mutagenesis tools is their ability to generate genetically-encoded libraries. Within living cells, a library of gRNAs can be used to target these tools across an entire gene to introduce mutations. When combined with an appropriate selection or screen, these libraries can be used to identify mutations that give rise to specific phenotypes. Both TAM and CRISPR-X have been used to identify protein variants that give rise to chemotherapeutic resistance.[18,19] The ability to rapidly identify mutations that cause drug resistance (or another specific phenotype of interest) in a cellular context is a powerful tool that holds great promise in the areas of personalized medicine and reverse genetics.

Cytidine deaminase-derived base editing tools such as BE1-4[13,21] and Target-AID[14] can also be used to create predictable C•G to T•A mutations by manipulating cellular DNA repair pathways. Specifically, the bacteriophage protein uracil glycosylase inhibitor (UGI) can be physically linked[22] to or co-expressed[23] with the base editor. UGI protects the uracil intermediate from excision by reversibly binding to UDG[24] and significantly decreases C•G to non-T•A mutation rates.[21] Alternatively, the uracil intermediate can be avoided entirely by directly deaminating a methylated cytosine, which results in a direct conversion to thymine.[25] To manipulate DNA repair pathways, the catalytic activity of a single endonuclease domain in dCas9 can be restored to create a Cas9 nickase (Cas9n).[26] During base editing, Cas9n creates a nick in the DNA strand opposite the uracil. This nick marks the strand as "newly synthesized", resulting in the mismatch repair pathway (MMR) repairing the U•G mismatch using the uracil-containing strand as a template.[27] By manipulating the MMR pathway in this way, the guanine of the original C•G base pair is replaced with an adenine, solidifying the desired edit before uracil can be excised. These features can be combined (as in BE3, BE4, and Target-AID) to maximize C•G to T•A base editing efficiency.

The ability to precisely introduce single or multiple uracil lesions at a pre-defined location of the genome of living cells makes base editors valuable tools for both therapeutic and research purposes. Beyond using cytidine deaminase base editors to correct disease-causing point mutations and create genetically-encoded libraries, these tools have also unearthed surprising new properties of uracil repair. For example, it was discovered that single G•U lesions are repaired through a different, more error-prone mechanism than multiple clustered G•U lesions.[21] Though the mechanism of this differential processing is still not fully understood, its elucidation will likely be aided by using cytidine deaminase base editors as precision DNA damaging tools. Likewise, base editors capable of installing other types of DNA damage will help further our understanding of other mechanisms of DNA repair.

Cytosine methylation and 5-methylcytosine oxidative demethylation

Methylation of cytosines at the 5 position results in 5-methylcytosine (5mC), a naturally occurring modified DNA base (**Figure 1.1**). The presence of 5mC in promoter regions (where it occurs mostly at CpG sites) has a significant influence on gene expression at the transcriptional level. As a general rule, hypermethylation of these regions results in gene silencing.[28,29] The natural methylation of cytosines and demethylation of 5mC's are facilitated by DNA methyltransferase (DNMT) and ten-eleven translocation methylcytosine dioxygenase (TET) enzymes, respectively (**Figure 1.1**). Modulation of 5mC levels is particularly essential during development as shifting methylation levels control cell differentiation.[30] Inappropriate silencing due to hypermethylation can have devastating effects, as observed in several cancers and fragile X syndrome.[31] Therefore, tools that manipulate methylation levels at target regions are greatly beneficial for studying and developing treatments for a variety of diseases.

Two types of base editors have been created that allow researchers to modulate 5mC levels in target regions of the genome.[32–39] A DNMT3A-dCas9 fusion protein was developed as a tool to increase methylation levels. DNMT3A recognizes cytosines in dsDNA and uses S-Adenosyl methionine (SAM) as a methyl donor to create 5mC (**Figure 1.1**). To create a tool that decreases methylation levels, the enzymatic domain of TET1 was fused to dCas9. TET1 recognizes 5mC's and uses α-ketoglutarate to oxidize the methyl group to form 5-hydroxymethylcytosine as the first step of cytosine demethylation. 5-hydroxymethylcytosine is then further oxidized and subsequently excised by thymine DNA glycosylase to ultimately yield unmethylated cytosine (**Figure 1.1**).[40] As the natural substrates of these enzymes are dsDNA, the activity of these epigenome editors is not constrained within the ssDNA region of the R-loop. This allows the enzymes to access and modify a large window of nucleobases surrounding the binding site of the editor. This window is further widened due to the tight coiling of chromatin, which can allow the enzyme to access genomic DNA that is potentially thousands of base pairs away from its target site, but spatially very close.

These epigenome editors are quite valuable for studying the effects of methylation in untranslated regions (UTR) of the genome. For example, they have been instrumental in understanding methylation effects in fragile X syndrome.[41] This genetic disease is caused by a CGG trinucleotide repeat expansion in the 5' UTR of the *FMR1* gene, which codes for fragile X mental retardation protein (FMRP).[42] Hypermethylation of this UTR leads to silencing of *FMR1* and loss of FMRP expression. By targeting the dCas9-TET1 editor to this region, demethylation was observed and FMRP expression was recovered, leading to alleviation of the phenotype in post-mitotic neurons.

Adenine deamination

Cytosine is unique among the DNA bases as it is modified by several naturally occurring enzymes. While this facilitated the creation of the first base editors, the development of base editors that modify DNA bases other than cytosine is complicated by the lack of natural enzymes that can be repurposed to perform this chemistry. Fortunately, RNA nucleobases are extensively post-transcriptionally modified by natural enzymes.[43] By evolving one of these RNA-modifying enzymes to accept ssDNA as a substrate, an adenine base editor (ABE) that deaminates target adenines to inosines was engineered.[17] Adenosine deamination substitutes the amino group for a keto group and alters the base pairing properties of the nucleobase to match those of guanine.[44] As such, ABE catalyzes an overall A•T to G•C edit in genomic DNA at a locus programmed by the gRNA (**Figure 1.1**).

Intriguingly, A•T to G•C editing with ABE exhibits far lower rates of random mutagenesis than uracil-derived base editors, suggesting significant differences in cellular excision efficiencies of inosine (by the DNA-3-methyladenine glycosylase enzyme, AAG, and homologs)[45,46] as compared to uracil. As spontaneous adenosine deamination is 50-times slower than cytosine deamination,[47] the cellular repair machinery for uracil must be more efficient than that of inosine to preserve the integrity of the genome. Consequently, ABE-installed inosines are excised less efficiently than uracils, resulting in a more consistent mutagenic outcome for ABE. This highlights another interesting feature of DNA repair discovered through the use of base editors, and suggests that less common types of DNA damage can be used to create genome editing tools with more predictable outcomes.

While making modifications to DNA nucleobases can create a permanent mutation or modification in a cell, transiently mutating or modifying the transcriptome may be desirable in certain situations. RNA Base editing can be achieved using SNAP-ADAR,[48] λN-ADAR,[49] and RNA Editing for Programmable A to I Replacement (REPAIR).[50] All three of these technologies rely on a fusion complex of the adenosine deaminase acting on RNA 2 (ADAR2) enzyme with a gRNA molecule (SNAP-ADAR and λN-ADAR) or the dCas13b, which uses a gRNA to bind to target RNA sequences. The gRNA is designed to base pair with a target mRNA sequence and form a C•A mismatch (with the adenosine on the mRNA strand). ADAR2 deaminates the mismatched adenosine to inosine more efficiently[51] than any well-matched adenosines in the protospacer, providing single nucleotide resolution. The resulting inosine exhibits the base pairing properties of guanine during translation, and thus RNA base editing can be used for the transient expression of mutant proteins.

Along with transient base editing, SNAP-ADAR, λN-ADAR, and REPAIR open up the opportunity for better understanding the roles of RNA modifications. RNA modifications occur across all domains of life and affect the activity, localization, and stability of RNA. Over 100 types of modifications have been identified, and they have been found to exist in all types of RNAs.[43] Tools that allow researchers to install these modifications throughout the transcriptome will significantly aid in illuminating their functions and mechanisms.

Conclusions and Future Outlook

Most importantly, base editing has been shown to work in a variety of in vivo contexts. Successful C•G to T•A base editing has been accomplished in plants, including rice, wheat, maize, and tomato,[52–54] as well as a variety of animals including silkworms,[55] zebrafish,[56,57] mice,[56,58] and human embryos. ABE has been used in rats,[59] mice,[60] rice,[61–63] and wheat.[61] Furthermore, base editing has been performed in post-mitotic sensory cells, showing that this technology does not require cellular replication.[64] These experiments show that base editing is compatible with a large variety of cell types and organisms, making it valuable in many different areas of research.

ABE and cytosine deamination base editors can be used to study protein mutants and disease-associated intronic and non-coding mutations in an endogenous manner. Current strategies to study protein mutants involve knocking-out or silencing the corresponding gene, followed by overexpressing a variant to observe phenotypic effects. This method often results in protein expression levels that are significantly different from the endogenous system. Additionally, this strategy is difficult to apply to the study of point mutations that occur in intronic and non-coding regions of the genome. While using wtCas9 and a donor template can be used to edit endogenous genomic loci,[65] base editors can introduce point mutations with higher efficiency and fewer genome editing byproducts than DSB-reliant methods. This allows for the more rapid study of multiple variants in parallel, and presents the potential to efficiently introduce mutations at different sites throughout the genome in a given cell. Furthermore, base editors can be used to knock out genes of interest by introducing early stop codons via a method termed CRISPR-STOP.[66] This method has advantages over traditional, DSB-reliant methods for gene knock out, including less cytotoxic genome editing intermediates and more predictable genome editing outcomes. Finally, base editors can be used to mutate conserved splice acceptor sites within introns, in effect facilitating exon skipping, in a method called

CRISPR-SKIP.[67] This method has the potential to modulate expression of different protein isoforms and skip mutation-containing exons such as in Duchenne muscular dystrophy and Huntingon's disease.[68]

Base editors provide a potentially less cytotoxic and more efficient method of point mutation introduction than DSB-reliant methods. Nevertheless, this technology faces several limitations as it continues to grow. For example, the current deaminating base editors are only able to facilitate C•G to T•A and A•T to G•C base pair conversions. To be more universally applicable, new base editors must be developed that catalyze additional point mutation changes. Effective intermediates for future base editors will likely be found in less common naturally occurring DNA modifications, as demonstrated by the more predicable editing outcomes observed with base editing using inosine intermediates compared to uracil.[17] Off-target effects are a large concern with genome editing agents in general, and indeed off-target editing is observed with current base editors.[69] The majority of off-target base editing sites overlap with those of wild-type Cas9. Consequently, base editors derived from Cas9 variants with increased specificity have been shown to alleviate unwanted editing at these sites.[56] However, off-target sites unique to BE3 have been identified, indicating that creative new solutions are needed in the future to increase base editing specificity. The base editors described here allow for the direct chemical modification of target nucleobases in a programmable manner.

These tools hold tremendous potential for regulating the central dogma. Cytosine and adenosine deamination DNA base editors facilitate the efficient installation of point mutations throughout the genome, allowing researchers to alter the identity or expression levels of proteins and functional RNAs. Cytosine methylation and demethylation DNA base editors allow researchers to modulate gene transcription levels. Adenosine deamination RNA base editors open the door for the targeted modification of RNAs which can ultimately modulate all aspects of transcription and translation. As base editing technologies continue to develop, so will our ability to manipulate the contents of the cell.

Acknowledgements

Chapter 1 has been reproduced, in full, with permission, from: **Ranzau, B. L.** & Komor, A. C. Genome, Epigenome, and Transcriptome Editing via Chemical Modification of Nucleobases in Living Cells. *Biochemistry* **58**, 330–335 (2019). The dissertation author was the primary author on all reprinted materials.

Chapter 2 Cas9-independent RNA editing by ABE variants

Introduction

ABE7.10 was developed by linking the tRNA-modifying enzyme TadA to dCas9 and performing directed evolution to improve the DNA-editing ability of TadA. When the wild-type TadA-nCas9 base editor (ABE0.1) was tested at six different genomic sites in HEK293T cells, A-to-G editing was not observed. Only after the first round of evolution enriched the D108N mutation, and this mutation was included in ABE1.1, was base editing observed in mammalian cells. This new editing activity was further increased as more mutations were enriched from each round of selection.

Despite the focus on DNA-editing during ABE7.10 development, the evolved base editor still displays high levels of RNA-editing. When ABE7.10 is over expressed in mammalian cells, high levels of RNA-editing have been observed transcriptome-wide, and most highly in UACG motifs.[70] This activity likely stems from the natural sequence recognition properties of TadA, which uses the UACG motif to recognize its target tRNA.[71,72]

The development of new base editors has been slowed by the lack of DNA-modifying enzymes that have been found in nature. The evolution of TadA to accept DNA as a substrate shows that other tRNA-modifying enzymes could be evolved to accept DNA and used to introduce a wide variety of intermediates into DNA for base editing. Understanding the changes to TadA at a molecular level could provide insight into how other tRNA-modifying enzymes could be mutated so that they can be used as base editors. In this chapter, we explore using sequence conservation as a tool for predicting mutations that may boost the DNA-editing ability of RNA-modifying enzymes. We then evaluate this prediction method by measuring the RNA-editing activity of ABE0.1 and several variants.

Measuring gRNA-independent RNA-editing by ABE variants

An analysis of sequence entropy (amino acid conservation) was performed using the wild-type TadA sequence and other closely related homologs. This analysis aligned the primary sequences of 35 TadA-related proteins to compare the amino acids that occur at each position across different species. A sequence entropy score was generated for each residue in TadA, with a high score indicating a position that is often mutated, and a low score indicating high conservation of a specific amino acid. More information on this sequence alignment can be found in our publication.[73]

Each mutated position in ABE7.10 was sorted based on their entropy score, with most of these mutations having a score >50%, and only the L84F mutation occurring at a highly conserved position.[73] This observation suggested that making mutations to positions that are not well conserved may enable a tRNA-modifying enzyme to accept DNA as a substrate. However, all of the proteins in this analysis are natural RNA-modifying enzymes, so it is likely that entropy score may be useful for predicting how mutations impact RNA editing. To test this hypothesis, we aimed to analyze the RNA-editing ability of ABE0.1, along with representative mutations from both a highly conserved position and a non-conserved position. As the mutation that enabled DNA-editing by TadA the D108N mutation was selected as the non-conserved position, and the L84F mutation was selected as the only mutation to occur at a well-conserved position. The RNA-editing activity of ABE7.10 was also analyzed to see how the many mutations to non-conserved residues impacted activity.

We first tested the ability of ABE0.1 (as both monomeric and dimeric wtTadA fused to nCas9) to introduce A-to-I edits in mRNA in a gRNA-independent manner. HEK293T cells were transfected with constructs encoding monomeric ABE0.1, dimeric ABE0.1, or heterodimeric ABE7.10 (wtTadA-TadA7.10-nCas9), mRNA was extracted after 36 h, and high-throughput sequencing was used to quantify A-to-I editing at six different sites throughout the transcriptome that had previously been shown to be edited by ABE7.10 in a gRNA-independent manner.[70] We observed >50% A-to-I RNA-editing efficiencies at all six sites by both wild-type constructs (**Figure 2.1**). Moreover, consistent with the recent report comparing the kinetics of ABEs on RNA substrates,[74] the RNA-editing activity of dimeric ABE0.1 was on average 21% higher than ABE7.10, highlighting the remarkable shift in substrate preference of the wtTadA enzyme due to the many mutations that were found through directed evolution for ABE7.10.

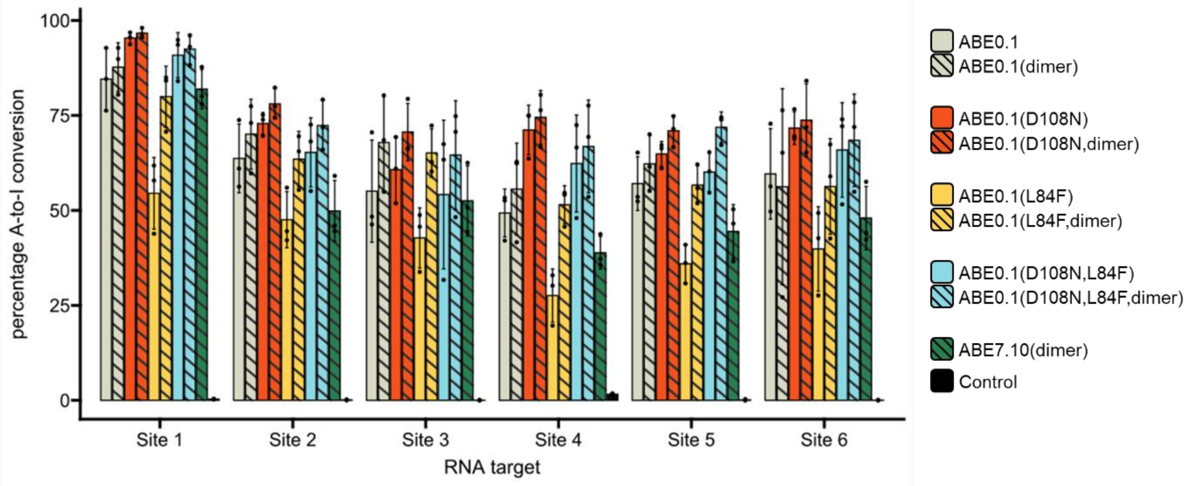


Figure 2.1 : A-to-I base editing efficiencies of various ABE mutants at RNA off-target sites

The indicated base editor was transfected into HEK293T cells. Cells were lysed and RNA was harvested after 37 hours, and target site amplicons were sequenced with high-throughput sequencing. Values and error bars reflect the mean and standard deviation of three independent biological replicates performed on different days.

The D108N mutation lead to a modest 11.2% (range 5.7–21.8%) increase in the A-to-I RNA-editing activity of ABE1.1 compared to ABE0.1. Moreover, the L84F mutation leads to a 25% (range 19.8–31.7%) or almost 1.7-fold decrease in RNA-editing efficiency of the enzyme compared to ABE0.1 across the six different RNA sites that were analyzed. Interestingly, the loss of activity due to the L84F mutation could be rescued either by adding a wtTadA subunit to create a heterodimeric base editor, or by combining the L84F and D108N mutations in the same TadA subunit. In the case of the ABE0.1(L84F, heterodimer), the increase in RNA editing is likely due to the addition of the wtTadA subunit, which is capable of efficient RNA editing on its own. However, the combination of the L84F and D108N mutations is particularly interesting as it displays how mutations can have non-additive effects on enzyme function. When comparing the editing efficiency of the ABE1.1(L84F) variant with the editing efficiencies of ABE0.1(L84F) and ABE1.1, the combination variant displays activity more similar to ABE1.1. This suggests that the D108N mutation in this context appears to change the TadA structure to accommodate the L84F mutation without any negative effects on RNA-editing activity.

Surprisingly, the RNA-editing activity of ABE7.10 was 1.2-fold lower than the that of ABE0.1. This suggests that the 12 other mutations, alongside D108N and L84F, cause an overall decrease in RNA-editing activity, while increasing DNA-editing activity. Since all of these other mutations occurred at non-conserved residue positions, we cannot conclude that the entropy score is a reliable predictor of how different mutations may affect either RNA- or DNA-editing by RNA-modifying enzymes. Instead, we observed that each mutation likely changes the landscape of the protein, making the prediction process more difficult with each additional mutation.

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. ABE variants were cloned via USER cloning following New England Biolabs (NEB) protocols, by introducing the mutation of interest via primsers that annealed to the *TadA* gene in the ABE-P2A-EGFP plasmid (Addgene plasmid #112101).

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 transfection, DNA mixtures were prepared with: 750 ng of base editor plasmid and 250 ng of a non-targeting gRNA plasmid. These 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.

RNA extraction and reverse transcription

36 hours after transfection, cells were lysed with 300ul RNA lysis buffer and RNA was extracted with either a Zymo Quick-RNA Miniprep kit (Zymo Research # R1054) or Qiagen RNeasy mini kit (Qiagen # 74004) following the manufacturers' instructions. RNA was reverse transcribed to produce cDNA using SuperScript III First-Strand Synthesis System (Invitrogen #18080-051) following the manufacturer's instructions.

High-throughput amplicon sequencing of reverse transcribed cDNA

RNA 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. PCRs followed the manufacturer's protocol, using 1 µl of cDNA 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[75]. 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-to-I edits were calculated using the nucleotide frequency at the target site, by dividing the number of G (which replaces I in DNA) 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.

Acknowledgements

Chapter 2 has been adapted, in part, with permission, from Rallapalli, K. L., **Ranzau, B. L.**, Ganapathy, K. R., Paesani, F. & Komor, A. C. Combined Theoretical, Bioinformatic, and Biochemical Analyses of RNA Editing by Adenine Base Editors. *The CRISPR Journal* **5**, 294–310 (2022). The dissertation author was the primary author on all reprinted materials.

Chapter 3 DNA editing by ABE0.1 in TACG motifs

Introduction

Base editing is a genome editing technique that enables the targeted introduction of single nucleotide variants (SNVs) without using double strand breaks (DSBs)[76,77]. 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 3.1A**)[78]. 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[79], which then modifies the target nucleotide(s) within the protospacer (**Figure 3.1A**). 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 3.1A**).

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[76,80–82]. ABEs were therefore developed by evolving a natural tRNA-modifying enzyme (*E. coli* TadA) to accept DNA as a substrate (**Figure 3.1B-D**)[77]. 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

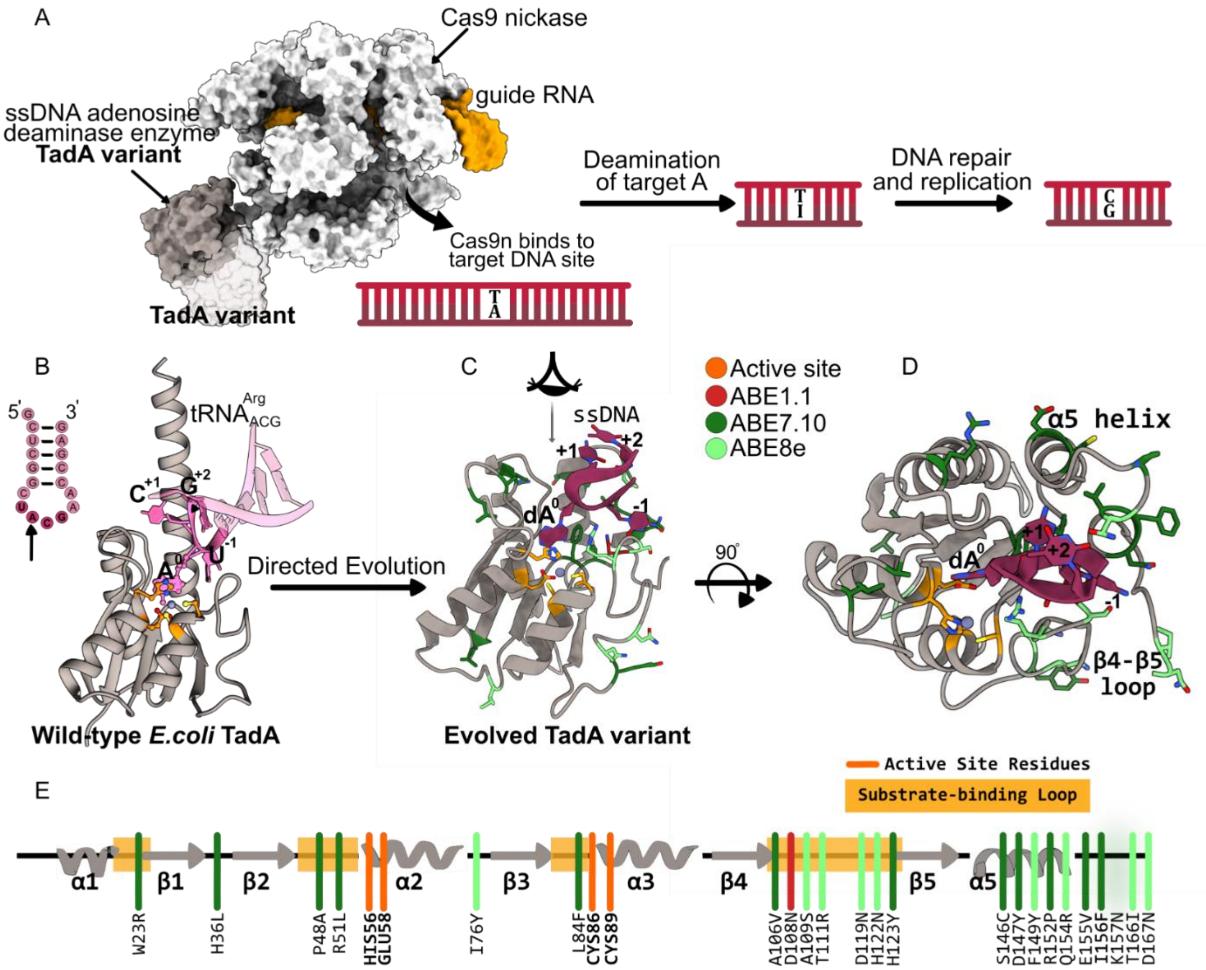


Figure 3.1: Overview of Adenine Base Editor (ABE) mechanism and evolution

(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)**.

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 3.1C**). Additional rounds of directed evolution have since been undertaken to engineer ABE8 (**Figure 3.1C-E**) and ABE9 editors, which perform base editing with even higher efficiencies and faster kinetics[83–85].

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 3.1C-D**), which is sufficient and imperative for imparting TadA with detectable DNA editing activity[73,86].

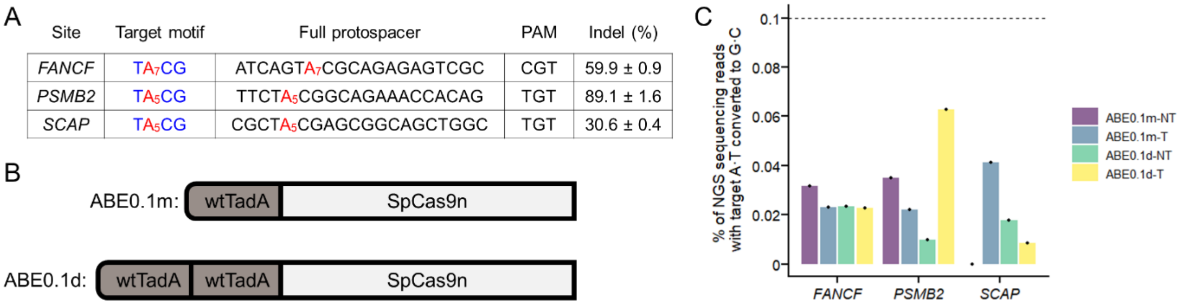
In nature, tRNA editing enzymes have evolved to recognize specific RNA structures or sequences[87]. As a result, they modify only a specific nucleotide on one or multiple tRNAs within the cell. Specifically, *E. coli* TadA recognizes the UACG motif within the anticodon loop of the $\text{tRNA}^{\text{Arg}}_{\text{ACG}}$ and deaminates the indicated adenosine, which is in the wobble position **(Figure 3.1B**)[71]. 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[70,73,88,89]. 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 $\text{tRNA}^{\text{Arg}}_{\text{ACG}}$ from other RNAs in the cell[73].

While initial experiments concluded that ABE0.1 was incapable of editing DNA, none of the tested protospacers contained a TACG target (**Figure 3.2A**)[77]. 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[73]. 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 3.3A**)[77]. We therefore identified three genomic loci containing this motif and designed protospacers that place the target A in position 5 or 7 (**Figure 3.2A**). 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 3.2A**). 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 (**Figure 3.2B**)[77]. However, later generations of ABEs utilized dimeric TadA fused to Cas9n, as these demonstrated increased editing efficiency[77]. Here, we tested both monomeric and dimeric ABE0.1 constructs



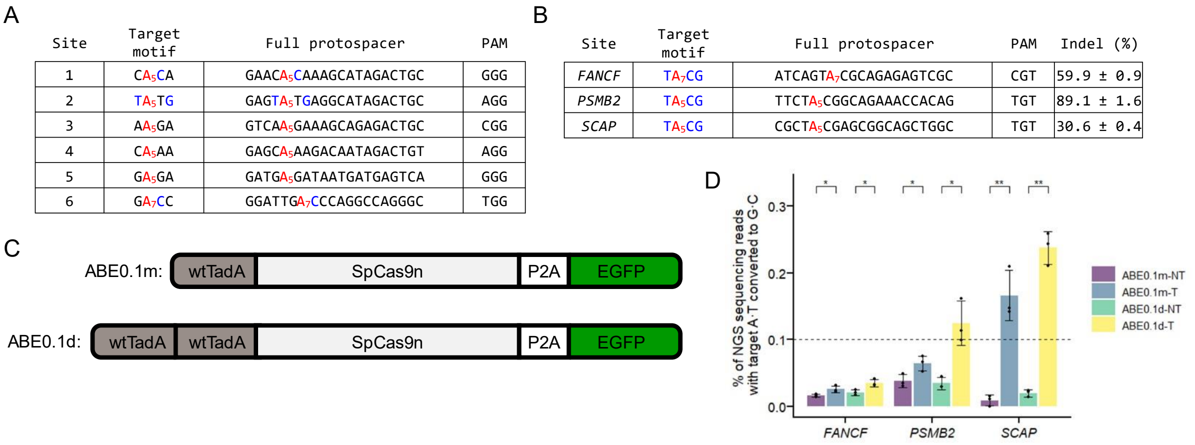
Site	Target motif	Full protospacer	PAM	Indel (%)
FANCF	TA$_7$CG	ATCAGTA$_7$CGCAGAGAGTCGC	CGT	59.9 ± 0.9
PSMB2	TA$_5$CG	TTCTA$_5$CGGCAGAAACCACAG	TGT	89.1 ± 1.6
SCAP	TA$_5$CG	CGCTA$_5$CGAGCGGCAGCTGGC	TGT	30.6 ± 0.4

Figure 3.2: ABE0.1 editing is below the limit of detection in unenriched cells

(A) Protospacers and PAM sequences of genomic loci at which ABE0.1 base editing was evaluated in this work, with target motifs indicated. Nucleotides in blue match the UACG motif that wtTadA targets in its native tRNA^{Arg}$_{ACG}$ target. **(B)** Architectures of ABE0.1 constructs used in **(C)**. **(C)** A•T to G•C base editing efficiencies by ABE0.1m and ABE0.1d at the three genomic sites from **(A)**. HEK293T cells were transfected with plasmids encoding ABE0.1m or ABE0.1d and gRNA. 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 in **(A)** represent the mean and standard deviation for n = 2 biological replicates. The results of one replicate are marked in **(C)**.

(ABE0.1m and ABE0.1d, respectively), as wtTadA is known to dimerize when editing its native $tRNA^{Arg}{}_{ACG}$ substrate, but the mechanism of DNA editing by evolved TadA enzymes is not currently known[71,72,90,91]. 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 3.1A**)[70,74,86,89]. These initial experiments did not show A•T to G•C editing levels above background (**Figure 3.2C**), 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 3.3C**), 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[92]. 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. 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 (**Figure 3.4B**). 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 3.3D**): 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 3.3D**). Editing above background levels was not observed at the *FANCF* site for any of the editors (**Figure 3.3D**), 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

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 3.3: 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 $\text{tRNA}^{\text{Arg}}_{\text{ACG}}$ 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 $\text{tRNA}^{\text{Arg}}_{\text{ACG}}$ 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: $p \geq 0.05$ not significant (ns), p = 0.01–0.05 significant (*), p = 0.001 – 0.01 very significant (**).

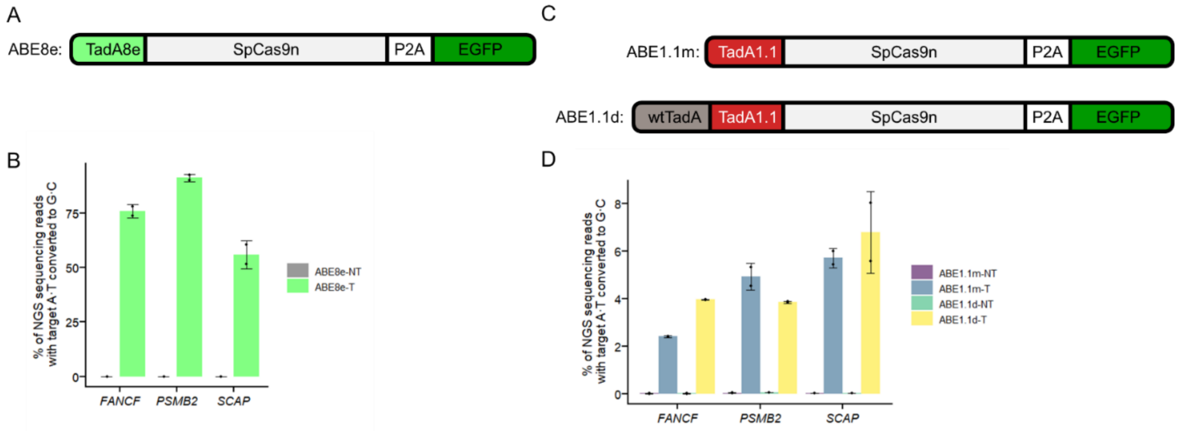


Figure 3.4: ABE1.1 and ABE8e editing at gDNA sites after enrichment.

(A) Architecture of the ABE8e construct used in **(B)**. **(B)** A•T to G•C base editing efficiencies by ABE8e at the three genomic sites from **Figure 2b**. **(C)** Architectures of ABE1.1 constructs used in **(D)**. **(D)** A•T to G•C base editing efficiencies by ABE1.1m and ABE1.1d at the three genomic sites from **Figure 2b**. **(B and D)** HEK293T cells were transfected with plasmids encoding the ABE variants indicated 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 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 **(B)** and **(D)**.

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[93,94]. 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[95,96]. 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 3.5A**)[82]. 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 3.5A**), 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[73]. 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 3.5A**), 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, **Figure 3.6A**)[97,98]. 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.

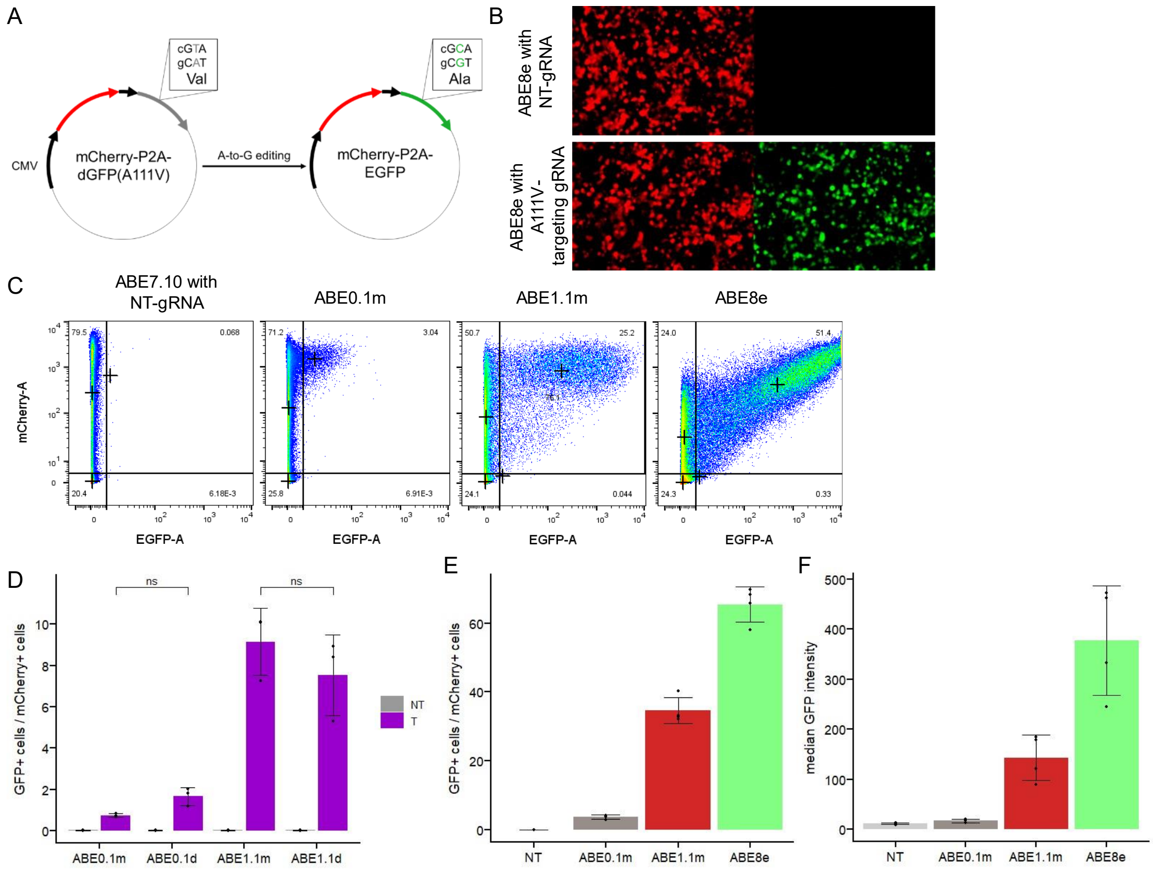


Figure 3.5: An EGFP reporter assay enhances editing detection by low-efficiency ABEs

(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).

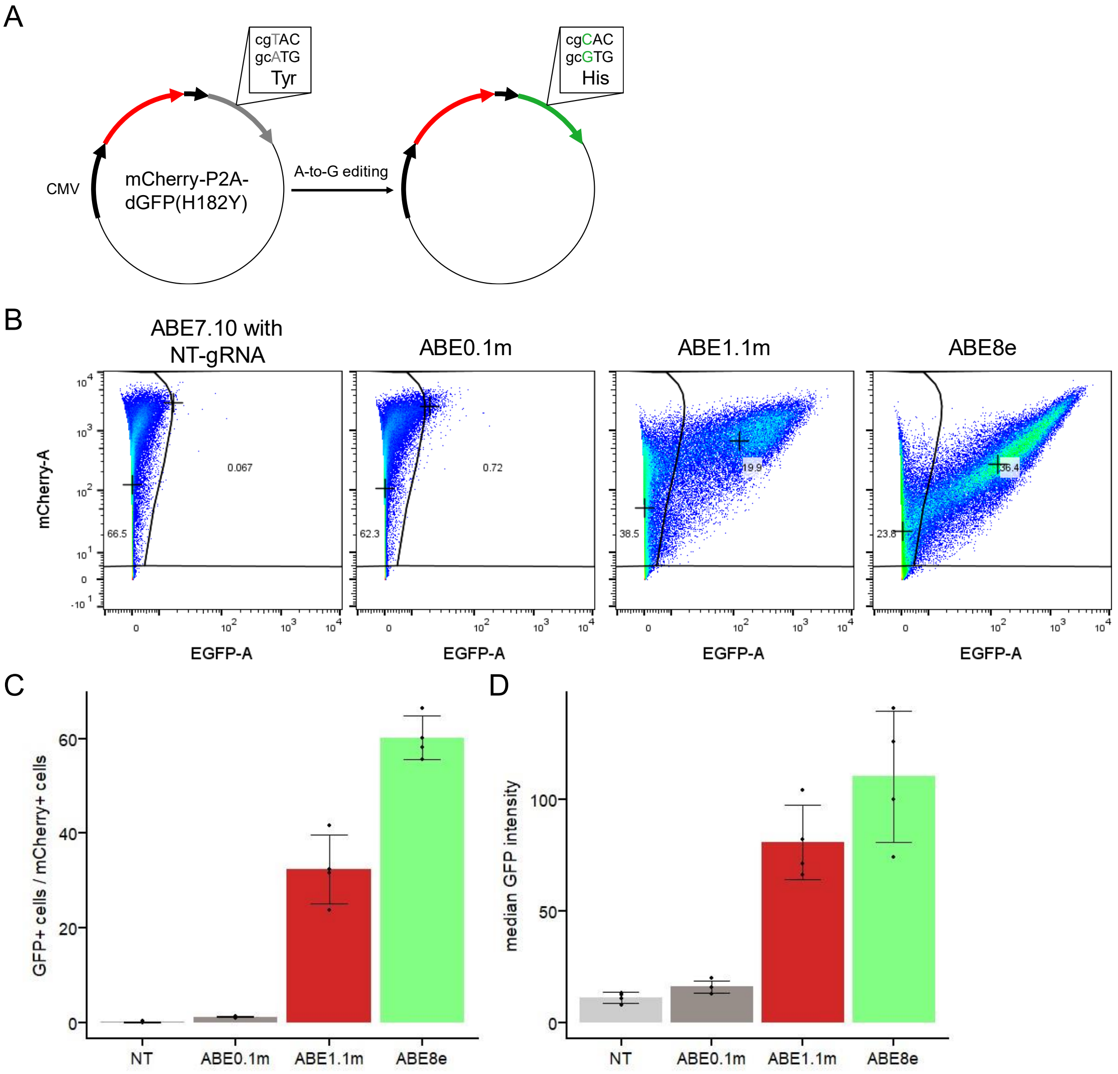


Figure 3.6: Detection of base editing by H182Y reporter

(A) Schematic overview of the H182Y reporter. **(B)** Representative flow cytometry graphs showing cells transfected with the H182Y reporter and the indicated ABE variant. mCherry intensity is shown on the y-axis, while EGFP intensity is shown on the x-axis. "+"'s indicate the median fluorescence intensities of mCherry and EGFP for each gated population. HEK293T cells were transfected with plasmids encoding the H182Y reporter, the indicated ABE variant, and a NT or targeting gRNA; and were analyzed by flow cytometry after 72 hours. **(C)** Shown are the percent of transfected cells (determined based on mCherry fluorescence) with EGFP fluorescence, using the gating strategy and ABE variants shown in **(B)**. **(D)** Shown are the median EGFP fluorescence intensity of mCherry, EGFP double positive cells shown in **(B)** and **(C)**. Values and errors in **(C-D)** represent the mean and standard deviation for n = 4 biological replicates. Each replicate is marked individually.

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 3.5B**) and analyzed by flow cytometry (**Figures 3.5C and 3.5E-F,** and **Figure 3.6B-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 (**Figure 3.6B**). 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 3.5E** and **Figure 3.6C**), 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 3.5C-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 3.5D**). 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 3.5D**). 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 (**Figure 3.4D**). 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 3.5F** and **Figure 3.6D**).

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 (**Figure 3.7A**). 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 (**Figure 3.7A**). 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, **Figure 3.7B**). 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, **Figure 3.7B**) using this analysis method, potentially due to a combination of the higher background signal and reduced voltage used to observe this reporter (**Figure 3.7B**). 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.

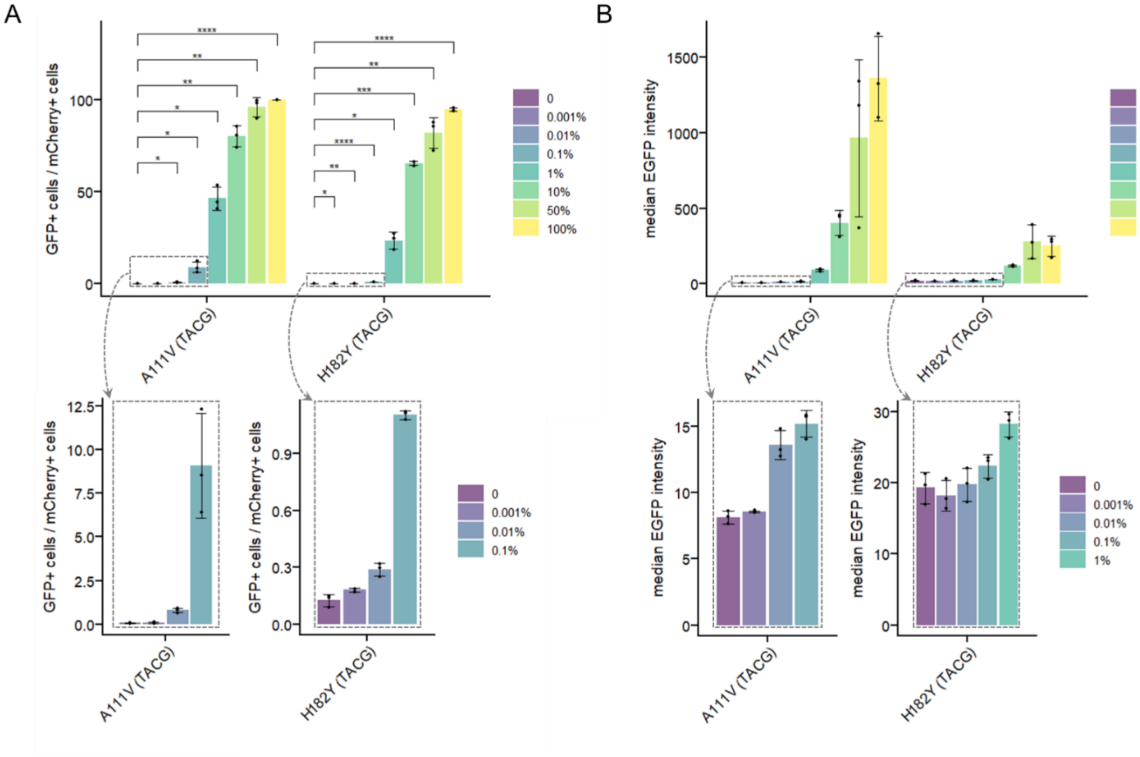


Figure 3.7: Flow cytometry can detect down to 0.01% of active reporter

Plasmid mixtures with either the A111V or H182Y reporter, and decreasing amounts of the equivalent wild-type EGFP reporter (from 100% to 0% wild-type) were prepared and transfected into HEK293T cells. After 72 hours, cells were analyzed by flow cytometry. **(A)** Shown are the percent of transfected cells (determined based on mCherry fluorescence) with EGFP fluorescence for each percentage of wild-type reporter plasmid. **(B)** Shown are the median EGFP fluorescence intensity of mCherry, EGFP double positive cells shown in **(A)**. Lower graphs in **(A-B)** show zoomed-in looks at the samples containing lower percentages of wild-type reporter plasmid. Values and errors in **(A-B)** represent the mean and standard deviation for n = 3 biological replicates. Each replicate is marked individually. Data were analyzed with one-tailed T-tests using the 0% wild-type reporter results as a reference and p-values are marked as following: p = 0.01–0.05 (*), p = 0.001 – 0.01 (**), p = 0.0001- 0.001 (***) p < 0.0001 (****).

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 3.5A**), and the H182Y reporter enables mutation of the C^{+1} as this resides at the wobble position of a Ser codon (**Figure 3.6A**).

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 **Figure 3.9A-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 (**Figure 3.9A**), suggesting that editing levels of these samples are very close to the limit of detection of the reporter.

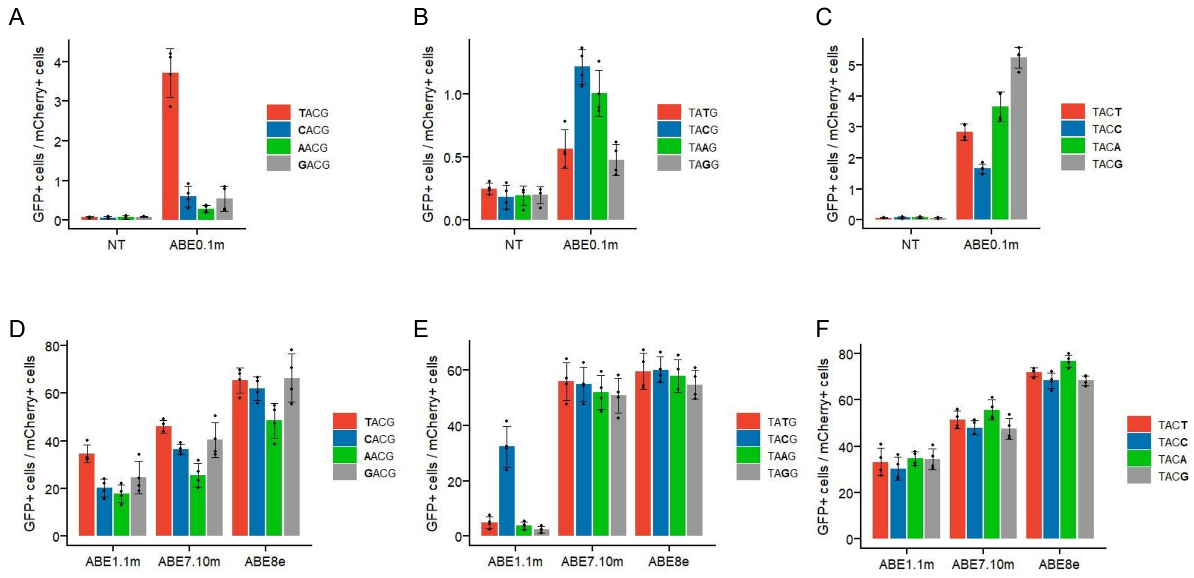


Figure 3.8: ABE0.1 shows strict sequence requirements for DNA base editing

(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 3.5C**) 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 3.5C**) for all samples indicated. Samples in **(D)** were transfected with A111V-derived reporters, with mutations at the -1 position.

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 3.8C**). 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).

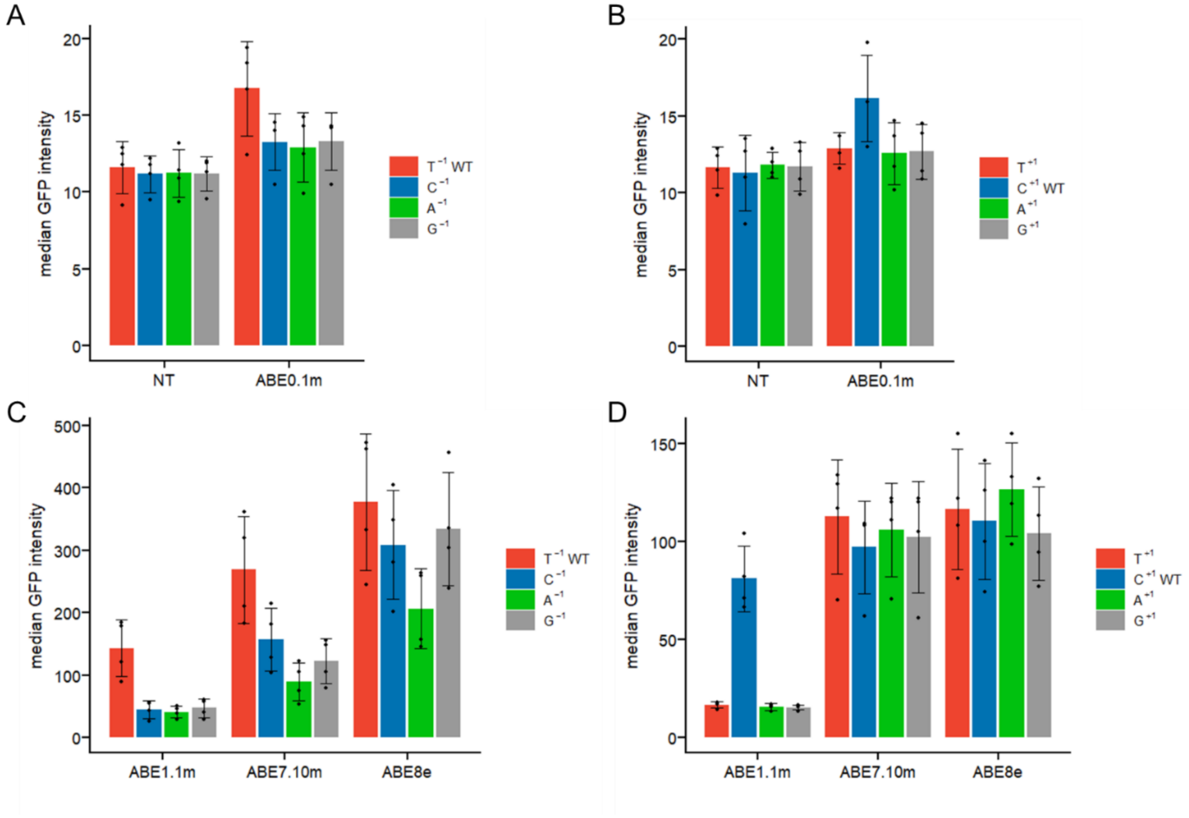


Figure 3.9: Median GFP intensity of reporters treated with ABE variants

Median GFP intensities are shown for the mCherry, EGFP double positive cells of the samples in **Figure 4A-B** and **Figure 4D-E**. HEK293T cells were transfected with plasmids encoding the fluorescent reporter indicated, the ABE variant indicated, and a NT or targeting gRNA. After 72 hours, cells were analyzed by flow cytometry. Samples in **(A)** and **(C)** were transfected with A111V-derived reporters, with mutations at the -1 position. Samples in **(B)** and **(D)** were transfected with H182Y-derived reporters, with mutations at the +1 position. Values and error bars represent the mean and standard deviation for n = 4 biological replicates. Each replicate is marked individually.

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[86]. 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[77]. Furthermore, reversion of this mutation back to wild-type caused the ABE7.10 construct to lose nearly all activity[86]. 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 3.8A** and **3.8D**). Further, the MFI of EGFP-positive cells treated with ABE1.1m were 8.6 ± 3.2-fold higher than those treated by ABE0.1m (**Figure 3.9A** and **3.9C**). 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 3.8D**). Analysis of the MFI of EGFP-positive cells of these samples revealed the same trend (**Figure 3.9C**). Interestingly, mutations at the G^{+2} position did not cause statistically significant changes in editing activity by ABE1.1m (**Figure 3.8F**).

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 3.8E**). Notably, this represents a 26.6 ± 6.7-fold increase when comparing ABE1.1m to ABE0.1m (**Figure 3.8B** and **3.8E**). 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 (**Figure 3.9B** and **3.9D**). 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 3.8E** and **Figure 3.9D**). 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 3.8E**).

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 3.3A**)[77]. 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^{-186}.

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[77,84]. 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 3.8D**). Similar to ABE1.1, mutations to the G^{+2} position did not significantly affect editing efficiencies (**Figure 3.8F**). 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**

3.8E). These same trends were also observed when analyzing the MFI of EGFP-positive cells (**Figure 3.9C-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 3.8D**). 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 (**Figure 3.9C-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.[77]

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 (**Figure 3.10A**). In light of this observation, we analyzed the EGFP MFI’s from our earlier gDNA editing experiment (**Figure 3.3**), 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 (**Figure 3.10B-C**). These data suggest that despite the significant improvement in editing efficiency with ABE8e, the accumulated mutations may have reduced its stability or expression.

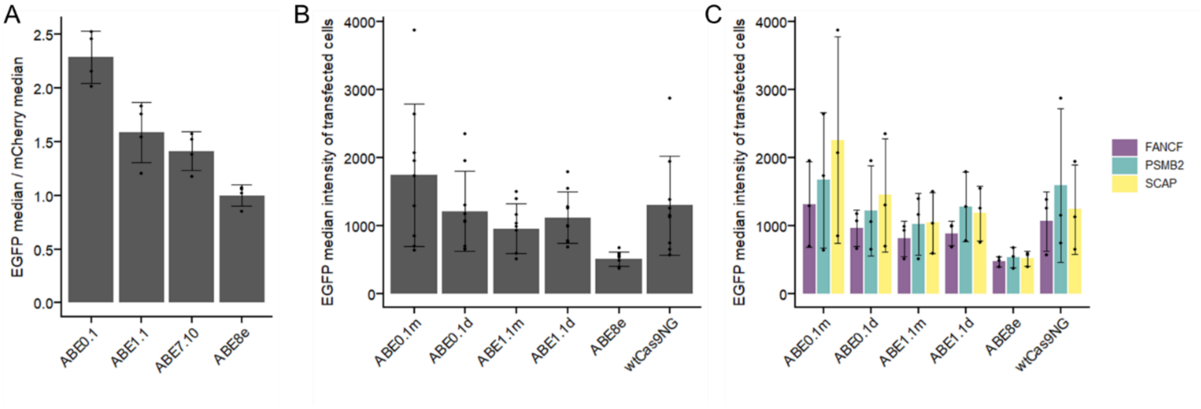


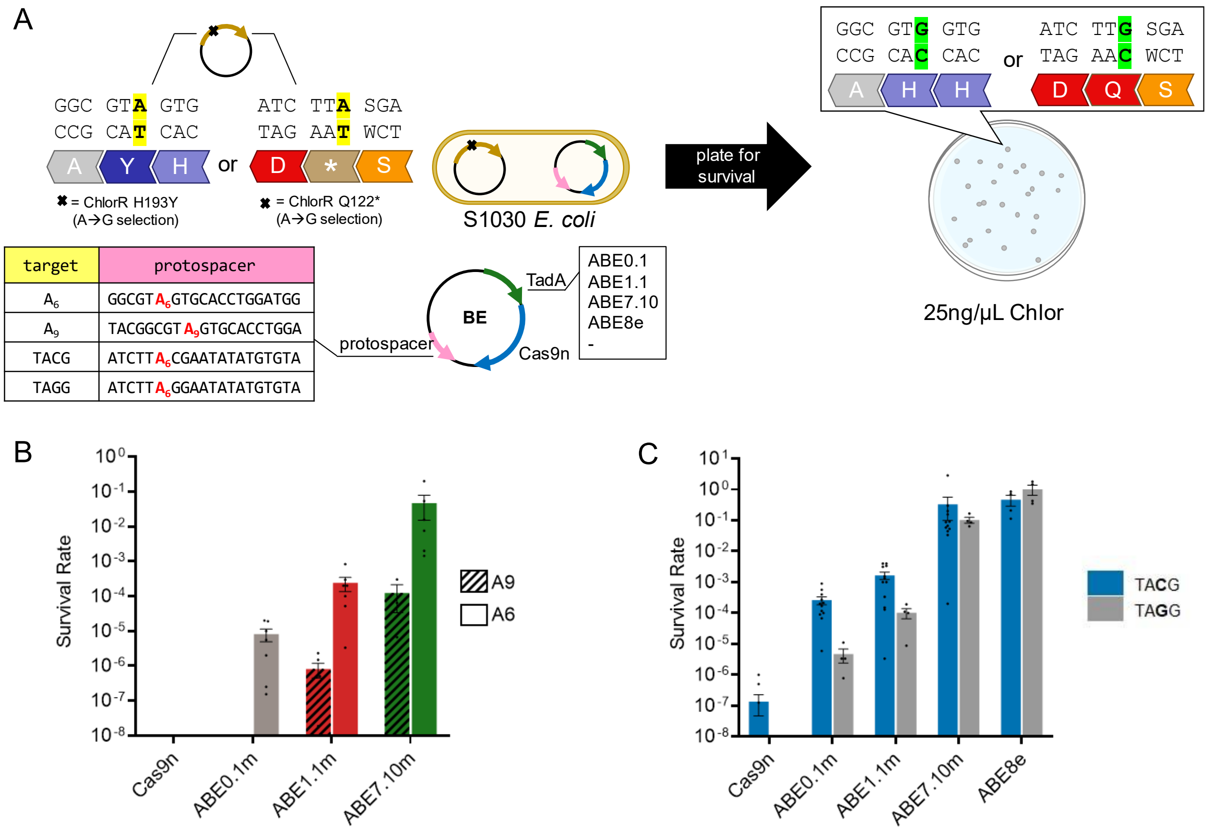
Figure 3.10: Intracellular ABE concentrations decrease with later generation ABEs

(A) HEK293T cells were transfected with mCherry-P2A-ABE-EGFP plasmids encoding the indicated ABE. After 72 hours, cells were analyzed by flow cytometry. Gates were drawn around transfected cells showing both mCherry and EGFP intensities, and MFI's were determined for this population. Shown are the EGFP MFIs for each ABE, normalized to the mCherry MFI. **(B)** The experiment from **Figure 3.3** was re-analyzed for EGFP MFI's. Using the sorting gates as shown in SI Figure 2, the EGFP MFI was determined for sorted cells transfected with each of the ABE-P2A-EGFP constructs. **(C)** The data from **(B)** is broken down by target site (see **Figure 3.3B** for sequences). Values and error bars represent the mean and standard deviation for n = 3 biological replicates in **(A)** and **(C)**, n = 9 biological replicates in **(B)**. Each replicate is marked individually.

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[77]. 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 T<u>A</u>GT motif in the inactive chloramphenicol resistance gene CmR H193Y (**Figure 3.11A**). 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[77]. 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 3.11A**)[99]. 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 3.11B**), consistent with what was observed during the original development of the ABEs[77]. We then used the PAM-relaxed Cas9n-NG variant to shift the protospacer by three bases and placed the target A at position 6[100]. We repeated the experiment and observed detectable activity by



target	protospacer
A_6	GGCGTA_6GTGCACCTGGATGG
A_9	TACGGCGTA_9GTGCACCTGGA
TACG	ATCTTA_6CGAATATATGTGTA
TAGG	ATCTTA_6GGAATATATGTGTA

Figure 3.11: 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 T<u>A</u>GT 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 T<u>A</u>CG (labeled as C^{+1} WT) and T<u>A</u>GG (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.

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 3.11B**). 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 3.11A**), and repeated the experiment with Cas9n, ABE0.1, ABE1.1, ABE7.10, and ABE8e (**Figure 3.11C**). 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 3.11C**). 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 (**Table 3.1**). In our previous simulation studies of ABEs, we used minimalistic models (specifically, we simulated only the TadA variant bound

Table 3.1: List of ABE variants and sequence contexts modelled and simulated in this study.

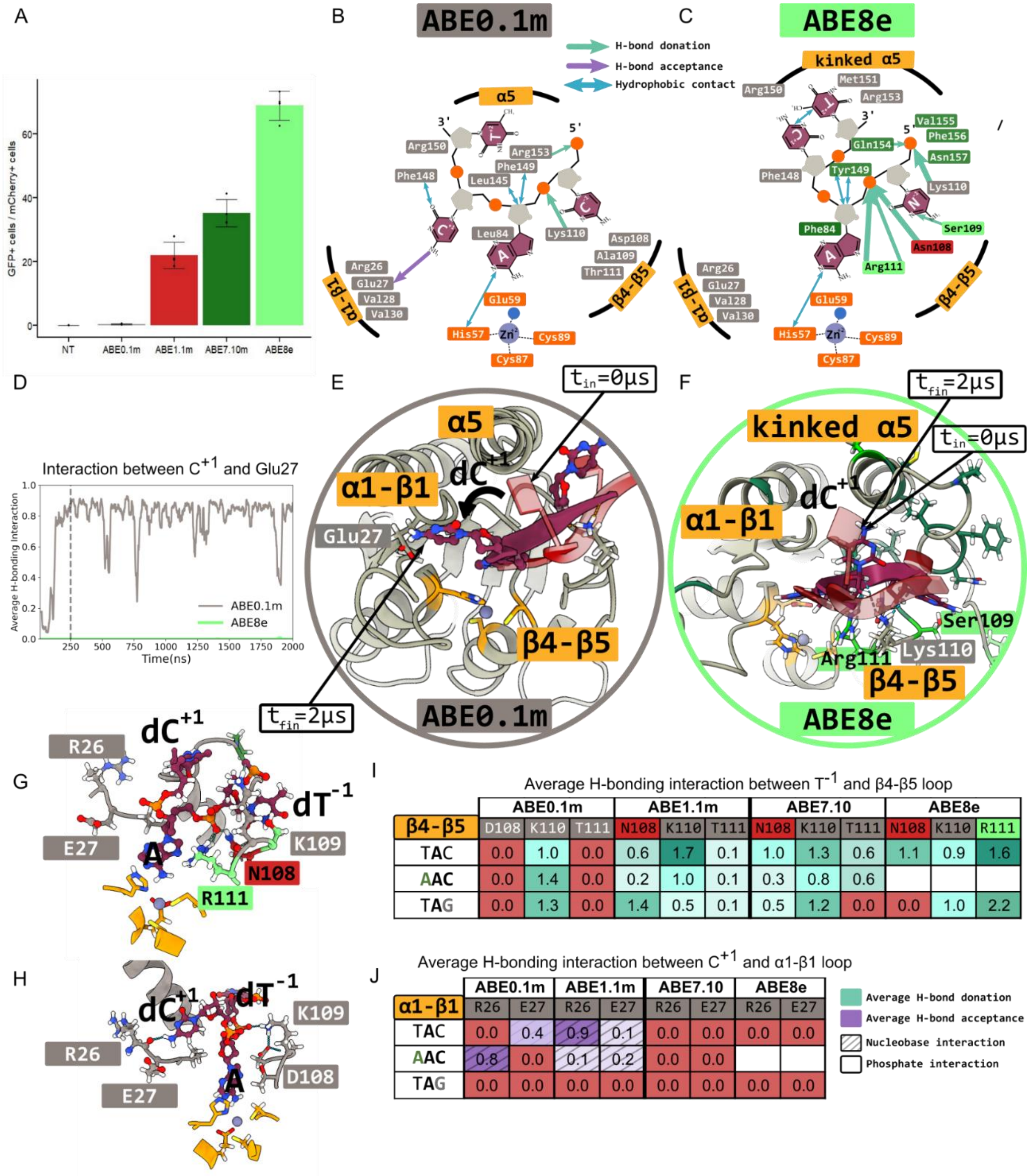
System		Number of atoms	Simulation Length (ns)
ABE variant	Target ssDNA sequence		
ABE0.1m	5′-GTTCCACTTT-3′	301933	2000
	5′-GTTCTACCTT-3′	301933	400
	5′-GTTCAACGTT-3′	301930	400
	5′-GTTCTAGGTT-3′	301936	400
ABE1.1m	5′-GTTCTACCTT-3′	301937	400
	5′-GTTCAACGTT-3′	286667	400
	5′-GTTCTAGGTT-3′	301940	400
ABE7.10	5′-GTTCTACCTT-3′	305071	400
	5′-GTTCAACGTT-3′	305068	400
	5′-GTTCTAGGTT-3′	336322	400
ABE8e	5′-GTTCCACTTT-3′	348601	2000
	5′-GTTCTACCTT-3′	313863	400
	5′-GTTCAACGTT-3′	317349	400
	5′-GTTCTAGGTT-3′	313857	400

to either its substrate ssDNA or tRNA)[73,86]. 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)[74]. 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 3.1A**). 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 3.12A**) 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 3.8A** and **3.8D**).

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

Figure 3.12: 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[74]. 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.



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

ssDNA strand, particularly the CAC target motif (**Figure 3.1D**)[74]. 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 3.12B-C**). In these "interaction maps" (**Figure 3.12B-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 3.12B-C**).

A comparison between the interactions maps of ABE0.1m (**Figure 3.12B**) and ABE8e (**Figure 3.12C**) 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 3.12B-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 3.12B-C**). This H-bond formed spontaneously during the course of the ABE0.1 simulation, and once formed remained stable throughout the trajectory (**Figure 3.12D**). 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 3.12E**). 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 3.1B**)[72]. 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 3.12F** and **3.1D**). 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 3.12B, E**) as well as in the wtTadA-RNA structure (**Figure 3.1B**)[72,77].

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 3.8**). Informed by the dynamics of the CAC systems which underwent conformational transitions within 200 ns (**Figure 3.12D**), 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 3.12I-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 3.12I**). 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 3.8A**), 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 3.12B-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 3.12G-H, J**). However, consistent with the experimental observations (**Figure 3.8**), 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 (**Figure 3.13**).

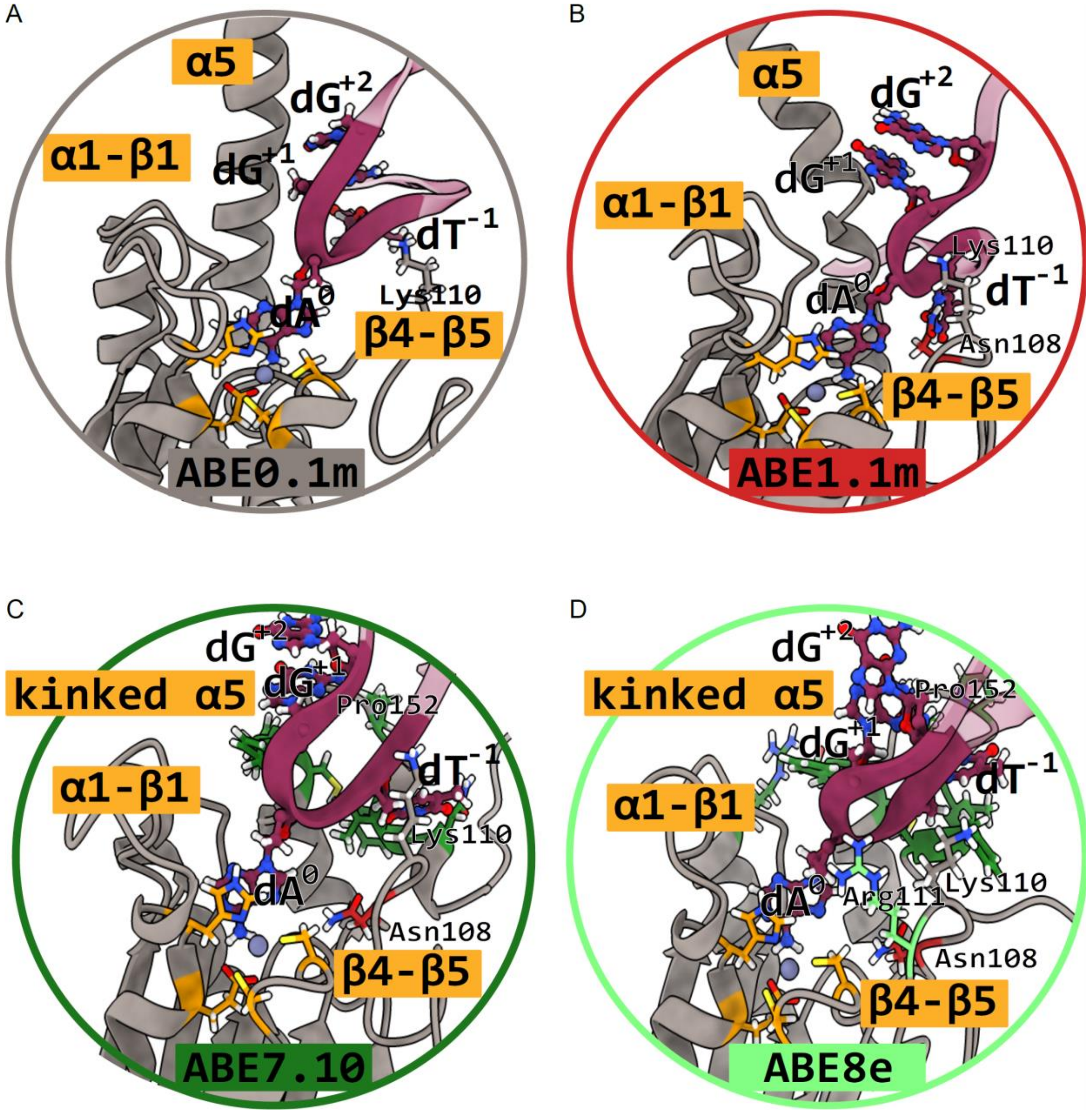


Figure 3.13: Steric clashes with the α5 helix prevent early generation ABEs from editing substrates with G^{+1}

All-atom molecular simulations of full-length **(A)** ABE0.1, **(B)** ABE1.1, **(C)** ABE7.10, and **(D)** ABE8e bound to gRNA and target DNA with the TAG motif were performed. The first 200 ns of simulation was excluded from analysis. Shown are simulation snapshots of the T$\underline{A}$G target bound to the ABE variants indicated, highlighting the conformation adopted by the G^{+1} base relative to the secondary structure elements of TadA. Specifically, in ABE0.1 and ABE1.1, the non-kinked α5-helix sterically clashes with this base, while the kink in the α5-helix in ABE7.10 and ABE8e avoid this unfavorable interaction.

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 3.12**). Overall, these mutations enhance editing activity while simultaneously relaxing the T^{-1} requirement as these residues bind the target through its phosphate backbone (**Figure 3.8A** and **3.8D**). 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 3.1C-D** and **3.12**). 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 3.8E**).

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 3.1E**). Given the apparent importance of the α5 helix on defining the sequence specificity of ABEs at the +1 position (the C^{+1} position in T$\underline{A}$CG), 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 3.14A**). 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 3.14B**). 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 3.14B**). 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 3.9**)

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 3.14C**). Interestingly, ABE1.1m(8e α5) displayed much higher editing than ABE1.1m when the C^{+1} base was mutated to G (**Figure 3.14C**). 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

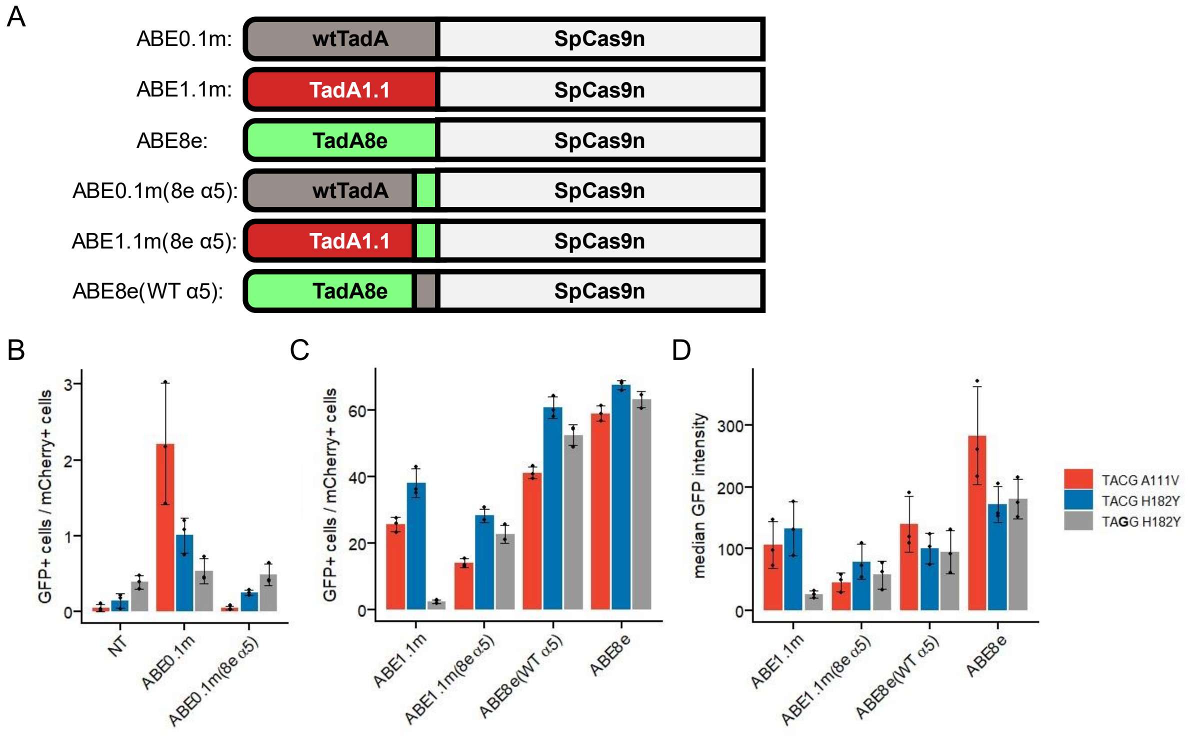


Figure 3.14: Mutations to the α5-helix affect C^{+1} requirements

(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 3.5C**) 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.

increase). These general trends were also observed when cells were analyzed for MFI of EGFP-positive cells (**Figure 3.14D**). 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.

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 3.14C-D**). However, overall editing activity by ABE8e(WT α5) was slightly reduced compared to ABE8e at all three reporters tested (**Figure 3.14C-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 $T^{-1}\underline{A}^{0}C^{+1}G^{+2}$ 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 (**Figure 3.15**). 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[86].

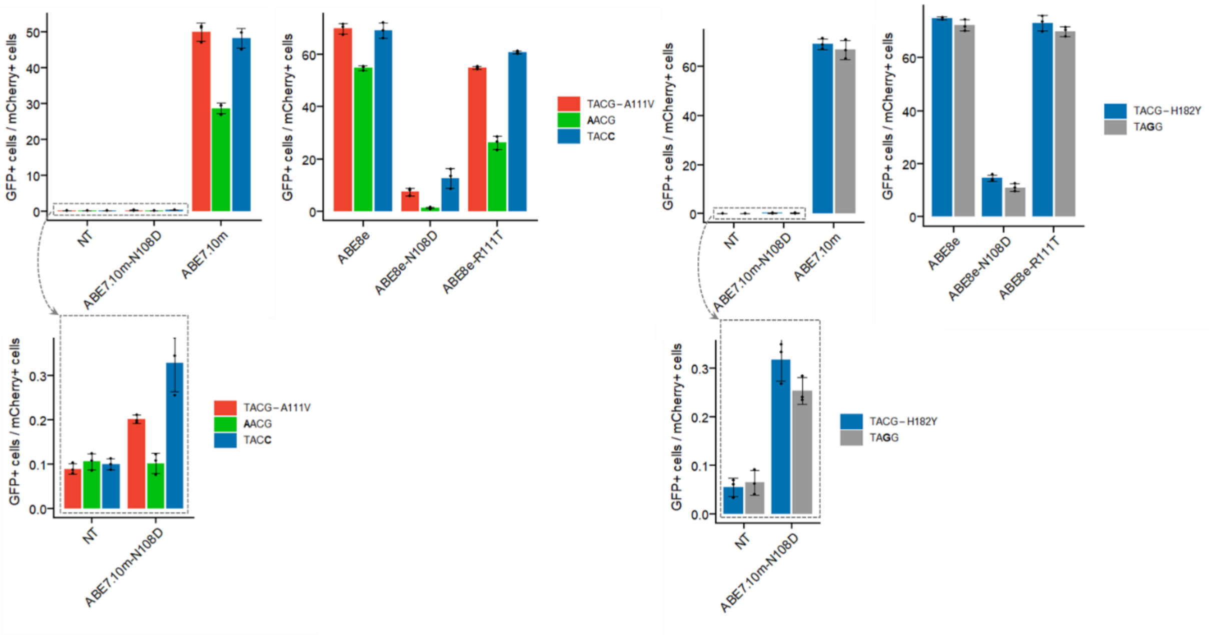


Figure 3.15: Reversion mutations in the β4-β5 loop reduce editing efficiencies by evolved ABEs

HEK293T cells were transfected with plasmids encoding the fluorescent reporter indicated, the ABE variant indicated (ABE7.10, ABE7.10-N108D, ABE8e, ABE8e-N108D, or ABE8e-R111T), and a NT or targeting gRNA. After 72 hours, cells were analyzed by flow cytometry. Shown are the percentage of transfected cells (determined based on mCherry fluorescence) that also display EGFP fluorescence. Values and error bars represent the mean and standard deviation for n = 3 biological replicates. Each replicate is marked individually.

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, **Figure 3.15**). 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 (**Figure 3.15**) 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, **Figure 3.15**). 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 (**Figure 3.15**). 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 (**Figure 3.15**). 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[74]. 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 of ABE editing of TACG motifs

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[77]. 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 (TACG) 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[82,101,102].

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}.[77]

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[72,91]. 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[74].

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.

Dimerization-impaired TadA variants

During the development of ABE7.10, the addition of a catalytically dead wild-type TadA monomer to the ABE molecule was found to increase base editing efficiency.[77] However, subsequent studies have shown that ABE7.10 and further evolved variants can be expressed with just the evolved TadA.[70] An explanation arose for this with a cryoEM structure of ABE8e, which showed the evolved TadA dimerizing with a separate ABE molecule.[74] As the editing efficiency of the ABE has increased with successive variants, it has remained unclear whether dimerization is a requirement for TadA activity.

The TadA dimerization interface is maintained in part by electrostatic interactions between residues in the α3 helix and the preceding loop (**Figure 3.16A**). On both subunits, the D53 and T55 residues interact with the R64 residue on the opposing subunit. To test if these interactions are crucial for DNA editing by TadA, ABE0.1 variants were cloned with each of these residues individually mutated to alanine, as well as a variant with all three residues mutated to alanine. These dimerization-impaired ABE0.1 variant plasmids were transfected into HEK293T cells along with the fluorescent TACG A111V reporter and corresponding gRNA plasmids. After three days, flow cytometry was used to detect mCherry and EGFP fluorescence in these cells (**Figure 3.16B**). The most drastic effect was caused by the D53A mutation which caused 20-fold reduction in cells displaying EGFP fluorescence, compared to ABE0.1m. The rate of EGFP turn-on in this sample was similar to the triple mutant (0.03%) and the NT gRNA control (0.06%). The other two single mutants also showed reduced editing, with the T55A mutation causing a 2.5-fold reduction in EGFP turn-on, and the R64A mutation causing a 1.6-fold reduction. The reduction in editing observed for each of these mutants suggests that DNA editing by ABE0.1 is dependent on TadA dimerization.

TadA variants evolved for DNA editing, such as ABE7.10, ABE8e, and ABE8.20 are hypothesized to be capable of acting as monomers. To gain insight into the dimerization requirement of these evolved TadAs, the dimerization-impairing mutations were cloned into ABE8e, individually and in a triple mutant combination. These variants were also analyzed using the TACG A111V reporter and the same protocol as the prior variants (**Figure 3.16C**) The D53A mutation again showed the largest effect, causing a 1.3-fold reduction in EGFP turn-on compared to the parental ABE8e. The T55A variant showed EGFP turn-on in 53.5% of transfected cells, while the R64A variant showed EGFP turn-on in 49.9% of transfected cells. Both of these rates were similar to the parental ABE8e which showed EGFP turn-on in 51.8% of transfected cells. The triple mutant ABE8e showed EGFP turn-on in 41.8% of transfected cells, which was similar to the D53A variant. This editing data suggests that ABE8e is capable of acting as a monomer, and is likely doing so.

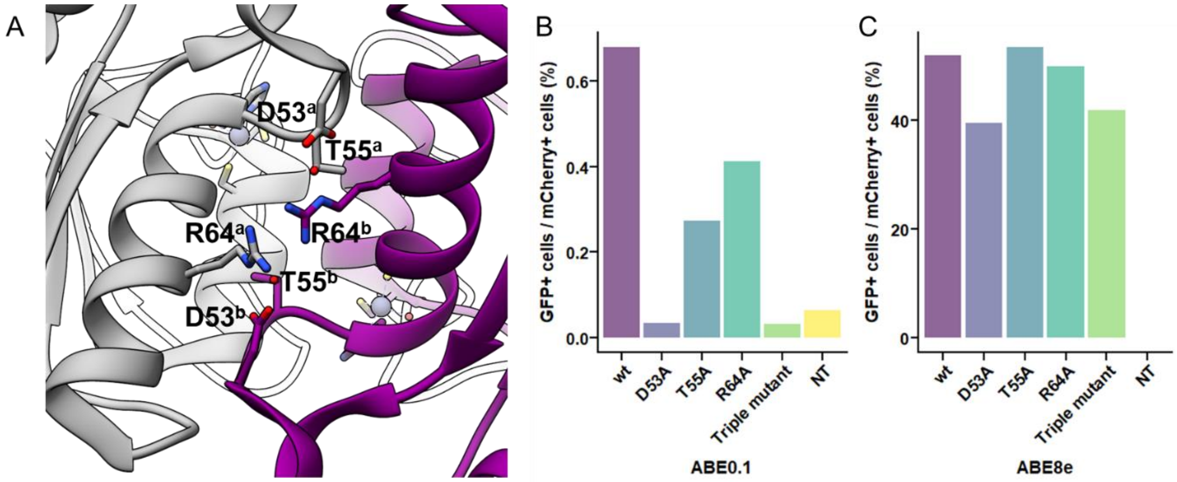


Figure 3.16: TACG A111V reporter editing by dimerization-impaired TadA variants

(A) (PDB ID: 1Z3A) The TadA homodimerization interface is shown, with the D53, T55, and R64 sidechains labelled on each subunit. **(B-C)** Plasmids expressing the indicated ABE variants were transfected into HEK293T cells along with the TACG A111V reporter plasmid and corresponding gRNA. After 72 hours, cells were analyzed by flow cytometry. Shown is the percentage of mCherry-positive, transfected cells that also show EGFP fluorescence. ABEs in **(B)** were made from ABE0.1, while variants in **(C)** were made from ABE8e. Values are for n = 1.

The D53 residue appears to be the most important residue for TadA dimerization, as an alanine mutation at this position abolishes editing activity by ABE0.1. Interestingly, this mutation was the only one that reduced editing by ABE8e, suggesting that this evolved TadA might still be dimerizing in cells, though this has a minimal effect on editing activity. In order to confirm this dimerization-dependent effect on editing, more replicates of this experiment are required. Additionally, more experiments should be done to confirm the dimerization ability of these TadA variants. Previously, we've observed significant gRNA-independent editing by ABE0.1m at UACG motifs in RNA.[73] These same sites could be observed for editing during the expression of these dimerization-impaired ABE variants. Finally, the dimerization state of these variants should be directly observed *in vitro*, to confirm that the lack of editing observed is a result of impaired dimerization.

Transfection efficiency affects the editing efficiency of fluorescent reporters

While using the panel of reporter plasmids to better understand the sequence-dependent editing of ABE0.1, significantly reduced editing efficiencies were occasionally observed. After the collection of many replicates, we observed that these results occurred in cells that showed low transfection efficiencies (**Figure 3.17**). For example, cells transfected with ABE1.1 showed as low as 7.3% EGFP turn-on of the A111V-TACG reporter with an 18% transfection efficiency, up to a 36.8% EGFP turn-on with an 81% transfection efficiency. In cells transfected with ABE8e and the A111V-TACG reporter, EGFP turn-on ranged from 37.4% to 63.3% with transfection efficiencies of 38.8% and 79.2%, respectively.

These large differences indicate that the rate of transfection plays a large role in determining editing efficiencies of these reporters, even though non-transfected cells are not included the calculation of editing efficiency. The reasoning for this effect may be linked to the observation that cells in our assays showing the highest intensity of mCherry fluorescence also show the highest intensity of EGFP fluorescence. This is likely due to these cells having a high number of total plasmids, leading to high expression of both the base editor and the fluorescent genes, and resulting in increased base editing occurring within these cells. As a result, we hypothesize that low transfection efficiencies indicate that fewer total plasmids enter these cells and reduce the number of opportunities for base editing. For future experiments with fluorescent reporters of base editing it is crucial to maintain similar transfection efficiencies across experiments, and it may be beneficial to develop an additional normalization procedure for this variable in the future.

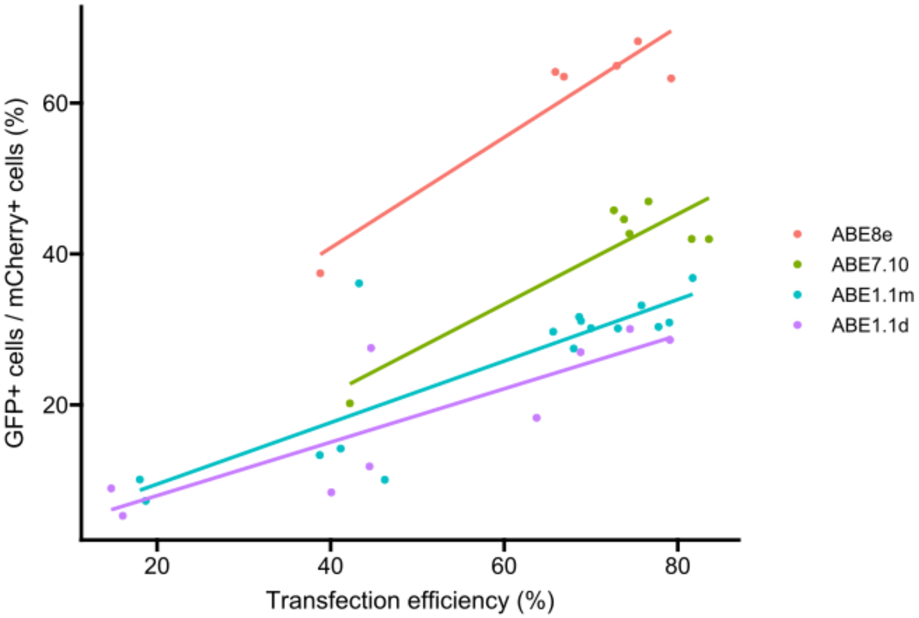


Figure 3.17: Transfection efficiency impacts reporter editing efficiency

Transfection efficiencies were determined across different fluorescent reporter experiments that used the TACG A111V reporter. These were plotted along with the corresponding editing outcomes and grouped by the indicated ABE variant.

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, 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[103]. 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[103]

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)[104]. 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 Quant-IT^{TM} dsDNA Assay Kit, high sensitivity (ThermoFisher Scientific #Q33120) on a Qubit fluorometer. Samples were then sequenced on an Illumina MiniSeq according to the manufacturer's protocol.

Analysis of Illumina HTS was performed with CRISPResso2[75]. 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 G 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[105,106]. Statistics tests were performed in R Studio using the "rstatix" package or with GraphPad Prism[107]. 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[108]. Single colonies were inoculated into Kanamycin containing culture, from which chemically competent target plasmid-containing S1030 stocks were developed[110]. 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)[74]. 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[109]. 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[111]. All titratable residues were protonated using the H++ server employing the default settings[112,113].

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′)[114]. 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[115]. 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[90]. 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[116]. 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[117–122]. 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[73,123].

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[122,124,125]. 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[126,127]. The visualization of the MD trajectories was rendered using ChimeraX, and data were plotted using Matplotlib[128–130].

Acknowledgements

Chapter 3 has been reproduced, in full, with permission, from **Ranzau, B. L.**, Rallapalli, K. L., Evanoff, M., Paesani, F. & Komor, A. C. The Wild-Type tRNA Adenosine Deaminase Enzyme TadA Is Capable of Sequence-Specific DNA Base Editing. *ChemBioChem*, e202200788 (2023). The dissertation author was the primary author of this paper.

Chapter 4 Directed evolution of new C-to-T base editors in bacteria

Introduction

Cytosine base editors (CBEs) use naturally occurring DNA-modifying enzymes to catalyze the deamination of cytosine to uracil.[76,81] When uracil is used as a template during DNA repair or replication, it is replaced with a thymine, resulting in a permanent C-to-T mutation. However, uracil is also specifically recognized and excised by the DNA-repair protein uracil N-glycosylase (UNG), leaving an abasic site which is resolved during DNA repair as any of the four canonical nucleotides. In current CBEs, this excision is inhibited by linking the uracil N-glycosylase inhibitor protein (UGI) to the C-terminus of the base editor. This inhibition is not perfect though and uracil excision can still occur. Further, the overexpression of UGI can also prevent UNG from protecting the genome from spontaneous deamination events that occur on the order of 200 times per day per cell.[131] Though current CBEs are effective at creating C-to-T mutations at target locations, they lack the precision to be used as therapeutics.

With adenosine base editors (ABEs) these same drawbacks are not observed. The inosine intermediate, compared to uracil, occurs 50-times less often in DNA and it is not excised by any known proteins.[132] This suggests that uncommon DNA modifications may evade excision by DNA repair proteins and might function better as base editing intermediates than more common modifications.

Two RNA modifications were chosen as potential intermediates for new C-to-T base editors: 3-methylcytosine (m^3C) and lysidine. Both of these modifications are found in tRNAs and occur along the base pairing edge of cytosine. m^3C has been shown to significantly disrupt the C:G base pair in favor of an m^3C:A base pair.[133,134] Similarly, lysidine is used to reprogram the wobble position of the <u>C</u>AU anticodon to have the same base pairing properties of a <u>U</u>AU anticodon.[135,136]

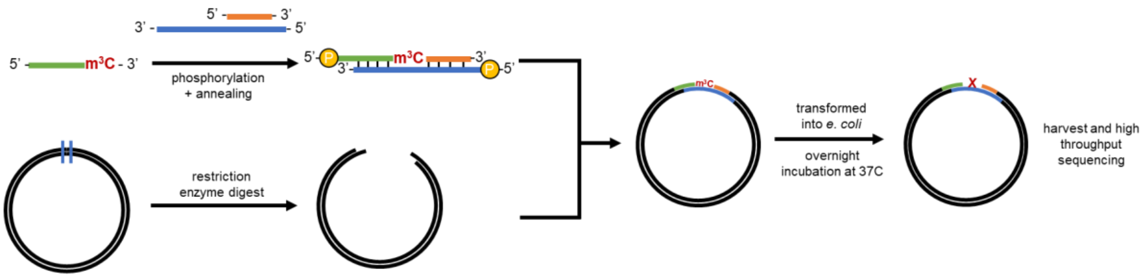
The incorporation of m^3C and L are catalyzed by two proteins: Trm140 and TilS, respectively. Trm140 is a SAM-dependent, tRNA-methyltransferase discovered in *S. cerevisiae* that installs m^3C_{32} in the anticodon loop of $tRNA^{Thr}$ and $tRNA^{Ser}$.[137,138] TilS is an ATP-dependent, tRNA-modifying enzyme found in *e. coli* that installs the L modification at position C_{34} in $tRNA^{Ile(CAU)}$. While structural studies have not been performed with Trm140, TilS has been shown to contain three discreet domains; the N-terminal domain is the catalytic domain (NTD), while two C-terminal domains (CTD1 and CTD2) recognize the $tRNA^{Ile}$ substrate.[139,140] The two recognition domains of TilS interact with vastly different parts of the tRNA, with CTD1 interacting with nucleotides in the anticodon loop and CTD2 interacting with nucleotides in the acceptor stem. In this chapter, base editors were developed with these two proteins and were tested for the ability to catalyze C-to-T editing in bacterial DNA.

m3C resolves to thymine

3-methylcytosine(m^3C) has been shown to mispair with adenosine to cause C-to-T mutations.[134,141] In order to test this finding, a DNA oligo was purchased containing an m^3C nucleotide and was ligated into a plasmid so that the m^3C was base paired with a G. This plasmid was transformed into *e. coil* and the bacteria were incubated for 24 hours so that the m^3C nucleotide could be used as a template in DNA replication. After the growth, the replicated plasmid was harvested and analyzed using high-throughput amplicon-sequencing (**Figure 4.1**).

Analyzing the results of this experiment, the m^3C nucleotide resolved to a T in 0.138% of the reads (**Figure 4.1**). This is less efficient than U, which resolved to a T in 0.302% of the reads but is within a similar magnitude and shows that m^3C may serve as a useful intermediate for C-to-T editing. For non-T outcomes, m^3C resolved to A in 0.038% of the reads and to G in 0.083% of the reads. As a comparison, uracil resolved to A in 0.363% of the reads and to G in 0.117% of the reads. Both of C-to-A and C-to-G outcomes were higher when using the U intermediate, especially C-to-A editing which was 10-fold higher than with the m^{3}C intermediate. These results show that m^{3}C can be used as an intermediate for C-to-T editing in bacteria and may result in higher product purity than a U intermediate.

A



B

X=	U	m3C	C
A	0.363	0.038	0.045
C	99.217	99.741	99.87
G	0.117	0.083	0.018
T	0.302	0.138	0.067
Reads	189819	330746	405650

Figure 4.1: Resolution of an m^{3}C intermediate in a bacterial plasmid

(A) Scheme for installing m^{3}C into a plasmid. An oligo containing m^{3}C was annealed with other oligos to form a ds DNA insert, which was ligated into a plasmid and transformed into *e. coli*. Plasmid was harvested after an overnight incubation, then analyzed with high-throughput amplicon sequencing. **(B)** High-throughput sequencing results. The top row shows the incorporated nucleotide, while the lower rows show the percentage of the indicated base that was read at the incorporation position, as well as the number of reads that were analyzed.

Trm140 and TilS base editors perform C-to-T editing in bacteria

To test Trm140 and TilS for C-to-T editing activity in DNA, bacteria-expression plasmids containing either Trm140 or TilS linked to dCas9 were cloned. The entire Trm140 gene was used to make a Trm140 base editor (Trm140-BE), while two TilS base editors were developed. The first TilS base editor contained only the catalytic NTD (TilS-BE), while the other included CTD1 (TilS(+CTD1)-BE). Since this recognition domain interacts with the anticodon loop, it may increase the affinity of a TilS base editor for specific sequence motifs. As a negative control, a deactivated Trm140-BE (dTrm140-BE) was also cloned by making a

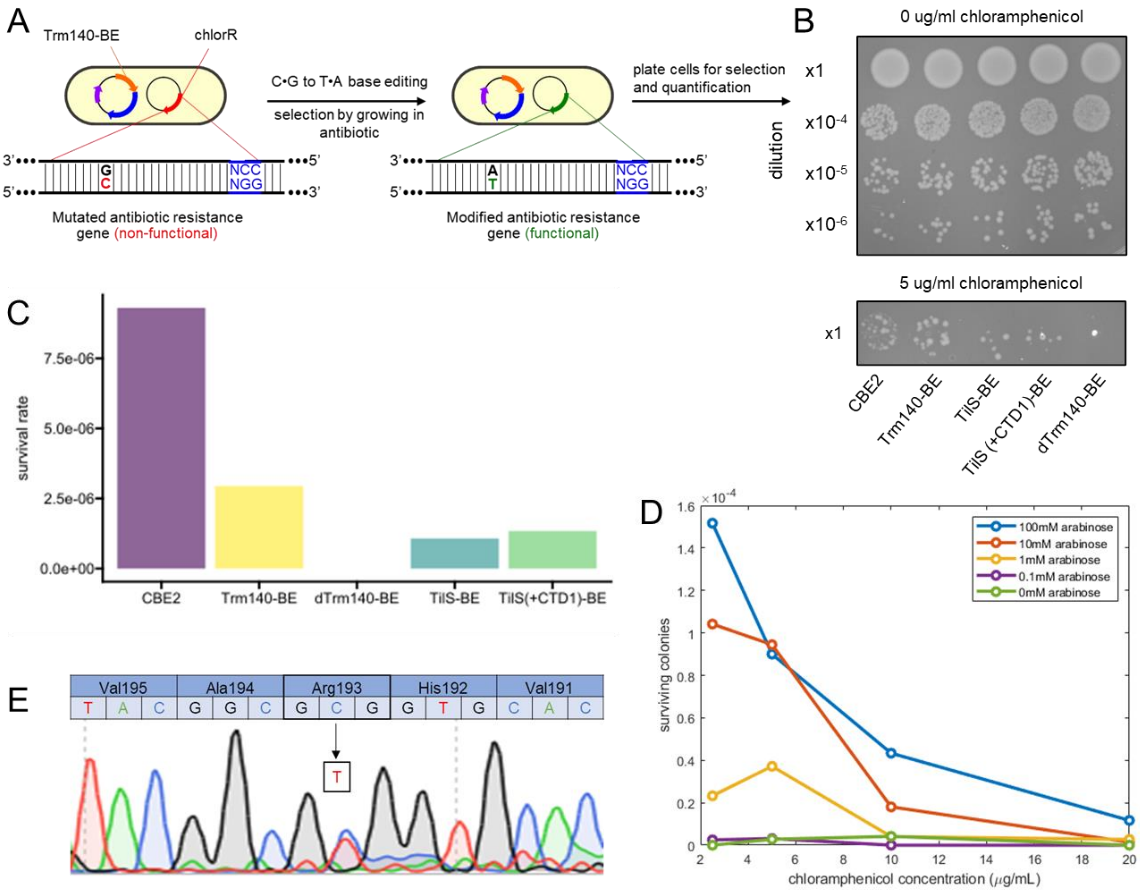


Figure 4.2: Trm140- and TilS- base editors rescue bacteria survival

(A) Scheme for an antibiotic resistance assay to detect C-to-T editing activity by Trm140 in bacteria. Bacteria are transformed with a selection plasmid containing a catalytically-inactive chloramphenicol resistance gene. A base editor plasmid is also transformed into these bacteria and targeted to reactivate the chloramphenicol resistance gene with a C-to-T mutation. Following arabinose-induced base editing, bacteria are spotted on media plates containing chloramphenicol. The number of colonies surviving selection are compared to the total number of colonies counted in aliquots not treated with chloramphenicol. **(B)** Representative example of bacteria colonies surviving antibiotic selection following base editing by the indicated base editors. **(C)** Quantification of the colony survival rate as observed in **(B)**. **(D)** Bacteria were treated with different concentrations of arabinose before selection. Colony survival increased with increasing concentrations of arabinose. **(E)** The selection plasmid was harvested from one of the colonies surviving chloramphenicol selection and sequenced to confirm the C-to-T edit that restores chloramphenicol resistance.

D190A mutation in the Trm140 gene.[138] Each base editor was tested in comparison to CBE2 using in a previously described antibiotic resistance assay (**Figure 4.2A**).[77] This assay is performed with a selection plasmid containing a chloramphenicol resistance gene that has been catalytically deactivated by a T-to-C mutation. Reversion of this mutation by a base editor restores chloramphenicol resistance in the bacteria. Each base editor was transformed into bacteria along with a selection plasmid and incubated in liquid media overnight to give time for base editing to occur. Bacteria were then spotted on media plates containing either 0 or 5 μg/ml chloramphenicol and incubated overnight. The survival rate was measured by counting the number of colonies that survived chloramphenicol treatment and normalizing that count using the number of colonies that grew on plates without chloramphenicol. Bacteria expressing CBE2 showed a survival rate of 9.3 x 10^{-6}, while bacteria expressing dTrm140-BE showed no survival in the presence of chloramphenicol. The new base editors showed lower survival rates than the CBE2, with bacteria expressing Trm140-BE showing a survival rate of 2.95 x 10^{-6}, TilS-BE showing a survival rate of 1.08 x 10^{-6}, and TilS(+CTD1)-BE showing a survival rate of 1.23 x 10^{-6} (**Figure 4.2B and 4.2C**). To confirm that a C-to-T edit was being introduced into the DNA sequence of the chloramphenicol gene, the selection plasmid was harvested from one of the colonies that survived selection with Trm140-BE. The H193R mutation in chloramphenicol was sequenced, showing that a C-to-T mutation had occurred in the selection plasmid to reactivate chloramphenicol resistance (**Figure 4.2E**). These data show that both Trm140-BE and TilS-BE are capable of editing a DNA sequence to rescue antibiotic resistance in bacteria.

Since base editor expression from the bacterial plasmids is controlled by an arabinose-inducible promoter, bacteria survival was tested with various concentrations of arabinose during the induction step (**Figure 4.2D**). Bacteria transformed with Trm140-BE were incubated with arabinose concentrations between 0.1 mM and 100 mM and the induced bacteria were plated on 2xYT plates containing either 2.5, 5, 10, or 20 μg/ml chloramphenicol. Bacterial survival was dependent on at least 1 mM arabinose and the survival rate increased with the arabinose concentration (**Figure 4.2D**). However, bacteria that were induced at 1 mM arabinose were unable to survive past 10 µg/ml chloramphenicol. Increasing the arabinose concentration to 10 mM enabled bacteria to survive in 10 µg/ml chloramphenicol, while 100 mM arabinose enable survival in 20 µg/ml chloramphenicol. Since a higher concentration of arabinose enabled a more stringent selection with higher concentrations of chloramphenicol, further bacterial experiments were performed using 100 mM arabinose for base editor induction.

Directed evolution of Trm140 and TilS base editors

Smaller base editors are preferable for therapeutic applications due to the packaging limitations of the viral vectors currently used for delivery. Since the two TilS-BEs tested showed similar editing results, further experiments with TilS only include the smaller NTD-alone variant. To see if Trm140 could also be shortened by a C-terminal deletion, two Trm140 variants were cloned: Trm140(Δ326-345)-BE and Trm140(Δ326-352). The antibiotic survival assay was repeated to compare these Trm140-BE variants with the wild-type Trm140-BE. All base editors were transformed into bacteria, induced with 100 mM arabinose, and plating bacteria with 0, 5, 10, or 20 µg/ml chloramphenicol (**Figure 4.3**).

The dTrm140-BE was unable to rescue bacteria survival under all conditions, again showing that survival is linked to C-to-T editing by Trm140-BE. The larger deletion variant, Trm140(Δ326-352), only showed activity at the 10 µg/ml chloramphenicol. At 5.9 x 10^{-7}, this variant showed the lowest survival rate in the experiment, suggesting that this deletion impairs DNA editing by Trm140. Interestingly, the Trm140(Δ326-345) variant showed higher survival than wild-type at all concentrations. At 5 µg/ml chloramphenicol Trm140-BE showed 2.9 x 10^{-6} ± 3.0 x 10^{-7} survival, while Trm140(Δ326-345)-BE showed 5.7 x 10^{-6} ± 6.9 x 10^{-6} survival. At 10 µg/ml chloramphenicol Trm140-BE showed 3.4 x 10^{-6} ± 3.3 x 10^{-6} survival, while Trm140(Δ326-345)-BE showed 9.2 x 10^{-6} ± 1.1 x 10^{-5} survival. Finally, at 20 µg/ml chloramphenicol Trm140-BE showed 0 survival, while Trm140(Δ326-345)-BE showed 1.9 x 10^{-6} ± 2.3 x 10^{-6} survival. While bacteria expressing Trm140(Δ326-345)-BE averaged higher survival rates across all chloramphenicol concentrations, there was a significant difference between replicates. To better understand the effects of deleting amino acids 326-345, more experiments are necessary. This region may be important for activity however, based on the lack of bacteria survival when using the Trm140(Δ326-352) variant. As a result, the full-length Trm140 was used for further experiments.

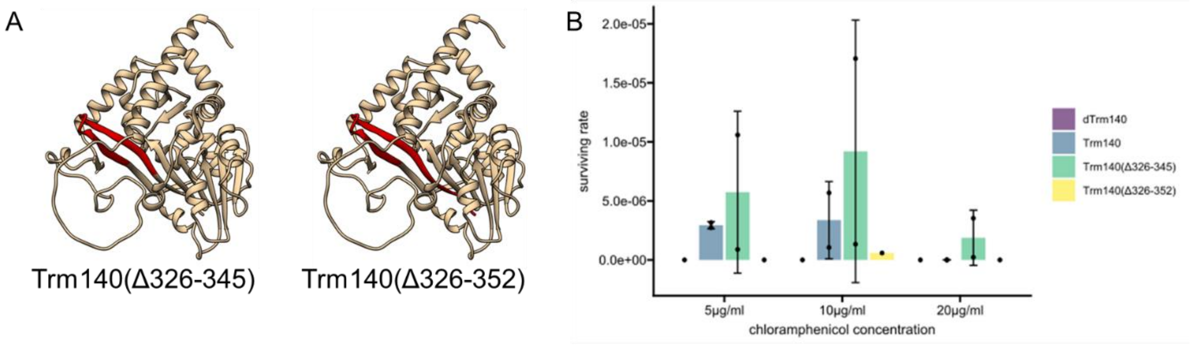


Figure 4.3: Bacteria survival with Trm140 N-terminal deletion variants

(A) An alpha-fold generated Trm140 structure with the deleted residues shown in red. **(B)** The indicated Trm140 variants were transformed in *e. coli* along with the chloramphenicol selection plasmid. Bacteria were grown in solution media overnight while expressing the base editor, then plated on media plates containing either 0, 5, 10, or 20 µg/ml chloramphenicol. Reported are the number of colonies surviving on the chloramphenicol plates, normalized to the number of colonies present on the no-chloramphenicol plate. Values and error bars represent the mean and standard deviation for n = 2 biological replicates. Each replicate is marked individually.

Since a crystal structure of TilS with its target tRNA was available, a high quality library of mutants could be generated at positions that interact with the substrate.[140] In order to reduce RNA-specific interactions, mutations were made to residues interacting with either a 2’- hydroxyl group (E90, R94, R97, and D186) or with the rigid secondary structure of a tRNA (R130, S132, and S144). A variant library was created by performing PCR with primers

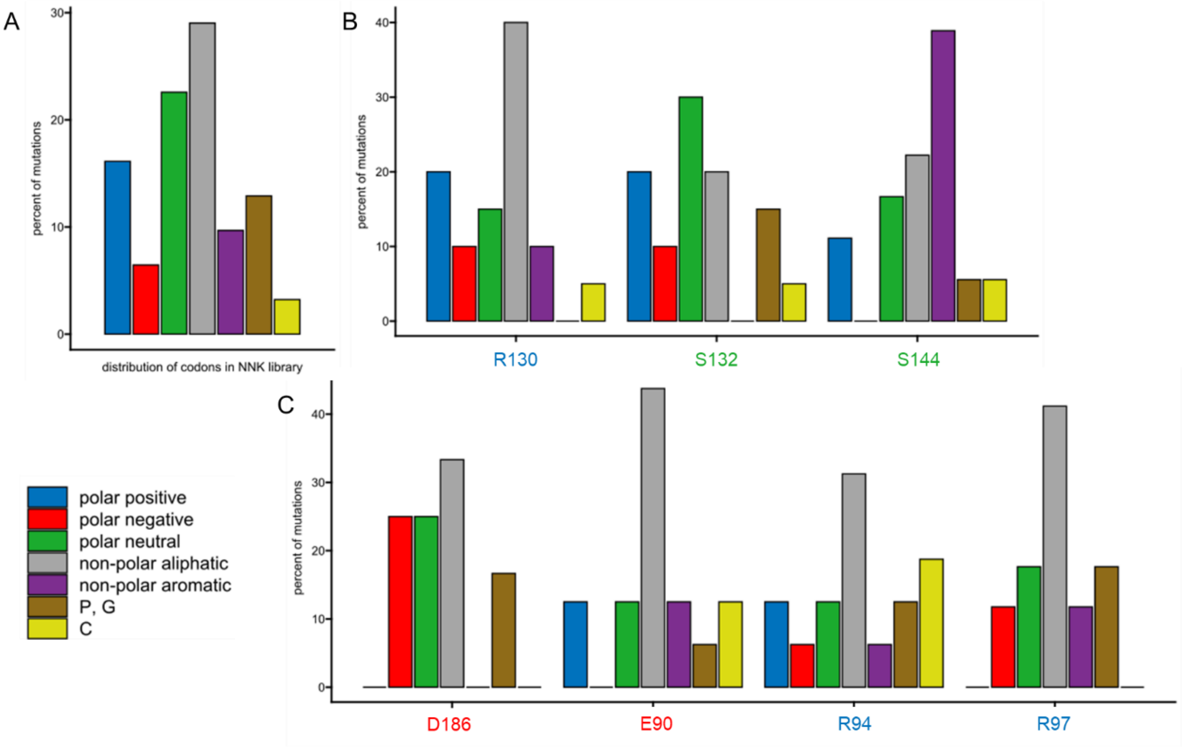
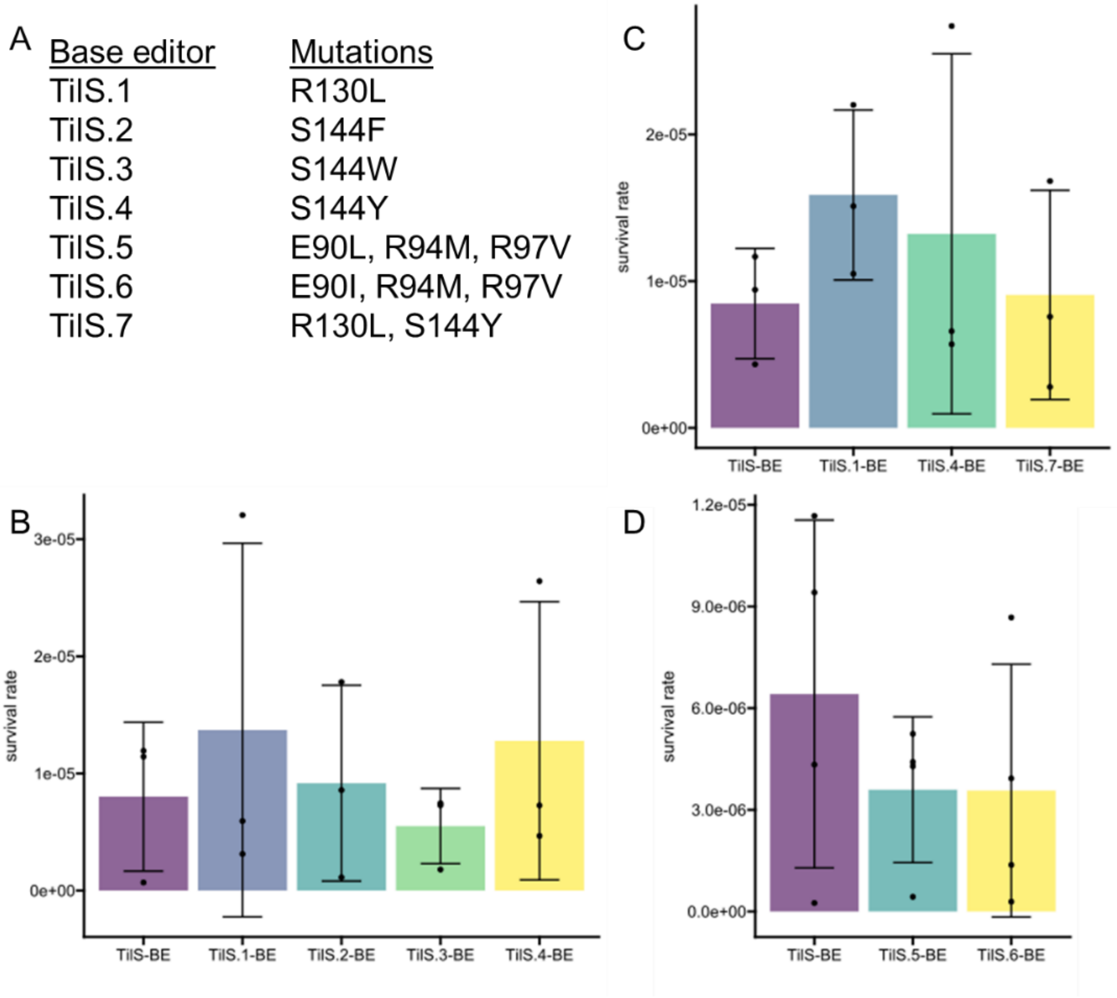


Figure 4.4: Distribution of amino acids in NNK libraries before and after selection

(A) The theoretical of distribution of codons within an NNK codon. **(B-C)** Following selection at 10 µg/ml chloramphenicol, 20 colonies were selected for sequencing. The distribution of amino acids at each position is reported here as a percentage of the total colonies sequenced.

containing degenerate NNK codons at each position of interest. These primers contained a variety of codons at the NNK position, allowing all possible amino acids to be introduced, while minimizing the number of stop codons in the mixture. To facilitate library creation, the seven residues of interest were split into four libraries, with close positions being grouped together (1: R130 and S132, 2: S144, 3: D186, 4: E90, R94, and R97). Each library was transformed into bacteria with a selection plasmid, base editor expression was induced overnight, and cultures were plated on 10 µg/ml chloramphenicol plates for selection. For each library, 20 surviving colonies were sequenced to determine which mutations were enriched. The mutations were sorted according to their side chain properties and compared to the expected distribution of these properties in an NNK library (**Figure 4.4**). R130 mutants showed an enrichment of non-polar aliphatic residues, the R130L mutation specifically was highly enriched. S144 mutants showed a strong selection for non-polar aromatic residues, with all three of these amino acids (Phe, Trp, and Tyr) all showing similar enrichment levels. The triple mutant library of E90, R94, and R97 mutants showed strong enrichment of non-polar aliphatic residues at all three positions. Specifically, a variant containing E90(I/L), R94M, and R97V mutations was highly enriched in this selection. Based on these enrichment results, six variant were selected for further characterization: TilS(R130L), TilS(S144F), TilS(S144W), TilS(S144Y), TilS(E90I, R94M, R97V), and TilS(E90L, R94M, R97V). To compare the editing efficiency of the TilS variants to the wild-type enzyme, each variant was tested in the antibiotic survival assay. Each variant was transformed into bacteria with a selection plasmid, base editor expression was induced overnight, and bacteria were spotted on media plates containing 5 μg/ml chloramphenicol (**Figure 4.5**). In the first experiment, wild-type TilS-BE was compared to the R130L and S144(F/W/Y) variants. The wild-type TilS-BE showed a survival rate of 8.0 x 10^{-6} ± 6.4 x 10^{-6}, while each of the variants showed similar survival rates. Nevertheless, the TilS.1 and TilS.4 variants (R130L and S144Y, respectively) showed higher survival rates at 1.3 x 10^{-5} ± 1.6 x 10^{-5} and 1.3 x 10^{-5} ± 1.2 x 10^{-5}, respectively (**Figure 4.5B**). The two mutations in these variants were combined to make TilS.7-BE to see if a significant increase in survival could be observed. The wild-type TilS-BE showed a similar survival rate as before at 8.5 x 10^{-6} ± 3.76 x 10^{-6}, which was also similar to all three variants. TilS.1 again showed the highest survival at 1.6 x 10^{-5} ± 5.8 x 10^{-6}, followed by TilS.4 at 1.3 x 10^{-5} ± 1.2 x 10^{-5}. The combination of these two variants into TilS.7-BE resulted in a survival rate near the wild-type level at 9.1 x 10^{-6} ± 7.1 x 10^{-6} (**Figure 4.5C**). Finally, the triple mutant variants, containing E90(I/L), R94M, and R97V mutations, were compared to the wild-type TilS-BE. Bacteria in this experiment treated with wild-type TilS-BE showed a survival rate of 6.4 x 10^{-6} ± 5.1 x 10^{-6}. Both TilS variants showed similar but lower survival rates, with TilS.5

showing 3.6 x 10^{-6} ± 2.2 x 10^{-6} and TilS.6 showing 3.6 x 10^{-6} ± 3.7 x 10^{-6} (**Figure 4.5D**). From these experiments, it appears that the R130L mutation may improve DNA editing by TilS, but not significantly.



A

Base editor	Mutations
TilS.1	R130L
TilS.2	S144F
TilS.3	S144W
TilS.4	S144Y
TilS.5	E90L, R94M, R97V
TilS.6	E90I, R94M, R97V
TilS.7	R130L, S144Y

Figure 4.5: Bacteria survival with TilS NNK variants

(A) TilS variants that were enriched from selection and tested for increased editing activity. **(B-D)** Testing each of the TilS variants in an antibiotic survival assay. Bacteria were transformed with the indicated base editor and the chloramphenicol selection plasmid. After inducing base editor expression overnight, bacteria were plated on media plates containing either 0 or 5 µg/ml chloramphenicol. Reported are the number of colonies surviving on the chloramphenicol plates, normalized to the number of colonies present on the no-chloramphenicol plate. Values and error bars represent the mean and standard deviation for n = 3 **(B-C)** or n = 4 **(D)** biological replicates. Each replicate is marked individually.

Optimization of selection variables for increased stringency

Since the TilS variants did not appear to significantly raise the DNA editing efficiency of TilS, the selection stringency for future experiments was optimized so that variants with a wild-type level of editing would drop out and mutations beneficial to DNA editing would be enriched. The antibiotic resistance assay can be easily used as a selection method with two variables that can modify selection stringency. The base editor induction time can be shortened to restrict how much time a base editor has to restore the antibiotic resistance gene. Also, the concentration of the selection antibiotic can be increased to require a higher editing efficiency to survive the selection. Optimization of these variables to prevent the survival of wild-type Trm140-BE and TilS-BE should allow only more efficient variants to survive.

The bacterial antibiotic survival experiment was repeated with CBE2, Trm140-BE, TilS-BE, and dTrm140-BE. After transformation of each base editor, bacteria were diluted into media and base editor expression was induced with 100 mM arabinose. Aliquots were pulled from each sample after 5, 6, and 7 hours of induction, then plated with various concentrations of chloramphenicol for selection (10, 25, or 50 µg/ml chloramphenicol). Surviving colonies were counted after an overnight incubation on these selection plates (**Figure 4.6A**).

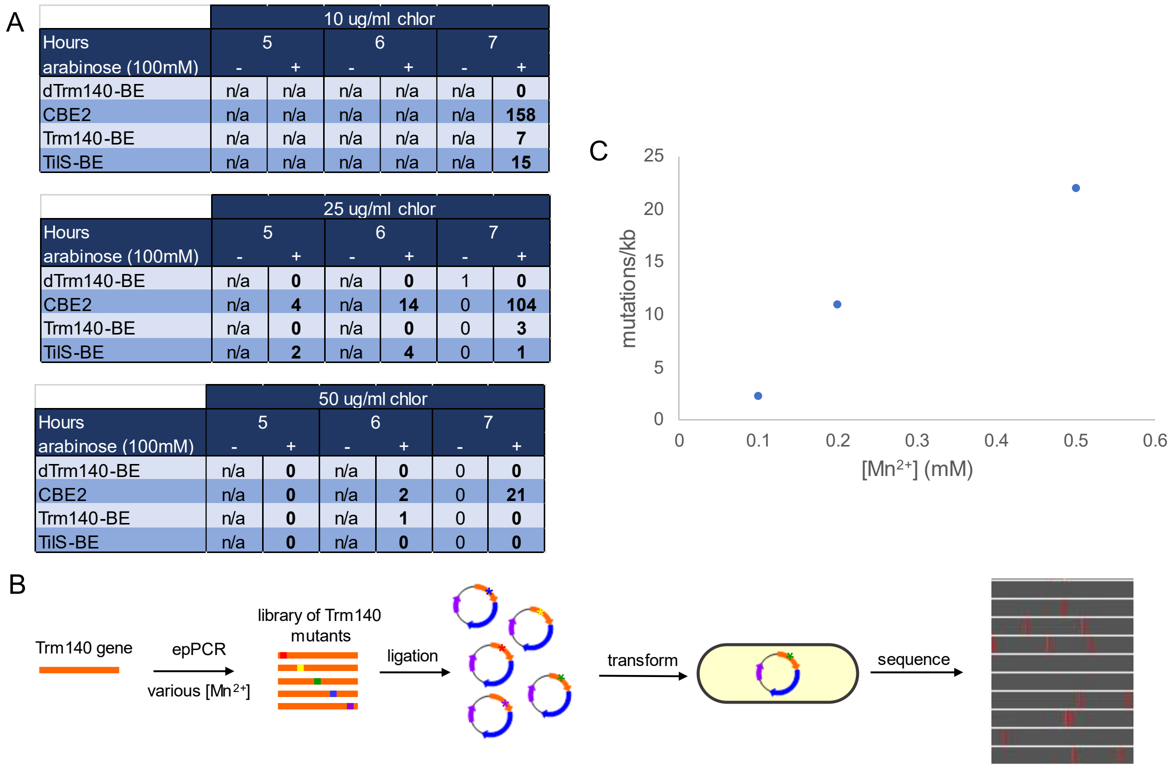
With only 5 hours of induction, colony survival was minimal for all base editors. When selected with 25 µg/ml chloramphenicol: 4 colonies survived with CBE2, 0 colonies survived with Trm140-BE, and 2 colonies survived with TilS-BE. The low survival with CBE2 expression suggests that 5 hours is likely too short for even an efficient base editor to correct the selection plasmid. The 6-hour induction led to increased survival with CBE2, with 14 colonies surviving at 25 µg/ml chloramphenicol and 2 colonies surviving at 50 µg/ml chloramphenicol. The new base editors performed relatively worse: at 25 µg/ml chloramphenicol 0 colonies survived with Trm140-BE and 4 colonies survived with TilS-BE, and at 50 μg/ml chloramphenicol 1 colony survived with Trm140-BE and 0 colonies survived with TilS-BE. The 7-hour induction increased survival even further. At 10 μg/ml chloramphenicol: 158 colonies survived with CBE2, 7 colonies survived with Trm140-BE, and 15 colonies survived with TilS-BE. At 25 μg/ml chloramphenicol: 104 colonies survived withCBE2, 3 colonies survived with Trm140-BE, and 1 colony survived with TilS-BE. At 50 μg/ml chloramphenicol: 21 colonies survived with CBE2 and 0 colonies survived with both Trm140-BE and TilS-BE.

These results show that a 7-hour induction is sufficient for the new base editors to perform DNA editing (**Figure 4.6A**). Bacteria expressing either of the new base editors were able survive at low concentration of chloramphenicol when induced for this time frame. At higher concentrations of chloramphenicol though, survival was more difficult and fewer colonies were observed. The combination of the 7-hour induction, followed by selection with 50 μg/ml chloramphenicol should be more stringent than the previous selection, which used an overnight induction and 10 μg/ml chloramphenicol.

Library production and selection

To generate a library of Trm140 and TilS variants for directed evolution, error-prone PCR (epPCR) was used to create mutations throughout the gene. This technique is performed with Taq polymerase supplemented with Mn^{2+}, which increases the error rate of Taq during DNA extension depending on the concentration of Mn^{2+}.[142] To calibrate this technique for generating ~2 mutations per variant, different concentrations of Mn^{2+} were tested during PCR (**Figure 4.6B**). The resulting amplicons were then ligated into the base editor backbone plasmid and transformed into bacteria. Each transformation was plated and 10 colonies from each Mn^{2+} condition were picked for sequencing. From this experiment, 0.1

mM Mn^{2+} created an average of 2.3 mutations per kb, which should create many mutations in a library, while maintaining enzyme activity in each variant (**Figure 4.6C**).



	10 ug/ml chlor					
Hours	5		6		7	
arabinose (100mM)	-	+	-	+	-	+
dTrm140-BE	n/a	n/a	n/a	n/a	n/a	**0**
CBE2	n/a	n/a	n/a	n/a	n/a	**158**
Trm140-BE	n/a	n/a	n/a	n/a	n/a	**7**
TilS-BE	n/a	n/a	n/a	n/a	n/a	**15**

	25 ug/ml chlor					
Hours	5		6		7	
arabinose (100mM)	-	+	-	+	-	+
dTrm140-BE	n/a	**0**	n/a	**0**	1	**0**
CBE2	n/a	**4**	n/a	**14**	0	**104**
Trm140-BE	n/a	**0**	n/a	**0**	0	**3**
TilS-BE	n/a	**2**	n/a	**4**	0	**1**

	50 ug/ml chlor					
Hours	5		6		7	
arabinose (100mM)	-	+	-	+	-	+
dTrm140-BE	n/a	**0**	n/a	**0**	0	**0**
CBE2	n/a	**0**	n/a	**2**	0	**21**
Trm140-BE	n/a	**0**	n/a	**1**	0	**0**
TilS-BE	n/a	**0**	n/a	**0**	0	**0**

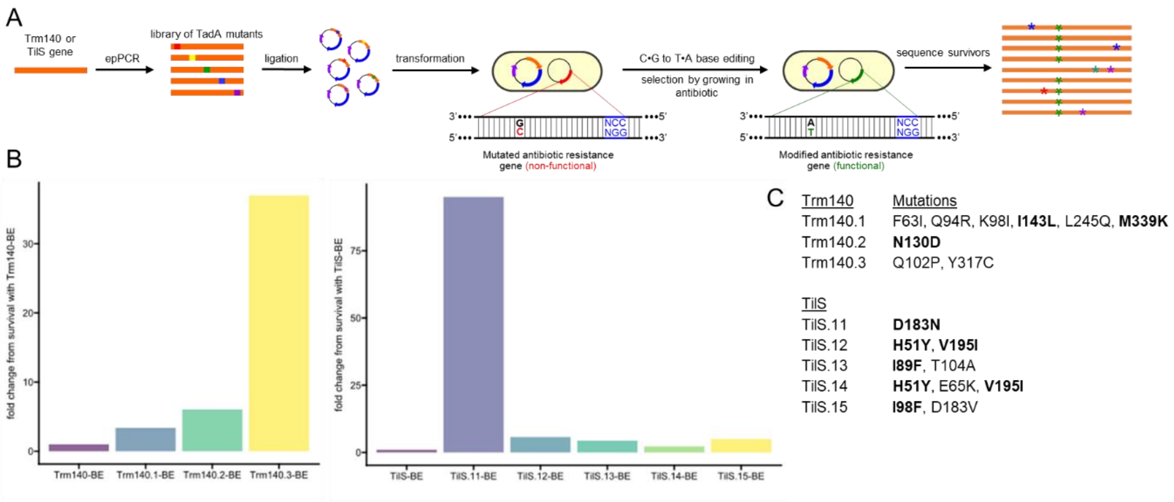
Figure 4.6: Optimization of bacterial directed evolution parameters

(A) Results of testing different arabinose induction times and different chloramphenicol concentrations for the selection of more active Trm140 variants. The indicated base editors were transformed into bacteria along with the selection plasmid containing a deactivated chloramphenicol resistance gene. Transformed bacteria were diluted into liquid media and incubated at 37C with shaking for the indicated amount of time, while base editor expression was induced with 100 mM arabinose. After induction, aliquots of each sample were plated on media plates containing the indicated concentration of chloramphenicol. Non-induced samples were incubated for the full 7 hours before being plated on the selection plates. Colonies on each plate were counted and reported here. **(B)** Scheme for creating a library of Trm140 variants and analyzing the number of mutations in each library member. The Trm140 gene was used as a PCR template with Taq polymerase supplemented with Mn^{2+}. Amplicons were ligated into a base editor backbone plasmid and transformed into bacteria. Bacteria colonies were picked and sequenced to determine the number of mutations in each colony. **(C)** Results of using different concentrations of Mn^{2+} to generate libraries of Trm140 variants. 20 colonies were analyzed from each library and the average number of mutations is reported.

With the selection optimized, libraries of Trm140 and TilS variants were produced using epPCR. The resulting amplicons were ligated and transformed into bacteria for harvesting. Plating a small aliquot from each transformation showed that ~1 x 10^6 different variants were present in both the Trm140 library and the TilS library. The harvested libraries were transformed into bacteria along with the selection plasmid. Induction was performed with 100 µg/ml arabinose for 7 hours, while a small non-induced aliquot was separated to ensure colony survival was dependent on base editor induction. Following induction, two aliquots were taken for selection at either 25 µg/ml or 50 µg/ml chloramphenicol (**Figure 4.7A**).

As controls, CBE2, Trm140-BE, TilS-BE, and dTrm140-BE were all separately transformed into bacteria and followed the same selection protocol as the Trm140 library. Non-induced samples of the bacteria transformed with these plasmids did not survive the chloramphenicol selection at either concentration. Further, only the bacteria expressing CBE2 were able to survive at 50 µg/ml chloramphenicol, suggesting that the Trm140 and TilS variants surviving at this concentration are likely showing higher activity than the wild-type base editors.

At 25 µg/ml and 50 µg/ml chloramphenicol respectively, 16 and 3 colonies survived from the Trm140 variant library, while 20 and 25 colonies survived from the TilS variant library. The editing enzyme was sequence from all colonies and a wide range of mutations were observed in each survivor. There were several mutations that occurred in two colonies, but enrichment beyond that was not observed, so the variants containing these mutations were selected for further testing (**Figure 4.7C**).



Trm140	Mutations
Trm140.1	F63I, Q94R, K98I, **I143L**, L245Q, **M339K**
Trm140.2	**N130D**
Trm140.3	Q102P, Y317C
TilS	
TilS.11	**D183N**
TilS.12	**H51Y**, **V195I**
TilS.13	**I89F**, T104A
TilS.14	**H51Y**, E65K, **V195I**
TilS.15	**I98F**, D183V

Figure 4.7: Directed evolution of Trm140 and TilS base editors

(A) The indicated Trm140-BEs were transformed into bacteria along with the deactivated chlorR gene to run the antibiotic survival assay. Following base editor induction, bacteria were spotted in duplicate on plates containing the indicated concentration chloramphenicol and incubated overnight at 37°C. Surviving colonies were counted and reported here as technical replicates. **(B)** The colony survival of each Trm140 variant at 10 µg/ml chloramphenicol from **(A)** were compared to the wild-type Trm140-BE and the fold-change graphed here. **(C)** Trm140 variants discovered during directed evolution. Mutations in bold were observed in 2 colonies.

To compare the editing activity of the new variants with the wild-type enzymes, each variant was cloned into individual plasmids and analyzed with the antibiotic survival assay. Each variant was plated on media plates containing 10 µg/ml chloramphenicol. All three of the Trm140 variants showed increased colony survival, but the Trm140.3-BE was particularly notable with a 37-fold increase in colony survival compared to the wild-type Trm140-BE. (**Figure 4.7B**). The TilS-BE variants also showed increased colony survival compared to the wild-type, with TilS.11-BE showing a 95-fold increase in colony survival compared to the wild-type TilS-BE. However, the colony counts throughout this experiment were lower than observed in previous experiments, which likely skewed the normalization calculations. Further experiments should be performed to confirm these results, but they suggest that the directed evolution of Trm140 and TilS can reveal mutations that improve DNA editing efficiency.

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. Trm140 and TilS genes were ordered through IDT as gBlocks and cloned into a pBAD-CBE1 plasmid, replacing the *APOBEC1* gene, via USER cloning following New England Biolabs (NEB) protocols. Trm140 truncation variants were also cloned with USER cloning, with primers that annealed within Trm140 and formed a junction that excluded the deleted sequence. Variants that were discovered by directed evolution were amplified directly out of the surviving colonies with primers that attached USER junctions to the end of the *Trm140* or *TilS* so that they could be cloned back into the base editor backbone.

m^{3}C incorporation into a plasmid

An oligo containing m^{3}C at the 3' end was ordered through IDT. This oligo was annealed with two other oligos to form a double-stranded insert with 4 bp overhangs on the 5' end of both strands of DNA. A receiver plasmid was designed with two BsaI cut sites that generated sticky ends to match the overhangs of the insert. This plasmid was first digested with BsaI then purified using a QIAquick PCR purification column (QIAGEN # 28104), following the manufacturer's protocol. The insert and digested backbone were ligated using T4 DNA ligase and purified using QIAquick PCR purification column (QIAGEN # 28104).

100 ng of plasmid containing the m^{3}C insert was transformed into *e coli* by heat shock. Bacteria was directly diluted into liquid 2xYT media, containing 100 µg/ml carbenicillin to maintain the m^{3}C plasmid, and incubated overnight at 37°C with shaking. The next morning, the plasmid was harvested from the bacteria culture using an E.Z.N.A. Plasmid

DNA mini kit (Omega Bio-Tek # D6942-00S). Plasmid DNA was used as a template for high-throughput amplicon sequencing.

The m^{3}C incorporation site was PCR amplified using Phusion High-Fidelity DNA Polymerase, and primers that bind to this site. PCRs followed the manufacturer's protocol, using 1 µl of plasmid 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 Quant-IT^{TM} dsDNA Assay Kit, high sensitivity (ThermoFisher Scientific #Q33120) on a Qubit fluorometer. Samples were then sequenced on an Illumina MiniSeq according to the manufacturer's protocol.

Analysis of Illumina HTS was performed with CRISPResso2[75]. 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.

Bacterial antibiotic survival assay

10ng of antibiotic selection 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 µg/mL Kanamycin maintenance antibiotic[108]. Single colonies were inoculated into Kanamycin containing culture, from which chemically competent target plasmid-containing S1030 stocks were developed[110]. 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 2xYT media. 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 µg/mL 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 2xYT media supplemented with 50 µg/mL kanamycin, 50 µg/mL carbenicillin, and 0 or 100 mM arabinose (to induce ABE expression, which is controlled by a pBAD arabinose-inducible promoter). Cultures were incubated at 37°C while shaking for 16 hours. Saturated cultures were then diluted 1 to $1x10^7$-fold in 2xYT, 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, 10, 25, or 50 µg/mL 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 10, 25, or 50 µg/mL chloramphenicol plates by the number of colonies that survived on the 0 µg/mL chloramphenicol plates.

Generation of Trm140 and TilS variant libraries

TilS NNK libraries were generated using Gibson Assembly.[145] Briefly, primers were ordered from IDT containing NNK codons at the positions being mutagenized. Each of these primers were used to amplify the TilS base editor plasmid along with a reverse primer that overlapped the forward primer up to the first NNK codon. Amplicons were gel purified from a 1% agarose gel with QIAquick Gel Extraction Kit (Qiagen #28704). Ligations were performed using Gibson Assembly master mix (NEB #E2611), according to the manufacturer’s protocol.

Error prone PCR (epPCR) libraries for Trm140 and TilS were generated using USER cloning. Briefly, epPCR was performed using Taq polymerase (NEB #M0273), supplemented with 0.1 mM Mn^{2+}. Primers amplified the *Trm140* or *TilS* genes and added overhang sequences that were used as USER junctions during the USER reaction. The mutagenized insert was ligated with a USER reaction into a backbone plasmid containing the remaining base editor components (generate by PCR using the Phusion-U protocol described earlier).

For both techniques, ligated plasmids were transformed into *e. coli* via heat shock. 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 µg/ml carbenicillin. The remaining culture was diluted to 40 ml with 2xYT supplemented with 50 µg/ml carbenicillin and incubated overnight at 37°C with shaking. Plasmid libraries were then harvested using ZymoPURE^{TM} II Plasmid Midiprep kit (Zymo Research #D4200).

Directed evolution selection

Selection of base editor variants was performed as described in the bacterial antibiotic survival assay. Following selection, variants were sequence by using the surviving colonies for templates in a colony PCR. Briefly, individual colonies were picked and lysed by mixing into 20 µl of water and heating at 95°C for 5 minutes. PCRs were performed using 0.5 µl of the lysed colonies as template and primers that amplified the *Trm140* or *TilS* gene. The resulting amplicon was submitted for sanger sequencing without clean up.

Chapter 5 Detecting DNA-editing by RNA-modifying enzymes in mammalian cells

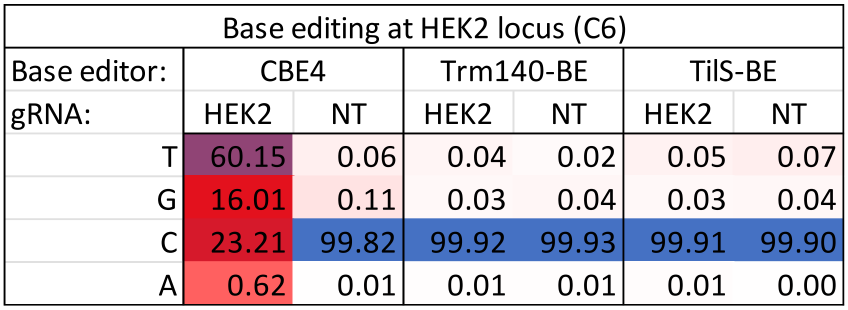
Introduction

Many RNA-modifying enzymes evolved to recognize specific sequences, so that modifications would only be applied to specific nucleotides within the target RNA.[87] Previously, the natural ability of TadA to recognize UACG motifs in RNA was exploited to observe that ABE0.1 could also recognize and modify TACG motifs in DNA.[143] This observation suggests that other RNA-modifying enzymes may also edit DNA sequences that match the RNA context in which they naturally operate.

In this chapter, multiple tRNA-modifying enzymes were linked to nCas9 to form new base editors and their activity was tested in mammalian cells. To better detect their activity, these new base editors are targeted to edit DNA sequences that match the contexts of their natural tRNA targets. As potential new C-to-T base editors, Trm140 and TilS were both tested with GFP reporters optimized with the respective recognition motifs of both enzymes. These same reporters were used to detect C-to-T editing in DNA by the tRNA adenine deaminase ADAT2, which is a finding that has been previously reported.[144] Finally, base editors are created with TrcP or Pus7 to test if psuedouracil can be incorporated into DNA in order to avoid excision by UNG.

Trm140 and TilS base editing in genomic DNA

Having previously confirmed that the Trm140-BE and TilS-BE both show DNA editing in bacteria, they were each cloned into a mammalian expression vector (which contains nCas9(NG) instead of the dCas9 used in bacteria) to test for editing activity at a genomic site in mammalian cells. The HEK2 site is a common target for base editor characterization as it is amenable to high levels of base editing, which may allow for detectable levels of editing by Trm140-BE.[76] HEK293T cells were transfected with either Trm140-BE, TilS-BE, or CBE4 and either a HEK2-targeting gRNA or a non-targeting gRNA. After 3 days, cells were lysed and the HEK2 locus was sequenced using high throughput amplicon sequencing (**Figure 5.1**). Efficient base editing was observed with CBE4, out of all the reads 60.13% showed C-to-T edits, 16.01% showed C-to-G edits, and 0.62% showed C-to-A edits. The Trm140-BE and TilS-BE however, did not show any evidence of base editing above 0.1%, with 99.92% and 99.91% of reads showing a C at the target site, respectively.



Base editing at HEK2 locus (C6)						
Base editor:	CBE4		Trm140-BE		TilS-BE	
gRNA:	HEK2	NT	HEK2	NT	HEK2	NT
T	60.15	0.06	0.04	0.02	0.05	0.07
G	16.01	0.11	0.03	0.04	0.03	0.04
C	23.21	99.82	99.92	99.93	99.91	99.90
A	0.62	0.01	0.01	0.01	0.01	0.00

Figure 5.1: gDNA editing by Trm104 and TilS base editors

The indicated base editors were transfected into HEK293T cells along with the indicated gRNA. After 3 days, cells were lysed and the HEK2 locus was amplified for high-throughput amplicon sequencing. Samples were sequenced and the base reads reported here.

Designing fluorescent reporters for detecting Trm140 and TilS base editing

Transiently expressed, fluorescent reporters of base editing have been shown to improve the detection of low editing levels.[143] Initially, an EGFP reporter containing a Y67C mutation was tested with Trm140-BE (**Figure 5.2A**). This reporter has previously shown efficient editing when targeted by APOBEC-based CBEs but is not optimized with a specific target sequence. This reporter plasmid and the corresponding gRNA plasmid were transfected into HEK293T cells along with the Trm140-BE variants. After 3 days, cells were analyzed for EGFP fluorescence by flow cytometry. BE3b-treated cells showed clear evidence of editing with 27.7 ± 3.8% of transfected cells also showing EGFP fluorescence (**Figure 5.2B**). Negative controls in these experiments ranged from 0.03 – 0.08% EGFP turn-on, with the highest editing being observed when Cas9 or nCas9 were targeting the reporter mutation. The Trm140-BE did not show editing above these negative controls, with Trm140-BE showing 0.0451 ± 0.0009% EGFP turn-on (**Figure 5.2C**). These results show that

Trm140 still requires significant directed evolution to reach the editing efficiency displayed by APOBEC1-based CBEs.

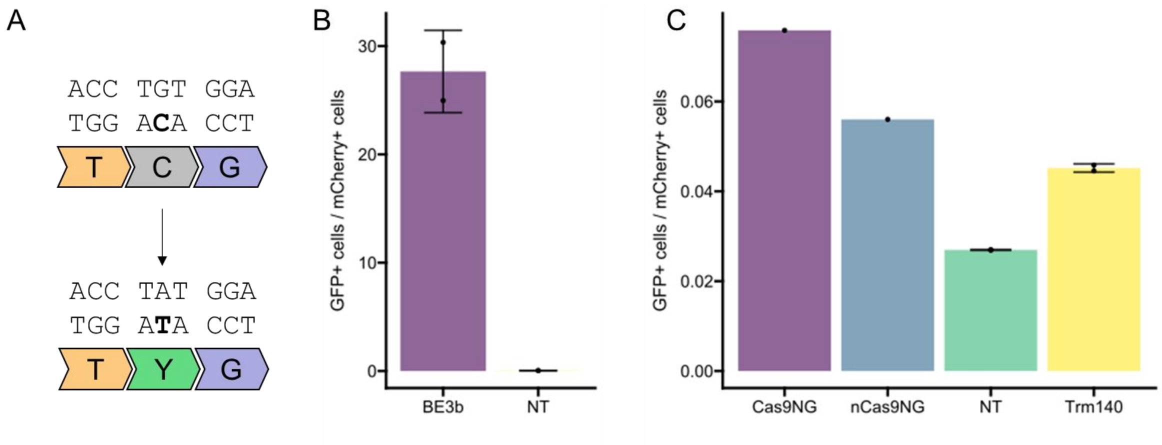


Figure 5.2: Trm140-BE editing with a Y67C fluorescent reporter

(A) Target site of the dGFP(Y67C) fluorescent reporter. A C-to-T edit at the bold nucleotide restores EGFP fluorescence. **(B-C)** The indicated base editors were transfected into HEK293T cells along with the EGFP(Y67C) fluorescent reporter and corresponding gRNA. Transfected cells were analyzed with flow cytometry to determine the percentage showing EGFP fluorescence, which is reported here. NT samples were treated with a non-targeting gRNA and BE4 in **(B)** and Trm140 in **(C)**. All other samples used a targeting gRNA along with the indicated base editor/Cas9. Values and error bars represent the mean and standard deviation for n =1 for Cas9NG and nCas9NG samples, and n = 2 for all other samples (biological replicates). Each replicate is marked individually.

Though DNA editing with Trm140-BE wasn't observed with a general fluorescence reporter, a reporter that utilizes a Trm140 recognition motif at the target site may improve editing to a detectable level.[143] A report on Trm140 target recognition showed that a G_{35}-U_{36}-t^6A_{37} motif in the target tRNA is recognized by Trm140 to methylate a C_{32}.[146] Based on this finding, a new reporter was developed that uses dGFP(Y93H) and contains a **C**acGTA motif at the mutation site (this reporter and motif was also described in a previous study and used for characterizing different APOBEC-BEs).[82]

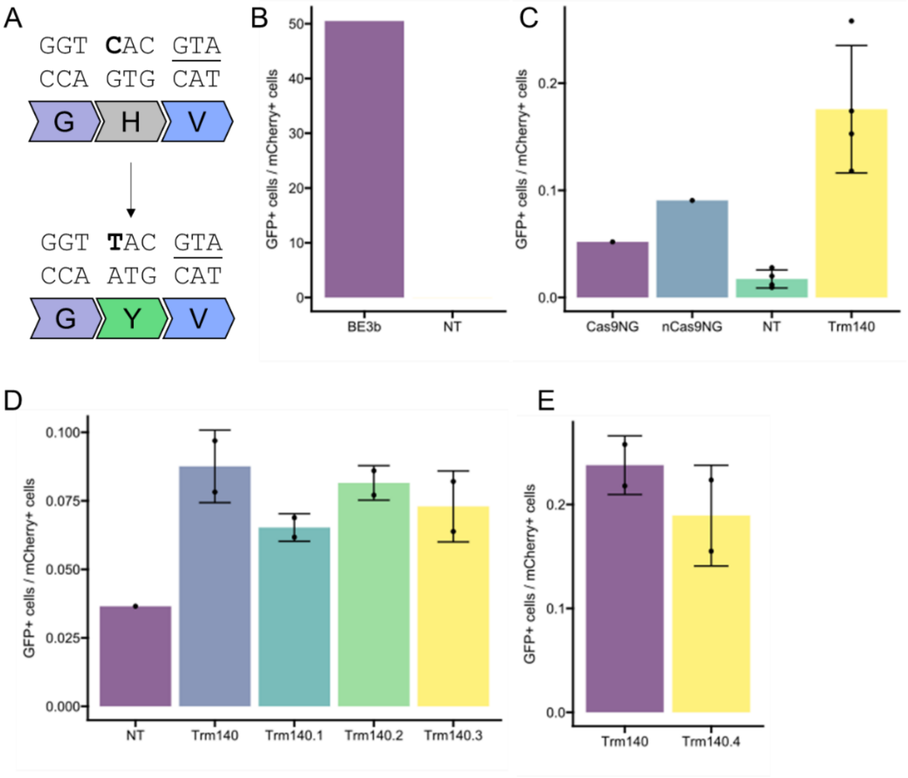


Figure 5.3: Trm140-BE editing with an optimized reporter

(A) Target site of the dGFP(Y93H) fluorescent reporter. A C-to-T edit at the bold nucleotide restores EGFP fluorescence. The underlined nucleotides indicate the Trm140 recognition motif. **(B-E)** The indicated base editors were transfected into HEK293T cells along with the EGFP(Y93H) fluorescent reporter and corresponding gRNA. Transfected cells were analyzed with flow cytometry to determine the percentage showing EGFP fluorescence, which is reported here. NT samples were treated with a non-targeting gRNA and BE3b in **(B)** and **(D)**, and with a non-targeting gRNA and Trm140 in **(C)**. All other samples used a targeting gRNA along with the indicated base editor/Cas9. Values and error bars represent the mean and standard deviation for **(B)** n = 1. **(C)** n = 1 for Cas9NG and nCas9NG, n = 4 for NT and Trm140. **(D)** n = 1 for NT, and n = 2 for Trm140 variants. **(E)** n = 2

The Trm140-optimized fluorescent reporter was transfected into HEK293T cells along with BE3b and either a T- or NT-gRNA. After 3 days, cells were prepared for flow cytometry and editing efficiency was determined by measuring the percentage of transfected cells showing EGFP fluorescence. As with the previous pACK243 reporter, BE3b showed efficient editing with the new reporter, with 50.5% of transfected cells also showing EGFP fluorescence (**Figure 5.3B**). The experiment was repeated, this time using the Trm140-BE, and EGFP turn-on was detected in 0.18 ± 0.06% of transfected cells. In the negative controls

EGFP turn-on ranged from 0.017 – 0.09%, showing that editing activity by Trm140-BE can be detected using an optimized EGFP reporter (**Figure 5.3C**).

With an optimized reporter for Trm140-BE activity, the experiment was repeated once again using the Trm140 variants that survived the bacterial selection (**Figure 5.3D**). In this set of experiments, the wild-type Trm140-BE showed EGFP turn-on in 0.087 ± 0.013% of transfected cells. Despite the improved editing observed in the bacterial assays, the Trm140 variants did not show improved editing over the wild-type (Trm140.1-BE: 0.065 ± 0.005%, Trm140.2-BE: 0.082 ± 0.006%, Trm140.3: 0.073 ± 0.013%). A Trm140(K283A) variant was also tested, which has been observed to increase m^{3}C abundance in *S. cerevisiae* (**Figure 5.3E**).[138] With the optimized fluorescence reporter though, this variant did not improve editing above the wild-type Trm140-BE (Trm140-BE: 0.24 ± 0.03% and Trm140.4-BE: 0.19 ± 0.05%). Despite observing increased survival with these variants in the bacterial antibiotic survival assay, in mammalian cells they all show similar editing as the wild-type Trm140-BE.

Seeing that Trm140-BE activity could be detected with an optimized fluorescent reporter, an additional reporter was designed to detect base editing by TilS-BEs. Studies looking at the tRNA recognition properties of TilS have shown that the CTDs of the enzyme primarily interact with nucleotides outside of anticodon loop.[139] However, the TilS-BE only includes the NTD and it is not known which nucleotides in the anticodon loop interact directly with this domain. Since TilS only interacts with tRNA^{Ile2}, a reporter was optimized using the entire single-stranded region of the anticodon loop (**Figure 5.4**). However, there was not a sequence within the EGFP gene where a TCATAA sequence could be silently introduced. The dGFP(Y93H) reporter was used again, but an additional V94N mutation was introduced to create the full TilS-recognition motif. Since the side chain of this residue is orientated outside of the GFP β-barrel, it is unlikely to impact fluorescence, and this was confirmed by transfecting the TilS-optimized reporter along with BE3b. Efficient EGFP turn-on was observed with the targeting gRNA, with 39.6% of transfected cells showing EGFP fluorescence, while the non-targeting gRNA only showed an EGFP turn-on rate of 0.095% (**Figure 5.4B**).

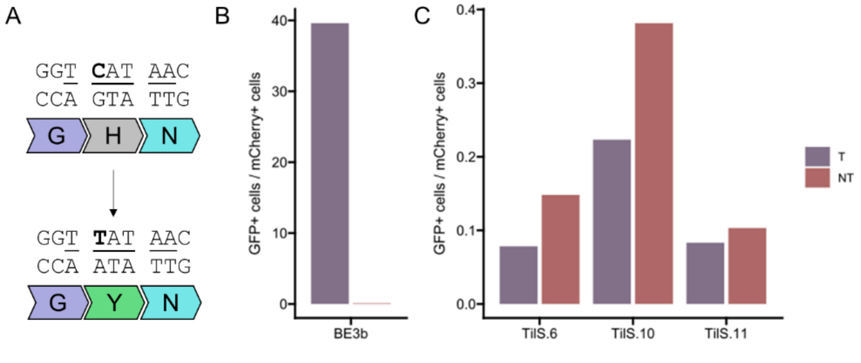


Figure 5.4: TilS-BE editing with an optimized reporter

(A) Target site of the dGFP(Y93H, V94N) fluorescent reporter. A C-to-T edit at the bold nucleotide restores EGFP fluorescence. The underlined nucleotides indicate the TilS recognition motif. **(B-C)** The indicated base editors were transfected into HEK293T cells along with the dGFP(Y93H, V94N) fluorescent reporter and corresponding gRNA. Transfected cells were analyzed with flow cytometry to determine the percentage showing EGFP fluorescence, which is reported here. Values represent the outcomes for n = 1 biological replicate.

To detect DNA editing by TilS-BEs, the new TilS-optimized reporter was transfected into HEK293T cells along with several TilS-BE variants. After three days, cells were analyzed by flow cytometry to observe EGFP fluorescence in transfected cells. With all three variants, EGFP turn-on rates were lower than the non-targeting controls (**Figure 5.4C**). The EGFP intensity observed in the EGFP-positive cells appeared to be significantly lower than what was observed in previous experiments using ABE0.1 or Trm140-BE, suggesting that none of the TilS-BEs were able to perform DNA editing on the optimized reporter.

Though DNA editing activity by TilS-BEs was observed in bacteria, it is possible that the bulky lysidine modification is efficiently detected and removed in mammalian cells. This hurdle could be overcome with methods that reduce lysidine repair, or by increasing the activity of TilS-BE so that it can compete directly with the DNA repair mechanisms of the cell.

To focus on creating a new C-to-T base editor though, the current work moved forward with the Trm140-BE, as it showed detectable levels of DNA editing within mammalian cells.

Testing Trm140-BE deletion variants

Following the release of AlphaFold, the software was used to generate a prediction for the structure of Trm140.[115] This structure showed that the N-terminus of Trm140 is highly disordered and may block the active site. To see if base editing would be affected by deleting this part of the protein, two Trm140 deletion-variants were cloned, removing residues 2-88 or 2-108 (**Figure 5.5**). We tested these variants with the optimized EGFP reporter assay and observed that neither of these deletions significantly changed the EGFP turn-on efficiency of the Trm140-BE (Trm140-BE: 0.10 ± 0.05%, Trm140(Δ2-88)-BE: 0.10 ± 0.06%, Trm140(Δ2-108): 0.13 ± 0.08%).

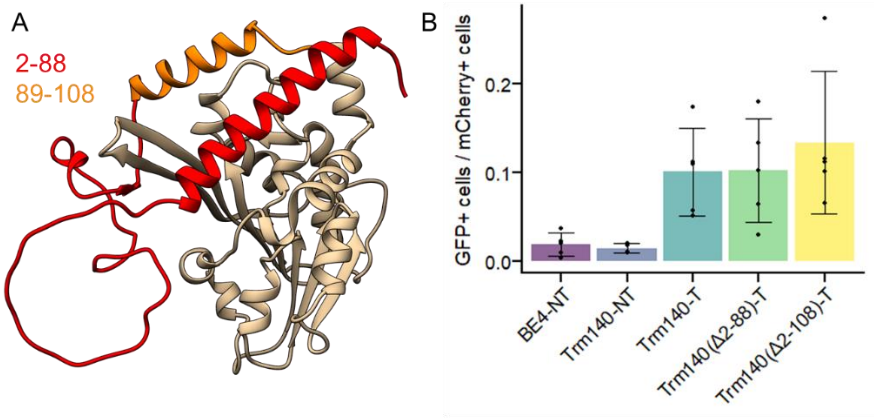


Figure 5.5: Trm140 N-terminal deletions

(A) A Trm140 structure predicted by AlphaFold, with the tested N-terminal deletions shown in red and orange. **(B)** The indicated base editors were transfected into HEK293T cells along with the Trm140-optimized fluorescent reporter and corresponding gRNA. Transfected cells were analyzed with flow cytometry to determine the percentage showing EGFP fluorescence, which is reported here. Values and error bars represent the mean and standard deviation for n = 5 biological replicates.

We were curious if the Trm140 AlphaFold structure could give us more information about how Trm140 might interact with DNA in the Cas9 R-loop. We took the ABE8e structure that was previously published (PDB ID: 6VPC) and replaced the TadA structure with Trm140, using structural data from other SAM-dependent methyltransferase enzymes to position the active site near the target nucleotide. Immediately, we observed a significant clash between the C-terminal residues of Trm140 and the Cas9 structure that would likely impede base editing (**Figure 5.6A**). These residues make two antiparallel β-sheets near the active site entrance of Trm140, and likely aid tRNA binding using the many basic residues found here.[138]

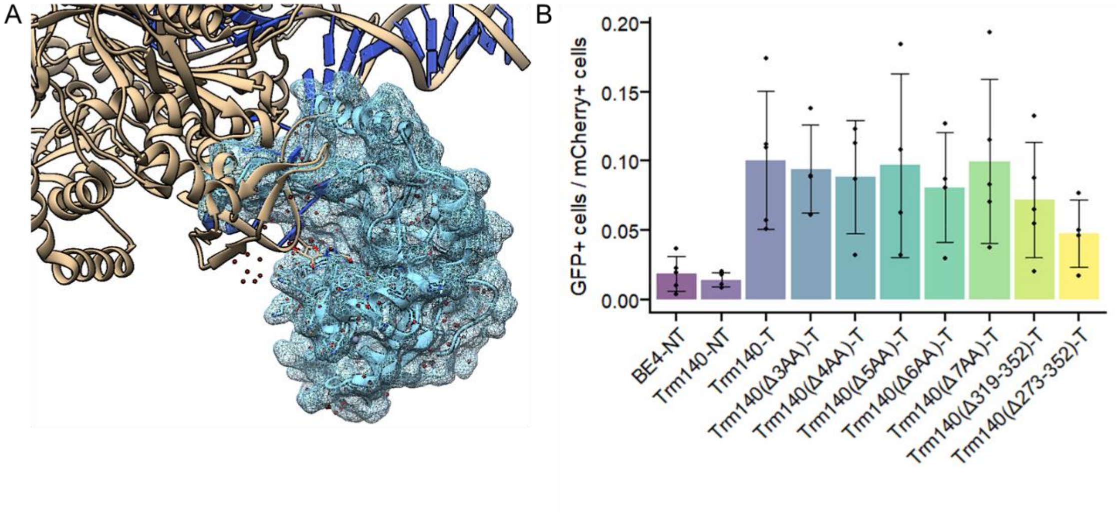


Figure 5.6: Trm140 active site clash deletions

(A) A Trm140 structure predicted by AlphaFold (blue), with the dCas9 structure from PDB ID: 6VPC. **(B)** The indicated base editors were transfected into HEK293T cells along with the Trm140-optimized fluorescent reporter and corresponding gRNA. Transfected cells were analyzed with flow cytometry to determine the percentage showing EGFP fluorescence, which is reported here. Values and error bars represent the mean and standard deviation for n = 5 biological replicates.

To see if we could reduce the clashes while maintaining the nucleic acid binding function of this domain, we began shortening these beta sheets, beginning with removing three amino acids from the end of each sheet (while leaving the connecting turn intact). We found that the three amino acid deletion did not significantly effect the EGFP turn-on efficiency of Trm140-BE (Trm140-BE: 0.10 ± 0.05% and Trm140(Δ3AA): 0.09 ± 0.03%), so we continued shortening these β-sheets by removing an additional amino acid from both sheet to make each variant (**Figure 5.6B**). Still, we did not observe any significant differences in EGFP turn-on by these variants. Finally, we created two variants in which the entire double-β-sheet domain was removed (Δ319-352 and Δ273-352). These variants showed EGFP turn-on rates that were lower, but not significantly lower, than those seen with the wild-type Trm140-BE (Trm140(Δ319-352): 0.07 ± 0.04% and Trm140(Δ273-352): 0.05 ± 0.02%). These results suggest that this tRNA-binding domain can be removed from Trm140 without impacting its ability to edit DNA.

ADAT and Tad1 deaminases

E. coli TadA is only one of many adenosine deaminases found in nature. Alongside homologs in other organisms, there are also deaminases that target adenosines within different types of RNA and within different sequence motifs. Similar to TadA, some of these enzymes may be able to accept DNA as a substrate as well as RNA, and may show base editing characteristics that are different from current ABEs. Three new base editors were developed containing adenosine deaminases that have not yet been used for base editing. ADAT2 and ADAT3 are both homologs of TadA and found in mammalian cells.[91,147] Structurally, they are significantly similar to TadA and form an ADAT3-ADAT2 heterodimer to edit the wobble base of target tRNA anticodons. This role is similar to TadA, however TadA edits only a single tRNA, while the ADAT3-ADAT2 complex recognizes many tRNAs as a target.[147] Further, endogenous ADAT3 localizes to the nucleus, so an ADAT2 base editor may be able to dimerize with cellular ADAT3 to perform base editing and shorten the length of the base editor sequence. The other enzymes to be tested as base editors are: mammalian ADAT1 and the yeast Tad1, both of which modify adenosine within a C$\underline{A}$ motif in tRNA while not modifying nearby T$\underline{A}$ motifs.[148,149] Neither of these enzymes are structurally similar to TadA, so they may have different characteristics from TadA-base editors that may make them advantageous in specific situations. ADAT3-ADAT2, ADAT2-alone, ADAT1, and Tad1 base editors were all cloned into mammalian expression vectors for testing these enzymes in fluorescent reporter assays.

ADAT1 and Tad1 have different sequence recognition properties than TadA, so a new dGFP(T64M) reporter was designed that contains a GCATGG motif at the target site. While only the CA motif has been shown to be important for activity, the other nucleotides in this motif mimic the natural tRNA editing context, which may improve editing efficiency. The effectiveness of the CA-T64M reporter was confirmed by co-transfection with ABE7.10 and a targeting or non-targeting gRNA. With the CA-T64M reporter, EGFP turn-on with ABE7.10 was observed in 63.3% of transfected cells with the targeting gRNA and 0.009% of transfected cells with a non-targeting gRNA; showing that the T64M effectively knocks out EGFP fluorescence and can be restored by A-to-G editing (**Figure 5.7A**). With ADAT1- and Tad1-base editors however, the observed EGFP turn-on efficiencies were lower than those with the non-targeting controls. In transfected cells, ADAT1-BE showed 0.008% EGFP turn-on with a targeting gRNA, and 0.02% EGFP turn-on with a non-targeting gRNA. The Tad1-BE showed 0.02% EGFP turn-on with a targeting gRNA, and 0.04% EGFP turn-on with a non-targeting gRNA (**Figure 5.7B**). These data suggest that these enzymes cannot act on a DNA substrate. Other explanations for these results might be poor protein stability or expression, so further optimization of these base editors may yet show DNA base editing by these enzymes.

As homologs of TadA, the ADAT3-ADAT2 and ADAT2 base editors could be targeted to the A111V and H182Y reporters developed for detecting ABE0.1 activity (**Figure 5.7C-D**). As expected, ABE7.10 displays high editing efficiency when targeting these reporters. Cells transfected with ABE7.10 and the TACG-A111V reporter showed 44.0% EGFP turn-on with a targeting gRNA, and 0.04% EGFP turn-on with a non-targeting gRNA. With the TAGG-H182Y reporter, transfected cells showed 65% EGFP turn-on with a targeting gRNA, and 0.01%

EGFP turn-on with a non-targeting gRNA. With the ADAT2- and ADAT3-ADAT2-base editors, EGFP turn-on efficiencies were again not significantly greater than the non-targeting controls.

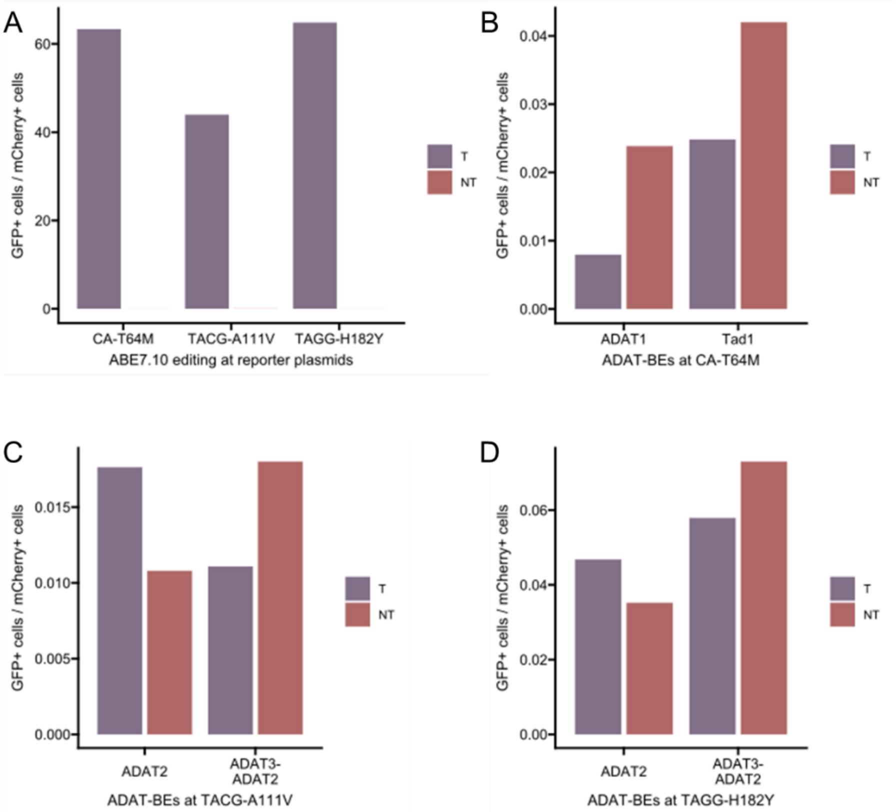


Figure 5.7: Editing of fluorescent reporters by ADAT and Tad1 base editors

(A-D) The indicated base editors were transfected into HEK293T cells along with the indicated fluorescent reporter and corresponding gRNA. Transfected cells were analyzed with flow cytometry to determine the percentage showing EGFP fluorescence, which is reported here. Values represent the outcome for n = 1 biological replicate. **(A)** ABE7.10 editing data with each of the reporters used in this figure. **(B)** ADAT1 and Tad1 base editors editing with the CA-T64M reporter. **(C)** ADAT2 and ADAT3-ADAT2 base editors editing with the TACG-A111V reporter. **(D)** ADAT2 and ADAT3-ADAT2 base editors editing with the TAGG-H182Y reporter.

With targeting gRNAs, the ADAT2-BE displayed 0.018% EGFP turn-on of the TACG-A111V reporter and 0.047% EGFP turn-on of the TAGG-H182Y reporter. These turn-on efficiencies were higher than the non-targeting controls for ADAT2-BE, which were 0.011% for the TACG-A111V reporter and 0.035% for the TAGG-H182Y reporter, but not higher than the non-targeting controls of the ADAT3-ADAT2-BE. This base editor with a non-targeting gRNA showed 0.018% turn-on with the TACG-A111V reporter, and 0.073% turn-on with the TAGG-H182Y reporter. With the targeting gRNA, ADAT3-ADAT2-BE displayed 0.012% turn-on with the TACG-A111V reporter and 0.058% turn-on with the TAGG-H182Y reporter. Again, these data show that ADAT2 base editors cannot perform DNA editing, however optimization of the sequences used to express these base editors may increase activity to a detectable level.

Trm140 and ADAT2 mixing reveals C-to-T editing by ADAT2

In *T. brucei* the tRNA editing activities of Trm140 and ADAT2/3 have been shown to be dependent on the expression of both proteins.[150,151] This may be due to the proteins interacting with each other, so we speculated that the presence of both proteins in a cell may improve base editing by Trm140. To quickly test this, Trm140-BE and each of the ADAT2-BEs were co-transfected into HEK293T cells along with the Y93H-Trm140 fluorescent reporter (**Figure 5.8**). As a baseline, each base editor was also transfected alone with the reporter so that any synergistic effects could be observed. No such effects were observed, with both the Trm140-BE + ADAT2-BE and Trm140-BE + ADAT3-ADAT2-BE combinations showing similar or lower EGFP turn-on rates as cells that were expressing Trm140-BE alone with the targeting gRNA. With the targeting gRNA, cells expressing Trm140-BE showed EGFP turn-on in 0.29 ± 0.04% of transfected cells, while cells expressing ADAT2-BE or ADAT3-ADAT2-BE along with Trm140-BE showed EGFP turn-on in 0.22% or 0.26% of transfected cells, respectively. Interestingly, EGFP turn-on was also observed in cells expressing the targeting gRNA with either ADAT2-BE or ADAT3-ADAT2-BE, but without Trm140-BE (**Figure 5.8B**). In these samples, ADAT2-BE displayed EGFP turn-on in 0.16 ± 0.026% of transfected cells, and ADAT3-ADAT2-BE displayed EGFP turn-on in 0.23 ± 0.00002% of transfected cell. In comparison, ADAT3-ADAT2-BE with a non-targeting gRNA displayed EGFP turn-on in 0.026% transfected cells. These results, along with the lack of editing observed in A-to-G reporters, suggests that these adenosine deaminases can perform C-to-T editing, but not A-to-G editing, in DNA. This finding has also been observed with these same proteins in *T. brucei*.[144]

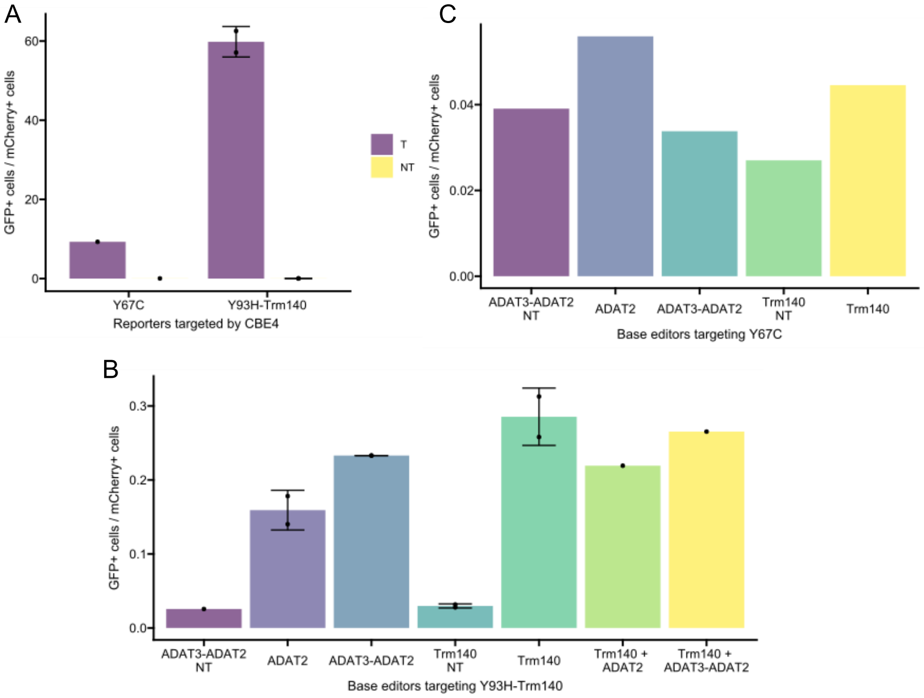


Figure 5.8: Editing of fluorescent reporters with a mixture of Trm140 and ADAT base editor plasmids

(A-C) The indicated base editors were transfected into HEK293T cells along with the indicated fluorescent reporter and corresponding gRNA. Transfected cells were analyzed with flow cytometry to determine the percentage showing EGFP fluorescence, which is reported here. **(A)** CBE4 editing data with each of the reporters used in this figure. n = 2. **(B)** Editing efficiencies of the indicated base editors with the Y93H, Trm140-optimized, C-to-T reporter. n = 1 for base editor mixtures and n = 2 for other samples. **(C)** Editing efficiencies of the indicated base editors with the Y67C C-to-T reporter. n = 1.

To gather more information on C-to-T editing in DNA by ADAT2, these base editors were transfected into HEK293T cells again, this time with the Y67C reporter which is not optimized for any specific protein (**Figure 5.8C**). ADAT2-BE showed the highest activity in this experiment with 0.056% of transfected cells displaying EGFP turn-on. Though, this value is barely larger than when ADAT3-ADAT2-BE was expressed with a non-targeting gRNA, which showed EGFP turn-on in 0.039% of transfected cells. Notably, cells expressing CBE4 and a targeting gRNA only showed Y67C correction in 9.29% of transfected cells, which is much lower compared to the 59.8 ± 3.8% EGFP turn-on rate observed with the Y93H reporter used previously. These results suggest that C-to-T editing in DNA by ADAT2 is inefficient, but requires more experiments to further understand.

ADAT2 variants

As a TadA homolog, its possible that the DNA editing ability of ADAT2 could be increased with the same mutations found to increase the DNA editing ability of TadA. The sequences and structures of TadA7.10 and ADAT2 were aligned and conserved residues in TadA7.10 that differed from the corresponding residue in ADAT2 were identified for mutation. Four promising mutations were identified (V103F, C125V, E128A, and F130T) and variants pf the ADAT3-ADAT2-BE were cloned containing either: one of these mutations or a combination of them (FVAT or VAT). Due to the previous observation of C-to-T editing activity by the ADAT3-ADAT2-BE, each variant was tested with both the A111V reporter (optimized for A-to-G editing by TadA) and the Y93H reporter (optimized for C-to-T editing by Trm140). ABE0.1m and CBE4 were each used as positive controls for their respective target reporter, with ABE0.1m showing 1.9% EGFP turn-on with the A111V reporter and CBE4 showing 25.8% EGFP turn-on with the Y93H reporter (**Figure 5.9A** and **5.9C**). Each was also used as a negative control for the other reporter, with ABE0.1m showing 0.1% EGFP turn-on with the Y93H reporter and CBE4 showing 0.05% EGFP turn-on with the A111V reporter. With the A111V (A-to-G) reporter none of the ADAT3-ADAT2-BEs showed turn-on rates that were above the non-targeting control, and were similar to the rate observed with CBE4 and a targeting gRNA (**Figure 5.9B**). With the Y93H (C-to-T) reporter, the ADAT3-ADAT2-BEs showed turn-on rates that were higher than the non-targeting control, however the ABE0.1m with a targeting gRNA also showed a similar turn-on rate, suggesting that the C-to-T editing observed with ADAT3-ADAT2-BE might be caused by a mechanism other than deamination by the ADAT3-ADAT2 complex (**Figure 5.9D**).

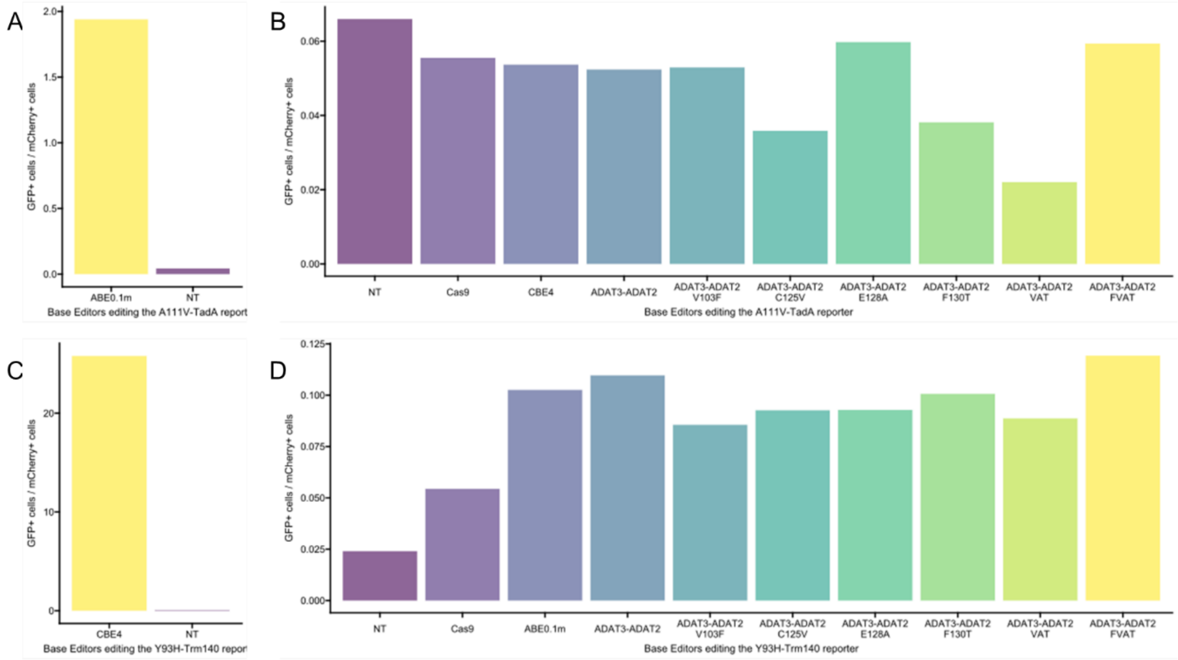


Figure 5.9: Editing of fluorescent reporters with ADAT2 base editor variants

(A-D) The indicated base editors were transfected into HEK293T cells along with the indicated fluorescent reporter and corresponding gRNA. Transfected cells were analyzed with flow cytometry to determine the percentage showing EGFP fluorescence, which is reported here. **(A-B)** Editing efficiencies of the indicated base editors with the A111V, TadA-optimized, A-to-G reporter. **(C-D)** Editing efficiencies of the indicated base editors with the Y93H, Trm140-optimized C-to-T reporter. Values are the outcome of n = 1 biological replicate.

dCas9 base editors

Uracil can be introduced spontaneously into DNA by hydrolytic deamination, which can result in spontaneous C-to-T mutation if the uracil is used as a template in DNA repair or replication.[152] During CBE development, it was discovered that nicking of the unedited strand during base editing drastically improves editing efficiency by recruiting DNA repair pathways to use the uracil intermediate as a template.[76] Based on the earlier observation of C-to-T editing by TadA and ADAT base editors, which deaminate adenosine, its possible that low levels of C-to-T editing could be detected with the Y93H reporter by DNA nicking alone. This would result in transfected cells displaying EGFP fluorescence with a targeting gRNA and any base editor, regardless of the attached DNA-modifying enzyme.

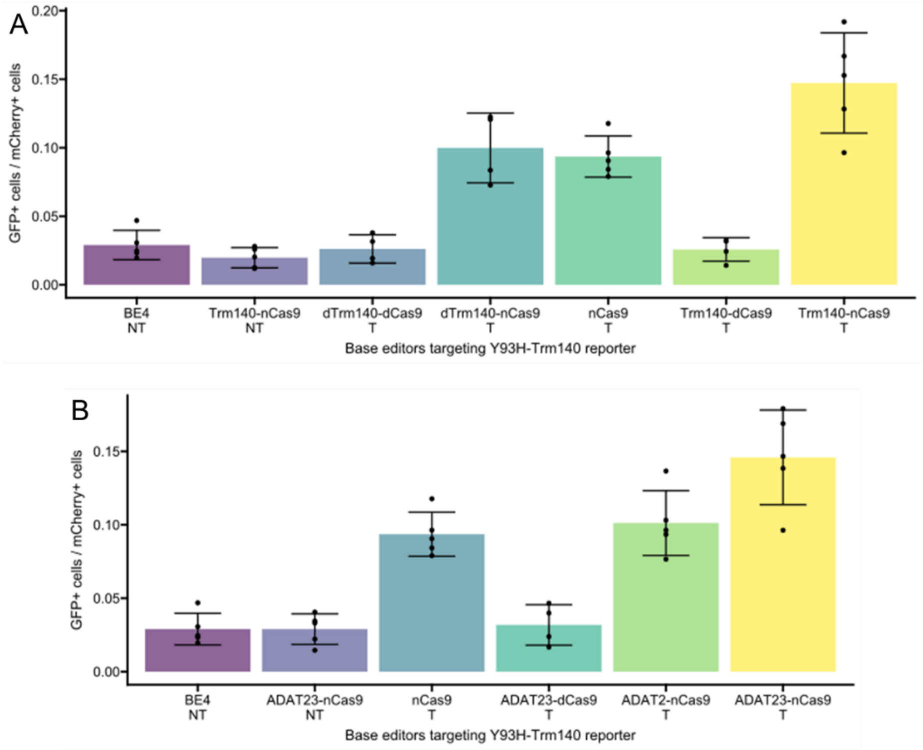


Figure 5.10: Editing of fluorescent reporters with ADAT2 base editor variants

(A-B) The indicated base editors were transfected into HEK293T cells along with the indicated fluorescent reporter and corresponding gRNA. Transfected cells were analyzed with flow cytometry to determine the percentage showing EGFP fluorescence, which is reported here. **(A)** Editing efficiencies of Trm140 base editors with the Y93H, Trm140-optimized C-to-T reporter. **(B)** Editing efficiencies of ADAT3-ADAT2 base editors with the Y93H, Trm140-optimized C-to-T reporter. Values and error bars are the means and standard deviation of n = 5 biological replicates.

To see if nicking activity was responsible for C-to-T editing observed with the Y93H reporter, an nCas9-alone plasmid was cloned, as well as dCas9 versions of Trm140-, ADAT2-, and ADAT3-ADAT2- base editors. These were transfected into HEK293T cells with the Y93H-Trm140 reporter and analyzed with flow cytometry to determine editing efficiency (**Figure 5.10**). The new nCas9 control showed a 3-fold higher turn-on rate than the non-targeting controls (nCas9 T: 0.094 ± 0.015% vs BE4 NT: 0.029 ± 0.011% and Trm140-nCas9 NT: 0.020 ± 0.007%), suggesting that nicking alone can cause significant levels of EGFP turn-on in C-to-T reporters. Likewise, dTrm140-nCas9 showed similar levels of editing as the nCas9 control, with 0.10 ± 0.025% of transfected cells also showing EGFP fluorescence. Despite these high levels of editing observed with nicking alone, Trm140-nCas9 did show a higher rate of turn-on, with 0.147 ± 0.037% of transfected cells showing EGFP fluorescence. Notably, in every replicate of this experiment Trm140-nCas9 displayed higher editing activity than both of the nicking controls, suggesting that Trm140 activity boosts the C-to-T editing observed with this reporter (**Figure 5.10A**).

The ADAT2 base editors displayed similar results in this experiment (**Figure 5.10B**). Cells expressing ADAT2-nCas9 showed EGFP turn-on in 0.10 ± 0.02% of transfected cells, which was very similar to the rate observed with nCas9 alone. Cells expressing ADAT23-nCas9 led to a marginal increase in editing, with 0.146 ± 0.032% of transfected cells showing EGFP fluorescence. This rate is similar to that of Trm140-nCas9, suggesting that the ADAT3-ADAT2 complex may also increase DNA editing efficiency.

The dCas9 versions of each base editors were also analyzed in this experiment to see if base editing could occur without nicking (**Figure 5.10**). However, both the Trm140-dCas9 and ADAT23-dCas9 base editors displayed similar activity levels as the non-targeting controls, with EGFP fluorescence observed in 0.026 ± 0.01% and 0.032 ± 0.014% of transfected cells, respectively. Collectively, these results show that Trm140 and the ADAT3-ADAT2 complex may be able to accept DNA as a substrate, but their base editing ability is dependent on the nicking of the unmodified DNA strand.

Conclusions for a Trm140 base editor

Base editors using methyltransferases like Trm140 face additional hurtles to enzymatic activity compared to the deaminases that are currently used for base editing. First is the requirement of co-factors used in the methyltransferase reaction. Every methyltransferase requires a separate molecule in the active site that serves as the methyl donor for the reaction. While hydrolytic deaminases like APOBEC1 and TadA are limited only be how efficiently they can catalyze the deamination reaction, Trm140 activity is further limited by the intracellular concentration of SAM. Second, m^3C is the target of DNA repair enzyme AlkB in *e. coli*. Three homologs of this enzyme have been identified in mammalian cells as well, with all three enzymes being reported to repair m^3C in DNA.[153] hABH2 specifically, is upregulated during the S-phase and localizes near the replication of fork.[153] This may explain the nicking-dependent activity of the Trm140-BE that we observed (figure **Trm140-dCas9**), as m^3C may be efficiently repaired during DNA replication.

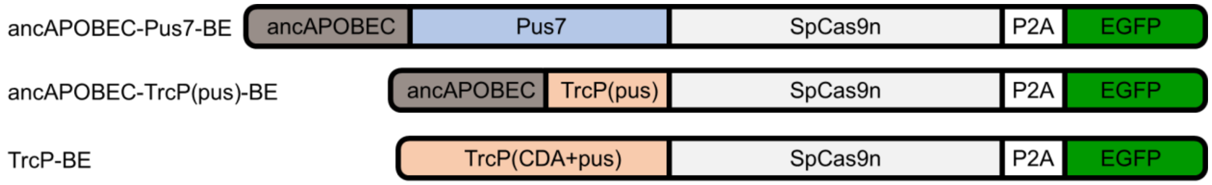
While the ABE was able to be developed by increasing the enzymatic activity of TadA, without any need to worry about protecting the inosine intermediate. Such a path may not be enough for a Trm140 base editor due to the SAM-dependent nature of its activity. It's likely that an evolved Trm140 base editor will require an additional strategy for protecting the m^3C intermediate from repair, similar to the UGI used for C-to-T editing with a U intermediate. Even so, the m^3C intermediate is preferable to U, as the demethylation reaction that repairs m^3C simply restores the original C instead of creating an abasic site (which is seen in U excision). This avoids the potential of creating C-to-non-T outcomes, allowing a Trm140-BE to make more precise edits than current CBEs.

Pseudouridine synthases

Pseudouracil is another intermediate that could be used for more precise C-to-T editing. Chemically, this base is the same as uracil and would likely have the same base pairing properties in DNA, but it is attached to the sugar phosphate backbone through a carbon-carbon bond instead of the usual glycosidic bond. Since the glycosidic bond is cleaved during uracil removal by UNG, psuedouracil is unable to be removed by UNG.[154] Since C-to-non-T editing by CBEs are a result of UNG activity, psuedouracil should prevent base excision during C-to-T editing. Furthermore, the isomerization of uracil to psuedouracil does not require co-factors, which should allow these base editors to quickly catalyze the reactions.

Two psuedouridine synthases were selected to test as base editors: Pus7 and TrcP. Pus7 is enzyme that accepts many different RNAs as a substrate. It's activity has been observed in UGUAR motifs in many different single-stranded RNA.[155] TrcP is more similar to the other RNA-modifying enzymes that have been tested, in that it is a tRNA-modifying enzyme found in *V. cholerae* that introduces psuedouracil into the anticodon loop of tRNA^{Tyr}.[156] Interestingly, TrcP contains two catalytic domains, with one deaminating a target cytidine to uracil before passing the substrate to the other domain for isomerization to psuedouracil.[156]These two domains are split by a long helical domain that binds to the target tRNA and transfers it from the cytidine deaminase domain (CDA) to psuedouracil synthase domain (PUS). While the sequence recognition properties of TrcP have not been directly studied, the anticodon loop contains a CUGUAAA sequence.

Both of these psuedouracil synthases were cloned into mammalian expression vectors along with nCas9 to create base editors (**Figure 5.11**). Since Pus7 does not contain a deaminase domain to make the initial C-to-U edit, it was combined with the ancAPOBEC1 used in ancBE4max to make ancAPOBEC1-Pus7-nCas9.[157] A similar base editor was developed with the PUS domain from TrcP, which was termed ancAPOBEC1-TrcP(PUS)-nCas9. Since it is not clear if the TrcP(PUS) domain can function separately from the CDA domain, a second TrcP base editor was developed with both of the TrcP catalytic domains, termed TrcP-nCas9.



Editor	Site	Full protospacer
Pus7	FANCF	GCTG$\underline{C}_5$AGAAGGGATTCCATG
		GTGCTG$\underline{C}_7$AGAAGGGATTCCA
	HEK site 3	GTG$\underline{C}_4$AGGAGCTGCACATACT
	HEK293T site 2	AAGTG$\underline{C}_6$AGAATATCTGATGA
TrcP	PSMB2	GT$\underline{C}_3$TGTAAATTGCATTTTCA
	PTBP2	GGGCA$\underline{C}_6$TGTGAAAGCATTTA
	RNF2	AT$\underline{C}_3$TGTAAAGTCAGTGAGGC
		ATAAT$\underline{C}_6$TGTAAAGTCAGTGA
	SAP30BP	CTT$\underline{C}_4$TATAAAAGAAACTGAC

Figure 5.11: Pseudouracil base editors and gDNA target sties

The architecture of the Pus7 and TrcP base editors, and the gDNA sites targeted with each base editor. Target C's are underlined, while nucleotides recognized by the specified editor are shown in blue.

To measure the editing outcomes of these new C-to-T base editors, each one was targeted to edit gDNA sequences containing the motifs present in their natural RNA substrates. Four targets were selected for the TrcP base editors (PSMB2, PTBP2, RNF2, and SAP30BP) and three targets were selected for the Pus7 base editor (FANCF, HEK293T site 2, and HEK site 3). When applicable, a second gRNA was designed for each site to measure how a shifted editing window affects the editing efficiency. In total, nine gRNAs targeting seven genomic sites were tested (**Figure 5.11**). Following transfection with the base editor and sgRNA plasmids, HEK293T cells were lysed and the gDNA targets were amplified for targeted amplicon sequencing by high-throughput sequencing. If the psuedouracil base editors are able to efficiently isomerize the uracil intermediate to psuedouracil, then a drop in C-to-non-T editing should be observed when compared to the BE3b control.

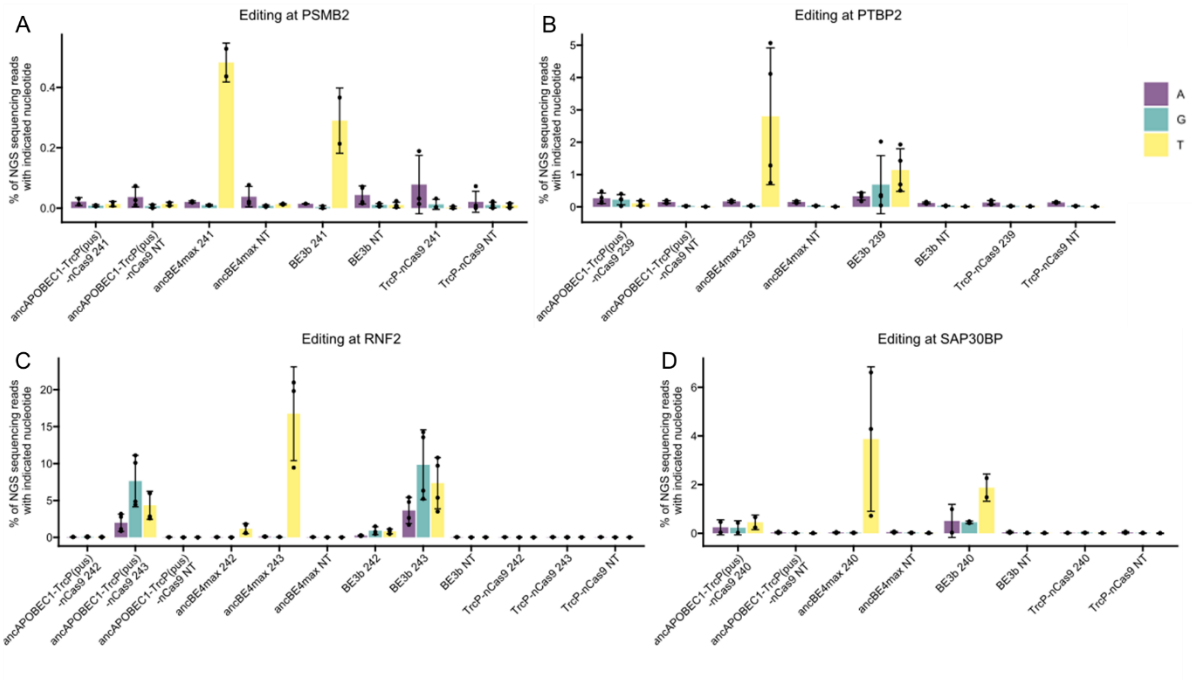


Figure 5.12: TrcP base editing at gDNA target sites

HEK293T cells were transfected with plasmids encoding the indicated base editor and gRNA. After three days, cells were lysed and the target site was amplified for high throughput targeted amplicon sequencing. Shown are the percentage of each nucleotide observed at the target base, as indicated in **Figure 5.11**. **(A)** Editing at the PSMB2 target. **(B)** Editing at the PTBP2 target. **(C)** Editing at the RNF2 target. **(D)** Editing at the SAP30BP target. Values and error bars are the means and standard deviation of n = 3 biological replicates.

At each of the TrcP target sites, low levels of editing were observed (**Figure 5.12**). The highest editing was observed at the RNF2 site, with ancBE4max showing 16.7 ± 6.4 C-to-T editing and 0.16 ± 0.06% combined C-to-non-T editing (**Figure 5.12C**). BE3b showed significant levels of C-to-non-T editing at this site, with 9.8 ± 4.7% C-to-G editing and 3.6 ± 1.8% C-to-A editing compared to 7.3 ± 3.5% C-to-T editing. However, a similar ratio of outcomes were observed with ancAPOBEC1-TrcP(pus)-nCas9 but with a lower number of overall edits, with 7.62 ± 3.47% C-to-G editing and 1.98 ± 1.17% C-to-A editing, compared to 4.35 ± 1.9% C-to-T editing. This shows that the TrcP(Pus) domain is likely not interacting with the uracil intermediate, but in extending the linker length between ancAPOBEC1 and the nCas9, the overall editing efficiency is decreased. When looking at the TrcP-nCas9 editor, which uses the TrcP CDA instead of ancAPOBEC1, the levels of base editing observed were not significantly different from the non-targeting controls. These data show that both domains of TrcP are likely not efficient enough at DNA base editing to be detected with this method.

Similarly low levels of editing were observed at the Pus7 target sites, with the highest editing observed at the HEK293T site 2 (**Figure 5.13B**). At this site, ancBE4max showed 26.9 ± 10.5% C-to-T editing and combined 0.1 ± 0.06% C-to-non-T editing. BE3b showed primarily C-to-T editing at this site, with 15.9 ± 2.8% C-to-T editing, 4.37 ± 0.37% C-to-G editing, and 2.83 ± 0.16% C-to-A editing. The ancAPOBEC1-Pus7-nCas9 showed reduced editing compared to BE3b, with 9.11 ± 4.67% C-to-T editing, 2.86 ± 2.0% C-to-G editing, and 1.3 ± 0.9% C-to-A editing. Both editors showed a similar ratio of outcomes, showing that Pus7 also does not significantly interact with the uracil intermediate, and decreases editing efficiency by extending the linker length.

The direct sequencing of a gDNA target allows for direct observation of the base editing outcome. As shown previously, this method is not as sensitive for low levels of editing as a fluorescent reporter. One step that may improve this method however, is pairing base editor expression with GFP expression, so that transfected cell can be enriched from the cell population. In the current experiment this step was difficult to the large combination of base editors and targeting gRNAs. Since base editing efficiency was highest at the RNF2 and HEK293T site targets, repeating the experiment with only these gRNAs would reduce the number of samples to a manageable amount for enrichment. Employing an optimized fluorescent reporter for each base editor may also improve the detection of base editing, however additional care must be taken to ensure that C-to-T edits can discerned from other editing outcomes and compared.

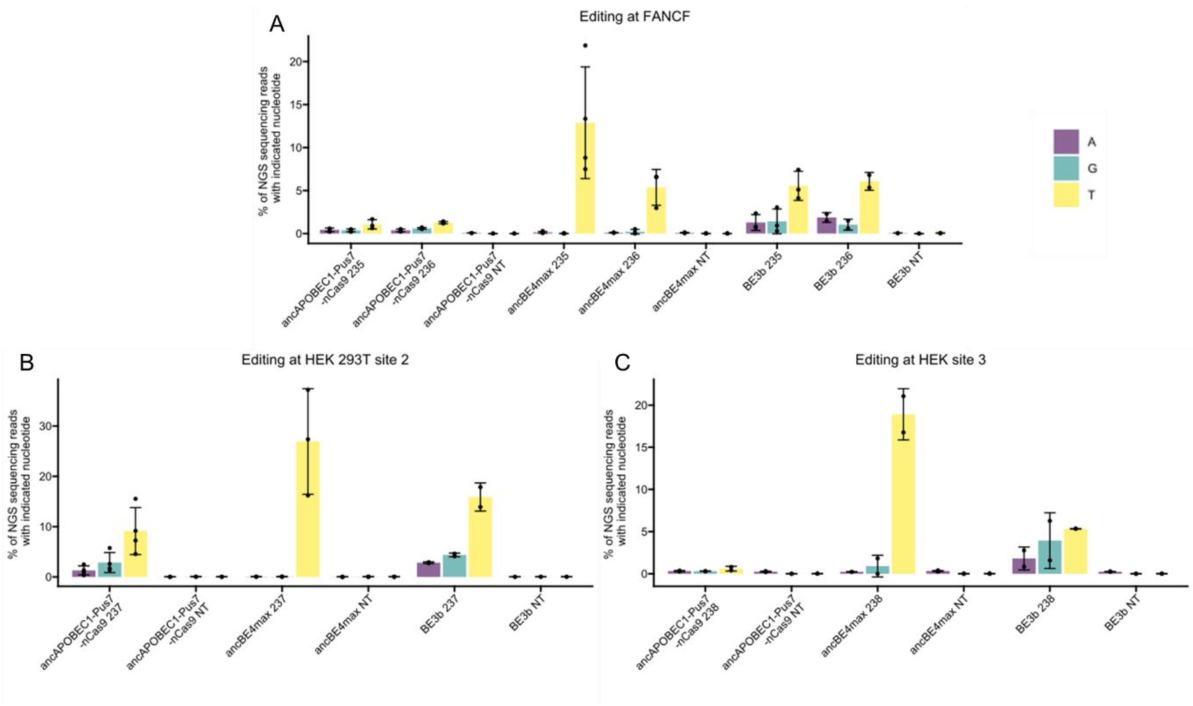


Figure 5.13: Pus7 base editing at gDNA target sites

HEK293T cells were transfected with plasmids encoding the indicated base editor and gRNA. After three days, cells were lysed and the target site was amplified for high throughput targeted amplicon sequencing. Shown are the percentage of each nucleotide observed at the target base, as indicated in **Figure 5.11**. **(A)** Editing at the FANCF target. **(B)** Editing at the HEK 293T site 2 target. **(C)** Editing at the HEK site 3 target. Values and error bars are the means and standard deviation of n = 3 biological replicates.

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. All variations of the mCherry-P2A-dGFP reporter plasmid were cloned by site directed mutagenesis[103]. Trm140 truncation variants were cloned with USER cloning, with primers that annealed within Trm140 and formed a junction that excluded the deleted sequence. Variants that were discovered earlier by directed evolution were amplified from the bacterial expression plasmid with primers that anneal to the base editing enzyme and add BsaI digestible overhangs. This amplicon was purified using a QIAquick PCR purification column (QIAGEN # 28104), then cloned into a mammalian expression plasmid by following the BsaI-HFv2 Golden Gate Assembly protocol from NEB. A similar destination plasmid was used to clone mammalian cell gRNA plasmids as previously described.[103]

The genes used to develop new base editor plasmids were either ordered as gBlocks from IDT (*Tad1, ADAT2, Pus7,* and *TrcP*), or amplified from HEK293T cDNA (*ADAT1* and *ADAT3*). cDNA was obtained by lysing HEK293T cells with 300ul RNA lysis buffer and extracting RNA with either a Zymo Quick-RNA Miniprep kit (Zymo Research # R1054) or Qiagen RNeasy mini kit (Qiagen # 74004) following the manufacturers' instructions. RNA was reverse transcribed to produce cDNA using SuperScript III First-Strand Synthesis System (Invitrogen #18080-051) following the manufacturer's instructions. All genes were amplified using primers that added BsaI digestible overhangs and cloned into mammalian expression plasmids using the BsaI-HFv2 Golden Gate Assembly protocol.

All plasmids used for transfection were harvested using a ZymoPURE^{TM} II Plasmid Midiprep kit (Zymo Research #D4200), following the manufacturer's protocol.

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 the Trm140 and ADAT2 combination experiments, DNA mixtures contained: 375 ng of each base editor plasmids (750 ng total), 500 ng of mCherry-P2A-EGFP reporter plasmid, and 250 ng of gRNA plasmid. For gDNA editing experiments, DNA mixtures were prepared with: 750 ng of base editor plasmid and 250 ng 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.

Flow cytometry

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)[104]. 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

High-throughput amplicon sequencing of genomic DNA

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. 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. 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 Quant-IT^{TM} dsDNA Assay Kit, high sensitivity (ThermoFisher Scientific #Q33120) on a Qubit fluorometer. Samples were then sequenced on an Illumina MiniSeq according to the manufacturer's protocol.

Analysis of Illumina HTS was performed with CRISPResso2[75]. 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). C•G to T•A edits were calculated using the nucleotide frequency at the target site, by dividing the number of T reads by the total reads.

Chapter 6 Designing a directed evolution technique for base editors in mammalian cells

Introduction

Currently, the directed evolution of new base editors has been done entirely in bacteria, which possess different pathways and proteins for DNA replication and repair than mammalian cells. This poses a problem for developing new base editors to be used in mammalian cells, as a bacterial selection or screen will optimize the base editor for bacterial use. For example, the A142N mutation in TadA was highly enriched during round 4 of the ABE directed evolution but was found to decrease editing at several genomic targets in mammalian cells, suggesting this mutation has a different effect depending on the cell type. Similarly, the use of bacterial selection techniques may fail to enrich beneficial mutations that only have an effect in mammalian cells. To ensure that evolved base editors are best adapted to working in mammalian cells, we set out to develop a mammalian cell directed evolution technique (**Figure 6.1**).

Our observations of ABE0.1 editing showed that base editors using wild-type RNA-editing enzymes will likely have very low DNA-editing efficiency in mammalian cells. This led us to consider using a screening method instead of a selection, because a low survival rate would likely inhibit surviving cells growing well following a selection. Further, our fluorescence-based reporter of base editing could be easily adapted for screening.
Since each base editor variant will be screened at the cellular level, another consideration for our screen is ensuring each cell contains a single variant. A lentiviral transduction at a low multiplicity of infection (MOI) provides an effective method for integrating a single copy of a base editor into the cell's genome. Further, integrations are likely to happen in open chromatin regions, so base editors should be expressed similarly across all treated cells (integration often happens in actively transcribed genes).[158] To develop new cell lines with integrated ABEs, a lentiviral transfer plasmid was developed (**Figure 6.2A**). The integration

insert contains an SFFV promoter driving expression of a TadA-nCas9(NG)-P2A-puroR, where integration and expression of the ABE can be monitored by the puromycin resistance conferred by the puroR gene. Expression of this insert for virus packaging is controlled by a hybrid CMV/LTR promoter, which shows increased virus production relative to an LTR sequence alone.

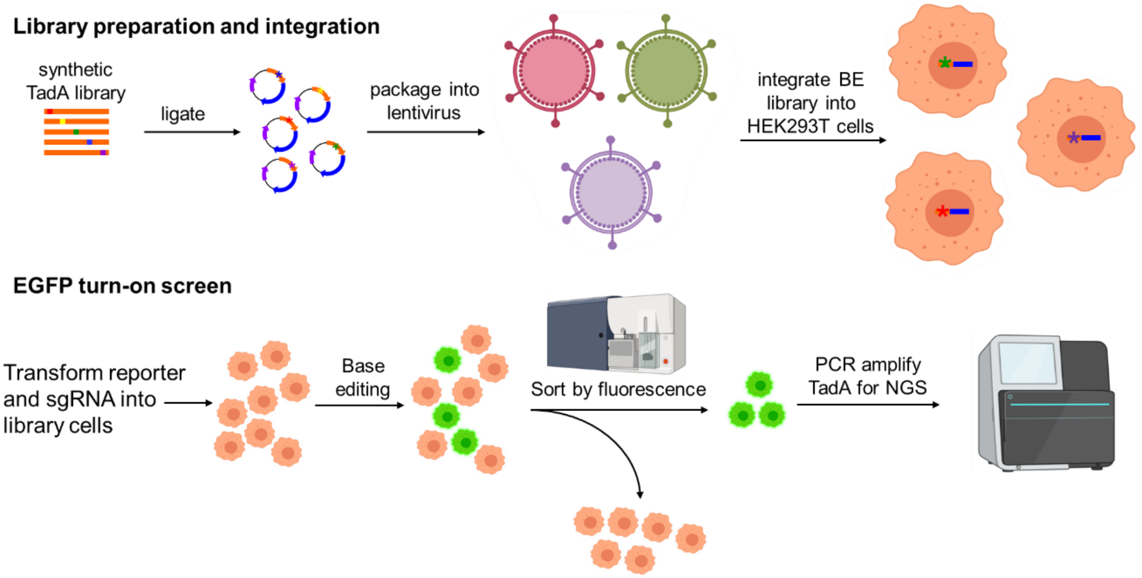


Figure 6.1: Protocol for directed evolution of base editors in mammalian cells

In this work TadA is used as a model to ensure that this directed evolution scheme is feasible. First, a synthetic library is ligated into a lentivirus transfer vector and packaged into lentivirus. HEK293T cells are transduced with the base editor lentivirus at a concentration that ensures a single integration per cell. Base editing is then detected in the transduced cells using an EGFP turn-on screen. Cells displaying EGFP fluorescence are sorted using FACS, and the TadA gene in this population is sequenced using high-throughput sequencing. Beneficial mutations are determined by comparison to an unsorted population, which shows the mutations that were enriched by the screen.

This chapter describes the optimizations made to each step of the directed evolution technique, using ABE0.1 and ABE1.1 as models to test the feasibility of the technique. Initially, different promoters were tested for ABE expression to ensure that base editing of a fluorescent reporter could be detected when using an integrated ABE. Next, cells with integrated ABE0.1 or ABE1.1 were mixed and transfected with a fluorescent reporter. Cells showing editing of the reporter were sorted and analyzed to see if cells containing the integrated ABE1.1 could be enriched using this scheme. Finally, the virus production protocol was optimized to ensure that a large number of base editor variants could be efficiently packaged into virus for integration.

Comparison of editing activity with different promoters

ABE0.1 and ABE1.1 show drastically different DNA editing activity with the previously described TACG fluorescent reporter, providing a convenient method to screen ABE variants. To check the activity of these integrated ABEs, cell lines were developed by integrating ABE1.1m using a low MOI transduction and were transfected with the TACG-reporter and corresponding gRNA plasmids to compare editing efficiencies using flow cytometry (**Figure 6.2B**). In this initial experiment, ABE1.1m caused EGFP turn-on in 0.32% of transfected cells. When compared to previous experiments in which ABE1.1m was transiently expressed, the integration of ABE1.1m caused a ~108-fold decrease in turn-on efficiency. With the non-targeting gRNA control, EGFP turn-on was observed in 0.17% of transfected cells. This experiment showed that an integrated ABE shows activity in the fluorescence reporter assay, but with dramatically lower efficiency.

Different promoters can drive different levels of expression in cells, with the CMV and EF1α promoters showing higher expression relative to the SFFV promoter.[159] Since the CMV promoter was being used to express the integration transcript for virus packaging, new lentiviral transfer plasmids were designed containing the EF1α promoter as a replacement for the SFFV promoter (**Figure 6.2A**). Use of this promoter often includes an additional 'intron a' sequence just downstream of the promoter that boosts expression. However, lentivirus packaging efficiency is impacted by large integration sequences, so an expression boost by the 1kb intron might come at the cost of reduced virus packaging. To observe the effect of the intron sequence on ABE editing activity, two integration plasmid vectors were developed, one with EF1α promoter and downstream intron and the other with just the core EF1α promoter.

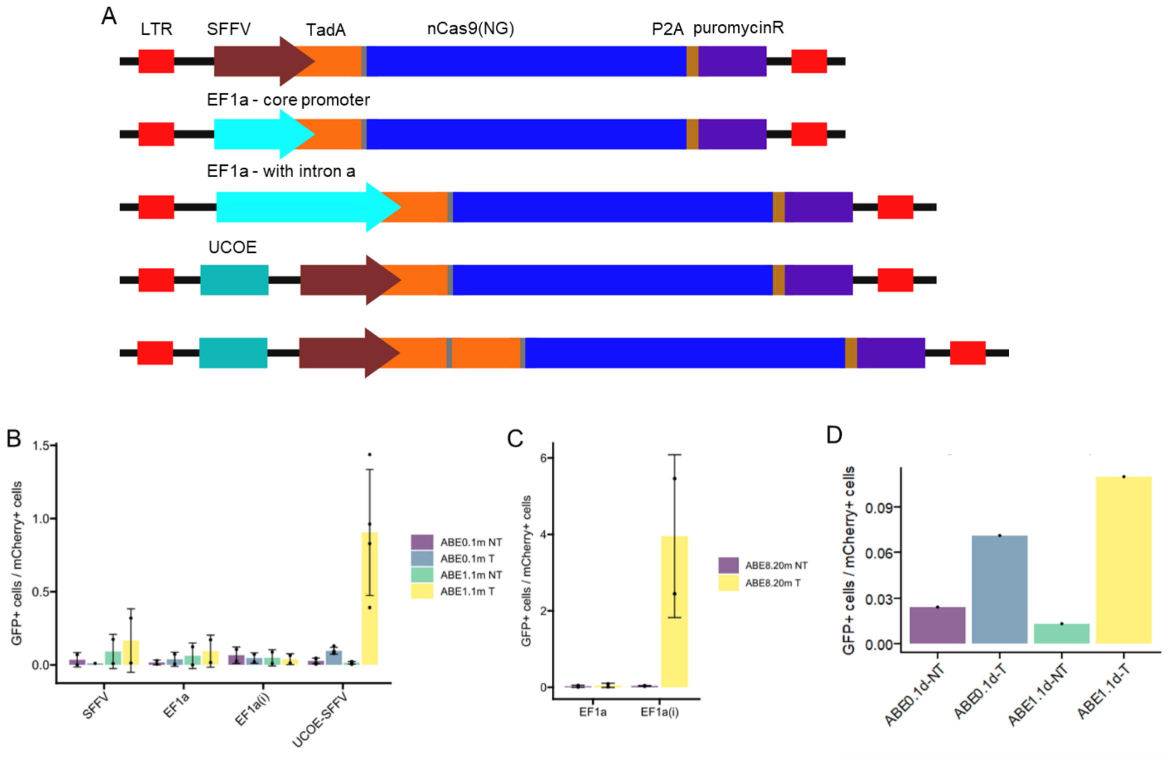


Figure 6.2: Base editor optimization for EGFP turn-on

(A) Integration sequences tested in this work. **(B-D)** The indicated base editors were integrated into HEK293T cells to produce new cell lines. After a puromycin selection, cell were transfected with plasmids expressing the TACG-A111V fluorescent reporter and corresponding gRNA. Transfected cells were analyzed with flow cytometry to determine the percentage showing EGFP fluorescence, which is reported here. **(B)** Cell lines using the indicated promoter to express either ABE0.1 and ABE1.1 were tested for EGFP turn-on efficiency. **(C)** Cell lines using the indicated promoter to express ABE8.20 were tested for EGFP turn-on efficiency. **(D)** Cell lines using the UCOE-SFFV promoter to express either ABE0.1d or ABE1.1d were tested for EGFP turn-on efficiency. Values and error bars indicate the means and standard deviation for biological replicates: n = 4 for UCOE-SFFV samples and n = 2 for all other samples in **(B-C)**, and n = 1 for **(D)**.

ABE0.1, ABE1.1, or ABE8.20 were cloned into both of these EF1α promoter backbones to create six integration plasmids, and were used to produce lentivirus and cell lines expressing an integrated ABE through the EF1α promoter. Transfecting these cell lines with the fluorescent reporter plasmids, EGFP turn-on was observed only in the EF1α(i)-ABE8.20 cell line, with 4.0 ± 2.1% of transfected cells showing EGFP fluorescence (**Figure 6.2B** and **6.2C**). In previous experiments using the SFFV promoter, cell lines expressing integrated ABE8.20 displayed low cell viability and did not survive long following transduction, likely due to high levels of off-target editing occurring within essential genes. The observation of cell survival with ABE8.20 expression, along with the lack of editing observed with ABE1.1, suggests that the EF1α promoter has lower expression than the SFFV promoter.

Though virus integration occurs in open chromatin regions, it's possible that DNA methylation following integration may reduce base editor expression. DNA sequences such as the ubiquitous chromatin opening element (UCOE) have been used to avoid the silencing effects of DNA methylation and maintain high levels of expression from nearby sequences.[160] To test if chromatin silencing was affecting the activity of integrated ABEs, a new lentiviral transfer vector was cloned with a UCOE sequence upstream of the SFFV promoter. ABE0.1m or ABE1.1m were cloned into this new vector and the resulting plasmids were used to produce ABE-integrated HEK293T cell lines. Testing the UCOE-containing ABE0.1m- and ABE1.1m-cell lines with the fluorescence reporter assay showed EGFP turn-on in 0.1 ± 0.02% and 0.9 ± 0.4% of transfected cells, respectively (**Figure 6.2B**). The non-targeting controls using these same cell lines showed EGFP turn-on in 0.03 ± 0.02% and 0.014 ± 0.009% of transfected cells, for ABE0.1m and ABE1.1m respectively. While both turn-on rates are ~38-fold lower than the transiently expressed counterparts (found in **Figure 3.8A** and **3.8D**), they show that the UCOE helps boost the activity of integrated ABE.

Integrated ABE dimers

TadA acts as a natural homodimer to deaminate tRNA.[71] However, ABE0.1m has been shown to edit DNA with a similar efficiency as an ABE0.1 using a TadA homodimer. This editing is likely driven by *in-trans* dimerization of two ABE molecules, facilitated by high expression from transient plasmids. It is possible that an integrated ABE0.1m is expressed at significantly lower levels, making *in-trans* dimerization more difficult and lower editing efficiency. To test if an *in-cis* TadA dimer would improve editing by integrated ABE, an additional wtTadA gene was cloned into the UCOE-SFFV transfer plasmids containing ABE0.1 and ABE1.1 (making ABE0.1d and ABE1.1d). New cell lines were created using these plasmids and transfected with the TACG-fluorescent reporter plasmids. Using these ABE0.1d and ABE1.1d cell lines, EGFP turn-on was observed in 0.07% and 0.11% of transfected cells, respectively (**Figure 6.2D**). Each of these values were lower than the editing observed with their monomeric counterparts using the same vector backbone, though more replicates are needed to confirm this observation.

ABE1.1 enrichment

The feasibility of a directed evolution method is dependent on how reliably beneficial mutations can be enriched from a population of variants. Since the D108N mutation in ABE1.1 increased ABE activity considerably, this mutation should be enriched by the fluorescent reporter screen from a mixture of ABE0.1m and ABE1.1m cells. A 1:1 mixture of these cell lines was plated and transfected with the fluorescent reporter plasmids (**Figure 6.3A**). Using FACS, cells were sorted into three aliquots: 1) cells that passed the scatter gates, 2) cells that were transfected (displayed mCherry fluorescence), and 3) cells that were edited (displayed high mCherry and any EGFP fluorescence). These cells were lysed and a portion of the TadA gene was amplified for HTS. All three aliquots were analyzed for enrichment of the D108N mutation: singlets showed 45.5%, transfected cells showed 52.7%, and edited cells showed 80.9% (**Figure 6.3B**). This shows that beneficial mutations can be enriched with this screen, with the D108N mutation showing a 1.5-fold enrichment from the transfected cells to the edited cells. This level of enrichment may not be high enough to reliably screen for mutations that increase DNA editing. However, the gates used to sort EGFP+ cells were very lenient in order to collect enough cells for this initial analysis. Since cells can display higher EGFP intensity when expressing more efficient base editor variants, a higher enrichment of beneficial mutations can likely be achieved by gating for cells with higher EGFP intensity (i.e. the “sorts” gate in **Figure 6.3C**).

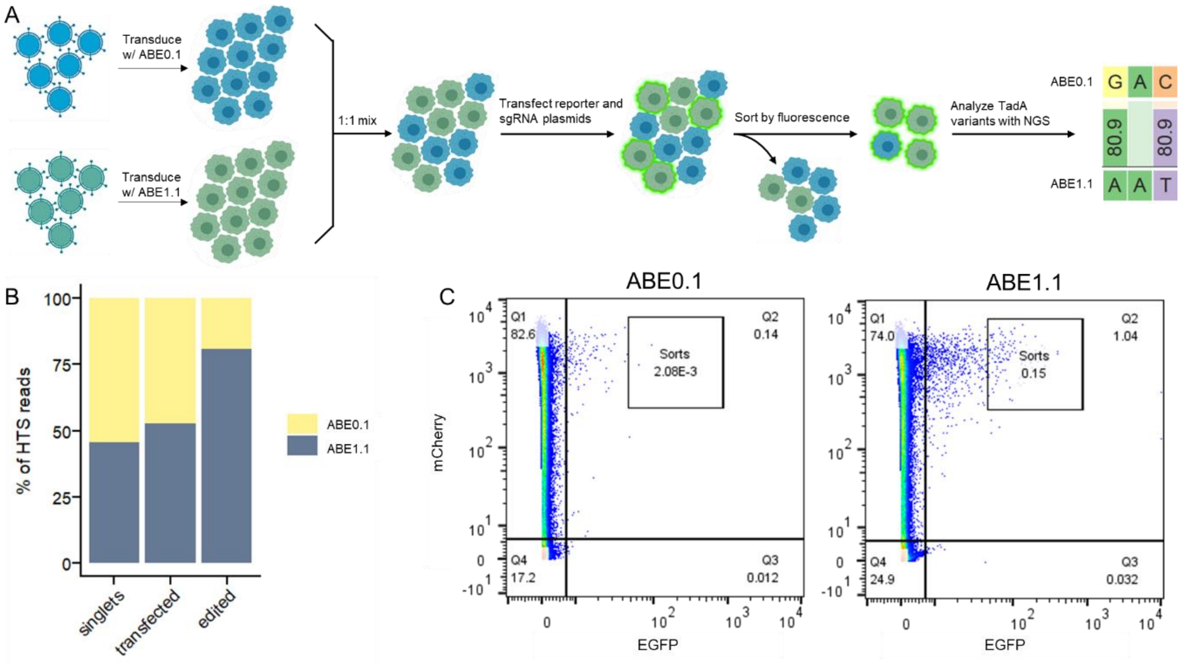


Figure 6.3: Mutation enrichment from a mixture of ABE0.1 and ABE1.1

(A) Protocol for mixing ABE0.1 and ABE1.1 cell lines and using the EGFP turn-on screen to enrich for ABE1.1 cells. **(B)** The percentage of cells containing the indicated ABE variant, based on HTS reads. **(C)** Representative flow cytometry graphs from cells containing either integrated ABE0.1 or integrated ABE1.1.

Transduction / virus production optimizations

One issue with using a lentiviral transduction method with base editors is the large size of the integration sequence. During virus production, the number of functional virus units has been observed to decrease as the size of the integration sequence increases.[161,162] To estimate the virus titer, the ABE1.1m integration plasmid and lentiviral packaging plasmids were transfected into HEK293T cells for virus production. The produced virus was used in increasing volumes to transduce wells of HEK293T cells. Following viral integration, wells were split into paired samples, where puromycin was added to one sample for the selection of integrated cells, while the paired well did not receive puromycin. Following puromycin selection, cells were counted in each of the sample and used to quantify the number of cells that were transduced. This titer quantification process shows the number of functional virus molecules in the solution based on the number of cells that are infected (**Figure 6.4A**). This initial virus titer experiment produced 1 ml of virus media with a viral titer of 1,175 TU/ml. For comparison, there are 3,320 possible single mutations at the amino acid level of TadA, with 10-times this amount of virus titer units required to ensure that all of the variants are present. The virus produced in this experiment reaches about 3.5% of this number, showing that making optimizations to the virus production protocol will facilitate the screening of many base editor variants in the future (**Figure 6.4B**).

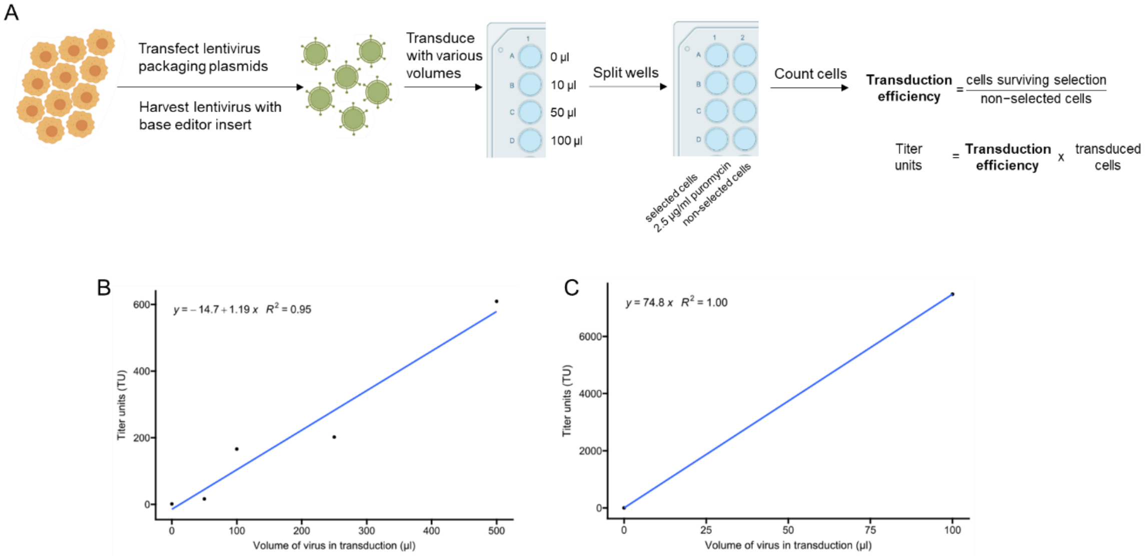


Figure 6.4: Initial transductions

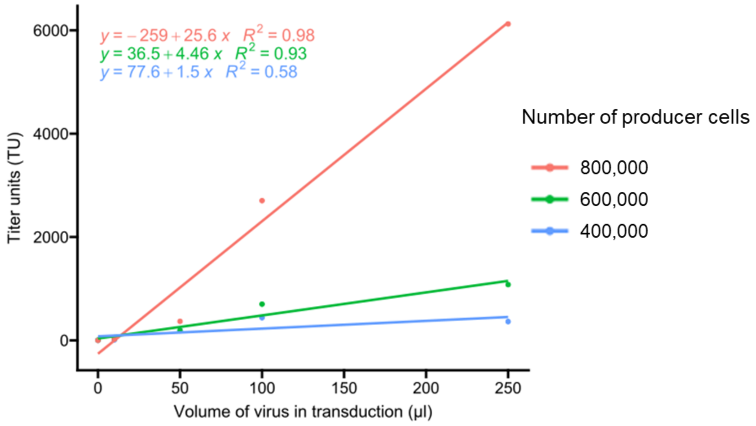
(A) Protocol to produce and quantify functional virus. HEK293T cells were transfected with three lentiviral packaging plasmids and grown for three days. The media is harvested and a range of volumes are used to transduce HEK293T cells. The transduced cells are split into two wells, with one of the wells being grown normally and the other being treated with puromycin to select for the cells expressing integrated base editors. The cells are counted and those that survive selection are compared to the non-selected cells to determine the transduction efficiency. This value is multiplied with the number of cells originally used in the transduction to determine the number of viral particles used in the transduction. The range of volumes used in the transduction are used to create a standard line to determine the total virus produced (as seen in **(B)** and **(C)**). **(B)** 1 ml of virus media was produced using a single well of a 6-well plate. **(C)** 6 ml of virus media was produced using 6 wells of a 6-well plate. This media was concentrated 10-fold, then used for serial transductions.

The virus production was repeated, this time using every well in a 6-well plate for virus production. The resulting virus media was concentrated 10-fold before repeating the virus titer experiment (**Figure 6.4C**). This scaled-up experiment showed a much higher virus titer of 74,800 TU/ml (though only 600 µl of virus media was produced), however only the lowest volume transduced well survived selection. This showed that the integrated ABE1.1 might be toxic to cells in high amounts. It also showed that there was room for optimization of the virus production method to consistently create the high virus titer required for the screen.

Optimizing producer cell density

To further optimize the virus production, the density of producer HEK293T cells was varied between ~40,000 and ~80,000 cells/cm^2. Virus media was harvested from each sample and used to transduce cells to determine the viral titer of each. The sample with 400,000 producer cells serves as the baseline sample, as previous experiments used this many cells, and produced ~3,528 functional virus molecules in this experiment. Both the 600,000 and 800,000 producer cell samples produced more virus than the baseline, producing ~10,295 and ~50,941 virus molecules, respectively. This experiment showed that virus production could be increased by transfecting a higher density of cells (**Figure 6.5**).

In order to integrate enough cells with base editor variants, transduction will have to be performed in large flasks. All transduction experiments up to this point had been performed in 48-well plates with a surface area of 1.1 cm^2. To see if the transduction efficiency from these experiments can be maintained at a higher scale, a virus titering experiment was set up and performed alongside transducing cells in a T75 flask (**Figure 6.6**). Lentivirus was produced and harvested using the optimized producer cell density. Increasing volumes of this virus media was used to transduce a set of wells, as in other experiments, to measure the viral titer. The T75 flask was transduced using a volume-to-cell ratio of 1 µl per 100 cells, which would match 400 µl on the trend line. The set of smaller scale transductions revealed a virus titer of ~2,841 TU/ml. With 400 µl volume of virus at this scale estimated to have ~ 1,136 virus molecules. The actual T75 showed ~22,634 infections for 1 x 10^6 cells, which would scale down to ~905 infections for 4 x 10^4 cells at the 48-well scale (**Figure 6.6B**). While the efficiency of the T75 transduction was lower than expected, this experiment showed that transductions can be scaled up to a T75 flask with an efficiency that matches a 48-well plate.



Producer cells	Total virus produced
800,000	~50,941
600,000	~10,295
400,000	~3,528

Figure 6.5: Producer cell density optimizations

Three samples of virus were produced in parallel, with each using a single well of a 6-well plate and the indicated number of producer cells. Transduction were performed to measure the number of functional viral particles in each sample, and the outcomes and standard lines were graphed. The total number of virus produced with each number of cells is listed.

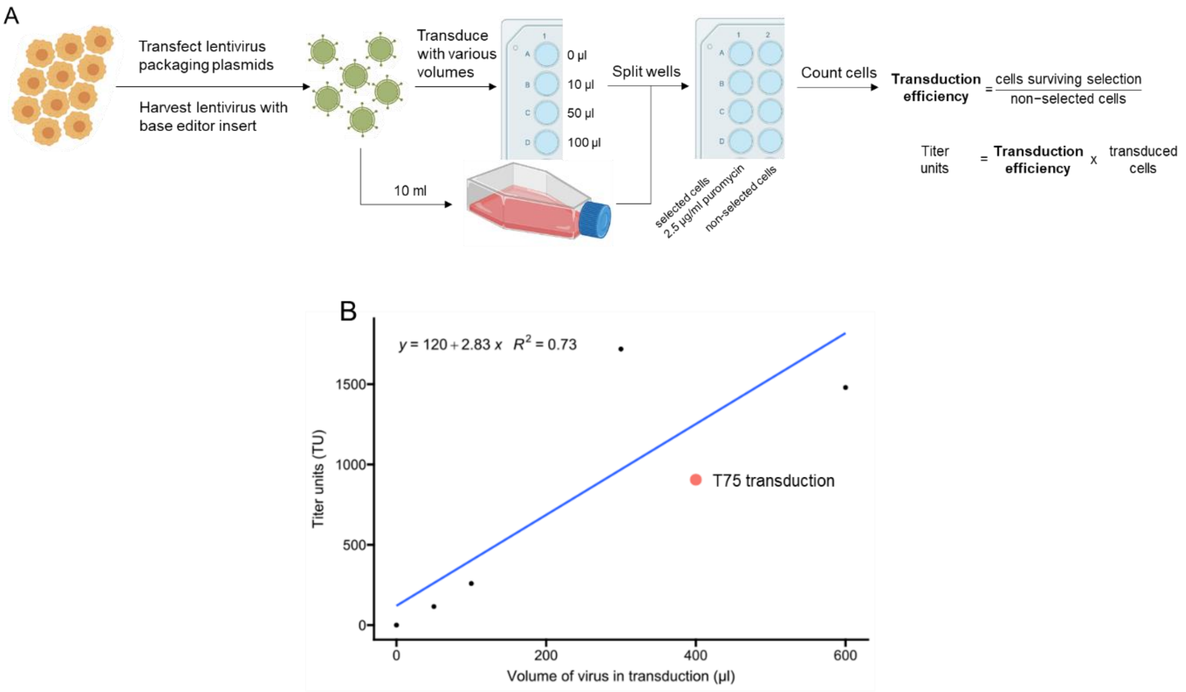

A
Transfect lentivirus packaging plasmids
Harvest lentivirus with base editor insert
Transduce with various volumes
A
B
C
D
1
0 µl
10 µl
50 µl
100 µl
10 ml
Split wells
1
2
selected cells
2.5 µg/ml puromycin
non-selected cells
Count cells
Transduction efficiency = cells surviving selection / non−selected cells
Titer units = Transduction efficiency x transduced cells
B
y = 120 + 2.83 x R² = 0.73
1500
1000
500
0
Titer units (TU)
T75 transduction
0
200
400
600
Volume of virus in transduction (µl)

Figure 6.6: Scaling up virus transduction

(A) Six wells in a 6-well plate were seeded with HEK293T cells and transfected with lentivirus packaging plasmids. ~12 ml of virus was harvested and combined for transductions, 10 ml of which was used to transduced cells in a T75 flask. The number of functional viral particles in each transduction was measured and a standard line was graphed. **(B)** Titer units are displayed for each transduction volume that was tested. The volume used for T75 transduction would be equivalent to 400 µl in a 48-well plate and is displayed here in red.

Standard virus production methods recommend two virus harvests on days 2 and 3 after transfection, as opposed to just a single harvest on day 3 (**Figure 6.7A**). To test if an additional harvest would boost virus production, cells were plated in a T75 at a density of ~80,000 cell/cm^2 and were transfected with the lentiviral packaging plasmids. After 2 days, virus media was collected and stored at 4°C overnight and fresh media was added to the producer cells. The next day, virus media was collected again and combined with the previously harvested media for a total virus media volume of 50 ml. To maintain a high virus concentration in the media, 40 ml of the media was incubated with lentivirus precipitation solution and concentrated to 1 ml. Both the concentrated and unconcentrated virus media were used to transduce HEK293T cells and measure the virus titer of both samples (**Figure**

6.7B). The unconcentrated virus media was used to determine the total virus produced in the full 50 ml of virus media and was measured to contain 236,510 virus particles. This measurement suggested that the concentrated media should have contained 189,200 total virus particles, however the virus titration showed only 25,587 total virus particles, showing that potentially 86.5% of virus particles were lost with this virus precipitation method.

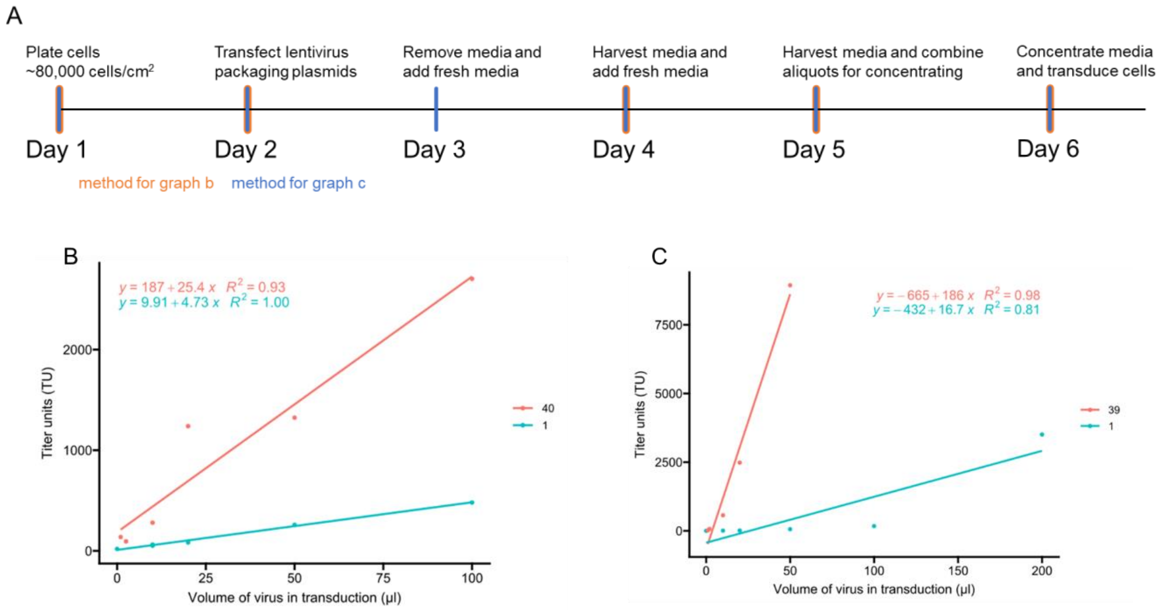


Figure 6.7: Optimization of the virus harvest

(A) Timeline of the double harvest protocol from a T75. Day 3 contains an additional media change that was used in **(C)** and for future virus production. **(B-C)** A T75 was seeded with HEK293T cells and transfected with lentivirus packaging plasmids. 50 ml **(B)** or 40 ml **(C)** of virus was harvested and the indicated volumes were concentrated to 1 ml for transduction. The number of functional viral particles in each transduction was measured and a standard line was graphed. The red points in each graph indicate the concentrated virus, with the number showing the volume (in ml) that was concentrated to 1 ml. The blue points indicate unconcentrated virus.

As a virus that targets mammalian cells, the lentivirus coat protein VSV-G is optimized for an environment with physiological pH levels. Outside of a pH range of 7.0 – 7.4, lentivirus shows drastic drops in infectivity. During the virus production in a T75, the phenol red pH indicator in the media indicated that the media in the first harvest was acidic. This drop in the pH level may have decreased virus viability, and an additional media change before harvest may boost the amount of virus harvested. The virus production in a T75 was repeated as before, this time with an additional media change the day after transfection of the lentivirus packaging plasmids. A 48-well plate was transduced using either concentrated virus (concentrated 39:1) or an unconcentrated aliquot, with a range of volumes to determine the viral titer (**Figure 6.7C**). The unconcentrated virus media showed ~667,568 total virus particles were produced. The concentrated virus media showed 185,335 total virus particles, showing a 72.2% loss during virus concentration. The virus harvest was larger than the previous experiment however, showing viral titer increases of 2.8-fold for the unconcentrated, and 7.2-fold for the concentrated sample. These findings suggest that the additional media change before the first virus harvest boosts virus production significantly, likely due to maintaining a physiological pH level in the media.

Further optimizations of virus production

The large integration sequence of a base editor makes virus production a large bottleneck for performing directed evolution in mammalian cells. By optimizing the density of producer cells and maintaining the pH of their media, virus production was greatly increased, allowing for 100,000’s of base editor variants to be integrated and screened. Additional replicates of the current method should be performed to confirm that these results can reliably predict the coverage of a base editor screening experiment.

Further optimization might be found in the lentivirus precipitation technique, which appears to be losing a significant amount of virus particles. The precipitation requires a 4°C incubation for at least 4 hours but was performed overnight in each of these experiments. Since virus particles lose infectivity over time, this extended incubation might be detrimental. Optimization of this step may improve the amount of functional virus particles for transduction.

Table 6.1: Virus production throughout all experiments

experiment (figure #)	total virus produced (TU)	surface area used (cm^2)
initial transductions (6.4B)	1,175	9.6
initial transductions (6.4C)	44,880	57.6
cell density transductions (6.5), 400k	3,528	9.6
cell density transductions (6.5), 600k	10,295	9.6
cell density transductions (6.5), 800k	50,941	9.6
T75 transduction (6.6)	32,382	57.6
optimized harvest (6.7B), 1	236,510	75
optimized harvest (6.7B), 40	25,587	75
optimized harvest (6.7C), 1	667,568	75
optimized harvest (6.7C), 39	185,335	75

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 lentivirus integration plasmid was cloned using the pHR-SFFV-dCas9-BFP-KRAB plasmid (Addgene #46911). The plasmid was modified in several ways: the H840 was reintroduced to change dCas9 into nCas9, the *BFP* and *KRAB* genes were deleted and replaced with a *P2A-Puro(R)* sequence, the 3' LTR was replaced with a chimeric CMV/LTR promoter to increase expression of the integration sequence, and finally a BsmBI insert site and *XTEN linker* were inserted 3' of the *nCas9* gene to facilitate the insertion of *TadA* variants. ABE integration plasmids were developed from this plasmid with the BsmBI Golden Gate Assembly protocol from NEB, where the *TadA* variant from a donor plasmid was excised and ligated into the integration plasmid by the enzymes in this reaction.

Integration plasmids using different promoters for ABE expression were cloned using the integration plasmid with the BsmBI insert site and used USER cloning to replace the SFFV promoter with the EF1α promoter. *TadA* genes were inserted into this plasmid as described earlier.

Integration plasmids expressing ABE dimers were cloned with USER cloning, using the ABE0.1 or ABE1.1 integration plasmids. Mammalian cell gRNA plasmids were cloned as previously described.[103]

All plasmids used for transfection were harvested using a ZymoPURE^{TM} II Plasmid Midiprep kit (Zymo Research #D4200), following the manufacturer's protocol.

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).

Virus production

This is the final optimized virus production protocol. 12-16 hours before transfection, in either a 6-well plate or a T75 flask, HEK293T cells were plated at a density of ~8.3 x 10^4 cells/cm^2 (~4 x 10^5 per well or ~6 x 10^6 cells, respectively). For a transfection in a 6-well plate, DNA mixtures were prepared with: 1.5 μg of lentivirus transfer plasmid, 1.35 μg of the dR8.2 packaging plasmid, 165 ng of the pMD2-G packaging plasmid, and 200 ng of EGFP-pMAX plasmid. For each transfection, 7.5 μl of *Trans*IT®-LT1 (Mirus #2300) was diluted with 300 μl of Opti-MEM (Gibco #31985-070), then combined with the DNA mixture. The mixture was incubated at room temperature for 15 minutes to form transfectant:DNA complexes, then was carefully pipetted onto the prepared HEK293T cells. For a transfection in a T75 flask, DNA mixtures were prepared with: 11.4 μg of lentivirus transfer plasmid, 10.26 μg of the dR8.2 packaging plasmid, 1254 ng of the pMD2-G packaging plasmid, and 200 ng of EGFP-pMAX plasmid. For each transfection, 57 μl of *Trans*IT®-LT1 (Mirus #2300) was diluted with 1900 μl of Opti-MEM (Gibco #31985-070), then combined with the DNA mixture. The mixture was incubated at room temperature for 15 minutes to form transfectant:DNA complexes, then was carefully pipetted onto the prepared HEK293T cells.

The following day, the media was removed from the transfected cells and replaced with 20 ml of fresh media supplemented with 40 μl of viral boost (500x) (Alstem #VB100). At 48 hours post-transfection, virus media was harvested and kept at 4°C overnight. 20 ml of fresh media was added to the producer cells. At 72 hours post-transfection, virus media was harvested again and combined with the previous harvest. The virus media was centrifuged at 300 x g for 10 minutes to pellet cell debris, then the supernatant was passed through a 0.45 µm filter and retained.

Concentration of virus was performed using Lentivirus Precipitation Solution (Alstem #VC150), as described by the manufacturer's protocol and performed overnight at 4°C.

Transductions and virus titer measurement

The day of the transduction, 3 x 10^4 HEK293T cells per well were plated in a 48-well plate. Each well received a different volume of virus media and was diluted to a final volume of 300 µl using fresh media and supplemented with 8 µg/ml polybrene. Transduced cells were passaged daily for three days. All cells were kept during passaging and moved to larger wells as needed to maintain a healthy confluency.

After three days, transduced cells were split into two wells, with one of the wells being treated with 2.5 µg.ml puromycin to select for transduced cells, and the other being maintained without any selection. Selection was also performed with non-transduced HEK293T cells to determine when all non-transduced cells died in the experimental samples. All cells were passaged as needed during selection. After 3-5 days of selection (when all non-transduced cells died), cells in each well were counted, and the transduction efficiency was determined by comparing the number of cells remaining after selection to the number of cells that did not undergo selection.

Transfections

For fluorescent reporter assays, 12-16 hours before transfection, 5 x 10^4 HEK293T cells/cm^2 were plated in a 48-well plate or a T75 flask, in 250 µl or 10 ml DMEM media. For assays in a 48-well plate, DNA mixtures were prepared with: 500 ng of mCherry-P2A-EGFP reporter plasmid and 250 ng of gRNA plasmid. 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). For assays in a T75 flask, DNA mixtures were prepared with: 34 µg of mCherry-P2A-EGFP reporter plasmid and 17 µg of gRNA plasmid. DNA mixtures were brought to a total volume of 854 µl using Opti-MEM (Gibco #31985-070) and then combined with a 854 µl solution comprised of 75 µl of Lipofectamine 2000 (Invitrogen #11668-019) and 779 µl Opti-MEM (Gibco #31985-070). For all transfections, the resulting DNA/Lipofectamine mixture was then added to the cells. 24 hours after transfection, fresh DMEM media was added to each sample. Cells were then incubated for 48 additional hours before harvesting for flow cytometry.

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)[104]. 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. ~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 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.

The integrated *TadA* gene was 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 Quant-IT^{TM} dsDNA Assay Kit, high sensitivity (ThermoFisher Scientific #Q33120) on a Qubit fluorometer. Samples were then sequenced on an Illumina MiniSeq according to the manufacturer’s protocol.

Analysis of Illumina HTS was performed with CRISPResso2[75]. 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). ABE0.1 variants were detected by the percentage of GAC codons at amino acid position 108, while ABE1.1 variants were detected by the percentage of AAT codons.

Chapter 7 Development of a GFP reporter for base editing demonstrations

Introduction

GFP and other color-forming molecules are great tools for explaining molecular biology techniques to a wider audience.[163] As such, we set out to develop a reliable and straightforward demonstration of base editing using GFP reporters. The goal of this demonstration being to communicate the general mechanism and applications of base editing to students who have minimal experience with biological techniques. To best demonstrate base editing, the GFP turn-on method required a significant and reliable difference between the negative and positive controls. Since the rate of GFP turn-on in fluorescent reporters is dependent on the target site of the reporter and the base editor, we set out to identify the best pairing to demonstrate base editing in bacteria.

Initially, an ABE8.20m and a dGFP(R35STOP) reporter plasmid were paired and co-transformed into bacteria. While this combination could display fluorescence in bacterial colonies, the effect was often inconsistent throughout a colony and often not a reliable demonstration of base editing (**Figure 7.1A**). Picking the bacteria colonies from these transfections revealed that an additional edit was occurring alongside the target A, which resulted in an L54P mutation, likely knocking out fluorescence despite reverting the R53STOP mutation. Since the A111V reporter used previously did not contain any additional As nearby the target, this mutation was installed into the EGFP bacterial expression plasmid and the appropriate gRNA was cloned into the ABE8.20 expression plasmid.[143] Unfortunately, the new ABE8.20 plasmid showed evidence of A-to-G editing happening to the gRNA sequence itself (**Figure 7.1B**). The arabinose-inducible pBAD promoter being used for ABE expression has been shown to have leaky expression, so this editing likely resulted from ABE8.20 recognition and editing of the TACG motif.[164]

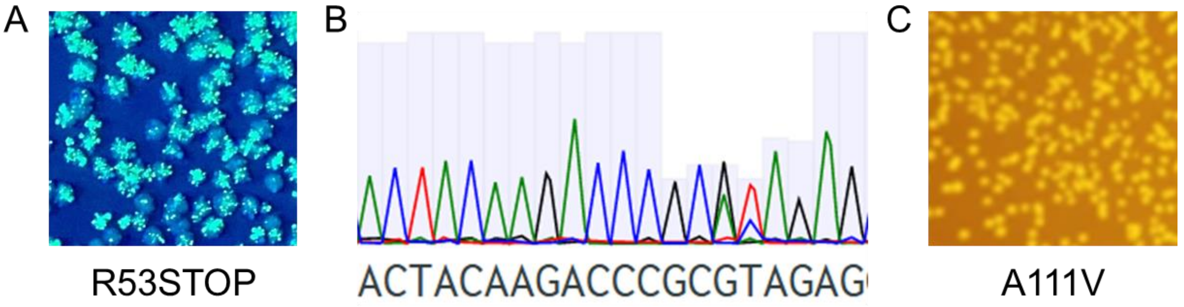


Figure 7.1: Optimization of EGFP turn-on in bacteria

(A) Bacteria colonies expressing ABE8.20, the dGFP(R35STOP) reporter, and corresponding targeting gRNA. EGFP fluorescence is observed as individual points throughout a single colony. **(B)** Sequencing spectra of the A111V gRNA protospacer, which is expressed from the same plasmid as ABE8.20. At the GT motif, peaks for A (green) and C (blue) can be observed, respectively. Indicating the CA motif on the opposite strand was likely deaminated by ABE activity. **(C)** Bacteria colonies expressing ABE8.20, the dGFP(A111V) reporter, and corresponding targeting gRNA. EGFP fluorescence is observed throughout entire colonies.

Editing of the gRNA sequence may reduce the ability of ABE8.20 to correct the A111V mutation in the GFP reporter plasmid. To avoid this off-target editing, the induction of base editing was moved from the transcriptional level to the translational level by including a theophylline-responsive riboswitch. The A111V gRNA was cloned into this version of the ABE8.20 expression plasmid and editing of the gRNA sequence was not observed. Co-transfection of the new ABE8.20 plasmid and the dGFP(A111V) reporter plasmid resulted in colonies that fluoresced brightly under UV light, clearly demonstrating base editing in the transformed cells (**Figure 7.1C**).

The full base editing outreach was designed as a 3-day experience, with the transformation of base editing plasmids into bacteria occurring on day 2. During this day, we also facilitate a discussion with the students about defining genetic diseases and the ethical dilemmas in the genome editing field. The goal of this discussion is to prompt students to think about how genetics impact someone’s identity, and how this connection between genetics and identity should be considered when applying base editing as a therapeutic tool.

More information about the Genome Editing Technologies program that was developed using GFP reporters can be found in our publication.[165]

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. Cloning of the ABE8.20-A111V gRNA plasmid was performed using blunt end cloning.

Fluorescent reporter assay in bacteria

E. coli were co-transformed with the A111V-TACG reporter and the ABE8.20-A111V gRNA plasmids, and allowed to recover for 1 hour 37°C in 2xYT media. Recoveries were plated on 2xYT agar plates supplemented with 50 ng/µL kanamycin and 50 µg/mL carbenicillin (BE plasmid maintenance antibiotic). Plates were then incubated for overnight at 37°C. EGFP fluorescence was detected in cells by observing them under a UV light.

Acknowledgements

Chapter 7 has been adapted, in part, with permission, from Vasquez, C. A. Evanoff, M., **Ranzau, B. L.**, Gu, S., Deters, E., Komor, A. C. Curing "GFP-itis" in Bacteria with Base Editors: Development of a Genome Editing Science Program Implemented with High School Biology Students. *The CRISPR Journal* **6**, 186–195 (2023).

References

1 M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, E. Charpentier, *Science* **2012**, *337*, 816.

2 O. O. Abudayyeh, J. S. Gootenberg, P. Essletzbichler, S. Han, J. Joung, J. J. Belanto, V. Verdine, D. B. T. Cox, M. J. Kellner, A. Regev, E. S. Lander, D. F. Voytas, A. Y. Ting, F. Zhang, *Nature* **2017**, *550*, 280.

3 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.

4 D. C. Swarts, J. van der Oost, M. Jinek, *Molecular Cell* **2017**, *66*, 221.

5 B. Zetsche, J. S. Gootenberg, O. O. Abudayyeh, I. M. Slaymaker, K. S. Makarova, P. Essletzbichler, S. E. Volz, J. Joung, J. Van Der Oost, A. Regev, E. V. Koonin, F. Zhang, *Cell* **2015**, *163*, 759.

6 S. Shmakov, O. O. Abudayyeh, K. S. Makarova, Y. I. Wolf, J. S. Gootenberg, E. Semenova, L. Minakhin, J. Joung, S. Konermann, K. Severinov, F. Zhang, E. V. Koonin, *Molecular Cell* **2015**, *60*, 385.

7 O. O. Abudayyeh, J. S. Gootenberg, S. Konermann, J. Joung, I. M. Slaymaker, D. B. T. Cox, S. Shmakov, K. S. Makarova, E. Semenova, L. Minakhin, K. Severinov, A. Regev, E. S. Lander, E. V. Koonin, F. Zhang, *Science* **2016**, *353*.

8 S. Konermann, P. Lotfy, N. J. Brideau, J. Oki, M. N. Shokhirev, P. D. Hsu, *Cell* **2018**, *173*, 665.

9 W. X. Yan, S. Chong, H. Zhang, K. S. Makarova, E. V. Koonin, D. R. Cheng, D. A. Scott, *Molecular Cell* **2018**, *70*, 327.

10 M. Adli, *Nature communications* **2018**, *9*, 1911.

11 L. S. Qi, M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin, W. A. Lim, *Cell* **2013**, *152*, 1173.

12 G. T. Hess, J. Tycko, D. Yao, M. C. Bassik, *Molecular Cell* **2017**, *68*, 26.

13 A. C. Komor, Y. B. Kim, M. S. Packer, J. A. Zuris, D. R. Liu, *Nature* **2016**, *533*, 420.

14 K. Nishida, T. Arazoe, N. Yachie, S. Banno, M. Kakimoto, M. Tabata, M. Mochizuki, A. Miyabe, M. Araki, K. Y. Hara, Z. Shimatani, A. Kondo, *Science* **2016**, *353*, aaf8729.

15 X. Li, Y. Wang, Y. Liu, B. Yang, X. Wang, J. Wei, Z. Lu, Y. Zhang, J. Wu, X. Huang, L. Yang, J. Chen, *Nature Biotechnology* **2018**, *36*, 324.

16 Y. B. Kim, A. C. Komor, J. M. Levy, M. S. Packer, K. T. Zhao, D. R. Liu, *Nature Biotechnology* **2017**, *35*, 371.

17 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.

18 G. T. Hess, L. Frésard, K. Han, C. H. Lee, A. Li, K. A. Cimprich, S. B. Montgomery, M. C. Bassik, *Nature Methods* **2016**, *13*, 1036.

19 Y. Ma, J. Zhang, W. Yin, Z. Zhang, Y. Song, X. Chang, *Nature Methods* **2016**, *13*, 1029.

20 V. H. Odegard, D. G. Schatz, *Nature Reviews Immunology* **2006**, *6*, 573.

21 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.

22 W. Jiang, S. Feng, S. Huang, W. Yu, G. Li, G. Yang, Y. Liu, Y. Zhang, L. Zhang, Y. Hou, J. Chen, J. Chen, X. Huang, *Cell Research* **2018**, *28*, 855.

23 L. Wang, W. Xue, L. Yan, X. Li, J. Wei, M. Chen, J. Wu, B. Yang, L. Yang, J. Chen, *Cell research* **2017**, *27*, 1289.

24 S. E. Bennett, D. W. Mosbaughs, *The Journal of biological chemistry* **1992**, *267*, 22512.

25 X. Wang, J. Li, Y. Wang, B. Yang, J. Wei, J. Wu, R. Wang, X. Huang, J. Chen, L. Yang, *Nature Biotechnology* **2018**, *36*, 946.

26 G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, *Proceedings of the National Academy of Sciences* **2012**, *109*, E2579.

27 A. Pluciennik, L. Dzantiev, R. R. Iyer, N. Constantin, F. A. Kadyrov, P. Modrich, *Proceedings of the National Academy of Sciences* **2010**, *107*, 16066.

28 P. A. Jones, S. B. Baylin, *Nature Reviews Genetics* **2002**, *3*, 415.

29 A. Bird, *Genes & development* **2002**, *16*, 6.

30 R. Jaenisch, A. Bird, *Nature Genetics* **2003**, *33*, 245.

31 K. D. Robertson, *Nature Reviews Genetics* **2005**, *6*, 597.

32 X. S. Liu, H. Wu, X. Ji, Y. Stelzer, X. Wu, S. Czauderna, J. Shu, D. Dadon, R. A. Young, R. Jaenisch, *Cell* **2016**, *167*, 233.

33 A. Vojta, P. Dobrinic, V. Tadic, L. Bockor, P. Korac, B. Julg, M. Klasic, V. Zoldos, *Nucleic Acids Research* **2016**, *44*, 5615.

34 X. Xu, Y. Tao, X. Gao, L. Zhang, X. Li, W. Zou, K. Ruan, F. Wang, G. Xu, R. Hu, *Cell Discovery* **2016**, *2*, 16009.

35 Y.-H. Huang, J. Su, Y. Lei, L. Brunetti, M. C. Gundry, X. Zhang, M. Jeong, W. Li, M. A. Goodell, *Genome Biology* **2017**, *18*, 176.

36 C. Pflueger, D. Tan, T. Swain, T. Nguyen, J. Pflueger, C. Nefzger, J. M. Polo, E. Ford, R. Lister, *Genome Research* **2018**, *28*, 1193.

37 P. Stepper, G. Kungulovski, R. Z. Jurkowska, T. Chandra, F. Krueger, R. Reinhardt, W. Reik, A. Jeltsch, T. P. Jurkowski, *Nucleic Acids Research* **2017**, *45*, 1703.

38 J. I. McDonald, H. Celik, L. E. Rois, G. Fishberger, T. Fowler, R. Rees, A. Kramer, A. Martens, J. R. Edwards, G. A. Challen, *Biology Open* **2016**, *5*, 866.

39 T. Xiong, G. E. Meister, R. E. Workman, N. C. Kato, M. J. Spellberg, F. Turker, W. Timp, M. Ostermeier, C. D. Novina, *Scientific Reports* **2017**, *7*, 6732.

40 S. Cortellino, J. Xu, M. Sannai, R. Moore, E. Caretti, A. Cigliano, M. Le Coz, K. Devarajan, A. Wessels, D. Soprano, L. K. Abramowitz, M. S. Bartolomei, F. Rambow, M. R. Bassi, T. Bruno, M. Fanciulli, C. Renner, A. J. Klein-Szanto, Y. Matsumoto, D. Kobi, I. Davidson, C. Alberti, L. Larue, A. Bellacosa, *Cell* **2011**, *146*, 67.

41 X. S. Liu, H. Wu, M. Krzisch, X. Wu, J. Graef, J. Muffat, D. Hnisz, C. H. Li, B. Yuan, C. Xu, Y. Li, D. Vershkov, A. Cacace, R. A. Young, R. Jaenisch, *Cell* **2018**, *172*, 979.

42 M. Avitzour, H. Mor-Shaked, S. Yanovsky-Dagan, S. Aharoni, G. Altarescu, P. Renbaum, T. Eldar-Geva, O. Schonberger, E. Levy-Lahad, S. Epsztejn-Litman, R. Eiges, *Stem Cell Reports* **2014**, *3*, 699.

43 I. A. Roundtree, M. E. Evans, T. Pan, C. He, *Cell* **2017**, *169*, 1187.

44 M. Yasui, E. Suenaga, N. Koyama, C. Masutani, F. Hanaoka, P. Gruz, S. Shibutani, T. Nohmi, M. Hayashi, M. Honma, *Journal of Molecular Biology* **2008**, *377*, 1015.

45 M. Saparbaev, J. Laval, *Proceedings of the National Academy of Sciences* **1994**, *91*, 5873.

46 I. Alseth, B. Dalhus, M. Bjørås, *Current Opinion in Genetics and Development* **2014**, *26*, 116.

47 P. Karran, T. Lindahl, *Biochemistry* **1980**, *19*, 6005.

48 T. Stafforst, M. F. Schneider, *Angewandte Chemie - International Edition* **2012**, *51*, 11166.

49 M. F. Montiel-Gonzalez, I. Vallecillo-Viejo, G. A. Yudowski, J. J. C. Rosenthal, *Proceedings of the National Academy of Sciences* **2013**, *110*, 18285.

50 D. B. T. Cox, J. S. Gootenberg, O. O. Abudayyeh, B. Franklin, M. J. Kellner, J. Joung, F. Zhang, *Science (New York, N.Y.)* **2017**, *358*, 1019.

51 S. K. Wong, S. Sato, D. W. Lazinski, S. K. E. E. Wong, S. Sato, D. W. Lazinski, *RNA* **2001**, *7*, 846.

52 Y. Zong, Y. Wang, C. Li, R. Zhang, K. Chen, Y. Ran, J. L. Qiu, D. Wang, C. Gao, *Nature Biotechnology* **2017**, *35*, 438.

53 J. Li, Y. Sun, J. Du, Y. Zhao, L. Xia, *Molecular Plant* **2017**, *10*, 526.

54 Z. Shimatani, S. Kashojiya, M. Takayama, R. Terada, T. Arazoe, H. Ishii, H. Teramura, T. Yamamoto, H. Komatsu, K. Miura, H. Ezura, K. Nishida, T. Ariizumi, A. Kondo, *Nature Biotechnology* **2017**, *35*, 441.

55 Y. Li, S. Ma, L. Sun, T. Zhang, J. Chang, W. Lu, X. Chen, Y. Liu, X. Wang, R. Shi, P. Zhao, Q. Xia, *G3* **2018**, *8*, 1701.

56 H. A. Rees, A. C. Komor, W.-H. Yeh, J. Caetano-Lopes, M. Warman, A. S. B. Edge, D. R. Liu, *Nature Communications* **2017**, *8*, 15790.

57 Y. Zhang, W. Qin, X. Lu, J. Xu, H. Huang, H. Bai, S. Li, S. Lin, *Nature Communications* **2017**, *8*, 118.

58 K. Kim, S. M. Ryu, S. T. Kim, G. Baek, D. Kim, K. Lim, E. Chung, S. Kim, J. S. Kim, *Nature Biotechnology* **2017**, *35*, 435.

59 Y. Ma, L. Yu, X. Zhang, C. Xin, S. Huang, L. Bai, W. Chen, R. Gao, J. Li, S. Pan, X. Qi, X. Huang, L. Zhang, *Cell Discovery* **2018**, *4*, 39.

60 S.-M. Ryu, T. Koo, K. Kim, K. Lim, G. Baek, S.-T. Kim, H. S. Kim, D. Kim, H. Lee, E. Chung, J.-S. Kim, *Nature Biotechnology* **2018**, *36*, 536.

61 C. Li, Y. Zong, Y. Wang, S. Jin, D. Zhang, Q. Song, R. Zhang, C. Gao, *Genome Biology* **2018**, *19*, 59.

62 F. Yan, Y. Kuang, B. Ren, J. Wang, D. Zhang, H. Lin, B. Yang, X. Zhou, H. Zhou, *Molecular Plant* **2018**, *11*, 631.

63 K. Hua, X. Tao, F. Yuan, D. Wang, J. K. Zhu, *Molecular Plant* **2018**, *11*, 627.

64 W.-H. Yeh, H. Chiang, H. A. Rees, A. S. B. Edge, D. R. Liu, *Nature Communications* **2018**, *9*, 2184.

65 F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, F. Zhang, *Nature Protocols* **2013**, *8*, 2281.

66 C. Kuscu, M. Parlak, T. Tufan, J. Yang, K. Szlachta, X. Wei, R. Mammadov, M. Adli, *Nature Methods* **2017**, *14*, 710.

67 M. Gapinske, A. Luu, J. Winter, W. S. Woods, K. A. Kostan, N. Shiva, J. S. Song, P. Perez-Pinera, *Genome Biology* **2018**, *19*, 1.

68 M. M. Evers, H.-D. Tran, I. Zalachoras, O. C. Meijer, J. T. den Dunnen, G.-J. B. van Ommen, A. Aartsma-Rus, W. M. C. van Roon-Mom, *Nucleic Acid Therapeutics* **2014**, *24*, 4.

69 D. Kim, K. Lim, S. T. Kim, S. H. Yoon, K. Kim, S. M. Ryu, J. S. Kim, *Nature Biotechnology* **2017**, *35*, 475.

70 J. Grünewald, R. Zhou, S. Iyer, C. A. Lareau, S. P. Garcia, M. J. Aryee, J. K. Joung, *Nat Biotechnol* **2019**, *37*, 1041.

71 J. Wolf, A. P. Gerber, W. Keller, *The EMBO Journal* **2002**, *21*, 3841.

72 H. C. Losey, A. J. Ruthenburg, G. L. Verdine, *Nat Struct Mol Biol* **2006**, *13*, 153.

73 K. L. Rallapalli, B. L. Ranzau, K. R. Ganapathy, F. Paesani, A. C. Komor, *The CRISPR Journal* **2022**, *5*, 294.

74 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.

75 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.

76 A. C. Komor, Y. B. Kim, M. S. Packer, J. A. Zuris, D. R. Liu, *Nature* **2016**, *533*, 420.

77 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.

78 M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, E. Charpentier, *Science* **2012**, *337*, 816.

79 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.

80 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.

81 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.

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

83 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**, *38*, 892.

84 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**.

85 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**, *19*, 101.

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

87 K. M. McKenney, M. A. T. Rubio, J. D. Alfonzo, in *The Enzymes*, ed. by Guillaume F. Chanfreau, Academic Press, **2017**, Vol. 41, pp. 51–88.

88 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.

89 H. A. Rees, C. Wilson, J. L. Doman, D. R. Liu, *Sci Adv* **2019**, *5*.

90 J. Kim, V. Malashkevich, S. Roday, M. Lisbin, V. L. Schramm, S. C. Almo, *Biochemistry* **2006**, *45*, 6407.

91 Y. Elias, R. H. Huang, *Biochemistry* **2005**, *44*, 12057.

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

93 H. Zhao, *Biotechnology and Bioengineering* **2007**, *98*, 313.

94 M. D. Lane, B. Seelig, *Current Opinion in Chemical Biology* **2014**, *22*, 129.

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

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

97 J. L. Fu, T. Kanno, S.-C. Liang, A. J. M. Matzke, M. Matzke, *G3 (Bethesda)* **2015**, *5*, 1849.

98 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.

99 S. A. Lynch, J. P. Gallivan, *Nucleic Acids Research* **2009**, *37*, 184.

100 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.

101 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.

102 P. Wang, L. Xu, Y. Gao, R. Han, *Molecular Therapy* **2020**, *28*, 1696.

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

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

105 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/, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, **2022**.

106 H. Wickham, *ggplot2*, 2nd ed. , Springer Cham, **2016**.

107 A. Kassambara, rstatix: Pipe-Friendly Framework for Basic Statistical Tests, **2022**.

108 J. C. Carlson, A. H. Badran, D. A. Guggiana-Nilo, D. R. Liu, *Nat Chem Biol* **2014**, *10*, 216.

109 B. Webb, A. Sali, *Methods Mol Biol* **2014**, *1137*, 1.

110 C. T. Chung, S. L. Niemela, R. H. Miller, *Proc Natl Acad Sci U S A* **1989**, *86*, 2172.

111 C. Anders, O. Niewoehner, A. Duerst, M. Jinek, *Nature* **2014**, *513*, 569.

112 J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath, A. Onufriev, *Nucleic Acids Res* **2005**, *33*, W368.

113 R. Anandakrishnan, B. Aguilar, A. V. Onufriev, *Nucleic Acids Res* **2012**, *40*, W537.

114 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.

115 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.

116 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926.

117 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J Chem Theory Comput* **2015**, *11*, 3696.

118 A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, M. Orozco, *Biophysical Journal* **2007**, *92*, 3817.

119 M. Zgarbová, M. Otyepka, J. Sponer, A. Mládek, P. Banáš, T. E. Cheatham, P. Jurečka, *J Chem Theory Comput* **2011**, *7*, 2886.

120 P. Banáš, D. Hollas, M. Zgarbová, P. Jurečka, M. Orozco, T. E. Cheatham, J. Šponer, M. Otyepka, *J Chem Theory Comput* **2010**, *6*, 3836.

121 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.

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

123 P. Li, K. M. Merz, *J Chem Inf Model* **2016**, *56*, 599.

124 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.

125 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., AMBER 2018, **2018**.

126 D. R. Roe, T. E. Cheatham, *J Chem Theory Comput* **2013**, *9*, 3084.

127 D. R. Roe, T. E. Cheatham, *J Comput Chem* **2018**, *39*, 2110.

128 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.

129 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.

130 J. D. Hunter, *Computing in Science & Engineering* **2007**, *9*, 90.

131 H. E. Krokan, F. Drabløs, G. Slupphaug, *Oncogene* **2002**, *21*, 8935.

132 P. Karran, T. Lindahl, *Biochemistry* **1980**, *19*, 6005.

133 S. Mao, P. Haruehanroengra, S. V. Ranganathan, F. Shen, T. J. Begley, J. Sheng, *ACS Chem. Biol.* **2020**.

134 A. Furrer, B. van Loon, *Nucleic Acids Research* **2014**, *42*, 553.

135 T. Suzuki, K. Miyauchi, *FEBS Letters* **2010**, *584*, 272.

136 C. Köhrer, D. Mandal, K. W. Gaston, H. Grosjean, P. A. Limbach, U. L. RajBhandary, *Nucleic Acids Res* **2014**, *42*, 1904.

137 S. D'Silva, S. J. Haider, E. M. Phizicky, *RNA* **2011**, *17*, 1100.

138 A. Noma, S. Yi, T. Katoh, Y. Takai, T. Suzuki, T. Suzuki, *RNA* **2011**, *17*, 1111.

139 K. Nakanishi, S. Fukai, Y. Ikeuchi, A. Soma, Y. Sekine, T. Suzuki, O. Nureki, *Proceedings of the National Academy of Sciences* **2005**, *102*, 7487.

140 K. Nakanishi, L. Bonnefond, S. Kimura, T. Suzuki, R. Ishitani, O. Nureki, *Nature* **2009**, *461*, 1144.

141 J. C. Delaney, J. M. Essigmann, *Proceedings of the National Academy of Sciences* **2004**, *101*, 14051.

142 R. C. Cadwell, G. F. Joyce, *Genome Res.* **1992**, *2*, 28.

143 B. L. Ranzau, K. L. Rallapalli, M. Evanoff, F. Paesani, A. C. Komor, *ChemBioChem* **2023**, *n/a*, e202200788.

144 M. A. T. Rubio, I. Pastar, K. W. Gaston, F. L. Ragone, C. J. Janzen, G. A. M. Cross, F. N. Papavasiliou, J. D. Alfonzo, *Proceedings of the National Academy of Sciences* **2007**, *104*, 7821.

145 S. Thomas, N. D. Maynard, J. Gill, *Nat Methods* **2015**, *12*, i.

146 L. Han, E. Marcus, S. D'Silva, E. M. Phizicky, *RNA* **2017**, *23*, 406.

147 E. Ramos-Morales, E. Bayam, J. Del-Pozo-Rodríguez, T. Salinas-Giegé, M. Marek, P. Tilly, P. Wolff, E. Troesch, E. Ennifar, L. Drouard, J. D. Godin, C. Romier, *Nucleic Acids Research* **2021**, *49*, 6529.

148 A. P. Gerber, H. Grosjean, T. Melcher, W. Keller, *The EMBO Journal* **1998**, *17*, 4780.

149 S. Maas, A. P. Gerber, A. Rich, *Proceedings of the National Academy of Sciences* **1999**, *96*, 8895.

150 K. M. McKenney, M. A. T. Rubio, J. D. Alfonzo, *RNA* **2018**, *24*, 56.

151 M. A. T. Rubio, K. W. Gaston, K. M. McKenney, I. M. C. Fleming, Z. Paris, P. A. Limbach, J. D. Alfonzo, *Nature* **2017**, *542*, 494.

152 T. Lindahl, *Nature* **1993**, *362*, 709.

153 M. P. Westbye, E. Feyzi, P. A. Aas, C. B. Vågbø, V. A. Talstad, B. Kavli, L. Hagen, O. Sundheim, M. Akbari, N.-B. Liabakk, G. Slupphaug, M. Otterlei, H. E. Krokan, *Journal of Biological Chemistry* **2008**, *283*, 25046.

154 S. S. Parikh, G. Walcher, G. D. Jones, G. Slupphaug, H. E. Krokan, G. M. Blackburn, J. A. Tainer, *Proceedings of the National Academy of Sciences* **2000**, *97*, 5083.

155 M. K. Purchal, D. E. Eyler, M. Tardu, M. K. Franco, M. M. Korn, T. Khan, R. McNassor, R. Giles, K. Lev, H. Sharma, J. Monroe, L. Mallik, M. Koutmos, K. S. Koutmou, *Proc Natl Acad Sci U S A* **2022**, *119*, e2109708119.

156 S. Kimura, V. Srisuknimit, K. L. McCarty, P. C. Dedon, P. J. Kranzusch, M. K. Waldor, *Nat Commun* **2022**, *13*, 5994.

157 B. W. Thuronyi, L. W. Koblan, J. M. Levy, W.-H. Yeh, C. Zheng, G. A. Newby, C. Wilson, M. Bhaumik, O. Shubina-Oleinik, J. R. Holt, D. R. Liu, *Nat Biotechnol* **2019**, *37*, 1070.

158 J. Demeulemeester, J. De Rijck, R. Gijsbers, Z. Debyser, *Bioessays* **2015**, *37*, 1202.

159 P. Fonseca, A. R. Bonny, G. R. Kumar, A. H. Ng, J. Town, Q. C. Wu, E. Aslankoohi, S. Y. Chen, G. Dods, P. Harrigan, L. C. Osimiri, A. L. Kistler, H. El-Samad, *ACS Synthetic Biology* **2019**, 14.

160 U. Müller-Kuller, M. Ackermann, S. Kolodziej, C. Brendel, J. Fritsch, N. Lachmann, H. Kunkel, J. Lausen, A. Schambach, T. Moritz, M. Grez, *Nucleic Acids Res* **2015**, *43*, 1577.

161 N. P. Sweeney, C. A. Vink, *Molecular Therapy - Methods & Clinical Development* **2021**, *21*, 574.

162 M. Kumar, B. Keller, N. Makalou, R. E. Sutton, *Human Gene Therapy* **2001**, *12*, 1893.

163 H. Ziegler, W. Nellen, *Methods* **2020**, *172*, 86.

164 M. Simcikova, K. L. J. Prather, D. M. F. Prazeres, G. A. Monteiro, *Vaccine* **2014**, *32*, 2843.

165 C. A. Vasquez, M. Evanoff, B. L. Ranzau, S. Gu, E. Deters, A. C. Komor, *The CRISPR Journal* **2023**, *6*, 186.