Base Editing: A New Approach to Genome Editing

Citation

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:41127124

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Base Editing: A New Approach to Genome Editing

A dissertation presented
by
Yongjoo Kim
to
The Committee on Higher Degrees in Chemistry

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Chemistry

Harvard University
Cambridge, Massachusetts
January 2018
Abstract

Genome editing technologies, such as the recently discovered CRISPR/Cas9 system, allow the researcher to site-specifically modify genes of interest in living cells and organisms, and have the potential to yield new therapeutics to cure genetic diseases. Traditional Cas9-based strategies, however, rely on the generation of a double-stranded DNA break (DSB) and its subsequent repair. This method is inefficient at introducing precise nucleotide changes, and generates an excess of random insertions or deletions (indels). Considering that over half of known genetic diseases are caused by single nucleotide changes in DNA, new technologies that can cleanly correct such mutations are needed.

This thesis outlines the development of base editing, a strategy for directly converting one base pair into another in living cells, without the generation of DSBs. Base editing uses a fusion protein composed of a cytidine deaminase, a catalytically impaired Cas9, and a base excision repair inhibitor to induce site-specific conversion of C to T or G to A in living cells. Base editing is able to correct a number of disease-relevant mutations in cell culture. New base editors with altered PAM specificities expand the number of disease mutations targetable in the human genome. Base editors were further engineered to improve their targeting precision and safety.
The last part of the thesis aims to discover ways to improve the cytosolic delivery of macromolecules, a bottleneck for realizing their therapeutic potential. Previously, proteins fused to highly anionic moieties, such as (-30)GFP, have been delivered into cells using cationic lipids. To identify a more potent tag for protein delivery, we conducted a screen of highly anionic proteins from the human proteome for their ability to deliver cargo into cells with cationic lipids. We demonstrate that a relatively unknown protein called ProTα outperforms (-30)GFP in protein delivery assays.
Vita brevis,
ars longa,
occasio praeceps,
experimentum periculosum,
iudicium difficile.
Acknowledgements

I thank Professor David R. Liu for his mentorship. I am fortunate to have been a part of a highly creative, fun, and intellectually-stimulating environment. Through David’s thoughtful supervision, I have grown as a scientist. I thank my graduate advising committee members, Professors Adam Cohen and Keith Joung, for their mentorship. I thank all the former and present lab members who have supported me in small and large ways.

I thank Alexis Komor. Without her pioneering work on base editing, and her generous support and kindness, this dissertation would not be.

I thank David Thompson for teaching me how to think creatively about science, which made the process so much fun.

I thank Kevin Zhao who showed me the spark of curiosity once again that had led me to research in the first place, and thus made working with him a joy.

I especially thank Holly Rees. Because of her I run a little bit faster toward my goal.

I thank Ahmed Badran, John Zuris, Kevin Davis, Basil Hubbard, and Jonathan Levy for their scientific knowledge and mentorship. I thank Dmitry Usanov, Kevin Davis, Chihui An, Xue Gao, Peng Wang, Zhen Chen, Ning Sun, Liwei Chen, Chris Podracky, Phillip Lichtor, Jia Niu, Weixin Tang, Alix Chan, Tim Roth, Ariel Yeh, Beverly Mok, Johnny Hu, Tina Zeina, Maarten Geurts, and Kevin Zhao for their generosity.

I thank my family.

I thank Natural Sciences and Engineering Council of Canada (NSERC) for financially supporting me through PGS-M and PGS-D scholarships.
Table of Contents

Abstract ............................................................................................................................................... iii
Acknowledgements ........................................................................................................................... vi
Table of Contents .............................................................................................................................. vii
List of Figures ..................................................................................................................................... ix
List of Tables ....................................................................................................................................... x
Abbreviations ..................................................................................................................................... xi

Introduction Part I: Genomic Editing ................................................................................................. 1

Introduction Part II: Cytosolic Protein Delivery ............................................................................... 2

Chapter One – Development of Base Editing ................................................................................ 4
  1.1 – Introduction ............................................................................................................................... 5
  1.2 – First-Generation Base Editor, BE1 .......................................................................................... 6
  1.3 – Second-Generation Base Editor, BE2 ..................................................................................... 9
  1.4 – Third-Generation Base Editor, BE3 ....................................................................................... 10
  1.5 – Correction of Disease-Relevant Mutations Using BE3 ........................................................... 12
  1.6 – Conclusion ............................................................................................................................... 14
  1.7 – Methods .................................................................................................................................. 15

Chapter Two – Improving the Targeting Scope of Base Editing .................................................. 24
  2.1 – Introduction: Targeting Scope in Base Editing ....................................................................... 25
  2.2 – Utilizing Cas9 Homologs for Base Editing .......................................................................... 25
  2.3 – Utilizing Cas9 Variants with Altered PAM-Specificities for Base Editing ......................... 32
  2.4 – Genomic Specificity of Base Editors with Altered PAM-Specificities ............................... 35
  2.5 – Conclusion ............................................................................................................................... 36
  2.6 – Methods .................................................................................................................................. 38

Chapter Three – Modulating the Activity Window of Base Editing ............................................ 43
  3.1 – Introduction: Selectivity in Base Editing ................................................................................ 44
  3.2 – Effect of Linker Length on Base Editing ................................................................................ 44
  3.3 – Effect of Guide RNA Length on Base Editing ....................................................................... 44
  3.4 – Attenuation of Deaminase Kinetics Influences Base Editing Window Width ................... 46
  3.5 – Development of Base Editors with Narrowed Activity Windows ................................... 49
  3.6 – APOBEC1 Mutations Cause a Shift in Base Editing Product Distributions .................... 51
  3.7 – The Narrowed Editing Window Profiles Are Stable over Multiple Generations ............. 53
  3.8 – Quantification of APOBEC1 Catalytic Impairment ............................................................... 53
  3.9 – Base Editors with Narrowed Activity Windows Are More Specific .................................. 55
  3.10 – Base Editors with Both Altered PAM-Sequences and Narrowed Windows ................ 56
  3.11 – Conclusion ............................................................................................................................. 58
  3.12 – Methods .................................................................................................................................. 60
Chapter Four – Improving the Safety of Base Editing ................................................................. 62
  4.1 – Introduction .......................................................................................................................... 63
  4.2 – Base Editors with Reduced Indel Rates ............................................................................. 63
  4.3 – Small Molecule-Inducible Base Editors ........................................................................... 64
  4.4 – Conclusion .......................................................................................................................... 66
  4.5 – Methods ............................................................................................................................. 66

Chapter Five – Efficient Cytosolic Delivery of Macromolecules via Cationic Liposomes ....67
  5.1 – Introduction ........................................................................................................................ 68
  5.2 – Protein Surface Charge on Cationic Lipid Delivery ........................................................... 69
  5.3 – Human Proteome Analysis ................................................................................................ 71
  5.4 – Identification of a Human Protein that Potently Facilitates Protein Delivery ............... 72
  5.5 – Comparison with Charge Analogs .................................................................................... 74
  5.6 – ProTα Truncation Analysis ............................................................................................... 75
  5.7 – Lycotoxin L17E Does Not Improve Cytosolic Protein Delivery ...................................... 78
  5.8 – ProTα-Mediated Delivery of Zinc Finger Nucleases ....................................................... 79
  5.9 – Conclusion ........................................................................................................................ 82
  5.10 – Methods .......................................................................................................................... 82

Chapter Six – Discussion and Future Directions ........................................................................ 85
  6.1 – Short-Term Limitations of Base Editing .......................................................................... 86
        6.1.1 – Variability in Base Editing Efficiencies ................................................................. 86
        6.1.2 – Within-Window Precision ........................................................................................ 86
  6.2 – The Future of Base Editing ............................................................................................... 88
        6.2.1 – Base Editing in the Era of the Genome-Editing Revolution .................................. 88
        6.2.2 – Eliminating Genomic Off-Target Base Editing ....................................................... 89
        6.2.3 – New Base Editing Transformations ......................................................................... 90
        6.2.4 – Beyond Base Editing ............................................................................................... 90
        6.2.5 – Beyond Genome Editing .......................................................................................... 91
  6.3 – Concluding Remarks ......................................................................................................... 92

References .................................................................................................................................. 93
List of Figures

Figure 1.1. Overview of base editing ................................................................. 6
Figure 1.2. Evaluation of different cytidine deaminase enzymes ........................ 7
Figure 1.3. Effect of different linker lengths and composition on deaminase-dCas9 fusion . 8
Figure 1.4. Editing efficiencies of BE1, BE2, BE3, and Cas9+HDR in genomic sites in HEK293T cells .................................................................................. 9
Figure 1.5. Indel generated by BE1, BE2, BE3, and Cas9+HDR ................................ 11
Figure 1.6. Permanence of base editing in human cells ........................................ 11
Figure 1.7. BE3-mediated correction of disease-relevant mutations in mammalian cells .............................................................................. 14
Figure 2.1. Base editors generated with Cas9 homologs and engineered variants ..... 26
Figure 2.2. Base editing with SaBE3 and SaKKH-BE3 ........................................... 27
Figure 2.3. BE3 activity at sites targeted by Cas9 homologs and engineered variants ........................................................................ 28
Figure 2.4. Base editing with CjBE3 and CjBE4 ....................................................... 30
Figure 2.5. Base editing with GeoBE3 and GeoBE4 .................................................. 31
Figure 2.6. CasX mammalian GFP-cleavage assay .................................................. 32
Figure 2.7. Base editing with VQR-BE3, EQR-BE3, and VRER-BE3 ........................ 34
Figure 2.8. Editing efficiencies of base editors with altered PAM-specificities in U2OS cells ........................................................................ 34
Figure 2.9. Off-target base editing by SaBE3, SaKKH-BE3, VQR-BE3, and EQR-BE3 ................................................................. 36
Figure 2.10. Disease-associated mutations in ClinVar correctable with base editing ........................................................................... 37
Figure 3.1. Effect of APOBEC1–Cas9 linker length on editing window width .......... 45
Figure 3.2. Effect of 5′-guide RNA truncations on editing window width ................ 45
Figure 3.3. Base editors with mutant cytidine deaminase domains exhibit narrowed editing windows .................................................................................. 48
Figure 3.4. Activity of base editors with APOBEC1 mutations on well-studied genomic sites ........................................................................ 50
Figure 3.5. Activity of base editors with narrowed deamination windows in U2OS cells ........................................................................ 51
Figure 3.6. Product distributions generated by base editors with narrowed activity windows ........................................................................ 52
Figure 3.7. Effect of APOBEC1 mutations on the editing efficiency of a target containing a single substrate C ................................................................. 52
Figure 3.8. Effect of post-treatment time on base editing activity window width ......... 54
Figure 3.9. Quantification of catalytic impairment conferred by APOBEC1 mutations ........................................................................ 55
Figure 3.10. Off-target activity of base editors with altered activity window widths .................................................................................. 56
Figure 3.11. Effect of window-narrowing mutations on VQR-BE3 and SaKKH-BE3 ................................................................. 57
Figure 3.12. Genetic variants in ClinVar that can be corrected in principle by base editors with a narrowed activity windows ........................................................................ 59
Figure 4.1. Fusion of Mu-Gam reduces the amount of indel generated during base editing ........................................................................ 64
Figure 4.2. Small molecule-inducible base editing .................................................. 65
Figure 5.1. Cationic lipid-mediated delivery of proteins using anionic protein tags ........................................................................ 69
Figure 5.2. GFP surface charge mostly correlates with delivery efficiency ............. 70
Figure 5.3. Analysis of the human proteome from the UniProt protein database .......... 71
Figure 5.4. ProTα enables potent delivery of Cre in multiple different cell lines ........ 74
Figure 5.5. ProTα potency cannot be recapitulated with poly-Asp/Glu tags ................. 75
Figure 5.6. Truncation analysis of ProTα ................................................................... 77
Figure 5.7. SV40 NLS has insignificant impact on Cre recombination ...................... 77
Figure 5.8. Lycotoxin L15E does not improve Cre delivery into HeLa-DsRed cells ........................................................................ 79
Figure 5.9. ProTα enables efficient delivery of ZFNs into HEK293T cells .................. 81
List of Tables

Table 5.1. List of highly anionic proteins chosen for investigation...............................................72
Table 5.2. Delivery potency of human proteins fused to Cre in HeLa-DsRed cells.........................73
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-HT</td>
<td>4-Hydroxytamoxifen</td>
</tr>
<tr>
<td>APOBEC1</td>
<td>Apolipoprotein B mRNA-editing enzyme 1</td>
</tr>
<tr>
<td>APOE</td>
<td>Apolipoprotein E</td>
</tr>
<tr>
<td>BER</td>
<td>Base excision repair</td>
</tr>
<tr>
<td>CRISPR</td>
<td>Clustered regularly interspaced short palindromic repeats</td>
</tr>
<tr>
<td>DSB</td>
<td>Double-stranded DNA break</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethyl sulfoxide</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>FBS</td>
<td>Fetal bovine serum</td>
</tr>
<tr>
<td>GFP</td>
<td>Green fluorescent protein</td>
</tr>
<tr>
<td>gRNA</td>
<td>Guide RNA</td>
</tr>
<tr>
<td>hAID</td>
<td>Human activation-induced cytidine deaminase</td>
</tr>
<tr>
<td>hAPOBEC3G</td>
<td>Human Apolipoprotein B mRNA-editing enzyme 3G</td>
</tr>
<tr>
<td>HDR</td>
<td>Homology directed repair</td>
</tr>
<tr>
<td>HEK</td>
<td>Human embryonic kidney</td>
</tr>
<tr>
<td>HPLC</td>
<td>High-performance liquid chromatography</td>
</tr>
<tr>
<td>HTS</td>
<td>High-throughput sequencing</td>
</tr>
<tr>
<td>Indel</td>
<td>Insertion or deletion mutation</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>MMR</td>
<td>DNA mismatch repair</td>
</tr>
<tr>
<td>NHEJ</td>
<td>Non-homologous end-joining</td>
</tr>
<tr>
<td>PAM</td>
<td>Protospacer adjacent motif</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein data bank</td>
</tr>
<tr>
<td>ProTα</td>
<td>Prothymosin alpha</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>Sodium dodecyl sulfate polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>sgRNA</td>
<td>Single guide RNA</td>
</tr>
<tr>
<td>siRNA</td>
<td>Small interfering RNA</td>
</tr>
<tr>
<td>TCEP</td>
<td>Tris(2-carboxyethyl)phosphine</td>
</tr>
<tr>
<td>TRIS</td>
<td>Tris(hydroxymethyl)aminomethane</td>
</tr>
<tr>
<td>UDG</td>
<td>Uracil-DNA Glycosylase</td>
</tr>
<tr>
<td>UGI</td>
<td>Uracil Glycosylase Inhibitor</td>
</tr>
<tr>
<td>ZFN</td>
<td>Zinc finger nuclease</td>
</tr>
</tbody>
</table>
Introduction Part I: Genome Editing

The clustered regularly interspaced short palindromic repeat (CRISPR) system has been used to edit the genomes of a variety of organisms\textsuperscript{1,2}. In particular, the Cas9 protein complexed with a single guide RNA (sgRNA) binds and cleaves the target DNA, generating a double-stranded DNA break (DSB)\textsuperscript{3}. In response to a DSB, non-homologous end joining (NHEJ), a predominant DNA repair pathway, occurs at the site of DNA cleavage, which results in random insertions or deletions (indels), effectively disrupting the gene\textsuperscript{4}. In the presence of a homologous DNA template, the region around the cleavage site may be replaced through homology-directed repair (HDR)\textsuperscript{4}. However, as NHEJ is more efficient than HDR during the resolution of DSBs, indels are generally more abundant outcomes than gene replacement\textsuperscript{2,4}. Unfortunately, most known genetic diseases are caused by a point mutation in the target locus\textsuperscript{5}. As such, precise correction of a single nucleotide, rather than stochastic disruption of the gene, is needed to study or address the underlying cause of the disease\textsuperscript{5}.

Many have sought to increase the HDR efficiency and suppress NHEJ\textsuperscript{6-9}. Yet, HDR remains inefficient under therapeutically relevant conditions (typically 5-10%), especially in unmodified, non-dividing cells\textsuperscript{6-9}. The shortcomings of HDR highlight the need for alternative approaches to correct point mutations in genomic DNA that do not require DSBs.

This thesis outlines one such technology, termed ‘base editing’. Base editing enables the conversion of one base into another, without generating DSBs in the process. We report various improvements to the base editing technology to improve its targeting scope, precision, spatiotemporal control, and safety.
Introduction Part II: Cytosolic Protein Delivery

Peptide and protein therapeutics, including cytokines, hormones, and monoclonal antibodies, have gained much support over the past two decades\textsuperscript{10-13}. There are over 60 approved peptide drugs worldwide, and over 100 ongoing clinical trials\textsuperscript{14}. By 2018, peptide drug market is expected to reach $25 billion\textsuperscript{10}.

Despite the fanfare, current macromolecular therapeutics are limited to targeting extracellular components since proteins cannot spontaneously enter cells\textsuperscript{11,15}. Their inability to translocate through the lipid bilayer has largely hampered the targeting of therapeutically relevant intracellular targets\textsuperscript{16,17}.

Tools that facilitate proteins to access the cytosol have been developed. Positively charged peptide tags such as TAT proteins or poly-Arginines\textsuperscript{18}, or more recently engineered superpositively charged GFP\textsuperscript{15}, allow the fused protein to localize to the cell membrane by interacting with the highly anionic peptidoglycans on the cell surface, followed by endocytosis\textsuperscript{19}. Yet, delivery of naked proteins has disadvantages, as they are susceptible to proteases, or be neutralized by serum proteins, blood cells, and the extracellular matrix\textsuperscript{20}.

Lipid membranes may be used to protect the complexed nucleic acids from degradation and antibody neutralization\textsuperscript{21}. Furthermore, liposomal endocytosis results in a more efficient endosomal escape into the cytosol\textsuperscript{22}. Traditionally, cationic lipids have been used to transfect DNA- or RNA-based reagents, such as plasmids and siRNAs. Proteins, unlike nucleic acids, usually lack the anionic charges that promote encapsulation by cationic lipids. To overcome this challenge, Zuris \textit{et al.} demonstrated that fusing an anionic protein tag, such as an engineered GFP with surface-exposed negative charges ((-30)GFP), to protein of interest enables its efficient encapsulation and delivery into mammalian cells\textsuperscript{23}. 
While (-30)GFP is a useful tool, a greater negative charge may enable even more efficient complexation and delivery using cationic lipids. Moreover, due to its bacterial origin, (-30)GFP elicits immunogenicity \textit{in vivo}, compromising the safety and utility of the method\textsuperscript{24}.

To address both limitations, we sought to identify a native human protein that is highly negatively charged, which may replace (-30)GFP as the fusion tag. Since it is of human origin, the risk of an immune response \textit{in vivo} is low. Chapter Five describes a screen of highly anionic human proteins for their ability to deliver cargo into the cytosol of the mammalian cell. We identify a relatively unknown protein, ProTα, that outperforms (-30)GFP in delivery assays.
Chapter One:  
Development of Base Editing

Alexis C. Komor, Y. Bill Kim, Michael S. Packer, John A. Zuris, and David R. Liu

Alexis Komor led the project. Alexis Komor generated constructs, performed experiments, and analyzed data. Bill Kim performed experiments and analyzed data. Michael Packer performed computational analyses.

1.1 Introduction

We envisioned that direct conversion of one DNA base to another at a programmable target locus without requiring DSBs could increase the efficiency of gene correction relative to HDR without introducing an excess of random indels. Catalytically dead Cas9 (dCas9), which contains Asp10Ala and His840Ala mutations that inactivate its nuclease activity, retains its ability to bind DNA in a guide RNA-programmed manner, but does not cleave the DNA backbone\(^3\). In principle, conjugation of dCas9 with an enzymatic or chemical catalyst that mediates the direct conversion of one base to another could enable RNA-programmed DNA base editing.

The deamination of cytosine (C) is catalysed by cytidine deaminases\(^25\) and results in uracil (U), which has the base-pairing properties of thymine (T). Most known cytidine deaminases operate on RNA, and the few examples that are known to accept DNA require single-stranded (ss) DNA\(^26\). Recent studies on the dCas9–target DNA complex reveal that at least nine nucleotides of the displaced DNA strand are unpaired upon formation of the Cas9:guide RNA:DNA ‘R-loop’ complex\(^27\). Indeed, in the structure of the Cas9 R-loop complex the first 11 nucleotides of the protospacer on the displaced DNA strand are disordered, suggesting that their movement is not highly restricted\(^28\). We reasoned that a subset of this stretch of ssDNA in the R-loop might serve as an efficient substrate for a dCas9-tethered cytidine deaminase to effect direct, programmable conversion of C to U in DNA (Figure 1.1).
Figure 1.1. Overview of base editing. Base editing enables programmable conversion of C to T (or G to A) in DNA. A base editor is composed of a modular DNA binding domain, such as a deactivated Cas9, fused with a single-stranded DNA-specific cytidine deaminase. Once dCas9 binds to the target site through gRNA-DNA interaction and PAM recognition, the deaminase accesses the single-stranded DNA that is not base paired with the gRNA (target strand) and catalyzes C to U conversion. Since U base pairs with A, subsequent DNA replication or repair generates an A-T base pair.

1.2 First-Generation Base Editor, BE1.

We screened four different cytidine deaminases (hAID, hAPOBEC3G, rAPOBEC1, pmCDA1) and optimized the linker architecture between the deaminase and dCas9. Rattus
norvegicus APOBEC1 (rAPOBEC1) resulted in the most efficient deamination in vitro (Figure 1.2), and a 16-amino acid stretch of a flexible linker, XTEN\textsuperscript{29}, provided the highest activity while confining the deamination to a ~5-nucleotide window within position 4-8 in the protospacer (NGG PAM is set to position 21-23) (Figure 1.3). This is termed the first-generation base editor, BE1.

We tested the ability of BE1 to convert C→T in human cells on 14 Cs in six well-studied target sites in the human genome. Although C→T editing in cells was observed for all cases, the efficiency of base editing was 0.8% to 7.7% of total DNA sequences (Figure 1.4).

---

**Figure 1.2. Evaluation of different cytidine deaminase enzymes.** Four candidate cytidine deaminases, hAID, hAPOBEC3G, rAPOBEC1, and pmCDA1, were in vitro transcribed and translated. They were subjected to a Cy3-tagged DNA substrate containing a target C. Deamination was detected by subsequent treatment with the USER enzyme mix (Uracil DNA glycosylase (UDG) and Endonuclease VIII). Only rAPOBEC1 resulted in significant conversion of C to U, generating a product that was cleaved by the USER reaction.
Figure 1.3. Effect of different linker lengths and composition on deaminase-dCas9 fusion. A gel-based deaminase assay showing the deamination window of base editors with deaminase-Cas9 linkers composed of (GGS)$_1$ (a), (GGS)$_3$ (b), XTEN (c), or (GGS)$_7$ (d). The position in which the target C is incorporated is shown above the lanes. XTEN is a flexible 16-amino acid linker described previously$^{29}$. 8U is a positive control sequence with a U synthetically incorporated at position 8. The XTEN linker provides the highest activity while having a relatively defined deamination window (approximately from position 4-8).
Figure 1.4. Editing efficiencies of BE1, BE2, BE3, and Cas9+HDR in genomic sites in HEK293T cells. Base editing allows for highly efficient conversion of Cs to Ts in multiple genomic loci in tissue culture. BE2 (orange) is on average ~3-fold more efficient than BE1 (red) and BE3 (green) further improves the efficiency by ~5-fold.

1.3 Second-Generation Base Editor, BE2

We hypothesized that the cellular DNA repair response to U:G heteroduplex DNA was responsible for the large decrease in base editing efficiency in cells. Uracil DNA glycosylase (UDG) catalyzes removal of U from DNA in cells and initiates base-excision repair (BER), with reversion of the U:G pair to a C:G pair as the most common outcome\(^\text{30}\). Uracil DNA glycosylase inhibitor (UGI), an 83-residue protein from *Bacillus subtilis* bacteriophage PBS1, potently blocks human UDG activity (IC\(_{50} = 12\) pM)\(^\text{31}\). In an effort to subvert BER at the site of base editing, we fused UGI to the C terminus of BE1 to create a second-generation base editor (BE2, APOBEC1–XTEN–dCas9–UGI) and repeated editing assays on all six genomic loci. Editing
efficiencies in human cells were on average threefold higher with BE2 than BE1, resulting in gene conversion efficiencies of up to 20% of total DNA sequenced (Figure 1.4).

1.4 Third-Generation Base Editor, BE3

Converting and protecting the substrate strand of a C:G base pair (bp) results in a maximum base editing yield of 50%. To augment base editing efficiency beyond this limit, we sought to further manipulate cellular DNA repair to induce correction of the non-edited strand containing the G. Eukaryotic mismatch repair (MMR) uses nicks present in newly synthesized DNA to direct removal and resynthesis of the newly synthesized strand\(^{32,33}\). We reasoned that nicking the DNA strand containing the unedited G would simulate newly synthesized DNA, inducing MMR to preferentially resolve the U:G mismatch into desired U:A and T:A products. We therefore restored the catalytic His residue at position 840 in the Cas9 HNH domain of BE2, resulting in the third-generation base editor (BE3, APOBEC1–XTEN–dCas9(A840H)–UGI) that nicks the non-edited strand containing a G opposite the edited U. BE3 retains the Asp10Ala mutation in Cas9 that prevents dsDNA cleavage, and also retains UGI to suppress BER.

Nicking the non-edited strand augmented base editing efficiency in human cells treated with BE3 by an additional two- to sixfold relative to BE2, resulting in up to 37% of total DNA sequences containing the targeted C to T conversion (Figure 1.4). Importantly, only a small frequency of indels, averaging 1.1% for the six tested loci, was observed from BE3 treatment (Figure 1.5). In contrast, when we treated cells with wild-type (WT) Cas9, sgRNA to target each of three loci, and a ssDNA donor template to mediate HDR, we observed C→T conversion efficiencies averaging only 0.5%, with much higher indel formation averaging 4.3%. The ratio of allele conversion to NHEJ outcomes averaged >1,000 for BE2, 23 for BE3, and 0.17 for wild-
type Cas9 (Figure 1.4 and Figure 1.5). We confirmed the permanence of base editing in human cells by monitoring editing efficiencies over multiple cell divisions in HEK293T cells at the HEK293 site 3 and 4 genomic loci (Figure 1.6). These results collectively establish that base editing can effect much more efficient targeted single-base editing in human cells than Cas9-mediated HDR, and with much less (BE3) or almost no (BE2) indel formation.

**Figure 1.5. Indel generated by BE1, BE2, BE3, and Cas9+HDR.** Base editing generates on average 1-5 % indel in HEK293T cells. On the other hand, Cas9+HDR generates a large amount of indel despite low HDR efficiencies.

**Figure 1.6. Permanence of base editing in human cells.** Base editing induced by BE2 or BE3 is maintained through multiple cell divisions and passaging in tissue culture.
1.5 Correction of Disease-Relevant Mutations Using BE3

We tested the potential of base editing to correct two disease-relevant mutations in mammalian cells. The apolipoprotein E gene variant APOE4 encodes two arginine residues at amino acid positions 112 and 158, and is the largest and most common genetic risk factor for late-onset Alzheimer’s disease. ApoE variants with Cys residues at these positions, including APOE2 (Cys112/Cys158), APOE3 (Cys112/Arg158), and APOE3r (Arg112/Cys158) have been shown or are presumed to confer lower Alzheimer’s disease risk than APOE4. We attempted to convert APOE4 into APOE3r in immortalized mouse astrocytes in which the endogenous APOE gene was replaced by human APOE4. We delivered into these astrocytes by nucleofection DNA encoding BE3 and an appropriate sgRNA placing the target C at position 5 relative to a downstream PAM. After two days, we isolated nucleofected cells and measured editing efficiency by HTS of genomic DNA. We observed conversion of Arg158 to Cys158 in 75% of total DNA sequencing reads (Figure 1.7). We also observed 50% editing of total DNA at the third position of codon 158 and 55% editing of total DNA at the first position of Leu159, as expected since all three of these Cs are within the base editing window. Neither of the other two C→T conversions, however, alters the amino acid sequence of the ApoE3r protein, as both TGC and TGT encode Cys (all C→T changes at the third position of a codon are silent), and both CTG and TTG encode Leu.

The efficiency of BE3-mediated editing of APOE4 demonstrates that a combination of suppressing BER and guiding MMR to repair the unedited strand enables base editing efficiencies to exceed the 50% maximum yield that would result from DNA replication alone. We observed 4.6% indels at the targeted locus following BE3 treatment (Figure 1.7). In contrast, identical treatment of astrocytes with wild-type Cas9 and donor ssDNA resulted in 0.1% APOE4
correction and 26% indels at the targeted locus, efficiencies consistent with previous reports of single-base correction using Cas9 and HDR\textsuperscript{36,37} (Figure 1.7). These results demonstrate that base editors can mediate highly efficient and precise single amino acid changes in the coding sequence of a protein, even when their processivity results in >1 nucleotide change in genomic DNA.

The dominant-negative p53 mutation Tyr163Cys is strongly associated with several types of cancer\textsuperscript{38} and can be corrected by a C→T conversion on the template strand, resulting in the translation of corrected protein even before the edited base is made permanent by DNA replication or DNA repair. We nucleofected a human breast cancer cell line homozygous for the p53 Tyr163Cys mutation (HCC1954 cells) with DNA encoding BE3 and an sgRNA programmed to correct Tyr163Cys. We observed correction of the Tyr163Cys mutation in 7.6% of nucleofected HCC1954 cells (Figure 1.7), with 0.7% indel formation. In contrast, treatment of cells with wild-type Cas9 and donor ssDNA resulted in no detectable TP53 correction (<0.1%) with 6.1% indels at the target locus (Figure 1.7). These results collectively represent the correction of disease-associated point mutations in mammalian cell lines with an efficiency and lack of other genome modification events that may not be achievable using previously described methods.
Figure 1.7. BE3-mediated correction of disease-relevant mutations in mammalian cells. Both APOE4 and TP53 Y163C could be reverted to APOE3r and WT TP53, respectively, using base editing, without generating an excess of indels.

1.6 Conclusion

The development of base editing advances both the scope and effectiveness of genome editing. The base editors described here offer researchers a choice of editing with very little (<0.1%) indel formation (BE2), or more efficient editing with < 1% indel formation (BE3). That the product of base editing is, by definition, no longer a substrate likely contributes to editing efficiency by preventing subsequent product transformation, which can hamper traditional Cas9 applications. By removing the reliance on dsDNA cleavage, donor templates, and stochastic DNA repair processes that vary by cell state and cell type, base editing has the potential to
expand the type of genome modifications that can be cleanly installed, the efficiency of these modifications, and the type of cells that are amenable to editing.

1.7 Methods

**Cloning.** PCR was performed using VeraSeq ULtra DNA polymerase (Enzymatics), or Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs). BE plasmids were constructed using USER cloning (New England Biolabs). Deaminase genes were synthesized as gBlocks Gene Fragments (Integrated DNA Technologies), and Cas9 genes were obtained from previously reported plasmids\(^39\). Deaminase and fusion genes were cloned into pCMV (mammalian codon-optimized) or pET28b (E. coli codon-optimized) backbones. sgRNA expression plasmids were constructed using site-directed mutagenesis. Briefly, the primers were 5’ phosphorylated using T4 Polynucleotide Kinase (New England Biolabs) according to the manufacturer’s instructions. Next, PCR was performed using Q5 Hot Start High-Fidelity Polymerase (New England Biolabs) with the phosphorylated primers and the plasmid pFYF1320 (EGFP sgRNA expression plasmid) as a template according to the manufacturer’s instructions. PCR products were incubated with DpnI (20 U, New England Biolabs) at 37 °C for 1 h, purified on a QIAprep spin column (Qiagen), and ligated using QuickLigase (New England Biolabs) according to the manufacturer’s instructions. DNA vector amplification was carried out using Mach1 competent cells (ThermoFisher Scientific).

**In vitro deaminase assay on ssDNA.** All Cy3-labelled substrates were obtained from Integrated DNA Technologies (IDT). Deaminases were expressed *in vitro* using the TNT T7 Quick Coupled Transcription/Translation Kit (Promega) according to the manufacturer’s instructions. Following protein expression, 5 µl of lysate was combined with 35 µl of ssDNA (1.8 µM) and
USER enzyme (1 unit) in CutSmart buffer (New England Biolabs) (50 mM potassium acetate, 29 mM Tris-acetate, 10 mM magnesium acetate, 100 µg ml⁻¹ BSA, pH 7.9) and incubated at 37 °C for 2 h. Cleaved U-containing substrates were resolved from full-length unmodified substrates on a 10 % TBE-urea gel (Bio-Rad).

Expression and purification of His₆–rAPOBEC₁-linker–dCas9 fusions. E. coli BL21 STAR (DE3)-competent cells (ThermoFisher Scientific) were transformed with plasmids encoding pET28b-His₆-rAPOBEC₁-linker-dCas9 with GGS, (GGS)₃, XTEN, or (GGS)₇ linkers. The resulting expression strains were grown overnight in Luria-Bertani (LB) broth containing 100 µg ml⁻¹ of kanamycin at 37 °C. The cells were diluted 1:100 into the same growth medium and grown at 37 °C to OD₆₀₀ = ~0.6. The culture was cooled to 4 °C over a period of 2 h, and isopropyl-β-D-1-thiogalactopyranoside (IPTG) was added at 0.5 mM to induce protein expression. After ~16 h, the cells were collected by centrifugation at 4,000g and resuspended in lysis buffer (50 mM tris(hydroxymethyl)-aminomethane (Tris)-HCl (pH 7.0), 1 M NaCl, 20% glycerol, 10 mM tris(2-carboxyethyl)phosphine (TCEP, Soltec Ventures)). The cells were lysed by sonication (20 s pulse-on, 20 s pulse-off for 8 min total at 6 W output) and the lysate supernatant was isolated following centrifugation at 25,000g for 15 min. The lysate was incubated with His-Pur nickel-nitrioloacetic acid (nickel-NTA) resin (ThermoFisher Scientific) at 4 °C for 1 h to capture the His-tagged fusion protein. The resin was transferred to a column and washed with 40 ml of lysis buffer. The His-tagged fusion protein was eluted in lysis buffer supplemented with 285 mM imidazole, and concentrated by ultrafiltration (Amicon-Millipore, 100-kDa molecular weight cut-off) to 1 ml total volume. The protein was diluted to 20 ml in low-salt purification buffer containing 50 mM tris(hydroxymethyl)-aminomethane (Tris)-HCl (pH 7.0), 0.1 M NaCl, 20% glycerol, 10 mM TCEP and loaded onto SP Sepharose Fast Flow
resin (GE Life Sciences). The resin was washed with 40 ml of this low-salt buffer, and the protein eluted with 5 ml of activity buffer containing 50 mM tris(hydroxymethyl)-aminomethane (Tris)-HCl (pH 7.0), 0.5 M NaCl, 20% glycerol, 10 mM TCEP. The eluted proteins were quantified by SDS–PAGE.

*In vitro transcription of sgRNAs.* Linear DNA fragments containing the T7 promoter followed by the 20-bp sgRNA target sequence were transcribed using the TranscriptAid T7 High Yield Transcription Kit (ThermoFisher Scientific) according to the manufacturer’s instructions. sgRNA products were purified using the MEGAclear Kit (ThermoFisher Scientific) according to the manufacturer’s instructions and quantified by UV absorbance.

**Preparation of Cy3-conjugated dsDNA substrates.** Sequences of 80-nt unlabelled strands were ordered as PAGE-purified oligonucleotides from IDT. The 25-nt Cy3-labelled primer is complementary to the 3’ end of each 80-nt substrate. This primer was ordered as an HPLC-purified oligonucleotide from IDT. To generate the Cy3-labelled dsDNA substrates, the 80-nt strands (5 µl of a 100 µM solution) were combined with the Cy3-labelled primer (5 µl of a 100 µM solution) in NEBuffer 2 (38.25 µl of a 50 mM NaCl, 10 mMTris-HCl, 10 mM MgCl2, 1 mM DTT, pH 7.9 solution, New England Biolabs) with dNTPs (0.75 µl of a 100 mM solution) and heated to 95 °C for 5 min, followed by a gradual cooling to 45 °C at a rate of 0.1 °C per s. After this annealing period, Klenow exo⁻ (5 U, New England Biolabs) was added and the reaction was incubated at 37 °C for 1 h. The solution was diluted with Buffer PB (250 µL, Qiagen) and isopropanol (50 µl) and purified on a QIaprep spin column (Qiagen), eluting with 50 µl of Tris buffer.

**Deaminase assay on dsDNA.** The purified fusion protein (20 µl of 1.9 µM in activity buffer) was combined with 1 equivalent of appropriate sgRNA and incubated at ambient temperature for
5 min. The Cy3-labelled dsDNA substrate was added to final concentration of 125 nM and the resulting solution was incubated at 37 °C for 2 h. The dsDNA was separated from the fusion by the addition of Buffer PB (100 µl, Qiagen) and isopropanol (25 µl) and purified on an EconoSpin micro spin column (Epoch Life Science), eluting with 20 µl of CutSmart buffer (New England Biolabs). USER enzyme (1 U, New England Biolabs) was added to the purified, edited dsDNA and incubated at 37 °C for 1 h. The Cy3-labelled strand was fully denatured from its complement by combining 5 µl of the reaction solution with 15 µl of a DMSO-based loading buffer (5 mM Tris, 0.5 mM EDTA, 12.5 % glycerol, 0.02 % bromophenol blue, 0.02 % xylene cyan, 80 % DMSO). The full-length C-containing substrate was separated from any cleaved, U-containing edited substrates on a 10 % TBE-urea gel (Bio-Rad) and imaged on a GE Amersham Typhoon imager.

**Cell culture.** HEK293T (ATCC CRL-3216) and U2OS (ATCC-HTB-96) were maintained in Dulbecco’s Modified Eagle’s Medium plus GlutaMax (ThermoFisher) supplemented with 10% (v/v) fetal bovine serum (FBS), at 37 °C with 5% CO2. HCC1954 cells (ATCC CRL-2338) were maintained in RPMI-1640 medium (ThermoFisher Scientific) supplemented as described above. Immortalized rat astrocytes containing the ApoE4 isoform of the APOE gene (Taconic Biosciences) were cultured in Dulbecco’s Modified Eagle’s Medium plus GlutaMax (ThermoFisher Scientific) supplemented with 10% (v/v) fetal bovine serum (FBS) and 200 µg ml⁻¹ Geneticin (ThermoFisher Scientific).

**Transfections.** HEK293T cells were seeded on 48-well collagen-coated BioCoat plates (Corning) and transfected at approximately 85% confluency. Briefly, 750 ng of BE and 250 ng of sgRNA expression plasmids were transfected using 1.5 µl of Lipofectamine 2000 (ThermoFisher Scientific) per well according to the manufacturer’s protocol.
Astrocytes, U2OS, HCC1954 and HEK293T cells were transfected using appropriate Amaxa Nucleofector II programs according to manufacturer’s instructions (basic glial cell, V, V, and V kits using programs T-020, X-001, X-005, and Q-001 for astrocytes, U2OS, HCC1954, and HEK293T cells, respectively). 40 ng of iRFP670 (Addgene plasmid 45457) was added to the nucleofection solution to assess nucleofection efficiencies in these cell lines. Astrocytes and HCC1954 cells were filtered through a 40 µm strainer (Fisher Scientific) after harvesting, and the nucleofected cells were collected on a Beckman Coulter MoFlo XDP Cell Sorter using the iRFP signal (abs. 643 nm, em. 670 nm). The U2OS and HEK293T cells were used without enrichment of nucleofected cells.

**High-throughput DNA sequencing of genomic DNA samples.** Transfected cells were harvested after 3 days and the genomic DNA was isolated using the Agencourt DNAAdvance Genomic DNA Isolation Kit (Beckman Coulter) according to the manufacturer’s instructions. On-target and off-target genomic regions of interest were amplified by PCR with flanking high-throughput sequencing primer pairs. PCR amplification was carried out with Phusion high-fidelity DNA polymerase (ThermoFisher) according to the manufacturer’s instructions using 5 ng of genomic DNA as a template. Cycle numbers were determined separately for each primer pair as to ensure the reaction was stopped in the linear range of amplification (30, 28, 28, 28, 32, and 32 cycles for EMX1, FANCF, HEK293 site 2, HEK293 site 3, HEK293 site 4, and RNF2 primers, respectively). PCR products were purified using RapidTips (Diffinity Genomics).

Purified DNA was amplified by PCR with primers containing sequencing adaptors. The products were gel-purified and quantified using the Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher Scientific) and KAPA Library Quantification Kit-Illumina (KAPA Biosystems). Samples were sequenced on an Illumina MiSeq as previously described.
Data analysis. Sequencing reads were automatically demultiplexed using MiSeq Reporter (Illumina), and individual FASTQ files were analysed with a custom Matlab script (see below). Each read was pairwise aligned to the appropriate reference sequence using the Smith-Waterman algorithm. Base calls with a Q-score below 31 were replaced with Ns and were thus excluded in calculating nucleotide frequencies. This treatment yields an expected MiSeq base-calling error rate of approximately 1 in 1,000. Aligned sequences in which the read and reference sequence contained no gaps were stored in an alignment table from which base frequencies could be tabulated for each locus.

Indel frequencies were quantified with a custom Matlab script (see below). Sequencing reads were scanned for exact matches to two 10-bp sequences that flank both sides of a window in which indels might occur. If no exact matches were located, the read was excluded from analysis. If the length of this indel window exactly matched the reference sequence the read was classified as not containing an indel. If the indel window was two or more bases longer or shorter than the reference sequence, then the sequencing read was classified as an insertion or deletion, respectively.

Base Calling Matlab Script.

```matlab
WTnuc='GCGGACATGGAGGACGTGCAGTACCGGCAGAGGTGACGCTGCGGAGGGCGCCGCTGGTGCAGTACCGCGGCGAGGTGCAGGCCATGCTCGGCCAGACGACCGAGGAGCTGCGGGTGCGCCTCGCCTCCCACCTGCGCAAGCTGCGTAAGCGGCTCCTCCGCGATGCCGATGACCTGCAGAAGCGCCTGGCAGTGTACCAGGCCGGGGCCCGCGAGGGCGCCGAGCGCGGCCTCAGCGCCATCCGCGAGCGCTGGGGCCCCTGGTGGAACAG';
% cycle through fastq files for different samples
files=dir('*.*.fastq');
for d=1:20
    filename=files(d).name;
    % read fastq file
    [header,seqs,qscore] = fastqread(filename);
    seqsLength = length(seqs); % number of sequences
    seqsFile = strrep(filename,'.fastq',''); % trims off .fastq
    % create a directory with the same name as fastq file
    if exist(seqsFile,'dir');
        error('Directory already exists. Please rename or move it before moving on.');
    end
    mkdir(seqsFile); % make directory
```
wtLength = length(WTnuc); % length of wildtype sequence  
%% aligning back to the wildtype nucleotide sequence
%  
% AlN is a matrix of the nucleotide alignment 
window=1:wtLength;

sBLength = length(seqs); % number of sequences  
% counts number of skips
nSkips = 0;

ALN=repmat(' ',[sBLength wtLength]); % iterate through each sequencing read
for i = 1:sBLength
  if you only have forward read fastq files leave as is
  %If you have R1 foward and R2 is reverse fastq files uncomment the %next four lines of code and the subsequent end statement
  % if mod(d,2)==0;
  %     reverse = seqrcomplement(seqs{i});
  % [score,alignment,start] = swalign(reverse,WTnuc,'Alphabet','NT');
  % else
  %     [score,alignment,start] = swalign(seqs{i},WTnuc,'Alphabet','NT');
  % end
  % length of the sequencing read
  len = length(alignment(3,:));
  % if there is a gap in the alignment , skip = 1 and we will % throw away the entire read
  skip = 0;
  for j = 1:len
      if (alignment(3,j) == '-' || alignment(1,j) == '-')
       skip = 1;
       break;
    end
    %in addition if the qscore for any given base in the read is %below 31 the nucleotide is turned into an N (fastq qscores that are not letters)
    if isletter(qscore{i}(start(1)+j-1))
    else
       alignment(1,j) = 'N';
     end
  end
  if skip == 0 && len>10
      ALN(i, start(2):(start(2)+length(alignment)-1))=alignment(1,:); 
  end
end
% with the alignment matrices we can simply tally up the occurrences of % each nucleotide at each column in the alignment these % tallies ignore bases annotated as N % due to low qscores
TallyNTD=zeros(5,wtLength);
for i=1:wtLength
TallyNTD(:,i)=[sum(ALN(:,i)=='A'),sum(ALN(:,i)=='C'),sum(ALN(:,i)=='G'),sum(ALN(:,i)=='T'),sum(ALN(:,i)=='N')];
end
we then save these tally matrices in the respective folder for further processing

```
save(strcat(seqsFile, '/TallyNTD'), 'TallyNTD');
dlmwrite(strcat(seqsFile, '/TallyNTD.txt'), TallyNTD, 'precision', '%.3f', 'newline', 'pc');
```

**INDEL Detection Matlab Script.**

```
WTnuc='GCGGACATGGAGGACGTGCGCGGCCCTGGTGCAGTACCGCGGCGAGGTGCAGGCCATGCTCGGCCAGA
GCACCCAGGAGGCTGGGCTGCCCTCGCCTCCACCTGGCAGGCGGCTAAGCGGCTCCTCCCGGATGCCGATGAC
CTGCAGAAGCGCCTGGCAGTGTACCAGGCCGGGGCCCGCGAGGGCGCCGAGCGCGGCCTCAGCGCCATCCGCGAGCG
CCTGGGGCCCTCGTGGGAAACAG';
% cycle through fastq files for different samples
files=dir('*.fastq');
% specify start and width of indel window as well as length of each flank
indelstart=154;
width=30;
flank=10;
for d=1:3
    filename=files(d).name;
    % read fastq file
    [header,seqs,qscore] = fastqread(filename);
    seqsLength = length(seqs); % number of sequences
    seqsFile = strcat(strrep(filename,'.fastq',''),'_INDELS');
    % create a directory with the same name as fastq file+_INDELS
    if exist(seqsFile,'dir');
        error('Directory already exists. Please rename or move it before moving on. ');
    end
    mkdir(seqsFile); % make directory
    wtLength = length(WTnuc); % length of wildtype sequence
    sBLength = length(seqs); % number of sequences
    % initialize counters and cell arrays
    nSkips = 0;
    notINDEL=0;
    ins={};
    dels={};
    NumIns=0;
    NumDels=0;
    % iterate through each sequencing read
    for i = 1:sBLength
        % search for 10BP sequences that should flank both sides of the "INDEL WINDOW"
        windowstart=strfind(seqs{i},WTnuc(indelstart-flank:indelstart));
        windowend=strfind(seqs{i},WTnuc(indelstart+width:indelstart+width+flank ));
        % if the flanks are found proceed
        if length(windowstart)==1 & length(windowend)==1
            % if the sequence length matches the INDEL window length save as not INDEL
```
if windowend-windowstart==width+flank
    notINDEL=notINDEL+1;
%if the sequence is two or more bases longer than the INDEL %window length save as an Insertion
elseif windowend-windowstart>=width+flank+2
    NumIns=NumIns+1;
    ins{NumIns}=seqs{i};
%if the sequence is two or more bases shorter than the INDEL %window length save as a Deletion
elseif windowend-windowstart<=width+flank-2
    NumDels=NumDels+1;
    dels{NumDels}=seqs{i};
%keep track of skipped sequences that are either one base %shorter or longer than the INDEL window width
else
    nSkips=nSkips+1;
end
%keep track of skipped sequences that do not possess matching flank %sequences
else
    nSkips=nSkips+1;
end

fid=fopen(strcat(seqsFile, '/summary.txt'), 'wt');
fprintf(fid, 'Skipped reads %i
not INDEL %i
Insertions %i
Deletions %i
', [nSkips, notINDEL, NumIns, NumDels]);
fclose(fid);
save(strcat(seqsFile, '/nSkips'), 'nSkips');
save(strcat(seqsFile, '/notINDEL'), 'notINDEL');
save(strcat(seqsFile, '/NumIns'), 'NumIns');
save(strcat(seqsFile, '/NumDels'), 'NumDels');
save(strcat(seqsFile, '/dels'), 'dels');
C = dels;
fid = fopen(strcat(seqsFile, '/dels.txt'), 'wt');
fprintf(fid, '"%s"
', C{:});
fclose(fid);
save(strcat(seqsFile, '/ins'), 'ins');
C = ins;
 fid = fopen(strcat(seqsFile, '/ins.txt'), 'wt');
fprintf(fid, '"%s"
', C{:});
fclose(fid);

end
Chapter Two:
Improving the Targeting Scope of Base Editing

Y. Bill Kim, Alexis C. Komor, Jonathan M. Levy, Michael S. Packer, Kevin T. Zhao, and David R. Liu

Bill Kim generated most constructs and conducted all experiments and analyzed all data. Alexis Komor and Jonathan Levy generated some constructs. Michael Packer assisted with computational analyses. Kevin Zhao conducted some experiments.

2.1 Introduction: Targeting Scope in Base Editing

Efficient editing by BE3 requires the presence of an NGG PAM that places the target C within a 5-nucleotide window near the PAM-distal end of the protospacer (positions 4–8, counting the PAM as positions 21–23)\(^{42}\). This PAM requirement limits the number of sites in the human genome that can be efficiently targeted by BE3, as many sites lack an NGG sequence 13–17 nucleotides downstream of the target C. This chapter describes the development of several new C:G to T:A base editors that target alternate PAM sequences and thereby substantially expand the targets suitable for base editing.

2.2 Utilizing Cas9 Homologs for Base Editing

We hypothesized that any Cas9 homolog that binds DNA and forms an ‘R-loop’ complex\(^ {28}\) containing a single-stranded DNA bubble could in principle be converted into a base editor. The Cas9 homolog from *Staphylococcus aureus* (SaCas9) can mediate efficient genome editing in mammalian cells and requires an NNGRRT PAM\(^ {43}\). We replaced the nickase form of SpCas9 with that of SaCas9 in BE3 to generate APOBEC1–SaCas9n–UGI (SaBE3), and transfected HEK293T cells with plasmids encoding SaBE3 and sgRNAs targeting six endogenous human genomic loci (Figure 2.1a, b). After 3 d, we used high-throughput DNA sequencing (HTS) to quantify base editing efficiency. SaBE3 very efficiently edited target Cs at human genomic loci, with conversion efficiencies of ~50–75% of total DNA sequences converted from C to T, without enrichment for transfected cells (Figure 2.2a). The efficiency of SaBE3 on NNGRRT-containing target sites in general exceeded that of BE3 on NGG-containing target sites\(^ {42}\). Perhaps due to greater solvent exposure of the strand not paired with the guide RNA\(^ {44}\), editing with SaBE3 can also result in detectable base editing at target Cs at positions
outside of the canonical BE3 activity window (Figure 2.2). By comparison, BE3 showed markedly reduced editing for the same non-NGG target sites (0–11 % editing), consistent with the PAM requirement of SpCas9 (Figure 2.3a)\(^45\). These data show that SaBE3 can mediate base editing at sites not accessible to BE3.

![Figure 2.1. Base editors generated with Cas9 homologs and engineered variants.](image)

**(a)** Architecture of BE3 (above) and BE4 (below). BE4 has an extended linker between APOBEC1 and CAS9n, and two copies of UGI for improved suppression of UDG. **(b)** A list of active base editors with altered PAM specificities developed in this thesis.
Figure 2.2. Base editing with SaBE3 and SaKKH-BE3. (a) Efficient base editing by SaBE3 and SaKKH-BE3 at sites containing the NNGRRT PAM. (b) Only SaKKH-BE3 displays efficient base editing at sites containing the NNHRRT PAM.
Figure 2.3. BE3 activity at sites targeted by Cas9 homologs and engineered variants. BE3 cannot efficiently edit genomic sites lacking an NGG PAM. BE3 activity at sites containing NNGRRT, NNHRRT (a), NGA, and NGCG PAMs (b).

Previously, efficient genome editing has been achieved using a Cas9 from 
Campylobacter jejuni (CjCas9)\textsuperscript{46,47}. Based on the previously reported PAM-discovery assays, we deduced the most stringent PAM requirement for CjCas9 to be NNNVRYAC\textsuperscript{46,47}. The nickase form of CjCas9 was cloned into the BE3 architecture to generate CjBE3. Recently, a base editor with increased efficiency and greater product purity, termed BE4, has been reported\textsuperscript{48}. BE4 has
two copies of UGI (instead of one in BE3) to more efficiently inhibit cellular UDG, and has a longer linker between APOBEC1 and nickase Cas9 (Figure 2.1a)\textsuperscript{48}. CjBE4 was constructed by inserting CjCas9 into the BE4 architecture. Plasmids encoding CjBE3 or CjBE4 were transfected into HEK293T cells with sgRNAs targeting genomic loci with NNNVRYAC PAMs. After 3 d, base editing efficiency was quantified using HTS. We observed efficient base editing up to 70\% of transfected cells using both CjBE3 and CjBE4, but did not observe a significant difference in the activity between the two (Figure 2.4a). Moreover, editing efficiencies among multiple genomic sites varied significantly. We suspect that this is due to the poor helicase activity of CjCas9 compared to \textit{S. pyogenes} or \textit{S. aureus} Cas9, which may prevent CjBE3/4 from gaining access to some sites. Furthermore, CjBE3 and CjBE4 have a large deamination window, as base editing frequently occurred outside of the canonical window (Figure 2.4b).
Figure 2.4. Base editing with CjBE3 and CjBE4. (a) CjBE3 and CjBE4 enables efficient base editing at multiple genomic sites containing the NNNVRYAC PAM. For *Campylobacter jejuni* base editors, BE4 architecture did not improve base editing compared to that of BE3. Compared to BE3 or SaBE3, CjBE3 and CjBE4 showed variable base-editing levels depending on the target site. (b) CjBE3 and CjBE4 have an enlarged deamination window on some sites compared to BE3 or BE4.

A thermostable Cas9 from *Geobacillus stearothermophilus* (GeoCas9) has the PAM specificity of NNNNCRAA\(^49\). We cloned in a nickase variant of GeoCas9 into BE3 and BE4 architectures to generate GeoBE3 and GeoBE4 respectively. We transfected plasmids encoding GeoBE3 or GeoBE4 with sgRNAs targeting sites containing the NNNNCRAA PAM in well-known genomic loci in HEK293T cells. Similar to CjBE3 and CjBE4, we observed a high variability in editing efficiency (~0%-35%) across multiple genomic loci (Figure 2.5).
suspect that GeoCas9, similar to CjCas9, has a weak helicase activity, which prevents its access to some genomic loci.

Figure 2.5. Base editing with GeoBE3 and GeoBE4. GeoBE3 and GeoBE4 enable base editing at sites containing the NNNNCAA PAM. Use of the BE4 architecture did not improve efficiency. Compared to BE3 or SaBE3, GeoBE3/4 exhibited variable activity levels depending on the target site.

CasX, which was discovered through a metagenomics search, is the first Cas9 discovered from Archaea, and it has a PAM specificity of TTCN at the 5'-end of the target site\textsuperscript{50}. We cloned a nickase form of CasX into the BE3 architecture to generate CasX-BE3. When CasX-BE3 was assayed in HEK293T cells with sgRNAs targeting the genomic loci with TTCN PAMs, no significant base editing was observed in all of the sites tested (data not shown). Since CasX has never been assayed in mammalian cells, we suspected that CasX may be unable to perform its function when ported from the Archaeal cellular environment to that of the mammalian cell. To ascertain whether CasX is functional in human cells, CasX nuclease activity was tested in a HEK293 GFP cleavage assay\textsuperscript{51}. We observed poor indel generation (0-4\%), compared to the amount generated by wild-type \textit{S. pyogenes} Cas9 (~50\%) (Figure 2.6). This strongly suggests
that CasX itself is not functional in mammalian cells, perhaps due to problems with protein folding, target search, or helicase activity.

Figure 2.6. CasX mammalian GFP-cleavage assay. CasX is not functional in mammalian cells. Plasmids containing CasX was co-transfected with a separate plasmid each encoding one of 12 different sgRNAs (CasX-1 to CasX-12) targeting EGFP into HEK293 cells stably expressing EGFP. Loss of green fluorescence was quantified using flow cytometry 5-days post-transfection.

Cas9 from a thermophilic bacterium, *Acidothermus cellulolyticus* (AceCas9), recognizes an NNNCC PAM. We sought to utilize AceCas9’s cytidine-rich PAM specificity for base editing. AceCas9 was cloned into the BE3 architecture to generate AceBE3. However, when assayed AceBE3 in HEK293T cells for base editing, we did not observe any C to T conversion across multiple genomic loci (data not shown). We suspect that CasX and AceCas9 are unable to function in mammalian cells due to improper protein folding or negligible target binding.

2.3 Utilizing Cas9 Variants with Altered PAM-Specificities for Base Editing

To further broaden the targeting range of base editors, we used recently engineered Cas9 variants with expanded or altered PAM specificities. Joung and co-workers recently reported SpCas9 mutants that accept NGA (VQR-Cas9), NGAG (EQR-Cas9), or NGCG (VRER-Cas9)
PAM sequences\textsuperscript{53}, as well as an engineered SaCas9 variant containing three mutations (SaKKH-Cas9) that relax the variant’s PAM requirement to NNNRRT\textsuperscript{54}. We replaced the SpCas9 portion of BE3 with these four Cas9 variants to produce VQR-BE3, EQR-BE3, VRER-BE3, and SaKKH-BE3, which should target NGAN, NGAG, NGCG, and NNNRRT PAMs, respectively. We transfected HEK293T cells with plasmids encoding these constructs and sgRNAs targeting several genomic loci for each new base editor, and measured C-to-T base conversions using HTS.

SaKKH-BE3 edited sites with NNNRRT PAMs with efficiencies up to 62% of treated, non-enriched cells (Figure 2.2b). As expected, SaBE3 was unable to efficiently edit targets containing PAMs that were NNHRRT (where H = A, C, or T) (Figure 2.2b). VQR-BE3, EQR-BE3, and VRER-BE3 exhibited more modest, but still substantial base-editing efficiencies of up to 50% of treated, non-enriched cells at genomic loci with the expected PAM requirements and an editing window similar to that of BE3 (Figure 2.7). Base-editing efficiencies of VQR-BE3, EQR-BE3, and VRER-BE3 closely paralleled the reported PAM requirements of the corresponding Cas9 nucleases; for example, EQR-BE3 was unable to efficiently edit targets containing NGAH PAM sequences (Figure 2.7). As expected, BE3 was unable to efficiently edit sites with NGA or NGCG PAMs (0–3 % efficiency) (Figure 2.3b). To confirm that the five new base editors functioned in multiple mammalian cell types, we assessed their performance in U2OS cells and observed robust editing, albeit with slightly lower editing and/or transfection efficiency than in HEK293T cells (Figure 2.8).
Figure 2.7. Base editing with VQR-BE3, EQR-BE3, and VRER-BE3. (a) VQR-BE3 and EQR-BE3 enable efficient base editing on several genomic sites targetable with the NGAG PAM. (b) VQR-BE3, but not EQR-BE3, efficiently base edits sites targetable with the NGAH PAM. (c) VRER-BE3 enables efficient base editing on several genomic sites targetable with the NGCG PAM.

Figure 2.8. Editing efficiencies of base editors with altered PAM-specificities in U2OS cells. SaBE3, SaKKH-BE3, VQR-BE3, EQR-BE3, and VRER-BE3 enable efficient base editing in the U2OS cell line.
2.4 Genomic Specificity of Base Editors with Altered PAM-Specificities

For each of the altered-PAM base editors, we selected two on-target loci that had been previously analyzed for Cas9 off-target cleavage and sequenced seven off-target loci for the SaBE3 constructs and ten off-target loci for the SpBE3 constructs\(^{53,54}\) (Figure 2.9). Consistent with our previous study\(^{42}\), we detected off-target base editing by SaBE3 and SaKKH-BE3 at a subset of known Cas9 off-target cleavage loci containing an appropriately placed target C (Figure 2.9). In contrast, we observed substantially less off-target base editing from VQR-BE3 or EQR-BE3 at ten known off-target loci of VQR SpCas9, suggesting that these base editors may offer enhanced specificity\(^{53}\) (Figure 2.9). Collectively, the properties of SaBE3, SaKKH-BE3, VQR-BE3, EQR-BE3, and VRER-BE3 establish that base editors behave in a modular fashion that facilitates our ability to repurpose Cas9 homologs and engineered variants for base editing.
Figure 2.9. Off-target base editing by SaBE3, SaKKH-BE3, VQR-BE3, and EQR-BE3. (a) Target and PAM sequences of genomic sites used to assess off-target base editing activity of SaBE3, SaKKH-BE3, VQR-BE3, and EQR-BE3. Four genomic loci which have been previously analyzed by GUIDE-Seq were chosen. Off-target sites containing at least one target C within the deamination window were analyzed using HTS. (b) SaBE3 and SaKKH-BE3 exhibited some off-target activity corresponding to the presence of matching PAMs. (c) VQR-BE3 and EQR-BE3 exhibited substantially reduced off-target base editing activity compared to the level expected from the GUIDE-Seq data.

2.5 Conclusion

The base editors behave in a modular fashion, tolerating incorporation of different Cas9 homologs and engineered variants. This expands the targeting scope of base editing through the use of SaCas9, CjCas9, and GeoCas9. However, not all homologs of Cas9 can generate functional base editors. In general, Cas9s whose helicase activities are thought to be poor, such
as CasX and AceCas9, also display poor base editing activity when ported into the BE3 architecture. This suggests that the helicase activity of Cas9 remains an essential functional part of the base editor.

Furthermore, mutations that alter or expand the PAM specificities of Cas9 can also be used to expand the targeting scope of base editing. VQR-BE3, EQR-BE3, and VRER-BE3, which contain mutations in the PAM interacting domain of respective Cas9s, allow for base editing at sites that were previously not targetable.

The base editors with altered PAM specificities described in this chapter together increased by ~3-fold the number of disease-associated mutations in the ClinVar database⁵ that can, in principle, be corrected by base editing (Figure 2.10).

![Percent of ClinVar pathogenic mutations targetable by C→T (or G→A) base editing (out of 3,408)](image)

**Figure 2.10.** Disease-associated mutations in ClinVar correctable with base editing. Using BE3, only 27% out of all possible mutations were targetable with C→T base editing. Using base editors with altered PAM specificities, up to 74% of all disease-associated C→T mutations are now targetable.
2.6 Methods

**Cloning.** PCR was performed using Q5 Hot Start High-Fidelity DNA Polymerase (New England BioLabs). Plasmids for BE and sgRNA were constructed using USER cloning (New England BioLabs) from previously reported plasmids\(^4^2\). DNA vector amplification was carried out using NEB 10beta competent cells (New England BioLabs). Site-directed mutagenesis of APOBEC1 variants was done using blunt-end ligation. Briefly, a primer with an overhang containing the desired point mutation was used to amplify the appropriate vector plasmid by PCR. KLD enzyme mix (New England BioLabs) was used to phosphorylate and circularize the PCR product before transformation.

**Cell culture.** HEK293T (ATCC CRL-3216) and U2OS (ATCC HTB-96) cells were cultured in Dulbecco’s Modified Eagle’s Medium plus GlutaMax (ThermoFisher) supplemented with 10% (v/v) FBS, at 37 °C with 5% CO\(_2\). Cell lines have been validated by vendors, and were tested using the MycoAlert PLUS Mycoplasma Detection Kit (Lonza).

**Transfections.** HEK293T cells seeded on 48-well collagen-coated BioCoat plates (Corning) were transfected at ~70% confluency. 750 ng of BE and 250 ng of sgRNA expression plasmids were transfected using 1.5 μl of LipoFectamine 2000 (ThermoFisher Scientific) per well according to the manufacturer’s protocol. 500 ng of BE and 250 ng of sgRNA expression plasmids were transfected into U2OS cells using a Lonza 4D-Nucleofector with the DN-100 program according to the manufacturer’s protocols.

**High-throughput DNA sequencing of genomic DNA samples.** Transfected cells were harvested after 3 d. The genomic DNA was isolated using the Agencourt DNAAdvance Genomic DNA Isolation Kit (Beckman Coulter) according to the manufacturer’s instructions. Genomic regions of interest were amplified by PCR with flanking HTS primer pairs. PCR amplification
was carried out with Phusion hot-start II DNA polymerase (ThermoFisher Scientific) according to the manufacturer’s instructions. PCR products were purified using RapidTips (Diffinity Genomics). Secondary PCR was performed to attach sequencing adaptors. The products were gel-purified and quantified using the KAPA Library Quantification Kit-Illumina (KAPA Biosystems). Samples were sequenced on an Illumina MiSeq as previously described.

**Analysis of pathogenic SNPs within the ClinVar database**

```python
In [ ]:
%matplotlib inline
import numpy as np
import scipy as sp
import matplotlib as mpl
import matplotlib.cm as cm
import matplotlib.pyplot as plt
import pandas as pd
pd.set_option('display.width', 500)
pd.set_option('display.max_columns', 100)
pd.set_option('display.notebook_repr_html', True)
import seaborn as sns
sns.set_style("whitegrid")

In [ ]:
import requests
import time
from bs4 import BeautifulSoup
import regex
import re
from Bio import SeqIO
import Bio

In [ ]:
ClinVar=pd.read_csv('clinvar.csv')
Phenotypes=pd.read_csv('DiseaseNames.csv')
PhenotypeDict=dict(zip(Phenotypes.CUI, Phenotypes.name))

In [ ]:
#count clinical variants to only SNPs
SNV=ClinVar[(ClinVar.Type=='single nucleotide variant')]
SNV.drop_duplicates('RS# (dbSNP)').shape

In [ ]:
#count those with capitalized Pathogenic in the significance field,
#this will not capture entries such as Likely pathogenic which have a lower case P
PathSNV=SNV['Pathogenic' in x for x in SNV.ClinicalSignificance]
PathSNV.drop_duplicates('RS# (dbSNP)').shape

In [ ]:
#open flanking sequence fasta files for all Y-type pathogenic human SNPs
#(includes both C>T and T>C ref>variant)
#downloaded as fasta file from:
#http://www.ncbi.nlm.nih.gov/snp/?term=((%22pathogenic%22%5BClinical+Significance%5D+AND+%22swith%22%5BSNP+Class%5D+AND+homo+sapiens%5BOrganism%5D)+AND(%22pathogenic%22%5BClinical+Significance%5D+AND+%22likely+pathogenic%22%5BSNP+Class%5D+AND+homo+sapiens%5BOrganism%5D)+AND(%22pathogenic%22%5BClinical+Significance%5D+AND+%22benign%22%5BSNP+Class%5D+AND+homo+sapiens%5BOrganism%5D)+AND(%22pathogenic%22%5BClinical+Significance%5D+AND+%22unknown%22%5BSNP+Class%5D+AND+homo+sapiens%5BOrganism%5D)+AND(%22pathogenic%22%5BClinical+Significance%5D+AND+%22not+pathogenic%22%5BSNP+Class%5D+AND+homo+sapiens%5BOrganism%5D)+AND(%22pathogenic%22%5BClinical+Significance%5D+AND+%22no+clinical+significance%22%5BSNP+Class%5D+AND+homo+sapiens%5BOrganism%5D)+AND(%22pathogenic%22%5BClinical+Significance%5D+AND+%22variant+of+uncertain+significance%22%5BSNP+Class%5D+AND+homo+sapiens%5BOrganism%5D)+AND(%22pathogenic%22%5BClinical+Significance%5D+AND+%22pathogenic%22%5BSNP+Class%5D+AND+homo+sapiens%5BOrganism%5D)+AND(%22pathogenic%22%5BClinical+Significance%5D)+AND(%22pathogenic%22%5BClinical+Significance%5D)
```
handle = open("YFasta.txt", "rU")
flanks = {}
# save as a dictionary keyed on rsID as an Integer with values being 25nt of flanking sequence on each side of the SNP
for record in SeqIO.parse(handle, "fasta"):
  flanks[int(record.id.split("|")[-1].strip('rs'))]=regex.findall('.{25}[^A,T,C,G].{25}', record.seq.tostring())
handle.close()
# merge flanking sequences to the CtoT frame on rsID
F=pd.DataFrame({'RS# (dbSNP)': flanks.keys(), 'Flanks': [x for x in flanks.values()]})
CtoT=F.merge(ClinVar, left_on='RS# (dbSNP)', right_on='RS# (dbSNP)', how='left')
# clinvar may refer to the opposite strand that was used in dbSNP; we want to allow clinvar reference alleles A and T with alternate alleles G and C respectively; we do not want to allow reference alleles G and C with alternate alleles A and T respectively; these Y-type SNPs must be removed
CtoT=CtoT[(CtoT.ReferenceAllele=='A') | (CtoT.ReferenceAllele=='T')].drop_duplicates('RS# (dbSNP)')

In [ ]:
# define window limits and the length of the pam including all N residues
windowstart=4
windowend=8
windowlen=windowend-windowstart+1
lenpam=3
# define a positional preference dictionary
priority=[5:0,6:1,7:2,8:3,4:3]
CtoT['gRNAs']=None
CtoT['gRNAall']=None
for i in range(len(CtoT)):
  if type(CtoT.ioc[i].Flanks)==list and CtoT.ioc[i].Flanks=[]:
    # define a potential gRNA spacer for each window positioning
    gRNAoptions=[test[(26-windowstart-j):(26-windowstart-j+lenpam+20)]
  for j in range(windowlen):
    # if there is an appropriate PAM placed for a given gRNA spacer
    # save tuple of gRNA spacer, the priority of the target C, and the position of off-target Cs in the window
    gRNA=[(gRNAoptions[k],priority[k+windowstart],[x.start()+1 for x in re.finditer('C',gRNAoptions[k]) if windowstart-l<x.start()+1<windowend+l])
  for k in range(len(gRNAoptions)) if regex.match('[A|C|T|G]GA', gRNAoptions[k][-lenpam:])]
  gRNAsingleC=[]
  for g,p,c in gRNA:
    # if the target C is the only C in the window save this as a single C site
    if g[windowstart-l:windowend].count('C')==0:
      gRNAsingleC.append(g)
    # OPTIONAL uncomment the ELIF statement if you are interest in filtered based upon position of off-target C
    # if the target C is expected to be edited more efficiently than the off-target Cs, also save as a single C Site
    #elif all([p<priority[x] for x in c]):
    #gRNAsingleC.append(g)
CtoT.gRNAs.iloc[i]=gRNAsingleC
CtoT.gRNAall.iloc[i]=[g for g,p,c in gRNA]

In [ ]:
#merge in phenotypes based upon MedGen IDs; remove redundant columns
CtoT=CtoT[['RS# (dbSNP)', 'GeneSymbol', 'Name', 'PhenotypeIDs', 'Origin', 'ReviewStatus', 'NumberSubmitters', 'LastEvaluated', 'gRNAs', 'gRNAall']]
ids=[re.findall('MedGen:.{7}', x) for x in CtoT.PhenotypeIDs.values]
CtoT['Phenotypes']=[[PhenotypeDict[y.istrip('MedGen:')] for y in x if y.istrip('MedGen:') in PhenotypeDict.keys()] for x in ids]
CtoT.drop('PhenotypeIDs', inplace=True, axis=1)

In [ ]:
#open flanking sequence fasta files for all R-type pathogenic human SNPs
#downloaded as fasta file from:
#http://www.ncbi.nlm.nih.gov/snp/?term=((%22pathogenic%22%5BClinical+Significance%22%5D+AND+%22snv%22%5Brsid%5D%5D+AND+type%5Bpathogenic%5D)+%22r%22%5BAllele%5D)
handle = open("RFasta.txt", "rU")
flanks={}
#save as a dictionary keyed on rsID as an Integer with values being 25nt of
#flanking sequence on each side of the SNP
for record in SeqIO.parse(handle, "fasta") :
    flanks[int(record.id.split('|')[-1].strip('rs'))]=regex.findall('.{25}[^A,T,C,G].{25}', record.seq.tostring())
handle.close()

#merge flanking sequences to the CtoT frame on rsID
F=pd.DataFrame({'RS# (dbSNP)': flanks.keys(), 'Flanks': [x for x in flanks.values()]})
GtoA=F.merge(ClinVar, left_on='RS# (dbSNP)', right_on='RS# (dbSNP)',
how='left')
#clinvar may refer to the opposite strand that was used in dbSNP;
#we want to allow clinvar reference alleles A and T with alternate alleles G
#and C respectively
#we do not want to allow reference alleles G and C with alternate alleles A
#and T respectively; these Y-type SNPs must be removed
GtoA=GtoA[(GtoA.ReferenceAllele=='A') | (GtoA.ReferenceAllele=='T')].drop_duplicates('RS# (dbSNP)')

In [ ]:
windowstart=4
windowend=8
windowlen=windowend-windowstart+1
lenpam=3
GtoA['gRNAs']=None
GtoA['gRNAall']=None
priority=[5:0,6:1,7:2,8:3,4:3]
for i in range(len(GtoA)):
    if type(GtoA.iloc[i].Flanks)==list and GtoA.iloc[i].Flanks=[]:
        test=GtoA.iloc[i].Flanks[0]
        gRNAoptions=[test[(25+windowstart)+j-20-lenpam:(25+windowstart)+j] for j in range(windowlen)]
        gRNA=[[gRNAoptions[k],priority[k+windowstart],[20+lenpam-x.start()] for x in re.finditer('G',gRNAoptions[k]) if windowstart-l<x.start()<(windowstart+x.start()+lenpam)] for k in range(len(gRNAoptions)) if regex.match('TC[A|C|T|G]', gRNAoptions[k][:lenpam])]
        gRNAsingleC=[]
        for g,p,c in gRNA:

41
if g[20+\text{lenpam}-\text{windowstart}-\text{windowlen}+1:20+\text{lenpam}-\text{windowstart}+1].\text{count}('G') == 0:
    \text{gRNAsingleC}.\text{append}(g)
# elif all([p<\text{priority}[x] for x in c]):
    \text{gRNAsingleC}.\text{append}(g)
\text{GtoA.gRNAs.i}[\text{loc}[i]]=\text{gRNAsingleC}
\text{GtoA.gRNAall.i}[\text{loc}[i]]=g \text{ for } g, p, c \text{ in } \text{gRNA}\
\text{In [ ]:}
\text{GtoA=GtoA}[['\text{RS} (\text{dbSNP})', '\text{GeneSymbol}', '\text{Name}', '\text{PhenotypeIDs}', '\text{Origin}', '\text{ReviewStatus}', '\text{NumberSubmitters}', '\text{LastEvaluated}', '\text{gRNAs}', '\text{gRNAall}']]\text{id}=\text{[re.\text{findall}('\text{MedGen}:C.\{7\}', x) for x in GtoA.\text{PhenotypeIDs}.\text{values}]}\text{GtoA['\text{Phenotypes}']=[[\text{PhenotypeDict[y.lstrip('\text{MedGen:}')]} for y in x if y.lstrip('\text{MedGen:}') in \text{PhenotypeDict.keys()}] for x in GtoA]}\text{GtoA.\text{drop}('\text{PhenotypeIDs}', \text{inplace=True}, \text{axis=1})}\
\text{In [ ]:}
\text{#number of duplicate entries due to +/- strand inconsistency between dbSNP and Clinvar}
\text{sum([i in CtoT['\text{RS} (\text{dbSNP})'] for i in GtoA['\text{RS} (\text{dbSNP})']])}\
\text{In [ ]:}
\text{CtoT.\text{to_csv('pathogenic_CtoT_all.csv')}}
\text{GtoA.\text{to_csv('pathogenic_GtoA_all.csv')}}\
\text{In [ ]:}
\text{pathogenic_CtoT_ha}[\text{PAM=CtoT}[[\text{type}(x)==\text{list} \text{ and x}!] for x in CtoT.gRNAall]]
\text{pathogenic_GtoA_ha}[\text{PAM=GtoA}[[\text{type}(x)==\text{list} \text{ and x}!] for x in GtoA.gRNAall]]
\text{In [ ]:}
\text{len(pathogenic_CtoT_ha[\text{PAM})}}
\text{In [ ]:}
\text{len(pathogenic_GtoA_ha[\text{PAM})}}
\text{In [ ]:}
\text{pathogenic_GtoA_ha[\text{PAM}.\text{to_csv('pathogenic_GtoA_ha[\text{PAM}.\text{csv')}}
\text{pathogenic_CtoT_ha[\text{PAM}.\text{to_csv('pathogenic_CtoT_ha[\text{PAM}.\text{csv')}}
\text{In [ ]:}
\text{pathogenic_CtoT_Sing}[eC=CtoT}[[\text{type}(x)==\text{list} \text{ and x}!] for x in CtoT.gRNAs]]
\text{In [ ]:}
\text{pathogenic_GtoA_Sing}[eC=GtoA}[[\text{type}(x)==\text{list} \text{ and x}!] for x in GtoA.gRNAs]]
\text{In [ ]:}
\text{len(pathogenic_CtoT_SingleC)}
\text{In [ ]:}
\text{len(pathogenic_GtoA_SingleC)}
\text{In [ ]:}
\text{pathogenic_GtoA_SingleC.\text{to_csv('pathogenic_GtoA_SingleC.csv')}}
\text{pathogenic_CtoT_SingleC.\text{to_csv('pathogenic_CtoT_SingleC.csv')}}
\text{In [ ]:}
\text{#duplicates in the two pathogenic tables}
\text{sum([i in pathogenic_GtoA_SingleC['\text{RS} (\text{dbSNP})'] for i in pathogenic_CtoT_SingleC['\text{RS} (\text{dbSNP})']])}\
\text{In [ ]:}
\text{with open("Summary.txt", "w") as text_file:}
    \text{text_file.write("singleC %s \n" % (len(pathogenic_CtoT_SingleC)+len(pathogenic_GtoA_SingleC)))}
    \text{text_file.write("hasPAM %s" % (len(pathogenic_CtoT_ha[\text{PAM})]+len(pathogenic_GtoA_ha[\text{PAM}))}
Chapter Three:
Modulating the Activity Window of Base Editing

Y. Bill Kim, Alexis C. Komor, Jonathan M. Levy, Michael S. Packer, Kevin T. Zhao, and David R. Liu

Bill Kim generated most constructs and conducted all experiments and analyzed data. Alexis Komor and Jonathan Levy generated some constructs. Michael Packer assisted with computational analyses. Kevin Zhao conducted some experiments.

3.1 Introduction: Selectivity in Base Editing

Base editors typically modify all Cs within the 5-nucleotide activity window to Ts with comparable efficiency\(^4\). The ability to modulate the width of this window is useful when it is important to edit only a subset of Cs present in the activity window, such as cases in which the target C is adjacent to other Cs that, if changed, would cause undesired coding mutations. This chapter describes new C:G to T:A base editors with narrowed editing window widths.

3.2 Effect of Linker Length on Base Editing

We previously noted that the length of the linker between APOBEC1 and dCas9 modulates the number of bases that are accessible by APOBEC1 \textit{in vitro}\(^4\). In HEK293T cells, however, varying the linker length did not significantly modulate the width of the editing window, suggesting that in the cellular milieu, the relative disposition of dCas9 and the cytidine deaminase are not strongly determined by linker length (Figure 3.1).

3.3 Effect of Guide RNA Length on Base Editing

We hypothesized that truncating the 5’ end of the sgRNA might narrow the base-editing window by reducing the length of single-stranded DNA accessible to the deaminase. Although for some target loci, truncated guide RNAs with 16- or 17-base spacers showed narrowed editing window widths, we observed no consistent changes in the editing window width when using truncated sgRNAs with 15- to 19-base spacers (Figure 3.2).
**Figure 3.1. Effect of APOBEC1–Cas9 linker length on editing window width.** BE3 activity in HEK293T cells on two genomic sites with multiple Cs in the protospacer. Varying the linker length does not alter the position at which base editing is favored.

**Figure 3.2. Effect of 5’-guide RNA truncations on editing window width.** BE3 activity in HEK293T cells on three genomic sites with multiple Cs in the protospacer. Truncating the guide RNA at the 5’ end usually does not alter the position at which base editing is favored.
3.4 Attenuation of Deaminase Kinetics Influences Base Editing Window Width

We envisioned that mutations to the deaminase domain might narrow the width of the editing window. Because the high activity of APOBEC1 likely contributes to the deamination of multiple Cs per DNA binding event\textsuperscript{42,55,56}, mutations that reduce the catalytic efficiency of the deaminase domain of a base editor might prevent it from catalyzing successive rounds of deamination before dissociating from DNA. Once any C:G to T:A editing event has taken place, the sgRNA no longer perfectly matches the target DNA sequence and re-binding of the base editor to the target locus should be less favorable. Alternatively, mutations to the deaminase domain might alter substrate binding, the conformation of bound DNA, or substrate accessibility to the active site in ways that reduce tolerance for non-optimal presentation of a C to the deaminase active site. We sought to test both possibilities to discover new base editors that distinguish among multiple cytidines within the original editing window.

Given the absence of an available APOBEC1 structure, we identified mutations previously reported to modulate the catalytic activity of APOBEC3G, a cytidine deaminase that shares 42\% sequence similarity\textsuperscript{26} of its active site–containing domain to that of APOBEC1. We incorporated the corresponding APOBEC1 mutations into BE3 and evaluated their effect on base-editing efficiency and editing window width in HEK293T cells at two C-rich genomic sites containing Cs at positions 3, 4, 5, 6, 8, 9, 10, 12, 13, and 14 (site A); or containing Cs at positions 5, 6, 7, 8, 9, 10, 11, and 13 (site B). For analysis purposes, we define the “editing window width” as the number of nucleotide positions at a given site for which editing efficiency exceeds the half-maximal value for that target site. The editing window width of BE3 for the two C-rich genomic sites tested was 4 (site A) and 6 (site B) nucleotides (Figure 3.3).
Mutating residues essential for deaminase activity, such as R118A, led to dramatic losses of base-editing efficiency (Figure 3.3). Other mutations, however, narrowed the editing window while maintaining substantial editing efficiency (Figure 3.3). W90 in APOBEC1, corresponding to W285 in APOBEC3G, is predicted to be crucial\textsuperscript{55,56} for the formation of a hydrophobic active site in APOBEC1. APOBEC3G mutant W285A has undetectable deaminase activity, and BE3 W90A similarly shows almost no base-editing activity (Figure 3.3). We hypothesized that W90Y or W90F might decrease the hydrophobicity of the active site while maintaining catalytic activity. Indeed, W90Y and W90F only modestly decreased base-editing activity while narrowing the editing window width at sites A and B to 3 or 2 nucleotides, respectively (Figure 3.3). These results demonstrate that mutations of the cytidine deaminase domain can narrow the width of the base-editing window.

Next, we tested whether mutations to residues involved in APOBEC1 substrate binding could also alter the editing window width. R126 in APOBEC1 is predicted to interact with the phosphate backbone of single-stranded DNA\textsuperscript{56} and mutation of the corresponding residue to Ala in APOBEC3G (R320A) decreases apparent activity by at least 5-fold\textsuperscript{56}. Notably, when introduced into APOBEC1 in BE3, R126A, and R126E maintained activity comparable to that of BE3 at the most strongly edited central positions (C5 and C6 for site A, C6 and C7 for site B), while decreasing editing activity at other positions (Figure 3.3). Both of these two mutations narrowed the width of the editing window at site A and site B to 3 nucleotides (Figure 3.3). R132 is near R126 and could also influence single-stranded DNA accessibility. R132E led to a decrease in editing efficiency but also narrowed the width of the editing window to 3 nucleotides for both sites A and B (Figure 3.3).
Figure 3.3. Base editors with mutant cytidine deaminase domains exhibit narrowed editing windows. BE3 activity in HEK293T cells on two genomic sites with multiple Cs in the protospacer. Modulating the catalytic activity of the deaminase domain through mutagenesis favors the deamination of substrates positioned near the center of the base editing window (C5 and C6 for site A and C6 for site B). Some mutations, such as R118A and W90A led to dramatic reduction in base editing efficiency. R126A/E, W90Y, and R132E mutations narrow the editing window yet maintain peak efficiency. Combinations of these mutations led to variants that have peak wild-type efficiency but significantly reduced window widths (W90Y+R126E for sites A and B).
3.5 Development of Base Editors with Narrowed Activity Windows

We combined W90Y/F, R126E, and R132E, the mutations that narrowed the editing window without drastically reducing base-editing activity, into doubly and triply mutated base editors. Combining the R126E and W90F mutations modestly narrowed the editing window compared to the W90F mutant, while combining W90F with R132E resulted in a dramatic loss of activity (Figure 3.3). The double mutant W90Y+R126E, however, resulted in a base editor (YE1-BE3) with BE3-like maximal editing efficiencies, but substantially narrowed editing window width of ~2 nucleotides for both site A and site B (Figure 3.3). The W90Y+R132E base editor (YE2-BE3) exhibited modestly lower editing efficiencies, but also narrowed the editing window width to ~2 nucleotides for both site A and site B (Figure 3.3). The R126E+R132E double mutant (EE-BE3) showed similar maximal editing efficiencies and editing window width as YE2-BE3 (Figure 3.3). The triple mutant W90Y+R126E+R132E (YEE-BE3) exhibited 2.9-fold lower average maximal editing yields but very little editing beyond the C6 position and an editing window width of ~2 nucleotides and 1 nucleotide for site A and site B, respectively (Figure 3.3). These data taken together indicate that mutations in the cytidine deaminase domain can strongly affect editing window widths, in some cases with minimal or only modest effects on editing efficiency.

Next, we compared the base-editing outcomes of BE3, YE1-BE3, YE2-BE3, EE-BE3, and YEE-BE3 in HEK293T cells targeting four well-studied human genomic sites that contain multiple Cs within the BE3 activity window. These target loci contained target Cs at positions 4 and 5 (HEK293 site 3), positions 4 and 6 (HEK293 site 2), positions 5 and 6 (EMX1), or positions 6, 7, 8, and 11 (FANCF). BE3 exhibited little preference for editing any of the Cs within the activity window of positions 4–8 (<1.2-fold average base preference for HEK293 sites
2, 3, and EMX1). In contrast, the narrow-window base editors exhibited increasing preference for editing one cytidine over others. For example, YEE-BE3 showed 31-fold preference for editing C6 over C4 at HEK293 site 2 (Figure 3.4). We also assessed editing of YE1-BE3, EE-BE3, YE2-BE3, and YEE-BE3 in U2OS cells, and observed similar alteration of base-editing window widths (Figure 3.5) as we observed in HEK293T cells. These findings establish that mutant base editors with narrowed editing windows can discriminate between adjacent Cs, even when both nucleotides are within the BE3 editing window.

Figure 3.4. Activity of base editors with APOBEC1 mutations on well-studied genomic sites. BE3 with APOBEC1 mutations bias the base editing outcome to prefer deamination of centrally placed cytidines over those that are near the edge of the window. Combining mutations led to YE1-BE3, EE-BE3, YE2-BE3, and YEE-BE3 that exhibit a progressively stringent preference for single-nucleotide editing within the protospacer.
Figure 3.5. Activity of base editors with narrowed deamination windows in U2OS cells. YE1-BE3, YE2-BE3, EE-BE3, or YEE-BE3 can be used to preferentially target a single nucleotide in U2OS cells.

3.6. APOBEC1 Mutations Cause a Shift in Base Editing Product Distributions.

We also analyzed the product distributions generated by each of the above base editors to assess their apparent processivity. BE3 generated predominantly T4-T5 (HEK site 3), T4-T6 (HEK site 2), and T5-T6 (EMX1) products in treated HEK293T cells, resulting in, on average, 7.4-fold more products containing two Ts, than products containing a single T (Figure 3.6). In contrast, YE1-BE3, YE2-BE3, EE-BE3, and YEE-BE3 showed substantially higher preferences for singly edited C4-T5, C4-T6, and T5-C6 products at these sites (Figure 3.6). The YEE-BE3 triple mutant strongly favored single-T products by an average of 14-fold across the three genomic loci (Figure 3.6). Interestingly, for the target site in which only one C is within the target window (HEK293 site 4, at position C5), all four mutants exhibited editing efficiencies comparable to those of BE3 (Figure 3.7). These findings together suggest that the decreased apparent processivity of these narrow-window base editors favors conversion of only a single C at target sites containing multiple Cs within the BE3 editing window. These data also suggest a
positional preference of C5 > C6 > C7 ≈ C4 for these mutant base editors, although we note that this preference could differ depending on the target sequence.

Figure 3.6. Product distributions generated by base editors with narrowed activity windows. Base editors with narrowed activity windows shift the product distribution of the experiment to predominantly generate singly modified species across multiple genomic sites in HEK293T cells.

Figure 3.7. Effect of APOBEC1 mutations on the editing efficiency of a target containing a single substrate C. HEK293T cells were transfected with plasmids expressing BE3 or its mutants and an appropriate sgRNA. Target Cs are shown in red, with a subscripted number denoting spacer position, and the PAM sequence is in blue. The BE3 point mutants do not substantially affect base editing efficiencies at HEK293 site 4, a site with only one target C. Values and error bars reflect the mean and standard deviation of at least two biological replicates.
3.7 The Narrowed Editing Window Profiles Are Stable Over Multiple Generations

To test if narrowed-window base editors maintain their editing window profiles following a single treatment, we treated HEK293T cells with YE1-BE3, EE-BE3, YE2-BE3, or YEE-BE3 and analyzed by HTS the distribution of base editing within the protospacer sequence 3 d, 6 d, or 9 d after treatment. We observed no substantial change in the positional distribution of base editing over 9 d at any of the four genomic loci tested (Figure 3.8).

3.8 Quantification of APOBEC1 Catalytic Impairment

We sought to quantify the degree of catalytic impairment conferred by individual mutations. We examined the mutations in the context of a rifampicin resistance assay. Briefly, deaminases can confer rifampicin resistance through mutagenesis of the *E. coli* RNA polymerase II gene. Mutations that impair APOBEC1 catalysis would generate fewer colonies resistant to rifampicin. We transformed plasmids encoding APOBEC1 variants under the arabinose promoter into BL21 cells, and induced deaminase expression overnight. *E. coli* containing WT APOBEC1 and APOBEC1 with mutations that are thought to have only a slight impact on catalysis (R126A and R126E) showed retarded growth, suggesting that, at this level of deaminase expression in BL21 cells, the deaminase activity was too toxic to allow for robust cell growth (Figure 3.9a). Cells that were successfully enriched via overnight growth were plated on rifampicin-containing agar plates for colony quantification. Cells containing single mutations in APOBEC1 exhibited ~10-fold more colonies resistant to rifampicin than those with two mutations, which in turn had ~100-fold more resistant colonies than those with three mutations (Figure 3.9b). These results suggest that the mutations that narrow the deamination window indeed suppress the deamination kinetics of the enzyme.
Figure 3.8. Effect of post-treatment time on base editing activity window width. Base editing outcome is stable after transient transfection of plasmids containing base editors and sgRNA over multiple passages.
Figure 3.9. Quantification of catalytic impairment conferred by APOBEC1 mutations.
Plasmids encoding unfused APOBEC1 under pBAD promoter is transformed into BL21 cells and grown before quantification of rifampicin resistance. (a) Bacterial growth was hampered when WT, R126A, or R126E APOBEC1 were present, which prevented them from use in the subsequent rifampicin-resistance assay. (b) Cells with plasmids encoding APOBEC1 with single mutations (R132E and W90Y) resulted in 10-fold more rifampicin-resistant colonies than those with double mutations. In turn, the cells containing the plasmid encoding the triple mutant generated ~100-fold fewer colonies than those with double mutations.

3.9 Base Editors with Narrowed Activity Windows Are More Specific

We examined the off-target activity of the base editors with narrowed activity windows. We assayed by HTS 21 known off-target loci corresponding to three on-target loci. We observed on average 3.6-fold less efficient off-target base editing per on-target editing event of the BE3 variants with narrowed activity windows when compared to BE3 across the 21 off-target sites (Figure 3.10). This improvement in DNA specificity may arise from a reduction in the residence time of base editors at off-target loci that further impedes deamination with catalytically impaired APOBEC1 domains.
Figure 3.10. Off-target activity of base editors with altered activity window widths. Three genomic loci and their off-target sequences which have previously been analyzed by GUIDE-Seq were assayed with BE3, YE1-BE3, EE-BE3, YE2-BE3, and YEE-BE3. Base editors with narrowed activity windows exhibit lower off-target base editing compared to BE3.

3.10 Base Editors with Both Altered PAM-Sequences and Narrowed Windows

We combined the window-modulating mutations in APOBEC1 with the Cas9 variants with different PAM specificities. For example, VQR-YE1-BE3 allowed editing with a narrowed activity window and greater positional selectivity of target sites containing an NGA PAM (Figure 3.11a). Not all window-narrowing mutations could be productively combined with altered PAM base editors, however. When window-narrowing mutations were installed in SaKKH-BE3, for example, we observed a decrease in base-editing efficiency without any obvious change in the width of the activity window, perhaps arising from differences in the
substrate accessibility of this base editor compared with that of BE3 and its variants (Figure 3.11b).

Figure 3.11. Effect of window-narrowing mutations on VQR-BE3 and SaKKH-BE3. (a) Mutations in the APOBEC1 domain are compatible with VQR-BE3, allowing the precise editing of sites targetable by NGA PAM. (b) However, incorporating APOBEC1 mutations into SaKKH-BE3 did not result in the narrowing of the deamination window without the loss of on-target activity, suggesting that the conformational constraints imposed by S. aureus Cas9 are different from that of S. pyogenes Cas9.
3.11 Conclusion

The deamination window of the base editor can be tuned. Changing the linker length and 5’ truncation of gRNA did not affect the position at which base editing is preferred. This highlights the difficulty of thermodynamically confining the productive conformations of the base editor towards one but not the other substrates in the vicinity by modulating the composition of the base editor. However, as base editors have limited residence time on target DNA before DNA replication or repair takes place, subtly slowing down the kinetics of APOBEC1 through the introduction of deleterious mutations prevents processive catalysis of multiple Cs within the target window. Such base editors have a greater chance of dissociating from the target site after only a single deamination event, resulting in an apparent narrowing of the base-editing window.

While the APOBEC1 mutations had a distinct window-narrowing effect on BE3 and VQR-BE3, both of which use SpCas9 as the binding domain, the effect was less pronounced for base editors that use SaCas9. This highlights the intricate relationship between the protein composition and the emergent property of the narrowed activity window arising from the kinetic impairment of the deaminase. The APOBEC1 mutations described in this thesis are tailored to the base editors composed of SpCas9 and the XTEN linker that connects the deaminase and Cas9. Changing either the origin of Cas9 or the length and composition of the linker may perturb the fine balance that is required for the apparent narrowing of the activity window without sacrificing the on-target base-editing efficiency at the target base. As such, the use of APOBEC1 mutations in CjBE3, GeoBE3, or BE4 that uses an extended linker may not result in narrowed activity windows.
The development of base editors with narrowed editing windows approximately doubled the fraction of ClinVar entries with a properly positioned NGG PAM that can be corrected by base editing without comparable modification of a non-target C (Figure 3.12).

**Figure 3.12. Genetic variants in ClinVar that can be corrected in principle by base editors with narrowed activity windows.** Out of 911 disease-associated mutations, BE3 may correct 31% of them without introducing additional mutations, due to multiple occurrences of cytidines within the 5-base window. Base editors with a narrowed activity window corrects 59% of the mutations without introducing unwanted additional mutations, assuming C5>C6>C7>C4~C8 selectivity.
3.12 Methods

See Section 2.9 for additional experimental methods.

**Rifampicin resistance assay.** Plasmids encoding non-fused APOBEC1 containing various mutations under arabinose induction were transformed into BL21 cells, and grew overnight. The cells were subcultured into Davis Rich Media (DRM) and grown overnight with arabinose (200 nM). Cells obtained from the overnight growth were quantified by serial dilution into non-rifampicin-containing LB-agar plates. The same cells were then serially plated into rifampicin containing LB-agar plates to quantify the number of colonies resistant to rifampicin. The ratio of the number of resistant colonies to the total number of colonies was used to determine the relative activity levels of the deaminases.

**Data Analysis.** Analyses of base-editing processivity were performed using a custom python script (see below). This program trims sequencing reads to the 20 nucleotide protospacer sequence as determined by a perfect match for the 7 nucleotide sequences that should flank the target site. These targets were then consolidated and sorted by abundance to assess the frequency of base-editing products. Bioinformatic analysis of the ClinVar database of human disease–associated mutations was performed in a manner similar to that previously described but with small adjustments. These adjustments enable the identification of targets with PAMs of customizable length and sequence. In addition, this improved script includes a priority ranking of target C positions (C5 > C6 > C7 > C8 ≈ C4), thus enabling the identification of target sites in which the on-target C is either the only cytosine within the window, or is placed at a position with higher predicted editing efficiency than any non-target C within the editing window.

**Base editing processivity script**

```python
%matplotlib inline
import numpy as np
import scipy as sp
```
import matplotlib as mpl
import matplotlib.cm as cm
import matplotlib.pyplot as plt
import pandas as pd
pd.set_option('display.width', 500)
pd.set_option('display.max_columns', 100)
pd.set_option('display.notebook_repr_html', True)
import seaborn as sns
sns.set_style("whitegrid")
sns.set_context("poster")
import requests
import time
from bs4 import BeautifulSoup
import regex
import re
import os
from Bio import SeqIO
import Bio

#BE processivity analysis, APOE
site='_APOE'
indir='/Users/michaelpacker/Desktop/Liu_Lab/MiSeqData/BK092916/fastq/APOE/
outdir='/Users/michaelpacker/Desktop/Liu_Lab/MiSeqData/BK092916/fastq/'
filenames=os.listdir(indir)
for i in range(len(filenames)):
    seqs={}
    for record in SeqIO.parse(indir+filenames[i], "fastq") :
        #split prior to spacer window
        split1=record.seq.tostring().split('ACCTGCA')
        if len(split1)==2:
            #take second item in first split
            #split again at the sequence right after the protospacer and take
            first item
            split2=split1[1].split('AGGCCGG')[0]
            #keep only 20 basepair long protospacers
            if len(split2)==20:
                seqs[record.id]=split2
        #store in a dataframe, group by the sequence; count occurrences and sort
        by frequency
    Counts=pd.DataFrame(seqs.items(),
columns=['ID','Window']).groupby('Window').count().sort('ID',
ascending=False)
    Counts.to_csv(outdir+filenames[i].strip('.fastq')+site+'.csv')
Chapter Four:
Improving the Safety of Base Editing

Y. Bill Kim, Kevin T. Zhao, Ahmed H. Badran, and David R. Liu

Bill Kim and Kevin Zhao generated constructs, conducted the experiments analyzed data. Ahmed Badran generated some constructs.

4.1 Introduction

Before base editing can be developed into a therapeutic, potential safety hazards need to be identified and addressed. For example, off-target activity or incorrect editing of the target sequence may cause unintended harm. Moreover, cellular and physiological effect on long term exposure to base editors is unknown (ie. during viral delivery). It is therefore desired to engineer ways to improve the safety of base editors. This chapter describes two improvements; increasing the product purity of base editing by reducing the incidences of indels, and engineering variants of base editors that are induced by a small molecule to enable control over when and where base editing takes place.

4.2 Base Editors with Reduced Indel Rates

A typical base editing experiment generates ∼1-5 % indel in cell culture. We suspected that the indels arise due to insufficient inhibition of cellular UDG. UDG generates an abasic site upon uracil excision, and the phosphodiester backbone is subsequently cleaved by AP-lyase to generate a nick on the target strand of base editing. Combined with a non-target strand nick generated by nickase Cas9, a DSB is created, which may lead to NHEJ-mediated indel formation.

We envisioned that localizing a DSB-binding protein, Mu-Gam, from bacteriophage Mu, may prevent NHEJ-mediated indel formation at the site of base editing. Mu-Gam competes with the Ku70/80 complex for DSB binding, a first step in the NHEJ pathway. Preventing the resolution of a DSB through Mu-Gam may be fatal, and those cells would be eliminated from the population.
We fused GAM to the N-terminus of BE3 and BE4 and transfected the plasmids encoding each with appropriate guide RNAs targeting well-studied genomic loci into HEK293T cells (Figure 4.1a). While the base editing efficiencies were similar to that of BE3 and BE4 (Figure 4.1b), BE3-Gam and BE4-Gam led to ~2-fold decrease in the amount of indel across multiple loci compared to their unfused counterparts (Figure 4.1c).

**Figure 4.1. Fusion of Mu-Gam reduces the amount of indel generated during base editing.** (a) Protein architectures of BE3-Gam and BE4-Gam. (b) The addition of Gam does not negatively affect base editing efficiencies. (c) The addition of Gam reduces the amount of indel generated, which results in an increase in editing-efficiency-to-indel-rate ratio.

4.3 Small Molecule-Inducible Base Editors

Inducible base editor allows for a greater control of its activity. This can be used to limit base editing to a localized area, or to a specific cellular or developmental timeframe. We sought to engineer base editors that are activated in the presence of a small molecule. Recently, a 4-hydroxytamoxifen (4-HT) inducible intein\(^1\) has been inserted into various positions in Cas9,
which enabled the control its nuclease activity using 4-HT\textsuperscript{62}. Since the intein disrupts the activity of Cas9 until it is spliced out, such a modality may also be applied to base editors.

We inserted 4-HT-inducible intein into S219 and C548 of Cas9 in BE3 to generate Intein-BE3 (IBE3) S219 and IBE3 C548, respectively. We transfected HEK293T cells with plasmids encoding the IBE3 variants and guide RNAs targeting HEK293 Site 3 and HEK293 Site 4, with or without 10 μM 4-HT for either 16 hr or 48 hr post transfection. In the presence of 4-HT, IBE3 and wild-type BE3 both resulted in high levels of base editing. In the absence of 4-HT, however, base editing was significantly reduced for IBE3 (Figure 4.2). Moreover, IBE3 exhibited greater base-editing-to-indel ratio than that of BE3 (Figure 4.2).

**Figure 4.2. Small molecule-inducible base editing.** For two genomic sites in HEK293T cells, IBE3 displays an increased base editing activity in the presence of 4-HT. IBE3 displays lower levels of indel formation, probably due to the reduced amount of base editor exposure to the cell. IBE3 GR denotes an IBE3 S219 variant with an additional G521R mutation in the intein that promotes its splicing in the presence of 4-HT, but not β-estradiol, which is natively present in the cell.
4.4 Conclusion

This chapter describes two innovations that enable cleaner base-editing outcome and improved control over base-editing activity in human cells. Fusion of Mu-Gam to BE3 and BE4 resulted in base editors that further suppress the formation of indels. On the other hand, a small-molecule inducible base editor has been developed by inserting a ligand-dependent intein into the base editor, allowing the researcher to precisely dictate when and where base editing takes place. These innovations improve the safety and versatility of base editing.

4.5 Methods

See Chapters 2 and 3 for Methods.
Chapter Five:
Efficient Cytosolic Delivery of Macromolecules via Cationic Liposomes

Y. Bill Kim, Kevin T. Zhao, David B. Thompson, and David R. Liu

Bill Kim generated constructs, conducted experiments, and analyzed data. Kevin Zhao helped out. David Thompson assisted with proteome analysis.
5.1 Introduction

Nucleic acids have an abundance of anionic charge, which enables efficient encapsulation by cationic lipid transfection agents. Proteins can be made net anionic by fusing moieties rich in aspartic acid and glutamic acid residues. It was recently discovered that, by mimicking the innately anionic nucleic acids, proteins can also be delivered into cells using cationic lipids (Figure 5.1)\textsuperscript{23}. (-30)GFP, which is an engineered sfGFP whose surface residues have been extensively mutated to contain net -30 charge, has been used to deliver genome-editing proteins using commercially available cationic lipids\textsuperscript{23}.

While (-30)GFP proved to be an efficient tag for mediating protein delivery, we reasoned that a better tag suitable for cationic lipid complexation is yet to be found. For example, (-30)GFP was incapable of incorporating any more acidic residues on the surface; addition of more charged residues led to stability losses\textsuperscript{23}.

In the human proteome, there exists many proteins with high theoretical net charge that could in theory replace (-30)GFP as a preferred tag. Such human proteins would have an additional benefit of minimizing potential immune responses when used \textit{in vivo}. In contrast, the use of GFP has been shown to cause mammalian immunogenicity that confound \textit{in vitro} experiments and hamper \textit{in vivo} usage\textsuperscript{24}. This chapter describes a screen of several negatively charged human proteins, which led to the discovery of a potent delivery agent that outperforms (-30)GFP.
Figure 5.1. Cationic lipid-mediated delivery of proteins using anionic protein tags. (a) Proteins that are neutral or positively charged can be rendered net anionic through the fusion of a supercharged tag, such as (-30)GFP. (b) Anionic proteins can be efficiently complexed into liposomes using commercially available cationic lipids via charge interaction. Liposomes are taken up through endocytosis, and the cargo is eventually released into the cytosol. Figure adapted from Zuris J.A. et al. Nat. Biotechnol. 33(1):73-80 (2015).

5.2 Protein Surface Charge on Cationic Lipid Delivery

Highly anionic proteins are encapsulated more efficiently by cationic lipids due to stronger charge interaction compared to proteins that are neutral or positively charged\textsuperscript{23}. To independently confirm this hypothesis, we expressed sfGFP variants whose surface residues are mutated into glutamic acids or aspartic acids to contain the net charge of -17, -20, -27, -30, and -40, to generate (-17)GFP, (-20)GFP, (-30)GFP, and (-40)GFP respectively. Each variant, in
addition to sfGFP which has a net -7 charge, was fused to the N-terminus of Cre and delivered into HeLa-DsRed cells using Lipofectamine 2000. Cre recombination was quantified after 3 d via flow cytometry. In general, higher surface charge resulted in greater Cre recombination (Figure 5.2). (-40)GFP was markedly less fluorescent than other charge variants, which may be a symptom of instability or misfolding, and explain its lower activity despite high negative charge (Figure 5.2). These results show that higher negative charge may facilitate better cationic lipid encapsulation and delivery, and that it may be possible to further improve on (-30)GFP as the fusion tag for protein delivery.

**Figure 5.2. GFP surface charge mostly correlates with delivery efficiency.** In general, an increase in the surface charge of sfGFP results in better lipid-mediated protein delivery assayed by a Cre-recombination assay. sfGFP is innately anionic with the net charge of -7. (-40)GFP exhibits impaired fluorescence, probably due to compromised folding or stability.
5.3 Human Proteome Analysis

The UniProt database was used to generate a list of human proteins, and they were ranked according to their theoretical charge-to-molecular weight ratio (Figure 5.3). Among the cohort of proteins with greater than 0.5 charge/MW, effort was made to identify proteins that (1) have crystal structures, (2) have a precedence of bacterial protein expression, (3) are identified as defective or missing in diseases, and (4) do not have extensive disulfide bonds or post-translational modifications. After manual screening of the proteins using these guidelines, 12 candidate proteins were chosen for further analysis (Table 5.1).

![Figure 5.3. Analysis of the human proteome from the UniProt protein database. All annotated human proteins were plotted in the graph. Green, yellow, and red dots represent proteins whose charge to molecular weight ratio exceed 0.8, 1.0, and 1.2 respectively. Blue dots represent various engineered sfGFPs for comparison.](image-url)
Table 5.1. List of highly anionic proteins chosen for investigation.

<table>
<thead>
<tr>
<th>#</th>
<th>Abbv.</th>
<th>Name</th>
<th>charge</th>
<th>MW</th>
<th>charge to MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IKBA</td>
<td>NF-kappa-B inhibitor alpha</td>
<td>-29</td>
<td>35605</td>
<td>-0.8</td>
</tr>
<tr>
<td>2</td>
<td>CAH8</td>
<td>Carbonic anhydrase-related protein</td>
<td>-21</td>
<td>32805</td>
<td>-0.6</td>
</tr>
<tr>
<td>3</td>
<td>AR2BP</td>
<td>ADP-ribosylation factor-like protein 2-binding protein</td>
<td>-20</td>
<td>17930</td>
<td>-1.1</td>
</tr>
<tr>
<td>4</td>
<td>MCFD2</td>
<td>Multiple coagulation factor deficiency protein 2</td>
<td>-17</td>
<td>13510</td>
<td>-1.3</td>
</tr>
<tr>
<td>5</td>
<td>DDIT3</td>
<td>DNA damage-inducible transcript 3 protein</td>
<td>-17</td>
<td>19154</td>
<td>-0.9</td>
</tr>
<tr>
<td>6</td>
<td>SIRT1</td>
<td>NAD-dependent protein deacetylase sirtuin-1</td>
<td>-59</td>
<td>76687</td>
<td>-0.8</td>
</tr>
<tr>
<td>7</td>
<td>PTMA</td>
<td>Prothymosin alpha (ProTa)</td>
<td>-44</td>
<td>12191</td>
<td>-3.6</td>
</tr>
<tr>
<td>8</td>
<td>PAIP2</td>
<td>Polyadenylate-binding protein-interacting protein 2</td>
<td>-25</td>
<td>14968</td>
<td>-1.7</td>
</tr>
<tr>
<td>9</td>
<td>TNNC1</td>
<td>Troponin C</td>
<td>-29</td>
<td>18382</td>
<td>-1.6</td>
</tr>
<tr>
<td>10</td>
<td>DPH3</td>
<td>DPH3 homolog</td>
<td>-14</td>
<td>9099</td>
<td>-1.5</td>
</tr>
<tr>
<td>11</td>
<td>RPAB2</td>
<td>DNA-directed RNA polymerases I, II, and III subunit RPABC2</td>
<td>-20</td>
<td>14462</td>
<td>-1.4</td>
</tr>
<tr>
<td>12</td>
<td>S100B</td>
<td>Protein S100-B</td>
<td>-14</td>
<td>10701</td>
<td>-1.3</td>
</tr>
</tbody>
</table>

5.4 Identification of a Human Protein that Potently Facilitates Protein Delivery

Each human protein was fused to Cre recombinase, and individually expressed and purified. Each Cre-fused human protein was complexed with Lipofectamine RNAiMAX, a commercially available lipid formulation that displayed the highest compatibility with protein delivery. The resulting lipid nanoparticles were delivered into HeLa cells with stably integrated floxed transcriptional terminator followed by a tdTomato gene (HeLa DsRed). Hence the delivery of Cre into the cell is fluorescently reported upon the elimination of the terminator sequence. Untagged Cre in the presence of lipid required ~70 nM to cause 50% of the cells to be fluorescent (EC50 = 70 nM) (Table 5.2). As expected, tagging (-30)GFP to Cre decreased its EC50 by ~10-fold (Table 5.2). Fusion of most human proteins did not enhance the recombination efficiency. However, the fusion of a protein called ProTa (from the PTMA gene) enabled approximately ~100-fold increase in recombination efficiency compared to that of untagged Cre. ProTa-Cre had EC50 of ~1 nM, which is ~10-fold lower even compared to (-30)GFP-Cre (Table 5.2 and Figure 5.4).
To demonstrate the generality of ProTα’s ability to deliver cargo across multiple different cell lines, we performed a lipid-mediated ProTα-Cre delivery into a BSR or a HEK293 cell line with a stably integrated RFP reporter. Consistent with the previous result, ProTα-Cre enabled approximately 100-fold better Cre delivery compared to untagged Cre, or 10-fold better delivery compared to (-30)GFP-tagged Cre (Figure 5.4).

Table 5.2. Delivery potency of human proteins fused to Cre in HeLa-DsRed cells. Each human protein was fused to the N-terminus of Cre, and delivered into HeLa-DsRed cells using Lipofectamine RNAiMAX at varying concentrations.

<table>
<thead>
<tr>
<th>Protein fused to Cre</th>
<th>Concentration at which 50% of cells are recombined (nM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTMA (ProTα)</td>
<td>1</td>
</tr>
<tr>
<td>PAIP2</td>
<td>289</td>
</tr>
<tr>
<td>TNNC1</td>
<td>57</td>
</tr>
<tr>
<td>DPH3</td>
<td>36</td>
</tr>
<tr>
<td>RPAB2</td>
<td>111</td>
</tr>
<tr>
<td>MCFD2</td>
<td>137</td>
</tr>
<tr>
<td>AR2BP</td>
<td>20</td>
</tr>
<tr>
<td>IKBA</td>
<td>63</td>
</tr>
<tr>
<td>DDIT3</td>
<td>89</td>
</tr>
<tr>
<td>CAH8</td>
<td>84</td>
</tr>
<tr>
<td>S100B</td>
<td>47</td>
</tr>
<tr>
<td>SIRT1</td>
<td>49</td>
</tr>
<tr>
<td>No fusion</td>
<td>73</td>
</tr>
<tr>
<td>-30GFP</td>
<td>10</td>
</tr>
</tbody>
</table>
Figure 5.4. ProTα enables potent delivery of Cre in multiple different cell lines. Different amounts of Cre, ProTα-Cre, and (-30)GFP-Cre were complexed with Lipofectamine RNAiMAX and delivered into HEK293-RFP, BSR-LNL-TdTomato, and HeLa-DsRed cell lines. In all cell lines tested, ProTα-Cre is 100-fold more efficient at delivery compared to unfused Cre, and 10-fold more efficient compared to (-30)GFP-Cre. Values and error bars reflect the mean and s.e.m. of three biological replicates performed on separate days.

5.5 Comparison with Charge Analogs

ProTα has -44 theoretical net charge and 3.8 charge-to-molecular weight ratio\textsuperscript{63,64}. To investigate whether ProTα’s delivery potency arises from simply having a higher net charge compared to (-30)GFP, we expressed two peptide tags composed mostly of aspartic acids and glutamic acids with theoretical net charges of -44 and -30 as fusions to Cre, termed (-44)PolyD/E-Cre and (-30)PolyD/E-Cre, respectively (Figure 5.5). We delivered into HeLa-DsRed cells both (-44)Poly-D/E-Cre and (-30)Poly-D/E-Cre with Lipofectamine RNAiMAX and assayed for their ability to induce Cre recombination. Interestingly, the two analogs failed to recapitulate ProTα’s potency, as neither tags were more potent than ProTα (Figure 5.5). The results show that ProTα’s delivery potency does not simply arise from its high anionic charge.
Figure 5.5. ProvTα potency cannot be recapitulated with poly-Asp/Glu tags. Simple conjunction of acidic residues to the N-terminus of Cre does not recapitulate the potent delivery properties of ProvTα-Cre. The chart on the left shows the sequence of the tags tested. The charge tags were compared against ProvTα-Cre for their ability to deliver Cre into HeLa-DsRed cells. Values and error bars reflect the mean and s.d. of three biological replicates performed on separate days.

5.6 ProvTα Truncation Analysis

We sought to identify regions of ProvTα responsible for enabling efficient liposome-mediated protein delivery. ProvTα is an unstructured peptide with multifaceted functionalities. ProvTα has sequence homology to multiple different protein domains, and different regions of ProvTα are thought to carry out different functions in both intracellular and extracellular environments. Intracellularly, ProvTα has been used as cancer markers, and is thought to be involved in cell cycle regulation. Extracellularly, it was shown to regulate the immune response. ProvTα can be broken down into three domains: N-terminal basic domain (residues 1-39), central acidic domain (39-100), and C-terminal NLS (100-109). The basic region (residue 1-28) has immunomodulatory functions, and contains a suspected receptor-binding domain. The first ~89 residues have also been implicated in various antiviral, antifungal, and anti-cancer activities. Residues 100-109 also have separate immune-modulatory functions, including activating

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProvTα (-44)</td>
<td>MGASDAAVDTSSEITTKDLKKEKKEVVEAENG RDAPANGNAENENEGQEADNEVDSEEEEG GEEEEEEDGGEEDGDEDEEASATGRRAA EDDDDVDTKKQKTDDED</td>
</tr>
<tr>
<td>(-44) poly D/E</td>
<td>MGADEESESDEELDEEELDEEEDTDEEGG DEEDEDELDEEDTDEEDEDEE</td>
</tr>
<tr>
<td>(-30) poly D/E</td>
<td>MGADEESESDEEELDEEELDEEEDTDEEGG DEEDELDEEDEEDEDEEDEE</td>
</tr>
</tbody>
</table>
macrophages, monocytes, natural killer cells, DC and neutrophils. Recently, ProTα has been shown to interact with histones, which facilitates their trafficking within the nucleus.

To further probe the mechanism behind ProTα’s ability to deliver cargos into cells, we truncated various regions of ProTα and expressed them as fusions to Cre (Figure 5.6). After formulating them into liposomes using Lipofectamine RNAiMAX, they were delivered into the BSR-TdTomato cell line. After 3d, the cells were analyzed for RFP fluorescence using flow cytometry. As expected, the two charge deletion variants (B3 and B5) showed a dramatic decrease in delivery efficiency (Figure 5.6). The N-terminal truncation variant (B4), missing the suspected receptor-binding activity of ProTα, decreased the EC50 of the assay by ~2-fold compared to the full length ProTα (Figure 5.6). Surprisingly, the deletion of the putative NLS (B2) from ProTα decreased EC50 about ~5-fold compared to full length ProTα (Figure 5.6). This is unexpected as Cre normally does not require an NLS. Consistently, the simple fusion of an SV40 NLS to the N-terminus of Cre did not improve its ability to catalyze recombination (Figure 5.7). These results show that multiple domains of ProTα are responsible for mediating potent liposomal protein delivery.
Figure 5.6. **Truncation analysis of ProTa.** The block diagram shows the full ProTa-Cre fusion and various truncations to the ProTa domains in black (B2-B5). Truncating the putative NLS resulted in a modest (~5 fold) decrease in delivery efficiency. Truncating the N-terminal domain resulted in a slight (~2 fold) decrease in delivery efficiency. Removal of the charge domain in both truncation variants (B3, B5) resulted in a dramatic reduction in delivery efficiency. Values and error bars reflect the mean and s.e.m. of five biological replicates performed on separate days.

Figure 5.7. **SV40 NLS has insignificant impact on Cre recombination.** Cre does not benefit from the addition of an SV40 NLS to its N-terminus when delivered using cationic lipids. Values and error bars reflect the mean and s.d. of three biological replicates performed on separate days.
5.7 Lycotoxin L17E Does Not Improve Cytosolic Protein Delivery.

M-Lycotoxin is an amphiphilic peptide with strong membrane-lytic activity\textsuperscript{70}. Recently, it was engineered to be activated only under acidic conditions, by introducing a glutamic acid residue in place of a leucine positioned in its hydrophobic face\textsuperscript{70}. The peptide, termed L17E, has been used to facilitate the lysis of endosomes during endosome maturation\textsuperscript{70}. We reasoned that L17E could have a synergistic effect when used with cationic lipids for protein delivery, potentially facilitating even more efficient cytosolic access.

To explore the compatibility of L17E’s endosome-disrupting ability with liposomal protein delivery, we fused L17E to the N-terminus of both Cre and ProTα-Cre to make L17E-Cre and L17E-ProTα-Cre, respectively. We delivered Cre, L17E-Cre, or L17E-ProTα-Cre into HeLa-DsRed cells, with or without lipid encapsulation. To replicate as much as possible the conditions described previously for L17E activity\textsuperscript{70}, we also delivered Cre, L17E-Cre, or L17E-ProTα-Cre doped with unconjugated L17E peptide into HeLa-DsRed cells. We then assayed the cells using flow cytometry after 3 d. We failed to observe a significant improvement in the cytosolic delivery of Cre and its variants in all conditions tested (Figure 5.8). We note that much higher doses of protein have been used in the previous publication (10-100 µM range), which may account for the discrepancies\textsuperscript{70}. 
Figure 5.8. **Lycotoxin L17E does not improve Cre delivery into HeLa-DsRed cells.** L17E was either fused to ProTα-Cre or Cre, and delivered using Lipofectamine RNAiMAX. (+L17E) denotes a condition where unfused L17E is treated at 20 µM. At the concentration regime tested in this experiment, L17E failed to show an improvement in all treatment conditions.

5.8 ProTα-mediated delivery of zinc finger nucleases.

Zinc finger nucleases (ZFNs) are a class of programmable DNA endonucleases\(^\text{71}\). Two sets of zinc-finger arrays are each fused to half of the heterodimeric FokI domain\(^\text{71}\). When the two zinc-finger arrays bind to the adjacent DNA sequence, two halves of the FokI domain dimerize and catalyze a double-stranded DNA break\(^\text{71}\). ZFN-based strategies have the potential to become effective therapeutics\(^\text{71}\). Genome-editing agents (such as ZFNs, Cas9, and base editors) benefit from transient exposure to target cells through protein delivery, since their prolonged presence in target cells or tissues cause elevated levels of off-target editing\(^\text{72}\).

We sought to demonstrate ProTα-mediated delivery of ZFNs into human cells. We fused ProTα to the N-terminus of each ZFN (ProTα-ZFN) designed to target AAVS1 site in human cells, a safe-harbor locus commonly used for gene integration\(^\text{73}\) (Figure 5.9a). We expressed and
purified ProTα-ZFN and ZFN protein pairs. Interestingly, fusion of ProTα to ZFNs markedly improved the yield and purity of the proteins, whereas ZFNs alone expressed poorly, consistent with literature. The fusion of ProTα to ZFNs did not hamper its ability to cut DNA \textit{in vitro} (Figure 5.9b). Next, we wanted to ensure that the fusion of ProTα did not affect its ability to cut the DNA once they are within the cell. When plasmids encoding either ProTα-ZFN pairs or ZFN pairs were transiently delivered to HEK293T cells, both treatments resulted in up to 78% of indel formation when assayed using HTS (Figure 5.9c). As expected, both the ‘left’ and the ‘right’ components of ZFNs were required for nuclease activity (Figure 5.9c). Next, we delivered ProTα-ZFN pairs and ZFN pairs, with and without lipid complexation, into HEK293T cells in serum-containing media. Whereas treatment of ProTα-ZFNs combined with lipid resulted in up to ~18% indel generation, treatment of unfused ZFNs with lipid resulted in much lower amount of indel formation (Figure 5.9d). Indel generation also required the presence of the lipid reagent (Figure 5.9d).

ZFNs, due to their high innate positive charge, have been shown to deliver into human cells without the use of lipid reagents. However, such a method requires serum-free media, low-micromolar protein concentrations, and repeated incubations to achieve modest genome editing. Furthermore, positively charged ZFNs are not compatible with cationic lipid-mediated protein delivery. On the contrary, ProTα-fused ZFNs are efficiently encapsulated within liposomes and delivered into cells, resulting in efficient DNA cleavage at mid nanomolar concentrations in the presence of serum media with single treatment.
Figure 5.9. ProTα enables efficient delivery of ZFNs into HEK293T cells. (a) ProTα-ZFN and ZFN gene architectures used for protein expression. (b) Fusion of ProTα to the N-terminus of ZFN does not impair its ability to cut DNA in test tube. (c) ProTα-ZFN and ZFN pairs can mediate efficient indel generation when expressed in HEK293T cells with transient plasmid transfection. (d) ProTα enables efficient lipid-mediated delivery of ZFNs at mid-nanomolar concentrations into HEK293T cells in presence of media containing 5% serum.
5.9 Conclusion

ProTα is a small human protein that can be used as a potent delivery tag in combination with cationic lipids. As ProTα expression is universal in all tissues\textsuperscript{67}, the body is less likely to mount an immune response when used \textit{in vivo}. The ProTα truncation analysis and comparison with the charge analogs show that different domains within ProTα contribute to its potent ability to be encapsulated by cationic liposomes and facilitate cytosolic delivery. Using cationic lipids, ZFNs tagged with ProTα can be efficiently delivered into human cells to enable genome editing. ProTα is a potent protein tag that may be used with cationic lipids to enable cytosolic delivery of many macromolecules.

5.10 Methods

**Cloning.** PCR was performed using Q5 Hot Start High-Fidelity DNA Polymerase (New England BioLabs). Candidate human protein DNAs were purchased from IDT as gBlock Gene Fragments. Bacterial expression plasmids encoding human protein fused to Cre were made using USER-cloning (New England BioLabs). Truncation of ProTα was done using blunt-end ligation to delete regions of ProTα. Following PCR, KLD enzyme mix (New England BioLabs) was used to phosphorylate and circularize the PCR product before transformation into NEB10beta cells (New England BioLabs).

**Protein Expression.** BL21 Star (DE3)-competent E. coli cells (ThermoFisher Scientific) were transformed with plasmids encoding the human proteins fused with Cre with a His\textsubscript{6} C-terminal purification tag. A single colony was grown overnight in 2xYT broth containing 50 \(\mu\text{g/ml}\) Carbenicillin at 37 °C. The cells were diluted 1:20 into 1 L of the same media and grown until OD600 ~ 0.5. The cultures were incubated on ice for 60 min and protein expression was induced.
with 0.5 mM isopropyl-b-D-1-thiogalactopyranoside (IPTG, GoldBio Sciences). Protein was expressed for 14–16 h with shaking at 16 °C. Cells were centrifuged at 10,000 rpm for 20 min, and then resuspended in a high salt buffer (100 mM tris(hydroxymethyl)-aminomethane (Tris)-HCl, pH 8.0, 1 M NaCl, 20% glycerol, 5 mM tris(2-carboxyethyl)phosphine (TCEP; GoldBio) with a protease inhibitor pellet (Roche). The cells were lysed using sonication and the supernatant was incubated with His-Pur nickel nitrioloacetic acid (nickel-NTA) resin (ThermoFisher) with rotation at 4 °C for 30 min. The resin was washed with the high salt buffer before the protein was eluted with an elution buffer (high salt buffer supplemented with 200 mM imidazole). The eluent was purified on a 5 ml Hi-Trap Q (GE Healthcare) anion exchange column using an AKTA Pure FPLC. The purified protein was quantified by a Pierce microplate BCA protein assay kit (Pierce Biotechnology) and snap-frozen in liquid nitrogen and stored at -80 °C until before use. ZFNs were purified according to previous literature\textsuperscript{74,76}.

**In vitro DNA cleavage assay.** AAVS1 locus containing the ZFN target sequence was amplified from purified HEK293T genomic DNA using PCR. ~350 bp PCR product was purified using Minelute columns (Qiagen). 100 ng of DNA substrate was incubated with 300 nM of ZFN or ProTα-ZFN pairs in Cutsmart buffer with 1 mM Arginine and 100 µM ZnCl\textsubscript{2} (New England Biolabs) at room temperature for 16 hours. The cleavage product was detected by running the mixture on an agarose gel.

**Cell culture.** HeLa DsRed, BSR TdTomato, and HEK293-loxP-GFP-RFP (GenTarget) Cre reporter cells were cultured in Dulbecco’s Modified Eagle’s Medium plus GlutaMax (ThermoFisher Scientific) supplemented with 10% (v/v) FBS, at 37 °C with 5% CO\textsubscript{2}.

**ZFN Transfections.** HEK293T cells were plated on a collagen-coated 48-well plate 1 d prior to experiment. At ~70% confluency, 500 ng of ‘left’ and ‘right’ CMV-ZFN and CMV-ProTα-ZFN
(1 µg total DNA content) were transfected using 1.5 µL Lipofectamine 2000 (ThermoFisher Scientific) according to manufacturer’s protocol.

**Protein delivery assays.** Protein was diluted to 12.5 µL in OptiMem and was complexed with 1.5 µL of Lipofectamine RNAiMAX (ThermoFisher Scientific) in 12.5 µL of OptiMem. The resulting complex was delivered to cells that had been seeded on a 48-well collagen-coated BioCoat plate (Corning) at ~70% confluency (250 µL final volume). After 3 d, the cells were trypsinized using TrypLE reagent (ThermoFisher Scientific), and resuspended in culture media before being analyzed using the CytoFlex flow cytometer (Beckman Coulter).

For ZFN experiments, equimolar amounts of ‘left’ and ‘right’ ZFNs was diluted to 12.5 µL in OptiMem and was complexed with 12.5 µL of OptiMem containing 3.5 µL of Lipofectamine RNAiMAX (ThermoFisher Scientific). The resulting complex was delivered to cells that had been seeded on a 48-well collagen-coated BioCoat plate (Corning) at ~70% confluency at the final volume of 100 µL per well. After 4 hr, the cells were incubated with fresh media for further 48 hours. Then, the cells were lysed and DNA was purified DNA was isolated using the Agencourt DNAdvance Genomic DNA Isolation Kit (Beckman Coulter) according to the manufacturer’s instructions. Genomic regions of interest were amplified by PCR with flanking high-throughput sequencing primer pairs. Then, DNA was further amplified by PCR with primers containing sequencing adaptors. The products were gel-purified and quantified using KAPA Library Quantification Kit-Illumina (KAPA Biosystems). Samples were sequenced on an Illumina MiSeq as previously described. Indel was quantified within the 30-base window surrounding the cleavage site among the high-quality reads (Q > 30) using the custom Matlab script (see Chapter 1).
Chapter Six: Discussion and Future Directions

Y. Bill Kim
6.1. Short-Term Limitations of Base Editing

Despite advances in base editing technology over the last two years, certain roadblocks prevent its wide adoption in research. These limitations arise from the imperfect parts that make up the base editor. Taking advantage of the modular nature of the base editor, engineering and evolution of its parts may address the limitations described in this section.

6.1.1. Variability in Base Editing Efficiencies

Base editing efficiencies vary, without clear logic, from site to site. This variability may be due to the epigenetic factors such as DNA methylation, hydroxymethylation, or histone modifications that prevent successful binding or catalysis by the base editor\(^77\). Approaches that map the epigenetic states in genomic loci can be used to understand why some sites are targetable by the base editor and some are not. If favorable chromatin and epigenetic states can be locally induced through a protein-based factor, fusing it to the base editor may expand its targeting scope even further.

6.1.2. Within-Window Precision

The precision of base editing, whether desired single nucleotide change is induced without modifying other bases in the vicinity, is important lest it introduce unintended amino acid changes in the gene. In Chapter 3, it was shown that the kinetic impairment of the deaminase narrows the apparent base-editing-window width. Nevertheless, this approach failed to sufficiently narrow the window width of SaKKH-BE3. In retrospect, this failure was not unexpected; BE3 (that uses \textit{S. pyogenes} Cas9) is particularly conducive to the kinetic approach of window-narrowing, unlike other base editors that use different Cas9 homologs.
The apparent narrowing of the base editing window via catalytic retardation of APOBEC1 depends on the fine tuning of two variables. First, the single-stranded DNA generated upon Cas9 binding needs to be only partially accessible. This depends on the structure of a specific Cas9. Having a large amount of exposed single-stranded DNA at the site of binding makes it unlikely for APOBEC1 to energetically favor binding to a specific base, as it may be able to accommodate, without significant energetic penalty, a variety of bases in the vicinity.

Second, the degree of freedom possessed by the base editor complex (the deaminase, Cas9, and the linker that connects between the two) needs not be excessive. Incorporation of an excessively long and flexible linker between APOBEC1 and Cas9 needlessly provides structural flexibility that enables the base editor to access multiple bases exposed in the single-stranded DNA with similar affinity.

Excess of either variables results in a state where the incorporation of kinetic mutations fail to narrow the apparent editing window, as it lacks the inherent base preference within the single-stranded DNA. To enable the apparent narrowing of the base editor window width, therefore, requires the minimum length of linker between Cas9 and APOBEC1 (that is still sufficient to induce maximal base editing).

Luckily, XTEN (the 16-amino-acid linker in BE3), contains enough length and flexibility to generate efficient base editing, as extending the linker (in BE4, for example) only minimally improves its efficiency\textsuperscript{42,48}. Yet, linkers that are shorter would have resulted in dramatic reduction in the efficiency (Figure 3.1). Also, \textit{S. pyogenes} Cas9 from BE3 contains protein domains that shield parts of the single-stranded DNA\textsuperscript{78}, limiting single-stranded DNA available to about 5 bases. Since BE3 has minimal structural flexibility and a small amount of exposed single-stranded DNA, each base within the 5-nucleotide window will have different energetic
cost associated with APOBEC1 binding. Hence, introducing a kinetic constraint to BE3 was able to narrow the apparent editing window width, as it allowed for the editing of only its highly-preferred base.

6.2. The Future of Base Editing

6.2.1. Base Editing in the Era of the Genome-Editing Revolution

The field of genome editing seeks the ability to manipulate the gene at will. To achieve this, however, researchers need to sufficiently address two questions; (1) ‘how can we target a specific stretch of DNA among millions of similar stretches of DNA with little difference in chemical composition?’ and (2) ‘how can we modify the DNA precisely?’ The discovery of programmable nucleases ushered in a new chapter in genome editing by demonstrating that proteins can easily be targeted to an arbitrary stretch of DNA specified by the researcher, addressing the former question with great efficiency.

Nonetheless, the prospect of actually deriving the desired change in DNA in vivo remained elusive. With programmable nucleases in hand, researchers sought to co-operate with the cellular machinery that deals with the product of the nuclease reaction: a double-stranded DNA break. Cellular machineries involved in DSB-repair pathways (such as NHEJ and HDR) are convoluted and cannot be easily manipulated. Their behavior is not general across different cell types. Sometimes heterogeneous product mixtures generated from NHEJ are sufficient for gene disruption. Yet, many other applications would benefit from the ability to modify DNA at a more precise level.

Base editing counters the notion that, because programmable nucleases are recruited to the target site, we must fix the resulting DSB into the desired product. Instead of relying on the
fickle cellular machinery for key transformation steps during the genome-editing event, the base editor contains all the components required for modifying DNA. With base editing, for the first time, the cell need not cooperate with the researcher to bring about a specific genetic change.

As base-editing technology matures, more sophisticated tools will emerge. The next generation of genome-editing tools will rely even less on the components provided from the cell. Below are a few possible examples.

6.2.2. Eliminating Genomic Off-Target Base Editing

It is not safe to harbor active genome-editing agents in the cell for long periods of time, as they may cause DNA damage\textsuperscript{80}. It is better to minimize the exposure of genome editing agents spatially and temporally to only where and when they are needed.

An approach hitherto unexplored is to generate a base editor that only works at the target site but not at off-target sites. This can be achieved by exploiting the mechanism of base editing. Base editors have finite residence time on DNA. Once bound to the target site, the increase in effective molarity of the deaminase in proximity to its substrate enables base editing to occur. On the contrary, binding at off-target sites is more transient, as a mismatched gRNA or incorrect PAM leads to a destabilized complex. Base editors could be designed in a way such that only the on-target binding of base editors allows for sufficient time to allow catalytic turnover by the enzyme. This idea seems to be corroborated by data in Chapter 3 where YE1-BE3 and other kinetic variants show markedly less off-target base editing compared to BE3.

Presumably, this strategic catalytic impairment would be equally useful in minimizing off-target RNA base editing, as wildtype APOBEC1 and most other deaminases natively act on RNA\textsuperscript{25}.
6.2.3. New Base Editing Transformations

Deaminases have been exclusively used in base editors thus far to transform C into U (using a cytidine deaminase\textsuperscript{42}) or, more recently, A into I (inosine, which base pairs with A, using an engineered adenine deaminase\textsuperscript{81}). Use of novel enzymes or chemical catalysts that perform different chemistry on DNA may extend the repertoire of base transformations on the genome.

Transversion reactions, for example, converting G to C or T, in addition to the deaminase chemistry, would enable virtually all possible single base nucleotide transformations. Yet transversions remain the most difficult, as a chemical means to replace a purine with a pyrimidine, and vice versa, is not straightforward. Strategically modifying the base or the sugar moiety could alter its base-pairing preference to induce such mutations, or the target base or nucleotide may be replaced as a whole.

6.2.4. Beyond Base Editing

Base editing is useful because it can correct single-nucleotide mutations that cause disease. However, it is useful to be able to modify a stretch of DNA, or a whole gene, which cannot be done through single-nucleotid changes. Although with poor efficiency, homology directed repair (HDR) is able to accomplish this. Eventually, newer genome-editing agents may be developed that bypass HDR but accomplish the same end product with better consistency and robustness without generating indels.

Recombinases are adept at manipulating DNA – inserting and removing chunks of DNA without requiring multiple protein complexes. However, direct DNA recombination with
researcher-specified target and substrate has not been widely successful to date. Naturally occurring recombinases have evolved to work at specific sequences of DNA, lest they wreak havoc on their own genome. Therefore, extensive engineering of the recombinase may be required to enable site-specific multi-base DNA manipulation.

It is possible that recombination of DNA is too delicate and complicated a process that engineering a general cut-and-paste module targeted to an arbitrary sequence is too difficult to achieve. An alternative may be to build a “recombinase complex” out of multiple proteins that behave modularly, akin to the machinery involved in biosynthetic pathways. Enzymes with various functions, such as nuclease, ligase, polymerase, can be brought in sequentially at the desired site. These “assembly-line” modules may be programmed to conduct coordinated multi-enzyme molecular surgery. An added benefit of this is that the “modules” are amenable for replacement and re-engineering to achieve a variety of transformations at the DNA level.

6.2.5. Beyond Genome Editing

Over the history of life, nature has generated a tremendous amount of diversity in genes and their products. Yet, throughout their genesis and evolution, constant selective pressure for survival has limited their possible solutions. As researchers gain mastery over DNA, they can imbue novel properties to the cell never encountered in natural history. Synthetic genes can be made to function alongside the natural genes. The synthetic genes will code not only for proteins with new chemical and biological function, aiding disease prevention or cure, but endow novel traits in humans. They may encode proteins with unnatural amino acids that can metabolize or polymerize new substances, or mediate information transfer between the biological body and the synthetic components that work outside of the body.
6.3. Concluding Remarks

Base editor is an unlikely fusion of three disparate proteins: a catalyst that modifies nucleotides, a protein that binds DNA sequence-specifically, and a binder that neutralizes an endogenous DNA repair pathway. The term ‘base editor’ aptly conveys the conceptual advance in the field that signals the possibility that we can take control of even the most challenging tasks as overwriting the genetic prose written by nature.

William Bateson, an English biologist who coined the term genetics, understood the inevitability of human intervention at the genetic level a century before it began. In 1905, Bateson remarked, “What ... will happen when ... the facts of heredity are as commonly known as those of bacteriology, for instance? One thing is certain: mankind will begin to interfere; perhaps not in England, but in some country more ready to break with the past and eager for “national efficiency.” ... Ignorance of the remoter consequences of interference has never long postponed such experiments ... The science of heredity will soon provide power on a stupendous scale; and in some country ... that power will be applied to control the composition of a nation. Whether the institution of such control will ultimately be good or bad for that nation or for humanity at large is a separate question.”

It is certain that the mastery of the gene confers humanity great power, the power to alter fundamental parts that define our identity. Great responsibility lies with humanity to shape the new generation of homo sapiens.
References


15. Cronican, J. J. *et al.* Potent delivery of functional proteins into mammalian cells *in vitro*


45. Zhang, Y. *et al.* Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA


49. Harrington, L. *et al.* A thermostable Cas9 with increased lifetime in human plasma. 1424 (2017) doi:10.1038/s41467-017-01408-4


73. Hockemeyer, D. *et al.* Efficient targeting of expressed and silent genes in human ESCs


