GONZALEZ, MICHELLE., M.S. Increasing CRISPR-Mediated Digital Data Storage in DNA
Using Engineered "Janus" Protospacers. (2023)
Directed by Dr. Eric A. Josephs. 56 pp.

DNA molecules can be used to store digital data if the sequence of their nucleotides is assigned to different values of '0' or '1'. Rather than having to synthesize new DNA molecules for data storage on demand, our laboratory and its collaborators have developed a method to "overwrite" the nucleotide sequences on a double-stranded DNA molecule by mutating them using CRISPR base editors *in vitro* so that new information can be stored on these existing DNA molecules. In this system, nucleotides on one strand in the sequences that are recognized by CRISPR effectors (the "protospacers") are changed from '0' (unmutated) to '1' (mutated). For this thesis, we sought to increase the data storage capacity in this system by engineering protospacer sequences where the CRISPR effectors could effectively recognize sequences on both of the strands ("Janus" protospacers) which is not generally or necessarily the case for a given protospacer sequence. We designed and experimentally validated several Janus protospacer sequences so that, for each of the protospacer sequences, it can potentially store a '0' (unmutated), a '1' (top strand mutated), a '2' (bottom strand mutated), or a '3' (both strands mutated). We were able to engineer Janus protospacers with sequence motifs that should be recognized by mutagens APOBEC3A or APOBEC3G, which have distinct mutational signatures and would further increase the capacity of the systems. In our work here, we attempt to use Janus protospacers so they could effectively double the information storage capacity of DNA molecules used for CRISPR-mediated DNA data storage, at minimum.

INCREASING CRISPR-MEDIATED DIGITAL DATA STORAGE IN DNA USING ENGINEERED "JANUS" PROTOSPACERS

by

Michelle Gonzalez

A Thesis
Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Greensboro

2023

Approved by

Dr. Eric A. Josephs
Committee Chair

DEDICATION

I would like to dedicate my research and thesis to my Luna family, in specifically, my dear mother, who has supported me through it all. Thank you for being the best mother I could ever have, for being my cheerleader and my shoulder to cry on. Gracias, for always being by my side. All the accomplishments I have achieved, I complete them with pride because behind of it all, will always be you. Your love, care, and prayers have embraced me to be better version of myself; to always love, care, share, and lead by example. My dear madre, you have always and continue to motivate me to challenge myself, work hard, and strive for success, saying that only the sky is the limit and that I can accomplish anything I put my mind to. It is because of you, I am where I am today, striving to be the best person I can be and following and working hard for my dreams and passions to come true. This is for you, mi madre querida.

APPROVAL PAGE

This thesis written by Michelle Gonzalez has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair ____________________________

Dr. Eric A. Josephs

Committee Members ____________________________

Dr. Dennis LaJeunesse

____

Dr. Daniel J.C. Herr

04/11/2023

Date of Acceptance by Committee

04/11/2023

Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to first acknowledge my adviser, Dr. Eric A. Josephs, for being very supportive and compassionate mentor and for wanting me to learn, grow, and succeed on this Master's thesis journey. I am inspired by Dr. Eric A. Josephs, for being very knowledgeable and having a creativity for different CRISPR applications projects. It has been a true honor to have been a part of and worked in the Dr. Josephs lab group, thank you for teaching and guiding me to become a better researcher. I have become very close to many of lab mates, who practically became my JSNN family, therefore, and I would acknowledge my lab group members, Ashley Herring Nicholas, Hillary Dimig, Merna Melad, Tasmia Islam, Tanjina Islam, Tinku Supakar, Naseem Salameh, Dr. Robert Glass, Dr. Rachel Tinker-Kulberg, Dr. Rammyani Bagchi, and Dr. Mohammad Salehin.

I would also like to acknowledge the whole faculty and staff of the Department of Nanoscience at the Joint School of Nanoscience and Nanoengineering for the tremendous support and opportunities offered through the past two years, while in the Nanoscience Master's program. A special thank you to my committee members Dr. Dennis LaJeunesse and Dr. Daniel J.C. Herr, whom have been supportive through each of the steps in this journey by motivating me, empowering me, and believing in me, when I sometimes did not.

This work has been supported the National Science Foundation (NSF) grant 2027738 and National Institutes of Health grant 5R35GM133483.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Nucleic Acids as a Medium for Digital Data Storage

Deoxyribonucleic acid (DNA) is the molecule responsible for storing the genetic information of a cell. In particular, DNA encodes the information to make for proteins responsible for performing cellular functions. DNA is a polymer composed of four different nucleotides: Adenine (A), Thymine (T), Cytosine (C), Guanine (G), that can be arranged in different sequences. DNA is mostly in the form of a double helix, where A nucleotides are paired with T nucleotides and vice versa, and C nucleotides paired with G nucleotides and vice versa, on the opposite antiparallel strand. Based on the order of the nucleotides, DNA can encode the information that can be used to make proteins, but synthetic DNA molecules can also be considered to hold digital information if the value of different nucleotides are assigned to have values of ‘00’, ‘01’, ’10, or ‘11’, for example, and read in the sequence order.

In this way DNA has been recognized to have a high storage density, to be molecularly stable, exhibit longevity, and to be able to store this information in a way that is energy efficient.[1,2,3] These features of DNA, which is the molecule of “genetic memory”, are attractive for applications of digital memory. Most digital memory that is used to store different kinds of data are made of silicon. The global memory demand is expected to exceed the silicon supply chain by the year 2040,[1] and so there is a need to identify non-traditional materials that can be used to store digital information.

The table in Figure 1 has a comparison of the current technologies that we use to store memory and the ability of DNA to store memory. The retention rate of hard disk and flash memory, such as USBs, on average is about 10 years while cellular DNA is over 100 years. The ability for cells to create new “data" by generating new DNA molecules has the equivalent reading and writing latency of approximately 100 µs per bit versus 3-5 ms per bit for a traditional hard disk. Compared with silicon-based data storage technologies, data writing in DNA using a DNA polymerase is considered a slow process and synthesizing oligonucleotides chemically is even slower.[1] Using DNA as a digital memory material is called NAM (Nucleic Acid Memory) products. The table demonstrates that the benefits of using cellular DNA as a digital memory material by having a having a higher read/write latency, a longer retention rate, and larger scale volumetric density.[1,4]

Figure 1. A comparison between different memory storage technologies with cellular DNA[1]

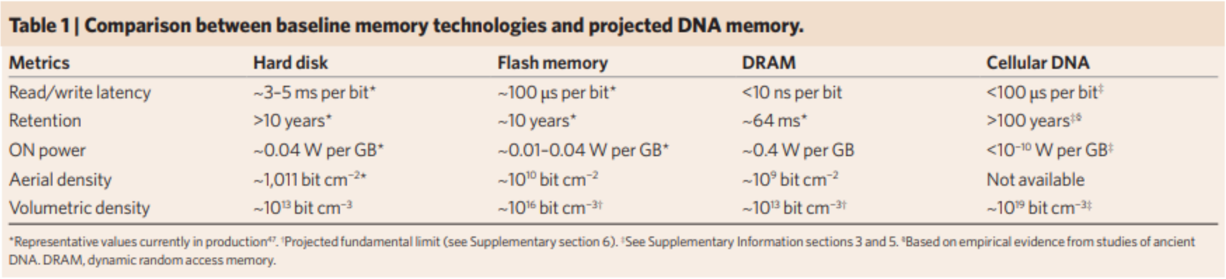


Table 1 | Comparison between baseline memory technologies and projected DNA memory.

Metrics	Hard disk	Flash memory	DRAM	Cellular DNA
Read/write latency	~3–5 ms per bit*	~100 µs per bit*	<10 ns per bit	<100 µs per bit‡
Retention	>10 years*	~10 years*	~64 ms*	>100 years‡§
ON power	~0.04 W per GB*	~0.01–0.04 W per GB*	~0.4 W per GB	<10^{-10} W per GB‡
Aerial density	~1,011 bit cm^{-2}*	~10^{10} bit cm^{-2}	~10^{9} bit cm^{-2}	Not available
Volumetric density	~10^{13} bit cm^{-3}	~10^{16} bit cm^{-3}†	~10^{13} bit cm^{-3}†	~10^{19} bit cm^{-3}‡

*Representative values currently in production[47]. †Projected fundamental limit (see Supplementary section 6). ‡See Supplementary Information sections 3 and 5. §Based on empirical evidence from studies of ancient DNA. DRAM, dynamic random access memory.

Note. The table contains a comparison of current technologies that are being used for digital memory storage and comparing it to the potential usage of cellular DNA, which holds a potential as a material for Nucleic Acid Memory (NAM).[1]

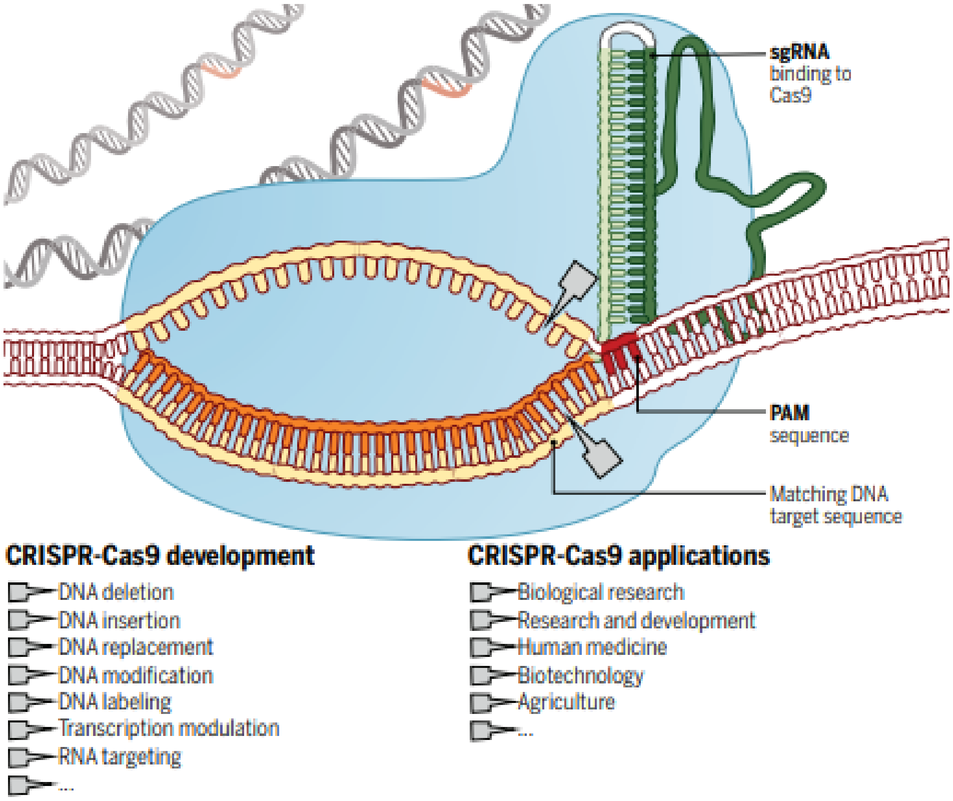
Despite there being several advantages to this alternative there also significant limitations to consider and overcome. The question that remains is how to practically write information into DNA for digital memory storage. Synthetic biology is a revolutionary field has the potential to address these limitations, and is a field which incorporates different areas and subjects such as therapeutics, industrial biomanufacturing, and the synthesis of biological materials. Using approaches from synthetic biology, novel DNA fragments can be generated by combining smaller oligonucleotides that can be chemically synthetized. This is known as gene synthesis and advances in this ability to generate DNA with nearly any nucleotide sequential order can be adapted for applications in digital data storage using DNA. These short genes, synthesized from fragments, are great for testing several different design possibilities, whether they are short sequences composed of a few base pairs (bp) or several kilobase pairs (over 1000 base pairs).[4,5,6,7,8,9] Designing and generating DNA fragments can cost up to almost $100 to $300 for a 1 kilobase (Kb) pair gene fragment.[5] While these costs allow for important applications of synthetic biology in the study of biological process and bioengineering, it should be recognized that the cost of potentially being able to store a 5-minute YouTube video onto a DNA molecule synthesized in this way would cost over 7 million dollars to produce, use around 100 KWh of energy, and could produce toxic waste in a volume of more than 15 liters.[5] In our lab, we use a biotechnology known as CRISPR systems to introduce targeted mutations into DNA, and these mutational signals can be used in a manner to encode the DNA with binary digits for coding without the need to chemically synthesize new DNA molecules in order to store new digital information.

Mechanistic Review of CRISPR Systems

First, it is important to understand what CRISPR is and how these systems work, and how scientist have come to use it. CRISPR, which stands for Clustered Regularly Interspaced Short Palindromic Repeats, and CRISPR-associated (Cas) proteins were originally discovered in bacteria and archaea, in which they would use these CRISPR-Cas systems as an adaptive immunity response to prevent reinfection by the same virus (bacteriophage). In a microbe with a CRISPR-Cas system, there is a segment of their genome that contains a series of palindromic repeats sequences, which are sequences that when read from either the top or bottom strand, they read as the same sequence.[10,11,12,13] For example, palindromic words that exist in English would be “mom”, “bob”, “Anna”, “level”, “wow”, etc. For DNA sequences, a palindromic sequence would be 5’- GAATTC-3’, which base pairs with 3’-CTTAAG-5’ or, when read 5’- to 3’-, as 5’-GAATTC’3. In between each of the palindromic repeats are “spacer” sequences that were derived from plasmid and viral genomic sequences. These spacer sequences are used by the CRISPR system so that it can recognize and destroy a plasmid or bacteriophage genome if it ever got infected by the same one.[12,13] When infected, the CRISPR locus in the microbe is transcribed into RNA along with the repeat sequence. The RNA is then processed and translated into CRISPR RNA (crRNA) and trans-activating CRISPR (tracrRNA), which in biotechnological applications they are fused together to make up what is called a single guide RNA (sgRNA). Either of these forms can be called “guide RNAs” or gRNAs.[13,14] The CRISPR effector, which in Type 2 CRISPR systems is known as Cas9, is a DNA nuclease that uses the crRNA:tracrRNA duplex or sgRNA molecule to recognize nucleotides that match the spacer sequence and introduce double-strand breaks into DNA at those sequences. Those sequences recognized by the Cas9 ribonucleoprotein (RNP) formed between the Cas9 and the gRNA are known as “protospacer” sequences. To prevent the Cas9 RNP from recognizing and destroying the spacer sequences themselves that are located in the microbe’s own CRISPR loci, the Cas9 RNP will only recognize protospacer sequences located next to specific nucleotide sequences found in the virus or plasmid but not next to the spacer in the CRISPR repeat. These sequences are known as protospacer adjacent motifs (PAM). Even if a DNA contains a protospacer sequence, it will not be recognized by the Cas9 enzyme if there is not a PAM sequence at its 3’ end. Our focus will be on using CRISPR-Cas9 proteins that are derived from *Streptococcus pyogenes,* which require a

‘NGG’ PAM sequence located immediately 3’- of the protospacer sequence, where ‘N’ is any nucleotide. Once the Cas9 has recognized the PAM sequence on the DNA, strand invasion by the spacer of the sgRNA occurs. Recognition of the protospacer begins when the Cas9 enzyme binds to the PAM and opens the DNA double helix at that site. The sgRNA will then “invade” the DNA double-helix by base-pairing its complementary strand, forming a structure known as an R-loop seen in Figure 2, nucleotide-by-nucleotide until the spacer is base-paired with its complementary sequence up to its 5’ end. This base-pairing activates the nuclease domains in the Cas9 enzyme to introduce double-strand breaks into the DNA. This RNA-guided nuclease activity used by the CRISPR-Cas9 enzyme is the foundation for many applications of targeted gene editing and many other purposes described in Figure 2.[12,13,14]

Figure 2. Illustration of Cas9 Enzyme and sgRNA binding complex [12]

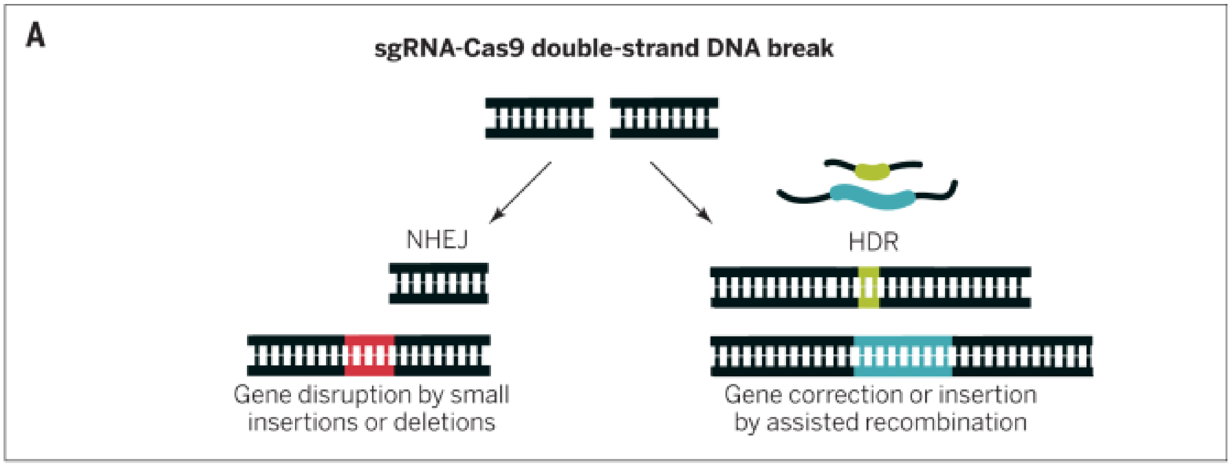


Note. Figure 2 is an illustration of the CRISPR-Cas9 complex (Cas9 enzyme in blue, the sgRNA (fused crRNA:tracrRNA) in green colors, and PAM sequence identified in red). The Cas9 identifies the PAM motifs and the sgRNA invades the DNA double-helix by base-pairing with its complementary DNA strand to form an “R-loop” structure that results in double-strand breaks by activating the nuclease domains. Below is a list of CRISPR-Cas9 applications of this RNA-guided nuclease activity.[12]

Targeted gene editing uses the nuclease activity of Cas9 and different cellular repair mechanisms to introduce mutations into genomic DNA at the sequences recognized by the spacer. In cells, double-strand breaks are differently repaired depending on the DNA repair mechanism: either Non-Homologous End Joining (NHEJ) and Homology Directed Repair.

Figure 3, provides an illustration of these DNA repair pathways. During the NHEJ pathway, the repair of a double strand break can disrupt certain genes with an insertion or deletion of a nucleotide. Homology Directed Repair pathway repairs breaks without introducing mutations by using a "template" DNA molecule as a basis to make sure the sequence is restored to its original form, but if exogenous DNA molecules with small differences to the sequence are also introduced into a cell with Cas9 those can be used as a template DNA during repair to introduce changes at those sites.[12,15]

Figure 3. Cas9-sgRNA complex double strand breaks and repair mechanism[12]


A
sgRNA-Cas9 double-strand DNA break
NHEJ
HDR
Gene disruption by small insertions or deletions
Gene correction or insertion by assisted recombination

Note. Figure 3 illustrates the two DNA repair pathways that can occur during a double strand DNA break (DSB). Non-Homologous End Joining (NHEJ) is a primary pathway in most mammalian cell cycles and can introduce mutations. Homology Directed Repair can also be used to introduce mutations if DNA molecules with small differences with the original sequence are used as a template for repair. [12]

For the experiments in this thesis, we want to be able to use the Cas9 enzyme to confirm whether the sgRNAs we design effectively recognize and form stable R-loops with their intended protospacers in the DNA molecules we generate. However, for the applications of digital data storage where we want to introduce mutations without double-strand breaks, we do not want to use the active Cas9 enzyme. Instead, we want to be able to use the deactivated version of Cas9 called endonuclease deficient Cas9 or dCas9. dCas9 is a mutant version form of Cas9, in which the endonuclease activity is eliminated as a result of mutations in the endonuclease domains of Cas9. The dCas9 still enables us to target and direct our mutations that may occur within the protospacer sequences without generating the double-strand DNA breaks.[16,17,18] Therefore, with the dCas9, we will able to still form a R-loop conformation at the targeted protospacer sequences using the corresponding sgRNAs.

“Base Editing" with CRISPR Systems

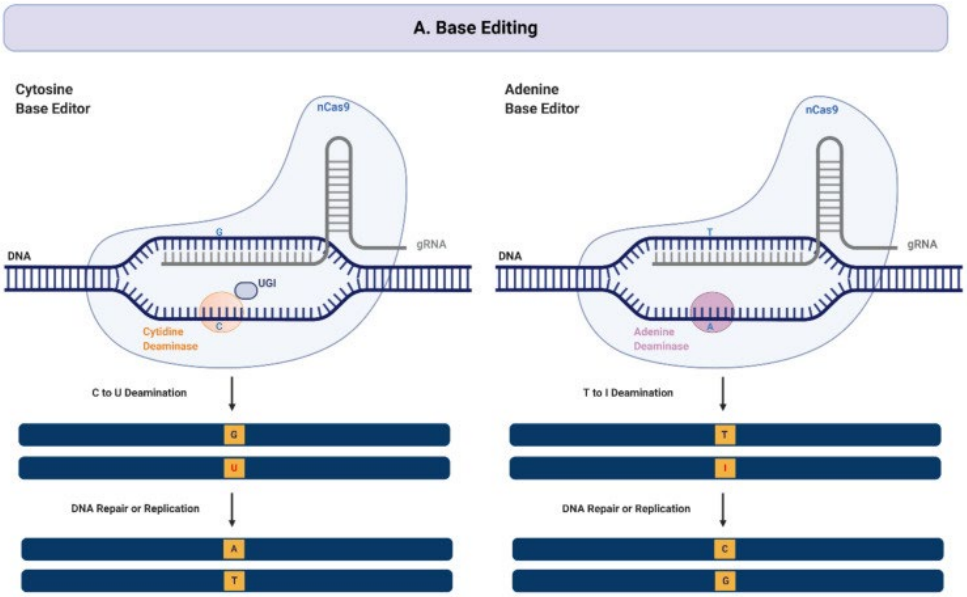
We wish to introduce mutations into DNA molecules without double-strand breaks for applications of digital data storage capacity, and to do so we use “base editor” proteins that chemically alter the nucleotides of DNA themselves. There are two types of base editors: cytosine base editors and adenine base editors. These base editors combine CRISPR proteins with other proteins called cytosine deaminases or adenosine deaminases that recognize single-stranded cytosines or adenines to convert them to uracil or inosine, respectively. These deaminases are strongly inhibited when the nucleotides they recognize are base-paired in a double-helix. If a CRISPR effector binds to specific protospacer and forms an R-loop, one strand of the DNA (the one that matches the sequence of the spacer) is no longer base-paired and susceptible to modification by the deaminases. Therefore, CRISPR can be used to form small regions where the chemical modifications of the nucleotides can occur within a double helix, as illustrated in Figure 4.[19] The cytosine base editor will deaminate the cytosines to the nucleotide uracil, that after DNA replication is recognized and used as a thymine to generate C > T mutations within protospacer sequence. Adenine base editors will deaminated the adenine to an inosine, and then after replication the inosine is recognized as a guanine to generate a A > G mutations within the protospacer sequence. This way, mutations can be introduced without double-strand breaks.[19]

There are a variety of specific cytosine and adenosine deaminases. One of the most used cytosine deaminases used in base editors are the APOBEC proteins. APOBEC proteins play an important role in human causation of illnesses and diseases. There are 11 different types of APOBEC proteins, each with a distinct function and structure. The types include the AID, A1, A2, A3A, A3B, A3C, A3D, A3F, A3G, A3H, and A4. Some of the APOBEC proteins can cause hypermutations when they are overexpressed. Many hypermutations and cancer resistance being built through APOBEC proteins are due to the cellular control of APOPEC activities.[20,21]

Two most common APOBEC proteins used in base editing applications to cause specific point mutations are APOBEC3A (A3A) and APOBEC3G (A3G). Both APOPEC3 (A3A and A3G) proteins have a preference for specific sequence motifs in which cytosines are most effectively modified. For example, in Figure 5, it illustrates A3A has a high affinity for deaminating cytosines that are right next to a thymine residue 5’ – T$\underline{\text{C}}$ – 3’, while A3G preferentially recognizes cytosines in a triple cytosine motif 5’ – $\underline{\text{CCC}}$ – 3’.[20,21,22,23]

The most popular adenine base editor (ABE) that is used in research is the ABE 7.10 generation base editor. There are other generational variants of ABE proteins. Each variant was created to have a high targeting efficiency. The ABE have a TadA component that was derived from *Escherichia coli*, which is a Transfer RNA (tRNA) deaminase. The version that we will be using for our experiments in this thesis is the ABE7.10; TadA–TadA*–nCas9 heterodimer. As displayed in Figure 4, the adenine base editor conducts adenosine to inosine conversions, and through the process of DNA base repair mechanism, the inosine is converted to a guanine. [23,24,25,26]

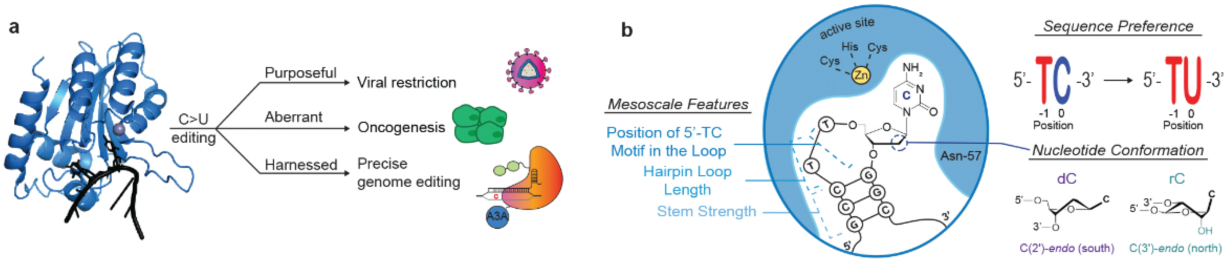
Figure 4. Schematic illustrating the mechanism of CRISPR base editing[19]



Note. Figure 4 illustrates the how cytosine (left) and adenine (right) base editors use deaminases that mutate single-stranded nucleotides in the R-loop formed by a CRISPR effector.[19]

As described below, CRISPR effectors and deaminase proteins will be used in this thesis to generate mutational changes at specific sites in a DNA molecule and that will be used to store digital data without requiring DNA synthesis on demand.

Figure 5. APOBEC3A (A3A) and sequence preferences for cytosine deamination.[21]

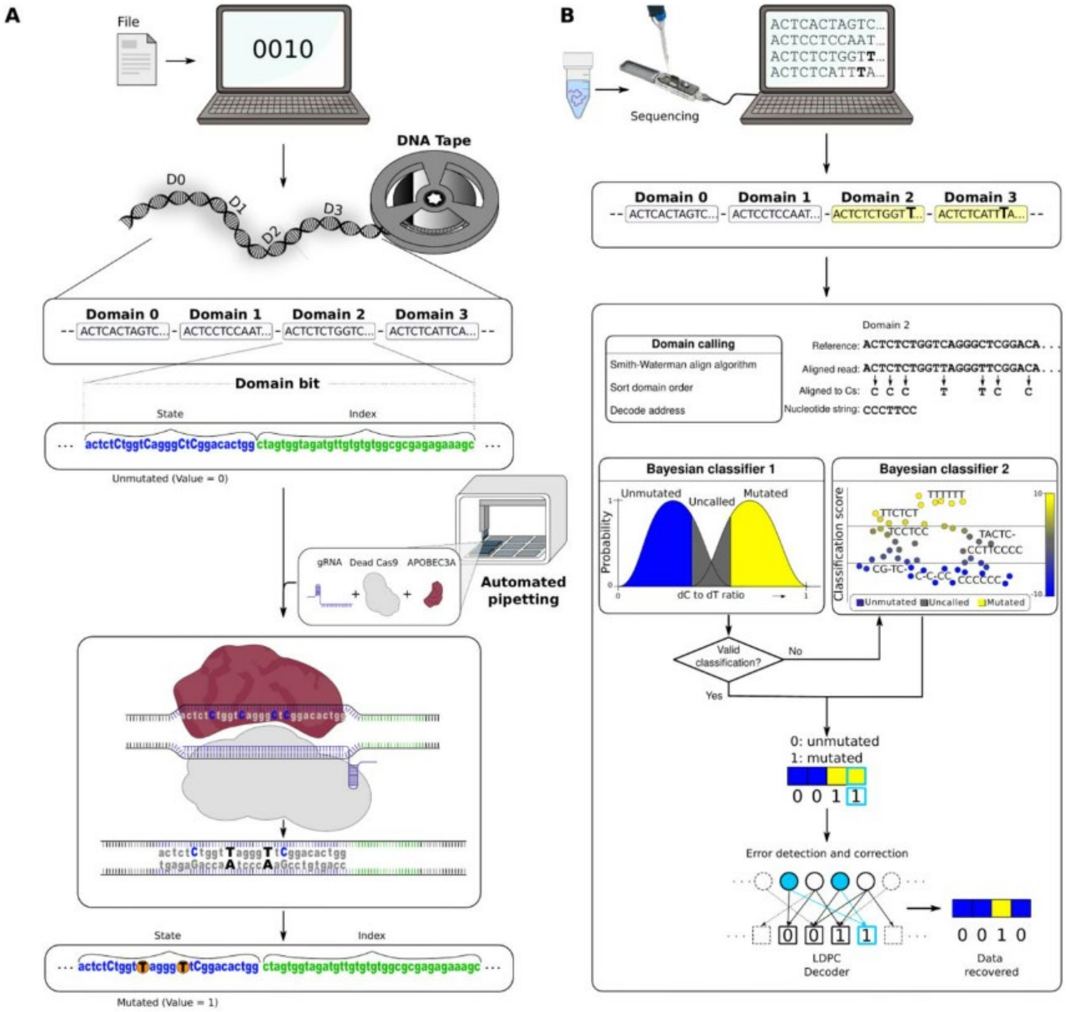


Note. Figure 5 demonstrates the biological and pathological background of APOBEC3A as well as the deamination that occurs at the TC motif.[21]

Using CRISPR systems for DNA data storage without requiring synthesis on demand

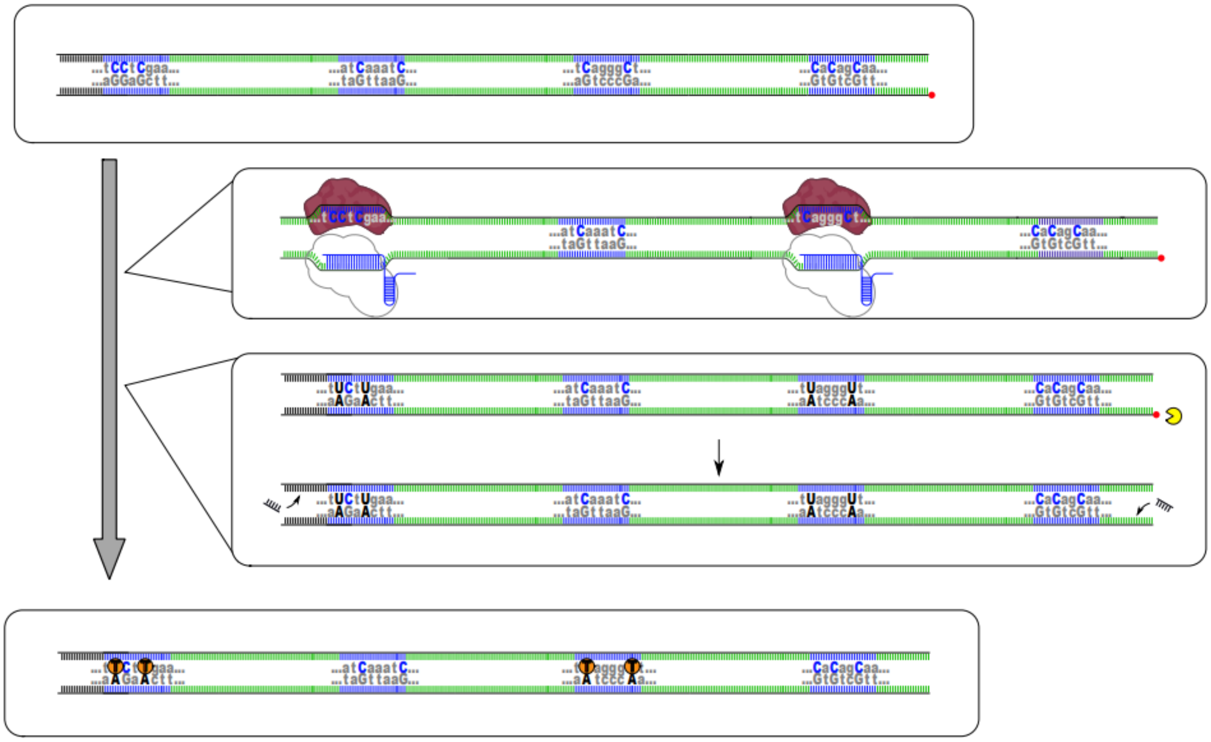
In the DNA Mutational Overwriting Storage (DMOS) project, a synthetic DNA molecule was created with a series of 16 protospacers that each contained multiple cytosines in TC motifs.[5] In this project, these DNA molecules, called "DNA tapes", were mutated *in vitro* using a CRISPR-dCas9 complex to recognize the 3 bp PAM motif along with cytosine deaminase APOBEC3A, which in specific has mutational motif preference of cytosines that are next to a thymine in DNA sequence (5' -TC -3') (Figures 6 and 7). The dCas9 enzyme recognizes that PAM motifs of the designed domain or registers, which will open the DNA double strand helix in to an R-loop confirmation with its dCas9-sgRNA complex. The spacer sequence of an sgRNA is 20 bp sequence that recognizes the "state" region of the DNA register in which will contain the at targeted cytosines. Whether or not the sequence of a particular "state" or protospacer was mutated could be read as “bits” (0’s or 1’s) by using DNA sequencing technologies, while the set of all the protospacers on a specific DNA molecule were called registers. This way, the same sets of DNA molecules or registers (the DNA “tape”) could be used to record digital data depending on which dCas9-sgRNA complexes were added to the reaction.[5]

Figure 6. A schematic review of the DMOS project to show DNA tapes can be written and read.[5]



Note. Figure 6 is schematic illustration of how the DMOS project was able to encode digital data onto DNA molecules without on-demand DNA synthesis. Their system was able to recognize and read the desired state regions of the DNA domain or bits, so that the system would read the states without a mutational protospacer (without cytosine mutated to thymine) as '0' digit in a binary coding system, while states with a mutated protospacer was read as a '1' digit.[5]

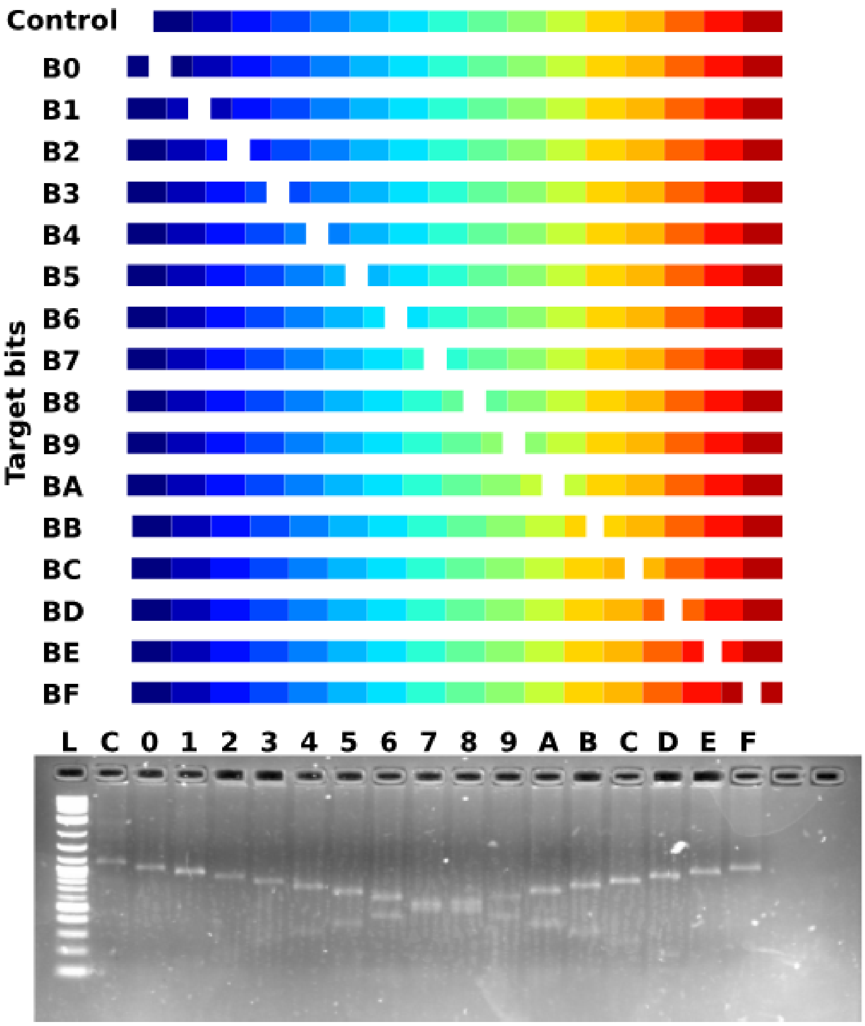
Figure 7. Schematic of reading and writing deaminase mutational signatures for binary coding using dCas9- APOBEC3A[5]



Note. Figure 7 demonstrates a schematic of how the APOBEC3A conducts cytosine deaminase on the top strand of the DMOS DNA type of the state regions of the protospacer, while the sgRNA intervenes in the R-loop by binding to the bottom strand.[5] After the dCas9 binds to the protospacer, the APOBEC3A enzyme mutates the cytosines in a TC motif within the R-loop. The unmutated strand is digested by lambda exonuclease and the mutated strand PCR'ed using uracil-tolerant polymerases to replicate the mutated strand.[5]

The sgRNA used in the DMOS project were designed and synthesized so that each protospacer domains (bits) would be recognized with their complementary designed sgRNA. The formation of a robust R-loop was verified using an active version of Cas9 to enable a double strand break, and the fragments of the DNA visualized using gel electrophoresis. Each protospacer domain (bits) had unique location on the DMOS DNA type, with a ribonucleotide protein complex (RNP) or "Cas9-sgRNA" the DNA type would introduce double-strand breaks or cut the DNA strand at the expected the protospacer when using the corresponding sgRNA. This double strand breaks would produce two fragments of different sizes that could be resolved using gel electrophoresis. Using a DNA ladder in the gel, a gel analysis can be conduct to visualize what size the DNA bands are and if they are the correct band length as the expected cut would be. The results revealed that each Cas9-sgRNA was very specific to its protospacer since there was no other fragments observed with the incorrect sizes, and also that they formed robust R-loops since the Cas9 were highly efficient in forming double-strand breaks.[5]

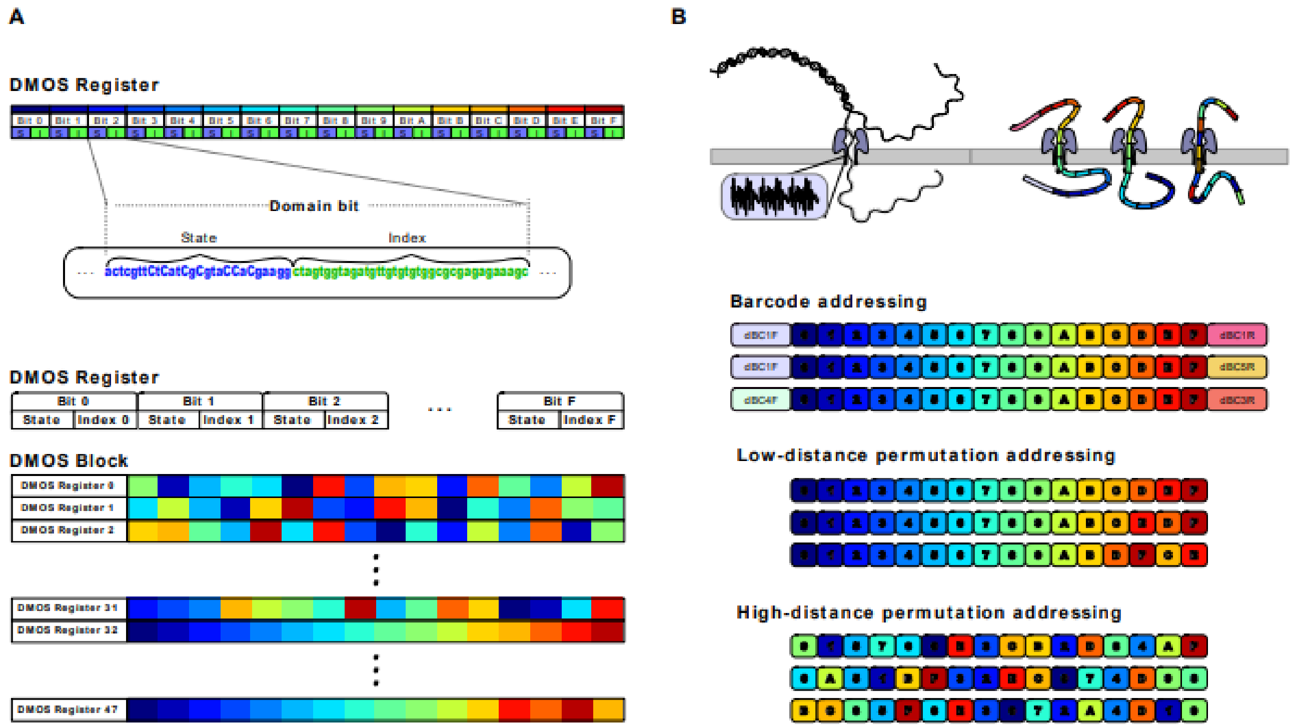
Figure 8. Agarose gel electrophoresis confirming on-target efficiency of sgRNAs of a DNA tape where each domain is a desire bit.[5]



Note. Figure 8 demonstrates a gel electrophoresis confirmation that the 16 guide RNAs for the DMOS experiments with Cas9 complexes were highly efficient in cleaving the DNA template at their precise domains.[5]

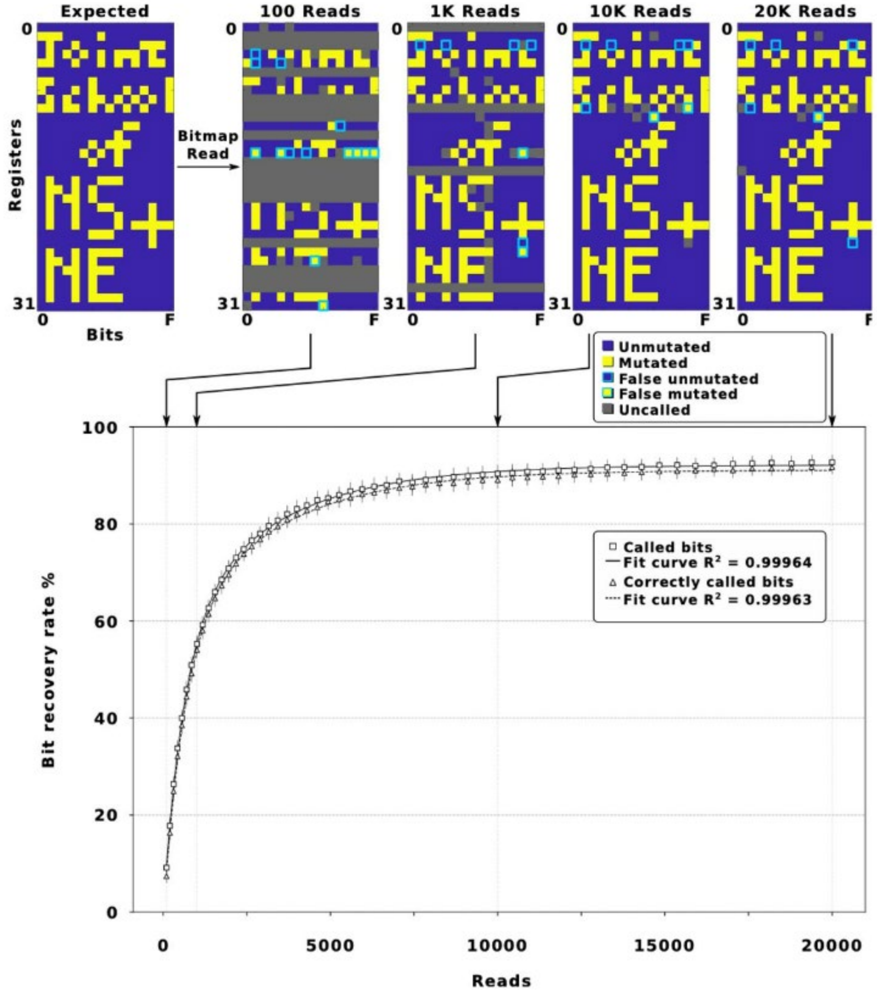
In the DMOS project, these mutations at each domain determined whether that bit was read as a digit '0' or a digit '1'. They were able to further establish that they could increase the data storage capacity in DNA by shuffling the DNA domains to create different combinations of register DNA blocks to encode data using the same sets of protospacers and gRNAs (Figure 9). In this way, they demonstrated over 1200 bits could be written into DNA molecules without on-demand synthesis and decoding using nanopore sequencing with >98% accuracy. With messages using error-correcting codes, they were able to decode short messages using this system with 100% accuracy.[5]

Figure 9. Design of DMOS DNA type using domain registers as a bit coding system.[5]



Note. Figure 9 demonstrates the design of the DMOS DNA type used in their experiments. It illustrates how the state in the DMOS DNA type is a nucleotide sequence of a 23 bp-length including the 3 bp PAM motif and has a 40 bp 'index' sequence. The figure also illustrates how the DNA is going through nanopore to be sequence and analyzed in nucleotide signatures (signals). In the DMOS project they were also able to increase data storage capacity by shuffling the order of the domains (which are read as bits) in each DMOS register creating a variety combination of these registers read as DNA blocks. In Figure 10, the DMOS project was able to display the JSNN logo, which is the abbreviation for the Joint School of Nanoscience and Nanoengineering, on a bitmap. This bitmap is composed of 32 DMOS DNA type registers and arranged in 512 pixels. After about 20k sequencing reads, they were able to recover >98% of the logo bit.[5]

Figure 10. Design of DMOS DNA type using domain registers as a bit coding system.[5]



Note. Figure 10 demonstrates the visual representation of 32 DMOS DNA type registers decode the JSNN logo (Joint School of Nanoscience and Nanoengineering).[5]

In this thesis, we wanted to further improve the data capacity of these systems by determining whether we could use gRNAs that instead recognized the opposite strand of each

DNA domain, which is not obvious given the mechanism of CRISPR recognition of DNA: no mutation on either strand would represent a digit ‘0’, when only one strand is targeted and mutated the top strand individually, it will represent a digit ‘1’. If the bottom strand is mutated individually then that would represent a ‘2’, and if both top and bottom were mutated simultaneously ‘3’. We attempt this in Chapter 2. Being able to selectively mutate both the top and bottom strands of the DMOS domains also could enable the ability to write, erase, and re-write data on the DMOS tapes with a combination of both cytosine base editors and adenine base editors. This is not something which is not easily able to be done with DNA data storage using synthesis on demand. We show our progress with this work in Chapter 3. In Chapter 4, being able to simultaneously introduce cytosine and adenine base editors would help to further increase data storage capacity, but to do so at the same time requires engineering fusion dCas9 proteins to restrict the deaminases to the specific domain of interest. Overall, the work of this thesis has helped to advance the DMOS project by improving data storage capacity on DNA molecules without requiring on-demand synthesis.

CHAPTER II: INTRODUCING MUTATIONS ONTO THE TOP AND BOTTOM STRANDS OF DNA FOR INCREASED MUTATIONAL DATA STORAGE

Aim

In the DMOS work, 16 different protospacers were mutated to write digital information onto DNA molecules.[5] The goal of this work is to determine whether the capacity for digital data storage could be increased on those DNA molecules if either none (‘0’), the top strand (C > T mutations, ‘1’), the bottom strand (G > A mutations, ‘2’), or both strands (C >T and G > A mutations, ‘3’) are mutated. If a second CRISPR RNP with a different gRNA that could target the bottom strands of the protospacers could be identified, we call protospacers with this property “Janus” protospacers. The “Janus” protospacer were designed to have PAM motifs at both ends of the protospacers so that R-loops can be initiated at either on their respective strand. R-loops are formed when an RNA molecule begins to displace one strand of a DNA molecule by base-pairing with the nucleotides on the other strand. Because these strand invasions are directional and initiate at one end near their PAM, it is not obvious that a protospacer that is receptive for R-loop formation with a gRNA that base-pairs with one strand would also be receptive to R-loop formation to a different gRNA that base-pairs with the other.

In these experiments, the sgRNAs that were designed and synthesized to determine whether the same protospacers used in the DMOS project, now modified with an additional PAM sequence on the other strand, could be targeted and mutated on their bottom strands. If they were able to be targeted, bound on both top and bottom strands, exposed on both top and bottom strands for base editing mutations with APOBEC3A (A3A), then this would in principle imply that we can increase DNA data from 1 bit per protospacer to 2 bits in a quaternary coding scheme. The first part of the experiment is to verify if the 16 sgRNA targeting the bottom strands of the 16 protospacers used in the DMOS project will bind to and cut at the corresponding locations of the protospacer regions with an active Cas9, a signal that R-loop formation is strong; while the second part is composed of testing the selected best dCas9 RNPs with APOBEC3A to see if they will also introduce the cytosine deaminase mutations within the protospacers on the other strand, generating an apparent G > A mutation on the top strand. These experiments allowed us to test the possibility of increasing what would be 16 bits in a binary coding system with 16 protospacers to selected 18 bits in a quaternary coding system with 9 domains found to be good targets for active Cas9 and with TC motifs on both strands.

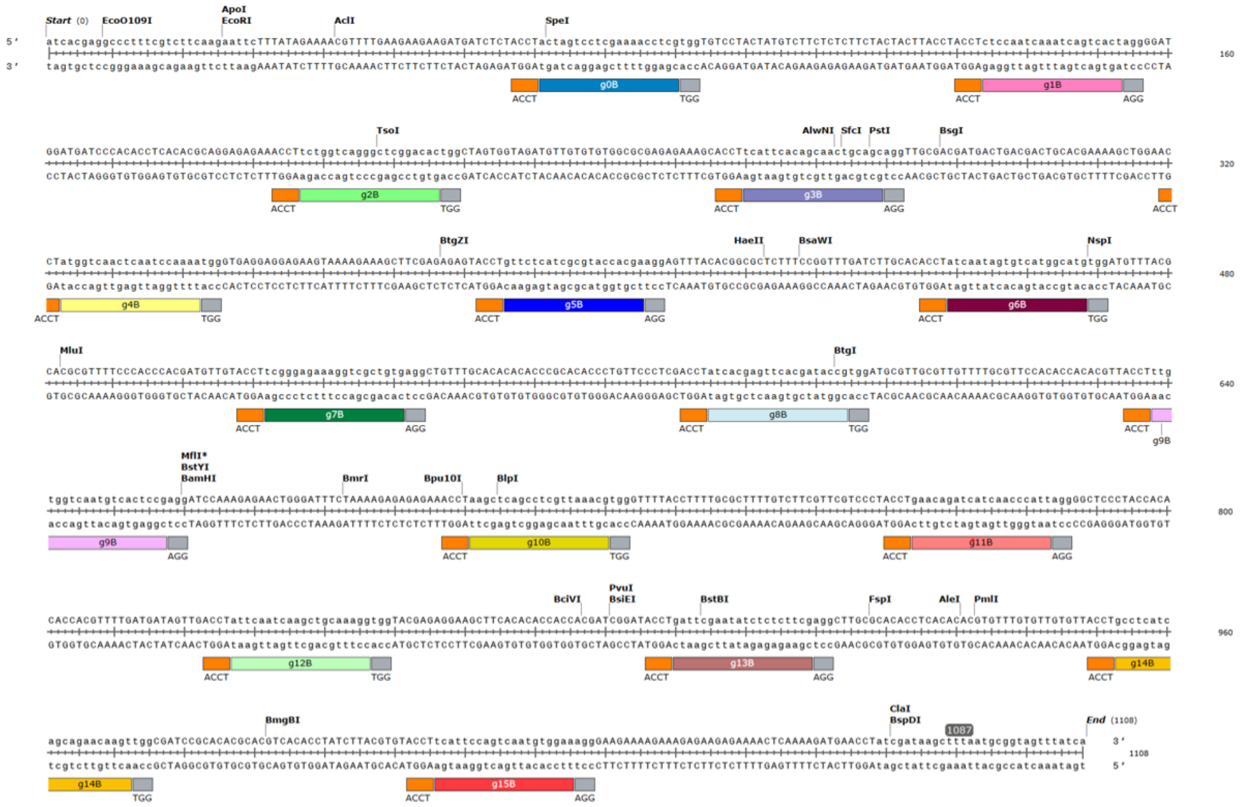
Methods

Synthesis of DMOS & Janus DMOS Template DNA and Purification

Design and Synthesis of the DMOS & Janus DMOS template

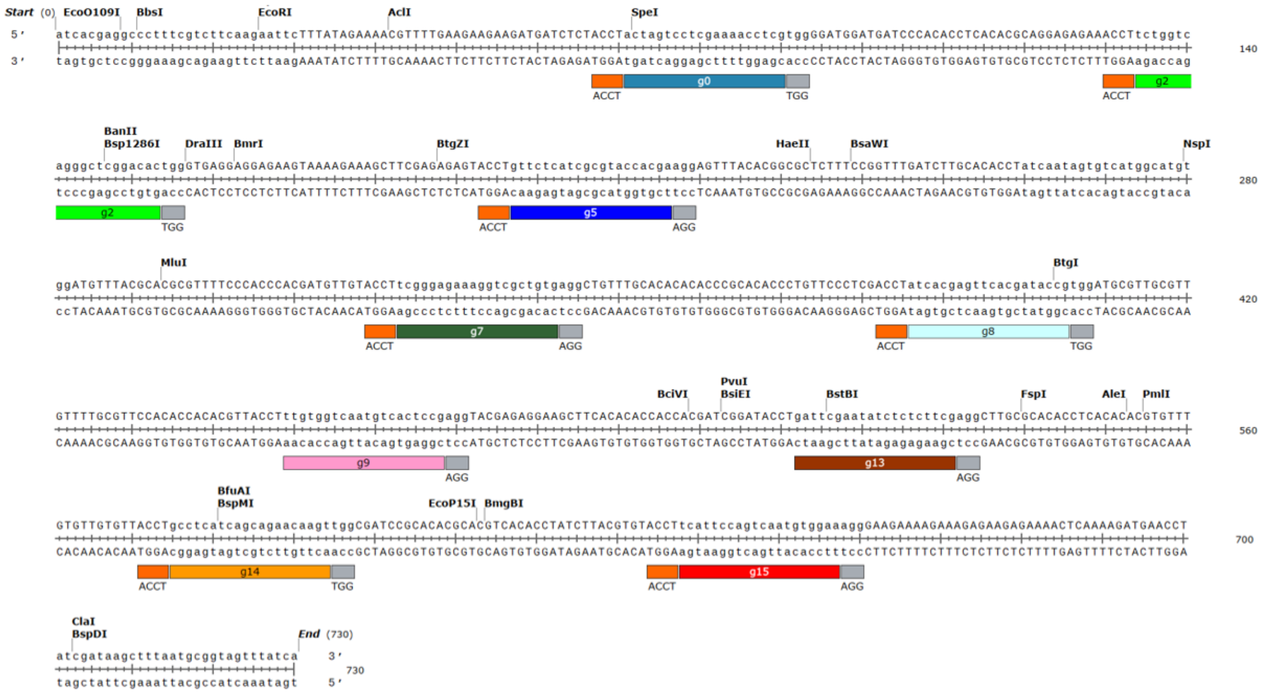
The creation of our new templates was based on the same one that was used for the DMOS project.[6] The differences between their DNA sequence and our designed DMOS DNA and Janus DNA is that we change the ACCT sequence on the 5'- of each protospacer so that it would add a PAM motif of NGG on the 3'- of the bottom strand. The ACCT was modified to ACTC in the DMOS DNA. Our templates, therefore, have a NGG of the 3’- on the bottom strand of the protospacer and an NGG on the 3’- end of the top strand of the protospacer; this modification is shown on Figures 11 and 12.

Figure 11. Schematic View of DMOS DNA Template and Protospacer domains (1108bp)



Note. Figure 11 is also a demonstration of the location of the protospacer domain sequences but in our DMOS DNA template (1108 bp) where the band cut sizes expected using synthesized Cas9-RNPs with sgRNA g0B 1015 bp/ 70 bp, g1B 952 bp/ 133 bp, g2B 889 bp/ 196 bp, g3B 826 bp/ 259 bp, g4B 763 bp/ 322 bp, g5B 700 bp/ 385 bp, g6B 637/448 bp, g7B 574 bp/ 511 bp, g8B 574 bp/ 511 bp, g9B 637/448 bp, g10B 700 bp/ 385bp, g11B 763 bp/ 322 bp, g12B 826 bp/ 259 bp, g13B 889 bp/ 196 bp, g14B 952 bp/ 133 bp, and g15B 1015 bp/ 70 bp.

Figure 12. Schematic View of Janus DMOS DNA Template and Protospacer domains (730bp)



Note. Figure 12 is also a demonstration of the location of the protospacer domain sequences but in our Janus DMOS DNA template (730 bp) where band cut sizes are expected, while using synthesized Cas9-RNPs with sgRNA g0B 640 bp/70 bp, g2B 577 bp/133 bp, g5B 514 bp/196 bp, g7B 388 bp/322 bp, g8B 385 bp/325 bp, g9B 448 bp/262 bp, g13B 511 bp/199 bp, g14B 574 bp/ 136 bp, and g15B 637 bp/ 73 bp.

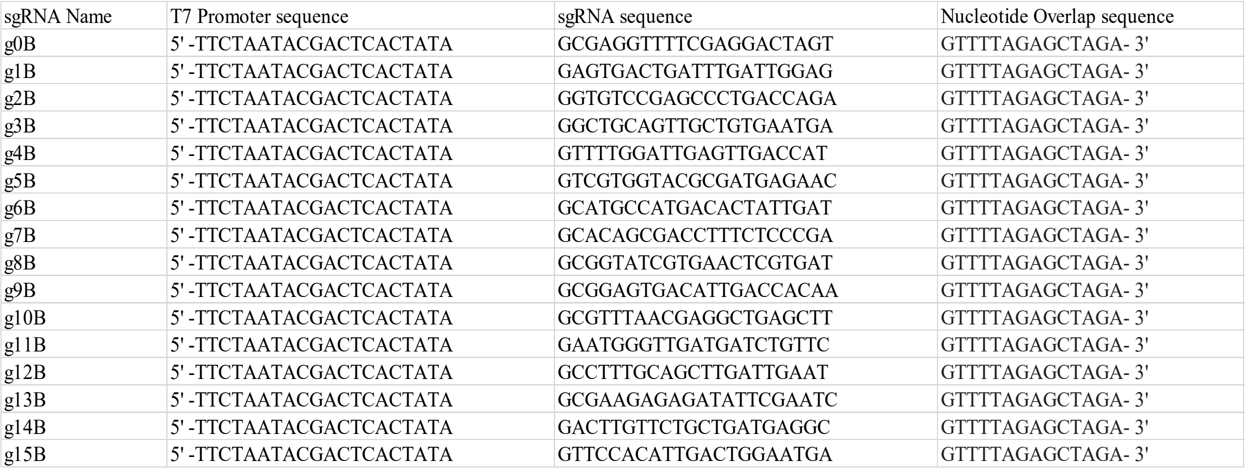
PCR Amplification of DMOS DNA

The DMOS Template DNA is a 1108 base-pairs long double-strand DNA. The DNA template was ordered through Twist Bioscience. Our DNA is similar to the DMOS DNA used in another experiment, the difference is that the template DNA has a PAM (NGG) motif on both top and bottom directions of the protospacer sequences. In this template we have 16 protospacer sequences. DNA arrived lyophilized and was eluted to a 1 ng/µl concentration with Nuclease Free water. Our DMOS DNA template was also amplified using the One Taq 2x Master Mix for PCR and followed the protocol; the annealing temperature for our DMOS DNA was 51˚C for 1 minute because the primers utilize. The extension settings were 68˚C for 1 minute and 7 seconds. The DNA was then purified using AMPure XP magnetic beads and concentration was recorded and measured using the nanodrop.

PCR Amplification of Janus DMOS DNA

The Janus DMOS template is a 730 base pair long double strand DNA, where compared to the DMOS template, out of the 16 guides used to target the 16 protospacers sequences, only nine sgRNAs were chosen as best sgRNAs from active Cas9 digestion experiments and the presence of TCR (R being a purine, A or G) motifs on the bottom strand. The DNA template was ordered through Twist Bioscience as well. The lyophilized DNA was eluted to a 1 ng/µl concentration with nuclease-free water. The Janus DMOS DNA template was also amplified using the One Taq 2x Master Mix for PCR and followed the same protocol; the annealing temperature for the Janus DMOS DNA was also 51˚C for 1 minute because all synthesized template DNA used in these experiments used the same forward and reverse primers for PCR. The extension settings were 68˚C for 44 seconds. The DNA was purified using AMPure XP magnetic beads just as all DNA templates were and concentration was recorded and measured using the nanodrop.

Table 1. Oligos for sgRNA synthesis for targeting protospacers in DMOS and Janus DMOS DNA template



sgRNA Name	T7 Promoter sequence	sgRNA sequence	Nucleotide Overlap sequence
g0B	5' -TTCTAATACGACTCACTATA	GCGAGGTTTTCGAGGACTAGT	GTTTTAGAGCTAGA- 3'
g1B	5' -TTCTAATACGACTCACTATA	GAGTGACTGATTTGATTGGAG	GTTTTAGAGCTAGA- 3'
g2B	5' -TTCTAATACGACTCACTATA	GGTGTCCGAGCCCTGACCAGA	GTTTTAGAGCTAGA- 3'
g3B	5' -TTCTAATACGACTCACTATA	GGCTGCAGTTGCTGTGAATGA	GTTTTAGAGCTAGA- 3'
g4B	5' -TTCTAATACGACTCACTATA	GTTTTGGATTGAGTTGACCAT	GTTTTAGAGCTAGA- 3'
g5B	5' -TTCTAATACGACTCACTATA	GTCGTGGTACGCGATGAGAAC	GTTTTAGAGCTAGA- 3'
g6B	5' -TTCTAATACGACTCACTATA	GCATGCCATGACACTATTGAT	GTTTTAGAGCTAGA- 3'
g7B	5' -TTCTAATACGACTCACTATA	GCACAGCGACCTTTCTCCCGA	GTTTTAGAGCTAGA- 3'
g8B	5' -TTCTAATACGACTCACTATA	GCGGTATCGTGAACTCGTGAT	GTTTTAGAGCTAGA- 3'
g9B	5' -TTCTAATACGACTCACTATA	GCGGAGTGACATTGACCACAA	GTTTTAGAGCTAGA- 3'
g10B	5' -TTCTAATACGACTCACTATA	GCGTTTAACGAGGCTGAGCTT	GTTTTAGAGCTAGA- 3'
g11B	5' -TTCTAATACGACTCACTATA	GAATGGGTTGATGATCTGTTC	GTTTTAGAGCTAGA- 3'
g12B	5' -TTCTAATACGACTCACTATA	GCCTTTGCAGCTTGATTGAAT	GTTTTAGAGCTAGA- 3'
g13B	5' -TTCTAATACGACTCACTATA	GCGAAGAGAGATATTCGAATC	GTTTTAGAGCTAGA- 3'
g14B	5' -TTCTAATACGACTCACTATA	GACTTGTTCTGCTGATGAGGC	GTTTTAGAGCTAGA- 3'
g15B	5' -TTCTAATACGACTCACTATA	GTTCCACATTGACTGGAATGA	GTTTTAGAGCTAGA- 3'

Note. Table 1 is a visual representation that demonstrates the full complete sequences of the DNA templates used to transcribe all 16 sgRNAs that were synthesized and used for the following experiments with our DMOS and Janus DMOS DNA template.

Single guide- RNAs synthesis and Purification

Synthesis of sgRNAs to expose BOTTOM strand of DMOS and Janus DMOS DNA templates

The oligos for g0B – g15B are a series of 16 guides, demonstrated on Table 1, were also designed and synthesized using the *EnGen® sgRNA Synthesis Kit, S. pyogenes* purchased from the New England Biolabs. They were also assembled at room temperature in an RNAse free zone hood. All 16 reactions were then incubated at 37˚C for 30 minutes in a thermocycler and then increased to a total volume of 50 µl by adding 30 µl of Nuclease free water and adding 2 µl DNase I treatment to remove any DNA, NTPs, or dNTP impurities. The 16 reactions were then incubated once again at 37˚C for an additional 15 minutes. The synthesized 16 sgRNAs were then purified using RNAClean XP magnetic beads protocol. All synthesized sgRNA concentrations were measured and recorded using the nanodrop.

Creating RNP Complexes with Cas9 and dCas9

Synthesis of RNP Complexes

All RNPs created for each experiment followed the same IDT Alt-R CRISPR-Cas9 system protocol for RNP complex creations. All RNP reactions were set to 50µl reactions, which is half of the total volume according to the protocol and were mixed with 44.2 µl of 1x PBS buffer solution and 0.8µl of 62 µM stock of Alt-R *S.p* Cas9 enzyme, and 5µl of 10µM Alt-R guide RNA. Depending on the actual concentrations of the synthesized gRNA, conversion may need to be calculated to make sure that everything is in equimolar amounts. The reactions incubate at room temperature for 10 minutes and right afterwards, 1 µl of RNase Murine Inhibitor was mixed into each RNP reaction.

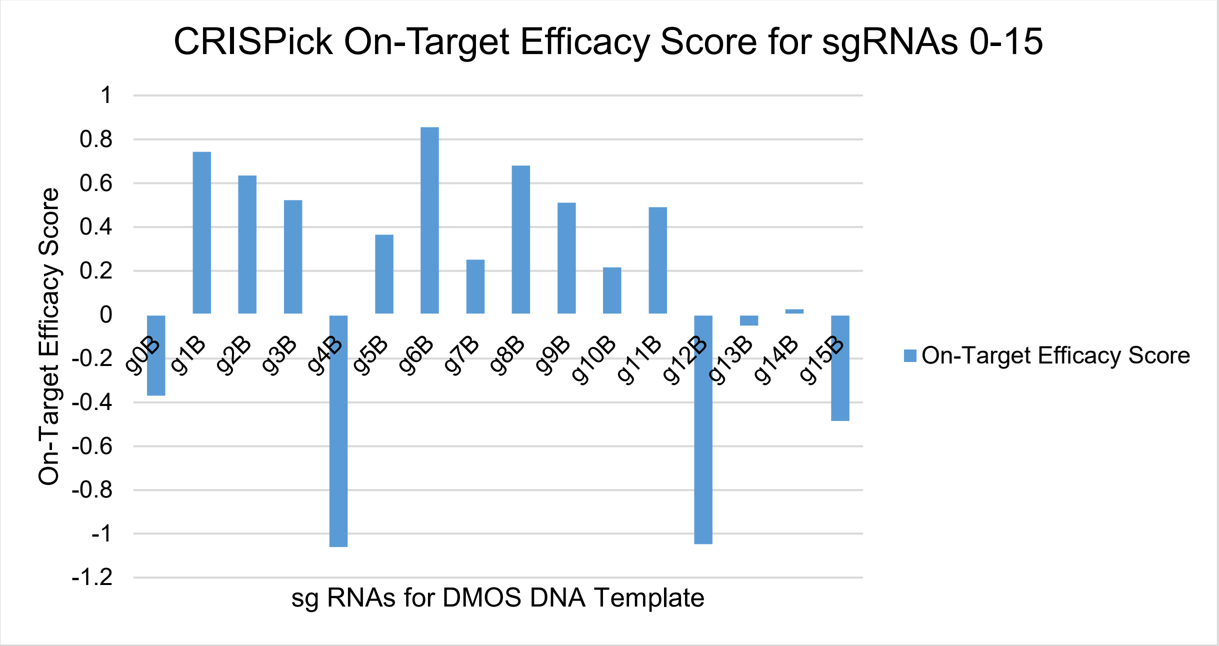
For the Janus DNA, we decided to pick out the selected 9 sgRNAs (0, 2, 5, 7, 8, 9, 13, 14, 15) that were prepared for RNP complex formations, but these RNPs will be formed using 62 µM stock of Alt-R *S.p* dCas9 enzyme. The RNPs created with dCas9 followed the same IDT Alt-R CRISPR-Cas9 system protocol that was mentioned for the RNP complex creations with Cas9. All RNP reactions were set to half of the total volume according to the protocol (50µl) and were mixed with 1x PBS buffer solution (44.2µl), and Alt-R *S.p* dCas9 enzyme (0.8 µl). The reactions were left to incubate at room temperature for 10 minutes and right afterwards, 1 µl of RNase Murine Inhibitor was mixed into each RNP reaction.

in silico Analysis of Predicted Cas9 Targeting efficiency for DMOS DNA[28,29]

Analysis of predicted Cas9 Targeting efficient for DMOS DNA to develop Janus DNA

In Figure 13, a bar graph that displays an analytical Cas9 cleavage efficacy, scores for each of the targeted sequences on the bottom strand. These scores were provided through the usage of the CRISPick website tool.[28,29] This tool allows one to insert their desired DNA template and it will generate all the possible outcome for sgRNAs and score each of them with an efficacy score. A positive score means they have a higher on-target activity while a negative would represent off-target activity. Once the CRISPick tool analyzed all the sgRNA outcomes for our DMOS DNA template, we were able to analyze the predicted efficacy score for the 16 sgRNAs we designed and used for our DMOS experiments. From the in silico analysis prediction we were able to concluded that out of the 16 designed guided RNAs, only 5 out of 16 had a negative score while 11 had a positive on target score. We used these predictions and our in vitro digestion, to design our Janus DMOS template.

Figure 13. Predicted On-Target Efficacy score for 16 sgRNA targeting protospacer domains[27,28]



Note. Figure 13 is a bar graph figure displaying the analytic review of the On-Target efficacy score given to each of the guides that were designed for our DMOS DNA template, using the CRISPick website tool.[27,28]

In vitro Cleavage digestion reaction

Performing an In vitro cleavage digestion reaction with DMOS and Janus DMOS DNA templates

Also following the IDT Alt-R CRISPR-Cas9 System protocol is the *in vitro* digestion reactions, these reactions were assembled at room temperature mixed with 10x Cas9 Nuclease Reaction Buffer that had a pH of 6.8, all RNPs were diluted to a 1µM concentration, and there were 16 reactions combined with 100 nM DNA substrate of our DMOS DNA template and there were 9 reactions set up and mixed with our Janus DMOS DNA template. These 9 reactions were selected from the original 16 RNPs as the best guides for the next experiment called Base Mutation Reactions.

DNA data storage / CRISPR mutation reaction

Base editing reaction protocol used with the selected 9 guides and Janus DMOS DNA template

The following guide RNAs were selected g0B, g2B, g5B, g7B, g8B, g8B, g13B, g14B, and g15B, along with their corresponding protospacers sequences were selected and are what the optimized Janus DNA is composed of. The gRNAs were selected as the best option for base mutation reactions, because they were sgRNAs that contained 5' – TC – 3', 3'- AG - 5', 5' – GA – 3', and 3'- CT- 5' motifs that were 8 nucleotides away from the PAM sequence in both directions of the R-loop. In this base editing reaction, using our Janus DMOS DNA template, we conducted 4 different reactions that were mixed with APOBEC 3A (A3A). We decided to use A3A as the base editor protein for the Em-Seq mutation reaction because it can recognize TC motifs in the Janus DMOS DNA and can conduct cytosine deamination. The 4 reactions were: #1 contains APOBEC3A and a combination of 1.5µl of all dCas9 RNPs developed for the Janus DMOS DNA template (g0B, g2B, g5B, g7B, g8B, g8B, g13B, g14B, and g15B), #2 contained APOBEC3A and a combination of the corresponding dCas9 RNPs that were used in the introductory DMOS project (g1, g3, g8, g9, g10, g14, g15, and g16) also 1.5µl each. For reactions # 3, we mixed 0.75µl of all 9 dCas9 RNPs used for the Janus DMOS DNA and all of the 9 corresponding dCas9 used in the original DMOS project. For reaction #4, we did not use any dCas9 RNPs it just had the template DNA, and it was set as our control experiment. All the reactions followed the adaptive Em-Seq mutation protocol from the original DMOS project and all reactions were in equimolar amounts. Each reaction had a second reaction as a trial, therefore, 8 reactions total. All 8 reactions were then purified using the AMPure XP magnetic beads and the concentrations were measured and recorded with the nanodrop instrument.

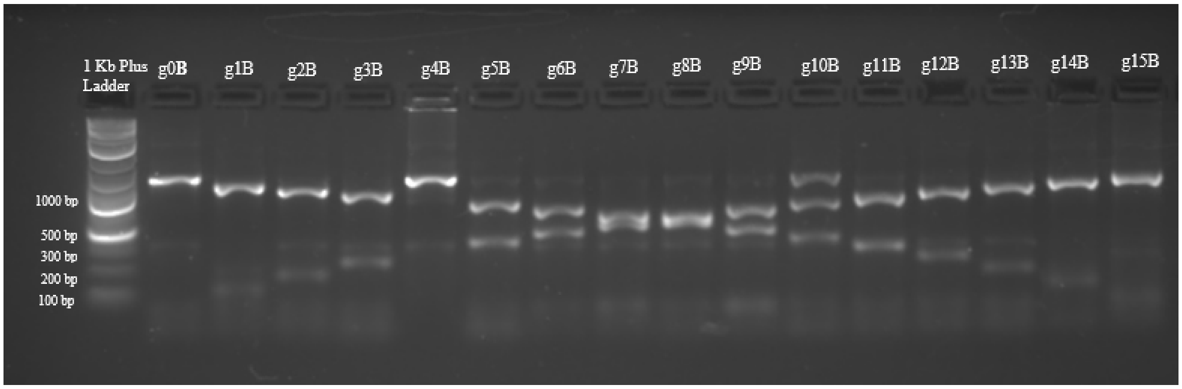
After the base editing reaction, all 8 purified reactions were amplified using the Q5U 2x Master Mix for PCR. For the PCR the same primers that were used to amplify DNA template #4 was used for this PCR reactions. The annealing temperature was set at 66˚C for 20 seconds and the extension temperature was set at 72˚C for 22 seconds. The procedure allows for us to amplify the DNA template that contains the mutational base changes from cytosine to thymine on both top and bottom strands of the targeted protospacers. Once PCR is complete, it is then purified again using AMPure XP magnetic beads sent for Sanger Sequencing to verify whether there were any C >T mutations or G > A mutations in the selected 9 protospacer regions.

Results and Discussion

In Figure 11, we are shown a visual representation of the template DNA and the corresponding 16 protospacer domains within. We performed an in silico analysis to predict protospacers where there might be high levels of Cas9 if the gRNA targeted the bottom strand. Estimates from the CRISPick predictor for Cas9 activity at a specific protospacer predicted, resulted in that only 10 out of the 16 sgRNAs designed had a positive efficacy score of over 0.2, ranging from 0.2 – 0.85 (g1B, g2B, g3B, g5B, g6B, g7B, g8B, g9B, g10B, and g11B); while 4 out of the 16 had a negative efficacy score ranging from -0.2 to approximately -1.1. Only 2 out of the 16, had an approximate 0 efficacy score (g13=-0.05B, g14B=0.025). Despite the CRISPick results we still decided to test all possible sgRNA that had the 5’ – TC – 3’, 3’- AG -

5’, 5’ – GA – 3’, and 3’- CT- 5’ motifs that were 8 nucleotides away from the PAM sequence in both directions of the R-loop. Those gRNAs were g0B, g2B, g5B, g7B, g8B, g8B, g13B, g14B, and g15B. We also tested 16 domains on the bottom strand using active RNP -Cas9 complexes as a proxy to test for robust R-loop production. It is possible that in the protospacers developed, some protospacer will have regions for possible Cytosine deaminase base pair mutation that are 8 bp at least away from both the distal PAM motifs. Therefore, in our results, shown in figure 14, the 2 % agarose gel electrophoresis displays that all the 16 protospacer domains were targeted by the sgRNA that were designed and cut into two DNA bands by the RNP-Cas9 complexes, and Figures 11 and 12, we decided that the best sgRNAs to use for target the template DMOS DNA were guides 0B, 2B, 5B, 7B, 8B, 9B, 13B, 14B, and 15B. These sgRNAs helped us design the next template DNA, which is the Janus DMOS DNA and is shown on Figure 13.

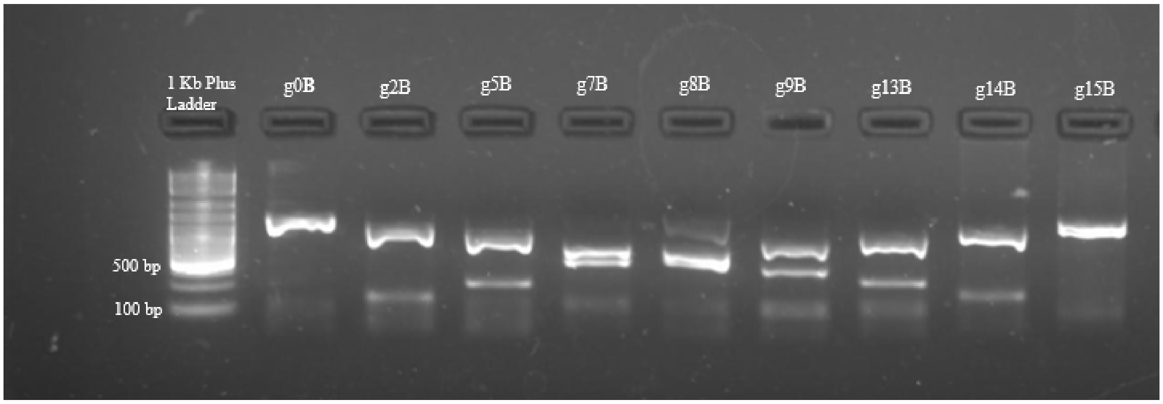
Figure 14. ***In vitro*** **Digestion of DMOS DNA template (1108bp)**



Note. Figure 14 is visual demonstration of the in vitro digestion reactions with DMOS DNA template in a gel electrophoresis. This is a 2% agarose gel with 5 µl Syber gold, that ran at 110 V for 55 minutes using 1 Kb Plus as the ladder marker.

In Figure 15, we were also able to demonstrate that the specific chosen 9 sgRNAs with active Cas9 were able to bind to and cut the top strand of the protospacers while exposing the bottom strand for potential cytosine deamination. The next experiment conducted was based off an adaptation of the NEBNext® Enzymatic Methyl-seq Kit, in specific the 1.8 Deamination of Cytosines protocol on the New England Biolabs website. All 9 RNP-dCas9 complexes used in this chapter and another 9 RNP-dCas9 complexes, which are based on the same protospacer sequences as DNA DMOS and Janus DMOS template and their corresponding guides, were used to test out 8 base mutational reactions using the EM-Seq protocol. The 9 RNP-dCas9 complexes were already shown and proved able to target and bind to the bottom strand exposing the top for cytosine deamination. We expected for the base mutational reaction to work and target the top strand of the protospacer bind to the top DNA strand of the protospacer and expose the bottom. This would imply that DNA as a digital storage source if cytosine base editors are used in a base mutation reaction and these sgRNAs can be reused to target both top and bottom simultaneously then this could be a representation of a digit # 3 in a ternary DNA coding system, which will allow us to further expand the possibilities of bits per DNA domain, meaning in a binary system 9 domains would represent 18 bits but with being able to mechanistically target both top and bottom strands of the engineered JANUS protospacers with a distal PAM motif on both ends, the domains can be either non mutated= 0, mutated top only = 1, mutate bottom only= 2, mutate top and bottom= 3. This expands the DNA digital capacity of the 9 domains to almost 18 bits. After we analyzed the sequencing results obtain from the base mutation reaction, it was reported inconclusive due to our control RNPs, which had previously demonstrated to conduct cytosine mutations, not having "significant" observed mutational signatures. Therefore, we cannot conclude whether this experiment worked or not. We will continue to test these outcomes for future works.

Figure 15. I*n vitro* Digestion of Janus DMOS DNA template (730 bp)



Note. Figure 15 is visual demonstration of the in vitro digestion reactions with Janus DMOS DNA in a gel electrophoresis. This is a 2% agarose gel with 5 µl Syber gold, that ran at 100 V for 55 minutes using 1 Kb Plus as the ladder marker.

Through this experiment we were able to develop and design protospacers along with their corresponding sgRNAs, so that they can target the protospacers regions containing T-C motifs for base mutational experiments. By conducting the base mutation reactions, we expected to see cytosine mutations in each of the 9 protospacers of the Janus DMOS DNA template. This allows us to increase the data capacity from 9 bits to potential 18 bits.

CHAPTER III: TARGETING TOP AND BOTTOM DNA STRANDS USING SYNTHESIZED SGRNAS FOR READING, WRITING, AND ERASING

Aim

The purpose of the following experiment was to be able to design and develop sgRNA that would target specific protospacer sequences in the template DNA, that have a triple cytosine sequence in the middle of the top DNA strand of the protospacer. (5' – CCC – 3'). Three different sgRNA named g1T, g2T, and g3T are specific sgRNA that would, theoretically, allow us to expose the top strand of the DNA to a cytosine base editor such as APOBEC3G which strongly recognizes the CCC motif, when it binds to the bottom DNA strand of the protospacer by being complementary to the bottom. In other words, if we first target the top sequence to mutate it from C > T it would produce a stretch of AAA on the bottom strand, and if we then used an adenine base editor, we could mutate them back (through A > G mutation) to CCC. This would allow writing, reading, and erasing the same DNA "tape" molecule. We also tested the template DNA with sgRNAs g1B, g2B, and g3B, the difference being they are complementary to the top DNA strand of the specific protospacers. This experiment included conducting and testing both top and bottom guides with an *in vitro* digestion, in which the Cas 9 enzyme in combination with the sgRNA creates a ribonucleotide protein (RNP) and cuts the DNA at the specific targeted protospacer sequence. This would allow us to verify whether the sgRNA designed in an RNP complex, would target and recognize the specific protospacer sequence, located in 3 different areas of the template DNA, making cuts in the DNA that could be differentiated by DNA fragment sizes when viewed on a gel electrophoresis. The design of these template DNA's containing the targeted protospacers and the development of the sgRNA's allowed us to test the possibility of writing, erasing, and potentially rewriting DNA, which in other terms can extend the possibility of increasing DNA data storage capacity.

Methods

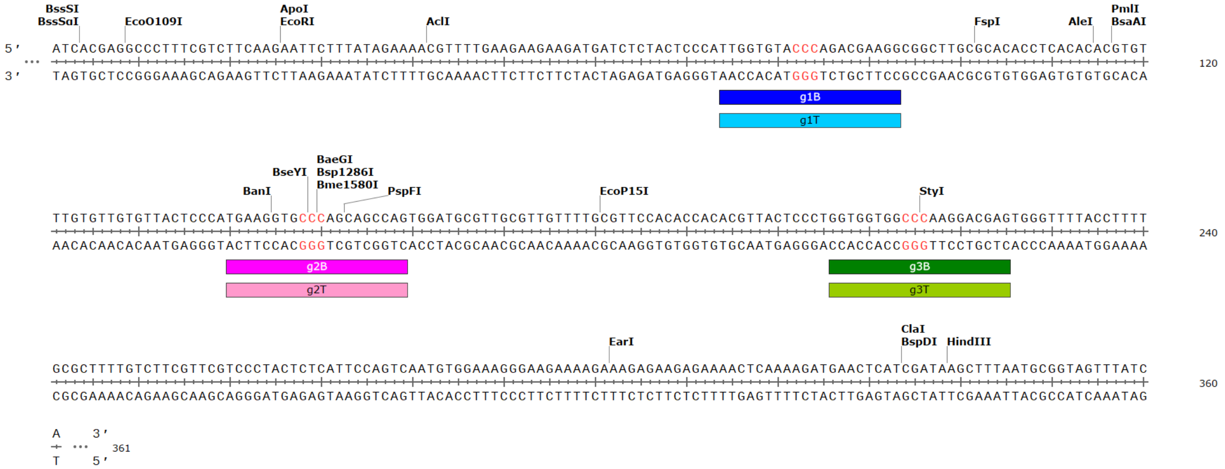
Synthesis of Janus Tri-C1 & Janus Tri-C2 DNA template, Amplification, and Purification

PCR Amplification of Janus Tri-C1 & Janus Tri-C 2

DNA Template Janus Tri-C1 is a 361 base pair long double strand DNA. The DNA template was ordered through IDT as a gBlocks gene fragment to allow for design flexibility in developing the protospacer sequences that are complementary to either sgRNA for top or bottom strand. The DNA came lyophilized and was eluted to a 1 ng/µl concentration with Nuclease Free water. DNA Janus Tri-C1 was amplified using the One Taq 2x Master Mix for PCR and followed the protocol; the annealing temperature was 51˚C for 1 minute and the extension settings were 68˚C for 22 seconds. The DNA was then purified using AMPure XP magnetic beads and concentration was recorded and measured using the nanodrop.

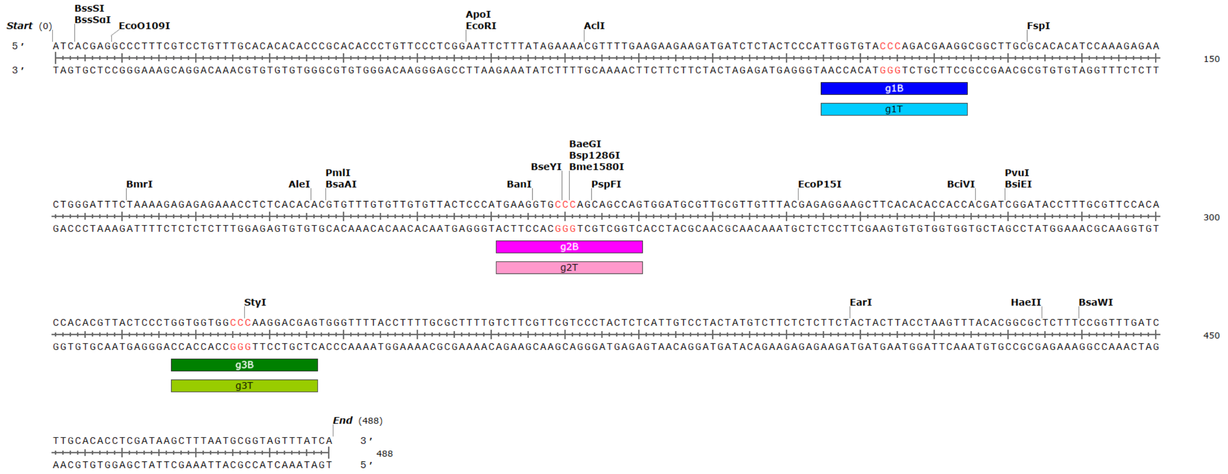
DNA Template Janus Tri-C2 is a 488 base pair long double strand DNA. The DNA template was ordered through Twist Bioscience as Non-Cloned Gene fragment. DNA Janus Tri-C2 was also designed for the protospacer sequences to have a complementary top and bottom strand for the same synthesized sgRNA that were initially used for DNA Janus Tri-C1. The DNA came lyophilized and was eluted to a 1 ng/µl concentration with Nuclease Free water. The DNA Janus TriC2 was also amplified using the One Taq 2x Master Mix for PCR and followed the protocol; the annealing temperature for Janus Tri-C2 was also 51˚C for 1 minute because both DNA Janus Tri-C1 and Janus Tri-C2 used the same forward and reverse primers for PCR. The extension settings were 68˚C for 29 seconds. The DNA was then also purified using AMPure XP magnetic beads and concentration was recorded and measured using the nanodrop.

Figure 16. Schematic View of Janus Tri-C1 and protospacer domains (361 bp)


BssSI
BssSαI
EcoO109I
ApoI
EcoRI
AclI
FspI
AleI
PmlI
BsaAI
5'
ATCACGAGGCCCTTTCGTCTTCAAGAATTCTTTATAGAAAACGTTTTGAAGAAGAAGATGATCTCTACTCCCATTGGTGTACCCAGACGAAGGCGGCTTGCGCACACCTCACACACGTGT
3'
TAGTGCTCCGGGAAAGCAGAAGTTCTTAAGAAATATCTTTTGCAAAACTTCTTCTTCTACTAGAGATGAGGGTAACCACATGGGTCTGCTTCCGCCGAACGCGTGTGGAGTGTGTGCACA
120
g1B
g1T
BaeGI
BseYI
Bsp1286I
Bme1580I
BanI
PspFI
EcoP15I
StyI
TTGTGTTGTGTTACTCCCATGAAGGTGCCCAGCAGCCAGTGGATGCGTTGCGTTGTTTTGCGTTCCACACCACACGTTACTCCCTGGTGGTGGCCCAAGGACGAGTGGGTTTTACCTTTT
AACACAACACAATGAGGGTACTTCCACGGGTCGTCGGTCACCTACGCAACGCAACAAAACGCAAGGTGTGGTGTGCAATGAGGGACCACCACCGGGTTCCTGCTCACCCAAAATGGAAAA
240
g2B
g2T
g3B
g3T
ClaI
EarI
BspDI
HindIII
GCGCTTTTGTCTTCGTTCGTCCCTACTCTCATTCCAGTCAATGTGGAAAGGGAAGAAAAGAAAGAGAAGAGAAAACTCAAAAGATGAACTCATCGATAAGCTTTAATGCGGTAGTTTATC
CGCGAAAACAGAAGCAAGCAGGGATGAGAGTAAGGTCAGTTACACCTTTCCCTTCTTTTCTTTCTCTTCTCTTTTGAGTTTTCTACTTGAGTAGCTATTCGAAATTACGCCATCAAATAG
360
A
3'
361
T
5'

Note. Figure 16 is a demonstration of the location of the protospacer domain sequences in DNA Janus Tri-C1 which has a total of 361 base pairs (bp) where band cut sizes expected using synthesized Cas9-RNPs with sgRNA g1T & g1B at 268 bp/ 73 bp, g2T & g2B at 202 bp/ 139 bp, and g3T & g3B at 205 bp/ 136 bp.

Figure 17. Schematic View of Janus Tri-C2 and protospacer domains (488 bp)


Start (0)
BssSI
BssSaI
EcoO109I
ApoI
EcoRI
AclI
FspI
5'
ATCACGAGGCCCTTTCGTCCTGTTTGCACACACACCCGCACACCCTGTTCCCTCGGAATTCTTTATAGAAAACGTTTTGAAGAAGAAGATGATCTCTACTCCCATTGGTGTACCCAGACGAAGGCGGCTTGCGCACACATCCAAAGAGAA
3'
TAGTGCTCCGGGAAAGCAGGACAAACGTGTGTGTGGGCGTGTGGGACAAGGGAGCCTTAAGAAATATCTTTTGCAAAACTTCTTCTTCTACTAGAGATGAGGGTAACCACATGGGTCTGCTTCCGCCGAACGCGTGTGTAGGTTTCTCTT
150
g1B
g1T
BmrI
AleI
PmlI
BsaAI
BanI
BseYI
BaeGI
Bsp1286I
Bme1580I
PspFI
EcoP15I
BciVI
PvuI
BsiEI
CTGGGATTTCTAAAAGAGAGAGAAACCTCTCACACACGTGTTTGTGTTGTGTTACTCCCATGAAGGTGCCCAGCAGCCAGTGGATGCGTTGCGTTGTTTACGAGAGGAAGCTTCACACACCACCACGATCGGATACCTTTGCGTTCCACA
GACCCTAAAGATTTTCTCTCTCTTTGGAGAGTGTGTGCACAAACACAACACAATGAGGGTACTTCCACGGGTCGTCGGTCACCTACGCAACGCAACAAATGCTCTCCTTCGAAGTGTGTGGTGGTGCTAGCCTATGGAAACGCAAGGTGT
300
g2B
g2T
StyI
EarI
HaeII
BsaWI
CCACACGTTACTCCCTGGTGGTGGCCCAAGGACGAGTGGGTTTTACCTTTTGCGCTTTTGTCTTCGTTCGTCCCTACTCTCATTGTCCTACTATGTCTTCTCTCTTCTACTACTTACCTAAGTTTACACGGCGCTCTTTCCGGTTTGATC
GGTGTGCAATGAGGGACCACCACCGGGTTCCTGCTCACCCAAAATGGAAAACGCGAAAACAGAAGCAAGCAGGGATGAGAGTAACAGGATGATACAGAAGAGAGAAGATGATGAATGGATTCAAATGTGCCGCGAGAAAGGCCAAACTAG
450
g3B
g3T
End (488)
TTGCACACCTCGATAAGCTTTAATGCGGTAGTTTATCA
3'
AACGTGTGGAGCTATTCGAAATTACGCCATCAAATAGT
5'
488

Note. Figure 17 is also a demonstration of the location of the protospacer domain sequences but in DNA Janus Tri-C2 (488 bp) where the band cut sizes expected using synthesized Cas9-RNPs with sgRNA g1T & g1B are 364 bp/ 104 bp, g2T & g2B at 258 bp/ 210 bp, and g3T & g3B at 316 bp/ 152 bp.

Single guide- RNAs synthesis and Purification

Synthesis of sgRNAs to expose TOP strand of Template DNA Janus Tri-C1 & Janus Tri-C2

Oligos for g1T, g2T, and g3T, presented in Table 2 were designed and synthesized using the *EnGen® sgRNA Synthesis Kit, S. pyogenes* purchased from the New England Biolabs. They were assembled at room temperature in an RNAse free area. The reaction was then incubated at 37˚C for 30 minutes in a thermocycler and then afterwards the 3 reactions total volume was increased to 50 µl by adding 30 µl of Nuclease free water and adding 2 µl DNase I to remove any DNA template, NTP, and dNTPS impurities, and were incubated once again at 37˚C for an additional 15 minutes. The synthesized sgRNAs were then purified using RNAClean XP magnetic beads protocol. All synthesized sgRNA concentrations were measured using the nanodrop and annotated.

Synthesis of sgRNAs to expose BOTTOM strand of Template DNA Janus Tri-C1 and Janus Tri-C2

The same sgRNA synthesis procedure for oligos for g1T, g2T, and g3T applies for oligos g1B, g2B, and g3B. They were designed and synthesized using the *EnGen® sgRNA Synthesis Kit, S. pyogenes* purchased from the New England Biolabs, assembled at room temperature in an RNAse free zone hood, and all the synthesized sgRNA concentrations were measured and recorded using the nanodrop.

Table 2. Oligos for sgRNA synthesis for targeting protospacer in Template Janus Tri-C1 and Janus Tri-C2

sgRNA Name	T7 Promoter sequence	sgRNA sequence	Nucleotide Overlap sequence
g1-T	5' -TTCTAATACGACTCACTATA	GTTGGTGTACCCAGACGAAGG	GTTTTAGAGCTAGA- 3'
g1-B	5' -TTCTAATACGACTCACTATA	GCCTTCGTCTGGGTACACCAA	GTTTTAGAGCTAGA- 3'
g2-T	5' -TTCTAATACGACTCACTATA	GTGAAGGTGCCCAGCAGCCAG	GTTTTAGAGCTAGA- 3'
g2-B	5' -TTCTAATACGACTCACTATA	GCTGGCTGCTGGGCACCTTCA	GTTTTAGAGCTAGA- 3'
g3-T	5' -TTCTAATACGACTCACTATA	GGGTGGTGGCCCAAGGACGAG	GTTTTAGAGCTAGA- 3'
g3-B	5' -TTCTAATACGACTCACTATA	GCTCGTCCTTGGGCCACCACC	GTTTTAGAGCTAGA- 3'

Note. Table 2 demonstrates the full complete sequences of all 6 sgRNAs that were synthesized and used for the following experiments using templates DNA Janus Tri-C1 and Janus Tri-C2. In red you can see the regions of CCC and GGG motifs, that will target and bind with their corresponding complementary strand.

Creating RNP Complexes with Cas9

Synthesis of RNP Complex utilizing Purified sgRNA Targeting TOP DNA strand

RNP complexes that would target and ideally leave the top strands of the protospacers exposed for DNA templates Janus Tri-C1 and Janus Tri-C2, where created using the IDT Alt-R CRISPR-Cas9 system protocol for RNP complex creations. We followed the same protocol as for when we developed the RNPs for our DMOS template and Janus DNA template. The guides used to make the RNPs for targeting and exposing top strand were sgRNAs g1T, g2T, and g3T. The protocol set for a total reaction volume of 100 µl, in our case we did half of the total volume for the RNP reactions, making sure everything was still in the same equimolar amounts. The sgRNAs were prepared to be equivalent to a 10µM concentration each guide RNA and each reaction was pipetted with 1x PBS buffer solution and 62 µM stock of Alt-R *S.p* Cas9 enzyme that was purchased through IDT. The reactions were incubated at room temperature for 10 minutes for the RNP complexes to form and afterwards 1 µl of RNase Murine Inhibitor was mixed into each RNP reaction.

Synthesis of RNP Complex utilizing Purified sgRNA Targeting BOTTOM DNA strand

The RNP complexes that would target and expose the opposite strands as g1T, g2T, and g3T, were sgRNAs g1B, g2B, and g3B; which would bind to the top strand of the protospacer exposing the bottom strand for potential base editing mutations. All RNPs were created using the same IDT Alt-R CRISPR-Cas9 system protocol for RNP complex creations. All RNP reactions were set to half of the total volume according to the protocol. The sgRNAs were all also prepared to be equivalent to a 10µM concentration each guide RNA and each reaction was pipetted with 1x PBS buffer solution and 62 µM stock of Alt-R *S.p* Cas9 enzyme that was purchased through

IDT. The reactions were incubated at room temperature for 10 minutes to maximize the formation of the RNP complexes and afterwards 1 µl of RNase Murine Inhibitor was mixed into each RNP reaction.

***In vitro* Cleavage digestion reaction**

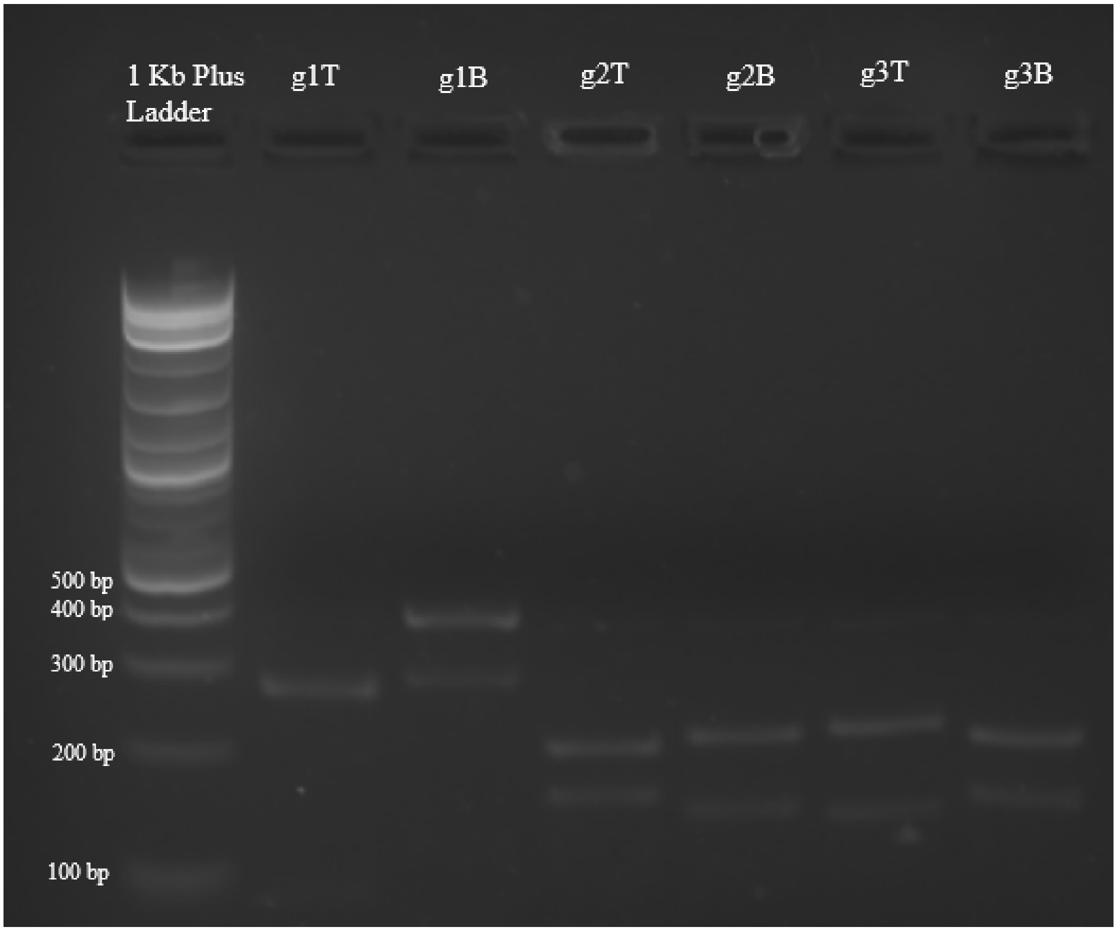
Performing an In vitro cleavage digestion reaction with DNA templates Janus Tri-C1 and Janus Tri-C2

IDT Alt-R CRISPR-Cas9 System protocol also provided one for the in vitro digestion reaction. The in vitro digestion reactions were assembled at room temperature mixed with 10x Cas9 Nuclease Reaction Buffer that had a pH of 6.8, all 6 RNPs were diluted to a 1µM concentration, and there were 6 reactions combined with 100 nM DNA substrate Janus Tri-C1 and another 6 reactions were mixed with DNA template Janus Tri-C2. Each RNP complex was tested with each template DNA. These two templates of DNA served as double trial experiment to verify whether the designed protospacer and the top and bottom complementary designed sgRNAs could target these specific locations. With the *in vitro* digestions we will be able to have a visual representation of whether the sgRNAs were able to target both top and bottom strands of the protospacers and have the Cas9 cut the DNA templates at the designed locations.

Results and Discussion

As shown in Figures 17 and in Figure 18, we are able to visualize the cleavage products of the template DNA Janus Tri-C1 and Janus Tri-C2 and their corresponding protospacer domains for each. These two sets of DNAs were used as trials since for both DNA templates the same 6 samples of RNP -Cas9 complexes were used for both template DNA in an *in vitro* digestion. According to the results obtained in the in vitro digestion reaction experiments shown in figure 18, in a 2 % agarose gel electrophoresis, we can visually see that each of the 3 protospacer domains were targeted by the sgRNA that were designed and cut into two DNA bands by the RNP-Cas9 complexes. These results show that we were able to design specific sgRNAs to target both the top and bottom strand of the protospacers individually, that contain mainly APOBEC3G motifs in the DNA sequence. Therefore, in principle APOBEC3G should be able to recognize them in R-loops, enabling both writing and erasing nucleotide sequences without DNA generation on demand. In this context we can theoretically conclude that if it targets the bottom DNA strand, we can demonstrate the ability to write, erase, and potentially rewrite in DNA. Like in the previous chapter, being able to target the bottom and top can also expand the number of bits that can be encoded in a single domain.

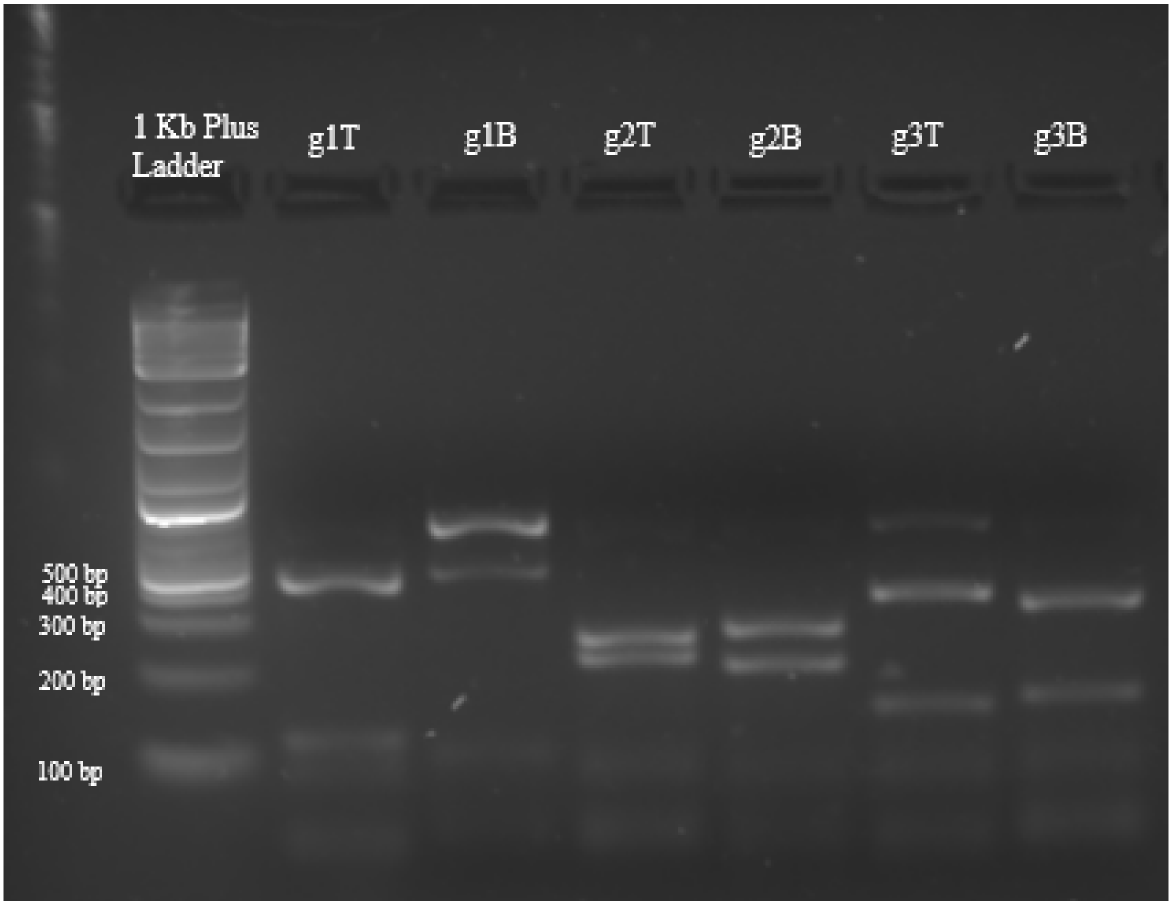
Figure 18. ***In vitro*** **Digestion of Template DNA Janus Tri-C1 (361bp)**



Note. Figure 18 is visual demonstration of the in vitro digestion reactions with DNA template Janus Tri-C1 in a gel electrophoresis. This is a 2% agarose gel with 5 µl Syber gold, that ran at 90 V for 40 minutes using 1 Kb Plus as the ladder marker.

The same is true with regards to Figure 19, where we were able to repeat the experiment using another design Janus Tri-C2 while keeping the protospacers sequences the same, so that the same RNP-Cas9 complexes used for DNA template Janus Tri-C1 could be used for this template DNA Janus Tri-C2. Again, we are visually able to see and confirm that the design protospacers and targeting sgRNA works and conducts a double strand break the specific protospacer locations.

Figure 19. ***In vitro*** **Digestion of Template DNA Janus Tri-C2 (488bp)**



Note. Figure 19 is visual demonstration of the in vitro digestion reactions with DNA template Janus Tri-C2in a gel electrophoresis. This is a 2% agarose gel with 5 µl Syber gold, that ran at 90V for 40 minutes using 1 Kb Plus as the ladder marker.

CHAPTER IV: CLONING BASE EDITOR FUSION PROTEINS FOR INCREASED DATA STORAGE AND FUTURE WORKS

Aim

In previous chapters we sought to increase digital data storage on DNA for DMOS, by targeting both the top and bottom strands of each domain. Because in the previous chapter we designed protospacers that included motifs optimized for APOBEC3G while the DMOS project used APOBEC3A, we realized that if we used fusion proteins, where dCas9 is connected to the deaminase at the gene level and expressed as a single protein, differences in the mutational signatures at those sites could also increase the number of bits stored per domain. In other words, we could target one domain with dCas9-APOBEC3A or dCas9-APOBEC3G and the resulting mutations could be differentiated, that could add an additional bit of information to each domain. With the 9 selected Janus DMOS protospacer, not only could we have domains potentially store a '0' (unmutated), a '1' (top strand mutated), a '2' (bottom strand mutated), or a '3' (both strands mutated), but with the fusion proteins we can further expand it to a '4' (both strands mutated with either base editor fusion protein combinations (A3A/ABE7.10, A3G/ABE7.10, A3A/A3G). Counting every different combinations of CRISPR effector targeting either or both strands (or neither) one at a time or simultaneously in sets of 2, 3 or 4, that would allow 256 different possible mutational signatures, or 8 bits / 1 byte per protospacer. For a 9 domain DNA template, that could potentially store nearly 72 bits. Fusing the proteins together is necessary so that the deaminase would remain close to the dCas9 RNP and not interfere with the mutations occurring at different domains.

We also had the idea that if there were multiple deaminases fused to the same dCas9, that would generate even more diversity in the mutational patterns in each domain that would therefore increase the number of bits per domain. The goal for this chapter was to explore the mutational capabilities if there a cytosine base editor and an adenine base editor were on the same dCas9 complex, and test the several possibilities of the combinations of these fusion base editor proteins. We tried to conduct this experiment by first attempting to clone the different base editor plasmid proteins in to one plasmid. Specifically, we tried to clone one cytosine base editor gene fragment from APOBEC3A (A3A) and one adenine base editor fragment (ABE7.10) into a plasmid with dCas9 and the second fusion protein plasmid would one cytosine base editor gene fragment from APOBEC3G (A3A) and one adenine base editor (ABE7.10) into a plasmid with Cas9. Both APOBEC3A and APOBEC3G are cytosine deaminase proteins but each once targets a different DNA motif as mentioned in the introduction. Therefore, the goal of this experiment is be able to explore if the fusion proteins can target and enable cytosine and adenosine deaminases simultaneously to increase the potential capacity of DNA data storage. We would then clone the fusion plasmid proteins, express them, then later apply them to base mutational reactions experiments to see whether they can alter cause cytosine and adenosine deamination at targeted locations of a Janus protospacers. In further detail, the experimental purpose is if we can successfully clone and express the pre-mix of different fusion proteins (A3A-ABE7.10 and A3G-ABE7.10), and have them both bind to the dCas9 complex in a base mutation reaction, then theoretically, each protospacer would have a different mutational signature and to mutate them all at once, can be used to further increase data storage density.

Methods

Primer design for isolating gene of interest in plasmid and Amplification

We used the NEBuilder Assembly Tool for developing forward and reverse primers to be able to amplified specific section of the plasmids. These primers were created so that each plasmid pCMV-hA3A-BE3 (Addgene #113410 plasmid, 8442 bp) and pCMV-hA3G-BE3 (Addgene # #113415 plasmid, 8997 bp), would be fully amplified except the UGI gene or region, which is about 249 bp. The Uracil Glycosylase Inhibitor (UGI) gene is an inhibitor of the Uracil-DNA glycosylase enzyme (UDG), which eliminates uracil nucleotides in DNA and initiates base-excision repair (BER) pathway. For the purpose of our cloning experiment, it is of no effect if we remove the UGI gene from the plasmids. Primers were also created to isolate the ABE7.10 gene (1191 bp) from the pCMV-ABE7.10 plasmid (Addgene #102919, 8727 bp). We then used the primers developed for each plasmid to amplify the regions of interest using the Q5 2x Master Mix purchased from NEB for PCR reactions. The plasmids, with their corresponding primers, had an annealing temperature of 64˚C for pCMV-hA3A-BE3, 65˚C pCMV-hA3G-BE3, 57˚C for pCMV-ABE7.10 plasmid for the PCR protocol, these reactions were made and prepared to be 25µl reactions. We were then able to purify the PCR products using the Monarch® PCR & DNA Cleanup Kit (5 µg) (NEB #T1030) and we would elute 20 µl of purified plasmid DNA. All plasmids were verified through a 2% agarose gel electrophoresis using SYBR Safe DNA gel stain.

Ligation of genes of interested into plasmid via HiFi DNA Assembly

Once all the desired regions of the genes of interest were amplified, we used the NEBuilder® HiFi DNA Assembly Master Mix protocol to be able to ligate the pCMV-hA3A-

BE3 plasmid with no UGI (A3A plasmid) along with ABE7.10 gene from the pCMV-ABE7.10 plasmid. As well as to conduct the ligation of pCMV-hA3G-BE3 plasmid with no UGI (A3G plasmid) along with ABE7.10 gene from the pCMV-ABE7.10 plasmid.

Bacterial Transformation using 10-Beta Competent Cells and Inoculation

Now that the DNA ligation and the segments were assembled, we then transformed them into 10-Beta Competent E Coli cells, using the High Efficiency Transformation Protocol using NEB® 10-beta Competent E. coli (High Efficiency) (C3019H/C3019I) protocol. We plated the bacterial transformations unto Luria-Bertani (LB) media plates prepared with Carbenicillin, which is an antibiotic and an ampicillin analog, and we let the bacterial plates incubate overnight (approximately 16 hours) at 37°C. After the 16 hours and there was bacteria growth display on the plates as colonies. From each plate, 5 isolated colonies were pick for inoculation (using liquid LB and the same antibiotic). The inoculated colonies were incubated at 37˚C overnight at 200 RPM.

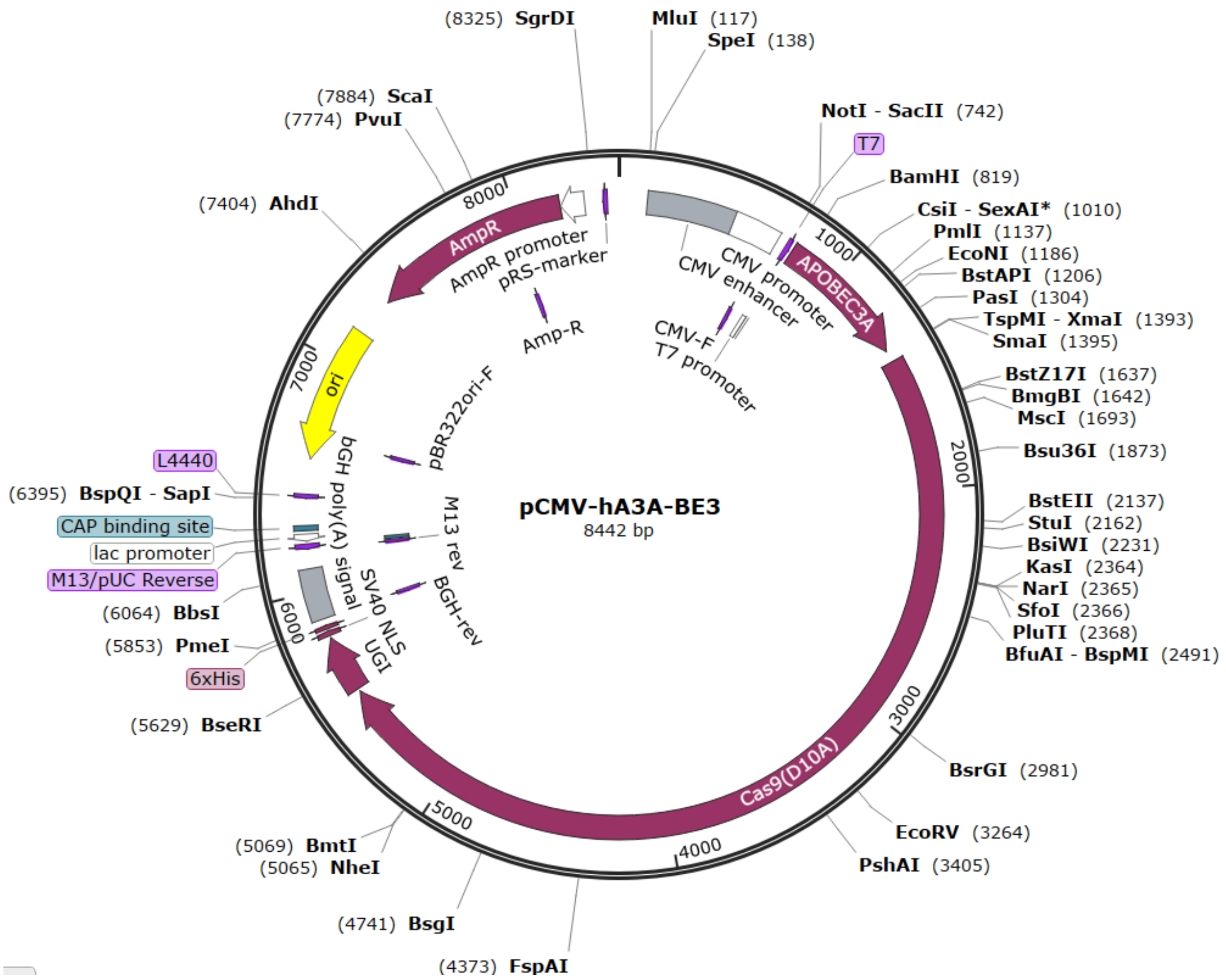
Colony PCR and Purification

We conducted a colony PCR using NEB PCR Protocol for OneTaq® DNA Polymerase (M0480). After several inoculations and colony PCR attempts, the last few colonies that were feasible to conduct a final PCR were colonies #2 and #5 from the A3A-ABE7.10 plate. For the PCR conditions, the A3A-ABE7.10 had an annealing temperature of 57˚C and the DNA from the colonies were picked with a sterile pipette tip. The PCR product was then purified using the Monarch® PCR & DNA Cleanup Kit (5 μg) (NEB #T1030). After the PCR products were purified, we ran an electrophoresis gel to verify if the plasmid DNA was present and was correctly amplified by comparing the DNA band lengths with the NEB 1 Kb Plus ladder. After analyzing the gel and conducted a second PCR and gel trial, we still got no visible DNA band at its corresponding DNA length. Therefore, we decided to send the samples for Sanger sequencing which would allow us to see if the ligated ABE7.10 gene sequence was even present in the A3A plasmid.

Results and Discussion

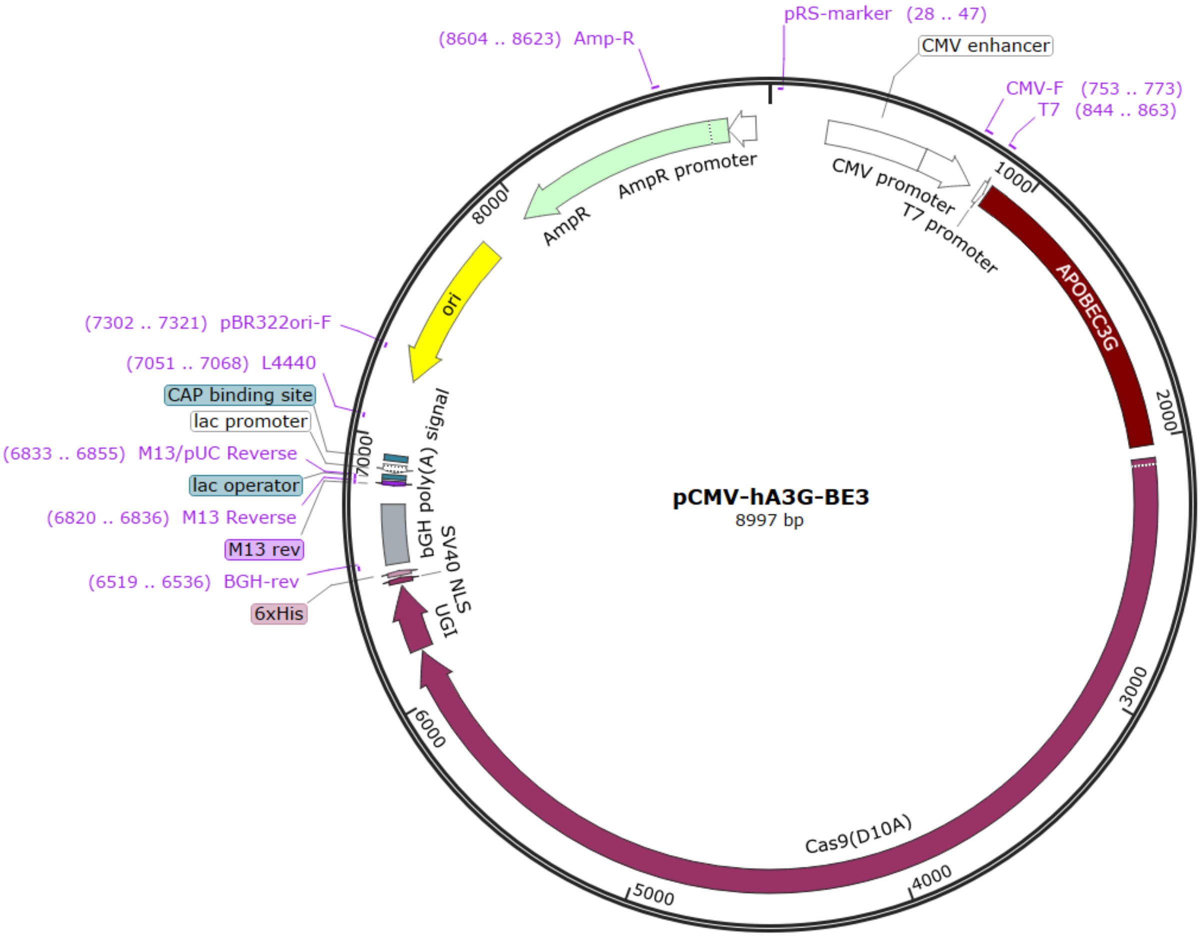
In result, although in despite of seeing bacterial growth from the 10-Beta competent cells overnight, after several trials of cloning, changing PCR and HiFi assembly conditions, and selecting different colonies on the plates, there was no promised data that the designed fusion proteins were successfully ligated and assembled. Even after conducting a gel electrophoresis of the colony PCR, there were still no visual of the large DNA band sizes around the 10kb. A fully ligated A3A-ABE7.10 would be 9384 bp and A3G-ABE7.10 would be 9939 bp. We sent them for sequencing to see if the plasmids were even ligated properly and unfortunately results came back negative, meaning the sequences of the plasmid DNA were still the original plasmid that we started with. Due to the lack of time, we were unable to further continue this work. However, we were able to successfully design and amplify the genes of interest from the pCMV-hA3A-BE3 and pCMV-hA3G-BE3 plasmids, along with the ABE7.10 gene (figures 20, 21, 22, and 23). This allowed us to acknowledge future works such as trying to ligate the plasmids into perhaps a smaller version of the plasmids, having a lower number of kilobase pairs nucleotides total (kb). The use of these fusion proteins in the future will increase the digital data stored at each DMOS domain, as they can be targeted combinatorically to either the top or bottom strands as well, which would further improve on the progress made in previous chapters.

Figure 20. Schematic image of the pCMV-hA3A-BE3 plasmid (8442 bp)



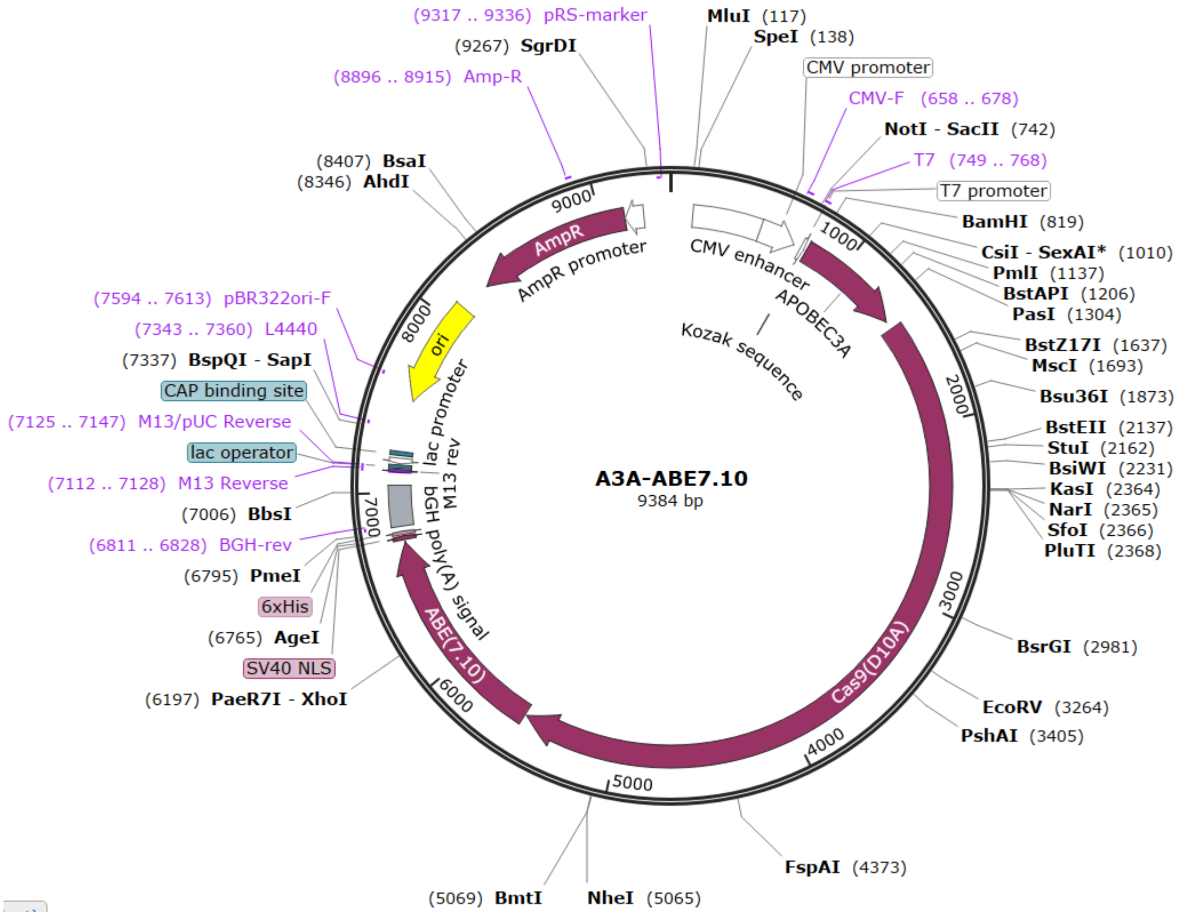
Note. Figure 20 is visual demonstration of the original APOBEC3A plasmid with Cas9. APOBEC3A is able to recognize 5’ -TC- 3’motifs.

Figure 21. Schematic image of the pCMV-hA3G-BE3 plasmid (8997 bp)



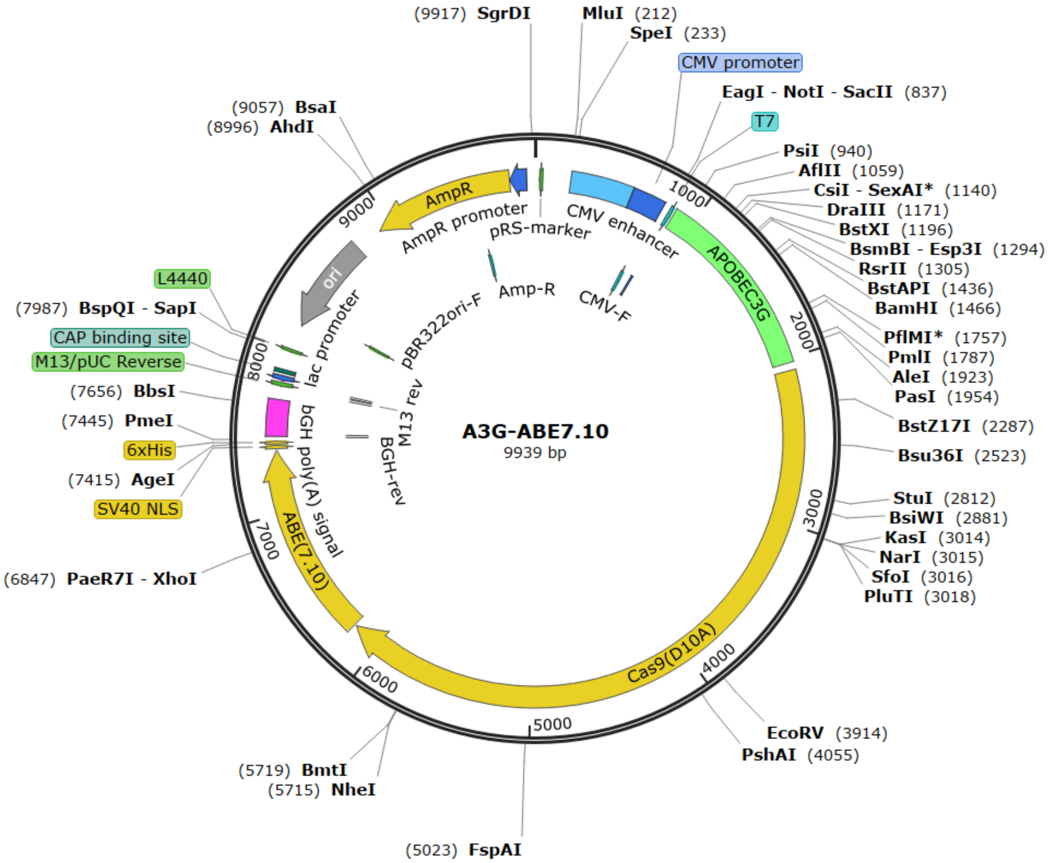
Note. Figure 21 is visual demonstration of the original APOBEC3G plasmid with Cas9. APOBEC3A is able to recognize 5' -CCC- 3'motifs.

Figure 22. Schematic designed A3A-ABE7.10 fusion plasmid (9384 bp)



Note. Figure 22 is visual demonstration of the designed APOBEC3A fused with the ABE7.10 gene, along with Cas9. APOBEC3A is able to recognize 5’ -TC- 3’motifs while the ABE7.10 (replaced the UGI gene in the original plasmid) conducts deamination of adenine to inosine.

Figure 23. Schematic designed A3G-ABE7.10 fusion plasmid (9939 bp)



Note. Figure 23 is visual demonstration of the designed APOBEC3G fused with the ABE7.10 gene, along with Cas9. APOBEC3A can recognize 5’ -CC- 3’motifs while the ABE7.10 (replaced the UGI gene in the original plasmid) conducts deamination of adenine to inosine.

CHAPTER V: SUMMARY AND FUTURE WORK

The work presented was performed to increase DNA data digital storage capacity utilizing CRISPR base editing, and was done by engineering novel “Janus” protospacers for CRISPR targeting where either the top or bottom strand of the DNA sequence could be mutated, “unmutated”/erased, or mutated with a different signature to increase the number of possibilities for data encoded at that DNA site. With the attempts to increase DNA data digital storage by writing and erasing, reading digital data from DNA molecules quickly also has its own challenges.

In Chapter 2, we were able to design sgRNAs and protospacer sequences into a template DNA that would have PAMs motifs on both the top and bottom ends of the protospacers, which is new and was not obvious that it would work from the mechanism of CRISPR target recognition. Having the double distal PAM motifs enable us to design the sgRNA to target and bind to both top and bottom strands of the protospacers simultaneously in a base mutation reaction. Although the sequencing results from the base mutation reactions of the Janus DMOS DNA template were unconclusive because we could not confirm the previously known positive controls that have contained mutations. For future work, we would need to repeat the experiment until our controls, which have previously demonstrated base mutation signatures, are positive for having mutational signatures, then only can we make an analysis and concluded whether the experiment worked or not. If it works, then we would also want to test this reaction with another base editor. In this case we only used APOBEC3A for the base mutational experiment, but it would be interesting to see if the APOBEC3G base editor would work in the same mutational experiment, since both APOBECs have a sequence motif preference of cytosine deamination.

In Chapter 3, we demonstrated that we could develop and design sgRNAs that would target both the top and bottom strand of a protospacer for APOBEC3G, which recognizes a CCC motif to generate an AAA motif on the bottom strand, which is a better sequence for recognition by adenine base editors. We were able to prove that the synthesized sgRNAs 1, g2, and g3 for Top (T) and Bottom (B) targeting protospacers with the CCC motif could be used with Cas9 to both bind and cut at the designated protospacer. Therefore, we could potentially mutate the CCC to TTT to write a '1', then later use the complementary sgRNA to mutate the bottom AAA back to GGG so the top strand would return to CCC or '0', thus writing then erasing a bit. Being able to bind and form an R-loop with both strands is a necessary condition for us to be able to represent more bits per DNA domain in a binary coding system and to enable writing, erasing, and re-writing of data.

Unfortunately, in terms of our work with trying to clone the cytosine and adenine base editors did not work. Although, the cloning experiment did not work, we remain optimistic that we can try this experiment again and make future modifications, as it would greatly expand the amount of information that could be stored within a single DMOS domain. A recommendation for future modification be to include using "Blue and White" screening with β-galactosidase to verify if any bacterial colonies contain the correctly cloned plasmid with the genes for fusion proteins ligated into regions of interested in smaller plasmids. We were able to accomplish in developing primers for targeting and amplifying the specific genes of interest. With this we further explore the possibilities to clone these plasmids into a smaller fusion plasmid. This enabled them to not be as large as almost 10 kb when ligated, which is what we would have obtained if the plasmids were successfully ligated. We could also try to utilizing different bacterial competent cells for ligating the plasmids.

Future work will enable the generation of new mutational signatures on DNA molecules to encode data without on-demand synthesis, increasing the capacity for data storage int this media without additional demands that synthesis requires. The work here shows how creative applications of DNA and design of biomolecular systems with engineered mutations can help us to reach that goal.

REFERENCES

1. Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M., & Hughes, W. L. (2016). Nucleic acid memory. Nature materials, 15(4), 366–370. https://doi.org/10.1038/nmat4594

2. Lin, K.N., Volkel, K., Tuck, J.M. et al. Dynamic and scalable DNA-based information storage. Nat Commun 11, 2981 (2020). https://doi.org/10.1038/s41467-020-16797-2

3. Organick, L., Chen, YJ., Dumas Ang, S. et al. Probing the physical limits of reliable DNA data retrieval. Nat Commun 11, 616 (2020). https://doi.org/10.1038/s41467-020-14319-8

4. Chen, YJ., Takahashi, C.N., Organick, L. et al. Quantifying molecular bias in DNA data storage. Nat Commun 11, 3264 (2020). https://doi.org/10.1038/s41467-020-16958-3

5. Digital data storage on DNA tape using CRISPR base editors. Afsaneh Sadremomtaz, Robert F. Glass, Jorge Eduardo Guerrero, Dennis R. LaJeunesse, Eric A. Josephs, Reza Zadegan. bioRxiv 2023.02.07.527074; doi: https://doi.org/10.1101/2023.02.07.527074

6. Heckel, R., Mikutis, G. & Grass, R.N. A Characterization of the DNA Data Storage Channel. Sci Rep 9, 9663 (2019). https://doi.org/10.1038/s41598-019-45832-6

7. Meiser, L.C., Antkowiak, P.L., Koch, J. et al. Reading and writing digital data in DNA. Nat Protoc 15, 86–101 (2020). https://doi.org/10.1038/s41596-019-0244-5

8. Koch, J., Gantenbein, S., Masania, K. et al. A DNA-of-things storage architecture to create materials with embedded memory. Nat Biotechnol 38, 39–43 (2020). https://doi.org/10.1038/s41587-019-0356-z

9. Hughes, R. A., & Ellington, A. D. (2017). Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. Cold Spring Harbor perspectives in biology, 9(1), a023812. https://doi.org/10.1101/cshperspect.a023812

10. Ma, E., Harrington, L. B., O'Connell, M. R., Zhou, K., & Doudna, J. A. (2015). Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes. Molecular cell, 60(3), 398–407. https://doi.org/10.1016/j.molcel.2015.10.030

11. Xue, C., & Greene, E. C. (2021). DNA Repair Pathway Choices in CRISPR-Cas9-Mediated Genome Editing. Trends in genetics: TIG, 37(7), 639–656. https://doi.org/10.1016/j.tig.2021.02.008

12. Ran, F., Hsu, P., Wright, J. et al. Genome engineering using the CRISPR-Cas9 system. Nat Protoc 8, 2281–2308 (2013). https://doi.org/10.1038/nprot.2013.143

13. Doudna, J. A., & Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science (New York, N.Y.), 346(6213), 1258096. https://doi.org/10.1126/science.1258096

14. Konermann, S., Brigham, M., Trevino, A. et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature 517, 583–588 (2015). https://doi.org/10.1038/nature14136

15. Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., Ishitani, R., Zhang, F., & Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell, 156(5), 935–949. https://doi.org/10.1016/j.cell.2014.02.001

16. Whinn, K.S., Kaur, G., Lewis, J.S. et al. Nuclease dead Cas9 is a programmable roadblock for DNA replication. Sci Rep 9, 13292 (2019). https://doi.org/10.1038/s41598-019-49837-z

17. Karlson, C. K. S., Mohd-Noor, S. N., Nolte, N., & Tan, B. C. (2021). CRISPR/dCas9-Based Systems: Mechanisms and Applications in Plant Sciences. Plants (Basel, Switzerland), 10(10), 2055. https://doi.org/10.3390/plants10102055

18. Wang, C., Qu, Y., Cheng, J.K.W. et al. dCas9-based gene editing for cleavage-free genomic knock-in of long sequences. Nat Cell Biol 24, 268–278 (2022). https://doi.org/10.1038/s41556-021-00836-1

19. Salter, J. D., Bennett, R. P., & Smith, H. C. (2016). The APOBEC Protein Family: United by Structure, Divergent in Function. Trends in biochemical sciences, 41(7), 578–594. https://doi.org/10.1016/j.tibs.2016.05.001

20. Kantor, A., McClements, M. E., & MacLaren, R. E. (2020). CRISPR-Cas9 DNA Base-Editing and Prime-Editing. International journal of molecular sciences, 21(17), 6240. https://doi.org/10.3390/ijms21176240

21. Barka, A., Berríos, K. N., Bailer, P., Schutsky, E. K., Wang, T., & Kohli, R. M. (2022). The Base-Editing Enzyme APOBEC3A Catalyzes Cytosine Deamination in RNA with Low Proficiency and High Selectivity. ACS chemical biology, 17(3), 629–636. https://doi.org/10.1021/acschembio.1c00919

22. Grünewald, J., Zhou, R., Lareau, C. A., Garcia, S. P., Iyer, S., Miller, B. R., Langner, L. M., Hsu, J. Y., Aryee, M. J., & Joung, J. K. (2020). A dual-deaminase CRISPR base editor enables concurrent adenine and cytosine editing. Nature biotechnology, 38(7), 861–864. https://doi.org/10.1038/s41587-020-0535-y

23. Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., & Liu, D. R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature, 551(7681), 464–471. https://doi.org/10.1038/nature24644

24. Gaudelli, N.M., Lam, D.K., Rees, H.A. et al. Directed evolution of adenine base editors with increased activity and therapeutic application. Nat Biotechnol 38, 892–900 (2020). https://doi.org/10.1038/s41587-020-0491-6

25. Zong, Y., Song, Q., Li, C. et al. Efficient C-to-T base editing in plants using a fusion of nCas9 and human APOBEC3A. Nat Biotechnol 36, 950–953 (2018). https://doi.org/10.1038/nbt.4261

26. Porto, E.M., Komor, A.C., Slaymaker, I.M. et al. Base editing: advances and therapeutic opportunities. Nat Rev Drug Discov 19, 839–859 (2020). https://doi.org/10.1038/s41573-020-0084-6

27. Doench, J., Fusi, N., Sullender, M. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol 34, 184–191 (2016). https://doi.org/10.1038/nbt.3437

28. Sanson, K.R., Hanna, R.E., Hegde, M. *et al.* Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat Commun* **9**, 5416 (2018). https://doi.org/10.1038/s41467-018-07901-8