GLASS, ROBERT F. Ph.D. Molecular Information Storage using CRISPR. (2023)
Directed by Dr. Eric Josephs. 89 pp.

DNA (deoxyribonucleic acids) can be used as a digital data storage medium. Using its nucleotide sequence as a basis for storing digital data, DNA has a significantly higher data density and longer lifespan compared to silicon-based digital data storage technologies. However, the drawbacks of synthesizing long sequences of nucleotides to record data on demand are so significant that they limit the use of DNA for this purpose. This dissertation will discuss methods to circumvent the requirement to synthesize nucleotides and proposes that we instead directly encode digital information onto a standardized double-stranded DNA (dsDNA) template *in vitro* through targeted mutations. This encoding scheme uses nucleobase (base) editing enzymes to alter DNA sequences at specific sites in pre-synthesized dsDNA templates as a way to store digital information. This document will explore the application of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) base editors to encode digital data into DNA molecules and how we can use DNA sequencing technologies like nanopore sequencing to extract data from these molecules.

We have developed an approach to apply CRISPR "base editing" reactions *in vitro*, where we chemically convert specific nucleotides from one to another onto pre-synthesized DNA "templates" to encode and extract >1250 bits, like they were a magnetic tape, using nanopore sequencing. After developing our *in vitro* biochemical encoding strategy and optimizing the nucleotide targets to be mutated, we demonstrate that we can direct CRISPR base editing reactions to perform cytosine mutations *in vitro* using the cytosine deaminase APOBEC3A and a form of the CRISPR enzyme Cas9 (dCas9) that binds to specific 20 bp sequences without cleavage. We then applied our strategy to generate controlled and detectable

mutations *in vitro* and that we can decode our intended data set with 100% accuracy. This document will conclude with future directions to improve our strategy to encode digital data with CRISPR base editing, such as ligating synthesized oligos into unique sequence arrangements and applying alternate deaminating enzymes to generate different mutation patterns. While other scientists have used CRISPR base editing to mutate genomic DNA in cells, this work demonstrates that this technique can be adapted to improve DNA as a medium to store digital data by directly encoding the digital information onto a standardized DNA template *via* targeted mutagenesis.

MOLECULAR INFORMATION STORAGE USING CRISPR

by

Robert F. Glass

A Dissertation
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
Doctor of Philosophy

Greensboro

2023

Approved by

Dr. Eric Josephs
Committee Chair

DEDICATION

I would like to dedicate this dissertation to my parents and family for their support and excitement for my work. They helped make me the individual I am today, and I believe that I would never have made it this far in my career without their enthusiastic support.

APPROVAL PAGE

This dissertation written by Robert F. Glass 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 Josephs

Committee Members

Dr. Dennis LaJuennesse

Dr. Reza Zadegan

Dr. Daniel Herr

March 22, 2023

Date of Acceptance by Committee

March 14, 2023

Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to thank Dr. Eric Josephs for guiding me throughout graduate school and being an influential advisor. You kept me on track and focused throughout my graduate school career, and it has been an honor to be a part of your lab group. I would like to thank Dr. Kulberg, Dr. Bagchi, and Dr. Salehin for their advice whenever I had trouble with my research. I would also like to thank all of the graduate student members of Josephs's lab for their encouragement and enthusiasm for this work.

From my years of an undergraduate, I would like to thank Dr. Pamela Lundin of High Point University for being the spark of my research career. You have been the one to introduce me to the Joint School and Nanoscience and Nanoengineering. You have encouraged me to push past my limits and given me the skills necessary to perform research, regardless of the field. I would also like to thank Mr. Ronald Obie of the Wood Coatings Research Group for hiring me part-time to expand my laboratory work knowledge in the textile industry.

I would like to thank the National Science Foundation (NSF) and the National Institute of Health (NIH) for funding this project. I also would like to thank Dr. Afsaneh Sadremomtaz and the Zadegan lab for collaborating on the Directed Mutagenesis Overwriting Storage (DMOS) project to help advance the integration of biology with semiconductor technologies. The coding in Python and C++ for data analysis was designed and performed by the Zadegan lab, influenced by the MATLAB decoding strategy from Dr. Eric Josephs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: LITERATURE REVIEW

Introduction

Society is using an ever-expanding amount of digital information storage every year, with the amount of digital information that is stored predicted to reach up to 175 zettabytes by 2025, according to a study presented in the Semiconductor Synthetic Biology conference in 2018.[1] As such, there is a desire to find alternatives to current solid-state digital storage devices for the purpose of archival digital data storage. Deoxyribonucleic acids (DNA) have the potential for greater data storage density and longevity (DNA sequences can be determined from molecules created centuries or thousands of years ago) compared to semiconductor-based memory devices.[2] DNA molecules therefore are excellent candidates for digital data storage, as one can assign the sequence of nucleotides to bit values.[3–9]

There have been significant advancements in how DNA has been utilized for data storage *via* encoding and error correction schemes.[10–15] Specific DNA sequences can be extracted from pools of oligonucleotides to be collected and translated individually to decode the digital data stored in their sequence *via* polymerase chain reactions (PCR), if they have been labeled with unique barcodes or fluorescent tags.[16–18] While DNA can store binary information such as 0's and 1's to write text, more complex forms of digital data, such as movies and music, can be encoded into nucleic acid templates, showing the versatility of DNA digital data storage[19,20] While DNA shows promise in its capacity to store digital material within their sequences, one of the medium's most significant obstacles is that its precise nucleic acid sequences must be created with near-perfect precision on demand to correctly translate the bit-level data.[21,22]

One method to overcome this obstacle is to perform precise mutagenesis on nucleic acid templates to directly change the DNA sequence of an existing DNA molecule without having to

synthesize a new one, much like semiconductor memory devices can change bits of information on the device without having to fabricate a new device as needed. The combination of CRISPR-Cas9 proteins and nucleobase editing enzymes like APOBEC are efficient in modifying the nucleic acid sequence of genomic DNA in a process called "base editing." Base editing has been used in human cells to change DNA sequences to treat developmental disorders such as Hutchinson-Gilford progeria syndrome, and there is great interest in using base editing technologies therapeutically.[23–26] The CRISPR-Cas9 system used in base editing has been optimized for these applications to improve their specificity in mutating specific nucleotide sequences, reducing off-target edits. Base editors, including those that mutate cytosine and adenine nucleotides to thymine and guanine nucleotides, respectively, have been engineered to increase the possibilities of targeted mutations performed on DNA sequences. Compared to other gene editing strategies, which rely on the cellular repair of double-strand breaks, base editing reduces the possibility of unwanted mutations. Because of the enhanced precision of base editing, this technique has been performed to mutate genomic DNA in both plant and mammalian cell lines.[27–32] Engineered fusion proteins that link CRISPR-Cas9 proteins with nucleotide-modifying enzymes can modify DNA therapeutically, but we show they can also be used to encode digital information into synthetic DNA molecules for data storage applications.

Using DNA as a digital storage medium is a potentially promising way to meet information storage demands if its data storage density can be increased without synthesizing new sequences. We show it is possible to encode digital information into synthetic DNA templates using CRISPR-Cas9 base editing to write bit states onto the molecules that can be deciphered using nanopore sequencing.[33–36] Specific bits of information can be written into DNA molecules since DNA-based digital data storage relies on the sequence of nucleotides on the

DNA template. Targeted mutation can be performed to introduce these bits at specific locations because nucleotide altering enzymes can only change the sequence of a DNA if the DNA is in a single-stranded conformation. Predominantly double-stranded DNA molecules will not be mutated unless a single-stranded segment of DNA is exposed at the site of CRISPR-Cas9 binding. These alterations to the nucleotide sequence can then be detected by sequencing the DNA molecules. Applying base editing methods to store digital data has multiple benefits compared to existing strategies where DNA must be synthesized *de novo*.[37–39] Without having to synthesize a new DNA molecule, using CRISPR base editing onto a DNA template can encode new information similarly to magnetic tape. Therefore, prefabricated structures can be altered in a way so that data can be reliably retrieved upon determining the structure and sequence of the information-storage units.

While base editing can enhance the capability of DNA as a storage medium, this review will discuss base editing mechanisms from existing gene editing strategies and determine how they can be applied to digital encoding. DNA as a data storage medium will also be reviewed and discussed to analyze potential pitfalls that need to be addressed and how base editing can alleviate those pitfalls. While base editing is a promising strategy *in vivo* for therapeutic applications, it can be a valuable asset to increasing the potential of DNA as a viable data storage medium.

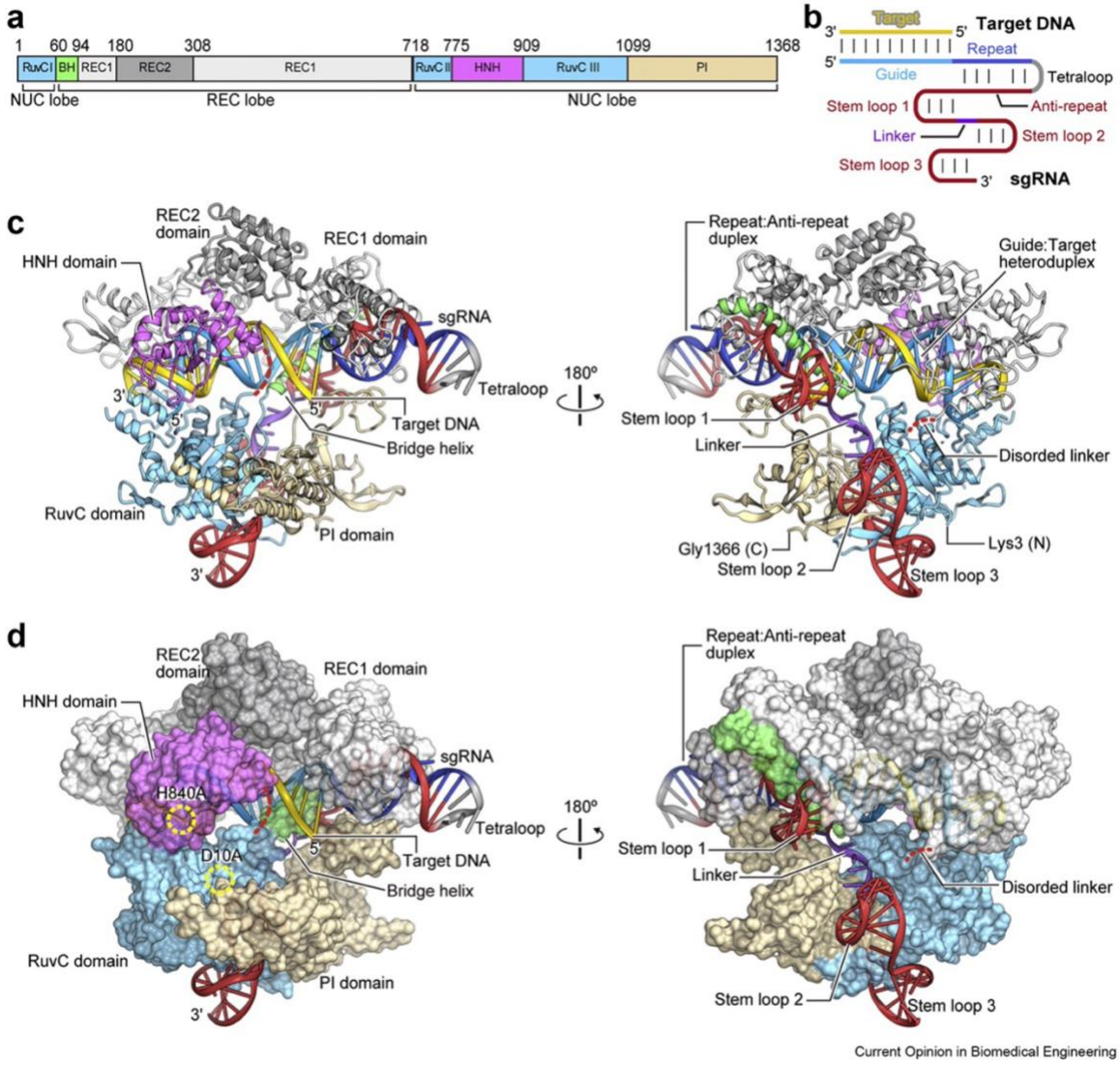
Type II CRISPR-Cas9 System

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are immunological defense mechanisms that protect prokaryotic organisms against the invasion of foreign DNA and other genetic material.[40] There are six recognized types of CRISPR-Cas systems that have been found in prokaryotes, with type II CRISPR systems being the most utilized for gene editing

applications.[41] Immunity is conferred from two components in the type II CRISPR-Cas9 system, the Cas9 effector nuclease and the CRISPR RNA (crRNA), which guides the binding of the Cas9 effector to specific sequences complementary to a 20 bp "spacer" sequence at the 5'- end of the RNA molecule to activate nuclease activity.

The Type II Cas9 effector is an endonuclease that cleaves DNA molecules that are complementary to crRNA spacer sequences, with the most commonly used effector being derived from *Streptococcus pyogenes* bacteria. *S. pyogenes* Type II CRISPR RNAs are derived from the genomic loci of that bacteria that contains a repeated array of spacer sequences that directs Cas9-based immunity acquired after surviving an infection.[42] As foreign DNA material is re-introduced to the cell; the spacer array is transcribed to form a precursor of CRISPR RNA sequences (pre-crRNA) complementary to the foreign invading nucleotide molecule. After processing, the crRNA spacers are bound to a non-coding RNA sequence called tracer RNA (tracRNA) to form a chimeric RNA that programs the effector to bind to foreign nucleotide sequences and initiate cleavage. The structure of the Cas9 protein binding site and cleavage domains are shown in Figure 1. This chimeric RNA sequence or guide RNA (gRNA) controls the binding location of the Cas9 effector. In biotechnological applications, the two RNA are fused into a single guide RNA (sgRNA) to direct Cas9 cleavage.

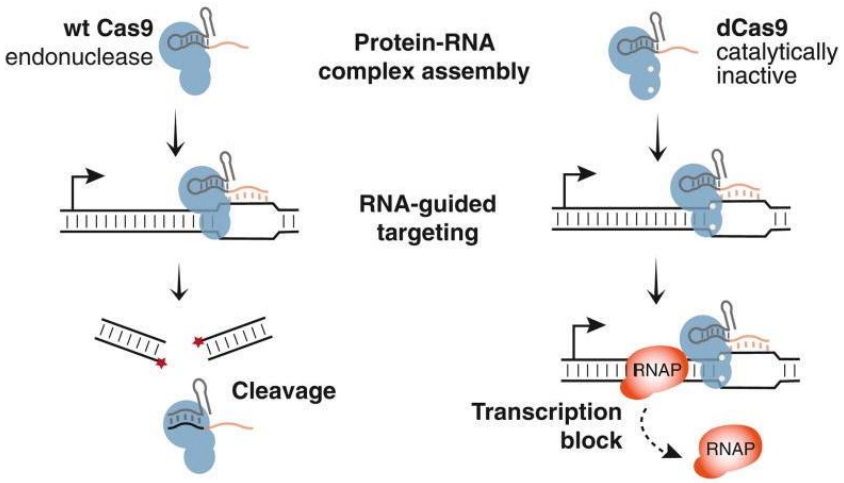
Figure 1 – Crystal Structure of Cas9 with sgRNA and DNA Complex[43]



Note. A) Domains of the Cas9 protein consist of nuclease (NUC) lobes and a recognition lobe (REC), which are required for cleaving the DNA target and recognizing the sgRNA, respectively. B) Schematic of sgRNA binding to the complementary target sequence. C) Ribbon representation of Cas9:sgRNA complex bound to DNA. D) Surface representation of spCas9:sgRNA complex bound to DNA. Cleavage activity is driven by the RuvC and HNH domains, and can be deactivated by edits in the H840A and D10A amino acids.

The Cas9-sgRNA complex, or ribonucleoprotein (RNP), then binds to the nucleotide sequence that is complementary to the spacer sequence (the "protospacer" DNA) that must be located adjacent to a short sequence named the protospacer adjacent motif (PAM). For *S. pyogenes* Cas9, the PAM is 5'-NGG-3', where N is any nucleotide, and it must be located 3'- of the sequence complementary to the spacer. After the gRNA spacer sequence has base-paired with the protospacer sequence, the effector cuts the DNA *via* two domains in its nuclease lobes, the HNH domain and the RuvC domain.[44] While the gRNA is base-paired with the protospacer sequence complementary to the 20 bp RNA spacer, a structure known as an R-loop is formed, in which the DNA is bent slightly to allow the cleaving domains of Cas9 to cut both backbones of the DNA target, and there is a single-stranded loop of DNA formed by the strand that is not hybridized with the gRNA. In our scheme for writing digital data onto DNA, we used an engineered Cas9 protein where the nuclease domains have been catalytically inactivated (dead Cas9 or dCas9). We applied dCas9 proteins to direct the mutations of nucleotide editing enzymes that alter DNA nucleotides in single-stranded R-loop formations. The DNA manipulating mechanisms between Cas9 and dCas9 are presented in Figure 2.

Figure 2 – CRISPR as an Editing Tool[45]

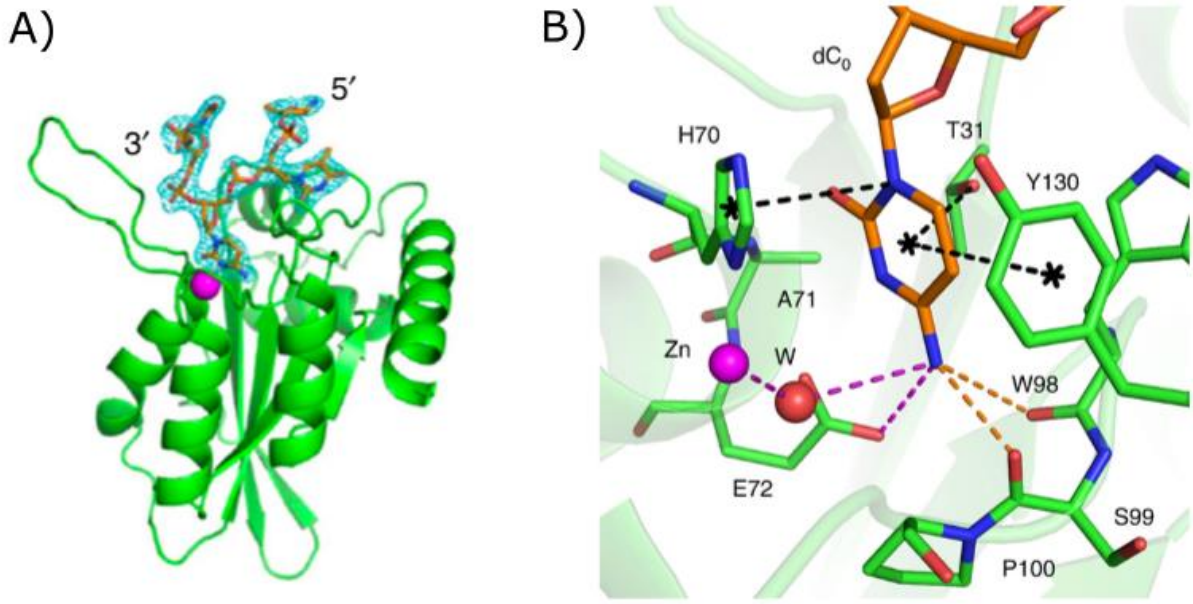


Note. In the presence of the Cas9 RNP, the RNP complex binds to the target sequence and cleaves both phosphate backbones of the DNA template. If the Cas9 protein is catalytically inactive, then the protein binds to the target with no backbone cleavage interactions, leading to applications such as blocking RNA transcription.

APOBEC3 and TadA Editing Functions

APOBEC (apolipoprotein B mRNA editing enzyme catalytic polypeptide) includes a domain that deaminates cytosine nucleobases, altering its chemical structure to uracil. This alteration is done by substituting amine with a carbonyl group through hydrolysis, which is then used as a complementary template for adenine nucleobases.[46] Thus, cytosines are effectively converted to thymines. APOBEC proteins are involved in various biological functions, such as demonstrating adaptive immune responses against HIV-like retroviruses, metabolizing fatty lipids, and found to correlate with tumor formations.[47–49] Up to 11 members of APOBEC proteins are present in human cells. Seven members are known to respond to viral invasions within eukaryotic cell lines.[50–52] These seven proteins belong to a subcategory of APOBEC proteins called APOBEC3 (A3A, A3B, A3C, A3D, A3F, A3G, A3H), which directly contribute to the deamination of cytosines in single-stranded polynucleotides in response to foreign RNA invasions.[53–55] Each APOBEC3 protein has different propensities to deaminate cytosines based on their sequence contexts. For example, the A3A proteins deaminate cytosines of the TC motif, while the A3G targets the CC motif.[56] The structure and active site of APOBEC3A are shown in Figure 3. The APOBEC3 enzymes only recognize cytosines if they are not base-paired, allowing us to exploit this property to mutate cytosines in double-stranded DNA only if they are part of the un-paired strand of an R-loop formed by dCas9 binding.

Figure 3 – APOBEC3A Structure and Active Site[26]

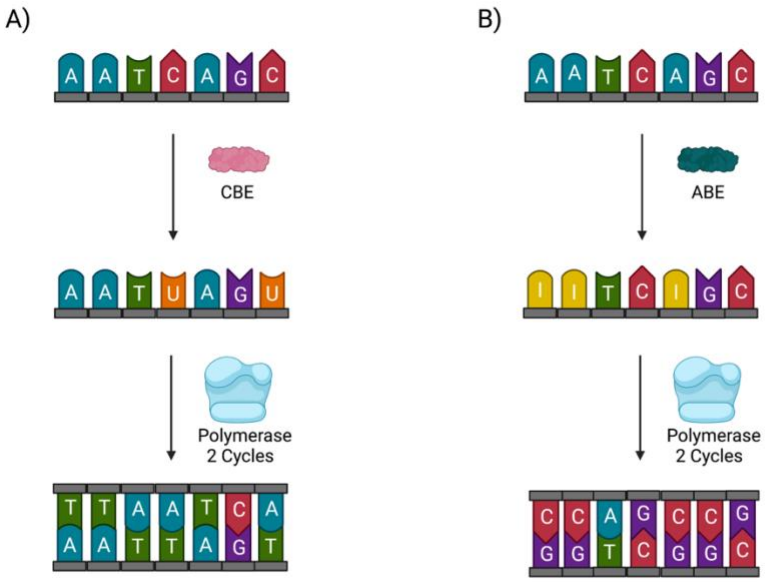


Note. A) APOBEC3A protein is depicted as a green ribbon structure with the bound DNA presented as a stick representation. B) Close-up view of DNA bound to the active site of A3A. Carbons and phosphates are colored orange, while nitrogen and oxygens are labeled blue and red, respectively.

Another nucleotide modifying enzyme used in base editing applications comes from the tRNA editing enzyme TadA, which explicitly targets tRNA adenine nucleobases.[57,58] TadA orthologues are found in mammalian and prokaryotic cells, like cytosine deaminases. Several alterations have been made to this enzyme to increase deamination efficiency in DNA while reducing off-target nucleotide alterations.[59] To enhance the TadA enzyme to recognize DNA, Dr. David Liu's work improved TadA's editing efficiency by altering 14 amino acids in the TadA protein and linking the engineered protein to an unaltered TadA enzyme.[60] These alterations changed the structure from a monomer into an engineered heterodimer that targets DNA strands more prominently than RNA molecules. Adenine conversions increased from 40% up to 80%

across various genomic sites in human kidney cells (HEK293) when this engineered deaminase was applied.[60] The deamination mechanism of both cytosine and adenine deaminase products is presented in Figure 4.

Figure 4 – Cytosine and Adenine Deamination Pathways



Note. A) Cytosine deaminating proteins convert the nucleobase cytosine into a carboxyl group-containing uracil, translated into thymine after two rounds of replication. B) Adenine deaminating proteins alter the purine structure of adenine, converting to inosine. After two rounds of replication, the inosine is replaced with guanine, changing the type of Watson-Crick pairing from A-T to C-G. (Created with BioRender.com)

While both proteins can readily alter nucleotides in single-stranded RNA molecules, DNA targets must also be in a single-stranded conformation to be a recognizable target for pyrimidine and purine modifications. They have been used in CRISPR base editing reactions in cells by linking CRISPR-Cas9 effector enzymes with these deaminating proteins *via* amino acid linkers to produce targeted base editing within the R-loops formed when Cas9 binds to its target.

Depending on the linked enzyme applied (APOBEC or TadA), these fusion proteins become cytosine base editors (CBE) or adenine base editors (ABE), respectively.

Targeted CRISPR Base Editing in DNA

CRISPR base editing systems that utilize fusion proteins linked to deaminating proteins generate chemical modifications of nucleotides directly at sites determined by the gRNA spacer sequence.[61] By linking an adenine or cytosine deaminating enzyme with a Cas9 protein *via* short amino acid chains and using the sequence of the short sgRNA spacer to control Cas9 binding, researchers can selectively introduce nucleotide edits within a eukaryotic genome by driving the deaminating protein towards targets specified by the sgRNA spacer. This linking has also been shown to reduce off-target activity, most likely the result of keeping the deaminating protein close to the DNA-binding Cas9. This has been demonstrated in human cells to model mitochondrial diseases and induce cytosine to guanine edits by incorporating both cytosine base editors and adenine base editors in human cells.[62,63] The use of base editors has also effectively altered human genomic DNA in HEK293 cells with minimal off-target mutations, showcasing the wide applications that base editing can exploit.

With this promising method of introducing precise genomic mutations, researchers have been developing tools and resources to enhance the specificity and efficiency of nucleotide conversions. It was found that deamination efficiency was enhanced by utilizing a nickase Cas9 (nCas9), in which the RuvC domain in Cas9 is edited so that the protein's cleavage activity is partially deactivated.[27,64–66] By breaking the phosphate backbone from the unhybridized strand, the deaminating proteins interact with the nucleobases that now have more conformational flexibility. This enables the DNA sequence to be edited and fixed without DNA double strand break (DSB) repair-associated mutations. This combination of nCas9-deaminating protein is

known as a “base editor” enzyme. Its efficiency also depends on the specificity of the sgRNA used in the reaction, as nicking of the phosphate backbone can only occur when the spacer is bound to the DNA template.

Since deamination reactions rely on the template being in a single-strand conformation, the gRNA spacer is a vital component to base editors. Not all gRNAs are equally effective in promoting R-loop formation, and it is a complex problem to predict which spacer sequences will result in stable dCas9 binding. CRISPR-dCas9 screens in bacteria were performed predict for on-target efficiency in *E.coli* for gRNA spacer sequences, as well as predicting gRNA spacers that are toxic to the bacteria based on 5 nucleotide sequences.[67,68] To improve gRNA protospacer activity prediction at both their intended and unintended sequences, several research groups have developed strategies to detect guides that exhibit off-target activity, such as Detect-seq and SWISS, and utilized them to measure off-target mutations in the genome.[29,69] Detect-seq’s screening mechanism enables the detection of off-target mutation across multiple spacers by identifying the presence of uracil on outside targets based on screening techniques such as GUIDE-seq.[70] These screening methods are important milestones that demonstrate that gRNA design is essential for base editing efficiency.

Overview of DNA as a storage medium

An organism's genome encodes all the biological information used to express phenotypic characteristics and holds an immense amount of genetic information within a single cell. The DNA that makes up the genome is made up of only four nucleotides but it is the sequence that encodes this information. These four nucleotides can also be interpreted as binary information, or bits (0 or 1) if DNA is used in molecular information storage.[71] According to estimations done by Grigoryev, a human diploid genome can store up to 1.5 Gb of digital data, or a whole

organism can store up to 150 zettabytes of information across all of their cells.[72] With automation, *de novo* synthesis of DNA molecules can be compared with solid-state media devices in being a long-term and stable storage medium that can overcome the disadvantages of magnetic tapes, floppy disks, etc.[73,74]

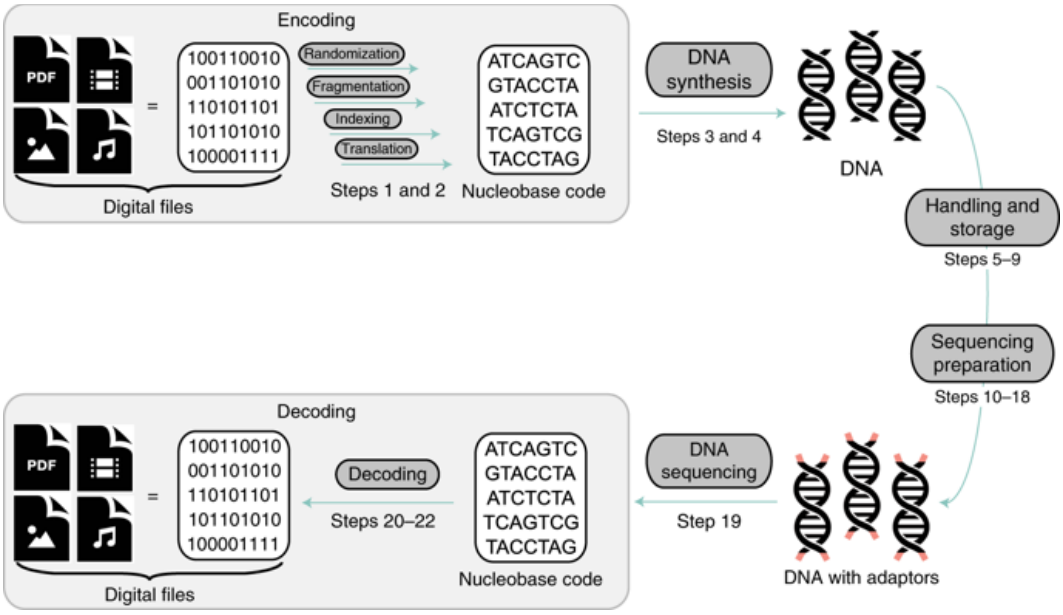
When oligonucleotides are translated into binary information, each nucleotide can be assigned as a pair of bits, indicating either 00, 01, 10, or 11, enough for each of the four nucleic acids in DNA. To encode information into DNA, this assignment can be used to generate an arrangement of nucleotides to be synthesized so that it can later be sequenced to retrieve the desired information. These templates can then be frozen or encapsulated in silica to be stored for millennia.[12,75] The overall schematic of translation between binary and nucleotide base codes is visualized in Figure 5.

Translating nucleotide signals of a DNA molecule into binary based requires DNA sequencing, a process which has been developed for a many other biological applications. In 2003, the Human Genome Project to sequence the full human genome (about 3 billion base pairs) cost about 2.7 billion dollars and took 13 years to complete.[76] However, thanks to advances in sequencing technology, it is now possible to sequence genomic DNA below $1,000 in less than a day. For DNA to advance as a proper storage medium, sequencing tools would need to accurately identify long arrangements of nucleotides in quick and cost-effective methods.

As mentioned previously, DNA nucleotides can be interpreted as binary codes. However, sequencing and DNA synthesis technologies are relatively error prone, and minor differences between sequencing signals will significantly alter the information decoded. To account for these potential source of errors, several error correction codes and mechanisms have been applied to nucleotide encoding schemes to help reduce errors.[77,78] Examples include Low-Density Parity

Checks (LDPC) and Reed-Solomon error codes.[79–81] If proper encoded schemes are employed, digital information stored in DNA that might have faulty sequencing can still be corrected to ensure that the correct data is extracted accurately, which is vital if synthetic DNA is going to be used as an archival storage medium.

Figure 5 – Overview of DNA Data Storage[82]



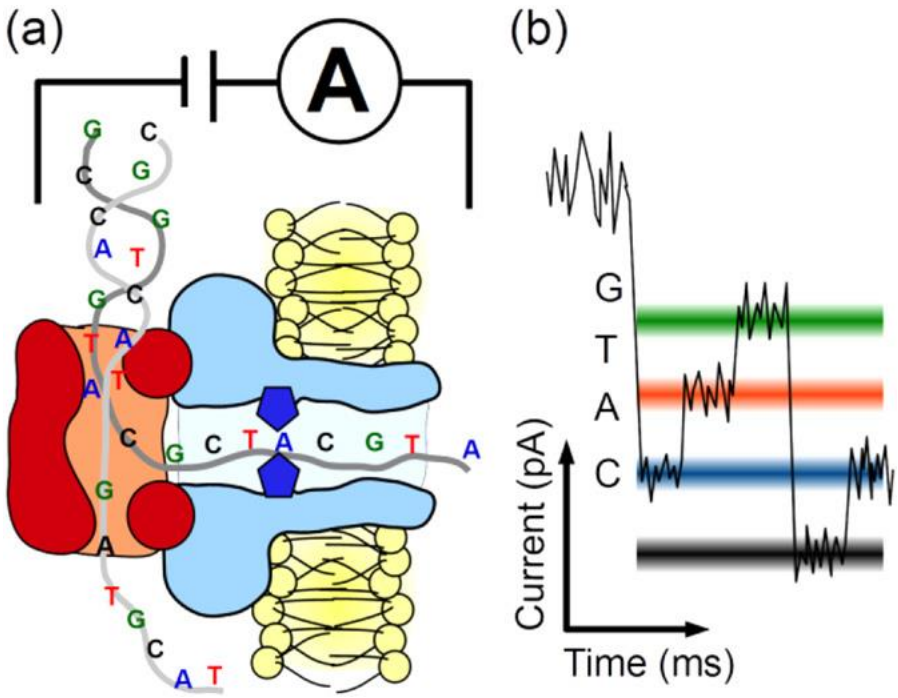
Note. Digital files are translated from binary information into a DNA sequence and then used to synthesize DNA molecules with that nucleotide sequence. The DNA molecules can then be "sequenced" and decoded to retrieve the digital file.

Next Generation Sequencing (NGS), or massive parallel sequencing, is a process to sequence small DNA fragments, that can later be assembled into larger contiguous DNA sequences with high confidence. A limitation is that during NGS with illumina technology, only DNA fragments of less than 500 bp can be read at one time.[83] Another issue that NGS can encounter is detecting smaller populations of DNA templates in pools with multiple DNA samples. One method in which this is applicable comes from a study done by Choi et al.'s study to eliminate two problems that derives from DNA data retrieval: the uneven population of

original and amplified DNA molecules after DNA modification, and the accessibility of distinct DNA sequences within a single medium.[38] By linking indexed 160 nt oligo templates onto lithographic micro-disks and applying PCR with selected reverse primers, extracting selected information among a pool of oligos.

Nanopore sequencing was developed to enable large-scale DNA sequencing of longer DNA molecules, up to 150 kilobases from a single molecule.[84] While nanopore sequencing gives longer reads compared than illumina sequencing, it's accuracy is lower for molecules longer than 500 bp. For sequencing DNA molecules *via* nanopore, a nanoscale pore is embedded in an electrically resistant polymer membrane that polymer materials travel through via an ionic current.[85] For DNA molecules to pass through the nanopore, a motor protein unwinds the double helix and drives one strand through the pore. As the strand passes through the nanopore, changes in electric current are recorded and characterized in sets of 5 or 6 nucleotides at a time. This process is visualized in Figure 6. Oxford Nanopore Technologies has developed and commercialized the technique in 2015, making it an accessible device for sequencing long DNA molecules as opposed to short-read sequencing methods.[86]

Figure 6 – Oxford Nanopore Sequencing Schematic[87]



Note. A) Nanopore Sequencing uses an electrically-resistant polymer nanopore membrane with a motor protein helicase. The helicase unravels double-stranded DNA molecules and drives the single-strand through the nanopore. B) Changes in ionic charge as the DNA molecules are moved through the pore are recorded as groups of nucleotides pass through the membrane, and from the electrical signal the sequence of the DNA molecule can be estimated with high confidence.

Conclusion

A major challenge for digital data storage using DNA is that DNA must be synthesized on-demand to store specific information to encode a piece of data. Using rewriteable media, like magnetic tapes and discs, have the advantage that they can be produced "blank" at scale then have data encoded into them as needed (without having to fabricate a new disk each time). In this dissertation we have worked to make DNA a rewritable media for digital data storage. Having

this ability by performing direct and accurate base edits can enhance the potential of digital data encoding without solid-state media.

For the rest of the dissertation we will show how we can use DNA as a rewritable data storage media through the use of CRISPR base editing reactions reconstituted *in vitro*. The later sections of this dissertation will cover the *in vitro* analysis of base editing of pre-synthesized DNA "templates" using dCas9 and APOBEC3A to encode digital information.

CHAPTER II: IN VITRO MUTAGENESIS VIA CRISPR AND APOBEC3A

Introduction

DNA templates as a digital storage medium requires precise synthesis and sequencing, as each nucleotide coding for a set of bits are necessary for preparing the desired digital file. As such, the stability of a DNA molecule's sequence is critical for accurate data retrieval, as an incorrect identification of nucleotides will produce an output of misarranged bit codes or corrupted data. While CRISPR has been used to alter DNA sequences for biological as well as digital DNA storage applications, these applications have mostly been performed in living cells.[88] *In vitro* mutagenic reactions could confer multiple benefits relative to mutation in cells because they have the advantage of strict biochemical control of the reaction. For example, designing and transfecting plasmids that express desired sgRNAs and fusion CRISPR proteins to target specific genes takes several days to perform, and mutagenic efficiency is often not high (below 10%). One advantage that *in vitro* mutagenesis confers is that it enables individual mutation reactions through direct application of purified deaminating proteins and prepared complexes of dCas9 and a gRNA or CRISPR ribonucleoproteins (RNPs). Users can directly apply the RNPs to target sequences where they want the mutations to occur, giving direct them control over the mutagenic process. To our knowledge, our work is the first to reconstitute the base editing reaction outside the cell, and here we use it for applications of digital data encoding.

As a demonstration of *in vitro* base editing, this chapter will discuss the application of catalytically inactive Cas9 enzymes (dCas9) with the APOBEC3A protein to demonstrate the functionality of controlled base editing of synthetic DNA templates. We designed a synthetic DNA "tape" or "template" that contains 16 individual 20 bp protospacers, each with a group of cytosines susceptible to APOBEC3A mutations. The dCas9 with a gRNA sequence will form an

R-loop in the DNA complementary to the gRNA's spacer sequence, while the APOBEC3A protein will target those R-loops for mutagenesis. Pilot studies used Sanger sequencing to detect mutations, and templates longer than 700 bp were sequenced via nanopore sequencing. Experimental samples were compared with non-mutated regions to determine the effectiveness of our *in vitro* base editing protocol. We iteratively optimized the gRNA sequences and protosspacer sequences as well as the mutagenesis protocol to further increase the mutation rate. The goals of these pilot experiments were to design an efficient encoding strategy with dCas9 and APOBEC3A and prepare a synthetic DNA template to facilitate the effectiveness of our encoding system.

Materials and Methods

Validating APOBEC activity *in vitro*

We designed and purchased a single-stranded DNA oligo from Integrated DNA Technologies (IDT) to measure the deamination rate of the APOBEC3A protein, which was purchased as part of the NEBNext® Enzymatic Methyl-seq Conversion Module kit from New England Biolabs. This 350 bp oligo was designed to base-pair with two complimentary ssDNA oligonucleotides that are complimentary to the 5' and 3' ends of the template. These double-stranded segments act as a barrier from deamination and prevent the primer-binding sites from being altered, which allows for the DNA molecules to be amplified by PCR. Therefore, only the cytosines in the unhybridized ssDNA template will be converted to uracil. This template was designed based on the CDA1 plasmid gene.[89] An aliquot of the ssDNA template and both oligos were mixed in 0.1x TAE buffer in a 1:1:1 ratio with a 50 uL volume and heated to 95°C for 5 minutes. The DNA was then cooled to 4°C with a ramp rate of 5°C per minute in a thermocycler.

The template was diluted to a concentration of 8.36 ng/uL to keep within the recommended range of DNA concentration according to the EM conversion module.

We used a commercially available NEB kit and the APOBEC3A protein, reaction buffer, and Bovine Serum Additive (BSA) provided to induce the deamination. For this pilot trial, we tested two conditions, one in which we expect mutation to occur with the presence of A3A and a negative control in which no mutation occurs. For the experimental reaction, 20 uL of the ssDNA template with a concentration of 8.36 ng/uL was mixed into a reaction consisting of 1 uL of BSA, 10 uL of 10x APOBEC reaction buffer, and 68 uL of nuclease-free water. 1 uL of APOBEC (434 ng/uL - recorded from Qubit fluorometer) was added last to bring the total volume to 100 uL per the standard protocol of the EM-seq Conversion Module from New England Biolabs. The control reaction was prepared the same method but with nuclease-free water substituted for APOBEC. Both reactions were incubated at 37°C for 3 hours in a thermocycler and heat-inactivated the APOBEC3A protein at 95°C for 2 minutes. After cooling the reactions to room temperature, the templates were amplified for 30 cycles using OneTaq polymerase with an annealing temperature of 54°C and an extension of 30 seconds. After purifying the PCR products, both samples were sequenced via Sanger Sequencing from GENEWIZ (Azenta Life Sciences).

In vitro dsDNA Mutational Activity

To test whether dCas9 and APOBEC3A can be simultaneously active in the same buffer conditions for targeted mutagenesis, we performed a targeted deamination experiment using a dCas9 RNP directed to a 20 bp sequence in plasmid pLY192, which was known to have high activity in Cas9 cleavage.[90,91] We performed two experimental reactions in which dCas9 RNP and APOBEC were present. The reactions were scaled down to preserve material, but the

concentration ratio between the RNP complex and the DNA template remained around 10:1. Therefore, the reactions were prepared: 19 uL of nuclease-free water, 3 uL of APOBEC reaction Buffer, 1 uL of BSA, 3 uL of the RNP complex, and 3 uL of the Target B fragment. For the experimental reaction, 1 uL of APOBEC (434 ng) was aliquoted and 1 uL of nuclease-free water was added for the control reaction. Both reactions were incubated at 37°C for 3 hours and left at 4°C overnight. The proteins in each reaction were denatured using 1 uL of proteinase K and incubated at 56°C for 10 minutes. After DNA purification, the templates were aliquoted and amplified using OneTaq polymerase with an annealing temperature of 50°C and an extension time of 2 minutes at 30 cycles.

Template 1.0 Mutation Trials

When we tested dCas9 and APOBEC activity, we designed a synthetic DNA "template" (later called "register") containing 12 domains, each with a unique target sequence that includes several cytosines in each domain. which was synthesized by TWIST© Bioscience. Each protospacer is separated by 40 bp indexes unique to each domain on the DNA template, as listed in Table 1. These index sequences were designed to help identify the mutated domain spacers after base calling. To prevent degradation and allow for amplification of the template, the nucleotide sequence of the template was assembled into the pBR322 vector in-between the Bsu15I and EcoRV digestion sites *via* Hi-Fi Assembly. The assembled plasmid was then transformed into NEB5-alpha bacteria and grown overnight in LB-agarose plates with Carbenicillin. Colonies with the cloned plasmid were extracted and inoculated overnight in Carbenicillin antibiotic LB media, then prepared a glycerol stock with the colony that has the inserted template.

Table 1 – Sequences of Domains and Indexes for Template 1.0

Segment	Sequence
Initiator	atcacgaggccctttcgtcttcaagaattc
Index	TTTATAGAAAACGTTTTGAAGAAGAAGATGATCTCT
Domain 1	ACTCgttctcatcgcgtaccacgaAGG
Index 1	TGTCCTACTATGTCTTCTCTCTTCTACTACTTACCT
Domain 2	ACTCggactgcttcacggtcaacgTGG
Index 2	GGATGGATGATCCCACACCTCACACGCAGGAGAGAA
Domain 3	ACTCaggtccgacgatcaccttcaTGG
Index 3	CTAGTGGTAGATGTTGTGTGTGGCGCGAGAGAAAGC
Domain 4	ACTCggaactcggagacactcgacTGG
Index 4	TTGCGACGATGACTGACGACTGCACGAAAAGCTGGA
Domain 5	ACTCgattcgaatatctctcttcgTGG
Index 5	GTGAGGAGGAGAAGTAAAAGAAAGCTTCGAGAGAGT
Domain 6	ACTCtcgggagaaaggtcgctgtgTGG
Index 6	AGTTTACACGGCGCTCTTTCCGGTTTGATCTTGCAC
Domain 7	ACTCactagtcctcgaaaacctcgTGG
Index 7	ATGTTTACGCACGCGTTTTCCCACCCACGATGTTGT
Domain 8	ACTCatcacgagttcacgataccgTGG
Index 8	CTGTTTGCACACACACCCGCACACCCTGTTCCCTCG
Domain 9	ACTCttgtggtcaatgtcactccgAGG
Index 9	ATGCGTTGCGTTGTTTTGCGTTCCACACCACACGTT
Domain 10	ACTCaagctcagcctcgttaaacgTGG
Index 10	ATCCAAAGAGAACTGGGATTTCTAAAAGAGAGAGAA
Domain 11	ACTCggcgatcacggccatcacagCGG
Index 11	GTTTTACCTTTTGCGCTTTTGTCTTCGTTCGTCCCT
Domain 12	ACTCgtattggttctcagcatcgcCGG
Index 12	GGCTCCCTACCACACACCACGTTTTGATGATAGTTG
Terminator	ACTCatcgataagctttaatgcggtagtttatca

Note. Lowercase nucleotides are involved in molecular interactions (i.e. PCR and dCas9 mutations) Uppercase nucleotides are unmodified to identify adjacent protospacers.

We synthesized sgRNA oligos complementary to each domain sequence using the commercially available enGen sgRNA synthesis kit and followed a standard protocol to express the sgRNA for each protospacer. The DNA oligo templates for each gRNA were ordered from Integrated DNA Technologies (IDT), each with a T7 RNA promotor and a sequence

complementary to the *S. pyogenes* crRNA sequence. Using the synthesized sgRNA, we formed the RNP using purified dCas9 proteins purchased from IDT by mixing the RNA and dCas9 in a 1:1 molar ratio in 1x PBS buffer and incubating the mixture at room temperature for 10 minutes. After forming the RNPs, we performed our deamination reactions by aliquoting 1.5 uL of 1 uM of RNP and 1.5 uL of 50 nM of our DNA template into master mixes of 8.5 uL of nuclease-free water, 1.5 of dCas9 buffer [200nM HEPES, 1M NaCl, 50 mM MgCl2, 1 mM EDTA, pH 6.9], 1 uL of RNAse inhibitor Murine, 0.5 uL of Bovine Serum Albumin. 0.5 uL (217 ng) of APOBEC was administered last to initiate the mutation reactions.

This process was repeated for 15 reactions, where 12 reactions were targeted by a different RNP and 3 are controls, in which APOBEC, dCas9, and both were substituted with NF H2O. The reactions were incubated at 37°C for 3 hours using a thermocycler with a heated lid at 105°C. The reactions were amplified using Q5U polymerase master mix using standard PCR settings expect we used a 69°C annealing temperature and a 45 second extension time. We use the Q5U polymerase for amplifying the mutated template, as it is a high fidelity polymerase but one that recognizes uracil as thymines during PCR. Unique barcode primer combinations were used to perform PCR for different samples and the combinations are listed in the supplementary material. After purifying each sample using AMPure XP magnetic beads, we aliquoted each sample and mixed them together in a 1:1 ratio with each reaction. The mixed samples were prepared for Oxford nanopore sequencing using SQK109 ligation kits. Using the unmodified template as a reference, we collected over 14,000 high quality reads, compared the presence of thymine at sequences in which we would expect the cytosines to be present, and measured the mutation rates across all guides and samples. The sequencing run was conducted on R9.4.1MinION Flongle flow cells from Oxford Nanopore Technologies at default settings on

MinKNOW. The fast5 raw signal files were base called using Guppy basecaller 6.1 basecalling on a laptop with Alienware m15 R4 1TB SSD with an Intel i7 10750H CPU, 16 GB of RAM and dedicated NVIDIA GeForce RTX 3060 GPU in the super high accuracy (sup) mode.[92] The generated FASTQ files were binned into pass or fail folders based on their q-scores. Only the reads that have passed the q-score threshold were analyzed. Our data was analyzed via MATLAB and the percentage of thymines where cytosines were expected. Each data set was organized via experiment #, where the first row indicates experiment 1, and so forth.

Template Optimization Reactions

Based on the results of our initial template, we altered our template design and protocol to enhance the rate of mutations. We included an exonuclease reaction in our targeted deamination experiment to remove the non-edited strand of our deaminated samples. To perform our optimization trials, we utilized the 20 bp target derived from the pYL192 plasmid used to determine the fidelity of our system. We designed and ordered primers that included a phosphate group at the nonmutated strand, making it susceptible to exonuclease digestion by lambda exonuclease. The target was digested using lambda exonuclease, an enzyme that degrades DNA sequences from 5' to 3' end directions, with a preference to phosphate groups. Therefore, we designed and ordered primers to amplify the targeted sequence with a phosphate group at the 3' end of the template, making it susceptible for exonuclease digestion. The forward 5' primer was modified to have 4 phosphothioate bonds to protect the mutated strand from degradation.

After performing the initial mutation reactions and the DNA purification step using AMPure XP magnetic beads, we elute the modified template in Nuclease-free H2O for the lambda exonuclease reaction. Then, we prepared the exonuclease reaction by adding 5 uL of 10x Lambda Exonuclease reaction buffer and 1 uL of Lambda Exonuclease, then increasing the

reaction volume to 50 uL with the addition of nuclease-free H2O and incubated the reaction inside of a thermocycler with the following settings: 37°C for 30 minutes, and then a 75°C incubation for 15 minutes to deactivate the exonuclease reaction. Then after deactivation, the ssDNA was purified and eluted to be amplified via Q5U PCR.

The mutation reactions were further optimized by testing different conditions. We determined if the amount of APOBEC3A protein used in the reaction would enhance the mutation efficiency. We performed our mutation reactions for this test following the protocol from our Template 1 experiment but focused on the amount of APOBEC used for our reactions . To test whether adding more APOBEC will increase mutational efficiency, we performed two reactions in which one reaction is performed with our normal mutation reactions. In contrast, our experimental sample includes two times the amount of APOBEC with our control. Therefore, we performed our mutation, exonuclease, and amplification reactions and sent the samples for Sanger sequencing.

Template 2.0 Design and Mutation Trials

Influenced by the results of our deamination trials and the motif affinity of APOBEC3A, designed 16 domains that are predicted to have a significant binding affinity in Cas9 interactions. The new template was designed to contain at least two TCR (R = A or G) motifs that the APOBEC3A protein will target. We designed and ordered the template dsDNA from TWIST and inserted the template into the pBR322 oligo using the protocol described in our deamination trials from Template 1. The sequences used are presented in Table 2. We performed our mutation reactions by individually targeting each of the sixteen domains by adding each RNP separately. Following our Template 1 experiment protocol, we induced mutations by adding 1 uM of the corresponding RNP to ~ 50 nM DNA template concentrations in a 1:1 volume ratio. These

experiments were performed with replicates and negative controls (no dCas9 or APOBEC), in which little mutation was expected to occur. After performing the mutation and post-experiment treatments, we amplified each experimental condition using primers with unique barcode overlaps to identify the conditions for each template. The results of these deamination experiments were sequenced via Oxford nanopore sequencing and base called via Guppy base calling protocols. We repeated this experiment to confirm whether the individual mutations did not affect adjacent mutations in our results, following the same conditions.

Table 2 – Sequences used for Template 2.0

Segment	Sequence
Initiator	atcacgaggcctttcgtcttcaagaattc
Index	TTTATAGAAAACGTTTTGAAGAAGAAGATGATCTCT
Domain 1	ACTCtcgccagatcgacaggatcaTGG
Index 1	TGTCCTACTATGTCTTCTCTCTTCTACTACTTACCT
Domain 2	ACTCctccaatcaaatcagtcactAGG
Index 2	GGATGGATGATCCCACACCTCACACGCAGGAGAGAA
Domain 3	ACTCtctggtcagggctcggacacTGG
Index 3	CTAGTGGTAGATGTTGTGTGTGGCGCGAGAGAAAGC
Domain 4	ACTCtcattcacagcaactgcagcAGG
Index 4	TTGCGACGATGACTGACGACTGCACGAAAAGCTGGA
Domain 5	ACTCatggtcaactcaatccaaaaTGG
Index 5	GTGAGGAGGAGAAGTAAAAGAAAGCTTCGAGAGAGT
Domain 6	ACTCgttctcatcgcgtaccacgaAGG
Index 6	AGTTTACACGGCGCTCTTTCCGGTTTGATCTTGCAC
Domain 7	ACTCatcaatagtgtcatggcatgTGG
Index 7	ATGTTTACGCACGCGTTTTCCCACCCACGATGTTGT
Domain 8	ACTCtcgggagaaaggtcgctgtgAGG
Index 8	CTGTTTGCACACACACCCGCACACCCTGTTCCCTCG
Domain 9	ACTCatcacgagttcacgataccgTGG
Index 9	ATGCGTTGCGTTGTTTTGCGTTCCACACCACACGTT
Domain 10	ACTCttgtggtcaatgtcactccgAGG
Index 10	ATCCAAAGAGAACTGGGATTTCTAAAAGAGAGAGAA
Domain 11	ACTCaagctcagcctcgttaaacgTGG
Index 11	GTTTTACCTTTTGCGCTTTTGTCTTCGTTCGTCCCT
Domain 12	ACTCgaacagatcatcaacccattAGG
Index 12	GGCTCCCTACCACACACCACGTTTTGATGATAGTTG

Domain 13	ACTCattcaatcaagctgcaaaggTGG
Index 13	TACGAGAGGAAGCTTCACACACCACCACGATCGGAT
Domain 14	ACTCctttcaagacctcaagaacgAGG
Index 14	CTTGCGCACACCTCACACACGTGTTTGTGTTGTGTT
Domain 15	ACTCgcctcatcagcagaacaagtTGG
Index 15	CGATCCGCACACGCACGTCACACCTATCTTACGTGT
Domain 16	ACTCtcattccagtcaatgtggaaAGG
Index 16	GAAGAAAAGAAAGAGAAGAGAAAACTCAAAAGATGA
Terminator	ACTCatcgataagctttaatgcggtagtttatca

Note. Lowercase nucleotides are involved in molecular interactions. (i.e. PCR and dCas9 mutations) Initiator and Terminator sequences are sites for primers to amplify the template. Indexes are included to identify initial domains after nanopore sequencing. Adjacent indexes are exclusive to each domain for our experiments.

To test whether our selected targets inhibit the activity of adjacent targets, we performed another experiment in which we target all 16 domains at once by introducing all 16 RNPS to measure the activity of simultaneous deamination on a single template. The deamination, single-strand digestion, and amplification steps were performed for this experiment as before. A mixture of all sixteen guides was prepared using 1 uL of each guide, diluted up to 20 uL with 1x PBS buffer, and administered to each experimental reaction. This reduced the concentration of all 16 guides in the mixture to 50 nM for each guide, in which we diluted the DNA template used to 5 nM to keep the ratio of Cas9 and DNA template within 10:1 ratio. We have also included a mutation reaction using a 20:1 ratio of RNP:DNA to determine if any visible improvements were administered.

Template 3.0 Design

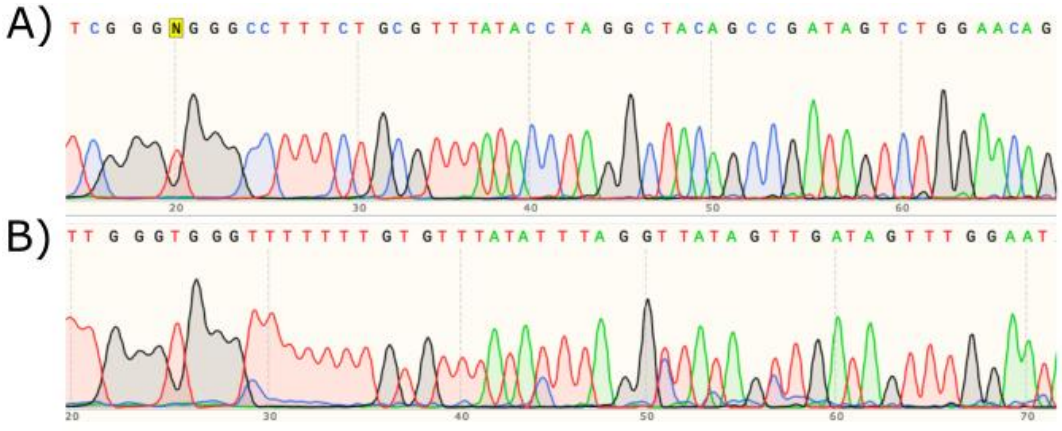
We designed and ordered a new 16 domain template with the most promising nucleotide spacers from Templates 1.0 and 2.0. In this design, we kept every domain from Template 2 but only replaced domains 1 and 14 with Template 1's domains 7 and 5, respectively. Using this

template, we performed pilot trials comparing our new template with our previous prototype using an all targeting dcas9 strategy. We mixed each dCas9 RNP in a 50 nM concentration. As such, we have had results from both templates 1 and 2 that demonstrate the potential deamination capacity of each domain. Therefore, in the next chapter, we applied this design as the final template to demonstrate CRISPR digital data encoding.

Experimental Results and Discussion

From our initial experiments with single-stranded DNA exposed to APOBEC *in vitro*, we found that when comparing our negative control and experimental samples, we found nearly all exposed cytosine express mutation rates from 59% to 100% shown in Figure 7. Therefore, we conclude that we can perform *in vitro* deamination with nucleotide modifying enzymes utilizing APOBEC3A outside of standard operating procedure of the kit. However, this high mutation rate was observed on a sample that was single-stranded. To control where our mutations occur in a double-stranded DNA molecule, we used a catalytically inactive Cas9 mutant (dCas9) to generate localized R-loops and direct where the deamination will occur.

Figure 7 – APOBEC Mutation on ssDNA Template



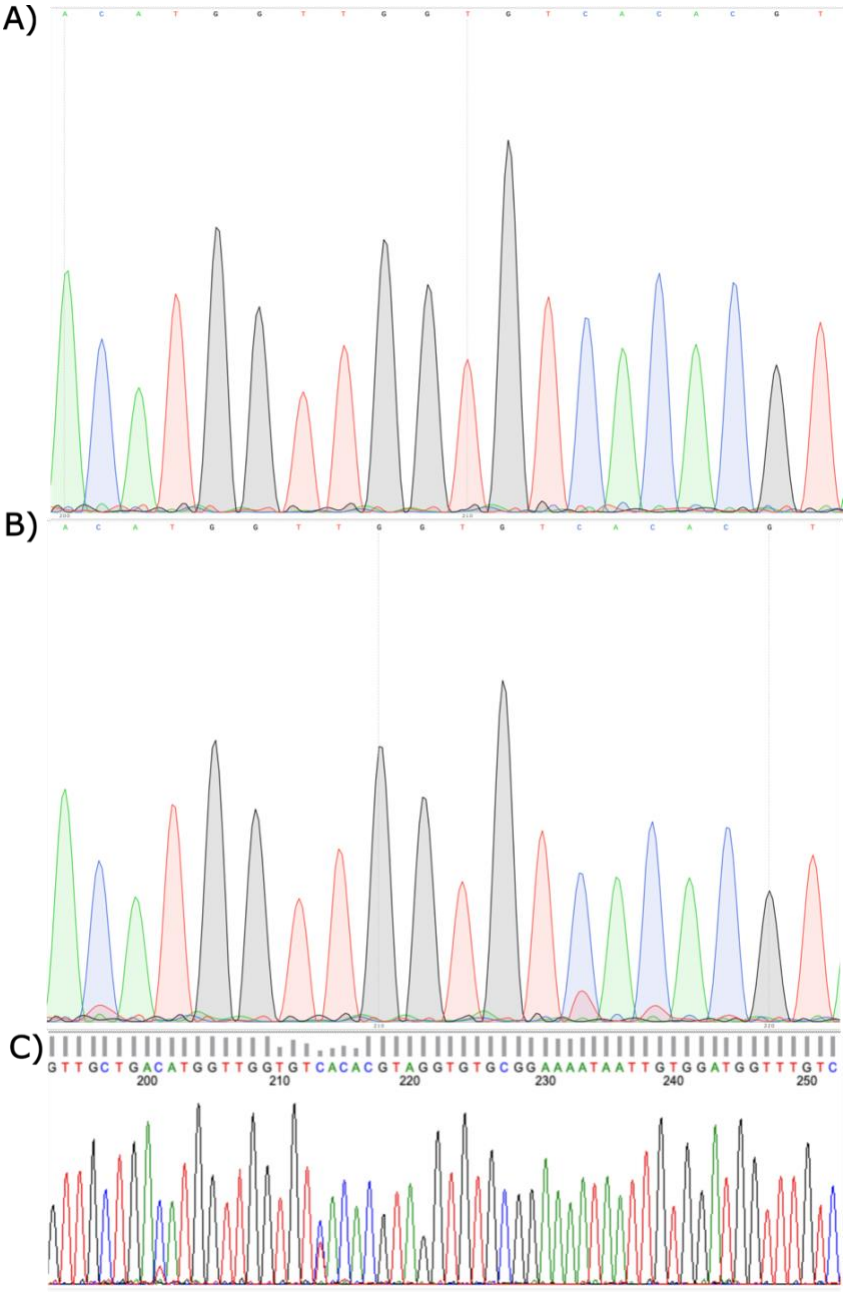
Note. A) Negative control from our ssDNA sample. We observed that the signals presented the complete sequence of our ssDNA oligo, indicating that no modifications have

occurred. We used this sequence to compare our experimental control. B) The ssDNA template was exposed to APOBEC3A in this reaction, in which we should expect to see high levels of thymine signals present where we expect cytosines in our unaltered template. Sanger Sequencing was performed by GENEWIZ (Azenta Life Sciences)

As previously discussed, dCas9 hybridization will form an R-loop at sequences complementary to the binding site of the sgRNA spacer. Therefore, the nucleotides that have been displaced by the RNA base-pairing are expected to be susceptible to deamination because they have a single-stranded conformation recognizable to the APOBEC3A protein. To test this, we attempted to perform targeted deamination onto a dsDNA molecule with a protospacer known to be readily recognized by Cas9 with its gRNA (plasmid YL192). For that target, we have observed increased presence of thymine at locations of cytosines but only in the protospacer region in exposed regions. In future optimization reactions using this template, we confirmed the absence of mutations "off-target" or not within the R-loop, which can also be seen in Figure 8C. This indicates an experimental correlation between the binding of complementary RNP and the mutation of sequences matching the sgRNA targeted areas. This result demonstrated that we could simultaneously reconstitute dCas9 and APOBEC3A protein activity to introduce targeted mutations within a dsDNA molecule *in vitro*.

To create a consistent protocol for our mutations, we utilized Cas9 buffer with a 6.9 pH level for our later experiments. Since the dCas9 RNP determines the site of mutation on our templates, we found that balancing the pH level of the buffer closer to the dCas9 conditions allowed us to observe significant mutation levels in our targeted regions. Therefore, we conclude that the presence of dCas9 with a gRNA complementary to a targeted site allows us to introduce targeted mutations *via* APOBEC3A in slightly acidic to neutral buffer conditions *in vitro*.

Figure 8 – dCas9/APOBEC Mutations in a dsDNA Target with Protospacer

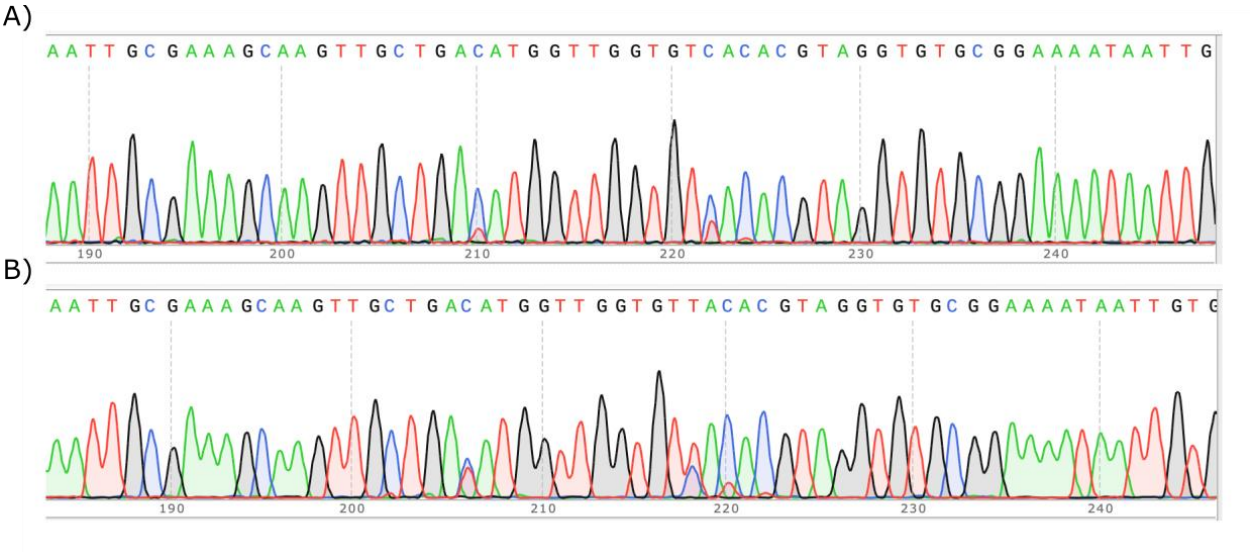


Note. A) Control target mutation in which APOBEC is not present. We have observed no alterations of B) Target nucleotides exposed to APOBEC in presence of dCas9. C) Mutation reactions including nucleotides outside of targeted domain, in which we observe no off-target mutations outside of the targeted region. This reaction was from an optimized mutation reaction.

Based on our findings, we conclude that the presence of dCas9 with the complementary sgRNA sequence will direct mutation of APOBEC3A when both are applied to our mutation reactions.

By increasing the amount of APOBEC used in our reactions, we observed a correlated increase in mutation rate, as shown in Figure 9. Since we have scaled the reaction volume down from 100 uL to 15.5 uL, we hypothesized that increasing the aliquot of APOBEC3A from 217 ng to 434 ng would increase mutation reactions due to the smaller amount of volume the protein is suspended in. As we doubled the amount of APOBEC in our experiment, the mutation rate also doubled. This observation was significant in modeling our encoding strategy, as we observed higher mutations adjacent to the PAM site in this experiment. As such, we increased the amount of APOBEC used in our experiment to 1 uL, or 434 ng.

Figure 9 – Effect of APOBEC Concentration on C>T Mutation Rates

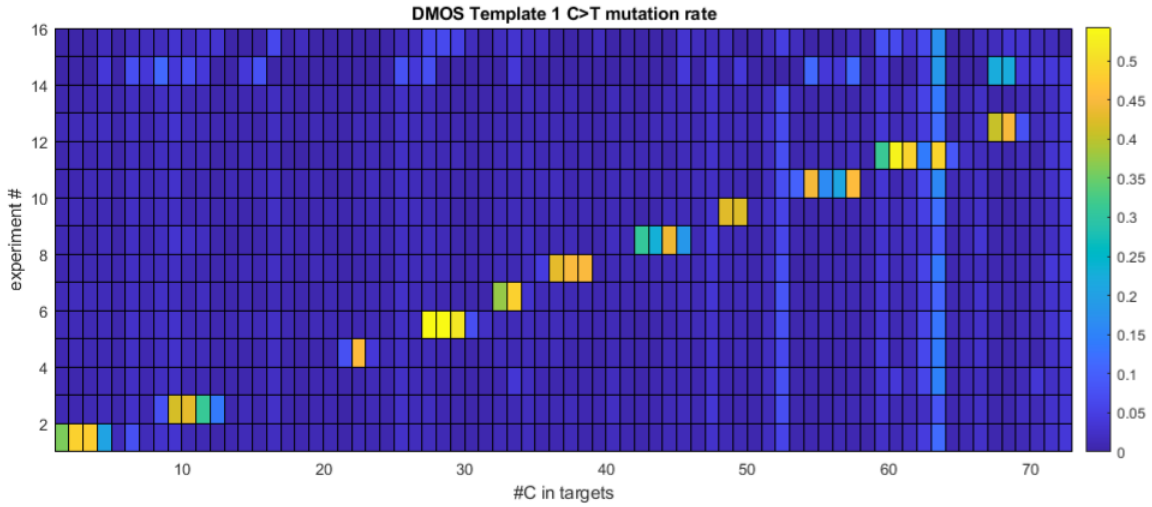


Note. A) Targeted sequence of TRV1 template was exposed of the standard concentration of APOBEC3A and analyzed across the expected sequence. The amount of APOBEC included in this reaction was 217 ng. B) The mutation protocol was altered to include 434 ng the amount of APOBEC from standard protocol and measured across the targeted sequence. We found that

increasing the concentration of APOBEC applied to our mutation reaction does increase the mutation rate of our reaction.

Having determined the conditions necessary to introduce targeted mutations in dsDNA *in vitro*, we sought to apply this approach to a 856 bp DNA molecule that contained 12 individually addressable protospacers. In the context of digital data storage, we referred to each protospacer as "domains" in our DNA templates. These results are presented in Figure 10, where each square indicates the position of the cytosines in the DNA molecule and the color represents the fraction of observed thymines at the positions of those cytosines from nanopore sequencing, for each of the 12 different gRNAs targeting their complementary domains. Dark blue sections indicate little to no mutation, where yellow squares indicate significant fraction of observed thymines at sites where cytosines were present in the original sequence. We have detected significant levels of mutation across most of our domains except for domain 3, which we expect to have been due to a missing component within that reaction or poor dCas9 binding. However, we found that our mutation strategy confers an "all-or-nothing" response, in which mutations will either occur across most of the cytosines in the targeted protospacer or not. We have also observed that our mutations were at their highest in cytosines within TCR (R=A or G) motifs, consistent with APOBEC3A mutation patterns,[93] and that cytosines less than 6 bp away from the protospacer adjacent motif (PAM) exhibited significantly lower mutation rates than the other affected cytosines. These observations influenced our design choice for the domain sequences so that going forward they would include at least two TCR motifs in each domain sequence to further enhance mutation efficiency in our reactions at least 6 bp away from the PAM site.

Figure 10 – Mutation Rate of Template 1.0

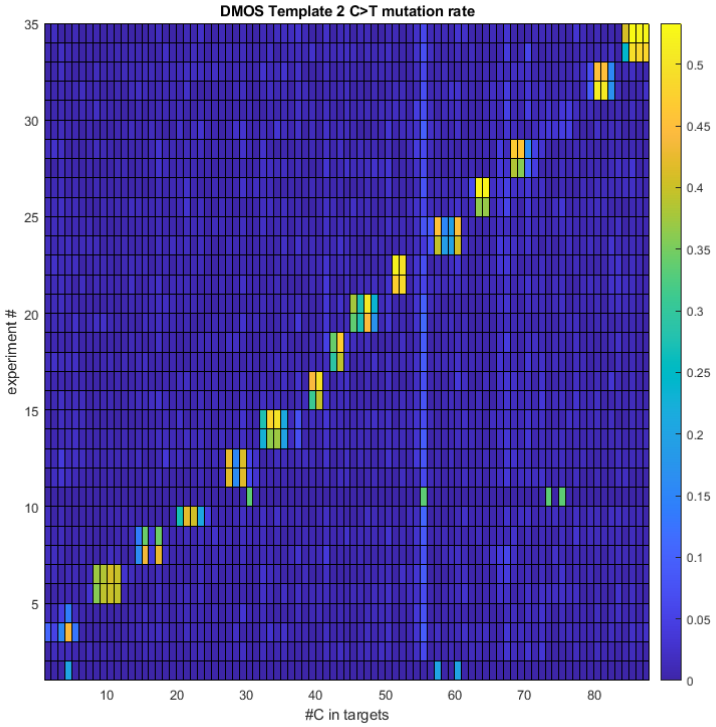


Note. Experiments 1 through 12 indicate mutation reactions that target only one domain across the entire template, i.e. experiment 1 indicates domain 1 and so forth. Experiments 13 through 15 are negative controls, omitting dCas9, APOBEC, and both respectively from the mutation reaction to distinguish mutations caused by APOBEC. The bitmap was generated using MATLAB code to detect the presence of thymine after guppy base calling and alignment. We find that we detected the presence of thymine in our targeted regions across each domain except for experiment 3.

We designed a second iteration of our DNA molecule that now included 16 domains (protospacers) with the above-stated design criteria (Template 2.0). From the mutation rate data presented in Figure 11 we observed significant mutation rates between 15% and 50% across all but two domains in our second template design. While this experiment was replicated to ensure the conclusions of our results are accurate, we find that guides 1 and 14 consistently exhibited low mutation rates and were later replaced. This was also expressed from our analysis in Figure 12, Therefore, we sought to replace these two domains in our third DNA design (Template 3.0)

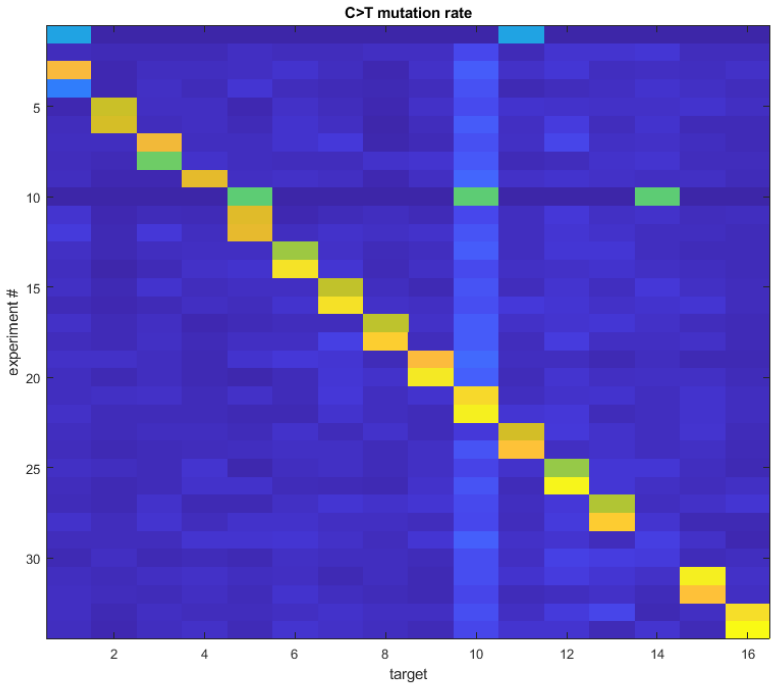
with protospacers 7 and 5 from Template 1.0, since they expressed high mutation rates from our Template 1 experiments.

Figure 11 – Mutation Rate of Template 2.0



Note. Experiment number indicates the conditions of each template. Each reaction was performed with a replicate and demultiplexed after nanopore sequencing using barcoded primers. Experiment 1 and 2 were negative controls where no APOBEC proteins were applied. Reactions were numbered to indicate targeted region, *i.e.* experiments 3 and 4 indicate protospacer/domain 1 was targeted by included a CRISPR RNP with the corresponding gRNA, experiments 5 and 6 show protospacer 2 was targeted and so forth. Each cytosine mutation was recorded and cross-referenced with the unmodified Template 2.0 to indicate percent of C > T mutations.

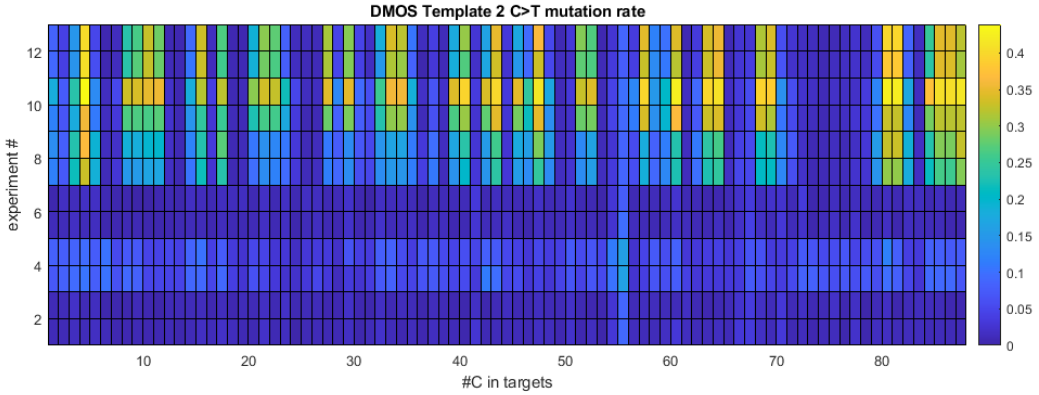
Figure 12 – Replicate of Individual Targeting of Template 2.0



Note. Experiments 1 and 2 were of negative controls, where no APOBEC and dCas9 were included in the mutation reaction. High background signal was present for protospacer 10 across all experiments. Each experiment and their replicate indicate a mutation at each protospacer target and numbered respectively. However, experiments 1 and 10 only had 3 and 5 reads respectively, while the other experiments expressed hundreds of reads each. Color scale was presented from lowest 0% to highest 53%, from blue to yellow. We also find that protospacer 14 expressed no mutations. We concluded that we saw that each mutation occurs individually and does not interfere with adjacent protospacers.

When we performed mutation reactions across all guides of template 2, we observed in Figure 13 that each mutation occurs independently and does not affect the mutation rate of adjacent domains. However, when we increased the ratio of dCas9 to the DNA template, we found no significant difference between 10:1 and 20:1 ratios. With this information, we determined that the encoding mechanism can maintain efficiency between these ratios, but further experiments would be needed to determine their full limitations.

Figure 13 – Mutation Rate across all Domains in Template 2.0



Note. Experiments 1 and 2 indicate negative control with no dCas9 or APOBEC. Experiments 3 and 4 indicate negative control with no dCas9. Experiments 5 and 6 indicate negative control with no APOBEC present. Experiments 7 – 10 were of Template 2 mutations across all 16 spacers with a 10:1 concentration ratio of RNP:DNA. Experiment 11 and 12 used a 20:1 concentration ratio of RNP:DNA. Mutation rates were determined based on presence of thymines in reads relative to with the unmutated template.

We realized there was a fundamental limit to our mutation rate since we were only mutating one strand of DNA. To increase our mutation rates from the *in vitro* reaction, we decided to degrade the unmutated strand prior to PCR amplification and sequencing. We did this by adding an additional exonuclease into our mutation protocol after APOBEC treatment.

Lambda exonuclease degrades DNA 5'- to 3' and with significantly enhanced rates at oligonucleotides with 5'- phosphates, and significantly lower rates at oligonucleotides with phosphorothioate bonds.[94] So, we modified the DNA template so that the strand mutated would have 5'- hydroxyl groups and a series of phosophorothioate bonds, while the unmutated strands had a 5'- phosphate. For the third iteration of our DNA molecule (later called a register in the context of data storage), we have designed our final register to exhibit high mutation rates through all 16 guides, and hypothesized that that introducing a phosphate group at the 5'- end of the unmutated strand would increase the mutational efficiency of our CRISPR system by removing the unmutated strand.

Influenced by our initial results, we sought to determine if our phosphorylated templates had a significant effect on the mutation rate of our samples. To present the effect that applying lambda exonuclease does in our experiment, we prepared both phosphorylated and unphosphorylated primers to amplify our template to showcase how removing the nonmutated strand enables modified PCR templates to populate the reaction for PCR. As such, we included equal molar amounts of both template forms to our mutation reactions to ensure the populations do not affect the PCR product results.

Thus, only the ssDNA mutated strand would be present for further amplification. Based on the sequence results from Figure 14, we conclude that introducing the lambda exonuclease into our mutation reactions significantly improves mutational efficiency compared to our controls by about two-fold, as expected. However, this step requires a nucleotide purification step after each reaction, as the entire template will be susceptible to mutation via APOBEC3A While the lambda exonuclease will degrade the nonmutated strand, the remaining strand will be vulnerable to APOBEC mutation from residual active proteins. Despite these added steps, we

determined that our encoding system allows for *in vitro* cytosine-to-thymine mutations that can be controlled and localized by introducing specific dCas9 RNPs.

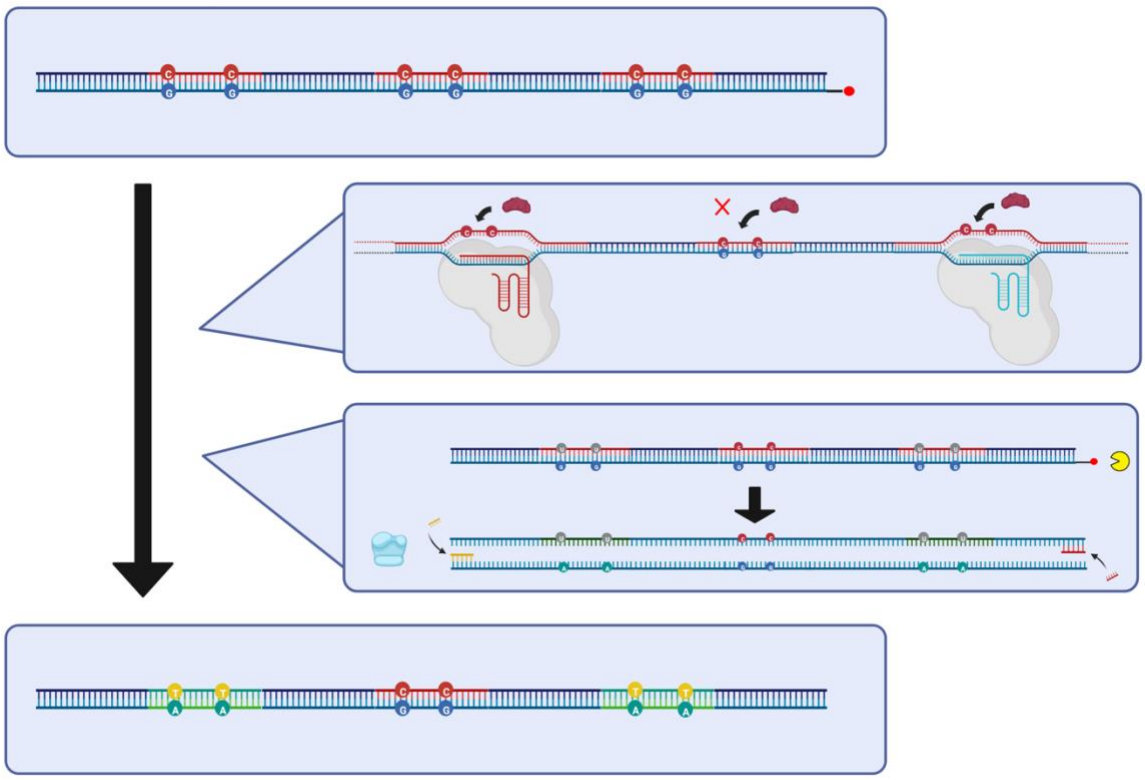
Figure 14 – Mutation Rate of Template 3.0 with and without Lambda Exonuclease

Note. Experiments 1 and 2 consist of reactions in which APOBEC is not present. Experiments 3 and 4 consist of reactions without dCas9. Experiments 5 and 6 indicate mutations on a template that does not have a phosphorylated 3' end, indicating no presence of exonuclease digestion when lambda exonuclease was applied to the template. Experiments 7 and 8 indicate the same conditions, but with a phosphorylated 3' end. Experiment 9 contains an unaltered template for reference, in which no APOBEC was applied to the reaction, to distinguish the mutated targets from nonmutated targets. We found that mutations significantly increased in templates that are phosphorylated, but we also noticed the necessity of performing a cleanup step after each reaction from experiments 3 and 4. We concluded that residual APOBEC in presence of the digested template can slightly induce mutation across the entire template, which can ruin our encoding strategy unless we purify the template before each reaction.

Conclusion

In summation, we demonstrated that dCas9 and APOBEC3A are active in slightly acidic buffer conditions and can induce controlled and localized mutations only when a dCas9 RNP containing a specific sgRNA is present. We have also optimized our DNA template and identified protospacers that exhibit strong recognition by their corresponding dCas9 RNP and by including motifs best recognized by the APOBEC3A protein at positions in the protospacer where they are most susceptible to mutation. We also developed and improved our mutation protocol through catalytic removal of the unmutated strand, for example. We designed and applied a nucleotide template with 16 targets that can each be utilized for digital data encoding using this CRISPR method presented in Figure 15.

Figure 15 – CRISPR Mutation Protocol



Note. The CRISPR encoding mechanism has a DNA molecule that is 5’ phosphorylated on only one end. The mutations were performed using dCas9 proteins to form the R-loop to initiate mutation by the base editor. The unedited strand is degraded via lambda exonuclease, as shown by the yellow enzymatic symbol, and amplified by Q5U polymerase, shown as the blue polymerase icon, to achieve the modified template. The final template is then sequenced using nanopore sequencing then assigned bit values of ‘0’ if its sequence is consistent with an unmutated domain or ‘1’ if its sequence contains a mutational signature of dCas9/APOBEC mutations. (Created with BioRender.com)

CHAPTER III: APPLYING CRISPR BASE EDITORS FOR DIGITAL DATA STORAGE ON DNA TAPE

Introduction

After preparing our enzymatic mutation system, as mentioned above we sought to use this mutational system—taking a cue from the semiconductor industry where digital data can be recorded on-demand as bits in state '0' or '1' at precise locations in mass-produced, blank disks/tapes rather than being hard-coded into the storage medium at creation—to rewrite digital data onto prefabricated DNA molecules. This chapter describes an application of our Template 3.0 design and mutation strategy to show that digital data can be effectively written at precise locations and read from pre-synthesized, 'blank' DNA molecules that we also refer to as DNA "tapes" or templates. This molecular storage system, called "DNA Mutational Overwriting Storage" (DMOS), does not require DNA synthesis outside of the synthesis of our unmodified DNA template and primers and can be performed *in vitro* allowing for precise control over biochemical conditions.

In the DMOS system, we introduced mutations at precise nucleotide regions using the CRISPR base-editing reactions. We assigned the bit value of 0 or 1 depending on the mutation state of each region on our templates. CRISPR proteins can recognize nearly any sequence of interest based on the spacer nucleotides of their gRNA co-factor and directly alter the nucleobase structures at the targeted domain in combination with the APOBEC mutagenic protein. We reconstituted this reaction *in vitro* and introduced different combinations of 16 gRNA co-factor into the base-editing mechanism to encode multiple bits into the DNA template at once.

To demonstrate CRISPR's capacity to encode digital information, we applied the design of Template 3.0, in which we tested to have high mutational efficiency in all 16 protospacers

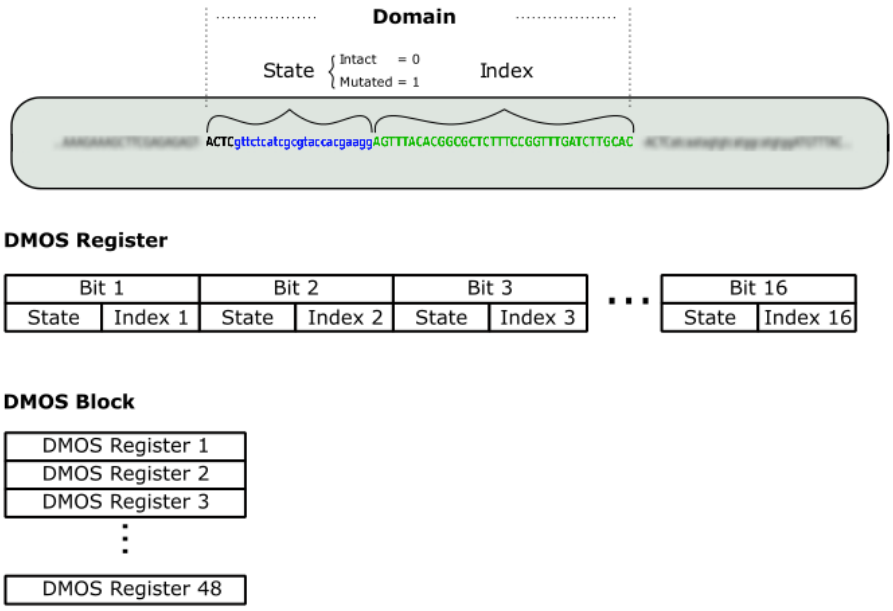
from our mutation tests in the previous chapter, and will be addressed as the DMOS template or a "register" of DNA bits. Based on the mutagenic states of their sequence we can determine whether they are in a state of 0's (unmutated, no CRISPR gRNA for this sequence added to the reaction) or 1's (mutated because a CRISPR RNP targeting that region was added) that are called based on changes of nucleotide sequence after deamination. We successfully recovered the mutational signature of the digital data encoded onto different combinations of domain arrangements ("registers") using long-read (nanopore) next-generation sequencing. Using this approach, we could encode >1250 bits across multiple DNA registers using the same 16 gRNA co-factors in parallel. By incorporating error-correcting strategies, we could recover messages with 100% accuracy encoded by the DMOS technique. With the same 16 gRNAs, our DMOS system could theoretically scale up to ~2.09×10^{13} (16!) different DMOS registers, and potentially store up to 38.06 TB of data.

Materials and Methods

Design of the Data Bit

Each bit recorded in our experiment consist of two sequences that make up each domain: the 'state' segment, or a protospacer that can be mutated, and the 'index.' As the bit is decoded, the bit is designated as either a 0 or a 1, depending on the condition of the state (whether or not it was mutated). The bit is classified as a 1 if mutations are present in the state sequence. To identify the location of the state sequence, each domain is assigned a unique index sequence that is always adjacent with the state. Each domain is considered an individual bit and can be arranged to form unique registers, sequences of bits that can be identified based on the positions the domains are arranged. The schematic view of how each data bit is arranged is shown in Figure 16.

Figure 16 – Data Bit Scheme



Note. Each data bit contains a spacer region in either a mutated or nonmutated state classified as either 0 or 1. The domain index acts as a marker to identify domains from sequencing results. Each sequence of bits is called a register in this experiment, while the collection of unique registers used is classified as a "block," where the entirety of the file is stored. This organization of protospacers and nucleotides make up the form of data we want to decode.

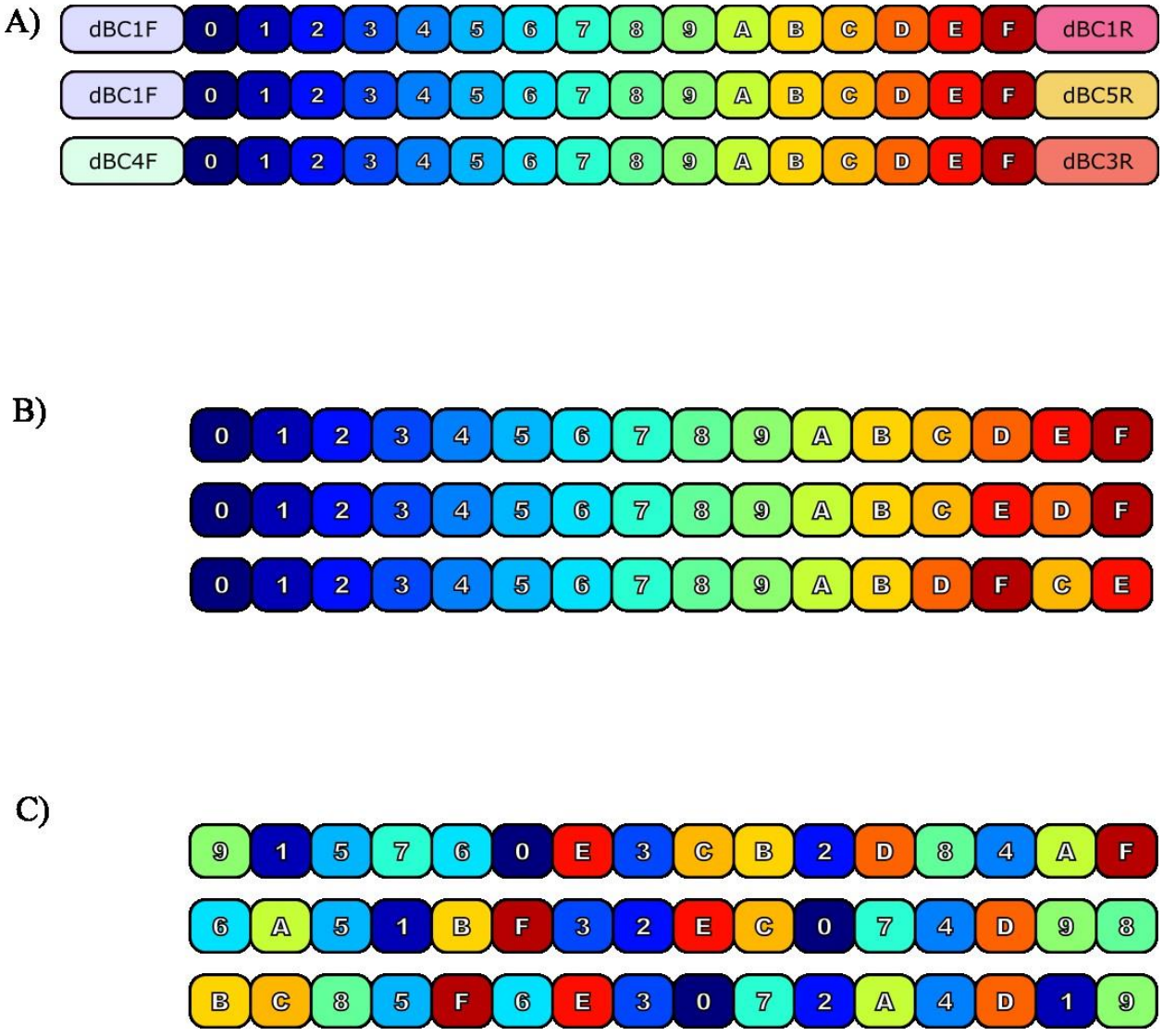
Design of DMOS registers

We defined a lexicographic representation of the order of domains/protospacers, where we map the domain positions as B0 = 0, B1 = 1, and up to B15=F using the notation of hexadecimal numerals. For instance, registers can be represented by 0123456789*ABCDEF* and 0123456789*ABCDFE*, respectively. The maximum theoretical number of combinations is ~2.09 $x10^{13}$, or 16 factorial (16!) because in a register each domain appears only once. Therefore, the DMOS encoding system can utilize the combinations of addresses to encode more digital

information using the same domain sequences. Following this scheme, we shuffled the last six domains of the registers. This shuffling enables minimum variations between the order of consecutive block addresses, but can lead to misaddressing during sequencing, which can be prone to errors. Utilizing a more randomized shuffling scheme increases variations among our registers and can increase accuracy of addressing. When permuted with greater randomization, it becomes less likely we will assign a register to one with another also that is also present during sequencing.

To enhance the variability of domain shuffling and distinguish our registers during nanopore sequencing, we applied a PRESENT S-box high-distance shuffling mechanism.[95] This cipher scheme generates high-distance permutations that ensure consecutive addresses are uncorrelated and as separable as possible. We designed a library of 32 DMOS registers in which the order of domains shuffled across the entire registers have high variability. We hypothesized that the higher address distance reduces the noise and increases the confidence in our measurements. Here, we defined one DMOS block as a blank drive containing 32 DMOS registers that are each permutated and addressing using the scheme derived from the PRESENT S-box. In other experiments, some DMOS registers used the standard permutation addressing scheme, which we later found resulted in higher rates of misaddressing. The permutation schemes used in our experiment are presented in Figure 17. For instance, register addresses 1 and 2 are represented by 915760E3CB2D84AF and 1E60D87BC932AF45, respectively (Figure 17C).

Figure 17 – Permutation Strategies for DMOS



Note. A) Registers were distinguished through unique barcode combinations *via* PCR of barcoded primers. B) Low-distance permutation, where domains were shuffled in low distances below six domains from each other in lexicographic order. C) High-distance permutations, where domains were shuffled across the entire register. These shuffling schemes enable the use of the same 16 domains to be used for our encoding strategy without including new spacer sequences.

Software development

Our DMOS encoder was developed using Python language and the Spyder IDE. The error-correcting layer uses the Protograph LDPC library (https://github.com/shubhamchandak94/ProtographLDPC).[96] To design our LDPC code, we selected the Protograph type AR4JA (accumulative repeat by four jagged accumulate) to define the Generator and Parity Check matrices, with a message-code ratio of 3/4, expansion factor 96.[97] We developed a Python interface to communicate with the LDPC library that allows the conversion of the intermediate binary files for input/output and captures the diagnostic signals of the LDPC decoder.

The DMOS software layer uses two main modules to retrieve the binary file: DMOS decoder and LDPC decoder. The DMOS decoder was written in C++ using the QtCreator IDE and uses the Smith-Waterman algorithm (https://github.com/mbreese/swalign) to align DNA sequences.[98] For the first pass of assignment, if a domain had a mutation rate above a threshold determined from the training data generated during the characterization of mutation reactions of each domain in the DMOS template or when either all domains or no domains were mutated, we calculated the threshold values used by the first Bayesian step through machine learning using our individual target mutations for each domain on the DMOS Template. The threshold values are listed in Table 3, and were used to determine the mutation sate of each domain across our registers. The threshold values were determined by extensive experiments performed during this chapter and the chapter before to determine “mutational signatures” of the cytosines in each domain when exposed to APOBEC with and without dCas9 binding.

Table 3 – Threshold Values for First Classifier[99]

Domain	Unmutated (t_u)	Mutated (t_m)
1	0.077050	0.144550
2	0.129758	0.178004
3	0.071650	0.151862
4	0.113244	0.168885
5	0.082900	0.166900
6	0.044773	0.139677
7	0.070100	0.157850
8	0.042900	0.094650
9	0.118000	0.178000
10	0.024487	0.102552
11	0.092800	0.160550
12	0.080100	0.147600
13	0.080100	0.147600
14	0.068250	0.135750
15	0.092200	0.177700
16	0.192950	0.288200

Note. Calculated values determining the mutation state of our input templates. This data was collected from previous mutation experiments to predict mutation activity. Mutation conversion values below the unmutated (t_u) thresholds are classified as unmutated, while the values above the mutated (t_m) threshold are classified as mutated. Values inbetween both thresholds are classified as uncalled.

The LDPC decoder uses a Parity-check Matrix to verify the validity of the codeword and the Generator Matrix to perform the inverse transformation and decode the message. We designed a DMOS encoding algorithm based on its state-specific coding maps (raw file) and the protected file with error correction codes. Our encoder processes the raw sequencing data and maps the sequencing data then binary data into each of the DMOS block by creating groups of

16 bits and associates them with the corresponding DMOS register. The encoder generates a file with the list of mutations addressed on the DMOS registers at the desired locations.

DMOS decoder design

We used a classification model to identify the DMOS register addresses and mutational states accordingly from a given set of nanopore sequencing reads. The decoder was developed and written in C++. We first use the Smith-Waterman align algorithm to perform the alignment.[98] Next, the decoder takes the FASTA sequences of the domains B_0, B_1, …, B_{15}, and determines (p_n,s_n, CIGAR) = swalign(B_n, Seq). Where "p" is defined as positions, "s" as score, "n" as the number of positions [0,1,2,...,15], "Bn" is the nucleotide sequence of the n-th domain, and "Seq" is a single nucleotide sequence obtained from the nanopore. Therefore, the algorithm generates a list of positions; [p_0, p_1, …, p_{15}], and sorts them to determine the order of the domains in the DMOS register.

Moreover, the decoder uses the CIGAR report (Compact Idiosyncratic Gapped Alignment Report) to generate a string of nucleotides aligned with the cytosines of the bit's states in the FASTA reference. Using maximum likelihood, the decoder determines each domain's position and alignment score on the sequenced strand. Using this domain order, the decoder generates the lexicographic representation of the address and converts it to an integer number. A valid address number should correspond to the addresses of our DMOS register library. If the address is valid, the decoder determines the values of the mutational states. The DMOS decoder then determines the mutation state using two Bayesian steps in a classification decision tree.

The DMOS decoder employs two Bayesian classifiers to determine the bit state of each domain across our registers.[100] The first Bayesian classifier counts all thymines and cytosines in

the nucleotide sequence of the domain's states, accumulates these values for all reads of the same states, and uses these values to calculate the C>T ratio as:

$$\mathrm{C} > \mathrm{T} = \sum(Tm)/(\sum(Tu) + \sum(Cu)) \quad (1)$$

The sigma symbol in equation 1 indicates the individual summation of cytosines and thymines called after base-calling. The population of thymines after mutation were designated at Tm, while thymines and cytosines present before mutation were measured as (T_u) and (C_u) respectively. We used our conversion ratios to show the likelihood of the states to be recognized as either intact or mutated. In situations in which the mutation rates were ambiguous (could be assigned with close to equal probability to being either mutated or unmutated), we designated those values as uncertain regions, in which the mutation state could not be determined from our first classifier. Therefore, we applied the threshold values to determine whether our second classifier is required using the following equations:

$$\text{Unmutated:} \quad 0 < (C > T) < t_u, \qquad CS = -10 \quad (2)$$

$$\text{Uncertain:} \quad t_u < (C > T) < t_m, \quad CS = \text{Use second classifier} \quad (3)$$

$$\text{Mutated:} \quad t_m < (C > T), \qquad CS = 10 \quad (4)$$

Where CS is the classification score (CS). If the first Bayesian classifier determines the state as uncertain, it would be addressed to the next Bayesian classifier. Otherwise, it will assign the CS values as -10 for unmutated (0) and 10 to mutated (1).

Next, we performed a classification training process for each combination of the nucleotides of all domain's state accordingly to the expected mutational state or mutational signature. As a result, we generated a library containing the score values per each combination to classify the domain's state as either unmutated or mutated. The second Bayesian classifier obtains the scores from the library using the nucleotide string from the CIGAR score:

$$(S_u, S_m) = CheckScores(individual\ position), \quad (5)$$

where S_u= intact score, S_m= mutated score, and iteratively adds the values to the classification score (CS), as follows:

$$uFrac_i = (S_u/(S_u + S_m)), \quad (6)$$

$$mFrac_i = (S_m/(S_u + S_m)), \quad (7)$$

$$CS = \sum_{n=i}^{k} \big(log(mFrac_i) - log(uFrac_i)\big) \quad (8)$$

The CS values are in the range [-10, 10]. CS values outside of the range are called either unmutated or mutated whether it is below -10 or above 10 respectively. The second classifier determines the mutational state according to:

Unmutated: CS < 0 (9)

Mutated: CS >= 0 (10)

If the CS values are between -5 and 5, the domains are also classified as uncertain and will need to be assigned the correct bit *via* LDPC error correction to confidently retrieve the message from the uncertain bit.

Uncertain: -5 < CS < 5 (11)

To transform the base calls into binary, the decoder uses the CS values and determines binary 0 and 1 for negative and positive CS values, regardless of whether they are within the "uncertain" range. The sequencing results compare the number of uncertain calls (UC) to the

total calls (TC) used in this experiment. The confidence value (CV) is calculated using the combination of the confidence obtained from the two Bayesian classifiers, as follows:

$$CBt = (tC-UCt)/TC \quad (12)$$

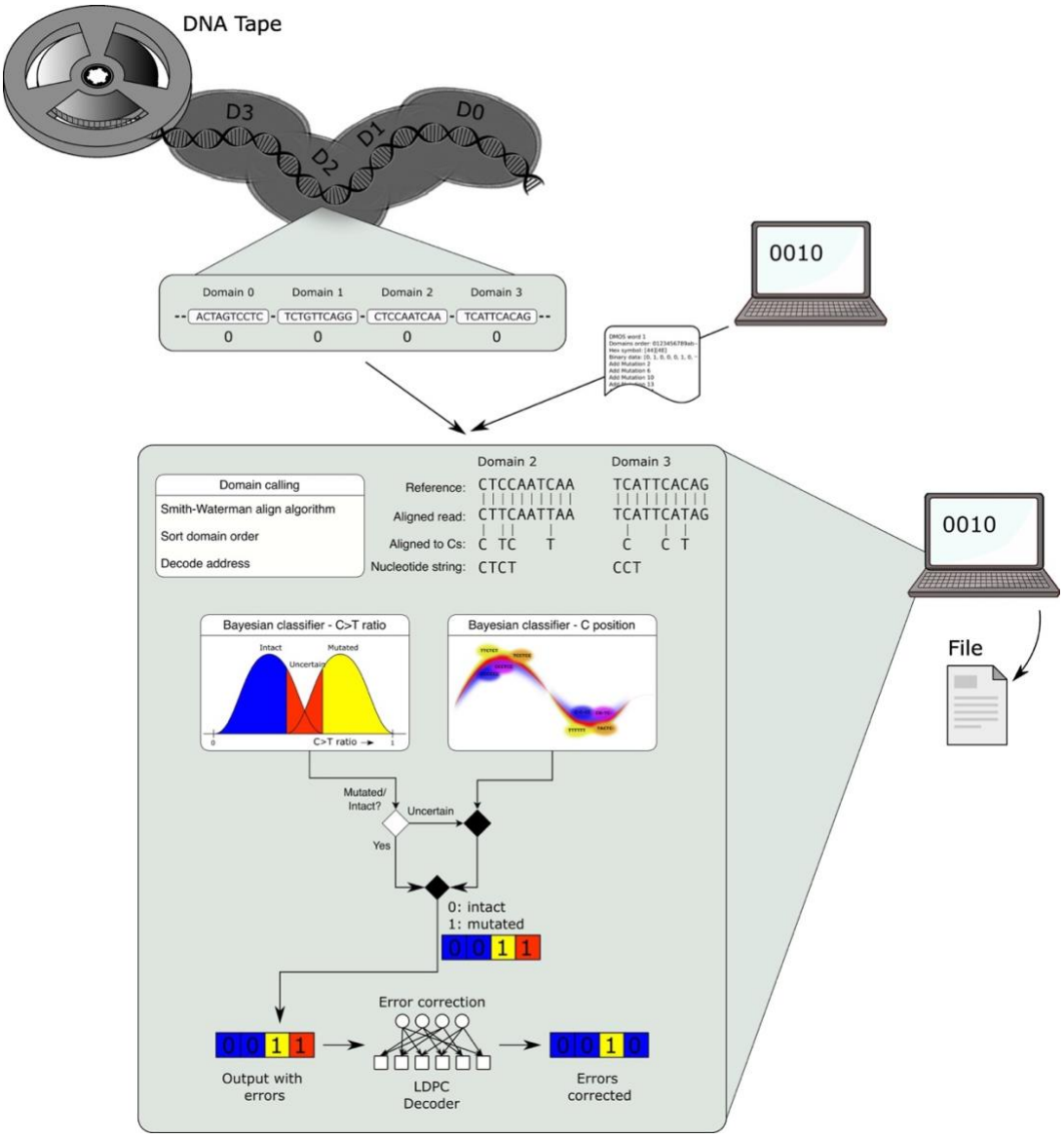
$$CB2 = (TC-UC2)/TC \quad (13)$$

$$CV(\%) = 100*((1-CB2)*CB2 + CB2*CBt) \quad (14)$$

CBt is the confidence calculated using the uncertain calls at the end of the composed classifier decision tree and CB2 is calculated using the uncertain calls from the second Bayesian classifier (UC2). The final value CV reveals the likelihood resulting from the second Bayesian classifier to be either mutated or unmutated during the first reads (CB2), and the confidence value converges to the final composed classifier results (CBt) after several iterations.

The DMOS decoder was written in C++, has a Python interface to generate a real-time visualization of snapshots, and is available on github.(https://github.com/SBMI-LAB/DMOSDecoder). Depending on whether the DMOS encoder with error correction is used, the Python script also calls the LDPC decoder algorithm and reports and if the original data was recovered. We next add error correction codewords to improve the data retrieval and accuracy of data storage. We designed a Protograph-LDPC code to protect data against bit errors.[101] During this process, we calculate the minimum code distance, and we use LDPC codes that can provide low encoding/decoding complexity. The entirety of this decoder system is visualized in Figure 18, in which we present the pathway in which we translate our file into a mutation schematic to encode our information inside of our DNA templates. The sequences of spacers used to test this encoding/decoding scheme is listed in Table 4.

Figure 18 – DMOS Encoding Schematic



Note. The DMOS encoding scheme treats the nucleotide template as a register, with each domain classified as a bit value. After applying the directed mutation / base editing protocol either manually or through a pipetting robot, each cytosine is called and aligned *via* Smith-Waterman alignment algorithm to detect the presence of thymine in where cytosine should be found in the reference sequence. Then, Bayesian classifiers call the state of each domain either mutated or unmutated based on the presence of mutational signals, designating either a 0 or 1.

Table 4 – Sequences for Experimental Template for DMOS

Segment	Sequence
Initiator	atcacgaggccctttcgtcttcaagaattc
Index	TTTATAGAAAACGTTTTGAAGAAGAAGATGATCTCT
State 1	ACTCactagtcctcgaaaacctcgTGG
Index 1	TGTCCTACTATGTCTTCTCTCTTCTACTACTTACCT
State 2	ACTCctccaatcaaatcagtcactAGG
Index 2	GGATGGATGATCCCACACCTCACACGCAGGAGAGAA
State 3	ACTCtctggtcagggctcggacacTGG
Index 3	CTAGTGGTAGATGTTGTGTGTGGCGCGAGAGAAAGC
State 4	ACTCtcattcacagcaactgcagcAGG
Index 4	TTGCGACGATGACTGACGACTGCACGAAAAGCTGGA
State 5	ACTCatggtcaactcaatccaaaaTGG
Index 5	GTGAGGAGGAGAAGTAAAAGAAAGCTTCGAGAGAGT
State 6	ACTCgttctcatcgcgtaccacgaAGG
Index 6	AGTTTACACGGCGCTCTTTCCGGTTTGATCTTGCAC
State 7	ACTCatcaatagtgtcatggcatgTGG
Index 7	ATGTTTACGCACGCGTTTTCCCACCCACGATGTTGT
State 8	ACTCtcgggagaaaggtcgctgtgAGG
Index 8	CTGTTTGCACACACACCCGCACACCCTGTTCCCTCG
State 9	ACTCatcacgagttcacgataccgTGG
Index 9	ATGCGTTGCGTTGTTTTGCGTTCCACACCACACGTT
State 10	ACTCttgtggtcaatgtcactccgAGG
Index 10	ATCCAAAGAGAACTGGGATTTCTAAAAGAGAGAGAA
State 11	ACTCaagctcagcctcgttaaacgTGG
Index 11	GTTTTACCTTTTGCGCTTTTGTCTTCGTTCGTCCCT
State 12	ACTCgaacagatcatcaacccattAGG
Index 12	GGCTCCCTACCACACACCACGTTTTGATGATAGTTG
State 13	ACTCattcaatcaagctgcaaaggTGG
Index 13	TACGAGAGGAAGCTTCACACACCACCACGATCGGAT
State 14	ACTCgattcgaatatctctcttcgAGG
Index 14	CTTGCGCACACCTCACACACGTGTTTGTGTTGTGTT
State 15	ACTCgcctcatcagcagaacaagtTGG
Index 15	CGATCCGCACACGCACGTCACACCTATCTTACGTGT
State 16	ACTCtcattccagtcaatgtggaaAGG
Index 16	GAAGAAAAGAAAGAGAAGAGAAAACTCAAAAGATGA
Terminator	ACTCatcgataagctttaatgcggtagtttatca

Note. Lowercase nucleotides are involved in molecular interactions. (i.e. PCR and dCas9 mutations) Initiator and Terminator sequences are sites for primers to amplify the template.

Indexes are included to identify initial domains after nanopore sequencing. Adjacent indexes are exclusive to each domain for our experiments.

DMOS Template Synthesis and Cloning

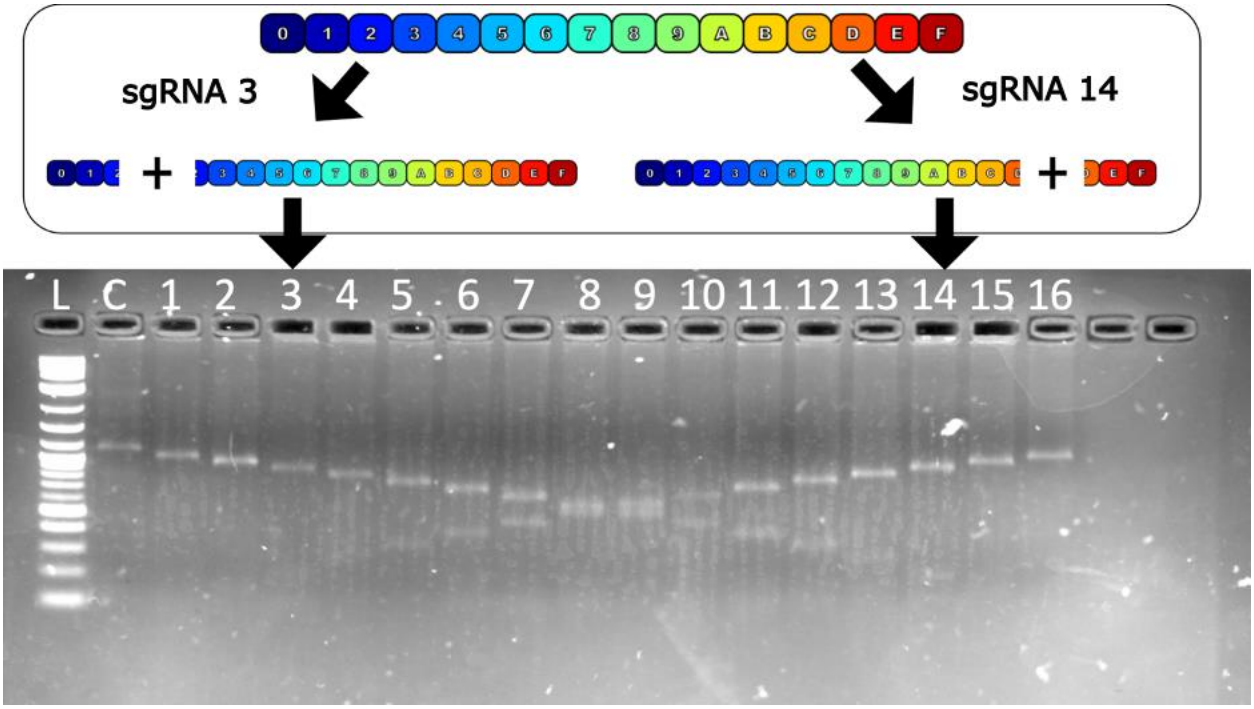
The DMOS registers are constructed by DNA sequences with different domain permutations synthesized by TWIST© Bioscience. The registers used for our orthogonality tests were assembled into the pBR322 plasmid (New England Biolabs) using the NEBuilder© HiFi Assembly Master Mix. The pBR322 plasmid was cleaved using FastDigest restriction enzymes Bsu15I and EcoRI to insert the register inside of the plasmid. The modified plasmid was transformed into NEB 5-α competent *E. coli* bacteria and grown under Carbenicillin antibiotic resistance. The assembled plasmid was extracted using the Monarch Plasmid Miniprep Kit (NEB). The register was amplified from the plasmid using Integrated DNA Technologies primers, with the forward primer containing four phosphorothioate bonds at the 5' terminus and a reverse primer with a phosphate group at the 5' end to condition the register for exonuclease digestion. The templates were purified using AMPure XP magnetic particles (Beckman Coulter) with 70% ethanol washes and eluted in Nuclease-Free water. The DNA templates for encoding were ordered from TWIST© Bioscience, but the templates were amplified directly.

Cleavage efficiency of gRNA to DMOS Domains

Cas9 cleavage requires strong binding of the Cas9 enzyme to its template sequence and stable formation of the R-loop.[102] To visualize the cleavage efficiency of the spacer sequences of gRNA and therefore estimate R-loop stability, we used active Cas9 proteins and confirmed the targeting of the writer system at the precise location of their targeted domains. Unlike the dCas9 proteins used in our writing system, the active Cas9 proteins will cleave the DMOS tape at the protospacer sequences. Each digestion splits the template into two fragments. Therefore, it

demonstrates that the RNP will generate a stable R-loop conformation required for APOBEC3A to recognize and deaminate the nucleotides within the loop. We validated the results of digested template fragments on 2% agarose gel to determine the length of the fragments compared to the control 1108 bp template labeled "C". We observed bands of a smaller length relative to the negative DMOS register control. This result indicates that Cas9 digestion only occurs in precise locations consistent with the sgRNA protospacers used. We also observed an "X" pattern across all digestion sites consistent with where we expect the Cas9 effector to cut (as each respective RNP cuts the DNA along the molecule at regular intervals), showing no significant off-target cleavage across the register in Figure 19. This result indicates that the RNPS are able to target and recognize the specific domain/bit only at the precise location of interest.

Figure 19 – Cas9 Digestion across DMOS template



Note. Lane L indicates the 1 kb plus ladder, while Lane C is a negative control to indicate the template when undigested. Each numbered lane presents a different cutting site affected.

Each domain is decoded from its lexicographic code to identify the initial sgRNA target values. In the schematic, the "2" represents protospacer/domain 3 targeted by RNP with gRNA 3 and the 14 represents the protospacer/domain targeted and cleaved by an RNP with gRNA 14.

Enzymatic writer Protocol

The deamination reactions were prepared by dispensing 1.5 uL of RNP and 1.5 uL of the DNA template in a 15.5 uL reaction with dCas9 buffer [200nM HEPES, 1M NaCl, 50 mM MgCl2, 1 mM EDTA, pH 7.4]. The reactions were prepared using 8.5 uL of nuclease-free H2O, 1.5 uL of 10x Cas9 buffer, 1 uL (40 units) of RNAse Inhibitor Murine, and 0.5 uL of BSA. The APOBEC3A and BSA materials were from the NEBNext© Enzymatic Methyl-seq Conversion Module kit from New England Biolabs (NEB). The deamination protocol has been scaled down to 15.5 uL reactions modified from the original deamination protocol from the kit. The reaction mix was centrifuged and incubated inside a MiniAmp thermocycler at 37 ℃ for 3 hours per the standard protocol.

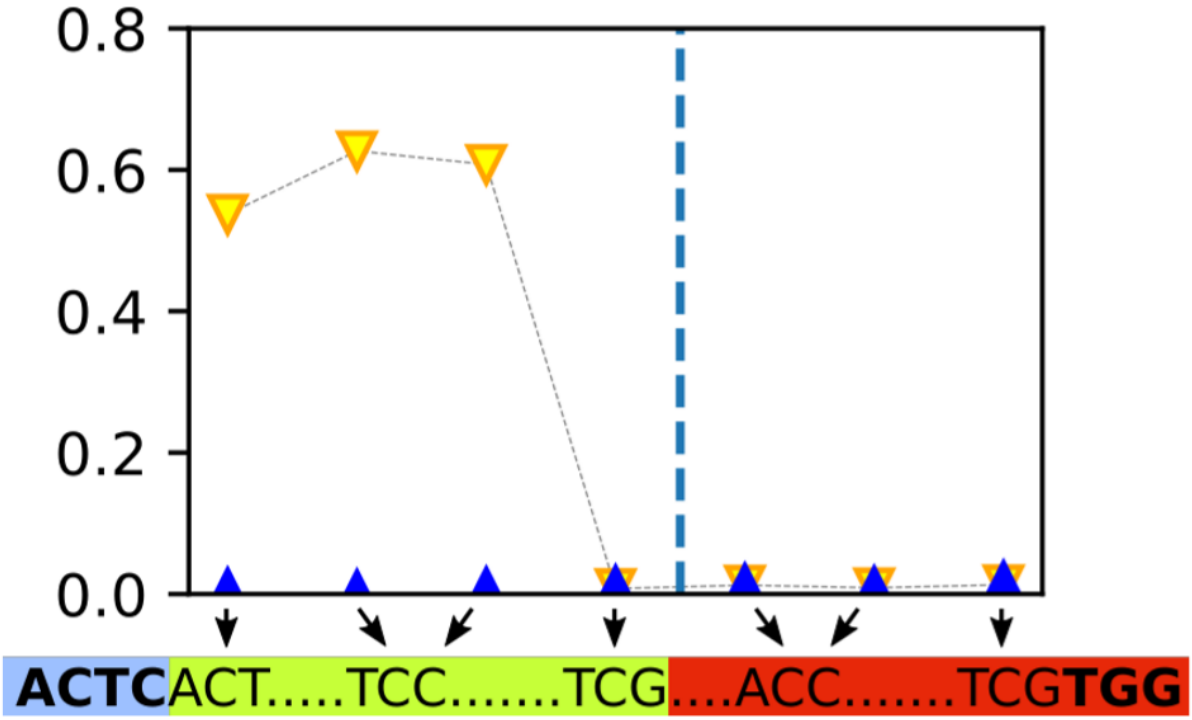
The resulting reaction was treated with 1 uL (0.8 units) of proteinase k and incubated at 56°C for 10 minutes. The samples were purified using AMPure XP magnetic beads following standard protocols. After eluting the sample from the beads, the DNA was treated with Lambda exonuclease to degrade the nonmutated strand from our templates using 1 uL of Lambda Exonuclease and 5 uL of commercial 10x Lambda Exonuclease buffer for a total reaction volume of 50 uL with nuclease-free H2O. The reaction was incubated at 37℃ for 30 minutes and heat-inactivated at 75 °C for 15 minutes. The reactions were purified using AMPure XP magnetic beads. For the PCR protocol, the mutated samples were amplified with addressing primers with an annealing temperature of 63 °C and an extension of 72 °C at 45 seconds. The

products were purified, mixed in femtomolar scale, and sequenced through an Oxford Nanopore device. This process was used for encoding our data into our synthetic templates.

Orthogonality tests on DMOS Template

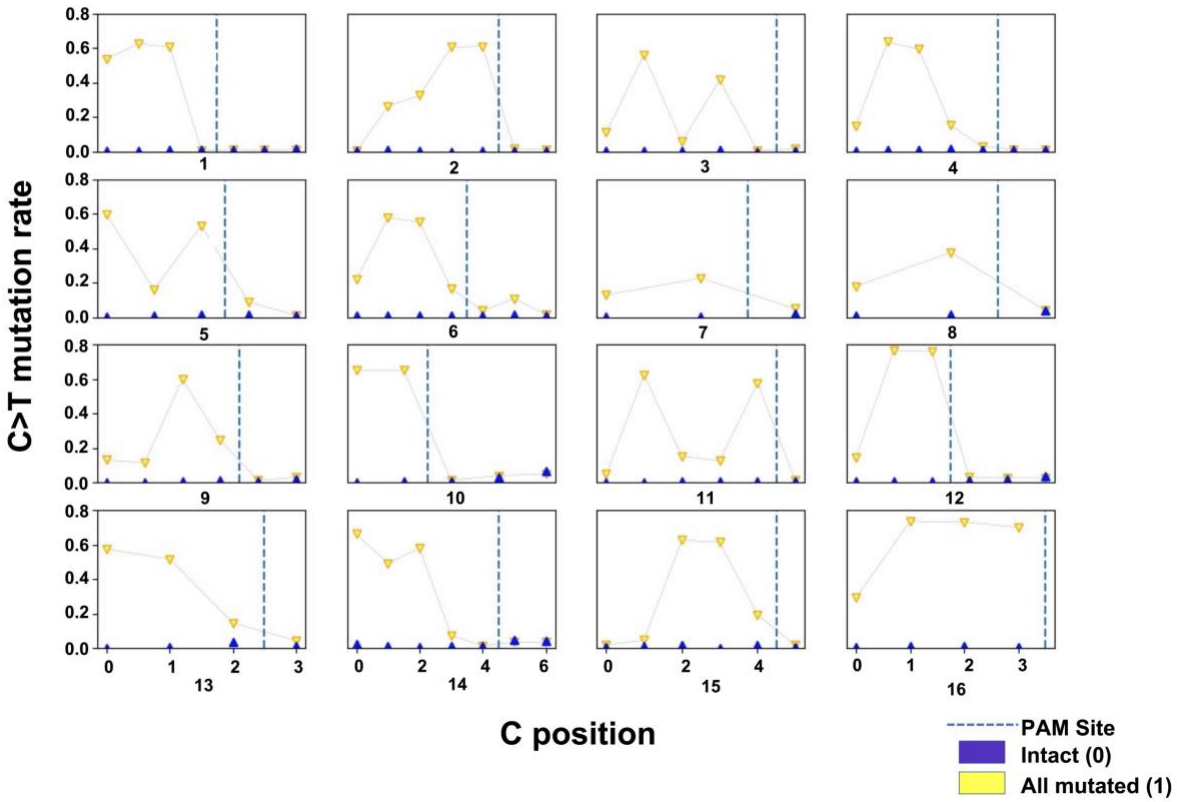
We tested orthogonality of dCas9 targeting by performing 16 reactions with two replicates, with a single RNP targeting each domain. Control templates with no dCas9, no APOBEC3A, and a reaction with neither material were included to distinguish between unmutated and mutated samples. Each reaction was performed manually with individual dCas9 RNPs (1 uM) with 2 replicates each for 32 reactions. The experimental and control reactions were incubated at 37°C for 3 hours inside of a thermocycler. Each reaction was treated with 1 uL of proteinase k (0.8 units) and purified using AMPure XP magnetic beads. The purified samples were treated with lambda exonuclease and purified using standard protocol. The registers were amplified using Q5U polymerase (NEB) with predetermined combinations of barcoded primers (Integrated DNA Technologies) to distinguish experimental conditions in a single nanopore sequencing run. Each reaction was purified and combined for nanopore sequencing. The results were analyzed and the experiment was repeated twice for 3 data sets. Individual domain analysis of cytosine mutation and analysis of all 16 domains are presented in Figures 20 and 21, respectively. A bit map with all experiment replicates was compiled and presented in Figure 22. Barcoded primers were used as the addressing scheme and are listed in Table 5. We performed our orthogonality tests for a total of 5 replicates, including controlled conditions in which no mutation was induced, and included an in-depth analysis of each cytosine position within each domains the results in Figure 23 to detect thymine conversions.

Figure 20 – Analysis of Cytosine Conversion in Domain #1



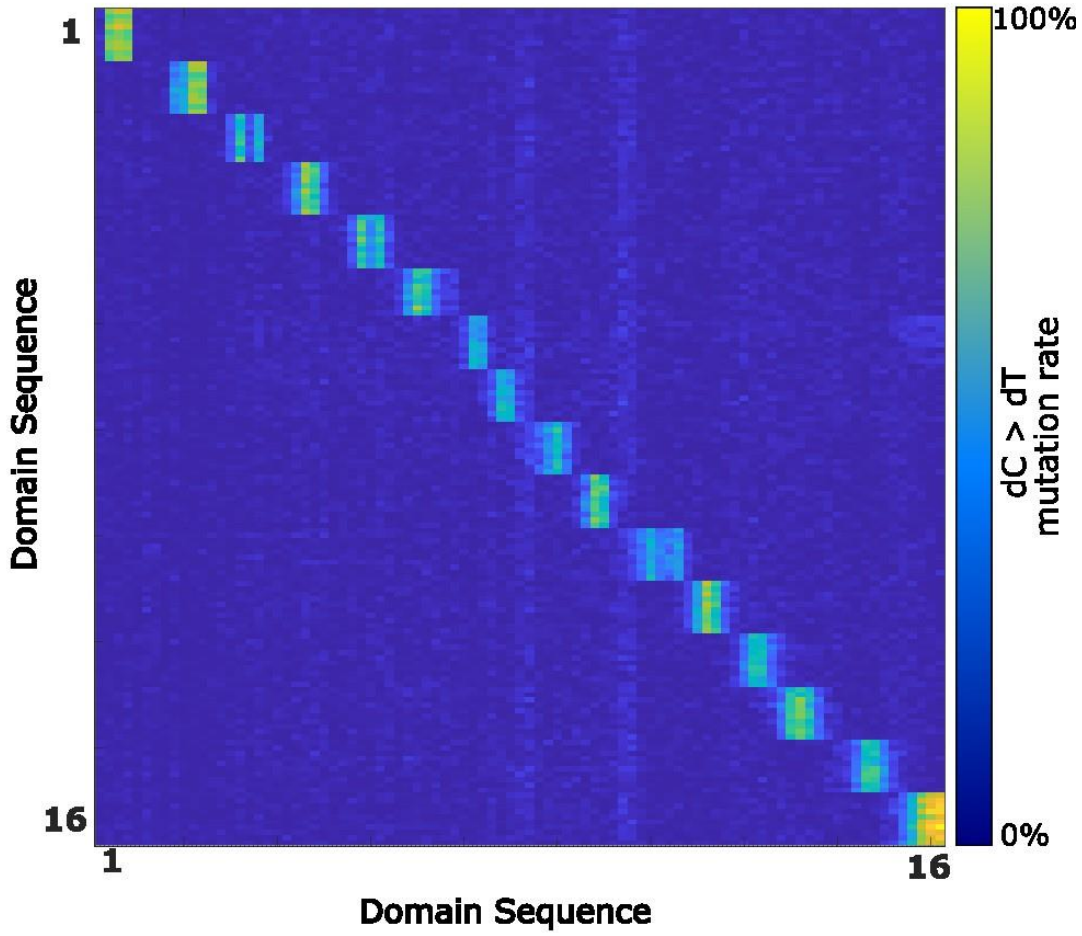
Note. Average cytosine conversion rate across individual cytosines in domain 1. Yellow upside-down triangles indicate mutation, while right-side blue triangles measure domains without APOBEC exposure. Mutations of each cytosine were analyzed to determine mutation patterns consistent with the APOBEC3A protein. Blue ACTC motifs were used as markers for domain identification. We observed that cytosines within 6 bp adjacent to the PAM site were protected from mutation regardless of being in the same spacer target, indicated in red, while affected cytosines are in green.

Figure 21 – Individual Cytosine Analysis across all domains



Note. Each cytosine was analyzed individually to determine patterns of mutation depending on position of cytosines within each spacer. We observe that the exposed cytosines within 6 nucleotides of the PAM site experienced no mutations, consistent with what we observed from APOBEC3A mutations.

Figure 22 – Orthogonality Analysis across each Domain in Template 3

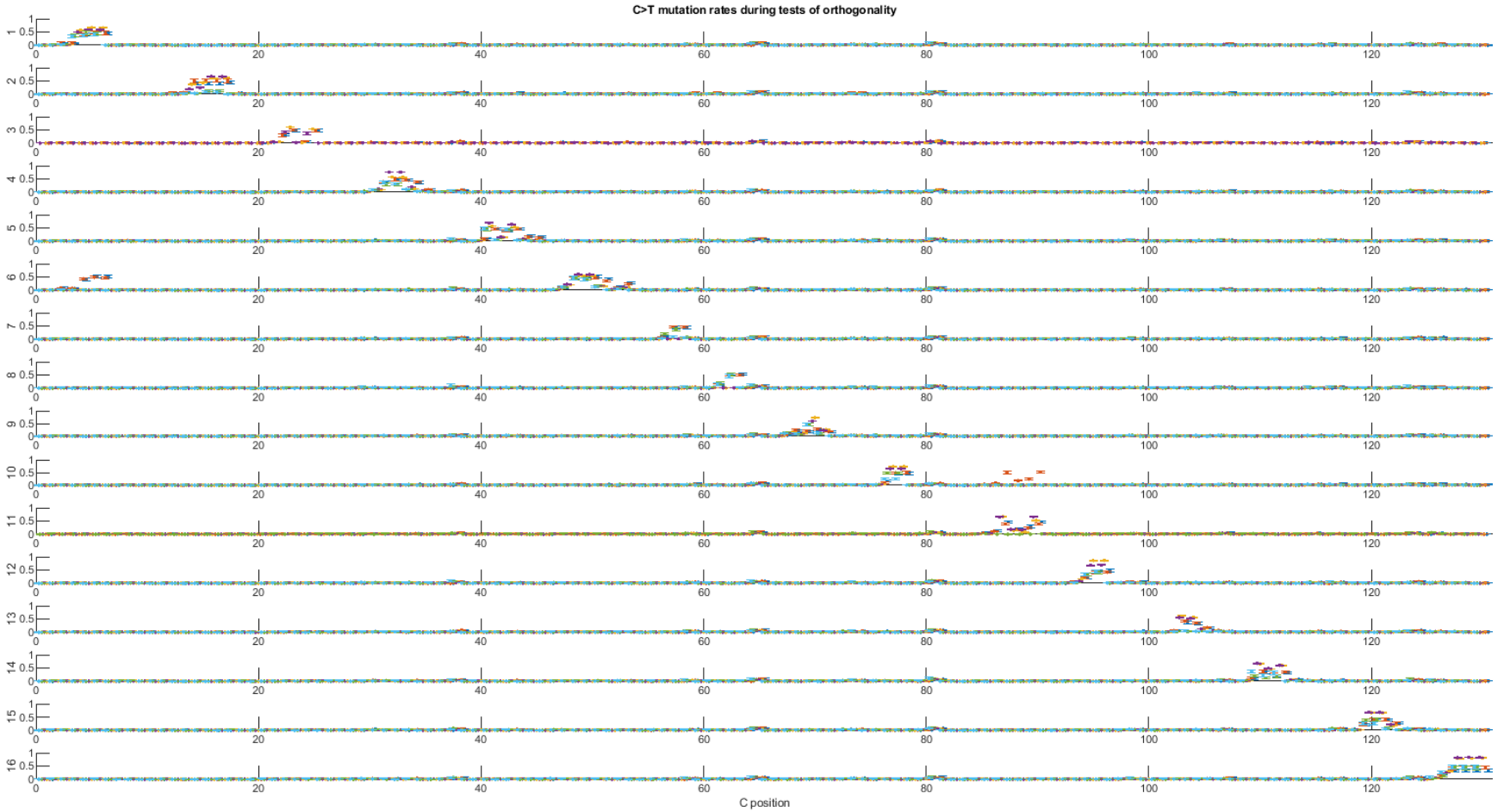


Note. Each domain was analyzed individually from our third template to determine cross-reactivity of each guide. This data was cumulated from five repeats of our mutation analysis experiments. Mutation rates were measured between the lowest of 40% up to a high of 99% These results conclude that we observe mutations directly where we expect them to occur based on the applied RNP spacer sequence, and that no off-target mutations occur across each spacer.

Table 5 – Barcoded Primers for Template Identification

Code	Sequence
dBC1F	TTTCTGTTGGTGCTGATATTGCGTTGTCGGTGTCTTTGTGATCACGA GGCCCTTTCG
dBC2F	TTTCTGTTGGTGCTGATATTGCCCGTTTGTAGTCGTCTGTATCACGA GGCCCTTTCG
dBC3F	TTTCTGTTGGTGCTGATATTGCTGTGTCCCAGTTACCAGGATCACGA GGCCCTTTCG
dBC4F	TTTCTGTTGGTGCTGATATTGCTTCTATCGTGTTTCCCTAATCACGA GGCCCTTTCG
dBC5F	TTTCTGTTGGTGCTGATATTGCCAGGGTTTGTGTAACCTTATCACGA GGCCCTTTCG
dBC6F	TTTCTGTTGGTGCTGATATTGCGAACAAACCAAGTTACGTATCACGA GGCCCTTTCG
dBC1R	TTGCCTGTCGCTCTATCTTCCCGTGGGAATGAATCCTTTGATAAACT ACCGCATTAAAGC
dBC2R	TTGCCTGTCGCTCTATCTTCCAAAGGCAGAAAGTAGTCTGATAAACT ACCGCATTAAAGC
dBC3R	TTGCCTGTCGCTCTATCTTCGCACAGCGAGTCTTGGTTTGATAAACT ACCGCATTAAAGC
dBC4R	TTGCCTGTCGCTCTATCTTCTGAAACCTTTGTCCTCTCTGATAAACT ACCGCATTAAAGC
dBC5R	TTGCCTGTCGCTCTATCTTCTCTATCGGAGGGAATGGATGATAAACT ACCGCATTAAAGC
dBC6R	TTGCCTGTCGCTCTATCTTCGAAAGAAGCAGAATCGGATGATAAACT ACCGCATTAAAGC

Figure 23 – Large Scale Orthogonality Test across Five Trials

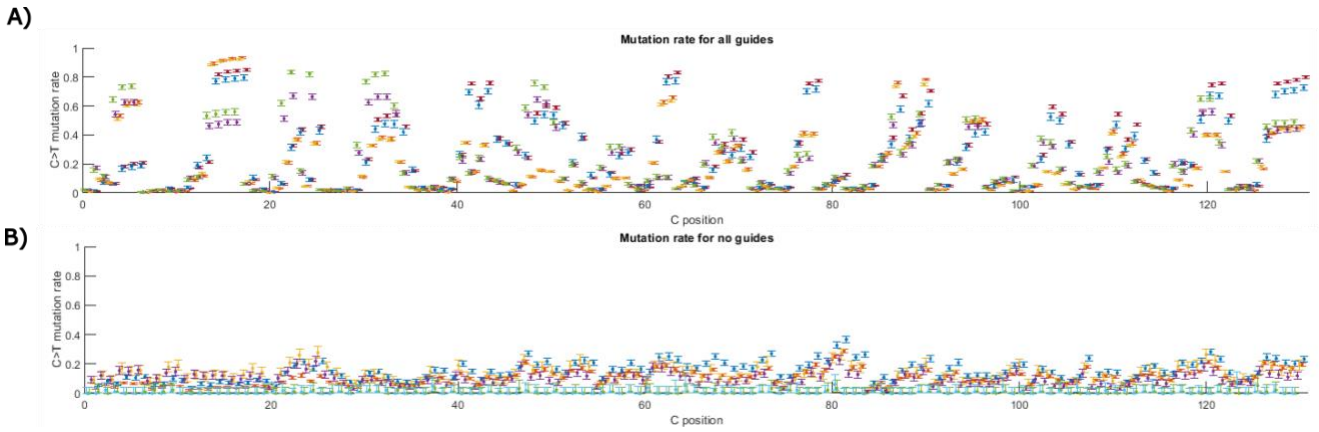


Note. Individual cytosine mutations across 5 separate trials, indicated via different colors. We observed that mutations were significantly high, from 40% to 60% across all domains in which the corresponding dCas9 spacer was applied to mutation reaction. Despite the misapplication of dCas9 from previous trials, we can induce consistent mutations with little off-target effects.

All-Domain Targeting of DMOS Template

We wanted to observe if multiple dCas9 directed mutations were possible with this template, so we prepared 50 nM mixtures of each of our RNP complexes and applied this mixture to our mutation reactions. Following our established mutation protocol, we applied the dCas9 RNP mix onto our DMOS template in a 10:1 ratio to measure mutational activity across all affected cytosines. We performed this experiment for a total of 5 separate trials and recorded our results in Figure 24.

Figure 24 – All-Guide Targeting of DMOS Template



Note. Mutation rates of all protospacer-occupying cytosines with the application of all 16 RNP spacers in each reaction. A) Mutation rate of cytosine in the presence of all 16 RNP spacers and APOBEC3A. We found variable levels of mutation across all expected domains, ranging from 60% to 90% in all affected regions. B) Mutation rate of cytosine in which no guides were present. While there is a small level of mutation present across each cytosine, we observe that our preliminary data presents a significant impact of the presence of dCas9 across each domain. Therefore, we conclude that we can include multiple dCas9 RNPs that will target different domains to our DMOS template, and be able to distinguish those results from our controls through machine learning.

Automation of Writing via OT-2 Pipetting Robot

To automate the "writing" protocol, an Opentrons OT-2 pipetting robot was programmed to introduce the dCas9 into each of our mutation reactions. This procedure requires the following plate preparations: The first plate contains the dCas9 library, which has been designed for the 16 RNPs, the second plate contains blank DNA registers in separate pools, each one labeled with the corresponding address, and the last plate contains the buffer to mix the contents as an intermediate step.

We developed Python scripts for the robot that reads the mutational list file and performs the following steps: For each target domain, it takes volumes of 1.5uL from the dCas9 library plate, and mixes them in the buffer mix plate. Once mixed, the robot takes 1uL from the buffer mix and deposits it into the selected register pool. After depositing the dCas9 into all the registers, we moved the templates for three hours in the thermocycler. Then, the samples were manually purified for post-mutation treatments and sequencing.

Encoding/Painting a File on DMOS Register using Controlled Mutations

The dCas9 RNP mixtures were prepared using an Opentrons pipetting robot followed by a predetermined message pattern. Each RNP mixture has a final concentration of 50 nM for each RNP to conserve a high RNP to register ratio. The mutation reactions were prepared manually using DNA registers with a concentration of 6.25 ng/µL, while the RNP mixtures were added based on the register in each reaction. The APOBEC3A was added last into each reaction and incubated for 3 hours at 37°C. The reactions were treated with 1 uL of Proteinase K to degrade the dCas9 and purified each DNA sample via AMPure XP beads and 70% Ethanol washes. The purified samples were each treated with Lambda Exonuclease and purified for amplification and storage. Each register was amplified using Q5U polymerase with 5 uL of the purified sample as a

template for each PCR with non-barcoded primer sequences. The PCR was performed with an annealing temperature of 63°C and an extension of 45 seconds for 30 cycles. The affected registers were purified using standard AMPure XP bead protocols and each template's concentration was recorded. The samples were mixed and sequenced using Oxford Nanopore Sequencing and left on overnight to collect results. Several thousand sequencing reads were collected and analyzed for mutations.

Encoding Full Message on DNA Tape

To demonstrate the encoding capabilities of CRISPR and the decoder, we turned to encoding the title of this work "**Digital data storage on DNA tape using CRISPR base editors"** across 48 DNA registers. We used Protograph LDPC codes to translate the digital file into ASCII code, (American Standard Code for Information Interchange), and then relayed the information into a list of required mutations to encode the message onto our 48 registers.[99] We programmed the Opentrons OT-2 pipetting robot to prepare the dCas9 mixtures and apply them at the respective registers after preparing the mutation reactions by hand. The mutation reactions were prepared using the same protocol as when we performed the painting experiment.

Results and Discussion

The DMOS bit Recording Strategy

The foundations of our data storage system are the "DMOS bits," or engineered DNA sequences that can be efficiently mutated during the CRISPR base-editing reaction to encode data, and that they can be assigned a '0' or '1' state after sequencing and recognition of the presence or absence of a "mutational signature."

By recognizing the presence or absence of a mutational signature of cytosines converted to thymines across the entire 'state' segment of a DMOS bit, we can determine whether or not

that bit should be assigned a 0 (unmutated) or 1 (mutated) as digital data. We identified and experimentally validated a set of 16 unique gRNAs and 'state' sequences where (1) dCas9 exhibited robust double-strand activity at that target, signaling stable R-loop formation, where each gRNA was highly specific to their target; (2) where each displaced DNA sequence would contain at least two 'TCR' nucleotide motifs, and (3) where those motifs were located in a mutagenic 'hot-spot' > 6 nt away from the PAM.

Since the 'state' segment of a DMOS bit requires a 20 bp sequence recognized by gRNA and a 3 bp protospacer adjacent motif 'NGG', we opted to optimize the data to be read using the DMOS system for long-read nanopore sequencing to increase the density of data per register. To help address and localize a bit along a register during long-read sequencing, we included the unique 40 bp 'index' sequences to each DMOS bit. These indexes served to space the bits out so dCas9 effectors could interact independently during our optimization experiments.

The DMOS DNA tape

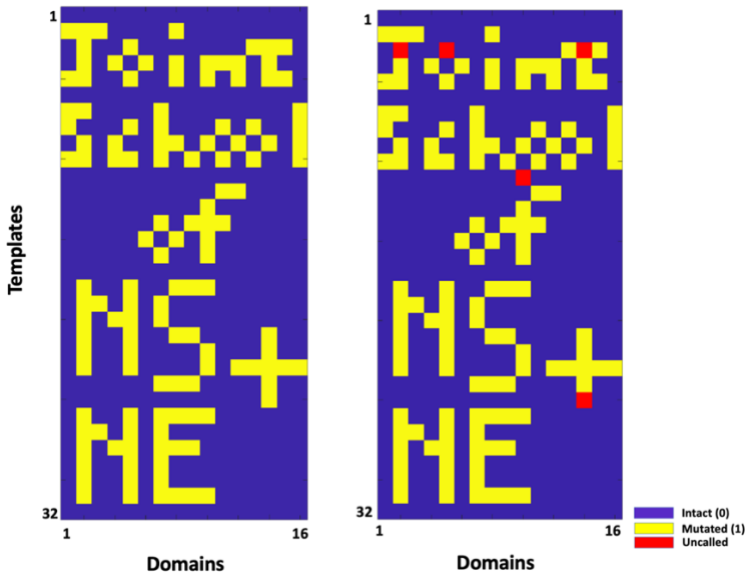
The 16 unique DMOS bits were assembled into "DMOS registers" with possible 2^{16} mutational states, determined by the presence or absence of each gRNA during an *in vitro* base-editing reaction performed. Each DMOS bit exhibited high levels of mutation when the corresponding RNP was present in the base-editing reaction. This activity was orthogonal, independent, and reproducible across technical and sequencing replicates.

We realized we could significantly increase the data storage capacity (amount of data that could be written in parallel or read simultaneously in the same sequencing run) using the same 16 DMOS bits and RNPs by generating different DMOS registers where the order of the DMOS bits was permuted, allowing up to 16! different DMOS registers. High-entropy permutations were deterministically generated and enumerated, and when the order of the DMOS bits was

determined from a molecule, that could be mapped back to the "address" or order of the DMOS register to organize them into "DMOS data blocks" or a "DMOS tape".

As a demonstration of manual cytosine conversions, we created a 16 x 32 bitmap (512 bits) representation of our school's logo, "Joint School of NS [Nanoscience] + NE [Nanoengineering]," across 32 DMOS registers, plotted out the desired locations of the 0's and 1's to illustrate the image and automated with pipetting robots to deliver each of the 16 gRNA/dCas9 complexes to each respective register. From the sequencing data, we determined the register location from each sequencing read. We assigned a digital state to each bit from each register using our encoding/decoding code. After 20,000 reads, sequencing, and analysis, we recovered our intended bitmap with 98.8% accuracy (6/512 errors) as seen in Figure 25. While the information retrieved from this experiment is not encoded with digital information per se, the expression of each template into a controlled image demonstrates that our system allows for individual control over where CRISPR mutations are applied, changing the bit state of our blocks.

Figure 25 – "Painting" of Joint School of Nanoscience and Nanoengineering Logo

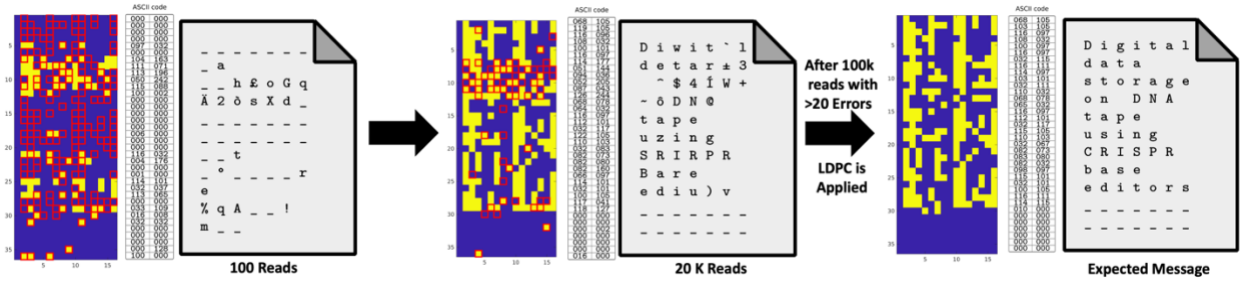


Note. (Left) Intended image that we wanted to demonstrate using our enzymatic puncher system and decoding strategy across 32 unique registers. (Right) Final message collected when applying the mutation and decoding strategy to unique 32 registers. It was found that out of 512 bits, we have achieved a 99% accuracy rate from our experiment, allowing us to apply our protocol to encoding actual data to our registers.

Writing and reading digital information on DNA tape

With an error rate of <1.2%, this implied that we could take advantage of error-correcting schemes to improve data reliability and to perfectly recover digital data from DMOS DNA tapes. To test this approach, we encoded the title of our upcoming manuscript describing the work, "**Digital data storage on DNA tape using CRISPR base editors**", requiring 72 bytes (576 bits) in ASCII, and LDPC creates a codeword to the system, which requires additional 96 bytes (768 bits total).[99] We encoded this sequence into 48 DMOS registers. After 100k reads, bit assignment, and error correction, we recovered the original message without corruption as shown in Figure 26 with 100% recovery and accuracy.

Figure 26 – Final Decoded Message after 100K Reads



Note. 100,000 reads in total were recorded with less than 20 errors. As we collected more reads from sequencing, we observed that our message was becoming more accurate. Our full message was retrieved when we applied error correction to our collected reads.

Conclusion

In this chapter, we applied an encoding scheme that incorporates mutations driven by CRISPR and can decode a full message using multiple registers based on our template 3.0 design. The number of potential permutations of DMOS registers can theoretically be up to 16! different registers. Of course, making use of all of those templates would require automation to eventually be feasible, and so here we also demonstrated that digital data could be written using our in vitro mutagenetic protocol using a pipetting robot. Our DMOS block that contains 48 registers weights 5.45×10^{-17} grams and stores 768 bits (96 bytes) of raw data, or 576 bits (76 bytes) of protected data that brings a bit density per nucleotide is 0.0144 bits/bp and a maximum theoretical bit density of 1.4×10^{19} bits per gram, equivalent to 1.5 exabytes/gram. Ultimately, we have demonstrated that we can perform digital data encoding into blank DNA "tapes" *in vitro* using CRISPR and mutagenic proteins and that using nanopore sequencing and a Bayesian classifier trained using sequencing results we generated during the course of this research that this data can then be decoded with 100% accuracy.

CHAPTER IV: FUTURE WORKS AND CONCLUSION

Future Works

Application of Multiple Base Editors

Since our encoding and decoding strategy relies on the generation of targeted mutational signatures into preformed DNA templates, we anticipate that adenine editing enzymes will allow for a broader range of mutations we can differentiate through sequencing. Therefore, we can implement greater numbers of mutational signatures or bits into each of our targeted domains by successfully expressing proteins with different nucleotide editing patterns. Our inspiration is derived from several works that utilize different types of base editors beyond cytosine base editors, such as adenine base editors, *in vivo*[59,103,104] where adenine is converted to inosine and is amplified as guanine. We look to utilize the previously discussed modified TadA enzyme in our system to include new modifications we can detect through nanopore sequencing. We would then be able to convert these distinguishable mutational signatures into unique bit values.[60,105]

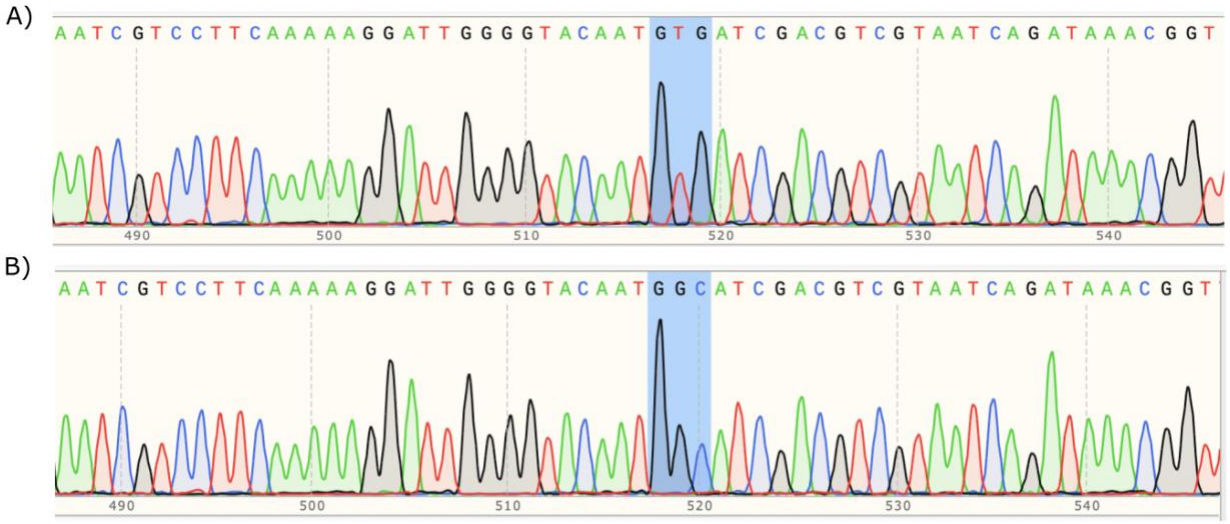
We also look to synthesize and prepare APOBEC3 proteins from the same family to analyze new mutational signatures that can be differentiated from the mutation patterns of APOBEC3A. To accomplish this, we look to utilize the APOBEC3G protein, which is found to induce mutations in 'CCC' motifs as previously discussed.[106] As such, we hypothesize that its mutational properties will be distinguishable, as we should see high rates of mutation across cytosine repeats instead of TCR motifs.

In preparation of these future applications, we obtained plasmids coding for genes used for base editing from Addgene that express base editors with A3A, A3G, and the synthetic adenine base editor, ABE7.10 that have all been expressed and tested through published works in cells.[60,106] Each plasmid is designed with a nCas9 nickase linked with a uracil glycosylase

inhibitor and their respective base editor. Before they can be used for *in vitro* analysis, these plasmids must be modified to instead express a dCas9 variant to deactivate nicking activity since our protocol relies on intact DNA molecules. We sought to limit the modification of the nucleotide backbone by editing the histidine 840 (GTG) amino acid codon. Therefore, we designed a protocol to replace the H840 amino acid with alanine (GCC) to deactivate the HNH domain. We performed this mutation *via* Gibson assembly with amplified fragments of each plasmid. We amplified three components of each plasmid, but one primer pair contains the alanine codon sequence instead of histidine, which should render the nCas9 fusion protein fully inactive when the reassembled plasmid is used as a template for transcription.

We tested our assembly on the hA3G-BE3 plasmid to test our modification strategy.[106] We amplified the plasmid sequence into thirds via PCR. Each fragment was ligated via Gibson assembly, cloned into *E.coli*, and sequenced to determine whether the point mutation was successfully administered. From the Sanger Sequencing results, we found that we were successful in mutating to the nCas9 to dCas9, as shown in Figure 27. With the mutation present in at least one colony of bacteria, we plan to have the next step be to express this protein and test it for activity using our DMOS encoding scheme.

Figure 27 – Codon Mutation of hA3G-BE3



Note. A) Sequence of hA3G-BE3 plasmid across histidine 840 amino acid without codon edit. This plasmid has not been edited to include the modification to the targeted codon. (highlighted in blue) B) Sequence of hA3G-BE3 plasmid that contains the alanine amino acid codon, showing a successful substitution of the targeted codon. Sequencing was performed by GENEWIZ (Azenta Life Sciences) using a reverse primer, showing the reverse-complimentary sequence of the targeted region.

However, one caveat to consider is that with both the CRISPR protein and base editor linked together, the formation of the RNP will need to be modified. We plan on adding RNA sequences alongside the mutation reactions as the temperature conditions for ribonucleic protein formation and mutations were at the same temperature and buffer conditions. To reduce the alterations to our current enzymatic mutation protocol, another avenue that can be pursued is to express the mutagenic proteins outside the fusion protein by amplifying the sequence coding for said protein and use it as a template for protein expression.

One method to express individual mutagenic proteins would be to use the codon sequence of APOBEC3G as a template, and express the protein using robust *in vitro* protein

expression kits. This is possible using the PURExpress *in vitro* protein express kit from New England Biolabs and can be reverse purified using NEBExpress Ni-spin columns. Using primers that amplify the human APOBEC3G protein from the T7 promotor region to the terminator region, we amplified the template that only codes for the APOBEC3G protein. From using 375 ng of the APOBEC3G template sequence, we incubated the DNA in a reaction of PURExress *in vitro* expression kit and found that we were able to express the APOBEC3G protein with this kit based on our comparison between our positive and negative controls, shown in Figure 27.

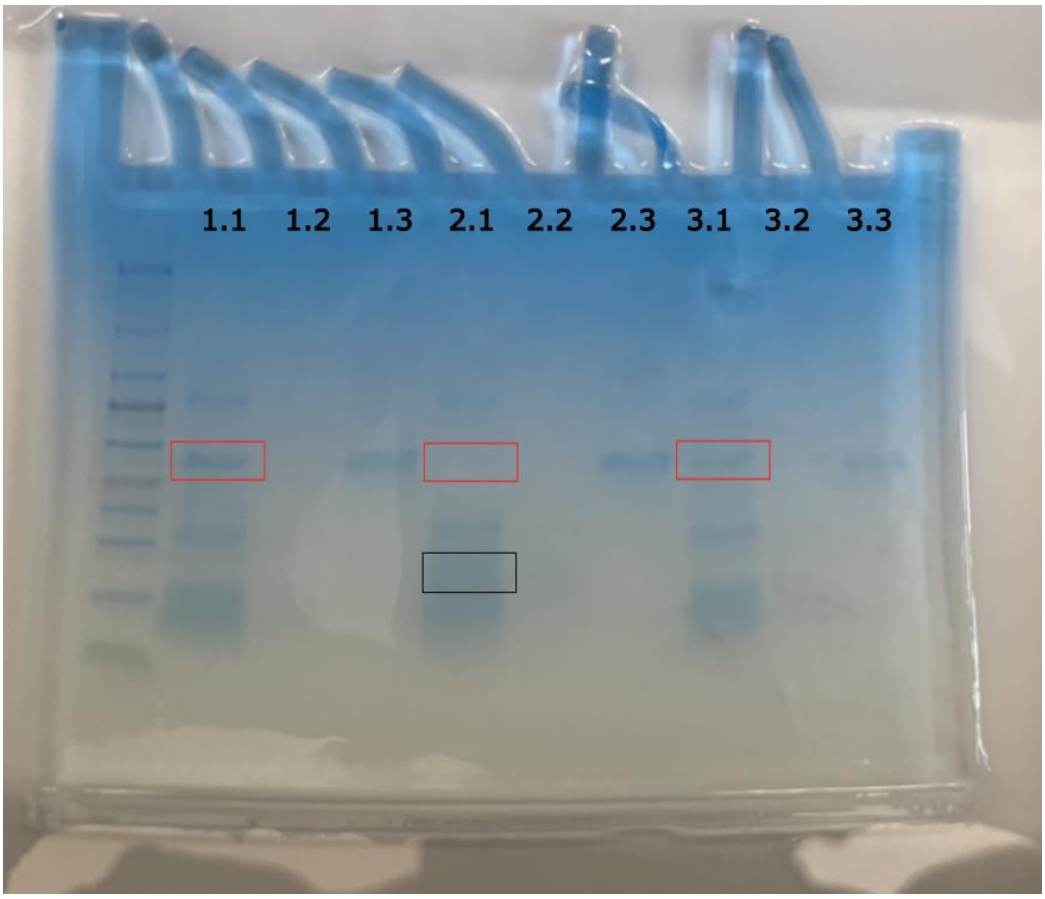
To perform our initial expression and purification experiments, we incubated 375 ng of A3G DNA and 250 ng of supplied DHFR plasmid into two separate PURExpress reactions following the kit's standard protocol. A third reaction in which neither material was added was included in the reaction. All three reactions were incubated at 37°C for three hours inside of a thermocycler. The reactions were then purified with Ni-spin columns through reverse purification, removing the his-tagged components of the PURExpress kit. Therefore, we expect the desired protein to be present in the supernatant, and not in the eluation. The supernatant, washes, and eluation phases were retained from each reaction to run through a SDS-PAGE 4%-20% precast protein gels. After running electrophoresis alongside a multicolor broad range protein ladder with the retained material, we suspended the gel in Coomassie gel stain overnight and acquired the image shown in Figure 28.

Based on the observation of the protein gel, we observed a dark band around the 45 kDa region of the reaction with the template DNA compared to the other two reactions, which is what we expect the size of the human APOBEC3G protein, which is around 46 kDa.[107] This leads us to conclude that there is a protein product being expressed, as no band was present in the same size in the positive control. We observed that there was a faint band present in the supernatant of

our negative control, but since the intensity of the his-tagged material was not consistent with the other two reactions, we conclude that the band may be leftover material that did not properly bind to the Ni-spin column.

In conclusion, we have observed that expression of the individual protein *in vitro* is possible. However, we cannot conclude on the activity of the protein unless we can isolate the protein and test mutational activity on our templates. This can help in the expansion of the DMOS protocol in future work to expand the "DMOS alphabet" of the number of bits that can be encoded within each domain *via* unique mutational signatures by different localized mutagens.

Figure 28 – SDS-PAGE gel of PURExpress Product after Reverse Purification



Note. Samples labeled 1.X indicate the supernatant (X.1), wash (X.2), and eluate (X.3) of the APOBEC3G protein expression reaction in sequence, while samples labeled 2.X and 3.X are

the positive and negative controls respectively. We observe a dark band around the 45 kDa region in our experimental reaction, indicating the presence of the desired protein. The red rectangles indicate the position of where the synthesized protein would be present. The positive control contains a band present around the 20 kDa region and no band is present in the 45 kDa region, showing that the positive control functioned as intended. No protein was expressed in the negative control, but a small band is present in the 45 kDa region, which we hypothesize that it is flow-through of his-tagged material that was remove prematurely due to the lower intensity of the his-tagged band in the eluation.

Domain Ligation using T4 Ligase

Another method of enhancing our information encoding density would be to generate new registers (orders of the 16 different domains) *in situ* without having to synthesize new templates from *de novo* synthesis. Using complementary overlaps and the T4 Ligase enzyme, we can scale up each sequence's bit information by ligating pre-designed domain sequences into longer registers than the 1108 bp templates we used in DMOS. For example, in a 96 well plate with each a unique domain sequence, we can potentially make more than 9,000 bits using 5760 bp templates. We can potentially include longer templates using this strategy, but it will require automating the process for this technique to become feasible and precision ligation to effectively prepare our own registers using short oligo protospacers.

Conclusion

In this dissertation, we demonstrated how DNA molecules can be mutated rather than synthesized *de novo* to storage digital information through targeted CRISPR base editing. Using deaminating enzymes such as APOBEC3A, we are able to encode digital information directly onto pre-synthesized DNA templates by changing the nucleobase structure within these

templates directed by specific CRISPR RNPs and read the data through sequencing and a computational base-calling strategy. Our approach to mutating DNA to store digital data can be easily automated as we have shown by writing our future publication's title across 48 individual templates. This process also has many promising avenues to expand the data storage capacity such as including different base-altering proteins that are distinguishable from APOBEC3A that can be differentiated by their unique mutational signatures—different signatures would therefore expand the DMOS alphabet so that in each domain we could encode, for example, 0, 1, or 2 (or even more) depending on the mutagenic protein added to the reaction. Ultimately, the development of a new *in vitro* scheme to chemically encode and decode DNA using sequencing enhances the field of digital data storage technologies with biological materials.

REFERENCES

1. V.V. Zhirnov & Rasic, D. 2018 Semiconductor Synthetic Biology Roadmap. (2018) doi:10.13140/RG.2.2.34352.40960.
2. Meiser, L. C. *et al.* Synthetic DNA applications in information technology. *Nat Commun* **13**, 352 (2022).
3. Biswas, S., Nath, S., Sing, J. K. & Sarkar, S. K. Extended nucleic acid memory as the future of data storage technology. 16.
4. Ceze, L., Nivala, J. & Strauss, K. Molecular digital data storage using DNA. *Nat Rev Genet* **20**, 456–466 (2019).
5. Chen, K. *et al.* Digital Data Storage Using DNA Nanostructures and Solid-State Nanopores. *Nano Lett.* **19**, 1210–1215 (2019).
6. Hossein TabatabaeiYazdi, S. M., Gabrys, R. & Milenkovic, O. Portable and Error-Free DNA-Based Data Storage. *Scientific Reports* **7**, (2017).
7. Dong, Y., Sun, F., Ping, Z., Ouyang, Q. & Qian, L. DNA storage: research landscape and future prospects. *Natl Sci Rev* **7**, 1092–1107 (2020).
8. Zakeri, B., Carr, P. A. & Lu, T. K. Multiplexed Sequence Encoding: A Framework for DNA Communication. *PLOS ONE* **11**, e0152774 (2016).
9. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**, 239 (2016).
10. Takahashi, C. N., Nguyen, B. H., Strauss, K. & Ceze, L. Demonstration of End-to-End Automation of DNA Data Storage. *Sci Rep* **9**, 4998 (2019).
11. Heckel, R., Mikutis, G. & Grass, R. N. A Characterization of the DNA Data Storage Channel. *Sci Rep* **9**, 9663 (2019).

12. Grass, R. N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angewandte Chemie International Edition* **54**, 2552–2555 (2015).
13. Jeong, J. *et al.* Cooperative sequence clustering and decoding for DNA storage system with fountain codes. *Bioinformatics* **37**, 3136–3143 (2021).
14. Ren, Y. *et al.* DNA-Based Concatenated Encoding System for High-Reliability and High-Density Data Storage. *Small Methods* **6**, 2101335 (2022).
15. Anavy, L. *et al.* Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nature Biotechnology* **37**, 1229–1236 (2019).
16. Bystrykh, L. V. Generalized DNA Barcode Design Based on Hamming Codes. *PLoS ONE* **7**, e36852 (2012).
17. Doroschak, K. *et al.* Rapid and robust assembly and decoding of molecular tags with DNA-based nanopore signatures. *Nat Commun* **11**, 5454 (2020).
18. Guo, J. *et al.* Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences* **105**, 9145–9150 (2008).
19. Naskar, P. K., Paul, S., Nandy, D. & Chaudhuri, A. DNA Encoding and Channel Shuffling for Secured Encryption of Audio Data. *Multimed Tools Appl* **78**, 25019–25042 (2019).
20. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345-349I (2017).
21. Lee, H. *et al.* Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage. *Nat Commun* **11**, 5246 (2020).

22. Goldman, N. *et al.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013).
23. Zhang, Y. *et al.* Programmable base editing of zebrafish genome using a modified CRISPR-Cas9 system. *Nature Communications* **8**, 1–5 (2017).
24. Wang, S. *et al.* Precise, predictable multi-nucleotide deletions in rice and wheat using APOBEC–Cas9. *Nature Biotechnology* **38**, 1460–1465 (2020).
25. Wu, Y. *et al.* CRISPR-BETS: a base-editing design tool for generating stop codons. *Plant Biotechnology Journal* **20**, 499–510 (2022).
26. Kouno, T. *et al.* Crystal structure of APOBEC3A bound to single-stranded DNA reveals structural basis for cytidine deamination and specificity. *Nat Commun* **8**, 15024 (2017).
27. Liu, Z. *et al.* Precise base editing with CC context-specificity using engineered human APOBEC3G-nCas9 fusions. *BMC Biology* **18**, 111 (2020).
28. Zong, Y. *et al.* Efficient C-to-T base editing in plants using a fusion of nCas9 and human APOBEC3A. *Nature Biotechnology* **36**, 950–953 (2018).
29. Li, C. *et al.* SWISS: multiplexed orthogonal genome editing in plants with a Cas9 nickase and engineered CRISPR RNA scaffolds. *Genome Biology* **21**, 1–15 (2020).
30. Xie, J. *et al.* ACBE, a new base editor for simultaneous C-to-T and A-to-G substitutions in mammalian systems. *BMC Biology* **18**, 1–14 (2020).
31. Kim, Y. B. *et al.* Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat Biotechnol* **35**, 371–376 (2017).
32. Qin, R. *et al.* Increasing fidelity and efficiency by modifying cytidine base-editing systems in rice. *The Crop Journal* **8**, 396–402 (2020).

33. Choi, Y. *et al.* High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Sci Rep* **9**, 6582 (2019).
34. Pan, C. *et al.* Rewritable two-dimensional DNA-based data storage with machine learning reconstruction. *Nat Commun* **13**, 2984 (2022).
35. Chandrasekaran, A. R., Levchenko, O., Patel, D. S., MacIsaac, M. & Halvorsen, K. Addressable configurations of DNA nanostructures for rewritable memory. *Nucleic Acids Research* **45**, 11459–11465 (2017).
36. Tabatabaei Yazdi, S. M. H., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A Rewritable, Random-Access DNA-Based Storage System. *Sci Rep* **5**, 14138 (2015).
37. Antkowiak, P. L. *et al.* Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. *Nature Communications* **11**, (2020).
38. Choi, Y. *et al.* DNA Micro-Disks for the Management of DNA-Based Data Storage with Index and Write-Once–Read-Many (WORM) Memory Features. *Advanced Materials* **32**, 2001249 (2020).
39. Chen, K., Zhu, J., Bošković, F. & Keyser, U. F. Nanopore-based dna hard drives for rewritable and secure data storage. *Nano Letters* **20**, 3754–3760 (2020).
40. Barrangou, R. *et al.* CRISPR Provides Acquired Resistance against Viruses in Prokaryotes. *Science* **315**, 1709–1712 (2007).
41. Koonin, E. V. & Makarova, K. S. Evolutionary plasticity and functional versatility of CRISPR systems. *PLoS Biol* **20**, e3001481 (2022).
42. Jinek, M. *et al.* A programmable dual RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).

43. Whinn, K. S., van Oijen, A. M. & Ghodke, H. Spy-ing on Cas9: Single-molecule tools reveal the enzymology of Cas9. *Current Opinion in Biomedical Engineering* **12**, 25–33 (2019).
44. Nishimasu, H. *et al.* Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* **156**, 935–949 (2014).
45. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
46. Lewis, C. A., Crayle, J., Zhou, S., Swanstrom, R. & Wolfenden, R. Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proceedings of the National Academy of Sciences* **113**, 8194–8199 (2016).
47. Martinez, T., Shapiro, M., Bhaduri-McIntosh, S. & MacCarthy, T. Evolutionary effects of the AID/APOBEC family of mutagenic enzymes on human gamma-herpesviruses. *Virus Evolution* **5**, vey040 (2019).
48. Powell, L. M. *et al.* A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **50**, 831–840 (1987).
49. Cannataro, V. L. *et al.* APOBEC-induced mutations and their cancer effect size in head and neck squamous cell carcinoma. *Oncogene* **38**, 3475–3487 (2019).
50. Insights into the Structures and Multimeric Status of APOBEC Proteins Involved in Viral Restriction and Other Cellular Functions. *Viruses* **13**, 497 (2021).
51. Conticello, S. G., Thomas, C. J. F., Petersen-Mahrt, S. K. & Neuberger, M. S. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Molecular Biology and Evolution* **22**, 367–377 (2005).

52. Ito, F., Fu, Y., Kao, S.-C. A., Yang, H. & Chen, X. S. Family-Wide Comparative Analysis of Cytidine and Methylcytidine Deamination by Eleven Human APOBEC Proteins. *J Mol Biol* **429**, 1787–1799 (2017).
53. Berger, G. *et al.* APOBEC3A Is a Specific Inhibitor of the Early Phases of HIV-1 Infection in Myeloid Cells. *PLoS Pathog* **7**, (2011).
54. Constantin, D., Dubuis, G., Conde-Rubio, M. del C. & Widmann, C. APOBEC3C, a nucleolar protein induced by genotoxins, is excluded from DNA damage sites. *The FEBS Journal* **289**, 808–831 (2022).
55. Warren, C. *et al.* Roles of APOBEC3A and APOBEC3B in human papillomavirus infection and disease progression. *Viruses* **2017**, (2017).
56. Logue, E. C. *et al.* A DNA Sequence Recognition Loop on APOBEC3A Controls Substrate Specificity. *PLoS One* **9**, e97062 (2014).
57. Kim, J. *et al.* Structural and Kinetic Characterization of *Escherichia coli* TadA, the Wobble-Specific tRNA Deaminase. *Biochemistry* **45**, 6407–6416 (2006).
58. Wolf, J. tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli. *The EMBO Journal* **21**, 3841–3851 (2002).
59. Jeong, Y. K. *et al.* Adenine base editor engineering reduces editing of bystander cytosines. *Nat Biotechnol* **39**, 1426–1433 (2021).
60. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
61. Jiang, G. *et al.* Molecular Mechanism of the Cytosine CRISPR Base Editing Process and the Roles of Translesion DNA Polymerases. *ACS Synth. Biol.* **10**, 3353–3358 (2021).

62. Cho, S.-I. *et al.* Targeted A-to-G base editing in human mitochondrial DNA with programmable deaminases. *Cell* **185**, 1764-1776.e12 (2022).
63. Kurt, I. C. *et al.* CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells. *Nature Biotechnology* **39**, 41–46 (2021).
64. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
65. Ran, F. A. *et al.* Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* **154**, 1380–1389 (2013).
66. Lee, S. *et al.* Single C-to-T substitution using engineered APOBEC3G-nCas9 base editors with minimum genome- and transcriptome-wide off-target effects. *Science Advances* **6**, eaba1773 (2020).
67. Calvo-Villamañán, A. *et al.* On-target activity predictions enable improved CRISPR–dCas9 screens in bacteria. *Nucleic Acids Research* **48**, e64 (2020).
68. Cui, L. *et al.* A CRISPRi screen in E. coli reveals sequence-specific toxicity of dCas9. *Nat Commun* **9**, 1912 (2018).
69. Lei, Z. *et al.* Detect-seq reveals out-of-protospacer editing and target-strand editing by cytosine base editors. *Nat Methods* **18**, 643–651 (2021).
70. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33**, 187–197 (2015).
71. Pauling, L. & Corey, R. B. Structure of the Nucleic Acids. *Nature* **171**, 346–346 (1953).

72. How Much Information is Stored in the Human Genome? Yevgeniy Grigoryev. — Ruslan Dmytrakovych. https://qqq.com.ua/articles-2/zaimstvovaniya/information-in-human-genome/.
73. Song, X. & Reif, J. Nucleic Acid Databases and Molecular-Scale Computing. *ACS Nano* **13**, 6256–6268 (2019).
74. Hao, Y., Li, Q., Fan, C. & Wang, F. Data Storage Based on DNA. *Small Structures* **2**, 2000046 (2021).
75. Paunescu, D., Puddu, M., Soellner, J. O. B., Stoessel, P. R. & Grass, R. N. Reversible DNA encapsulation in silica to produce ROS-resistant and heat-resistant synthetic DNA 'fossils'. *Nat Protoc* **8**, 2440–2448 (2013).
76. The Cost of Sequencing a Human Genome. *Genome.gov* https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.
77. Vishwakarma, R. High Density Data Storage In Dna Using An Efficient Message Encoding Scheme. *IJITCS* **2**, 41–46 (2012).
78. Zan, X., Xie, R., Yao, X., Xu, P. & Liu, W. *A robust and efficient DNA storage architecture based on modulation encoding and decoding*. http://biorxiv.org/lookup/doi/10.1101/2022.05.25.490755 (2022) doi:10.1101/2022.05.25.490755.
79. Chandak, S. *et al.* Improved read/write cost tradeoff in DNA-based data storage using LDPC codes. in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* 147–156 (IEEE, 2019). doi:10.1109/ALLERTON.2019.8919890.

80. Lenz, A. *et al.* Concatenated Codes for Recovery From Multiple Reads of DNA Sequences. in *2020 IEEE Information Theory Workshop (ITW)* 1–5 (2021). doi:10.1109/ITW46852.2021.9457675.
81. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
82. Meiser, L. C. *et al.* Reading and writing digital data in DNA. *Nature Protocols* **15**, 86–101 (2020).
83. Tan, G., Opitz, L., Schlapbach, R. & Rehrauer, H. Long fragments achieve lower base quality in Illumina paired-end sequencing. *Sci Rep* **9**, 2856 (2019).
84. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* **17**, 239 (2016).
85. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends in Genetics* **34**, 666–681 (2018).
86. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348–1365 (2021).
87. Steinbock, L. J. & Radenovic, A. The emergence of nanopores in next-generation sequencing. *Nanotechnology* **26**, 074003 (2015).
88. Wang, C. *et al.* dCas9-based gene editing for cleavage-free genomic knock-in of long sequences. *Nat Cell Biol* **24**, 268–278 (2022).
89. Komor, A. C. *et al.* Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Science Advances* **3**, eaao4774 (2017).

90. Liu, Y., Schiff, M., Marathe, R. & Dinesh-Kumar, S. P. Tobacco Rar1, EDS1 and NPR1/NIM1 like genes are required for N-mediated resistance to tobacco mosaic virus. *The Plant Journal* **30**, 415–429 (2002).
91. Bagchi, R. *et al.* Polyvalent guide RNAs for CRISPR antivirals. *iScience* **25**, 105333 (2022).
92. Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat Methods* **18**, 165–169 (2021).
93. Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet* **47**, 1067–1072 (2015).
94. Mitsis, P. G. & Kwagh, J. G. Characterization of the interaction of lambda exonuclease with the ends of DNA. *Nucleic Acids Research* **27**, 3057–3063 (1999).
95. Bogdanov, A. *et al.* PRESENT: An Ultra-Lightweight Block Cipher. in *Cryptographic Hardware and Embedded Systems - CHES 2007* (eds. Paillier, P. & Verbauwhede, I.) vol. 4727 450–466 (Springer Berlin Heidelberg, 2007).
96. Prabhakar, R., Chandak, S. & Tatwawadi, K. Implementation of Protograph LDPC error correction codes. (2020) doi:10.5281/zenodo.4016076.
97. Lv, G., Yang, M., Li, H., Li, M. & Liu, X. Two-step construction method of protograph-based AR4JA codes in deep space communication. in *2012 1st IEEE International Conference on Communications in China (ICCC)* 556–561 (2012). doi:10.1109/ICCChina.2012.6356946.
98. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197 (1981).

99. Sadremomtaz, A., Glass, R., Guerrero, J. *et al.* Digital data storage on DNA tape using CRISPR base editors. 2023.02.07.527074 Preprint at https://doi.org/10.1101/2023.02.07.527074 (2023).

100. Langley, P., Iba, W. & Thompson, K. An analysis of Bayesian classifiers. in vol. 90 223–228 (Citeseer, 1992).

101. Fang, Y., Bi, G., Guan, Y. L. & Lau, F. C. M. A Survey on Protograph LDPC Codes and Their Applications. *IEEE Communications Surveys & Tutorials* **17**, 1989–2016 (2015).

102. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).

103. Abdullah *et al.* Adenine Base Editing System for Pseudomonas and Prediction Workflow for Protein Dysfunction via ABE. *ACS Synth. Biol.* **11**, 1650–1657 (2022).

104. Rothgangl, T. *et al.* In vivo adenine base editing of PCSK9 in macaques reduces LDL cholesterol levels. *Nat Biotechnol* **39**, 949–957 (2021).

105. Kim, D. *et al.* Genome-wide target specificities of CRISPR RNA-guided programmable deaminases. *Nat Biotechnol* **35**, 475–480 (2017).

106. Wang, X. *et al.* Efficient base editing in methylated regions with a human APOBEC3A-Cas9 fusion. *Nature Biotechnology* **36**, 946–949 (2018).

107. APOBEC3G - DNA dC->dU-editing enzyme APOBEC-3G - Homo sapiens (Human) | UniProtKB | UniProt. https://www.uniprot.org/uniprotkb/Q9HC16/entry.

APPENDIX A: TEMPLATE 1.0 DNA SEQUENCE

5′-
ATCACGAGGCCCTTTCGTCTTCAAGAATTCTTTATAGAAAACGTTTTGAAGAAGAAGATGATCT
CTACTCGTTCTCATCGCGTACCACGAAGGTGTCCTACTATGTCTTCTCTCTTCTACTACTTACC
TACTCGGACTGCTTCACGGTCAACGTGGGGATGGATGATCCCACACCTCACACGCAGGAGAGAA
ACTCAGGTCCGACGATCACCTTCATGGCTAGTGGTAGATGTTGTGTGTGGCGCGAGAGAAAGCA
CTCGGAACTCGGAGACACTCGACTGGTTGCGACGATGACTGACGACTGCACGAAAAGCTGGAAC
TCGATTCGAATATCTCTCTTCGTGGGTGAGGAGGAGAAGTAAAAGAAAGCTTCGAGAGAGTACT
CTCGGGAGAAAGGTCGCTGTGTGGAGTTTACACGGCGCTCTTTCCGGTTTGATCTTGCACACTC
ACTAGTCCTCGAAAACCTCGTGGCTGTTTGCACACACACCCGCACACCCTGTTCCCTCGACTCA
TCACGAGTTCACGATACCGTGGATGCGTTGCGTTGTTTTGCGTTCCACACCACACGTTACTCTT
GTGGTCAATGTCACTCCGAGGATGTTTACGCACGCGTTTTCCCACCCACGATGTTGTACTCAAG
CTCAGCCTCGTTAAACGTGGATCCAAAGAGAACTGGGATTTCTAAAAGAGAGAGAAACTCGGCG
ATCACGGCCATCACAGCGGGTTTTACCTTTTGCGCTTTTGTCTTCGTTCGTCCCTACTCGTATT
GGTTCTCAGCATCGCCGGGGCTCCCTACCACACACCACGTTTTGATGATAGTTGACTCATCGAT
AAGCTTTAATGCGGTAGTTTATCA-3′

APPENDIX B: TEMPLATE 2.0 DNA SEQUENCE

5′-
ATCACGAGGCCCTTTCGTCTTCAAGAATTCTTTATAGAAAACGTTTTGAAGAAGAAGATGATCT
CTACTCTCGCCAGATCGACAGGATCATGGTGTCCTACTATGTCTTCTCTCTTCTACTACTTACC
TACTCCTCCAATCAAATCAGTCACTAGGGGATGGATGATCCCACACCTCACACGCAGGAGAGAA
ACTCTCTGGTCAGGGCTCGGACACTGGCTAGTGGTAGATGTTGTGTGTGGCGCGAGAGAAAGCA
CTCTCATTCACAGCAACTGCAGCAGGTTGCGACGATGACTGACGACTGCACGAAAAGCTGGAAC
TCATGGTCAACTCAATCCAAAATGGGTGAGGAGGAGAAGTAAAAGAAAGCTTCGAGAGAGTACT
CGTTCTCATCGCGTACCACGAAGGAGTTTACACGGCGCTCTTTCCGGTTTGATCTTGCACACTC
ATCAATAGTGTCATGGCATGTGGATGTTTACGCACGCGTTTTCCCACCCACGATGTTGTACTCT
CGGGAGAAAGGTCGCTGTGAGGCTGTTTGCACACACACCCGCACACCCTGTTCCCTCGACTCAT
CACGAGTTCACGATACCGTGGATGCGTTGCGTTGTTTTGCGTTCCACACCACACGTTACTCTTG
TGGTCAATGTCACTCCGAGGATCCAAAGAGAACTGGGATTTCTAAAAGAGAGAGAAACTCAAGC
TCAGCCTCGTTAAACGTGGGTTTTACCTTTTGCGCTTTTGTCTTCGTTCGTCCCTACTCGAACA
GATCATCAACCCATTAGGGGCTCCCTACCACACACCACGTTTTGATGATAGTTGACTCATTCAA
TCAAGCTGCAAAGGTGGTACGAGAGGAAGCTTCACACACCACCACGATCGGATACTCCTTTCAA
GACCTCAAGAACGAGGCTTGCGCACACCTCACACACGTGTTTGTGTTGTGTTACTCGCCTCATC
AGCAGAACAAGTTGGCGATCCGCACACGCACGTCACACCTATCTTACGTGTACTCTCATTCCAG
TCAATGTGGAAAGGGAAGAAAAGAAAGAGAAGAGAAAACTCAAAAGATGAACTCATCGATAAGC
TTTAATGCGGTAGTTTATCA-3′

APPENDIX C: TEMPLATE 3.0 DNA SEQUENCE

5′-
ATCACGAGGCCCTTTCGTCTTCAAGAATTCTTTATAGAAAACGTTTTGAAGAAGAAGATGATCT
CTACTCACTAGTCCTCGAAAACCTCGTGGTGTCCTACTATGTCTTCTCTCTTCTACTACTTACC
TACTCCTCCAATCAAATCAGTCACTAGGGGATGGATGATCCCACACCTCACACGCAGGAGAGAA
ACTCTCTGGTCAGGGCTCGGACACTGGCTAGTGGTAGATGTTGTGTGTGGCGCGAGAGAAAGCA
CTCTCATTCACAGCAACTGCAGCAGGTTGCGACGATGACTGACGACTGCACGAAAAGCTGGAAC
TCATGGTCAACTCAATCCAAAATGGGTGAGGAGGAGAAGTAAAAGAAAGCTTCGAGAGAGTACT
CGTTCTCATCGCGTACCACGAAGGAGTTTACACGGCGCTCTTTCCGGTTTGATCTTGCACACTC
ATCAATAGTGTCATGGCATGTGGATGTTTACGCACGCGTTTTCCCACCCACGATGTTGTACTCT
CGGGAGAAAGGTCGCTGTGAGGCTGTTTGCACACACACCCGCACACCCTGTTCCCTCGACTCAT
CACGAGTTCACGATACCGTGGATGCGTTGCGTTGTTTTGCGTTCCACACCACACGTTACTCTTG
TGGTCAATGTCACTCCGAGGATCCAAAGAGAACTGGGATTTCTAAAAGAGAGAGAAACTCAAGC
TCAGCCTCGTTAAACGTGGGTTTTACCTTTTGCGCTTTTGTCTTCGTTCGTCCCTACTCGAACA
GATCATCAACCCATTAGGGGCTCCCTACCACACACCACGTTTTGATGATAGTTGACTCATTCAA
TCAAGCTGCAAAGGTGGTACGAGAGGAAGCTTCACACACCACCACGATCGGATACTCGATTCGA
ATATCTCTCTTCGAGGCTTGCGCACACCTCACACACGTGTTTGTGTTGTGTTACTCGCCTCATC
AGCAGAACAAGTTGGCGATCCGCACACGCACGTCACACCTATCTTACGTGTACTCTCATTCCAG
TCAATGTGGAAAGGGAAGAAAAGAAAGAGAAGAGAAAACTCAAAAGATGAACTCATCGATAAGC
TTTAATGCGGTAGTTTATCA-3′