STRUCTURAL MECHANISM OF SUBSTRATE SPECIFICITY IN HUMAN CYTIDINE DEAMINASE FAMILY APOBEC3S

A Dissertation Presented

By

SHURONG HOU

Submitted to the Faculty of the
University of Massachusetts Graduate School of Biomedical Sciences,
Worcester in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

April 28th, 2020

BIOCHEMISTRY & MOLECULAR PHARMACOLOGY

STRUCTURAL MECHANISM OF SUBSTRATE SPECIFICITY IN HUMAN CYTIDINE DEAMINASE FAMILY APOBEC3S

A Dissertation Presented By
SHURONG HOU

This work was undertaken in the Graduate School of Biomedical Sciences
BIOCHEMISTRY & MOLECULAR PHARMACOLOGY PROGRAM

The signatures of the Dissertation Defense Committee signify completion and approval as to style and content of the Dissertation

_____
Celia A. Schiffer, Ph.D., Thesis Advisor

_____
Paul R. Thompson, Ph.D., Member of Committee

_____
Francesca Massi, Ph.D., Member of Committee

_____
Andrei A. Korostelev, Ph.D., Member of Committee

_____
Ivet Bahar, Ph.D., External Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

_____
Daniel N. Bolon, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the School.

_____
Mary Ellen Lane, Ph.D.
Dean of the Graduate School of Biomedical Sciences

APRIL 28th, 2020

Page #

1 Table of Contents

DEDICATION

This thesis is dedicated to the memory of my grandmother, Cuiying Sun. I love you and miss you every day. But I know you saw this process from heaven through to its completion, offering the support to make it possible, as well as plenty of friendly encouragement.

My thanks to …

…my parents, for dealing with me being worlds away, for supporting and inspiring me to become not only an independent scientist but also a better woman.

…my husband, for your kindness and extensive support through good times and bad, I love you and cherish you forever.

…my son, for bringing happiness and joy into my life, you will always be my sunshine.

ABSTRACT

APOBEC3s (A3s) are a family of human cytidine deaminases that play important roles in both innate immunity and cancer. A3s protect host cells against retroviruses and retrotransposons by deaminating cytosine to uracil on foreign pathogenic genomes. However, when mis-regulated, A3s can cause heterogeneities in host genome and thus promote cancer and the development of therapeutic resistance. The family consists of seven members with either one (A3A, A3C and A3H) or two zinc-binding domains (A3B, A3D, A3D and A3G). Despite overall similarity, A3 proteins have distinct deamination activity and substrate specificity. Over the past years, several crystal and NMR structures of apo A3s and DNA/RNA-bound A3s have been determined. These structures have suggested the importance of the loops around the active site for nucleotide specificity and binding. However, the structural mechanism underlying A3 activity and substrate specificity requires further examination.

Using a combination of computational molecular modeling and parallel molecular dynamics (pMD) simulations followed by experimental verifications, I investigated the roles of active site residues and surrounding loops in determining the substrate specificity and RNA versus DNA binding among A3s. Starting with A3B, I revealed the structural basis and gatekeeper residue for DNA binding. I also identified a unique auto-inhibited conformation in A3B that restricts access to the active site and may underlie lower catalytic activity compared to the highly similar A3A. Besides, I investigated the structural mechanism of substrate specificity and ssDNA binding conformation in A3s. I found an interdependence between substrate conformation and specificity. Specifically, the linear DNA conformation helps accommodate CC dinucleotide motif while the U-

shaped conformation prefers TC. I also identified the molecular mechanisms of substrate sequence specificity at -1' and -2' positions. Characterization of substrate binding to A3A revealed that intra-DNA interactions may be responsible for the specificity in A3A. Finally, I investigated the structural mechanism for exclusion of RNA from A3G catalytic activity using similar methods.

Overall, the comprehensive analysis of A3s in this thesis shed light into the structural mechanism of substrate specificity and broaden the understanding of molecular interactions underlying the biological function of these enzymes. These results have implications for designing specific A3 inhibitors as well as base editing systems for gene therapy.

PREFACE

Publications contained in this thesis:

- **Hou S**, Silvas TV, Leidner F, Nalivaika EA, Matsuo H, Kurt Yilmaz N, Schiffer CA. "Structural analysis of the active site and DNA binding of human cytidine deaminase APOBEC3B." *Journal of Chemical Theory and Computation* 15.1 (2018): 637-647.

- Silvas TV, **Hou S**, Myint W, Nalivaika EA, Somasundaran M, Kelch BA, Matsuo H, Kurt Yilmaz N, Schiffer CA. "Substrate sequence selectivity of APOBEC3A implicates intra-DNA interactions." *Scientific Reports* 8.1 (2018): 7511.

- Solomon WC, Myint W, **Hou S**, Kanai T, Tripathi R, Kurt Yilmaz N, Schiffer CA, Matsuo H. "Mechanism for APOBEC3G catalytic exclusion of RNA and non-substrate DNA." *Nucleic Acids Research* 47.14 (2019): 7676-7689.

- **Hou S**, Lee JM, Kurt Yilmaz N, Schiffer CA. “Structural mechanism of substrate specificity in human cytidine deaminase family APOBEC3s Z1 domains.” In preparation for submission to *The Journal of Biological Chemistry.*

- Maiti A, Myint W, Delviks-Frankenberry KA, **Hou S**, Rodriguez CS, Kurt Yilmaz N, Pathak VK, Schiffer CA, Matsuo H. “Crystal structure of full-length APOBEC3G with a dinucleotide bound reveals domain orientation and a likely ssDNA-binding channel.” Under review at *Nucleic Acids Research.*

Additional publications from my graduate study:

- Prachanronarong KL, Canale AS, Liu P, Somasundaran M, **Hou S**, Poh YP, Han T, Zhu Q, Renzette N, Zeldovich KB, Kowalik TF, Kurt Yilmaz N, Jensen JD, Bolon DNA, Marasco WA, Finberg RW, Schiffer CA, Wang JP. "Mutations in influenza A virus neuraminidase and hemagglutinin confer resistance against a broadly neutralizing hemagglutinin stem antibody." *Journal of Virology* 93.2 (2019): e01639-18.

- Avnir Y, Prachanronarong KL, Zhang Z, **Hou S**, Peterson EC, Sui J, Zayed H, Kurella VB, McGuire AT, Stamatatos L, Hilbert BJ, Bohn MF, Kowalik TF, Jensen JD, Finberg RW, Wang JP, Goodall M, Jefferis R, Zhu Q, Kurt Yilmaz N, Schiffer CA, Marasco WA. "Structural determination of the broadly reactive anti-IGHV1-69 anti-idiotypic antibody G6 and its Idiotope." *Cell Reports* 21.11 (2017): 3243-3255.

- Timm J, Kosovrasti K, Henes M, Leidner F, **Hou S**, Ali A, Kurt Yilmaz N, Schiffer CA. “Molecular and structural mechanism of pan-genotypic HCV NS3/4A protease inhibition by glecaprevir.” *ACS Chemical Biology* 15.2 (2020): 342-352.

- Cai EP, Ishikawa Y, Zhang W, Leite NC, Li J, **Hou S**, Kiaf B, Hollister-Lock J, Kurt Yilmaz N, Schiffer CA, Melton DA, Kissler S, Yi P. “Genome scale in vivo CRISPR screen identifies RNLS as a modifier of beta cell vulnerability in type 1 diabetes.” Under revision for second review by *Nature Medicine*

- Ejemel M, Li Q, **Hou S**, Schiller ZA, Wallace A, Amcheslavsky A, Kurt Yilmaz N, Toomey JR, Schneider R, Close BJ, Chen DY, Conway HL, Mohsan S, Cavacini LA, Klempner MS, Schiffer CA, Wang Y, “IgA MAb blocks SARS-CoV-2 Spike-ACE2 interaction providing mucosal immunity.” Submitted *to Nature Communications*

Chapter I describes the background of my thesis work.

Chapter II is a collaborative study that has been previously published as:

Hou S, Silvas TV, Leidner F, Nalivaika EA, Matsuo H, Kurt Yilmaz N, Schiffer CA. "Structural analysis of the active site and DNA binding of human cytidine deaminase APOBEC3B." *Journal of Chemical Theory and Computation* 15.1 (2018): 637-647.

Contribution from Shurong Hou:

I devised the concept of this manuscript. I performed the cloning, expressing, purifying A3B-CTD proteins and mutants for this study. I performed fluorescence anisotropy based binding assays and the analysis of the data with assistance from Ellen A. Nalivaika for this study. I performed all the molecular modeling and molecular dynamics simulations and the analysis of the data for this study. I created all figures and tables for this manuscript. I interpreted the data and wrote the manuscript with the assistance of Nese Kurt-Yilmaz and Celia A. Schiffer.

Chapter III is a collaborative study that is in preparation:

Hou S, Lee JM, Kurt Yilmaz N, Schiffer CA. “Structural mechanism of substrate specificity in human cytidine deaminase family APOBEC3s Z1 domains.” In preparation for submission to *The Journal of Biological Chemistry*.

Contribution from Shurong Hou:

I devised the concept of this manuscript. I performed all the molecular modeling and molecular dynamics simulations and the analysis of the data for this study. I created all figures and tables for this manuscript. I interpreted the data and wrote the manuscript with the assistance of Nese Kurt-Yilmaz and Celia A. Schiffer.

Chapter IV is a collaborative study that has been previously published as:

Silvas TV, **Hou S**, Myint W, Nalivaika EA, Somasundaran M, Kelch BA, Matsuo H, Kurt Yilmaz N, Schiffer CA. "Substrate sequence selectivity of APOBEC3A implicates intra-DNA interactions." *Scientific Reports* 8.1 (2018): 7511.

Contribution from Shurong Hou:

I contributed to the structural analysis and molecular modeling of the intra-DNA interactions based on A3A-ssDNA crystal structure for this study.

Chapter V is a collaborative study that has been previously published as:

Solomon WC, Myint W, **Hou S**, Kanai T, Tripathi R, Kurt Yilmaz N, Schiffer CA, Matsuo H. "Mechanism for APOBEC3G catalytic exclusion of RNA and non-substrate DNA." *Nucleic Acids Research* 47.14 (2019): 7676-7689.

Contribution from Shurong Hou:

I performed the molecular modeling and molecular dynamics simulations of A3G/A3A and the analysis of the data for this study. I created figure 5.7 and 5.8 in this study. I wrote the method for the molecular modeling and molecular dynamics simulations.

Chapter VI describes the conclusions and future directions.

LIST OF TABLES

Pages #

LIST OF FIGURES

Pages #

LIST OF THIRD PARTY COPYRIGHTED MATERIAL

Figure 1.3: Adapted from Nature, volume 533, pages 420–424. Alexis C. Komor et al. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Copyright 2016, with permission from Springer Nature and Copyright Clearance Center. (License Number: 4806770146684).

1 CHAPTER I INTRODUCTION

1.1 APOBEC3s

1.1.1 The human cytidine deaminase family APOBEC3

APOBEC3s (A3s, apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3) is a family of human cytidine deaminases that catalyze the deamination of cytosine to uracil in the single stranded DNA(ssDNA) or ssRNA[1-5]. The A3 family consists of seven members (A3A, A3B, A3C, A3D, A3F, A3G and A3H): three of these enzymes (A3A, A3C, A3H) have a single zinc-binding (Z) domain while the other four (A3B, A3D, A3F, A3G) have two Z domains. The two-domain A3s have a catalytically active C terminal domain (CTD) and a pseudo-catalytic N terminal domain (NTD) that binds to nucleic acids but does not have deaminase activity (**Figure 1.1A**) Although NTDs have no deamination activity, they appear to be important for regulating the catalytic activity through increasing ssDNA binding affinity and promoting oligomerization[6]. The Z domains of A3s can be separated into Z1, Z2, and Z3 phylogenetic groups, which are defined by conserved amino acid differences (**Figure 1.1C**). Specifically, Z1 group is comprised of A3A, A3B-CTD and A3G-CTD; Z2 has A3B-NTD, A3C, A3D-NTD, A3D-CTD, A3F-NTD, A3F-CTD, A3G-NTD and A3G-CTD; while Z3 has only A3H, which has seven different human haplotypes.

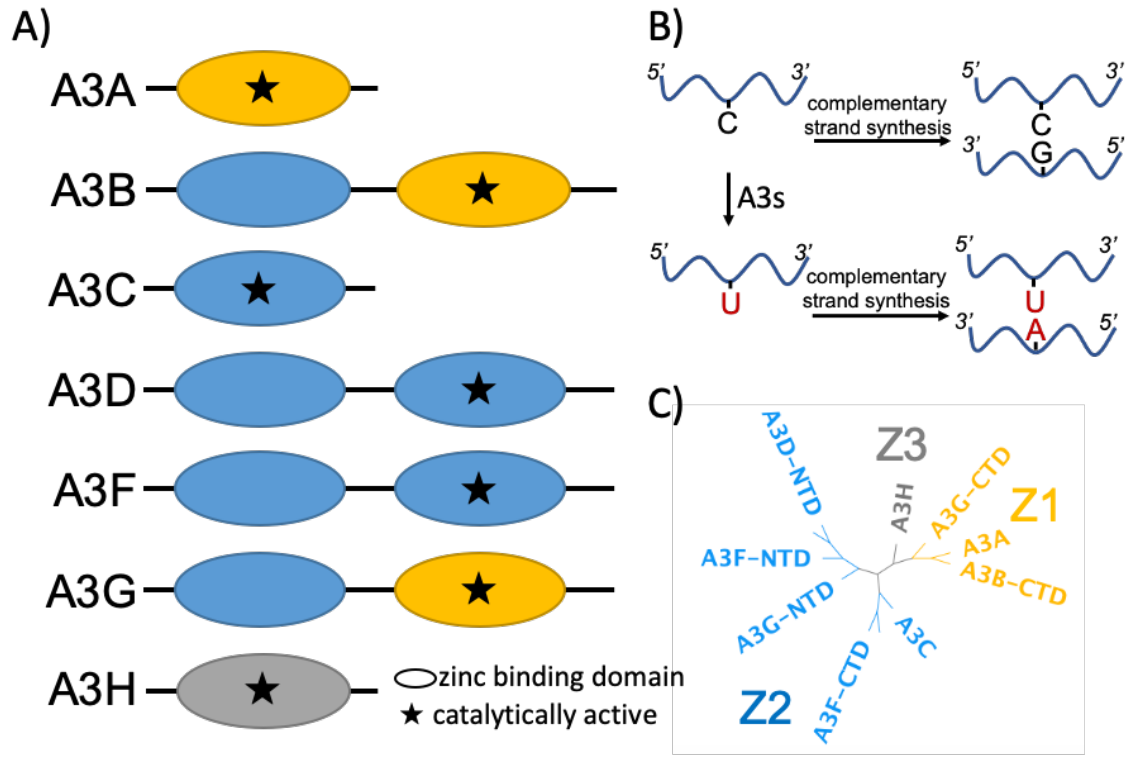

Figure 1.1: The APOBEC3 family of human cytidine deaminases.
A) Domain organization of the seven human APOBEC3s. N-terminal and C-terminal domains are represented by ovals. Active domains are marked with a star. Z1 domains are in orange, Z2 domains are in blue, and Z3 domain is in gray. B) A schematic cartoon for the outcome of A3 catalytic activity. A3s deaminate C to U on ssDNA and thus cause G to A mutations on the complementary strand. C) Phylogenic tree of A3 Z domains.

1.1.2 The human APOBEC superfamily

A3 family belongs to the APOBEC superfamily, which consists of 11 members, including activation-induced cytidine deaminase (AID), APOBEC1, APOBEC2, APOBEC3 and APOBEC4[2, 4, 7]. All these enzymes have a conserved zinc-binding motif (Cys/His)-Xaa-Glu-$Xaa_{23\sim28}$-Pro-Cys-$Xaa_{2\sim4}$-Cys, where X represents any amino acid. The active site zinc is tetrahedrally coordinated with the His and Cys residues and an additional water. The catalytic activity of the family is the deamination of cytosine to uracil in single strand polynucleotides (DNA/RNA)[2].

1.1.3 The conserved deamination reaction mechanism

The APOBEC superfamily shares a conserved deamination reaction mechanism (**Figure 1.2**). The reaction mechanism was proposed by studies of bacterial[1, 8] and yeast[9-11] cytidine deaminase. First, an active site water molecule is deprotonated by the catalytic glutamate carboxylic acid side chain, which acts as a general acid/base (Figure 1.2: step 1 to 2). The generated negatively charged hydroxide ion is stabilized by the positively charged zinc ion and attacks the C4 carbon on the pyrimidine ring of cytosine, yielding an unstable tetrahedral intermediate (Figure 1.2: step 3 to 4). A proton transfer reaction occurs from the protonated Glu carboxylic acid to the negatively charged N3 atom of the pyrimidine ring while in the tetrahedral intermediate state (Figure 1.2: step 4 to 5). The tetrahedral intermediate then collapses to give an uracil-nucleobase and an ammonia molecule (Figure 1.2: step 5 to 6). The catalytic glutamate is now deprotonated and ready to activate another water molecule to repeat the catalytic cycle (Figure 1.2: step 6 to 1).

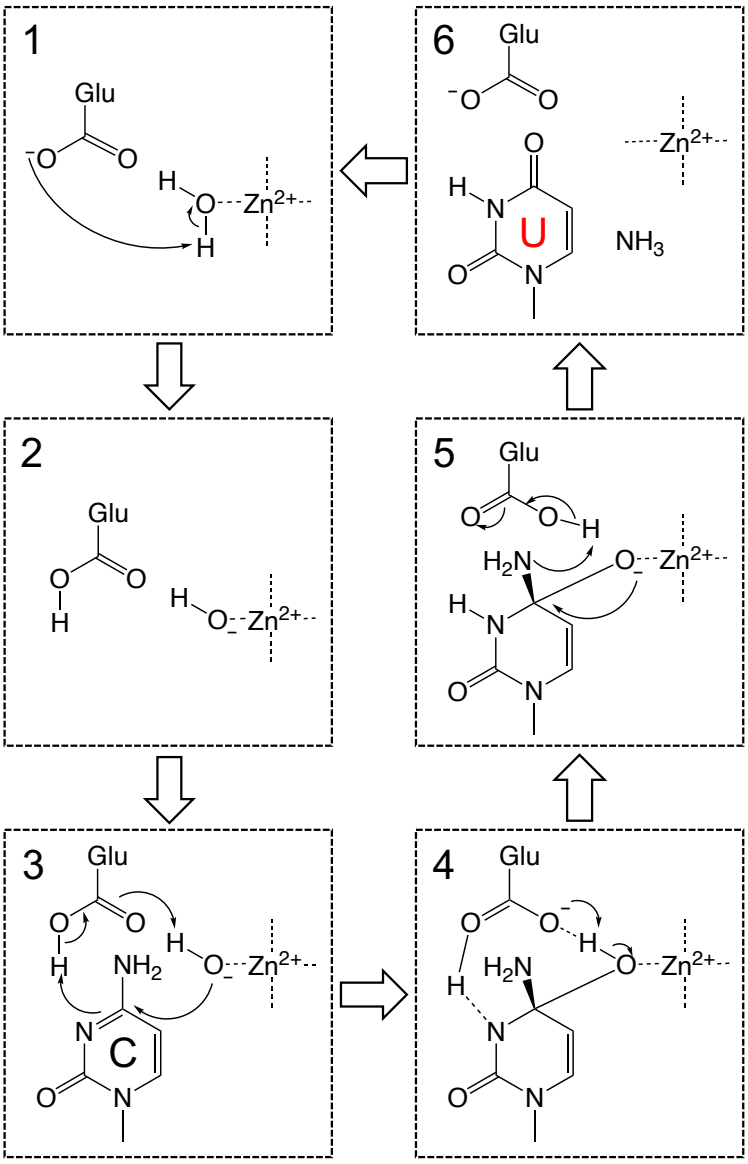

Figure 1.2: The deamination reaction mechanism of APOBEC family.

1.1.4 Other members of the human APOBEC superfamily

AID, which is encoded by the activation induced cytidine deaminase (AICDA) gene on human chromosome 12, plays an essential role in adaptive immune response. AID deaminates cytidines on ssDNA regions during the transcription of immunoglobulin genes. These deamination events regulate antibody diversification, specifically in the processes of class-switch recombination and somatic hypermutation[12]. AID can deaminate 5-methylcytosine (5meC) in CpG dinucleotides, suggesting that AID may be involved in epigenetic reprogramming and cell plasticity.

APOBEC1, which also lies on human chromosome 12, regulates lipid metabolism. APOBEC1 deaminates cytidine6666 in apolipoprotein B (apoB) mRNA transcript, which encodes a key player in lipid transport. This RNA editing creates a premature stop codon to produce a truncated apoB lipoprotein called apoB48 that is required for lipid transport from the intestines to other locations in the body[13-15]. In addition to editing the mRNA of apoB protein, APOBEC1 is also involved in DNA demethylation,[16, 17] and retrovirus restriction similar to A3s[18-20]. APOBEC2 (encoded by gene locus on chromosome 6) and APOBEC4 (encoded by gene locus on chromosome 1) have not been reported to have any deamination activity and their physiological function has remained elusive[7, 21].

1.2 The biological functions and applications of APOBEC3 family

1.2.1 A3s are critical for innate immunity against viral infections

A3s were first discovered through the identification of viral gene products that interfere with their function. For instance, A3G was first identified because of its role in restricting Vif (virion infectivity factor)-deficient HIV[22, 23]. Of the seven human A3s, A3D,

A3F, A3G and A3H can potently inhibit HIV-1 replication through hypermutation of the viral genome by deamination activity[24, 25]. Specifically, A3s deaminate cytosines to uracils on the single stranded (-) DNA synthesized during reverse transcription. The resulting uracils in the (-) DNA serve as a template for the reverse transcriptase during (+) DNA synthesis, leading to G to A mutations in (+) DNA (**Figure 1.1B**). The resulting hyper-mutated genome causes the virus to be defective for further replication. Inhibition of viral replication by the A3s has also been shown to occur through deamination-independent mechanisms. Specifically, A3s can directly bind viral genomic RNA or oligomerize on the template DNA during reverse transcription, resulting in a roadblock for the reverse transcriptase[26-29]. However, in the presence of Vif, A3s are targeted for proteasome degradation, which prevents A3s from restricting HIV replication. In addition to activity against retroviruses (including HIV), A3s are involved in the restriction of endogenous retrotransposons, especially LINE-1 elements. A3s also restrict DNA viruses including nuclear replicating ssDNA viruses such as adeno-associated virus[30] and dsDNA viruses such as hepatitis B virus, herpes viruses and HPV[31-34].

1.2.2 Consequences of mis-regulated cytidine deamination activity

A3 activity can be a double-edged sword. In addition to inducing mutations on single-stranded viral genomes, A3s can cause mutations in host genomes when localization and/or activity of A3s is mis-regulated. Although there is no known function of A3s that necessitates targeting genomic DNA, A3s are likely able to deaminate cytosines in single-stranded region of genomic DNA, such as the lagging strand of replication forks[35-38], the resected ends of double strand breaks[39-42], and non-transcribed strand during gene transcription. The A3 mutational signature, which is C to

T transition in TC context, has been observed in multiple cancer genomes[43-45]. These mutations may help promote tumor evolution and the development of therapeutic resistance[46]. Overexpressed A3s, especially A3A, A3B and A3H, have been shown to be a major endogenous source for mutations in various types of human cancer, such as breast, bladder, head and neck, cervical, and lung cancer[44, 45, 47, 48]. The A3s involved in cancer usually are able to localize to the nucleus: A3B and A3H haplotype I appear to localize to the nucleus, A3A and A3C are found throughout the cell while the other A3s remain in the cytoplasm[47, 49, 50]. Considering A3s' roles in cancer, discovering inhibitors that target A3s may benefit cancer therapeutics. The design and screening of first-in-class A3 inhibitors will be discussed in ***Appendix II***.

1.2.3 Using APOBEC3s as base editors to treat genetic diseases

The techniques to precisely and efficiently edit a specific DNA sequence have potential for use to correct disease-causing mutations in the genome of a living organism. Recently, CRISPR/Cas9, which plays a crucial role in bacterial defense against DNA viruses, has been modified as a powerful tool for genetic editing through the ability to create a dsDNA break (DSB) at a precise target location[51-58]. Cas9, a DNA endonuclease, binds single-guide RNA (sgRNA), forming CRISPR/Cas9 protein–RNA complexes. The CRISPR/Cas9 protein–RNA complexes generate a DSB at the locus specified by sgRNA. In response to DSB, cellular DNA repair processes, including non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ), can lead to gene disruption by introducing insertions, deletions, translocations and other DNA rearrangements at the DSB site[51, 59-61]. In the presence of a homology donor DNA template at the DSB site, the DNA surrounding the cleavage site can be replaced by

homology-directed repair (HDR), which can generate precise insertions, deletions, or any point mutation of interest[62, 63]. However, HDR is very inefficient (~0.5–5 %)[56, 64], and can lead to off-target mutations or unwanted changes such as indels, translocation, and rearrangements[65, 66].

To address these limitations, base-editors that can directly modify genomic DNA at single-base resolution without creating DSBs have been developed. These combine a modified Cas9 (catalytically inactive Cas9 (dCas9)/ nickase Cas9 (nCas9)) with a DNA-modifying enzyme. There are two classes of such base editors: cytosine base editors (CBEs)[67, 68] that alter a C•G base pair to a T•A base pair, and adenine base editors (ABEs)[69] that alter an A•T base pair to a G•C base pair. The first class of base-editors, CBEs have been constructed by linking a cytidine deaminase (AID, APOBEC1 or A3A) to dCas9/nCas9, together with a uracil DNA glycosylase inhibitor (UGI)[70] to disrupt the cellular uracil base excision repair pathway. Unlike CBEs, adenine base editors (ABEs) could not be developed by simply fusing an adenosine deaminase with dCas9/nCas9, because there is no known enzyme to deaminate adenine in DNA. Instead, an enzyme (TadA*), which can effectively deaminate adenine in ssDNA, has been engineered from *E. coli* tRNA adenosine deaminase (wtTadA) by performing extensive directed protein evolution. Heterodimeric TadA (wtTadA–TadA*) variants optimized to have high editing efficiency in human cells were fused with nCas9, creating ABEs. When these base editors are bound to their target DNA, dCas9(or nCas9) denatures the DNA duplex and generates an R-loop[52, 71] in which the DNA strand unpaired with the sgRNA exists as disordered single-stranded bubbles. The resulting ssDNA is targeted by the deaminase. Each deaminase causes a mutation (CBEs: a C•G to T•A mutation, ABEs: a A•T to G•C

mutation) in an ~5-bp window of ssDNA (positions ~4–8, counting the protospacer adjacent motif (PAM) as positions 21–23) generated by dCas9/nCas9.

Although both CBEs and ABEs have significantly improved editing efficiency and reduced indel formation compared to HDR, several studies have reported significant off-target effects of CBEs. In addition, CBEs have still problems in product purity and editing window length[51]. To overcome these problems, several versions of CBEs with the various scope and effectiveness of genome editing have been engineered by adding one more UGI for increasing product purity[72], generating high fidelity Cas9 for reducing off-target effects[73], or using different Cas nucleases[74-78] for narrower activity windows. However, product impurity and off-targeting by CBEs are mainly caused by unregulated activity (or overexpression) of cytidine deaminases. Therefore, additional studies to incorporate A3s with different specificities or to engineer A3s with desired specificity may allow to substantially reduce off-target effects and increase base editing efficiency.

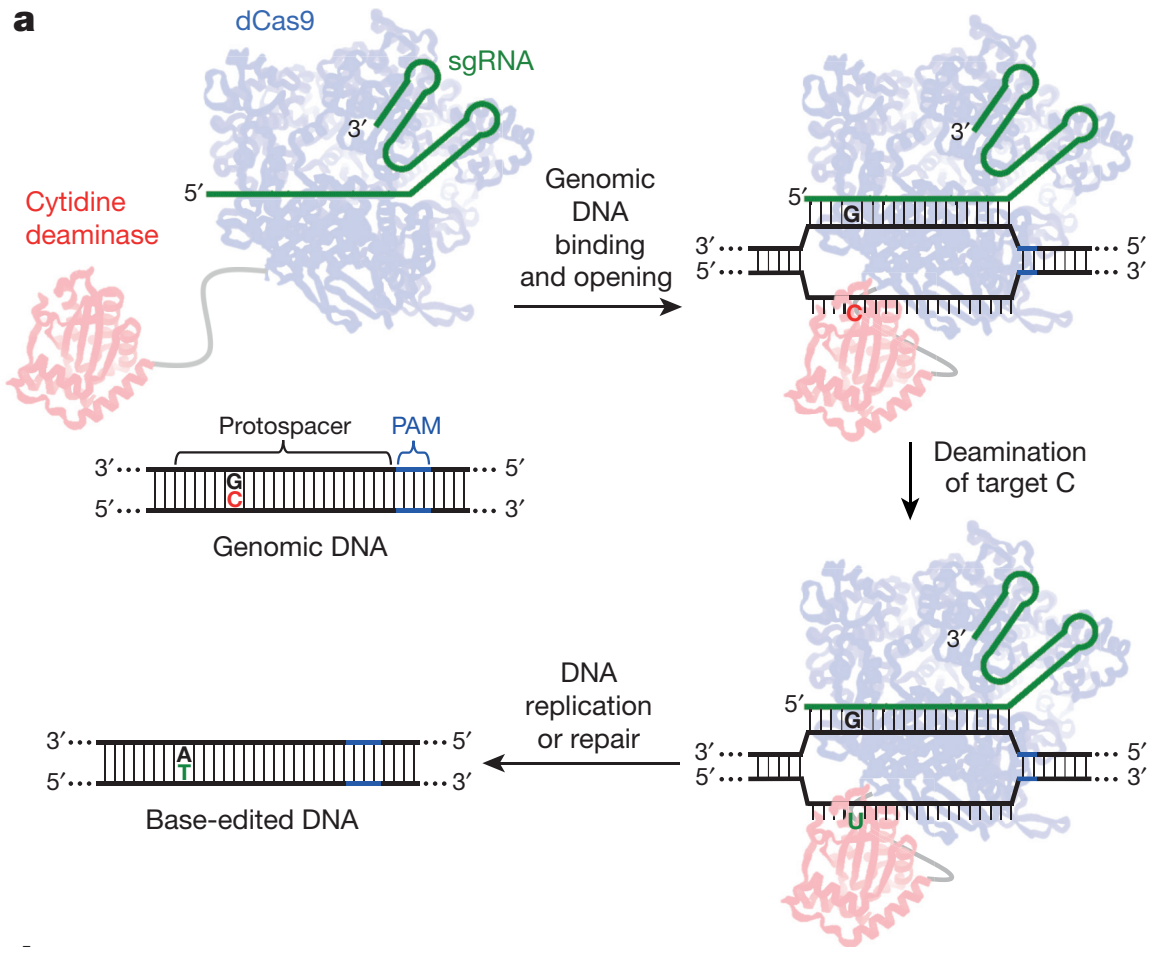

Figure 1.3: Generation of Cas9 fused cytosine base editors (CBE).

1.3 The structures of APOBEC3 family members

1.3.1 The apo structures

Over the past several years, crystal and NMR structures of human or primate A3 single domains (A3A, A3C, A3H; CTDs of A3B, A3F, A3G; NTDs of A3B, A3G) in the apo state have been determined by our group [79-85] and others [86-102]. In general, A3 proteins are structurally similar, including pseudo-catalytic NTDs. The overall A3 domain structure consists of six alpha-helices and five beta-strands with the zinc-binding region in the middle (**Figure 1.4**). The active site residues, for instance the catalytic glutamic acid and zinc coordinating residues, are highly identical among all A3 domains. Although the overall fold is conserved, subtle sequence differences among A3s have resulted in variations in loops length, structure, and flexibility as well as variations in surface charge, active site interactions, and oligomeric tendency. These variations underlie the functional characteristics of each A3 protein. Particularly, sequence differences in active site loops (loop 1, loop 3, loop 5 and loop 7) that surround the active site pocket of catalytically active domains mainly contribute to the differential substrate specificity, binding affinity and deamination activity for ssDNA, as well as the distinct physiological functions in A3s[103].

1.3.2 The nucleotide-bound A3 structures

Recently, our laboratory [104, 105], along with other groups, have determined the crystal structures of several A3–DNA complexes (A3A-DNA, chimeric A3B-CTD-DNA, A3G-CTD-DNA, A3F-DNA and rA3G-NTD-DNA) [99, 106-108]. Besides, three RNA-bound A3H structures[109-111] were solved. Among these structures, A3A-DNA (PDB: 5KEG; 5SWW), chimeric A3B-CTD-DNA (PDB: 5TD5), A3G-CTD-DNA (PDB: 6BUX) and

rA3G-NTD-DNA have ssDNA bound at the active site (**Figure 1.5**). These structures identified the substrate-binding conformation for deamination in the active site, revealed the critical residues for binding (e.g. gate-keeper residue His29 in A3A), and provided insights for substrate specificity at -1' position (e.g. the hydrogen bond interactions between Asp131 and -1' base in A3A). Active site loops (loop 1, 3, 5 and 7), which have direct contacts with ssDNA, have shown the most conformational changes compared to apo structures. The dynamics of these loops might be the key for defining the substrate specificities and functional variation among A3s.

In addition, these structures revealed differences in the conformation of bound ssDNA (U-shape in A3A and chimeric A3B; linear in A3G). Hence, the differences in the secondary structure of substrate DNA may provide fundamental insights into the mechanisms by which A3s recognize their specific substrates.

1.3.3 The structures of full-length double-domain A3 structures

Due to intrinsic tendency for oligomerization and poor solubility, the determination of full-length double-domain A3s structures is extreme challenging. Luckily, through engineering mutations guided by soluble CTDs and primate A3s, our group and another lab[112] have recently solved the structure of full-length A3G. Our group revealed the structure of human full-length A3G by mutating approximately 16% of residues to solubilize the full-length protein while the other group reported the structure of full-length A3G from rhesus monkey by mutating 4 amino acids to improve solubility. Combined with molecular modeling, these structures shed light on the role of pseudo-catalytic NTD and potential oligomerization in functions and mechanisms of wild type double-domain A3s. The detailed structural analysis of full-length A3G will be discussed in ***Appendix I***.

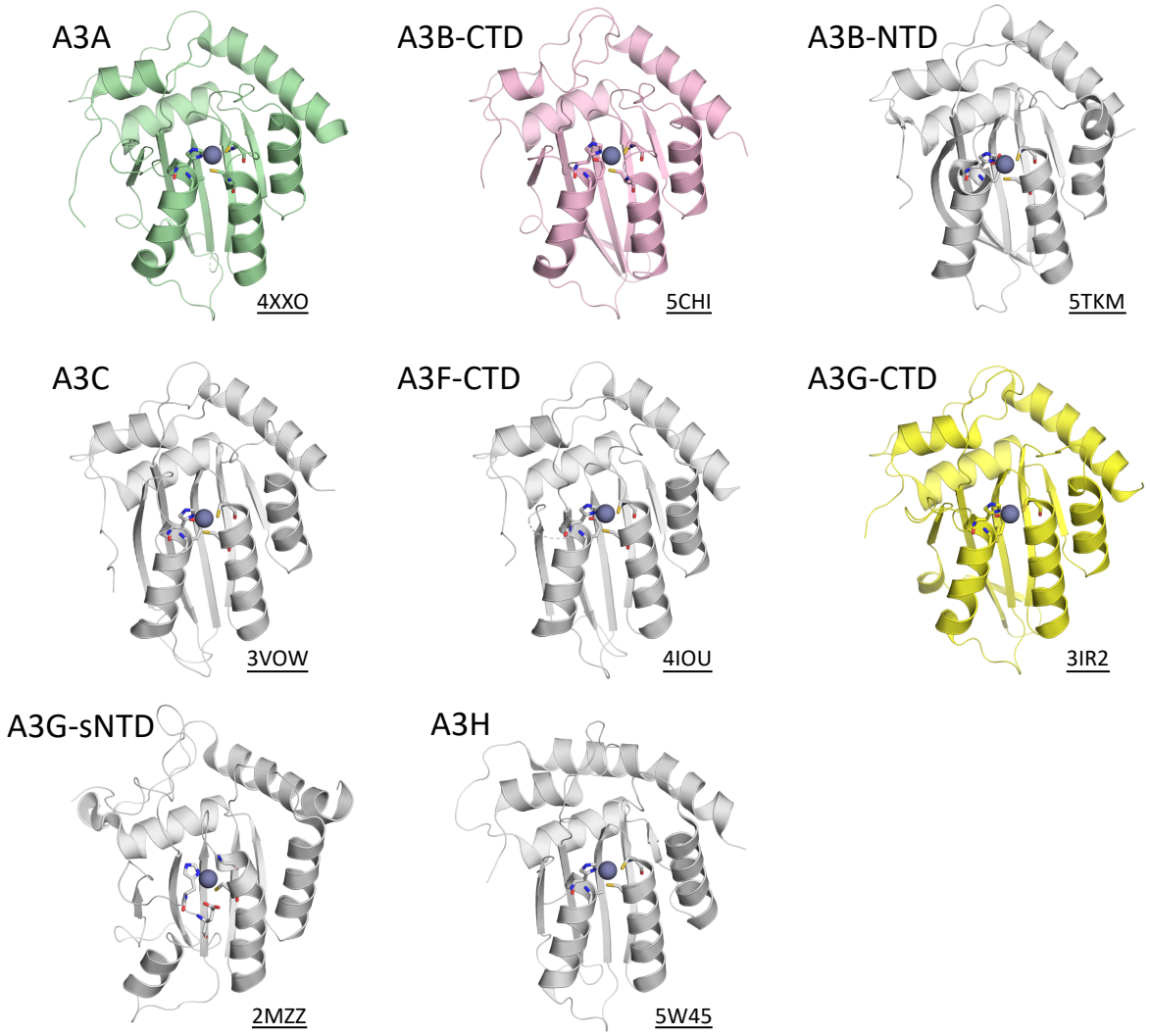

Figure 1.4: Apo APOBEC3 structures.
Currently determined apo structures of APOBEC3 proteins. A3A (PDB: 4XXO), A3B-CTD (PDB: 5CHI), A3B-NTD (PDB: 5TKM), A3C (3VOW), A3F-CTD (PDB: 4IOU), A3G-CTD (PDB: 3IR2), A3G-sNTD (PDB: 2MZZ) and A3H (5W45). The protein structures are shown in cartoon representation. The catalytic zinc is represented as a grey sphere. The zinc coordinating residues, Glu (inactive form has Ala for crystallization), His and two cysteines are shown as sticks. The three A3 proteins that are mainly discussed in this thesis are colored green (A3A), pink (A3B) and yellow (A3G).

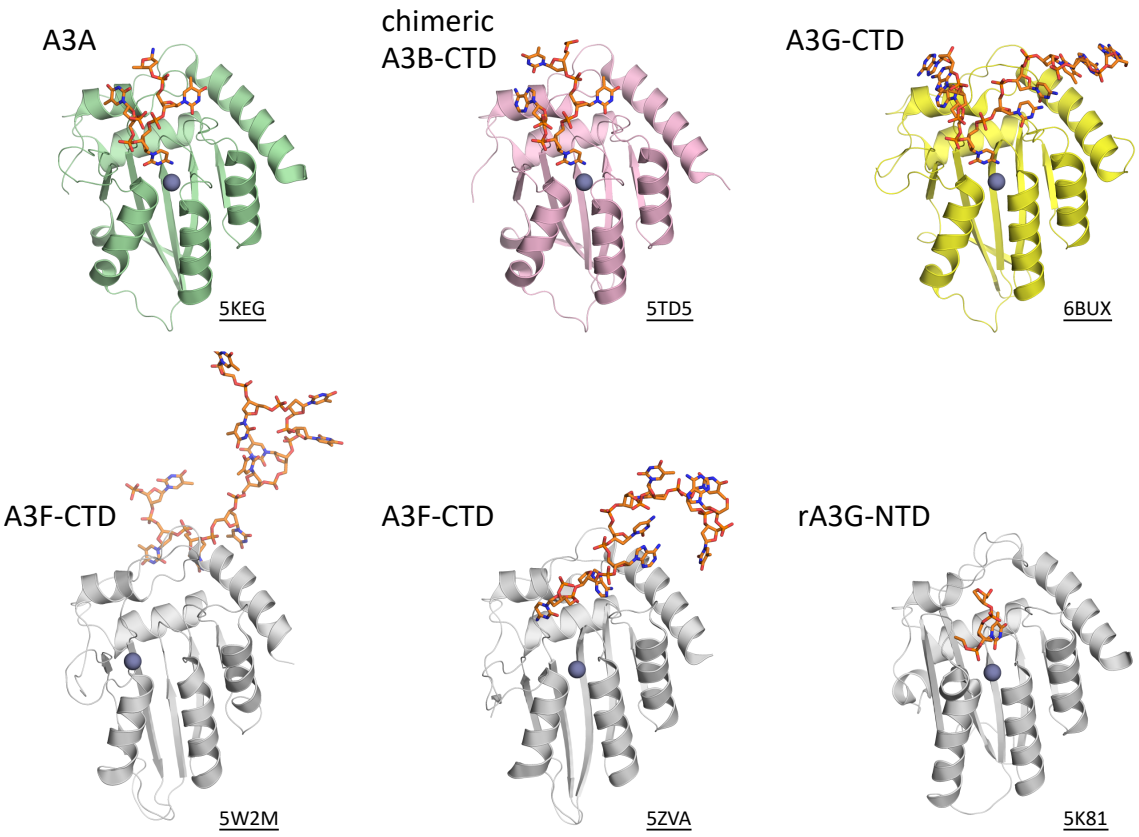

Figure 1.5: DNA-bound APOBEC3 structures.
Currently solved DNA-bound APOBEC3 structures. A3A (PDB: 5KEG), chimeric A3B-CTD-DNA (PDB: 5TD5), A3G-CTD-DNA (PDB: 6BUX), A3F-CTD-DNA (PDBs: 5W2M and 5ZVA) and rA3G-NTD-DNA (PDB: 5K83). A3 proteins are shown as cartoon representation while DNAs are depicted as orange sticks. Catalytic zinc is shown as grey sphere. Notice that for A3F-CTD DNA-bound structures, zinc and DNA are outside of the active site.

1.3.4 Substrate specificities of A3s

Although A3 domains share similar structure and overall fold, they have different catalytic activities, substrate preferences, and sequences specificities. First of all, not all A3 domains can deaminate cytidine. The NTDs alone have no catalytic activity despite having the conserved zinc binding motif as in CTDs. Besides, active A3s have different levels of deamination activity. The activity of A3A, which is the highest in A3 family, could be up to 5000-fold higher compared to the least active A3D[113].

In addition to cytidine, only certain A3s can deaminate methylated cytidines and thus may be involved in epigenetic regulation, particularly A3A, A3B and A3H[113]. Besides, similar to APOBEC1 which uses mRNAs as substrate, A3A and A3G of A3s have the ability and function to deaminate RNAs. A3A has been reported to bind[114] and deaminate ribo-cytidine with relatively lower activity compared to deoxyl-cytidine[115]. The deamination of RNA by A3G was observed in natural killer cells, lymphoma cell lines and CD8-positive T cells under specific conditions, such as cellular crowding and hypoxia, but not in cells under normal conditions[116]. However, the other A3s cannot deaminate RNA.

In general, A3s prefer to deaminate cytidine in a TC motif, except A3G which prefers CC. The preferred sequence motifs for different A3s including the flanking nucleotides next to the substrate cytidine according to currently published references are shown in **Table 1.1**[113, 117-120]. These substrate specificities may contribute to the differences in physiological functions among A3s and provide a mechanism for the evolution of the functionally overlapping but distinct APOBEC family.

Table 1.1: The preferred substrate sequence for deamination activity of human AID and APOBEC proteins.

	PREFERRED SUBSTRATE SEQUENCE
AID	(A/T)(A/G)**C**AA
APOBEC1	T**C**AA
APOBEC3A	(T/C)T**C**(A/G)
APOBEC3B	AT**C**(A/G)
APOBEC3C	(A/T)(C/T)**C**(A/G)
APOBEC3D	T**C**
APOBEC3F	TT**C**(A/T)
APOBEC3G	CC**C**(A/C/T)
APOBEC3H	T**C**

According to amino acid sequence alignment, active site loops (loop 1, 3, 5, and 7) have the most variation (**Figure 1.6**). These loops have not only different residues, but also different lengths. Swapping these loops among A3s has effects on both activity and specificity. For instance, exchanging loop 1 of A3A into A3B-CTD resulted in an order of magnitude increase in deamination activity; exchanging loop 7 of A3A into A3G-CTD altered substrate sequence preference from CC to more A3A-like TC. Therefore, comprehensive studies of the active site loops in A3s may reveal the structural mechanism of substrate specificities and activity.

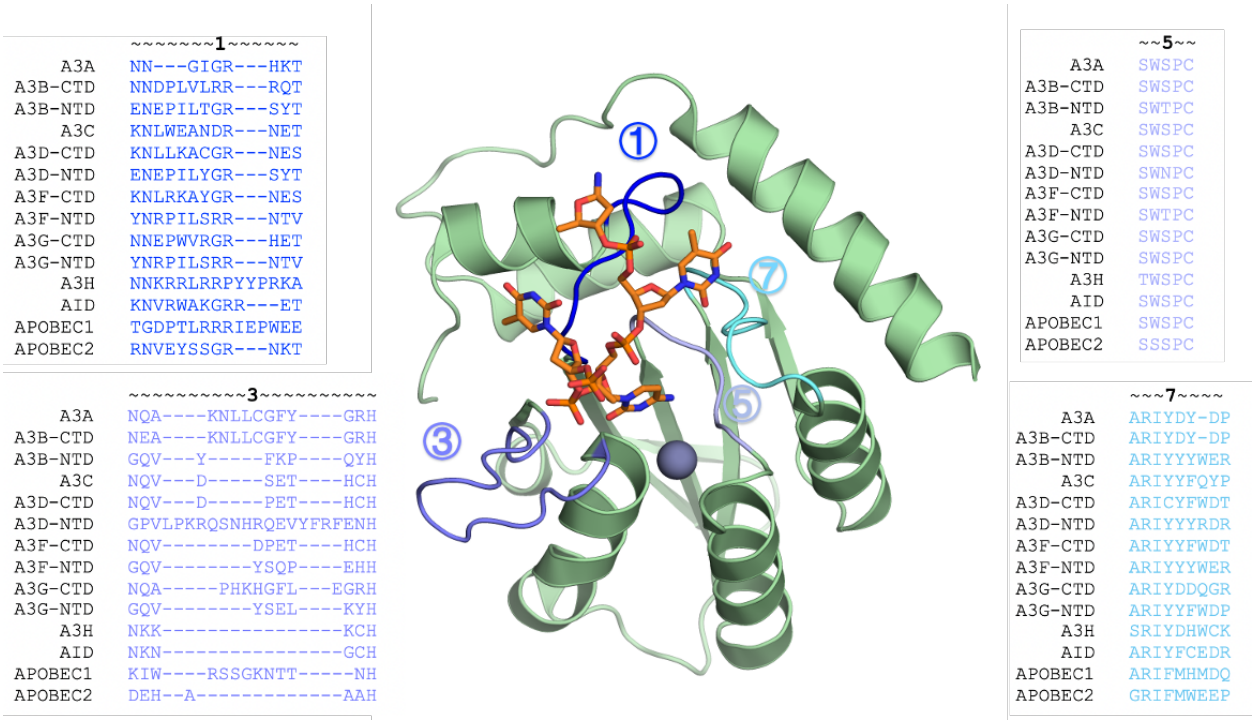

Figure 1.6: Sequence alignment and structural representation of active site loops in APOBEC3 family.

1.4 Protein modeling and dynamics in molecular recognition

1.4.1 Molecular modeling

The structure of proteins is the basis for understanding the molecular mechanism and interactions at the atomic level. There are multiple experimental methods to determine protein structures. The most common ones are X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and recently revolutionized cryogenic electron microscopy (cryo-EM). Using these methods, over 160 thousand structures (according to Protein Data Bank) have been solved and enabled breakthroughs in research and education. However, there are drawbacks of each of these methods: X-ray crystallography requires the formation of stable protein crystals; NMR spectroscopy requires high concentration/solubility and is largely limited to small proteins; cryo-EM is primarily suited for proteins with over 100 kilodalton of molecular weight. Improvements have been developed to overcome these problems. However, determining structures of proteins with poor solubility or of all mutants could still be very challenging and time consuming.

Molecular modeling or homology modeling, which is complementary to experimental methods discussed above, is a powerful tool to construct an atomic resolution model of a protein. According to amino acid sequence alignments, the method builds a protein structural model based on a related homolog that has experimentally determined three-dimensional structure[121, 122]. The quality or accuracy of the structural model is highly dependent on the quality of the sequence alignment and template structure[123]. Errors usually increase with decreasing sequence identity. The

regions without a template reference, for instance loop regions, are generally less accurate compared to the rest of the model[124, 125].

Molecular modeling provides valuable insights for studying the molecular properties of protein molecules and their interactions with binding partners (substrates, peptides, inhibitors and proteins). The findings or hypothesis derived from molecular models could be later verified through experimental studies. Molecular modeling in addition to experimental structural determination has significantly minimized the "structure knowledge gap" between the number of protein sequences and number of known structures[121]. The growing number of three-dimensional structures or models enables rational structure-based approaches in a broad range of applications in life science research, such as drug or antibody design, engineering protein specificities or protein-protein interactions[126]. For example, homology models help accelerate the high throughput virtual screening process in structure-based drug design. Virtual screening using homology models can provide insights to drug discovery before crystal structures are available or experimental high-throughput screening. In addition, molecular models may help guide the optimization of lead compounds in pharmaceutical development.

1.4.2 Molecular dynamics simulations

Proteins undergo conformational changes to perform their biological function. Hence, understanding protein dynamics is critical for understanding function, including molecular recognition. Molecular dynamics (MD) simulation is a computational method that enables studying protein dynamics by following conformational changes through a period of time. Proteins are typically simulated at the atomic level. The simplest simulation system usually consists of a single protein molecule solvated in water or with

other relevant molecules such as ligands or nucleotides. The potential energy of the system of atoms in MD simulations is calculated in terms of interatomic bonded (bond, angle, dihedral) and nonbonded potentials (van der Waals, electrostatics), which constitute the so-called force field.

$$E_{total} = E_{bonded} + E_{nonbonded}$$

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral}$$

$$E_{nonbonded} = E_{electrostatic} + E_{van\ der\ Waals}$$

There are many classic force fields that have been developed for studying biomolecules, for instance, CHARMM[127, 128], AMBER[129, 130], OPLS[131, 132]. These force fields assign parameters to each atom type and are usually paired with a particular solvent (e.g. water) model and simulation protocol[133].

MD simulations are most effective when coupled with cross-validations from experiments[134]. Besides, energy minimization in MD simulations allow the initial structural coordinates to adjust to minimize the energy of the protein[133], which is very helpful for assessing and optimizing models generated from molecular modeling. MD simulations have a broad range of applications in life science research, such as protein folding, substrate/inhibitor binding.

Nowadays, the speed by which MD simulations can be performed has been greatly increased thanks to the availability of supercomputing clusters and increased parallelization of calculations using powerful graphics processing unit (GPU) technology. Therefore, MD simulations can be not only run on much longer time scales, but also in replicates and for similar systems in parallel. Our lab has developed parallel MD simulations (*p*MD) as a method to characterize a series of related systems to inform

inhibitor design and understand resistance mechanisms in viral proteases[135-138], which we more recently applied to A3s[115, 139]. Specifically, pMD involves a series of parallel MD simulations on highly related systems (e.g. same enzyme with different substrates; same inhibitor with different protein variants, etc.) and thus allows detailed systemic comparisons among these systems. The trajectories generated in pMD are then analyzed to characterize and compare using metrics including root-mean-squared fluctuations (RMSFs), intermolecular interactions (hydrogen bonds, vdW contacts), electrostatics surface analysis. Examples of these analyses are shown in **Figure 1.7**. Detailed comparisons among pMDs enable generating novel hypotheses and proposing mechanisms, which can be supported or verified with experimental results. Thus, MD simulation is a valuable method in revealing the underlying molecular details of experimental observations in protein biology.

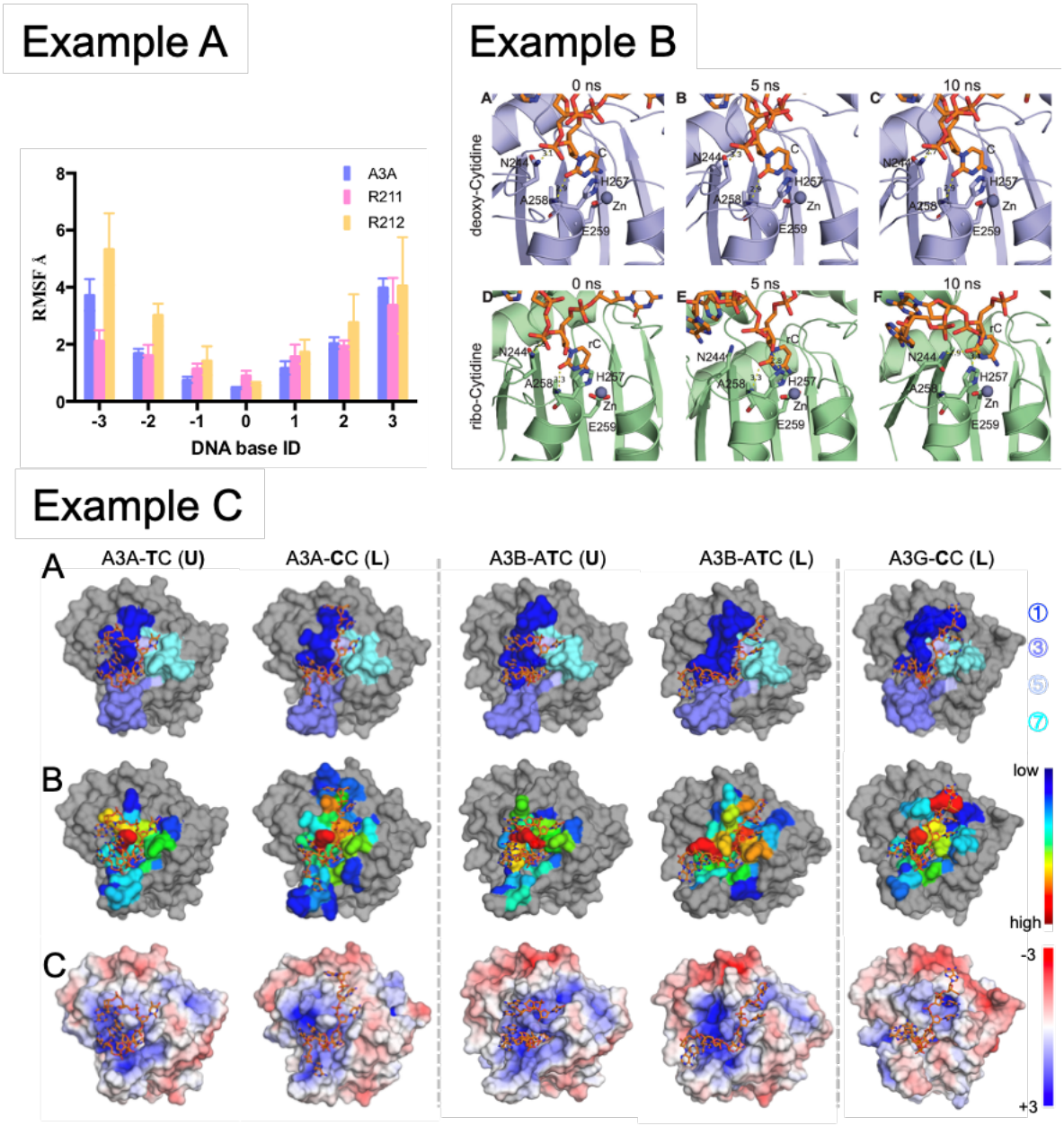

Figure 1.7: Examples of the types of analyses that can be performed with pMD. Example A. (Figure from Chapter II) RMSFs of ssDNA in A3B R211 or R212 models compared to A3A crystal structure. The results suggest that A3B uses R211 to coordinate ssDNA binding. **Example B**. (Figure from Chapter V) Hydrogen bond analyses between substrate deoxycytidine or non-substrate ribocytidine and A3G protein. **Example C**. (Figure from Chapter III) vdW and electrostatics surface analysis comparing A3A, A3B and A3G.

1.4.3 Molecular modeling and dynamics in understanding A3 specificity

Molecular dynamics simulations have been used for studying the biological functions and substrate binding in APOBEC superfamily (APOBEC1[140], A3B[141], A3G[142] and AID[143]). For instance, molecular modeling and simulations combined with experimental assays have helped reveal the possible DNA-bound conformations[143] as well as key loop for substrate recognition[144] in AID. The low solubility, tendency for oligomerization and low DNA affinity of certain A3 proteins have required introducing mutations to be able to structurally and biochemically characterize these proteins in vitro [79-81, 83-85, 92, 96, 99, 104, 106], or prevented such characterization especially for NTDs and full-length A3s. Hence, combining experimental structures with computational molecular modeling, verified by simulations and experimental analysis, can provide insights into DNA binding and specificity of A3s.

The active site loops of A3, which are critical for substrate binding and specificity, are very flexible. These loops undergo conformational changes to accommodate substrate binding, as shown in both A3A and A3G ssDNA-bound structures compared to apo structures. Therefore, studying the motion of these loops using molecular dynamics simulations may help reveal the underlying molecular mechanisms for substrate recognition as well as varying specificities.

1.5 SCOPE OF THE THESIS

The deamination activity of A3s contributes to restriction of viruses including HIV, but causes somatic mutation in many cancers. Due to the roles of A3s in viral infection and cancer, a better understanding of the mechanism by which A3s recognize target

nucleic acids and regulate their catalytic activity is critical for developing more effective antivirals and cancer therapeutics. The sequence and structural analysis of A3s have revealed that substrate specificity and binding affinity, and catalytic rate of A3s depend on differences in active site loops. However, the mechanism by which active site residues and loops regulate nucleotide specificity and catalytic activity remains still elusive. This thesis attempts to understand the structural mechanism of substrate specificity among A3s using molecular modeling of the available A3 structures, experimental mutational analysis, and parallel molecular dynamics (pMD). Specifically, the aims are to investigate the role of active site residues and surrounding loops that determine substrate specificity and RNA versus DNA binding, and to provide guidelines for developing specific inhibitors against A3s or engineering A3s with different specificities. Chapter I provides the current comprehension of the functional, biochemical, and structural mechanism for A3s.

Chapter II: Structural Analysis of the Active Site and DNA Binding of Human Cytidine Deaminase APOBEC3B: Using molecular modeling and simulations, further verified by experimental binding assays, this chapter elucidates the molecular mechanism of DNA recognition by A3B and how A3B-CTD structurally regulates its catalytic activity compared to the highly similar A3A.

Chapter III: Structural Mechanism of Substrate Specificity Among A3s: Using computational modeling and structural analysis of DNA-bound A3 complexes, this chapter reveals the structural mechanism of substrate specificities for A3A, A3B and A3G.

Chapter IV: Substrate Sequence Selectivity of APOBEC3A Implicates Intra-DNA Interactions: Using systematic studies of A3A binding with different substrate sequences with experimental binding assays, this chapter discusses the substrate sequence specificity and underlying molecular mechanism in A3A.

Chapter V: Mechanism for APOBEC3G Catalytic Exclusion of RNA and Non-substrate DNA: Using NMR and molecular dynamics simulations in combination with deamination assays, this chapter identifies the molecular mechanism for the exclusion of non-substrate ribo-cytidine compared to deoxy-cytidine in A3G-CTD.

2 CHAPTER II: Structural Analysis of the Active Site and DNA Binding of Human Cytidine Deaminase APOBEC3B

Chapter II is a collaborative study that has been previously published as:

Hou S, Silvas TV, Leidner F, Nalivaika EA, Matsuo H, Kurt Yilmaz N, Schiffer CA. "Structural analysis of the active site and DNA binding of human cytidine deaminase APOBEC3B." *Journal of Chemical Theory and Computation* 15.1 (2018): 637-647.

2.1 ABSTRACT

APOBEC3s proteins (A3s), a family of human cytidine deaminases, protect the host from endogenous retro-elements and exogenous viral infections by introducing hypermutations. However, overexpressed A3s can modify genomic DNA to promote tumorigenesis, especially A3B. Despite overall similarity, A3 proteins have distinct deamination activity. Recently determined A3 structures have revealed the molecular determinants of nucleotide specificity and DNA binding. However, for A3B, the structural basis for regulation of deamination activity and the role of active site loops in coordinating DNA had remained unknown. Using advanced molecular modeling followed by experimental mutational analysis and dynamics simulations, we investigated molecular mechanism of DNA binding by A3B-CTD. We modeled fully native A3B-DNA structure, identified Arg211 in loop 1 as the gatekeeper coordinating DNA and critical residues for nucleotide specificity. We also identified a unique auto-inhibited conformation in A3B-CTD that restricts access and binding of DNA to the active site. Our results reveal the structural basis for DNA binding and relatively lower catalytic activity of A3B and provide opportunities for rational design of specific inhibitors to benefit cancer therapeutics.

2.2 INTRODUCTION

APOBEC3s (A3s) are a family of cytidine deaminases that catalyze a zinc-dependent cytidine to uridine reaction on single strand DNA (ssDNA) or single strand RNA (ssRNA) [1-4]. The family comprises of seven members that have either one (A3A, A3C, A3H) or two (A3B, A3D, A3F, A3G) zinc-binding domain [5]. The two-domain A3s

have a pseudo-catalytic N-terminal domain (NTD) and a catalytically active C-terminal domain (CTD). A3s play a key role in innate immunity by protecting the host cell from exogenous viral infections and endogenous retro-elements through introducing G to A hypermutations [22, 145-149]. However, when overexpressed, their mutagenic activity can also cause modification of genomic DNA and thus promote tumorigenesis [41, 47, 150]. A3B has been identified as a significant enzymatic source of mutagenesis in a variety of cancers [44]. Endogenous A3B is involved in the restriction of retro-element LINE-1 [151] and HBV [152, 153]. However, when overexpressed, A3B can mutate the host genome to trigger cancer phenotypes [150]. The up-regulation of A3B in tumors is correlated with both dispersed and clustered high occurrence of cytidine mutations, p53 (tumor protein 53) inactivation, and poor patient outcome in cancer treatment [45, 150, 154-156]. In addition, the genomic mutations preferentially occur at 5' -TCA, 5' -TCG, and 5' -TCT trinucleotide motifs, which resemble the substrate preference of A3B in biochemical assays [45, 154, 157]. Unlike other cancer sources, A3B can actively create genomic mutations, which means that a growing number of DNA mutations will be created in cancer cells. This will further benefit cancer evolution, for instance, to help escape immune monitoring, outgrow, metastasize, and potentially acquire resistance to therapeutic treatments [158]. Hence, in addition to its non-essential nature [159], A3B represents a promising target for novel anti-cancer drug development.

Over the past several years, crystal and NMR structures of human A3s (A3A, A3C; CTDs of A3B, A3F, A3G) in the apo state have been determined by our group [79-85] and others [86-99]. In general, A3 proteins are structurally highly similar despite their distinct deamination activities. Even though full-length A3B has 5-6 fold higher activity

compared to A3B-CTD, A3B-CTD alone can deaminate cytidine in ssDNA but A3B-NTD is catalytically inactive [98, 160, 161]. Among all human A3s, A3B-CTD and A3A share the highest sequence identity (**Figure 2.1A**); however, A3A is about 15-fold more active compared to A3B-CTD [98]. The overall A3 domain structure consists of six alpha-helices and five beta-strands with the zinc-binding region in the middle. In fact, based on the amino acid sequence or even the available structures, it is not apparent what molecular mechanisms are responsible for varied ssDNA binding affinity and deamination activity among A3 domains, including the catalytically inactive NTDs. Recently, our laboratory [104], along with two other groups, have determined the crystal structures of three A3–DNA complexes (A3A-DNA, chimeric A3B-DNA and rA3G-DNA) [99, 106]. When A3A binds to DNA, two major changes occur at the active site involving the side chain of Tyr132, which stacks against the DNA, and the gatekeeper His29, which locks the DNA in the active site. To facilitate crystallization of DNA-bound A3B, loop 1 of A3A was swapped into A3B to determine the structure. However, differences in loop 1 between the two A3s are mainly responsible for the difference in catalytic activity, as swapping loop 1 of A3B-CTD by that of A3A increases A3B-CTD activity by 10-fold [98]. Loop 1 also exhibits the largest amino acid sequence difference between A3A and A3B-CTD (**Figure 2.1A**) and is longer in A3B with a three-residue insertion $_{206}$PLV$_{208}$. In addition, the DNA gatekeeper residue His29 in A3A[104] is missing in A3B and is likely replaced by one of the arginines within the unique triple arginine patch $_{210}$RRR$_{212}$. However, the role of loop 1, and identifying whether Arg211 or Arg212 might be the gatekeeper residue, in coordinating DNA and in regulating A3B's catalytic activity could not be revealed by crystal structures determined to date.

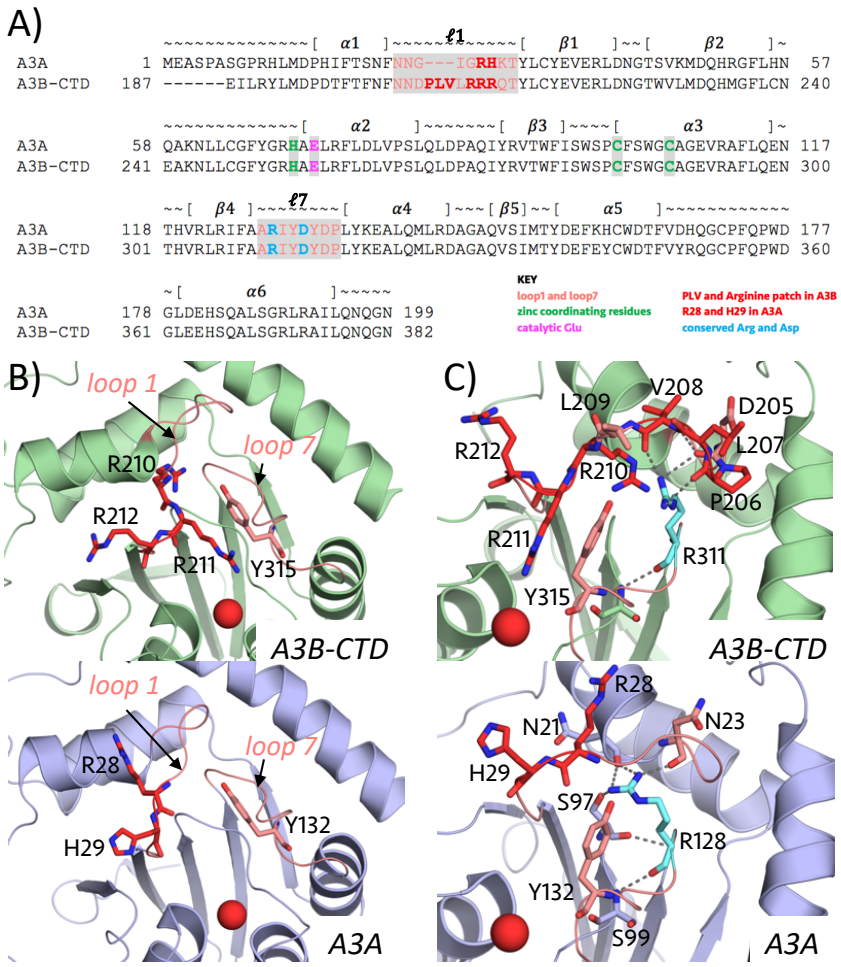

Figure 2.1: Protein sequence alignment and structure comparison between A3B-CTD and A3A.
A) Amino acid sequence alignment of A3A and the catalytic domain of A3B (A3B-CTD). B) A3B has a closed active site conformation while A3A has an open active site in crystal structures. C) Extra PLV residues alter the conformation of the conserved Arg311 in A3B through extensive hydrogen bond interactions. A3B-CTD (pdb: 5CQH) and A3A (pdb: 4XXO) are shown in cartoon representation (A3B in green; A3A in slate blue). The catalytic zinc is shown as red sphere. Loop 1 and loop 7 are colored salmon. $_{210}$RRR$_{212}$ and $_{206}$PLV$_{208}$ in A3B and 28RH29 in A3A are colored red. The conserved arginine is colored cyan. All the labelled residues are shown in stick representation.

To elucidate the molecular mechanism of A3B-DNA recognition and how A3B-CTD structurally regulates its catalytic activity compared to A3A, we used a combination of molecular modeling with molecular dynamics (MD) simulations and experimental mutational analysis followed by fluorescence-anisotropy based DNA binding assays. We identified the key role of loop 1 in A3B binding to DNA and down regulating its activity. We present a structural model of the A3B–DNA complex that elucidates the molecular mechanism and determinants of DNA binding to A3B-CTD. The model and mutational verification identified Arg211 as the gatekeeper for locking DNA into the active site, which is further stabilized by Arg212. We also identified an auto-inhibited conformation in A3B-CTD that is unique among human A3s, resulting from differences in loop 1 length and sequence, which explains the relatively low catalytic activity of A3B. Overall, our results shed light into the structural regulation of A3 activity and differences in loop 1 coordination around the bound DNA, which may potentially lead to discovering anti-cancer drugs to benefit cancer therapeutics.

2.3 RESULTS AND DISCUSSION

Despite overall similarities, the length and sequence differences of loop 1 between A3B-CTD and A3A (**Figure 2.1**) are responsible for alterations in the active site and likely the differences in DNA binding and deamination activity. To elucidate the mechanism of DNA binding by A3B-CTD, multiple MD simulations were performed of both apo and DNA-bound structures of fully wild-type (WT) A3B-CTD, and compared with A3A (**Table 2.1**). These mechanisms were further validated by a series of

experimental fluorescence anisotropy-based DNA binding assays of A3B-CTD variants (**Table 2.2**).

2.3.1 Molecular mechanism of A3B-DNA recognition

The role of loop 1 and molecular mechanism of DNA binding had remained unknown for A3B, as recently determined A3B-CTD DNA co-crystal structure is a chimera with the crucial loop 1 swapped from A3A [106]. Here careful molecular modeling was used using available crystal structures to answer for A3B: 1. How does DNA bind to A3B? 2. Which residue is the gatekeeper for latching DNA in the active site? 3. How does A3B define its substrate specificity for thymidine over cytidine at -1 position? To address these questions, WT A3B-CTD bound to substrate DNA containing a TCG trinucleotide motif was modelled based on the crystal structures of apo A3B-CTD and A3A–DNA complex (see Methods for details). The quality of the complex models was further examined through both computational analysis of 100 ns MD simulations (**Table 2.1**), and experimental DNA binding assays of inactive A3B-CTD variants (**Table 2.2**). All the MD simulations of DNA-bound structures converged and were stable over the 100 ns trajectory time.

Table 2.1: List of the molecular dynamics simulations that were performed in this study.
White shade represents the simulations of apo proteins while grey represents the simulations of the DNA bound proteins.

Construct	Simulation Time (ns)	Replicates
WT A3A	1000	1
WT A3B-CTD	1000	1
A3B-CTD ΔPLV	1000	1
A3B-CTD Y315F	1000	1
A3B-CTD P206G	1000	1
WT A3A – ssDNA TCG	100	3
WT A3B-CTD R211 – ssDNA TCG	100	3
WT A3B-CTD R212 – ssDNA TCG	100	3
WT A3B-CTD R211 – ssDNA CCG	100	1
A3B-CTD R212H – ssDNA TCG	100	1

Table 2.2: DNA binding affinity of A3B-CTD inactive (E255A) variants.
The binding affinities, represented as Kd, of linear DNA and hairpin DNA with TCG motif in the loop to A3B and variants, determined by fluorescence anisotropy-based assays. All A3B variants contain the E255A mutation to catalytically inactivate the enzyme and prevent substrate deamination. Kd >10 μM indicates weak binding but binding is detectable. NB indicates no detectable binding over the range of concentrations tested (up to 20 μM of A3B-CTD). Representative binding curves for binding (WT), weak binding (R210K) and no binding (P206G) are shown in Figure S2. Activity (fold) indicates catalytic activity relative to wild type A3B-CTD.

A3B-CTD	poly A_TCG	DNA hairpin	Activity
WT	5.4 ± 2.6	2.0 ± 0.5	1
Y315F	7.3 ± 4.6	1.8 ± 0.4	1[a]
R212H	*NB*	1.2 ± 0.3	2.5[b]
R212A	*NB*	1.4 ± 0.4	-
R211A	*NB*	*NB*	0.05[a]
R211H	*NB*	*NB*	-
R210A	*> 10*	*> 10*	-
R210K	*> 10*	*> 10*	-
P206G	*NB*	*NB*	-

Footnote:
[a] Shi K et al. (2015). J. Biol. Chem.
[b] Shi K et al. (2017). Sci. Rep. (fold change was estimated from gel band intensities using ImageJ on Figure 6B)

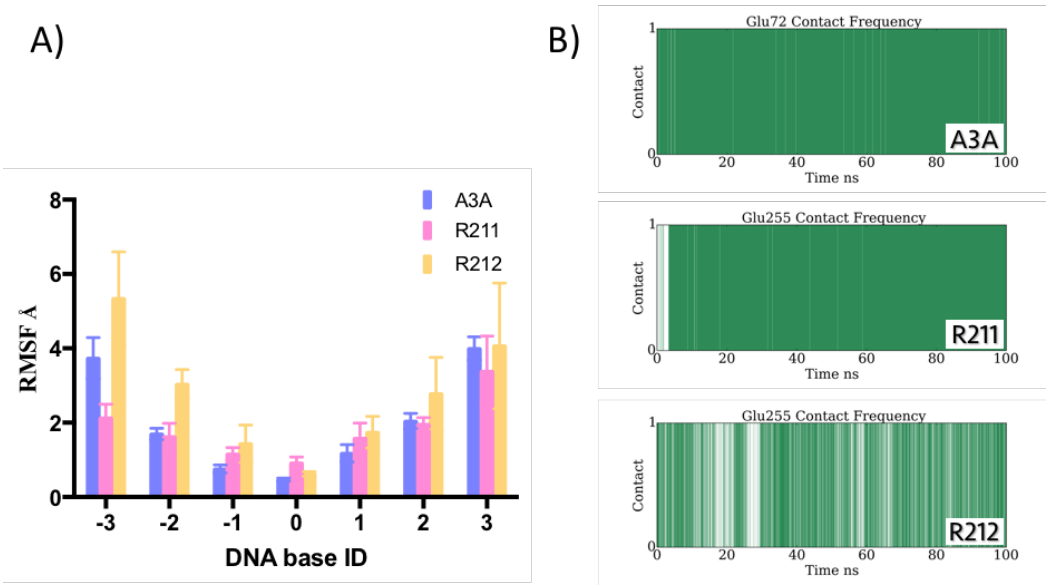

Figure 2.2: Comparison of A3B-DNA model structures with either R211 or R212 latching the DNA in the active site.
A) The root-mean-squared-fluctuations (RMSF) of individual bases of DNA molecule during MD simulations in A3A, A3B-CTD R211 and R212 models. B) The contact frequency of the intermolecular interactions between catalytic residue Glu255 and substrate cytidine base over the simulation time in A3A, A3B-CTD R211 and R212 models.

2.3.2 Arg211 is the gatekeeper residue sequestering DNA in the active site

As A3B-CTD has a unique $_{210}$RRR$_{212}$ patch in loop 1 instead of $_{28}$RH$_{29}$ compared to A3A (**Figure 2.1A**), either Arg211 or Arg212 could be the gatekeeper for DNA binding. To identify the critical residue, DNA-bound models were generated with either Arg211 or Arg212 latching DNA in the active site and performed and analyzed triplicate 100 ns MD simulations. The Arg212 model DNA complex was much less stable during the MD simulations compared to either A3A or Arg211 model as indicated by the considerably larger root-mean-square fluctuations (RMSFs) of bound DNA, especially at the two termini (**Figure 2.2A**). In both A3A structure and Arg211 model, Glu255 consistently interacted with substrate cytidine 98.94% and 96.29% of the simulation time, respectively (**Figure 2.2B**). However, in the Arg212 model, the contact frequency was decreased to only 60.21%, which suggests poor quality of the model. Thus, the stability over MD simulations indicated that Arg211 rather than Arg212 is the gatekeeper for DNA binding in A3B-CTD.

To further verify our results from computational analysis, we experimentally generated A3B-CTD R210A, R210K, R211A, R211H, R212A and R212H inactive variants and measured DNA binding using fluorescence anisotropy-based assay (**Table 2.2**). The low binding affinity and catalytic activity of WT A3B-CTD poses challenges in assessing changes in deamination rates and DNA binding. Hence, based on previous studies [84, 114], both TAMRA-labeled linear (in poly A background) and hairpin DNA with A3B preferred sequence TCG in the stem loop were tested in order to minimize background non-specific binding and promote binding affinity. Despite low affinity, the

A3B-CTD inactive variant could bind to both linear and hairpin DNA. Both R210A and R210K variants were able to bind DNA but with decreased binding affinity, which is in agreement with Arg210's role in stabilizing overall structure through the conserved hydrogen bond network in apo crystal structure (**Figure 2.3**). R212A variant bound to DNA with same affinity as wild type A3B, which confirmed that Arg212 is not the gatekeeper for DNA binding. In contrast, R211A mutant lost binding completely to both ssDNA and hairpin DNA. In agreement with DNA binding assay results, which we have shown to correlate with catalytic activity [114], recent A3B catalytic activity studies have also shown that R211A mutant lost deamination activity (**Table 2.2**). Thus, experimental DNA binding assay data were in agreement with our model and that Arg211 is the critical residue for DNA binding.

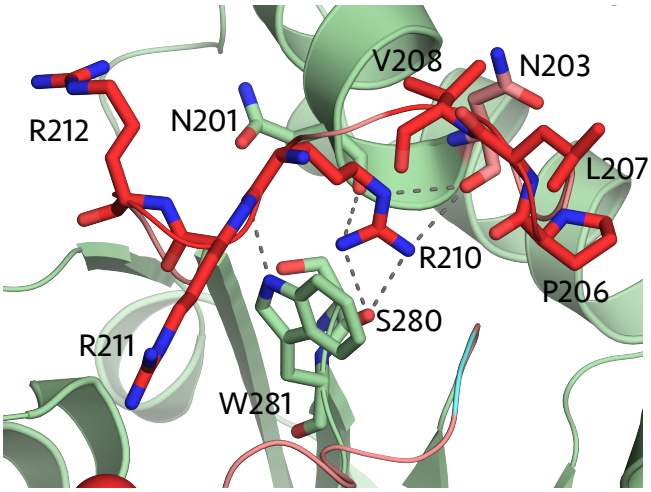

Figure 2.3: Hydrogen bond network of Arg210 in A3B-CTD apo crystal structure (PDB: 5CQH).
The catalytic zinc is shown as red sphere. Loop 1 and loop7 are colored salmon. $_{210}$RRR$_{212}$ and $_{206}$PLV$_{208}$ in A3B and 28RH29 in A3A are colored red. The conserved arginine is colored cyan. All the labelled residues are shown in stick representation.

In A3A, His29 is the gatekeeper for DNA binding through both hydrogen bond interactions to DNA backbone and stacking interactions to +1 base [104]. A3B-CTD R211H variant, however, showed no binding to DNA. Rather the two types of interactions His29 makes with bound DNA in A3A can be assigned separately to Arg211 and Arg212 in A3B, as Arg211 hydrogen bonds to DNA in the active site while Arg212 stabilizes DNA binding through either stacking or hydrogen bond interactions with the +1 DNA base. In agreement with this mode of binding, in contrast to R211H, the R212H variant was able to bind DNA with comparable affinity as WT A3B-CTD (**Table 2.2**). During the MD simulation for R212H in complex with DNA, residue R212H formed pi stacking and hydrogen bond interactions with +1 G base 49% and 47% of the time, respectively, which is actually slightly higher than Arg212 (43% and 44%, respectively). The cyclic histidine side chain may facilitate better stacking interactions, leading to the slightly improved binding and catalytic activity (**Table 2.2**). Therefore, our computational model of fully native DNA-bound A3B-CTD structure was verified by both MD simulations and experimental mutational analysis.

2.3.3 ssDNA binding to A3B-CTD

The DNA-bound model of A3B-CTD revealed the molecular mechanism as well as the role of loop 1 for DNA binding to A3B-CTD. Overall, ssDNA bound to A3B-CTD in a U-shape similar to A3A (**Figure 2.4A**). Compared to the apo crystal structure, loop 1 underwent major conformational changes to open up the active site for DNA binding (**Figure 2.4B**), especially at Arg211 which stacks against Tyr315 to close the active site in apo structure (**Figure 2.1B**). In addition, the side chain of Tyr315 rotated from a dihedral angle $\chi 1$ of ~180° to ~60° as in A3A to accommodate DNA binding (**Figure**

2.4B). In general, our model overlaid well with chimeric A3B-CTD DNA co-crystal structure but provided critical information on loop 1 conformation in DNA-bound form (**Figure 2.4C**), and conformational changes needed to open up the active site to allow DNA binding. The critical DNA binding residues were also examined in terms of intermolecular van der Waals (vdW) interactions (**Figure 2.4D, E**) and hydrogen bonds (**Figure 2.5**). The total vdW contacts between A3B-CTD and ssDNA is about -84.3 kcal/mol, which is predominantly contributed from the loops around the active site, loop 1 (-37.8 kcal/mol), loop 3 (-15.9 kcal/mol) and loop 7 (-15.9 kcal/mol). Significantly, loop 1, which is missing in the chimeric A3B-CTD DNA crystal structure, contributed 45% of the total vdW contacts. The most critical intermolecular interaction between A3B-CTD and ssDNA involved the gatekeeper Arg211. Arg211 coordinated DNA binding through both hydrogen bond interactions with the phosphate backbone of -1 T, 0 C and +1 G bases and hydrophobic interactions with DNA backbone (**Figure 2.5A, C**). Arg212 instead stabilized DNA binding through either stacking (**Figure 2.5A, B**) or hydrogen bond interactions with the +1 G base.

A3B should be able to bind pre-bent or hairpin DNA based on our model as observed for A3A [114] and likely with higher affinity as the entropic cost of bending the DNA would be decreased. Accordingly, we observed higher binding affinity to hairpin DNA (K_d = 2.0 μM) compared to linear DNA (K_d = 5.4 μM) for A3B (**Table 2.2**). Interestingly, unlike A3A [114, 162], A3B-CTD showed no binding to RNA hairpin (data not shown). These structural and functional differences between A3A and A3B-CTD despite their high sequence similarity might have implications for their biological function as well as cellular localization.

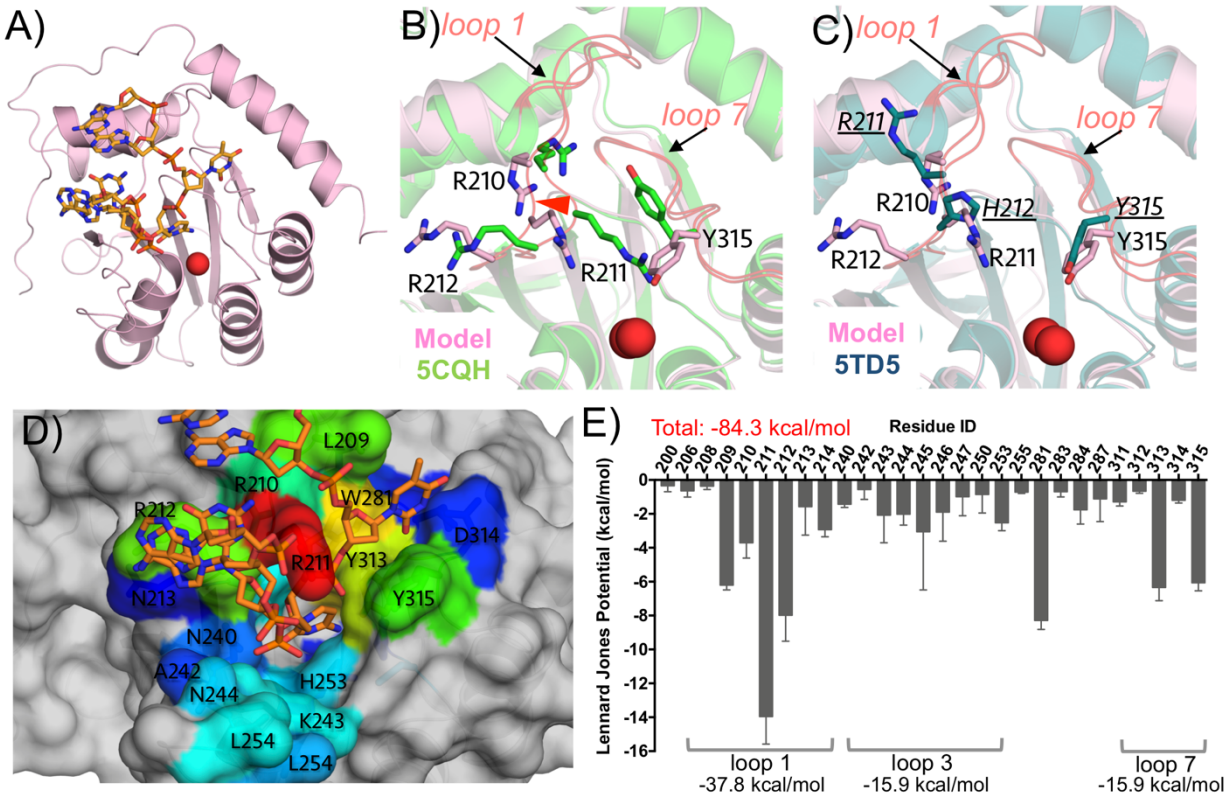

Figure 2.4: Structural model of A3B-CTD in complex with ssDNA.
A) Overall structure of DNA bound A3B-CTD model. B) Conformational changes of residues R210, R211, R212 and Y315 upon DNA binding, with side chains displayed in stick representation. Loop 1 and loop 7 are colored salmon and the conformational changes upon DNA binding are indicated by arrows. C) Structural comparison between the modeled fully native A3B-CTD and the chimeric A3B-CTD DNA crystal structure (PDB: 5TD5). D,E) Mean vdW contacts between protein and ssDNA calculated from triplicate MD simulations. The residues are colored on a rainbow scale from blue to red for increasing contacts; hence warmer colors indicate residues with the most contribution to the intermolecular contacts. The cut-off for the scale is -0.5 kcal/mol.

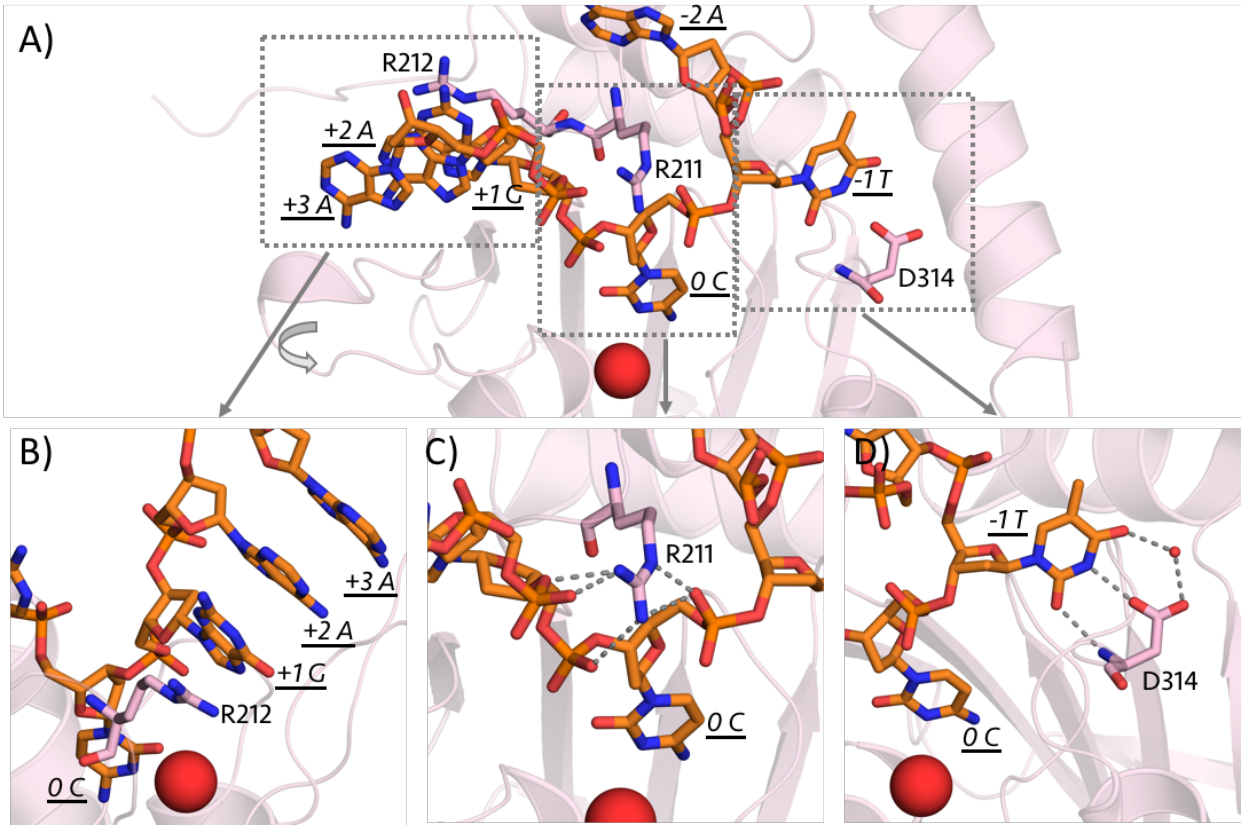

Figure 2.5: Intermolecular interactions between A3B-CTD and ssDNA.
A) Overview of key residues in binding to DNA. B) R212 can stack with downstream DNA bases. C) R211 forms extensive hydrogen bond interactions with DNA backbone. D) D314 makes extensive hydrogen bonds with -1 base that defines substrate specificity. The final frame of representative MD simulation is displayed. The protein is displayed as a pink-colored ribbon diagram and the bound DNA is in orange stick representation. The zinc ion at the active site is depicted as a red-colored sphere. The side chains of R210, R211, R212, D314 and Y315 are shown as sticks. The hydrogen bond interactions are indicated with grey dashed lines.

2.3.4 D314 defines substrate specificity for thymidine over cytidine at -1 position

In A3A, the substrate specificity for thymidine over cytidine at -1 position is determined through hydrogen bond interactions with Asp132 [104]. In our A3B–DNA model, the same hydrogen-bonding pattern between -1 T base and Asp314 as in A3A (**Figure 2.5A,D; Figure 2.6A**) was observed. Specifically, O2 atom of -1 T formed a direct hydrogen bond with Asp314 backbone, while OD1 and OD2 atoms of Asp314 had both direct and water-mediated hydrogen bonding with N3 and O4 of -1 T base. All these hydrogen bond interactions were stable during MD simulations (**Figure 2.6B**). In contrast, when thymidine was changed to a cytidine, the side chain hydrogen bond interactions between the -1 C and Asp314 were disrupted and the DNA was destabilized as indicated by increased dynamics in the active site (**Figure 2.6C,D**). Similarly, in the Arg212 model, which we deduced to be poor based on dynamics above, the hydrogen bonds of Asp314 with -1 T were destabilized and not reproducible among the replicate simulations (**Figure 2.7**). These findings suggest that A3B likely uses the same molecular mechanism to determine the substrate specificity as A3A at -1 position since Asp314 is conserved and the DNA interactions maintained between the two A3s.

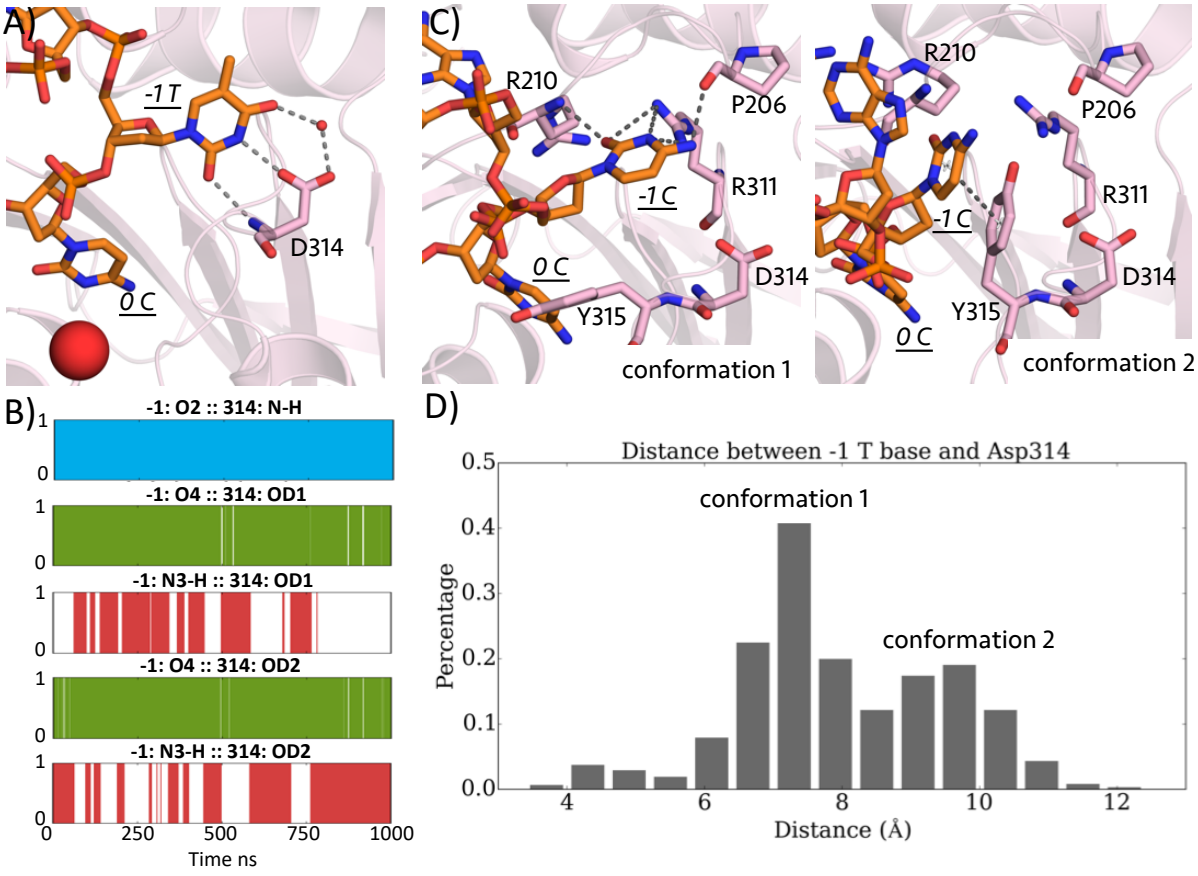

Figure 2.6: Comparison of TC versus CC binding by A3B-CTD.
A) D314 makes extensive hydrogen bond interactions with -1 T, stabilizing this base and contributing to TC specificity. B) The hydrogen bond interactions between D314 and -1 T are stable throughout the whole MD simulations. C) In contrast to T, -1 C has alternative conformations during the MD trajectory. D) The histogram of the distance between atom N3 of -1 C and CG atom of Asp314. The two peak conformations are shown in panel C.

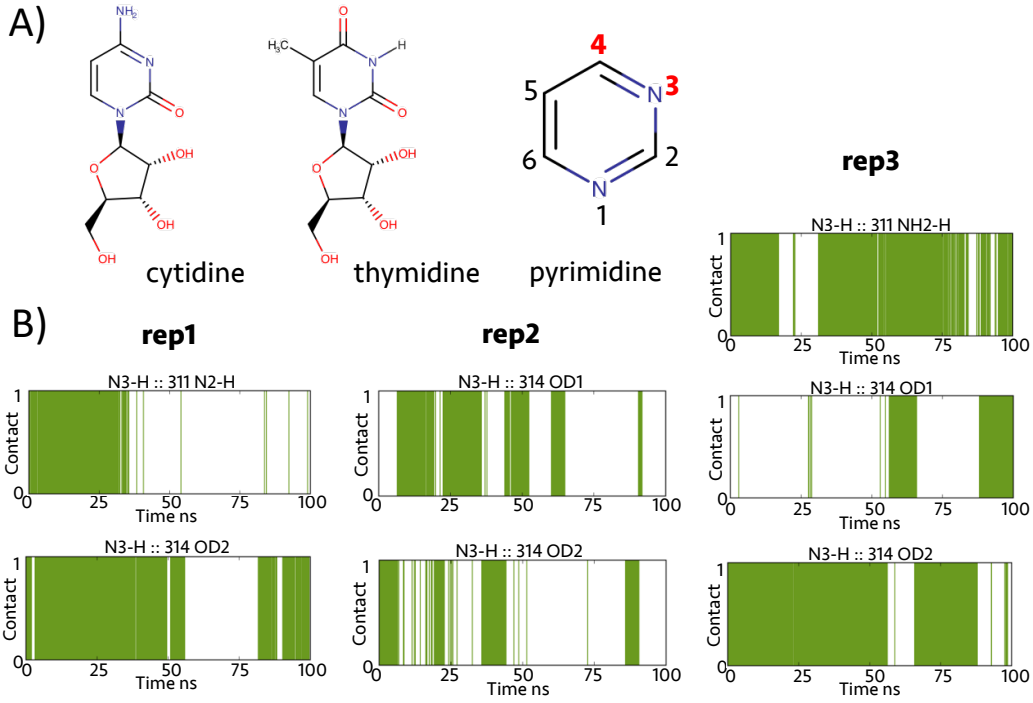

Figure 2.7: The hydrogen bond interactions between A3B-CTD protein and atom N3 and O4 of -1 thymidine in R212 model.
A) The schematic representation of cytidine, thymidine and pyrimidine. The main differences between cytidine and thymidine are at position 3 and 4. B) The hydrogen bond interactions of atom N3 and O4 of -1 thymidine in R212 model during the MDs. No direct hydrogen bonds found. Green represents water-mediated hydrogen bond interactions.

2.3.5 Structural mechanism of auto-inhibited conformation of apo A3B-CTD

The crystal structure of apo A3B-CTD has a closed active site conformation that results from the stacking interactions between Arg211 or Arg212 of loop 1 and Tyr315 of loop 7, in contrast to A3A [96, 100] (**Figure 2.1B**). Therefore, the active site of A3B-CTD has to open up from the closed conformation to accommodate DNA binding (**Figure 2.4B**). To elucidate the structural basis of the closed active site conformation in A3B-CTD, which can modulate DNA binding and thus deamination, we performed detailed structural analysis on the apo forms of A3B-CTD and A3A.

2.3.6 Closed active site conformation correlates with lower DNA affinity

A structural inspection of A3B-CTD in comparison to the closely related A3A and other A3 domains revealed that $_{206}$PLV$_{208}$ insertion in loop 1 forms a unique hydrogen bond network with Arg311 and Asp205 (**Figure 2.1C**). Specifically, the backbone carbonyl oxygen of Pro206 makes a hydrogen bond with NH2 atom of Arg311; Val208 forms two backbone hydrogen bonds with the backbone of Asp205 and NH1 atom of Arg311. The backbone carbonyl oxygen of Asp205 also has a hydrogen bond with NH1 atom of Arg311. Arg311 is conserved among all A3 domains (**Figure 2.8A**). However, when we superimposed all the available active apo A3 structures (A3A, A3B-CTD, A3C, A3F-CTD and A3G-CTD), the side chain of this conserved Arg was locked in a hydrogen bond network that was distinct from the conformation of Arg311 observed in A3B. (**Figure 2.1C; Figure 2.8B, C**). This network involves primarily the backbone atoms of conserved Ser97, Ser99, Asn21 (Gln in A3C and His in A3F-CTD and A3G-CTD) and Asn23 (Lys in A3C, A3F-CTD and A3G-CTD). In A3A, the side chain of Ser97 forms an additional hydrogen bond with this conserved Arg. Rather than with Arg311 in

A3B-CTD, these residues form an analogous hydrogen bond network with Arg210 in loop 1 (**Figure 2.3**). The distinct conformation of Arg311 in A3B-CTD and hydrogen bonding with the $_{206}$PLV$_{208}$ in loop 1 may contribute to the closed active site conformation of A3B-CTD as well as A3B's lower activity.

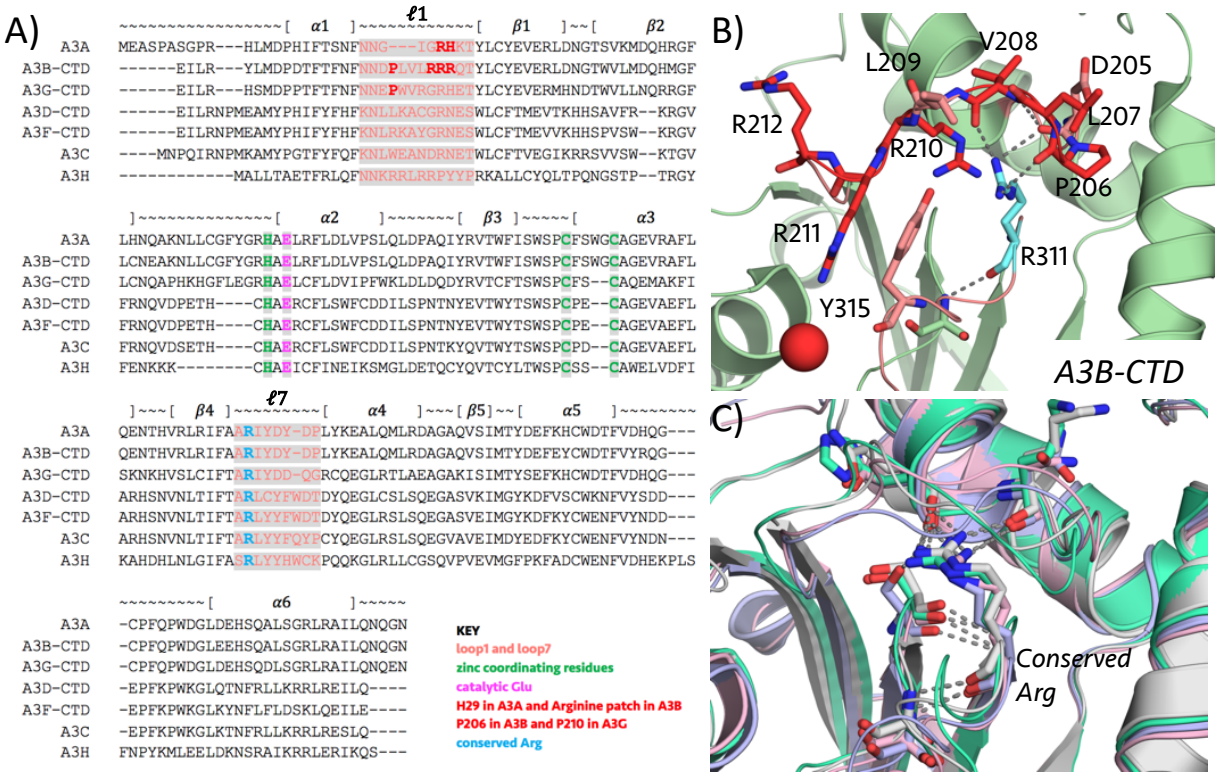

Figure 2.8: Amino acid sequence and structural differences between A3B-CTD and catalytically active A3 domains.
A) Protein sequence alignment of all catalytic active A3 domains. B) The conformation of the conserved Arg (Arg311) in A3B-CTD. C) The conserved arginine is locked in the same hydrogen bond network in crystal structures of catalytically active A3 domains except A3B-CTD. A3A (PDB: 4XXO) is in slate blue; A3C (PDB: 3VOW) is in white; A3F-CTD (PDB: 4IOU) is in cyan green; A3G-CTD (PDB: 3IR2) is in pink.

To investigate $_{206}$PLV$_{208}$'s role in regulating activity, 1 μs MD simulations were performed on WT A3A and A3B-CTD as well a variant where the $_{206}$PLV$_{208}$ sequence was deleted (A3B-CTD-ΔPLV). From the MD simulation trajectories, the stability of the closed active site conformation (**Figure 2.9**) was monitored. The closed active site conformation in WT A3B-CTD was stable during the MD simulations, as the distance between the side chain of Arg211 in loop 1 and Tyr315 in loop 7 varied around 6 Å, which is within the range of stacking interactions that close the active site. In contrast, the equivalent residues in wild type A3A, Arg28 and Tyr132 had a distance distribution around 15 Å, which indicates the active site is in the open conformation (**Figure 2.9A,C**). Interestingly, in A3B-CTD-ΔPLV, Arg211 lost the stacking interactions with Tyr315. The distance between the side chains of Arg211 and Tyr315 was more than 12 Å during the MD simulations. As a result, the active site conformation was altered into the open conformation, analogous to that observed in A3A. The more open active site correlates with higher activity in A3A and A3B-CTD-ΔPLV compared to WT A3B.

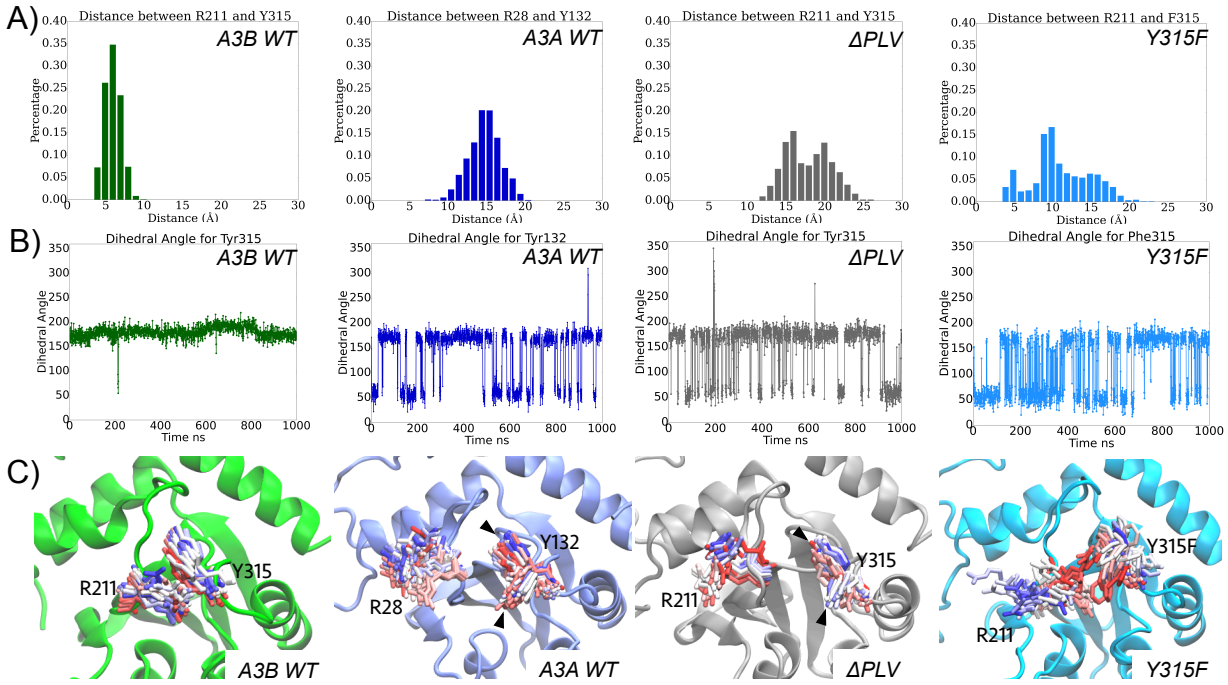

Figure 2.9: Dynamics of the active site in A3A, A3B-CTD and A3B-CTD mutants. A) The histogram of the distance between CZ atom of Arg311 (Arg28 in A3A) and center of benzene ring of Tyr315 (Tyr132 in A3A) in wild type A3B-CTD, A3A, A3B-CTD-ΔPLV and A3B-CTD Y315F variants during 1 μs MD simulations. B) The dihedral angle of the side chain (C, CA, CB and CG) of Tyr315 (Tyr132 in A3A) over 1 μs MD simulations in wild type A3B, A3A, A3B ΔPLV and Y315F mutants. C) The time series of the side chain conformations of Arg211 (Arg28 in A3A) and Tyr315 (Tyr132 in A3A) (shown as stick; colored based on the simulation time, start as red and end with blue) during 1 μs MD trajectory of wild type A3B-CTD (green), A3A (slate blue), A3B-CTD-ΔPLV (grey) and A3B-CTD Y315F (cyan) variants. Different conformations of Tyr315 are indicated with arrows.

The side chain conformation of Tyr315, which is analogous to Tyr132 in A3A, was also monitored during the MD simulations as an indicator for the compatibility to bind DNA (**Figure 2.9B, C**). The Tyr side chain has to undergo a conformer change to accommodate binding of DNA at the active site [104, 106] (**Figure 2.4B**). In WT A3B, the side chain dihedral angle $\chi 1$ of Tyr315 remained around 180° in 99.8% of the time during the MD simulations, in agreement with our finding that the closed active site conformation in A3B is not compatible for DNA binding. The same dihedral angle $\chi 1$ of Tyr132 in A3A changed from about 180° to 60° upon DNA binding in the crystal structure [104]. In the MD simulation, the side chain of Tyr132 in A3A sampled between the apo and DNA-compatible conformations (68% of the time in DNA-compatible conformation). In A3B-CTD-ΔPLV, the side chain of Tyr315 sampled two conformations (73% of time in DNA-compatible conformation) as in A3A. Thus, the high sampling frequency of DNA-compatible side chain conformation of Tyr315 in A3B-CTD-ΔPLV and A3A correlates with the higher DNA affinity and activity compared to WT A3B.

We also observed auto-inhibited conformation of WT A3B-CTD in the 1 μs MD simulation, which involved the $_{206}$PLV$_{208}$ hydrogen bond network with Tyr315 (**Figure 2.10A**). Specifically, the OH atom of Tyr315 interacted with both NH2 atom of Arg311 and backbone carbonyl oxygen of Pro206 through direct and water-mediated hydrogen bonds. As a result, the side chain of Tyr315 was locked in the DNA incompatible conformation (**Figure 2.9B, C**). These hydrogen bonds were stable throughout the MD simulations (**Figure 2.10B**). To verify the role of this auto-inhibited conformation in down-regulating A3B's activity, the Y315F variant was modeled and subjected to the same 1 μs MD simulations. Phe315 in the Y315F variant lost the ability to interact with

$_{206}$PLV$_{208}$ hydrogen bond network, as the hydroxyl group was lost, and was released from the auto-inhibited conformation. As a result, the side chain dihedral angle χ1 of Phe315 sampled the DNA-compatible conformation with 46.4% frequency in MD simulation (**Figure 2.9B**). The active site of Y315F variant somewhat opened up as the stacking interactions between Arg211 and Phe315 was disrupted during the MD simulation (**Figure 2.9A, C**). However, the extent of active site opening was less compared to A3B-CTD-ΔPLV and A3A. The distance between Arg211 and Phe315 (~10 Å) was smaller than that in A3B-CTD-ΔPLV (~18 Å) and A3A (~15 Å) (**Figure 2.9A, C**). Overall, this result suggests that the hydrogen-bonding network involving Tyr315 helps stabilize the closed active site conformation. In agreement with the MD results, the Y315F variant slightly gained DNA affinity in experimental binding assay relative to WT protein, especially for DNA hairpin (**Table 2.2**). Hence, disrupting the hydrogen-bonding network between residue 315 and $_{206}$PLV$_{208}$ destabilizes the closed active site conformation but is not enough to shift to a fully open active site, as the longer loop 1 with the PLV insertion is critical for the closed active site conformation and thus down-regulating A3B-CTD's activity. Recent studies have shown that removing PLV from loop 1 in A3B increases the enzyme's activity, similarly to the chimeric A3B-CTD with the whole loop 1 swapped from A3A (estimated from gel band intensities from reference [100]); this is in agreement with the more open active site conformation that we observed for A3B-CTD-ΔPLV. Thus, the closed active site conformation observed in modeling and simulations were in complete agreement with experimental binding and catalytic activity. Together these findings strongly suggest that the PLV insertion in loop 1 is the key for restricting and regulating A3B's deamination activity.

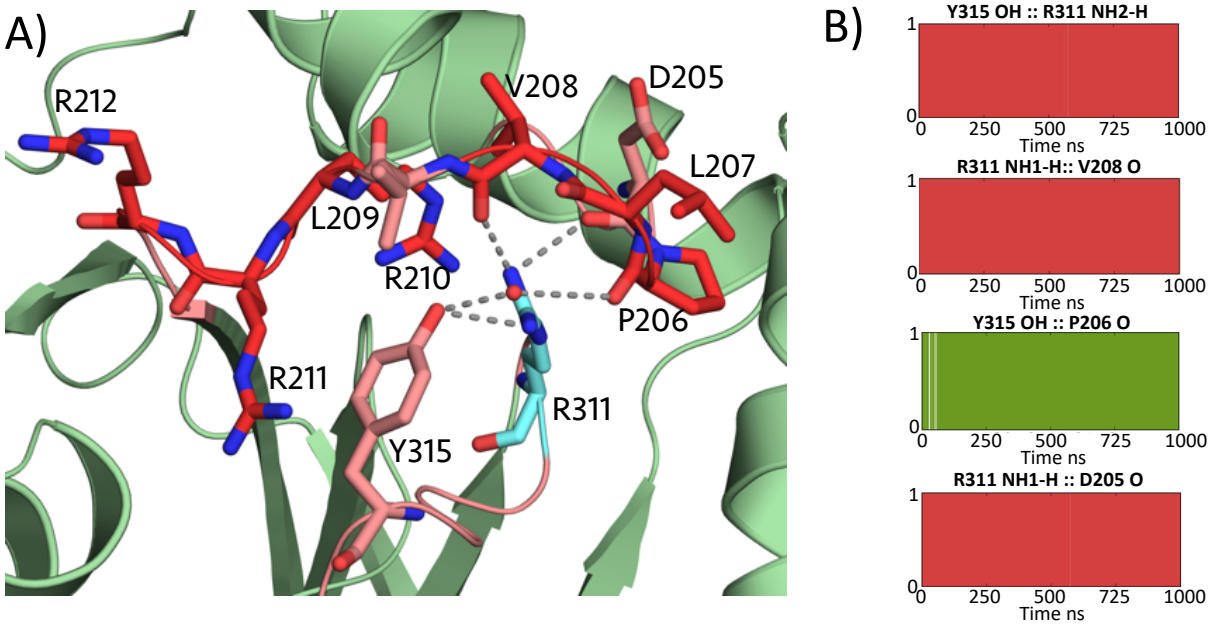

Figure 2.10: PLV hydrogen bond network locks Tyr315 in DNA-binding incompatible conformation.
A) A3B-CTD has an auto-inhibited mode that results from the hydrogen bond interactions of Tyr315 with PLV residues in loop 1. Loop 1 and loop7 are colored salmon. The catalytic zinc, PLV, RRR patch in A3B are shown in red. Conserved arginine is colored in cyan. Hydrogen bond interactions are shown as gray dashed lines. B) The hydrogen bond interactions during 1 μs MD simulation of apo A3B-CTD. Direct hydrogen bond interactions are colored red. Water-mediated hydrogen bond interactions are shown in green.

2.3.7 Proline stabilizes the conformation of the longer loop 1 in A3B-CTD for DNA binding

Among all A3s, A3A, A3B-CTD and A3G-CTD share the highest sequence similarity and belong to the Z1 group. Both A3B-CTD and A3G-CTD have a longer loop 1 that includes a proline residue, compared to A3A. Considering proline's unique geometry and rigidity compared to other amino acids, this extra residue may help stabilize the conformation of a longer loop 1. To test this hypothesis, we modeled A3B-CTD P206G variant and examined the RMSF of loop 1 as a measure of dynamics (**Figure 2.11**). Despite the differences in activity, WT A3B-CTD and A3A have similar levels of loop 1 dynamics with RMSF values varying around 4 Å, as well as WT A3G-CTD. The RMSF of loop 1 in A3B-CTD Y315F mutant and A3B-CTD-ΔPLV were also within 4 Å during the MD simulations. Loop 1 in A3B-CTD P206G variant, however, is highly dynamic with RMSF varying from 1 up to 8 Å indicating that loop 1 bearing the P206G mutation may not be able to stably coordinate DNA. Experimentally, the P206G variant lost the ability to bind both linear and hairpin DNA (**Table 2.2**). These results indicate the importance of proline in a longer loop1 and consistency of loop 1 dynamics in DNA binding.

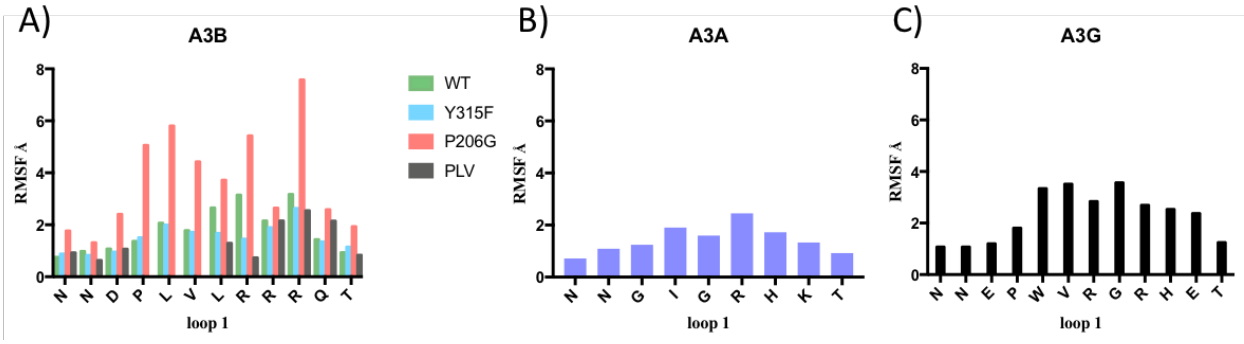

Figure 2.11: The dynamics of loop 1 during 1 μs MD simulations.
A) The root-mean-squared-fluctuations (RMSF) of individual residues of loop 1 in wild type A3B-CTD and A3B-CTD variants. B) The RMSF of all residues in loop 1 of wild type A3A. C). The RMSF of all residues in loop 1 of wild type A3G-CTD.

2.3.8 Conclusions and implications for DNA binding to other A3s

There are still several A3 domains whose structures are unknown, and no full-length two-domain A3 structure has been determined. The low solubility and DNA affinity of certain A3 proteins have required introducing mutations to be able to structurally and biochemically characterize these proteins in vitro [79-81, 83-85, 92, 96, 99, 104, 106], or altogether prevented such characterization especially for NTDs. For instance, WT A3B-CTD has poor solubility and low binding affinity towards DNA, which makes crystalizing native A3B-DNA complex extremely challenging. The available DNA-bound structure was that of an A3B chimera engineered to increase affinity and promote crystallization. While this structure did not inform on the role of loop 1 in DNA binding, the apo structure of A3B with the native loop1 and A3A–DNA structures enabled computational modeling of WT A3B bound to substrate DNA. Future studies focusing on the active site loops may elucidate differences between A3 family members, including the catalytically active and pseudo-catalytic A3 domains. Similarly, combining experimental structures with computational modeling, verified by simulations and experimental mutational analysis as in this study, can provide insights into the function and DNA binding of other A3s.

A3 proteins have the same overall fold, with highly conserved active sites and yet neither the available structures nor the amino acid sequences offer obvious insights into why they have highly varying catalytic activity, from totally inactive pseudo-catalytic NTDs to the highly active A3A. Instead the seemingly minor diversity in the loops 1, 3 and 7 around the active site may be responsible for regulating A3 activity, which could have implications in regulating the biological function in innate immunity and cancer

development. Our results suggest that the length and sequence differences in loop 1, which were missing in the DNA-bound A3B crystal structure [106], are key in regulating activity of Z1 A3 domains (**Figure 2.12**). A short loop 1, as in A3A or A3B-CTD-ΔPLV, results in high catalytic activity. A longer loop 1 as in A3G-CTD, which includes a proline to stabilize the overall conformation, can form molecular interactions with loop 7 to close the active site, and results in medium activity. Finally, having the auto-inhibited conformation due to the $_{206}$PLV$_{208}$ hydrogen bond network with Arg311 in addition to a longer loop 1 in A3B-CTD further restricts deamination activity. Thus, the detailed analysis of A3B-CTD structure here revealed insights into how amino acid differences in loops around the active site can structurally regulate the relative catalytic activity of A3s despite highly similar overall structure and conserved active site.

To date, design and development of inhibitors or activators for A3s has proven to be extremely challenging. Our results provide opportunities for drug design to specifically target A3B and thus benefit cancer therapeutics. Small molecules that stabilize the unique auto-inhibited mode of A3B might be able to allosterically inhibit A3B without cross-reacting with other A3s. Besides, the residue-specific information on regulation of auto-inhibition and closed active site conformation provides the starting point for engineering A3 domains to achieve varying catalytic efficiencies or distinct substrate specificity.

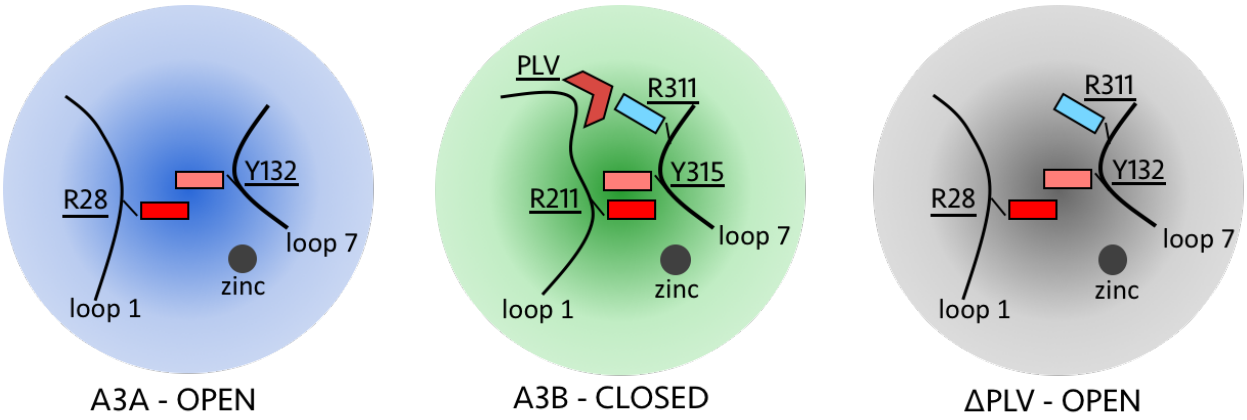

Figure 2.12: A schematic representation of the mechanism by which A3B-CTD regulates activity.
The structural features at the active site of A3A, A3B-CTD and PLV deletion variant that regulate catalytic activity. Loop 1 and loop 7 are shown as lines, and the catalytic zinc is represented as a grey sphere. The side chain of R28, Y132 in A3A and R211, Y315, R311 in A3B-CTD are shown as rectangles. PLV in A3B-CTD is represented as a wedge.

2.4 MATERIALS AND METHODS

2.4.1 Molecular modeling

A3B-CTD wild type apo structure was modeled based on human A3B-CTD isoform A sequence and A3B-CTD crystal structure (PDB: 5CQH) through program Modeller 9.15 using basic modeling. All DNA-bound structures were modeled based on both the apo crystal structure (PDB: 5CQH for A3B; PDB: 4XXO for A3A) and A3A–DNA co-crystal structure (PDB: 5KEG) through program Modeller 9.15 using advanced modeling. 5'-TCG motif was used in both modeling and fluorescence anisotropy-based binding assay as A3B-CTD shows highest deamination activity against this sequence.[163] The DNA molecule of the bound models and A3A–DNA co-crystal structures were modified through program Coot to the oligo sequence (AATCGAA) that was used in the fluorescence anisotropy-based binding assay. The phosphate groups of 5' A base were removed to prevent strong electronegative environment. AACCGAA was modeled similarly to test the molecular mechanism of substrate preference at -1 position. All DNA-bound structure models were then energy minimized through Protein Preparation Wizard from Schrodinger using default settings.

2.4.2 Molecular dynamics simulations

All molecular dynamics simulations were performed using Desmond[164] from Schrodinger. The models were first optimized using Protein Preparation Wizard. The simulation systems were then built through Desmond System Setup using OPLS3 force field [131]. We used SPC solvation model and cubic boundary conditions with 12 Å buffer box size. The final system was neutral and had 0.15 M sodium chloride. A multi-stage MD simulation protocol was used, which was previously described[165]. Briefly, the

system was initially relaxed for 100 ps/stage using Brownian Dynamics NVT (10 K) with gradually reduced restraints (500, 250, 50 force constant) on backbone heavy atoms to solute heavy atoms. This step was followed by simulations using NPT ensemble with gradually increased simulation time (24, 50 and 500 ps) and decreased restraints on the solute heavy atoms to no restraints. The final production stage was performed at 300 K and 1 bar with no restraints using NPT ensemble. 1 μs MD simulations were performed for all the apo structures to ensure final system convergence. The DNA-bound models (A3A/A3B-CTD R211/A3B-CTD R212) were simulated as triplicates to ensure reproducibility for 100 ns each. One round of 100 ns MD simulation was performed for other A3-DNA structures (A3B-CTD with CCG motif and A3B-CTD R212H with TCG motif) to compare with the final A3B-CTD DNA-bound model (**Table 2.1**).

2.4.3 Analysis of molecular dynamics simulations

The analysis of MD simulations was performed separately for each trajectory as well as the MD simulation triplicates, which help ensure reproducibility and conservation of the results among separate trajectories. The RMSD and RMSF of protein and DNA molecule as well as the protein-ligand contacts diagram were calculated using Simulation Interactions Diagram from Schrodinger. Hydrogen bonds occupancies over the trajectories were calculated using in-house modified Schrodinger trajectory analysis python scripts. Hydrogen bonds were determined for pairs of eligible donor/acceptor atoms using criteria set by Schrodinger: For a pair of heavy atoms to form a hydrogen bond, the distance between donor-hydrogen and acceptor had to be less than 2.8 Angstrom, the angle between donor, hydrogen and acceptor had to be at most 120 degrees and the angle between hydrogen, acceptor and the next atom had to be at

least 90 degrees. The residue vdW potential between A3B and DNA during the MD simulations was extracted from the simulation energies using Desmond. For both hydrogen bonds and vdW potential, errors were calculated using block averaging [166]. The distance histograms display the distance between the CZ atom of Arg211 (28 in A3A) and the benzene ring center of the side chain of Tyr315 (132 in A3A). CG, CA, CB and C of Tyr315 or Phe315 were used to determine the side chain dihedral angle. The time series representation of side chain conformations of Arg211 (28 in A3A) and Tyr315 (132 in A3A) in Figure 4C were generated with program VMD using 50 frames as time step (total 2000 frames).

2.4.4 Cloning and mutagenesis of inactive A3B constructs

Human A3B E255A gene was codon-optimized and synthesized by GenScript. This gene was then cloned into pGEX-6p-1 vector using BamHI and EcoRI restriction sites. The pGEX-6p-1 A3B E255A catalytically inactive overexpression construct was used for all experiments in this study. All the mutations were introduced using the Q5 site-directed mutagenesis kit (NEB), and the plasmids were sequenced to verify the mutation by Genewiz.

2.4.5 Protein expressions and purification

The pGEX-6p-1 A3B inactive mutant constructs were transformed into BL21 DE3 STAR E. coli strain for overexpression. Expression of GST-tagged A3B-CTD recombinant protein was performed at 17 °C for 22 hours in LB medium containing 0.5 mM IPTG and 100 µg/mL ampicillin. Cells were then pelleted, re-suspended in purification buffer (50 mM Tris HCl pH 7.4; 250 mM NaCl; 0.01% Tween 20 and 1 mM DTT) and lysed with the cell disruptor. The lysate was collected and the recombinant

protein was separated using a GST column. The GST tag was cleaved by the PreSecission Protease on the column at room temperature overnight. The flowthrough was then collected and further purified through size-exclusion chromatography using a HiLoad 16/60 Superdex 75 column (GE Healthcare).

2.4.6 Fluorescence anisotropy-based DNA binding assay

Fluorescence anisotropy-based DNA binding assays were performed as described (28) with minor modifications. We used 5'-TAMRA labeled oligonucleotides as the binding substrate. The linear oligonuclotide sequences used were 5'-AAA-AAA-AAA-AAA-AAA-3' (polyA) and 5'-AAA-AAA-AAT-CGA-AAA-3' (polyA TCG). The hairpin sequences used were 5'-GCC-ATC-ATT-CGA-TGG-G-3' (DNA hairpin) and 5'-rGrCrC-rArUrC-rUrArU-rCrGrA-rUrGrG-3' (RNA hairpin). The reaction buffer was 50 mM Tris buffer (pH 7.4), 100 mM NaCl, 0.5 mM TCEP. The concentration of APOBEC3 was varied from 0 to 20 μM in triplicate wells containing constant amount (10 nM) of substrate. Plates were incubated for an hour on ice before reading the plates. For all experiments, fluorescence anisotropy was measured using an EnVision plate reader (PerkinElmer), with excitation at 531 nm and detecting polarized emission at 579 nm wavelength.

Data analysis was performed using Prism 7 with least-square fitting of the measured fluorescence anisotropy values (Y) at different protein concentrations (X) with a single-site binding curve with Hill slope and constant background using the equation $Y = (B_{max} \times X^h)/ (K_d^h + X^h) + \text{Background}$, where K_d is the equilibrium dissociation constant, h is the Hill coefficient, and B_{max} is the extrapolated maximum anisotropy at

complete binding. The standard deviation was calculated for each measurement point from three independent repeats.

2.5 ACKNOWLEDGMENT

The authors would like to thank Drs. Brian Kelch, Mohan Somasundaran and Paul Thompson for their suggestions and helpful critiques.

3 Chapter III: Structural mechanism of substrate specificity in Z1 A3 domains

Chapter III is a collaborative study that is in preparation:

Hou S, Lee JM, Kurt Yilmaz N, Schiffer CA. “Structural mechanism of substrate specificity in human cytidine deaminase family APOBEC3s Z1 domains.” In preparation for submission to *The Journal of Biological Chemistry*.

3.1 INTRODUCTION

APOBEC3s (A3s) are a family of cytidine deaminases that have seven members in human [1-5] with functions in innate immunity and roles in cancer. All A3 domains share a conserved structural fold with an active site zinc tetrahedrally coordinated with catalytic His and Cys residues and an additional water. The human A3s have either one (A3A, A3C and A3H) or two zinc-binding domains (A3B, A3D, A3F and A3G). The two-domain A3s consist of a catalytically active C-terminal domain (CTD) and a pseudo-catalytic N-terminal domain (NTD) which binds to substrate but has no deamination activity. A3s deaminate cytosine to uracil on single strand DNA (ssDNA) and certain RNAs [114, 115] thus creating mutations.

Through deamination, A3s play crucial roles in innate immunity by mutating foreign pathogenic genomes and thus protecting host cells against retroviruses and retrotransposons [22, 23, 145-149]. Specifically, A3s deaminate cytosines to uracils on ssDNA during reverse transcription and thus create G to A hypermutations on the complementary strand. However, mis-regulated A3 deamination activity may promote cancer and the development of therapeutic resistance. Overexpressed A3s, especially A3A, A3B and A3H, have been shown to cause heterogeneities in multiple cancers, including breast, bladder, head and neck, cervical, and lung cancer[44, 45, 47, 48]. The A3 mutational signature, which is C to T transition in TC context, has been observed in multiple cancer genomes[43-45]. Moreover, study of human cancer cell lines has suggested A3s may be involved in the origination of cancer in human[167]. Recently, coupled with CRIPSR/Cas9, A3s are explored as novel base editors to treat genetic diseases[67, 68].

The structures of A3s provide the basis for understanding the underlying molecular mechanisms in A3 biology. Several crystal and NMR structures of human or primate A3 single domains (A3A, A3C, A3H; CTDs of A3B, A3F, A3G; NTDs of A3B, A3G) in the apo state have been determined by our group [79-85] and others [86-102]. The A3 domain fold consists of six alpha-helices and five beta-strands. The catalytic active site, which is also the zinc binding domain is in the middle (Figure 1A). Recently, our laboratory [104, 105], along with other groups, have solved the crystal structures of several A3–ssDNA complexes (A3A-DNA, chimeric A3B-CTD–DNA, A3G-CTD–DNA, A3F–DNA and rA3G-NTD–DNA) [99, 106-108]. These structures identified the binding conformation of DNA, revealed the critical residues for binding and provided insights into substrate specificity, especially at -1' position. Of these structures, three (A3A, A3B and A3G) has substrate DNA bound at the active site, with the target cytidine to be deaminated in essentially the same conformation. However, the rest of ssDNA can bind to A3 in different conformations; either in a U-shape as seen in A3A (PDB: 5KEG; 5SWW) or chimeric A3B-CTD (PDB: 5TD5) or a more extended linear (L) shape as seen in A3G-CTD (PDB: 6BUX). As these complex structures were determined with varying ssDNA sequences, the conformation of the bound DNA might be enzyme or substrate specific.

Although A3s share highly similar structural folds, they have varying levels of deamination activity and substrate specificity. For instance, the activity of A3A, which is the highest in A3 family, could be up to 5,000-fold higher compared to the least active A3D[113]. All A3 proteins deaminate deoxy-cytidines in ssDNA, but vary in their preferred hotspot sequences, 5'- (T/C)TC(A/G) for A3A, 5'-ATC(A/G) for A3B and 5'-CCC(A/C/T) for A3G[113, 117-120]. However, based on the amino acid sequence or even the available

structures, it is not apparent which molecular mechanisms are responsible for varied ssDNA binding affinity and deamination activity as well as substrate specificities among A3 domains. According to amino acid sequence alignment (**Figure 3.1B**), the loops (loop 1, 3, 5 and 7) surrounding the active site pocket of catalytically active domains are the most diverse. In addition, these active site loops undergo substantial conformational changes compared to apo structures upon ssDNA binding (**Figure 3.1A**). Therefore, detailed examination of the active site loops may help reveal the molecular mechanisms for the different substrate specificity, binding affinity and deamination activity for ssDNA, as well as the distinct physiological functions in A3s.

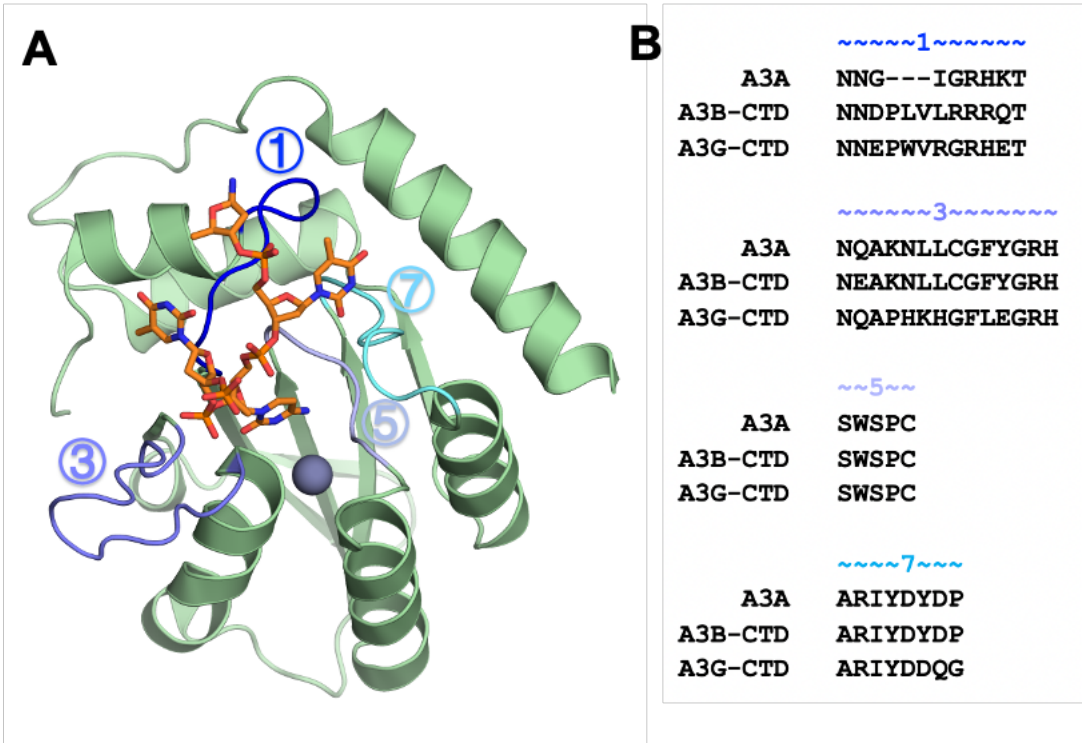

Figure 3.1:Structure and active site loops of A3s.
A. Cartoon representation of A3A bound to ssDNA (orange sticks) where the active site loops 1, 3, 5 and 7 are highlighted (PDB ID: 5KEG). **B.** Amino acid sequence comparison of active site loops in A3A, A3B-CTD and A3G-CTD.

In this study, we investigate the structural mechanism of substrate specificity and ssDNA binding conformation in A3s using a combination of molecular modeling, structural analysis, and parallel molecular dynamics (pMD) simulations. Three members of the human A3 family, namely A3A, A3B-CTD and A3G-CTD, were chosen for analysis as they have available structures and experimental characterization of substrate binding, in addition to high sequence similarity but varied substrate preference. These A3 domains were modeled with bound ssDNA of varying nucleotide sequences, and U or L-shape conformation, The results show an interdependence between substrate specificity and ssDNA conformation. Although the ssDNA was crystallized in a U-shape with A3A and (chimeric) A3B-CTD, we find that the wild-type enzymes can bind both U and L-shape to accommodate different substrates. Detailed analysis of inter-molecular interactions with the active site loops identified the molecular mechanisms of substrate sequence specificity at -1' and -2' positions. These results shed light into the structural mechanism of substrate specificity in A3s, which has implications for designing specific A3 inhibitors as well as base editing systems for gene therapy.

3.2 RESULTS

The three human A3s investigated, A3A, A3B-CTD and A3G-CTD, were modeled with substrate ssDNA bound either in a U or linear (L) shape (**Table 3.1**), with the DNA conformation based on that in the cocrystal structures with A3A and A3G-CTD, respectively. The preferred dinucleotide deamination motif is TC for A3A and A3B-CTD, and CC for A3G-CTD. Substrate DNA with either TC or CC motif, and either in U or L

shape was modeled bound to each of the three A3 proteins. The resulting models of A3–DNA pairs were subjected to energy minimization followed by fully solvated molecular dynamics simulations to analyze the stability of modeled complexes with a focus on inter-molecular interactions.

3.2.1 Substrate specificity and conformation correlate with overall dynamics in the simulations

The wilt type A3A–TC (U) model, which represents the complex in the cocrystal structure of A3A–DNA with the preferred substrate sequence, was stable during the MD simulations as expected. The bound DNA had relatively small root-mean-square fluctuations (RMSFs), less than 1.8 Angstroms (Å) for the central five nucleotides (**Figure 3.2A**). In addition, the key molecular interactions between the target cytidine and A3A observed in crystal structures were all well maintained: hydrogen bond interactions with His29, Thr31, Asn57, Ala71, Glu72, Ser99, Tyr130 and stacking interactions with Tyr130 and His70 (**Figure 3.3A; Figure 3.4**). Next, we examined whether A3A can bind ssDNA in a conformation similar to that bound in the A3G-CTD crystal structure. With the identical substrate ssDNA sequence, we modeled A3A-TC (L) with the DNA in a more linear conformation. This complex, had much higher ssDNA fluctuations compared to TC (U) according the RMSFs of each nucleotide (**Figure 3.2A**) Overall the RMSFs were about 2 folds higher compared A3A-TC (U) (2.9 Å compared to 1.6 Å at -3', 1.6 Å compared to 1.1 Å at -2', 1.1 Å compared to 0.7 Å at -1', 0.5 Å compared to 0.6 Å at 0', 1.8 Å compared at 0.9 Å at +1' and 2.8 Å compared to 1.8 Å at +2'). Besides, the gate-keeper residue His29, which is critical for stabilizing ssDNA

binding in the U-shape seen in the crystal structure, flipped back to apo conformation (**Figure 3.3A; Figure 3.4**). As a result, the stacking interactions with downstream nucleotides (+1, +2, +3) and hydrogen bonds with DNA backbone were lost. Together these results suggested that A3A prefers binding ssDNA with a TC motif in U-shape rather than L-shape. Then, we studied the dynamics of bound DNA with a CC sequence motif, which is ~3 fold less preferred compared to TC in A3A[114]. Interestingly, DNA with a CC motif in L-shape, A3A–CC (L) was stable similar to the A3A–TC (U) model, as indicated by both the RMSFs of bound DNA and stabilities of critical substrate interactions with A3A (**Figure 3.2A, 3.3A; Figure 3.4**). In the A3A-CC (U) model, zinc coordinating residue His70 lost stacking interactions with the target cytidine, which may result in destabilized active site coordination and lower activity. Nevertheless, A3A was able to stably bind ssDNA with both TC and CC motifs, in agreement with experimental data [114], but with the DNA in different binding conformations: U-shape helps stabilize TC while L-shape enables binding to CC motif.

Table 3.1: List of A3–DNA complexes for which MD simulations and analysis were performed in this study.

Protein	ssDNA	Abbreviation
A3A WT	AC**T**C**C**AAA (U)	A3A-TC (U)
	AC**T**C**C**AAA (L)	A3A-TC (L)
	AC**C**C**C**AAA (U)	A3A-CC (U)
	AC**C**C**C**AAA (L)	A3A-CC (L)
A3B-CTD WT	A**AT**C**C**AAA (U)	A3B-ATC (U)
	A**AT**C**C**AAA (L)	A3B-ATC (L)
	A**AC**C**C**AAA (U)	A3B-ACC (U)
	A**AC**C**C**AAA (L)	A3B-ACC (L)
	A**CT**C**C**AAA (U)	A3B-TC (U)
	A**CT**C**C**AAA (L)	A3B-TC (L)
	A**CC**C**C**AAA (U)	A3B-CC (U)
	A**CC**C**C**AAA (L)	A3B-CC (L)
A3G-CTD WT	AC**T**C**C**AAA (U)	A3G-TC (U)
	AC**T**C**C**AAA (L)	A3G-TC (L)
	AC**C**C**C**AAA (U)	A3G-CC (U)
	AC**C**C**C**AAA (L)	A3G-CC (L)
	A**AC**C**C**AAA (L)	A3G-ACC (L)
A3G-CTD2	AC**T**C**C**AAA (U)	A3G-CTD2-TC (U)
	AC**T**C**C**AAA (L)	A3G-CTD2-TC (L)
	AC**C**C**C**AAA (U)	A3G-CTD2-CC (U)
	AC**C**C**C**AAA (L)	A3G-CTD2-CC (L)

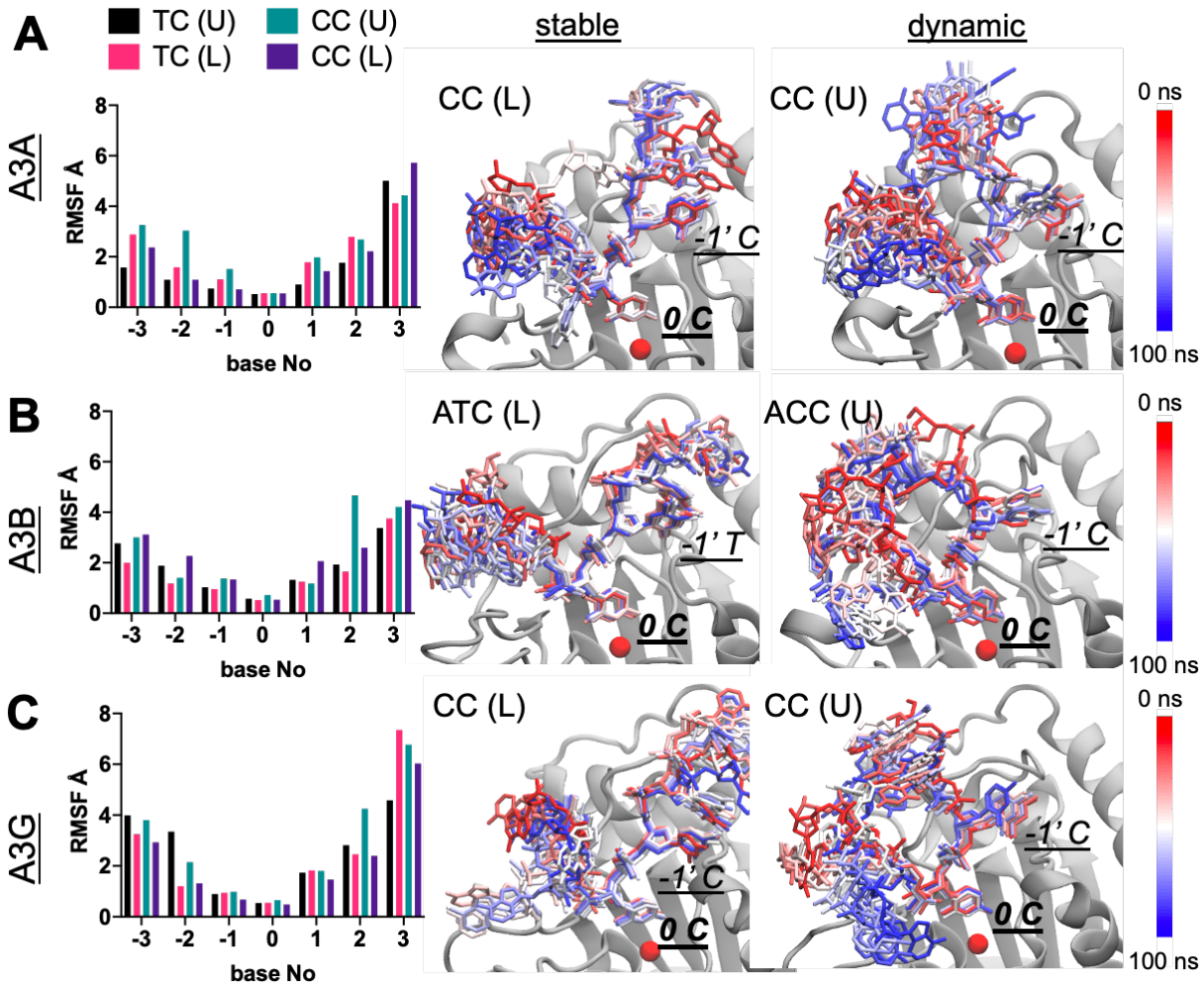

Figure 3.2: The dynamics of ssDNA in MD simulations.
The root-mean-squared fluctuations of each nucleotide are shown in the left column. Snapshots of ssDNA conformation from the simulations are superimposed and colored red to blue along the trajectory for examples of stable or dynamic complexes for **A**. A3A **B**. A3B-CTD **C**. A3G-CTD. The target (0' C) and -1' position nucleotide in the active site is labeled.

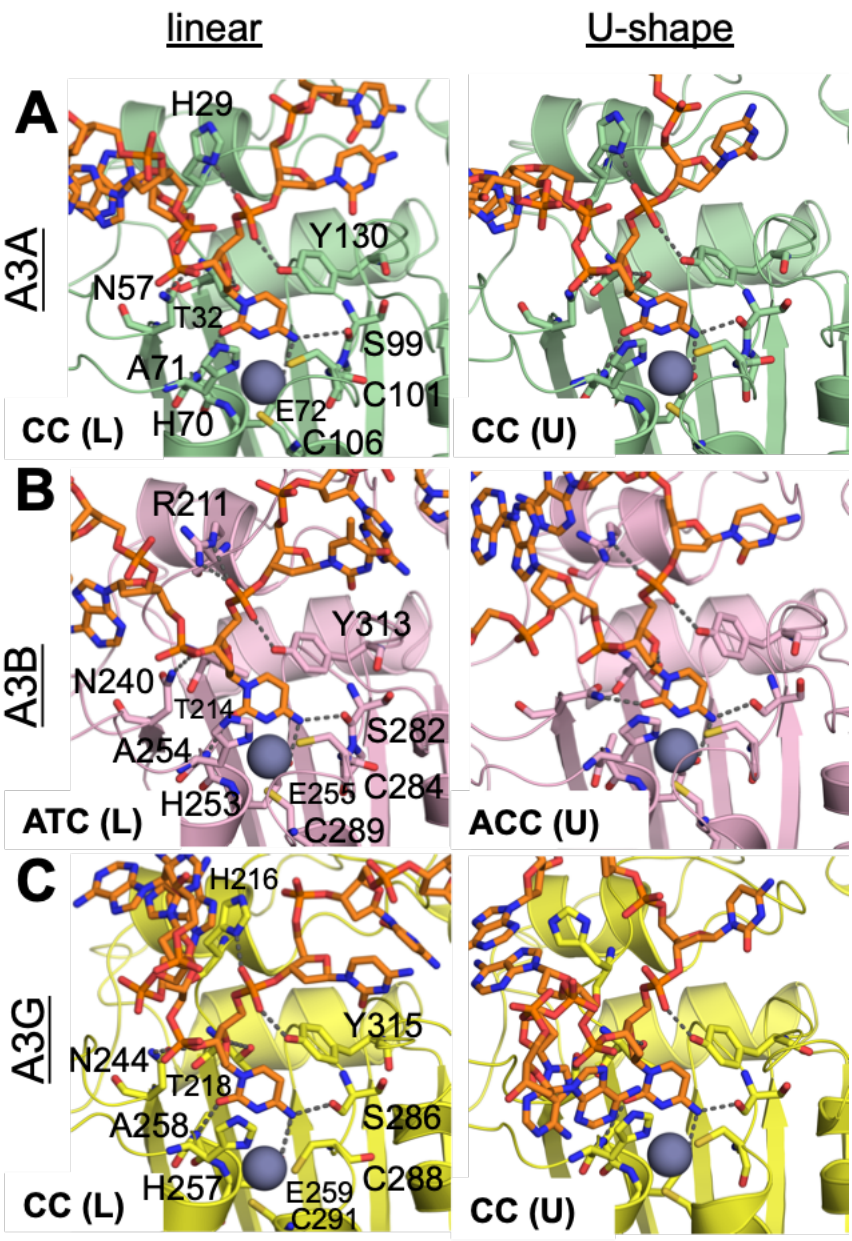

Figure 3.3: The interactions between target cytidine and active site residues in MD simulations of linear and U-shaped ssDNA.
A. A3A **B**. A3B-CTD **C**. A3G-CTD.

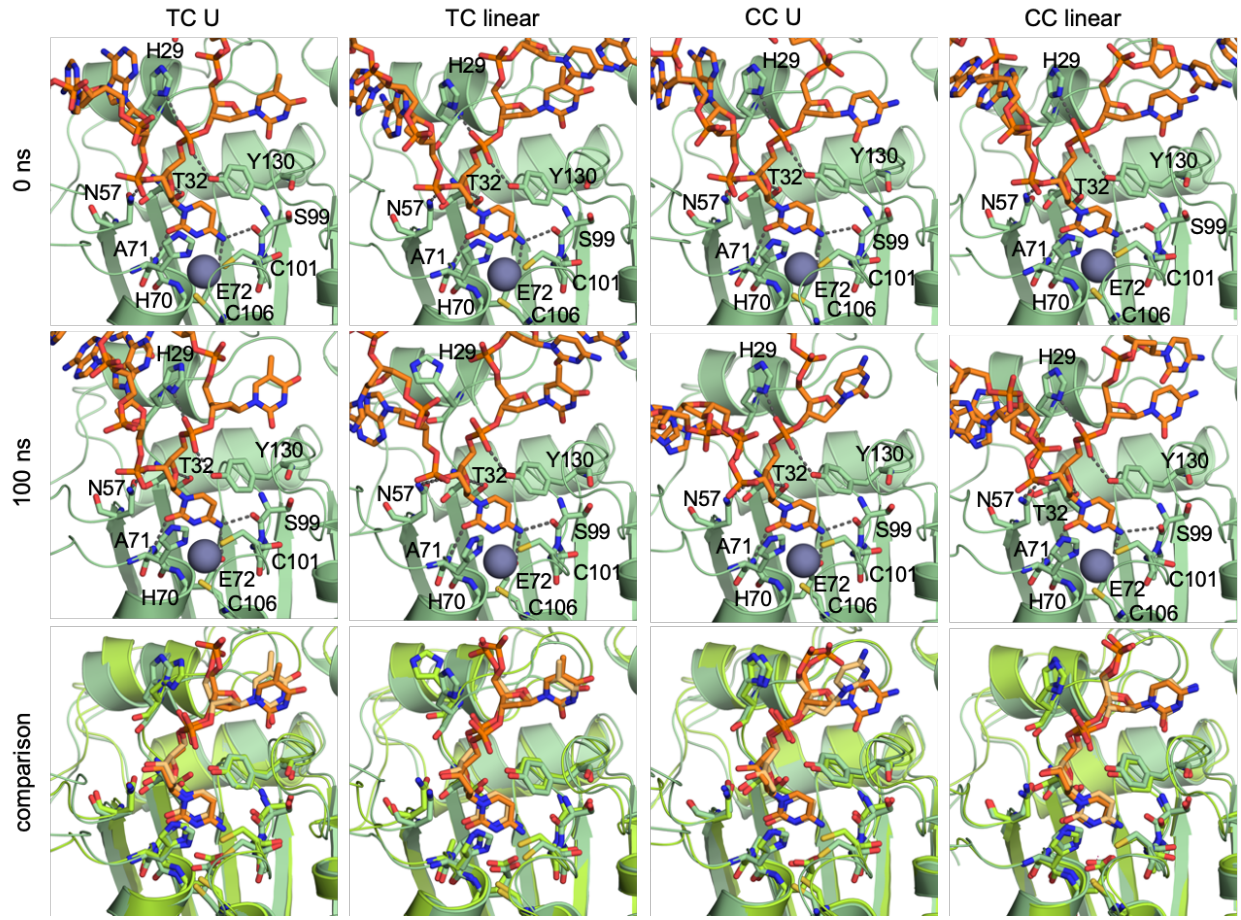

Figure 3.4: The comparison of the first and final frame from A3A simulations.

Similar to the case with A3A, the A3B-ATC (U) model, which is based on the crystal structure of chimeric A3B–DNA complex [106] and corresponds to our previously presented wild-type model [168], was stable during the MD simulations. However, unlike A3A which preferred U or L-shaped ssDNA depending on the target dinucleotide motif, A3B showed strong preference for TC over CC in MD simulations regardless of the DNA conformation. The RMSFs of bound DNA in A3B–ATC models, especially -1' and 0' nucleotides (less than 1 Å and 0.6 Å), were relatively small compared to others (**Figure 3.2B; Figure 3.5**). The molecular interactions between target cytidine and A3B were maintained in ATC models but not others (**Figure 3.3B; Figure 3.5**). For instance, in the A3B-ACC (U) model, the target cytidine was unstable in the active site. The conformation of side chain of active site residue Asn244 changed, similar to what was previously observed for A3G non-substrate (rC) simulations[115], which suggests that cytidine was not poised for deamination. In conclusion, A3B may accommodate both DNA binding conformations but not cytidine at -1' position.

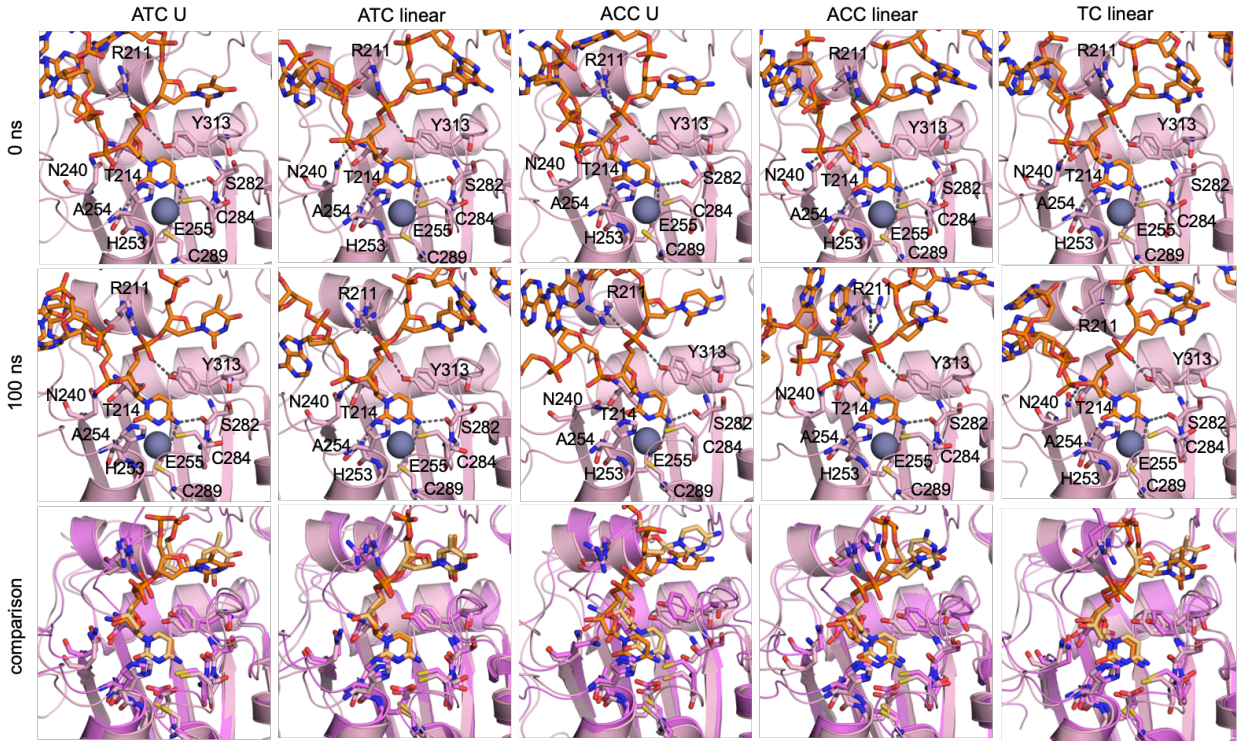

Figure 3.5: The comparison of the first and final frame from A3B simulations.

Overall, the MD simulations of A3G–DNA complexes were considerably more dynamic compared to A3A and A3B, which corelates with lower binding affinity (Supplementary Table 2). The wild type A3G–CC (L) model, which represents the crystal structure, was the most stable A3G complex during the MD simulations as indicated by relatively low fluctuations of bound DNA (especially the central 5 nucleotides: RMSF < 2.4 Å compared to 2.8 Å for TC (U), 4.3 Å for CC (U) and 2.5 Å for TC (L)) and the most stable interactions between target cytidine and protein (**Figure 3.2C; Figure 3.3C; Figure 3.6**). More specifically, the hydrogen bond interactions between His216 and ssDNA were lost in TC (U) and CC (U) models while the stacking interactions between His216 and downstream bases were lost in TC (L) and CC (U) models.

Overall out of the 12 models, 5 corresponded to stable A3–DNA complexes which were further analyzed and compared. These are A3A-TC (U), A3A-CC (L), A3B-ATC (U), A3B-ATC (L) and A3G-CC (L) models.

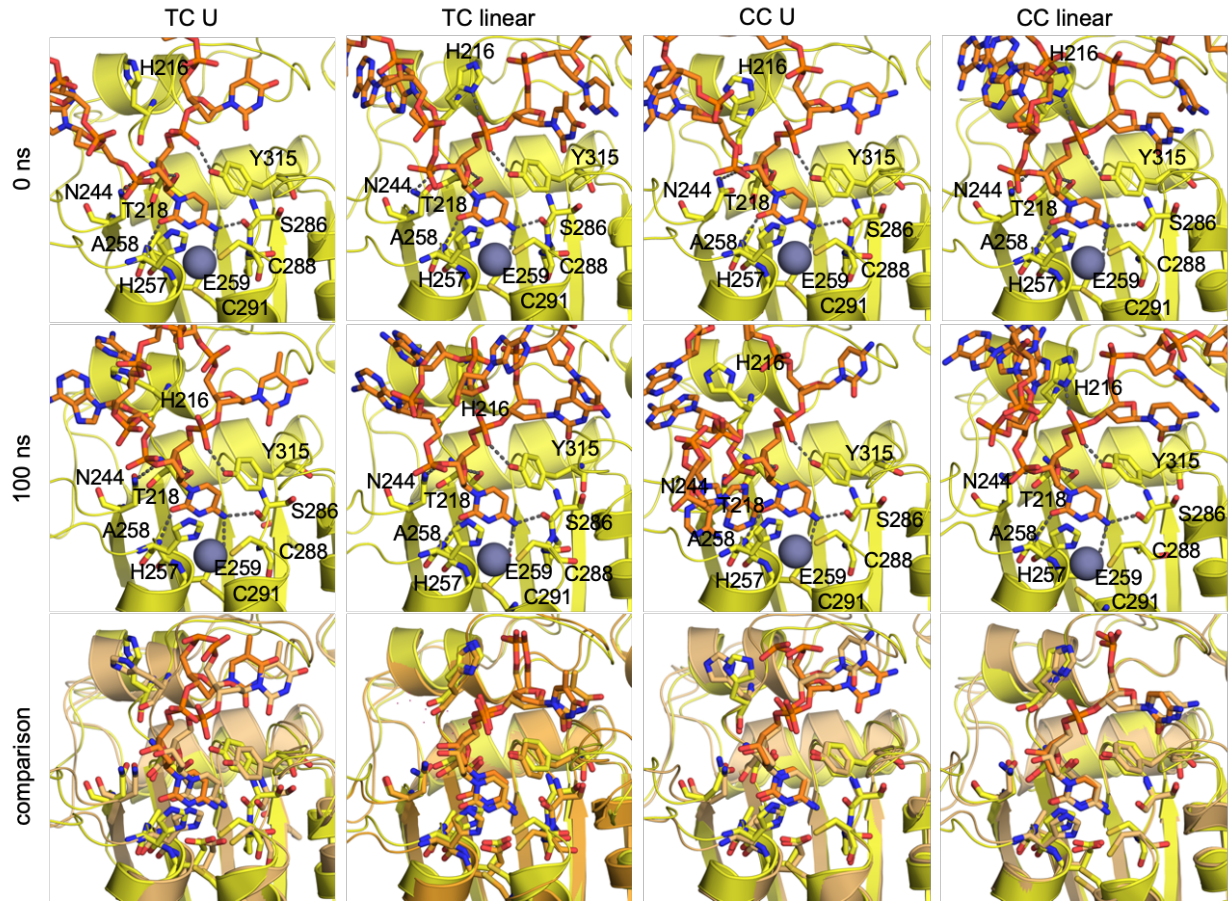

Figure 3.6: The comparison of the first and final frame from A3G simulations.

3.2.2 Loop 1 is important for defining the ssDNA binding conformation

Among all the active site loops, loop 1 made the most extensive molecular contacts with ssDNA. ssDNA either wrapped around (as seen in U-shape models) or extended along loop 1 (as seen in L-shape models) (**Figure 3.7A**). As a result, loop 1 contributed the most van der Waals (vdW) interactions with ssDNA to stabilize binding (**Figure 3.7B**).

The shorter loop 1 with a three-residue deletion in A3A may allow ssDNA to bind in both conformations. In A3A–TC (U) model, ssDNA wrapped around the gate keeper residue His29 in loop 1. Arg28 in loop 1 stacked with upstream bases (-2; -3) and thus stabilized the U conformation. As a result, His29 and Arg28 had the most vdW contacts with ssDNA. In A3A–CC (L) model, the three-residue deletion in loop 1 allowed ssDNA to reach out and make interactions with His182 in alpha-helix 6. Besides, Ile26 in loop 1 packed with upstream bases and thus had the second highest vdW interactions with ssDNA.

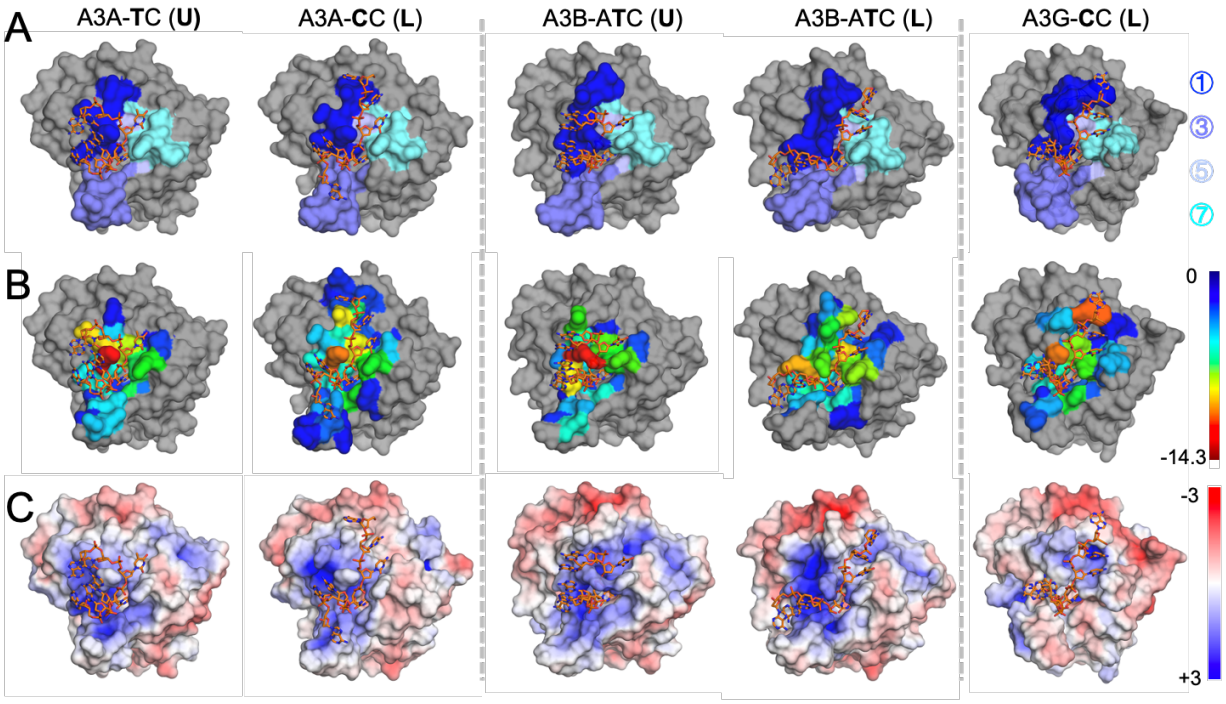

Figure 3.7: Active site loops and electrostatics of ssDNA–A3 complexes displayed for representative frames from MD simulations.
Proteins are shown as surface representation, and ssDNA is shown as orange sticks. **A**. Active site loops are colored blue and shown on protein surface. **B**. The residues that interact with ssDNA are colored blue to red for increasing vdW interactions. **C**. The electrostatics of protein surfaces, where red is negative and blue is positive charge.

Both A3B and A3G has a longer loop 1 compared to A3A. The Trp211 in the PWV insertion in A3G stacks with -3' base in the A3G–DNA co-crystal structure and thus stretches ssDNA binding into a more extended L-shape. Similarly, Pro206 in loop 1 packed with -3 nucleotide in A3B–ATC (L) model thus stabilizing the overall extended L-shape. The U-shape binding conformation of A3B may be defined by Arg210 in loop 1. Arg210 in A3B has a unique side chain conformation compared to other A3s[96, 168]: Arg210 instead of Arg313, which is conserved among A3s, stabilizes overall structure through the conserved hydrogen bond network as shown in the apo crystal structure. As a result, the side chain of Arg210 is oriented toward the core of the protein and thus results in a cavity next to the gate-keeper residue Arg211. This cavity may allow ssDNA to wrap around Arg211 and bind in U conformation as seen in A3B–ATC (U) model.

For A3G-CTD, Trp211 in loop 1 is critical for the L binding conformation of ssDNA. Trp211 formed strong stacking interactions with upstream nucleotides in both A3G L model and crystal structure and thus had the most vdW contacts. Interestingly, the active site of A3G-CTD is also less positively charged compared to A3A and A3B-CTD (**Figure 3.7C**). Less positively charged active site may decrease the binding affinity in A3G-CTD considering the strong negatively charged DNA backbone, correlating with the lower binding affinity (**Table 3.2**).

Table 3.2: Binding affinity (Kd) for linear and hairpin ssDNA with preferred sequence by A3s.

Protein	ssDNA/hairpin	Kd
A3A	AAA-ACC-AAA-AAA (Linear) [1]	250 ±14 nM
	AAA-ATC-AAA-AAA (Linear) [1]	145 ± 2 nM
	AAA-ATC-GAA-AAA (Linear) [1]	154 ± 2 nM
	AAA-CTC-AAA-AAA (Linear) [1]	85 ± 1 nM
	GCC-ATC-ATT-CGA-TGG-G (hairpin) [1]	26 ± 2 nM
A3B-CTD	AAA-AAA-AAT-CGA-AAA (Linear) [2]	5.4 ± 2.6 μM
	GCC-ATC-ATT-CGA-TGG-G (hairpin) [2]	2.0 ± 0.5 μM
A3G-CTD	AAT-CCC-AAA (Linear) [3]	160 μM

[1] Silvas et al, Scientific Reports 8.1 (2018): 7511.
[2] Hou et al, JCTC 15.1 (2018): 637-647.
[3] Maiti et al, Nat Commun 9.1 (2018): 2460.

3.2.3 Substrate specificity at -1' position

As explained above, A3A can bind to both TC and CC dinucleotide motifs through accommodating different DNA binding conformations. According to previous reports ([114], A3A can bind either thymidine or cytidine at -1' position with slight preference of T over C (K_d ~85 nM versus ~250 nM). Our MD simulations showed similar results: thymidine in A3A–TC (U) model had stable hydrogen bonds with the side chain of Asp131 and the backbone of Tyr132 (**Figure 3.8A** left). The same hydrogen bonds were also observed in A3A–DNA co-crystal structure [104, 106]. Cytidine in A3A–CC (U) model, however, showed more fluctuations compared to -1' T in the A3A–TC (U) model. The RMSF of cytidine in CC (U) model was about 2-fold higher compared to thymidine (**Figure 3.1A; Figure 3.2**). In addition, -1' C lost the hydrogen bond interactions with the backbone of Tyr132. Interestingly, -1' C in A3A–CC (L) model revealed stable interactions with A3A protein. The RMSF of -1' C in CC (L) model was about the same as -1' T in TC (U) model. Moreover, -1'C formed stable hydrogen bonds with the backbone of Tyr132 and side chain of Asp131 though different side chain conformation (**Figure 3.8A** right). Together these results suggested that A3A may use alternative DNA binding conformations to adapt different substrate sequences.

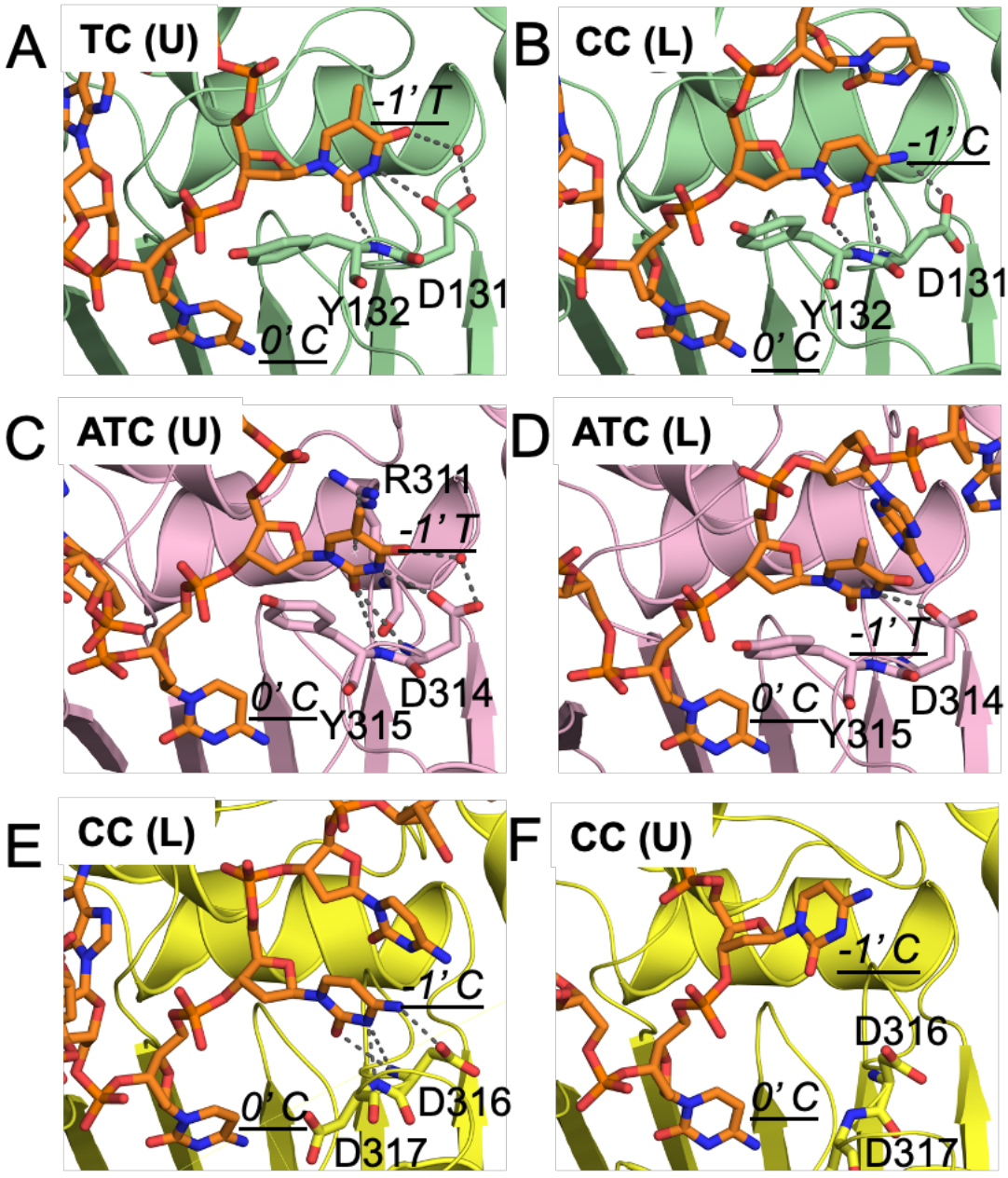

Figure 3.8: The molecular interactions between ssDNA and A3 active site at -1' position.
(**A**) A3A (**B**) A3B-CTD (**C**) A3G-CTD.

A3B exhibited a preference for TC rather than CC at -1' position. In A3B–ATC (U) model, -1' T formed stable hydrogen bonds with the backbone of Tyr315 and the side chain of Asp314, which was similar to A3A and consistent with our previous results [168] (**Figure 3.8B** left; **Table 3.2**). Interestingly, in the A3B–ATC (L) model, unlike A3A–CC (L), Asp314 maintained the same side chain conformation as in U-shape model and thus formed hydrogen bond that promoted T over C (**Figure 3.8B** right). When having cytidine at -2 position, A3B may bind cytidine at -1' position with ssDNA binding in L conformation. In the A3B–CC (L) model, Asp314 and Tyr315 maintained the same hydrogen bonds to -1' C as in A3A (hydrogen bond occupancy: 71%, 54%; 83%). These results suggest that the substrate specificity at -2' position may affect the specificity at -1' position in A3B; A3B prefers thymidine at -1' position but could bind ssDNA in different conformations for different substrates.

A3G binds ssDNA in L shape and thus prefers CC rather than TC at -1' position. A3G is the only A3 that prefers cytidine over thymidine at -1' position. From our MD simulations, the side chain of Asp316 and the backbone of Asp317 made stable hydrogen bonds with cytidine (**Figure 3.8C** left) in A3G-CC (L) model. These direct hydrogen bonds were lost in A3G–CC (U) model (**Figure 3.8C** right). All together our results suggest that there might be interdependent interactions between ssDNA binding conformation and substrate sequence specificities.

3.2.4 Substrate specificity at -2' position

Previously, we revealed that intra-DNA interactions in A3A may underlie the substrate specificity of T/C at -2' position[114]. However, how A3B or A3G defines its

substrate specificity had remained elusive. To address the mechanism of specificity at -2' position in A3B, we created additional models of A3B with CTC sequence to compare with the ATC described above, which is the preferred sequence for A3B [119]. In A3B–ATC (L) model, -2' A formed stable hydrogen bonds with the side chain of Arg311 (occupancy 38% during the simulations), Ile312 (occupancy 90%), Asp314 (occupancy 98%) and stacking interactions with Trp281 (occupancy 32%) (**Figure 3.9A** left). However, all these interactions were lost in the A3B–CTC (L) model (**Figure 3.9A** right). The vdW interactions between Trp281 and -2' nucleotide was also decreased from -9.2 kcal/mol in ATC model to -7.9 kcal/mol in the CTC model. Moreover, the side chain of the gate-keeper residue for DNA binding, Arg211, lost interactions with DNA backbone, which may impair the binding of target cytidine in the active site (**Figure 3.5**).

To study the specificity of -2' position, A3G was modeled with ACC sequence to compare with CCC, which is the preferred motif for A3G [86, 169-171]. In A3G–CC (L) model, -2' C was locked in an extensive hydrogen bonding network with residues Pro210 (water-mediated; occupancy 37%), Arg374 (water-mediated; occupancy 31%, 41%), Ile314 (water-mediated; occupancy 41%), Val212 (water-mediated; occupancy 34%) and Asp316 (occupancy 43%) (**Figure 3.9B** left). Interestingly, the larger base of A in ACC (L) model did not occupy the space where water-mediated hydrogen bonds were; instead -2' A had only one hydrogen bond with Asp316 (occupancy 42%) (**Figure 3.9B** right). Thus, A3G preferred to accommodate the smaller C at the -2 position through an extensive water-mediated hydrogen bonding network.

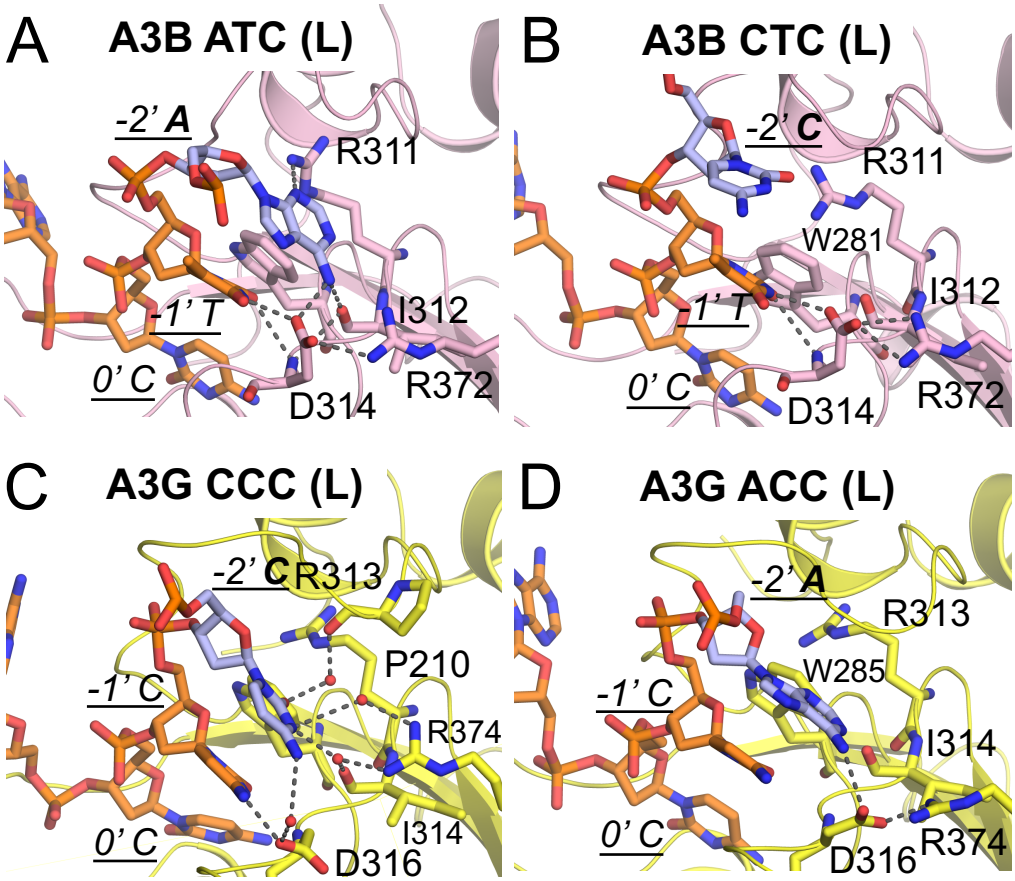

Figure 3.9: The molecular interactions between ssDNA and A3 active site at -2' position.
A. A3B-CTD, **B**. A3G-CTD.

3.2.5 Interdependent interactions between substrate specificities at nucleotide positions

Analysis of pMD performed on the A3 enzymes suggested interdependence between substrate specificities various nucleotide positions. First of all, the specificity at -1' position may affect the binding of the target nucleotide in the active site (0' position). Having a disfavored nucleotide at the -1' position destabilized the target nucleotide. Second, the specificity at -2' position may influence the specificity at -1' position. In A3B, the interactions between -2' A and Asp314 locked the side chain of Asp314 in the conformation that promotes thymidine over cytidine at -1' position (Figure 6A left). Similar interdependent interactions were also observed in A3G (Figure 6B). Finally, the ssDNA binding conformation (U or L shape) also impacted the substrate sequence specificity.

3.3 DISCUSSION

In this study, we investigated the structural mechanism for substrate specificities in A3s. Interestingly, we found an interdependence between substrate conformation and specificity. In addition to the U-shape binding conformation of substrate DNA observed in crystal structures, we found that A3A and A3B can bind DNA in a more linear conformation. Specifically, the linear conformation helps accommodate CC dinucleotide motif while the U-shape prefers TC. The active site loops play important roles in defining the overall binding surface and conformation for DNA binding to A3s. For A3A, A3B and A3G, loop 1 is critical with extensive interactions with DNA. The gate-keeper residues (His29 in A3A, Arg211 in A3B and His216 in A3G), which locks DNA in the active site,

are all in loop 1. Besides, the three-residue insertion in loop 1 of A3B and A3G compared to A3A seems to be important for defining the more extended L-shape through stacking interactions. Previous studies have shown that swapping loop 7 from A3G into A3B altered the substrate specificity in A3B from TC to CC[96]. Additionally, changing Asp317 of A3G into the corresponding residue of A3A (Tyr132) caused A3G to adopt a more A3A-like 5'-TC preference[172]. These results suggested the Tyr132/Tyr315/Asp316 in loop 7 might be important for substrate specificity at -1' position. However, in our analysis there is no stable side chain interactions between this residue and -1' base during the MD simulations. Instead the tyrosine might be important for stabilizing the U-shape of DNA, which is supported by the considerably higher vdW contacts with DNA compared to aspartic acid, and which may switch the preference to TC. Finally, we observed interdependence between various nucleotide binding sites. Unlike human CDAs, A3s require at least five nucleotides for stable binding of DNA. The interdependence between binding interactions around the active site suggests the sequence of nucleotides flanking the target cytidine is also critical for stable substrate binding.

Considering the roles of A3s in viral infections and cancer, a better understanding of the mechanism by which A3s recognize different oligonucleotides will be critical for developing therapeutics. Currently, combined with catalytically inactive Cas9 (dCas9), A3s are investigated as novel base editors for direct modification of genomic DNA at single-base resolution [67, 68]. As cytosine base editors (CBE), A3s can create mutations to potentially correct genetic diseases but still require improvements. Several studies have reported significant off-target effects of CBEs. In addition, CBEs have problems

with product purity and editing window length[51]. To overcome these problems, several versions of CBEs have been engineered, UGI added to increase product purity[72] and generate high fidelity Cas9 with reduced off-target effects[73], or different Cas nucleases[74-78] were used for narrower activity windows. However, modifications that have been implemented to improve the efficiency and fidelity of CBEs have not focused on the deaminase. Another major problem in applying CBEs to treat genetic diseases is that the target site must naturally exist in the preferred DNA sequence context for cytidine deaminase, which may not be the case for the desired modification. Therefore, having a library of A3s with different substrate specificities as context-dependent base editors would expand the toolkit available for base editing. Our results revealing the molecular mechanisms underlying A3 specificities may help guide engineering of A3s, especially with modifications to the active site loops, to rationally design A3s to adapt the desired sequence specificity.

3.4 EXPERIMENTAL PROCEDURES

3.4.1 Protein sequence alignment

Protein sequence alignment was generated by program Geneious 9.0.5 using default Multiple alignment.

3.4.2 Molecular modeling

All structure models in this study were first generated from program MODELLER9.23; then optimized using Protein Preparation Wizard in Maestro from Schrodinger suite. The optimization was performed at pH 7.0; H-bond assignment panel

with minimize hydrogens of alter species function; minimized using restrained minimization panel. The ssDNA-bound crystal structures of A3A (PDB: 5KEG as protein template; 5SWW as ssDNA template) and A3G-CTD2 (PDB: 6BUX) were used as templated for molecular modeling of wild type A3-ssDNA complexes. Different ssDNA sequences were mutated through program Coot. ssDNA-bound A3B-CTD structures were modeled using the crystal structures of both apo A3B-CTD (PDB: 5CQH) and A3A-ssDNA complex.

3.4.3 Molecular dynamics simulations

All molecular dynamics simulations were performed for 100 ns using program Desmond from Schrodinger suite. The simulation systems were built using SPC solvation model and cubic boundary conditions of 12 Å buffer box size with OPLS3 force field through system builder in Maestro. The final systems were neutral and had 0.15 M sodium chloride. A multi-stage MD simulation protocol was used, which was previously described.

3.4.4 Analysis of molecular dynamics simulations

The RMSD and molecular interactions (hydrogen bond; stacking interaction) occupancies over trajectories were calculated using Simulation Interaction Diagram in Maestro from Schrodinger. The per base RMSFs of ssDNA were calculated using Schrodinger python API. The residue vdW potential between A3s and ssDNA during the MD simulations was extracted from the simulation energies using Desmond.

The frame that closet to average RMSD was used as representative structure for each MD simulation. The electrostatic distributions were calculated using PDB2PQR server and Pymol with the APBS plugin; and visualized with contour levels positive (+3) and negative (-3). The time series representations of ssDNA were generated with program VMD using 2000 frames as time step (total 20000 frames for each MDs). All other structural graphics were made using program PyMol.

4 Chapter VI: Substrate Sequence Selectivity of APOBEC3A Implicates Intra-DNA Interactions

Chapter IV is a collaborative study that has been previously published as:

Silvas TV, **Hou S**, Myint W, Nalivaika EA, Somasundaran M, Kelch BA, Matsuo H, Kurt Yilmaz N, Schiffer CA. "Substrate sequence selectivity of APOBEC3A implicates intra-DNA interactions." *Scientific Reports* 8.1 (2018): 7511.

4.1 ABSTRACT

The APOBEC3 (A3) family of human cytidine deaminases is renowned for providing a first line of defense against many exogenous and endogenous retroviruses. However, the ability of these proteins to deaminate deoxycytidines in ssDNA makes A3s a double-edged sword. When overexpressed, A3s can mutate endogenous genomic DNA resulting in a variety of cancers. Although the sequence context for mutating DNA varies among A3s, the mechanism for substrate sequence specificity is not well understood. To characterize substrate specificity of A3A, a systematic approach was used to quantify the affinity for substrate as a function of sequence context, length, secondary structure, and solution pH. We identified the A3A ssDNA binding motif as (T/C)TC(A/G), which correlated with enzymatic activity. We also validated that A3A binds RNA in a sequence specific manner. A3A bound tighter to substrate binding motif within a hairpin loop compared to linear oligonucleotide, suggesting A3A affinity is modulated by substrate structure. Based on these findings and previously published A3A–ssDNA co-crystal structures, we propose a new model with intra-DNA interactions for the molecular mechanism underlying A3A sequence preference. Overall, the sequence and structural preferences identified for A3A leads to a new paradigm for identifying A3A's involvement in mutation of endogenous or exogenous DNA.

4.2 INTRODUCTION

The APOBEC3 (short for "apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like") family of human cytidine deaminases provides a first line of defense against many exogenous and endogenous retroviruses such as HIV-1 and the retro-

element LINE-1 [22, 145-149]. APOBEC3 (A3) proteins restrict replication of retroviruses by inducing hypermutations in the viral genome [23]. A3s deaminate deoxycytidines in ssDNA into uridines during reverse transcription. This results in G to A hypermutations, as adenosines are transcribed across from uridines during second strand DNA synthesis. While all A3 enzymes deaminate deoxycytidines in ssDNA, they have differential substrate specificities that are context dependent, resulting in altered frequencies of mutation for the deoxycytidines. Some A3s deaminate the second deoxycytidine in a sequence containing C<u>C</u> while others deaminate deoxycytidine in a T<u>C</u> context [24, 150, 173]. However, not every cognate dinucleotide motif (C<u>C</u> or T<u>C</u>) in the ssDNA of the HIV genome is deaminated [174]. Nevertheless, hypermutation in a viral genome results in defective proteins and proviruses, thus decreasing the probability of further viral replication [175].

Beyond restricting viral replication, the ability of A3s to deaminate deoxycytidines in ssDNA have made A3s a double-edged sword. When overexpressed, A3s can mutate the host genome resulting in a variety of cancers. The identities and patterns of the mutations observed in cancer genomes can define the source of these mutations. Recently, the search for the deaminase(s) responsible for kataegic mutations found in breast cancer was narrowed down to APOBEC3B, through the comparison of all known APOBEC mutational signatures and eliminating APOBEC3G and other deaminases from potential mutational contributors [41, 150]. Soon after, APOBEC3B was found to be correlated with a variety of other cancers such as ovarian, cervical, bladder lung, head and neck; signature sequence analysis was also a contributing factor that led to these conclusions [44, 154]. Most recently APOBEC3H, which has a different sequence

preference than APOBEC3B, has been identified to also play a role in breast and lung cancer [47]. Thus, defining A3 sequence specificity can be helpful in identifying A3's role in viral restriction and in cancer.

A3 signature sequences proposed for deaminating deoxycytidines range between di-nucleotide to quad-nucleotide motifs [24, 47, 94, 98, 106, 150, 173, 174, 176, 177]. Although A3s are known to have varied sequence preference, quantitative and systematic studies of sequence specificity are incomplete. Recently, crystal structures of APOBEC3A (A3A) and APOBEC3B-CTD (an active site A3A chimera) with ssDNA have been solved [104, 106]. However, despite these breakthrough structures, the molecular mechanism underlying substrate sequence specificity flanking the TC dinucleotide sequence remains unclear.

A3A is a single-domain enzyme with the highest catalytic activity among human APOBEC3 proteins [178] and a known restriction factor for the retroelement LINE-1 and HPV [179, 180]. A3A can also contribute to carcinogenesis with increased expression or defective regulation [181]. A3A is the only A3 where both the intact apo and substrate bound structures have been determined [84, 93, 94, 104, 106]. Initial substrate specificity studies have shown a preference for DNA over RNA, suggested by NMR chemical shift perturbation [94]. Since A3A is the best biochemically characterized A3 human cytidine deaminase and thus a critical benchmark within the family, we chose A3A to elucidate the extended characteristics of ssDNA specificity.

To determine the substrate specificity of A3A, we systematically quantified the affinity of A3A for nucleic acid substrates as a function of substrate sequence, length, secondary structure, and solution pH. We identified the A3A preferred ssDNA binding

motif, (T/C)TC(A/G) and found this sequence correlated with enzymatic activity. Also, we determined that A3A can bind RNA in a sequence specific manner. Surprisingly, A3A's signature sequence was necessary but not sufficient to account for A3A's high affinity for ssDNA. Significantly, A3A bound more tightly to the motif in longer oligonucleotides, and in the context of a hairpin loop. Using recently published structures of A3As complexed with ssDNA from our lab and others, we propose a structural model for the molecular mechanism for this enhanced affinity where inter-DNA interactions contribute to A3A recognition of the cognate sequence. This model provides insights into how the nucleotides flanking the canonical TC sequence may contribute to substrate sequence preference of A3A.

4.3 RESULTS

4.3.1 A3A binding to ssDNA is context dependent

To interrogate the substrate sequence preference of A3A, we systematically quantified the changes in binding affinity of catalytically inactive A3A bearing the mutation E72A to a library of labeled ssDNA sequences using a fluorescence anisotropy-based DNA binding assay [84]. First, to ensure that the affinity for substrate was due entirely to the sequence of interest and not due to nonspecific binding or undesired secondary structure effects, an appropriate control background sequence was identified. The dissociation constants (K_d's) for homo-12-mer ssDNA sequences, Poly A, Poly T, Poly C, were determined (**Figure 4.1A)**. Poly G was not tested due its propensity to form secondary structure elements. Poly T (750 ± 44 nM), which had previously been used in background sequences [84], bound to A3A with 2-fold higher

affinity than Poly C (1,600 ± 117 nM). Thus without a greater context for A3A to target, Poly C was only weakly bound. A3A had the lowest affinity for Poly A with a K_d of >11,00 nM (**Table 4.1**). For all subsequent assays, Poly A was used as the background, as there is no detectible binding affinity of A3A to Poly A.

The specificity of A3A for substrate versus product was measured by binding to Poly A with a single C versus Poly A with a single U (**Figure 4.1B**). Surprisingly, the presence of a single deoxycytidine in a Poly A background was not sufficient for binding with appreciable affinity. The affinity of A3A for the Poly A-C (5A-1C-6A) (>5,000 nM) is similar to the affinity for Poly A-U (5A-1U-6A) (>6,500 nM) and even the background Poly A. This is in contrast to A3A's specificity for binding a single C over U in a Poly T background, which is more than ten-fold (35 ± 2 nM and 500 ± 23 nM respectively) (**Figure 4.1C**), as we previously measured [84]. This strong context dependence differentiating substrate C versus product U within the background of Poly A versus Poly T indicates that A3A heavily relies on the identity of the surrounding nucleotide sequence to recognize and bind substrate deoxycytidine.

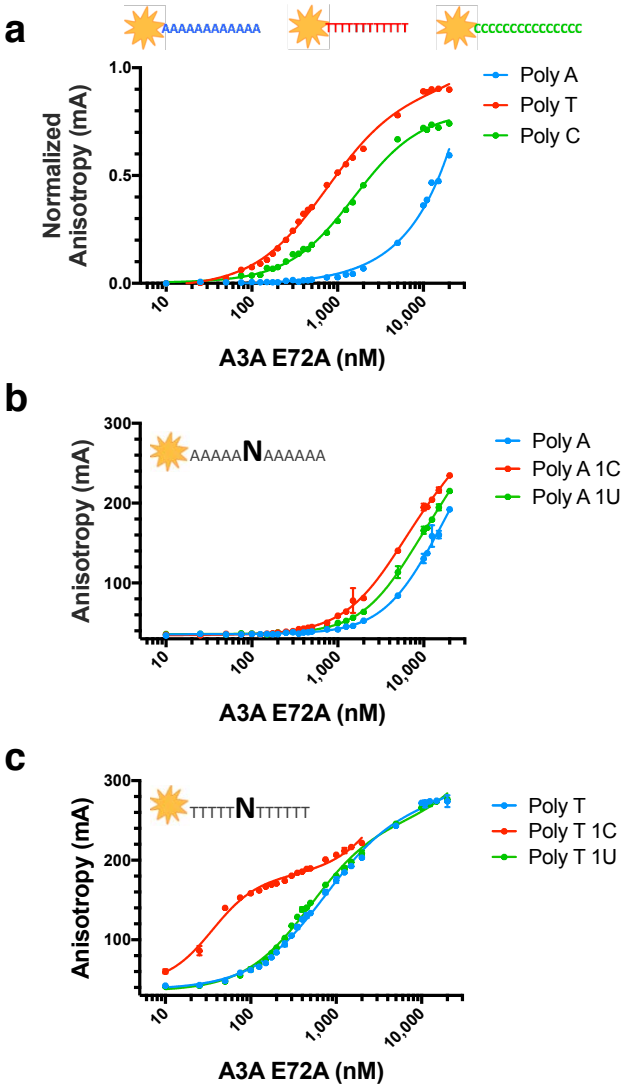

Figure 4.1: A3A specificity to ssDNA background and substrate.
Fluorescence anisotropy of TAMRA-labeled ssDNA sequences binding to A3A(E72A). A) Binding of A3A to poly nucleotide (12 mers): Poly A (blue), Poly T (red) and Poly C (green), B) Binding to Poly A (blue), 5A-C-6A (red), 5A-U-6A (green), C) Binding to Poly T (blue), 5T-C-6T (red), 5T-U-6T (green).

Table 4.1: A3A affinity for DNA sequences used in this analysis.

DNA sequence	KD (nM)
Poly C (12C)	1,568 ± 117
Poly T (12T)	748 ± 44
Poly T-C (5T-1C-6T)	35 ± 2
Poly T-U (5T-1U-6T)	499 ± 23
Poly A (12A)	>11,000
Poly A-C (5A-1C-6A)	>5,000
Poly A-U (5A-1U-6A)	>6,500
Poly A-TC (4A-TC-6A)	143 ± 4
Poly A-CC (4A-CC-6A)	250 ± 14
Poly A-GC (4A-GC-6A)	>6,500
Poly A-ATUA (3A-ATUA-5A)	>5,000
Poly A-ATUG (3A-ATUG-5A)	328 ± 42
Poly A-TTUA (3A-TTUA-5A)	306 ± 17
Poly A-ATCA (3A-ATCA-5A)	145 ± 2
Poly A-ATCT (3A-ATCT-5A)	209 ± 5
Poly A-ATCC (3A-ATCC-5A)	163 ± 3
Poly A-ATCG (3A-ATCG-5A)	154 ± 2
Poly A-TTCA (3A-TTCA-5A)	90 ± 1
Poly A-TTCT (3A-TTCT-5A)	127 ± 2
Poly A-TTCC (3A-TTCC-5A)	114 ± 2
Poly A-TTCG (3A-TTCG-5A)	92 ± 2
Poly A-CTCA (3A-CTCA-5A)	85 ± 1
Poly A-CTCT (3A-CTCT-5A)	122 ± 2
Poly A-CTCC (3A-CTCC-5A)	101 ± 2
Poly A-CTCG (3A-CTCG-5A)	86 ± 1
Poly A-GTCA (3A-GTCA-5A)	150 ± 3
Poly A-GTCT (3A-GTCT-5A)	218 ± 7
Poly A-GTCC (3A-GTCC-5A)	152 ± 2
Poly A-GTCG (3A-GTCG-5A)	150 ± 3
Hairpin-TTC (G-CCATC-ATTC-GATGG-G)	26 ± 2
Hairpin-AAA (G-CCATC-AAAA-GATGG-G)	676 ± 399

4.3.2 A3A affinity for ssDNA is pH dependent

A systematic measurement of A3A affinity in a broad range of pH values was performed to verify and quantify the pH dependence of A3A to substrate ssDNA [98, 181], and set a reference pH for subsequent experiments. The K_d of A3A for TTC in a Poly A background was determined at pH ranging from 4.0 to 9.0 in 0.5 pH increments (**Figure 4.2 and Table 4.2**). A3A had the highest affinity for Poly A-TTC at pH 5.5 with a K_d of 68 ± 3 nM. The isotherms for A3A binding ssDNA at pHs below 6.0 show some secondary binding event that may be due to non-specific binding or aggregation (**Figure 4.2A**). A steady decrease was also observed for the affinity of A3A for ssDNA when pH was increased above 6 (**Figure 4.2B**), in agreement with decreased deamination activity at higher pH [181]. A3A affinity also overall correlated with reported deamination activity determined using a different assay at pH 7.5 [182]. Interestingly, A3A had no appreciable affinity for Poly A-TTC above pH 8.0. Since A3A is stable at these higher pH values, the lower affinity for ssDNA with increased pH is likely not due to aggregation but due to the protonation of His 29, as previously described [181] and reported to be responsible for coordinating ssDNA [183]. Therefore, all of the subsequent binding experiments were performed at pH 6.0 to avoid any potential for secondary binding events or aggregation of the protein.

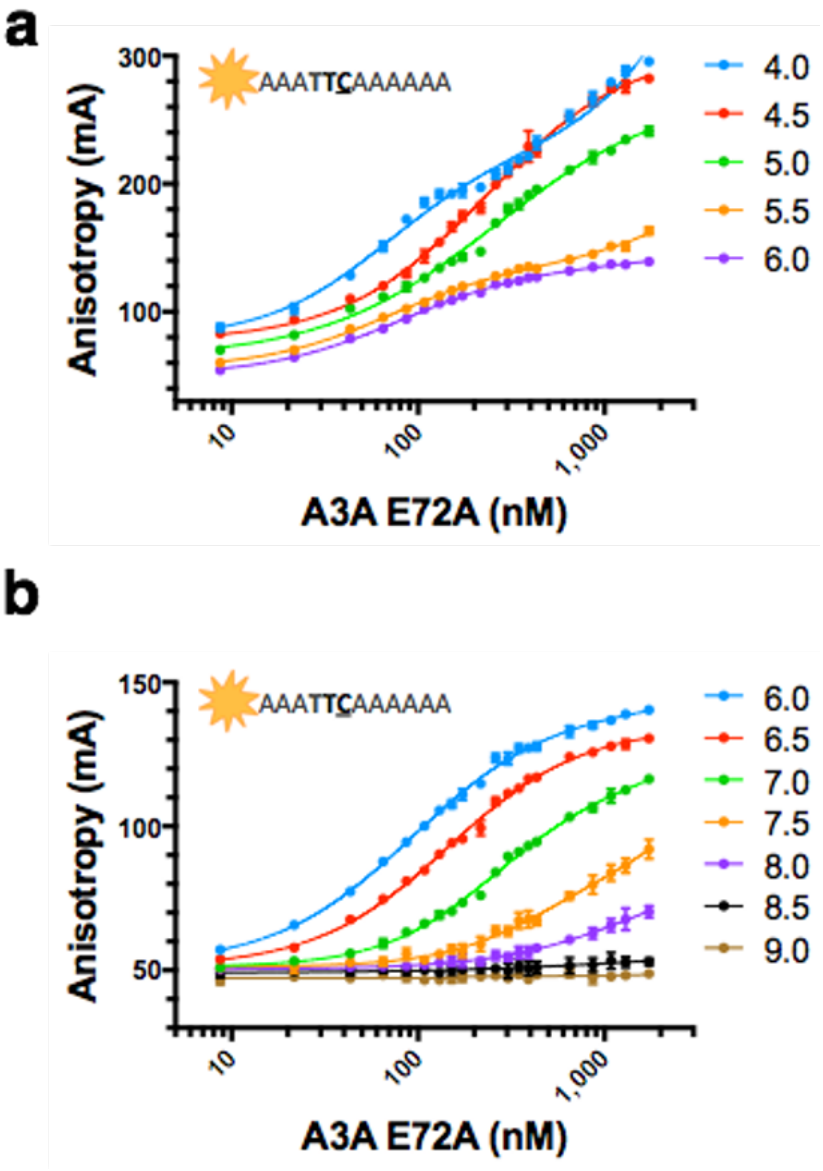

Figure 4.2: A3A affinity to ssDNA at different pHs.
Fluorescence anisotropy of TAMRA-labeled ssDNA 4A-TTC-6A binding to A3A(E72A). A) Binding of A3A to ssDNA at pH 6.0 (blue), 6.5 (red), 7.0 (green), 7.5 (orange), 8.0 (purple), 8.5 (black), 9.0 (brown). B) Binding of A3A to ssDNA at pH 4.0 (blue), 4.5 (red), 5.0 (green), 5.5 (orange), and 6.0 (purple).

Table 4.2: A3A affinity for ssDNA Poly A -TTC in a range of pHs.

pH	KD
4.0	Could not calculate
4.5	Could not calculate
5.0	Could not calculate
5.5	68 ± 3
6.0	90 ± 1
6.5	146 ± 8
7.0	260 ± 21
7.5	> 4,500
8.0	> 5,000
8.5	no detectable binding
9.0	no detectable binding

4.3.3 Substrate recognition is dependent on thymidine directly upstream of target deoxycytidine, with preference for pyrimidines over purines

To study the effect of the nucleotide identity at position -1 relative to target deoxycytidine (NC) on A3A affinity for substrate (**Figure 4.3A**), the Kd values of A3A for (4A)-TC-(6A), AC, CC, GC in a Poly A background were determined. A preference for TC (143 ± 4 nM), followed by CC (250 ± 14 nM) was identified. Interestingly, AC and GC had similarly very weak binding affinities for A3A (>5,000 and >6,500 nM respectively), validating a preference for pyrimidines (T or C) over purines (A or G) at -1 position with T as the strongest binder.

The effects of the sequence identity around the cognate dinucleotide deamination motif (TC) on affinity of A3A for ssDNA was determined by first testing the change in affinity for all nucleotide substitutions at -2 position (3A)-NTC-(6A). A3A has a preference for pyrimidine over purine at -2 position (**Figure 4.3B**) with TTC and CTC having similar affinities (90 ± 1 nM and 85 ± 1 nM respectively) compared to that of purines ATC and GTC (145 ± 2 nM and 150 ± 3 nM respectively). While not as strong as for -1 position, there is a preference for the smaller pyrimidines at position -2. Next, the effect of +1 position on affinity of A3A to TC was determined (**Figure 4.3C**). A3A did not demonstrate a strong preference for any particular nucleotide, although disfavoring T, at the +1 position (145 ± 2 nM for background versus 209 ± 5 nM).

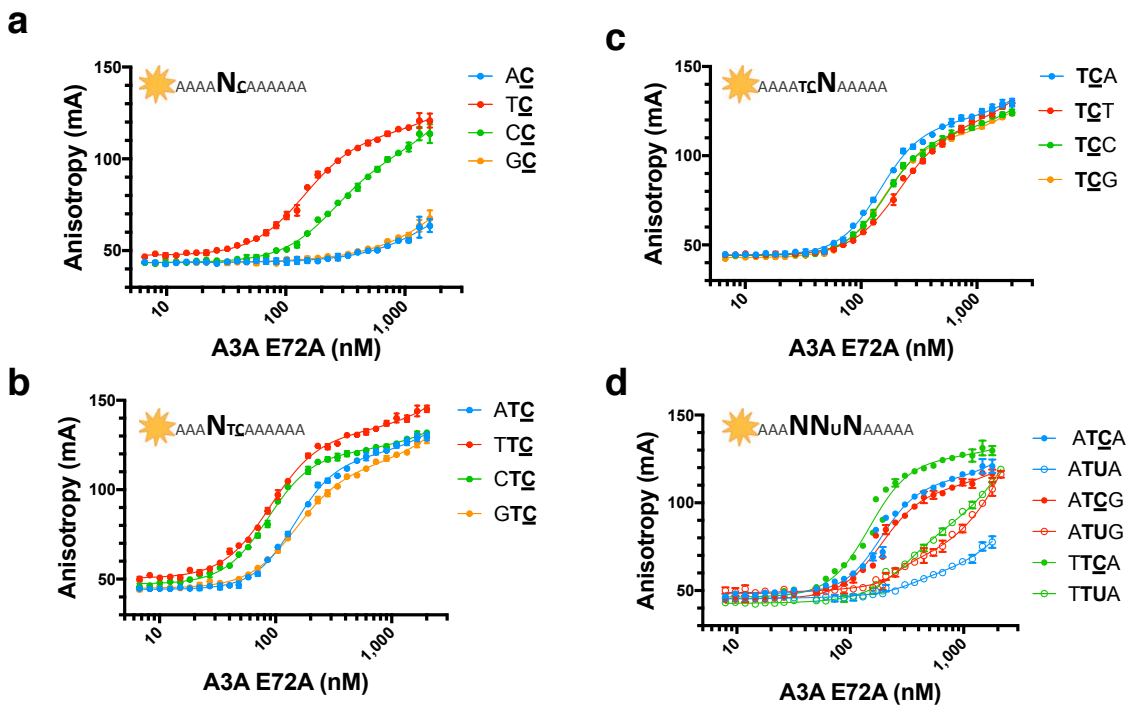

Figure 4.3: A3A specificity for nucleotides flanking substrate cytidine. Fluorescence anisotropy of TAMRA-labeled ssDNA sequences to A3A(E72A). A) Binding of A3A to ssDNA with changes at -1 position of substrate C and TU (purple) in a poly A background (12 mers): 4A-AC-6A (blue), 4A-TC-6A (red), 4A-CC-6A (green), and 4A-GC-6A (orange). B) Binding of A3A to ssDNA with changes at -2 position in a TC context in a Poly A background (12 mers): 4A-ATC-6A (blue), 4A-TTC-6A (red), 4A-CTC-6A (green), and 4A-GTC-6A (orange). C) Binding of A3A to ssDNA with changes at +1 position in a TC context in a Poly A background (12 mers): 4A-TCA-6A (blue), 4A-TCT-6A (red), 4A-TCC-6A (green), and 4A-TCG-6A (orange). D)Three substrate sequences, TTCA (green), ATCG (red) and ATCA (blue), in closed circles with the corresponding 3 product sequences TTUA, ATUG and ATUA in open circles.

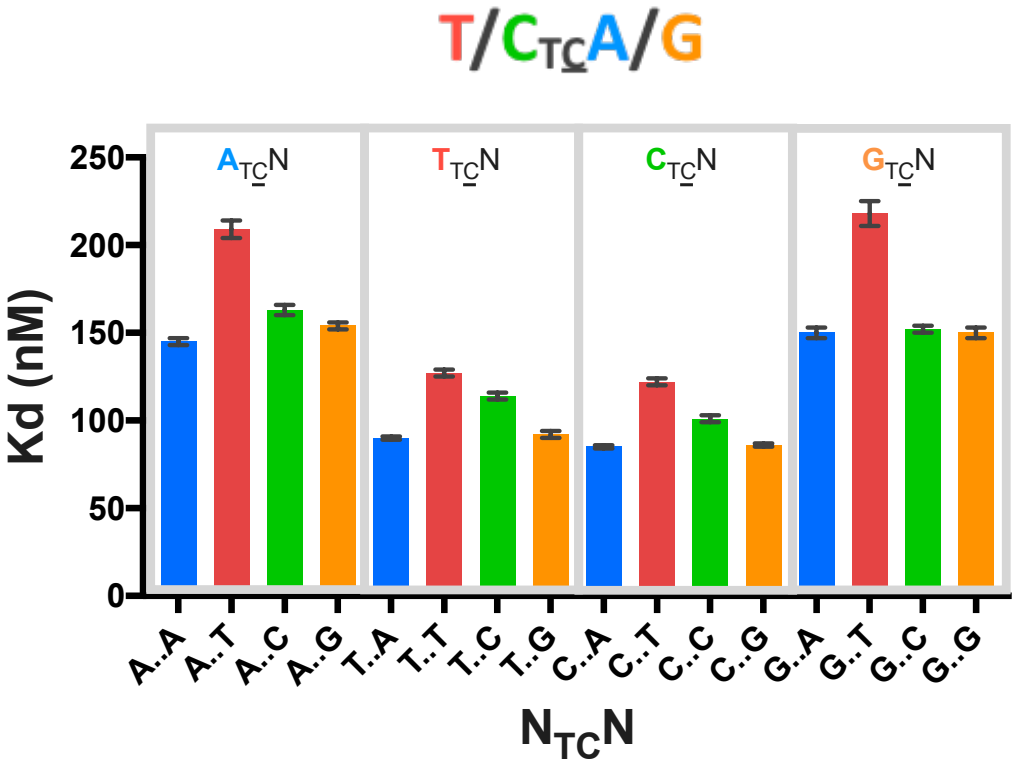

Figure 4.4: A3A specificity for poly A xTCx.
Binding affinity of A3A(E72A) to TAMRA-labeled ssDNA sequences in a Poly A background. Gray boxes bin sequences by -2 nucleotide identity. Colors represent +1 nucleotide identity: A (blue), T (red), C (green), G (orange). Consensus sequence derived from these Kd values is shown above the graph.

Finally, to identify if there was any interdependency between nucleotide identity at -2 and +1 positions, the affinity of A3A for (3A)-NTCN-(5A) was determined (**Figure 4.4, Table 4.1**). A3A displayed preference for pyrimidines at -2 position regardless of the nucleotide at +1. A3A also disfavored T at +1 position regardless of the nucleotide identity at -2. Most interestingly, A3A preferred a pyrimidine at -2 when there was a purine at +1 position. However, the reverse was not true; purine at -2 position with pyrimidine at +1 position did not result in comparable affinities. In fact, the worst binders (ATCT and GTCT) were those that contained purines at -2 with pyrimidines at +1 position. Thus, we have broadly have three classes of substrate binders high affinity (80-130 nM), medium affinity (150-165nM), and weak affinity (210-220 nM) and have identified (T/C)TC(A/G) as the preferred sequence for ssDNA recognition by A3A.

4.3.4 A3A preference for binding to substrate over product is context dependent

A3A's affinity for substrate C was compared to product U in the context of variations of the signature A3A substrate sequence (T/C)TC(A/G). The affinity of three substrate sequences, TTCA, ATCG and ATCA, were compared to the corresponding product sequences (**Figure 4.3D)**. For all three sequences, a substantial loss of binding affinity was observed for the corresponding TTUA, ATUG and ATUA, with the most substantial loss with ATUA. Thus, the decrease in affinity for product over substrate was context dependent.

4.3.5 Positive correlation between sequence preference of binding and enzymatic activity

Although enzymatic activity and binding affinity are not expected to be directly correlated, the trends for specificity would likely be similar. Thus A3A's deamination

activity was determined in the context of variations of the signature sequence (T/C)TC(A/G) using a ^{1}H NMR based A3 deaminase activity assay. High (TTCA and TTCG), medium (ATCA, ATCG, GTCA, GTCG, TTCT) and low (ATCT and GTCT) affinity sequences were tested (**Table 4.3**) to determine the correlation between binding and activity. Overall, activity by NMR has the same trend as affinity from the binding assay (**Figure 4.5)**. This indicates that in general those substrates sequences with varying binding affinity (high, medium and weak) are also processed in a similar order.

Table 4.3: A3A enzyme activity for DNA sequences.

	Activity (min-1) 40ºC
ATCA	27 ± 1
ATCG	28 ± 1
ATCT	30 ± 2
GTCA	38 ± 2
GTCG	31 ± 2
GTCT	22 ± 2
TTCA	52 ± 2
TTCG	36 ± 1
TTCT	37 ± 2

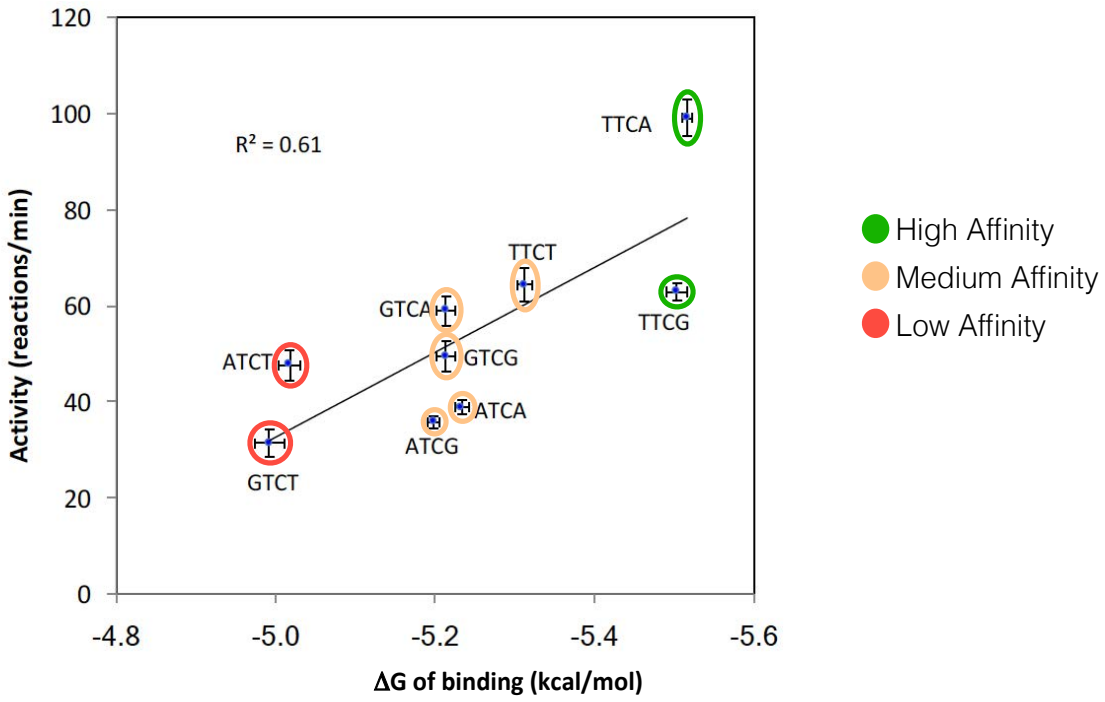

Figure 4.5: Binding affinity versus enzyme activity.
The enzyme activity of active A3A measured by NMR based deamination assay versus the free energy of binding calculated (ΔG =-RTln (Kd) from the binding affinity for nine 12-mers. These nine represent, 2 high binding (green), 5 medium binding (orange) and 2 weak binding (red) sequences.

4.3.6 Structural basis for A3A specificity for binding to preferred recognition sequence

To determine the structural basis for the A3A consensus sequence (T/C)TC(A/G), crystal structures of A3A bound to ssDNA recently determined by our group and others (PDB ID: 5KEG and 5SWW) were analyzed [104, 106]. The target deoxycytidine is well coordinated and buried within the active site of A3A (**Figure 4.6A**) in these structures. The thymidine at position -1 has extensive contacts with loop 7 (Y130, D131 and Y132), and van der Waals contacts with loop 5 (W98) (**Figure 4.6B**). The Watson-Crick edge of the thymidine base faces the loop 7 residues, and makes three hydrogen bonds: one with the backbone nitrogen of Y132 and the other two, one is water mediated, are with the D131 sidechain. The D131 side chain further forms a salt bridge to the R189, which stabilizes the overall hydrogen-bonding configuration of loop 7 to the thymine base. This coordination appears critical, as residue 189 is conserved as a basic residue (Arg/Lys) in catalytically active A3 domains. This coordination also explains why -1 must be the thymidine base. If the -1 position is modeled as a cytidine the N3 atom lacks the proton to hydrogen bond with D131 (**Figure 4.6C**) and wouldn't be as well coordinated thus would be less preferential. Residues Y130 and D131, in loop 7, physically would preclude a larger purine base from fitting in this position (as modelled **Figure 4.6D**). Thus, the T specificity at the -1 position is consistent with the crystal structures.

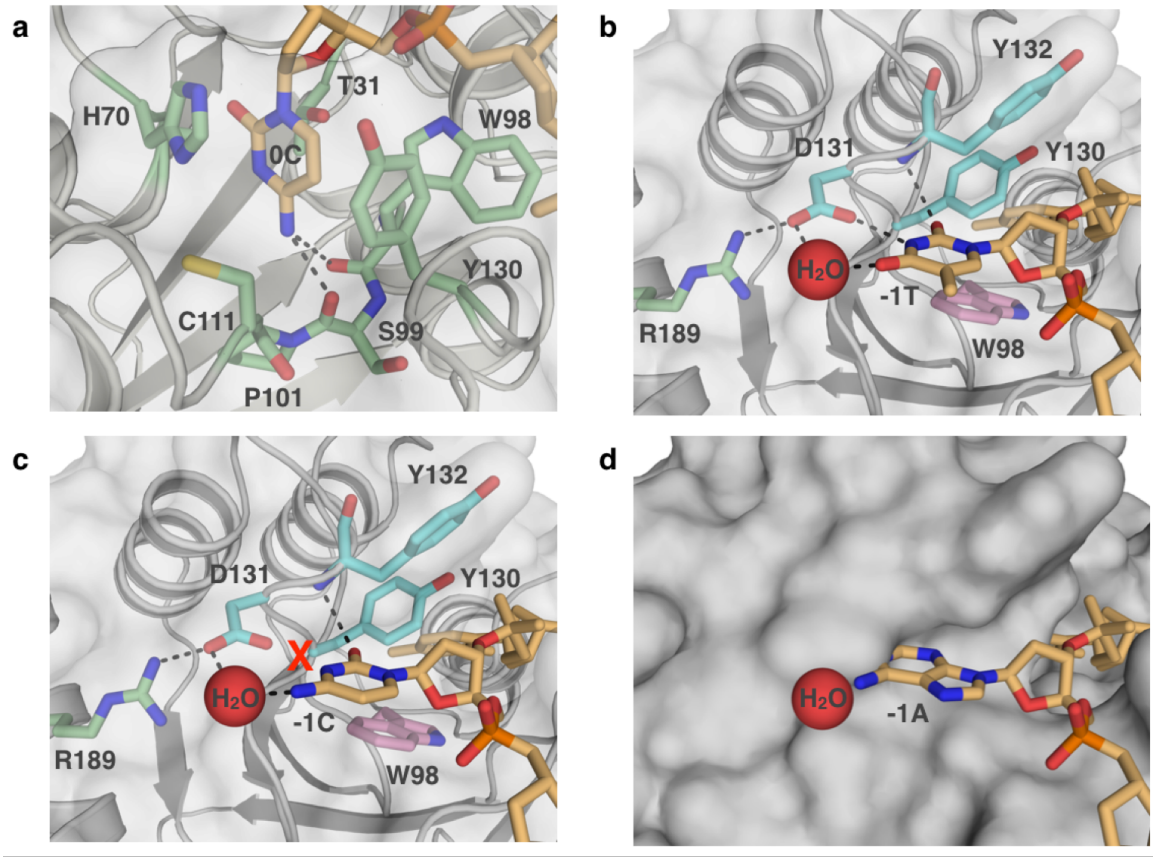

Figure 4.6: A3A recognition of substrate cytidine and pyrimidines at -1.
Crystal structure of A3A(E72A/C171A) shown in surface view (gray) bound to Poly T-1C ssDNA sequence represented as sticks (PDB ID: 5KEG). A) Substrate cytidine (orange sticks) is buried in active site of A3A. Residues interacting with cytidine are shown in green sticks. B) -1 nucleotide thymidine (orange sticks) surrounded by Y130, D131 and Y132 of loop 7 (light blue sticks), W98 of loop 5 (pink sticks), and R189 (green sticks). C) Cytidine modeled into -1 position (orange sticks). N3 atom lacks proton to hydrogen bond with D131 indicate with a red X. D) Adenosine modeled into -1 position (orange sticks) shows severe van der Waal clashes if occupying the same site as the pyrimidines. Other nucleotides are shown as orange sticks. Hydrogen bond and a salt bridges shown in dashes black lines. Water shown as red spheres. Nitrogen and oxygen of residues and nucleic acids are in blue and red respectively.

Although A3A has prefers (T/C)TC(A/G), neither of the co-crystal structures has the optimal nucleotide identity at the -2 and +1 positions [104, 106]. Specificity for purine at the -2 position was not evident in the available A3A–ssDNA structures, presumably as neither structure contains an optimal ssDNA sequence. For instance, even though the 5KEG structure contains a preferred pyrimidine in the -2 position, the thymidine is disordered in this complex. However, in both structures [104, 106], the base at +1 (pyrimidine T in 5KEG and a purine G in 5SWW) stacks with the critical histidine 29 (**Figure 4.7A,B**) [104, 106]. This type of histidine π-π stacking can occur with either a purine or a pyrimidine**.** However, protonated histidine prefers to stack with a purine base over pyrimidine, with thymidine stacking being the least preferred [184] at pH 6. Thus the base stacking potential with protonated histidine 29 provides strong rationale for the specificity for purines and the disfavoring of thymidine at the +1 position relative to substrate deoxycytidine observed in our biochemical assays (**Figure 4.4**).

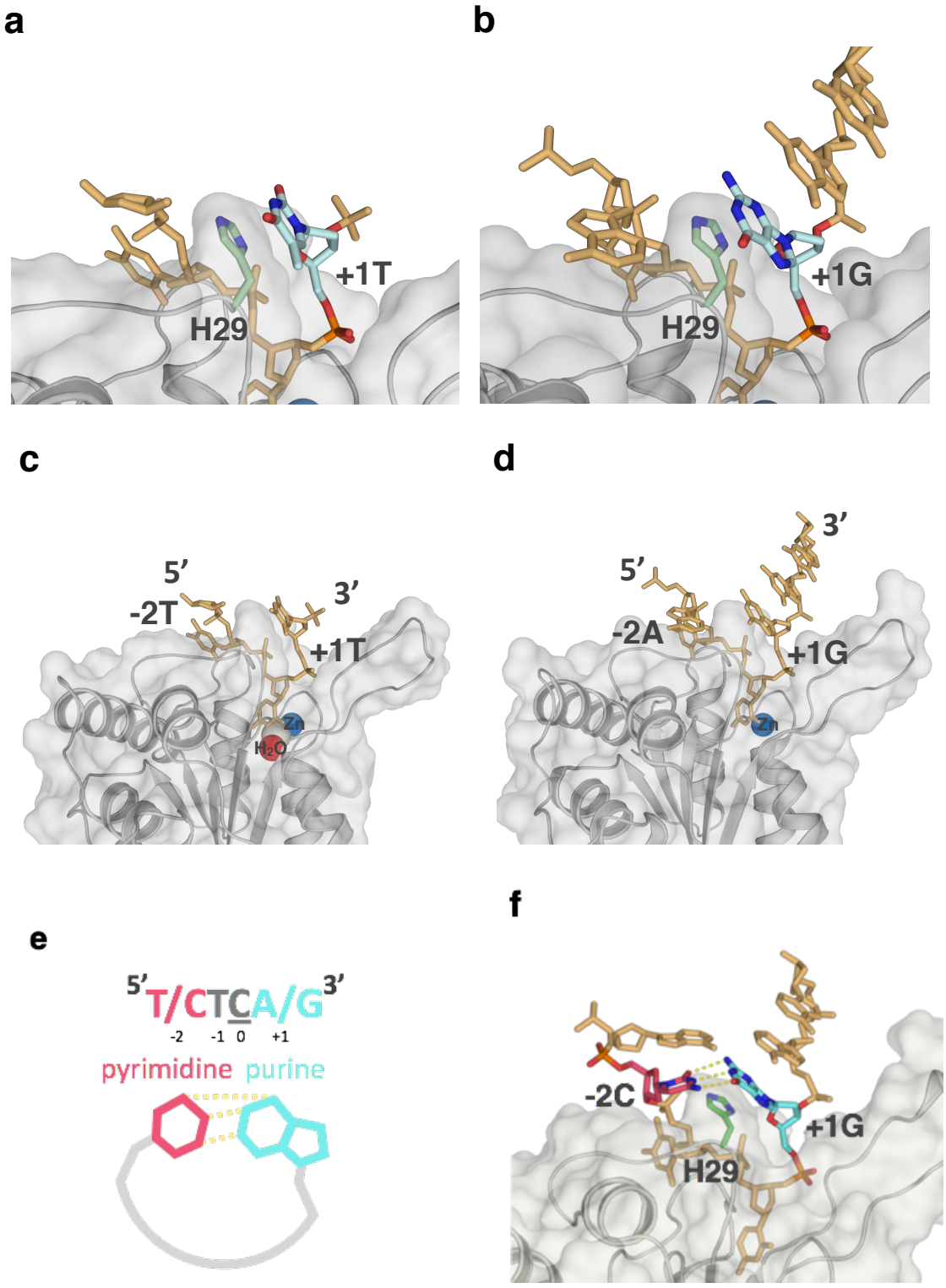
a
+1T
H29
b
+1G
H29
c
5’
-2T
3’
+1T
Zn
H2O
d
3’
5’
-2A
+1G
Zn
e
5’T/CTCA/G3’
-2 -1 0 +1
pyrimidine purine
f
-2C
+1G
H29

Figure 4.7: ssDNA is bent within the complex with A3A.
Crystal structure of A3A shown in surface and cartoon representation (gray) bound to ssDNA displayed as orange sticks; A) +1 thymidine (light blue) is interacting with His 29 (light green sticks) through aromatic stacking (PDB ID: 5KEG). B) +1 guanine (light blue) also interacting with His 29 through aromatic stacking (light green sticks) (PDB ID: 5SWW). C) A3A(E72A/C171A) with TTTTTTTTCTTTTTT (PDB ID: 5KEG) D) A3A(E72A) with AAAAAAATCGGGAAA (PDB ID: 5SWW). Other nucleotides are shown as orange sticks, while water (red), zinc (blue), and chloride (gray) in the active site are shown as spheres. Nitrogen and oxygen of residues and nucleic acids are in blue and red respectively. E) A schematic of hydrogen bonding between pyrimidine (pink) at -2 and purine (light blue) at +1 position via bending of the DNA by A3A upon binding. F) Model of inter-DNA base interactions through binding of A3A to ssDNA. A3A(E72A)–ssDNA complex (PDB ID: 5SWW) was used to model A3A signature sequence CTCG bound at the active site. A3A is shown as gray surface and cartoon, His29 as light green sticks, original ssDNA as orange sticks with +1G in light blue. Adenosine at -1 position was switched to cytosine (pink) with hydrogen bonds to +1G displayed as yellow dashes.

4.3.7 A3A bends ssDNA to potentially allow for intra-DNA interaction between -2 and +1 nucleotides

A common feature between the two A3A–ssDNA complex structures is that the ssDNA forms a “U” shape in the active site (**Figure 4.7 C,D**) [104, 106]. This U shape of the bound polynucleotide may be conserved among deaminases, including adenosine deaminases [106, 185]. In both A3A-ssDNA structures, the U shape of the ssDNA orients the -2 and +1 bases in close proximity to each other. Thus, we hypothesized that the observed sequence preference (**Figure 4.4**) for the -2 position is a result of intra-DNA interactions rather than specific interactions with the protein.

To determine the potential for intra-DNA interactions when A3A is bound to a (T/C)TC(A/G) signature sequence, molecular models were developed based on the crystal structures of A3A bound to ssDNA (PDB ID: 5KEG and 5SWW) [104, 106]. These models orient the bases of the -2 and +1 nucleotides so that they form hydrogen bonds, with the larger purine at +1 position stacking on His 29 and the smaller -2 pyrimidine coordinating the +1 base (**Figure 4.7 E,F**). The reversal of the nucleotides at +1 and -2 positions would not result in a fit nearly as well, which could explain the lower affinity of purine-TC-pyrimidine. Thus the structural model explains the preference for (T/C)TC(A/G) and suggests stabilizing the inter-DNA interactions may further increase the affinity.

4.3.8 Length of ssDNA affects affinity of A3A for substrate sequence

If the bending of the ssDNA is important for substrate recognition, dependence of binding affinity on substrate length may be expected. To determine if the DNA beyond the four-nucleotide signature sequence contributed to the binding, the length of the

ssDNA that contained the recognition sequence was varied in Poly A-TTC (AAA TTCA AAA AAA). A competition assay with different length oligonucleotides was performed to test the effect of ssDNA length on affinity for substrate (**Figure 4.8**). Length was varied from 1 nucleotide flanking each end of TTCA (TTCAA and ATTCA) to 3 nucleotides flanking each end, increasing by one nucleotide addition on either end. Surprisingly, a single nucleotide flanking TTCA signature sequence was not enough to permit binding (**Figure 4.8A**), and even three nucleotides on either side still did not bring A3A binding to original binding affinity as Poly A-TTC (AAA TTCA AAA AAA) (**Figure 4.8B**). Thus, binding affinity is impacted beyond the recognition motif to prefer longer sequences, although the additional nucleotides not expected to have any direct contacts with A3A, consistent with the model that intra-DNA interactions modulate A3A affinity.

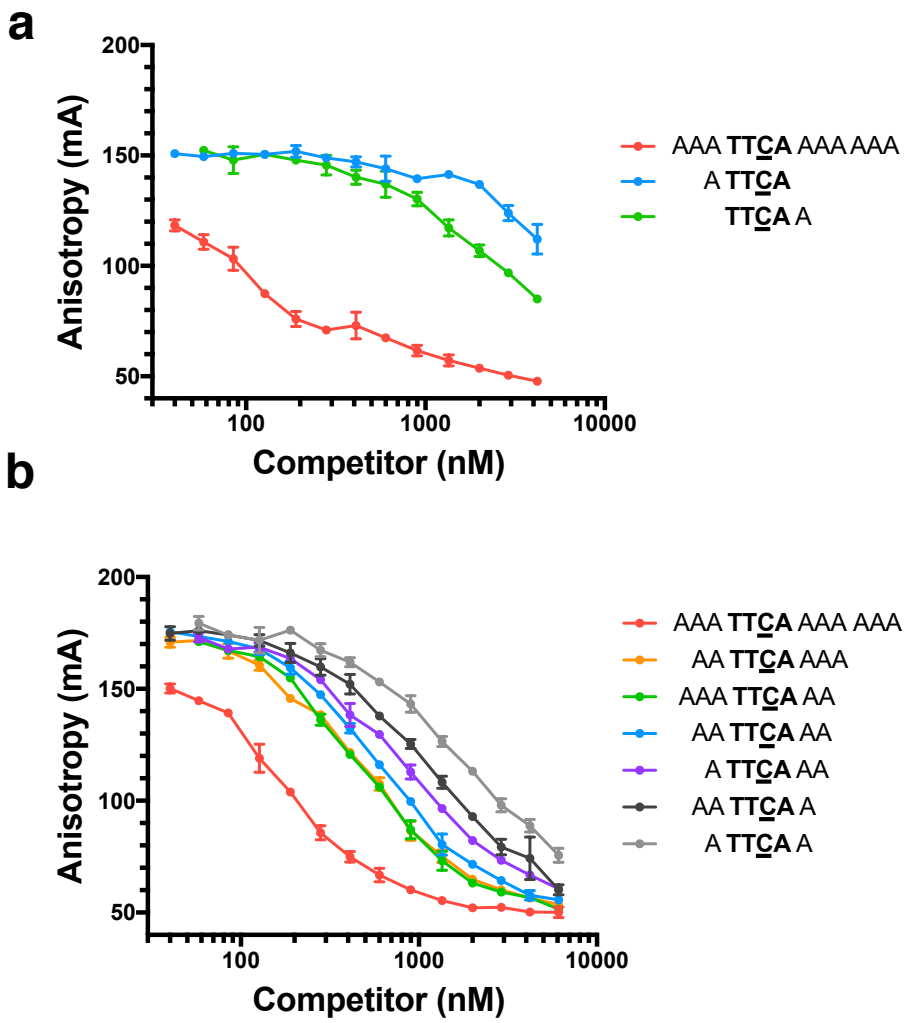

Figure 4.8: A3A affinity to ssDNA of varied lengths.
Fluorescence anisotropy of TAMRA-labeled ssDNA 3A-TTCA-6A to A3A(E72A) competing with unlabeled ssDNA of different lengths. A) Binding of A3A to labeled ssDNA preincubated with unlabeled 3A-TTCA-6A (red), 1A-TTCA (blue), and TTCA-1A (green). B) Binding of A3A to labeled ssDNA preincubated with unlabeled 3A-TTCA-6A (red), 2A-TTCA-3A (blue), 3A-TTCA-2A (green), 2A-TTCA-2A (blue), 1A-TTCA-2A (purple), 2A-TTCA-1A (black), and 1A-TTCA-1A (gray).

4.3.9 A3A prefers binding to target sequence in the loop of structured hairpins

Another implication of this model would be that pre-bent DNA could be a better substrate for A3A binding, as A3A would not have to pay the entropic cost of bending the DNA. This bending of DNA could be achieved either by the inter-DNA interactions modeled in (**Figure 4.7F)**, or when within a loop of a hairpin. To determine the significance of the bent U shape DNA structure in the mechanism of A3 binding, we tested A3A affinity to a target deoxycytidine in the loop region of a DNA hairpin. The hairpin sequence was based on a previously identified potential RNA substrate for A3A, from succinate dehydrogenase complex iron sulfur subunit B (SDHB)[162]. The affinity for TTC in the loop region of hairpin DNA was higher than that in linear DNA (26 nM vs 90–127 nM respectively). As expected, A3A had a higher affinity for the DNA hairpin with loop region containing TTC compared to one with AAA (26 nM vs ~676 nM respectively) (**Figure 4.9A**). Interestingly, the Kd value for the hairpin (26 nM) is comparable to that for a single C in a polyT background (35 nM) [84]. This may imply that the polyT DNA adopts a hairpin structure in solution, as has been reported [186].

A3A affinity to a target cytidine in the loop region of an RNA hairpin was also tested. The exact SDHB hairpin RNA sequence including UC in the loop of this hairpin versus a modified SDHB hairpin RNA replacing the AUC with AAA was compared. A3A had specific affinity for the hairpin RNA containing UC compared to AA (37 nM vs 202 nM respectively) (**Figure 4.9B**). In contrast to what has been previously proposed [94], we found that A3A has high affinity and specificity for RNA. Furthermore, A3A has a higher affinity for AUC in the loop region of a hairpin compared to UUC in a linear sequence (**Figure 4.10**). The potential UUC substrate sequence in linear RNA has no measurable

affinity, comparable to linear RNA without a potential substrate sequence. Overall, A3A has higher affinity for target sequence in the context of a pre-ordered loop region rather than linear DNA, and specific affinity for RNA hairpins with a substrate site.

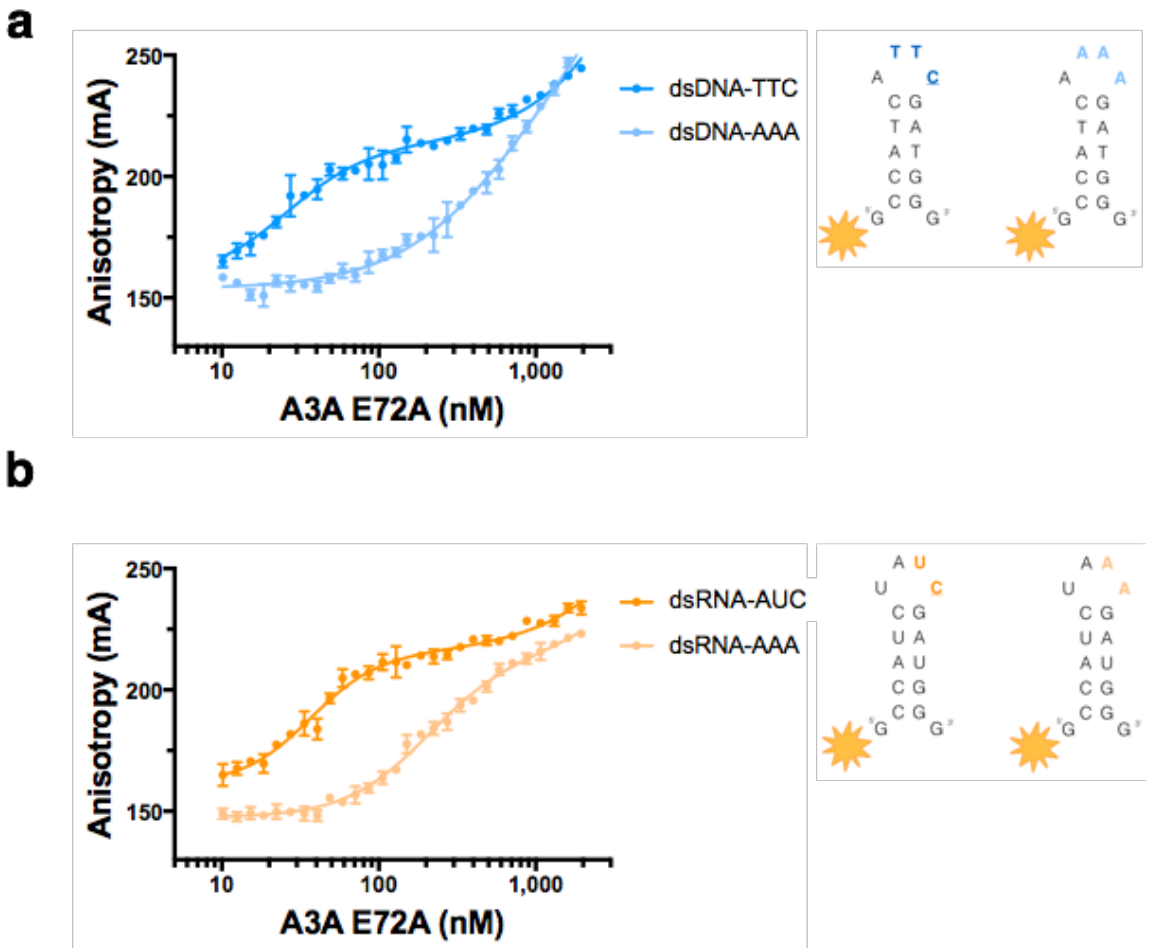

Figure 4.9: A3A specificity for substrate in loop region of stem-loop nucleic acids. Fluorescence anisotropy of TAMRA-labeled hairpin DNA and RNA to A3A(E72A). A) Binding of A3A to a DNA version of the hairpin SDHB RNA containing TTC (dark blue) and AAA (light blue) in the loop region. B) Binding of A3A to hairpin SDHB RNA (dark orange) and the same RNA sequence replacing the UC with AA in the loop region of the hairpin (light orange).

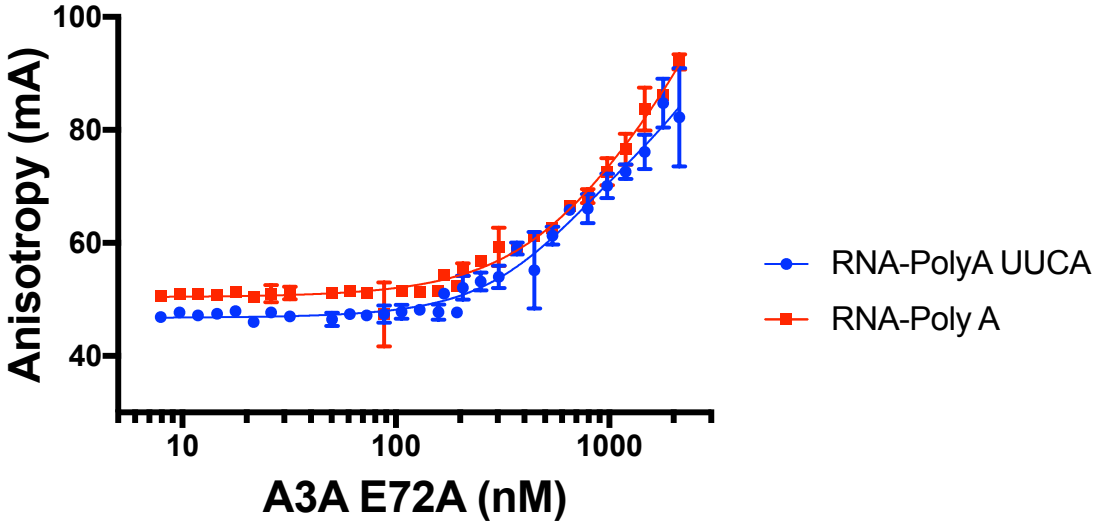

Figure 4.10: A3A affinity to ssRNA.
Fluorescence anisotropy of TAMRA-labeled ssRNA sequences to A3A(E72A). Binding of A3A to ssRNA with PolyA UUCA (blue) and ssRNA Poly A (red).

4.4 DISCUSSION

A3A is a single-domain enzyme with the highest catalytic activity among the human APOBEC3 proteins [178], a known restriction factor [179, 180], and also likely contributes to carcinogenesis [181]. In this study we quantified the ssDNA specificity of A3A, and identified the consensus signature sequence as (T/C)TC(A/G). The dinucleotide sequence preference for A3A, TC, which was previously found through activity assays [98, 106, 173] was confirmed and expanded to a preference for pyrimidine-TC-purine. Surprisingly context matters, in that the background nucleotide sequence impacts binding affinity, with essentially no binding observed for Poly A 1C **(Figure 4.1B**), while Poly T 1C binds with 35 ± 2 nM affinity [84]. Furthermore, the length of the ssDNA in which (T/C)TC(A/G) is imbedded within also modulates affinity (**Figure 4.8**). Structural analysis of the two A3A-ssDNA complexes containing two distinct, but suboptimal ssDNA sequences have led us to develop a model with intra-DNA interactions for the molecular mechanism for A3A's specificity to ssDNA. In contrast to previous results [93], which implicate the -2 position as defining specificity, the base at this position observed in both A3A–ssDNA co-crystal structures do not make any specific interactions with the protein. Rather, the hydrogen bonding edge of the -2 base is in close proximity to corresponding edge of +1 base, suggesting possible intra-DNA interactions as being determinants of preference. Our molecular modeling confirmed such interactions could stabilize the U-shaped DNA conformation within the A3A active site, explaining the -2 position specificity.

We found that A3A binds to RNA in a highly specific and structural context-dependent manner. Previous reports [94] suggested that A3A bound only weakly and did

not deaminate RNA. However, the potential substrate sequence was designed to lack secondary structure, which in light of our results on hairpin versus linear RNAs, may have inadvertently precluded RNA deamination. Recently, A3G and A3A were implicated in deaminating RNA in proposed RNA hairpins in whole cell lysates but the specificity was not quantified [162, 187]. Intriguingly, our data show that A3A binds RNA hairpins with similar affinity as for DNA hairpins, which suggests that RNA-editing activity of A3A might be more prevalent than previously anticipated. Future experiments will identify if A3A's catalytic efficiency is similar for DNA and RNA hairpins.

The comprehensive identification of A3A signature sequences and preference for loop structures will enable a more accurate evaluation of A3 activity based on sequence analysis. Previous studies used only a *single* identified A3 signature sequence to implicate A3's role in viral restriction or cancer progression. In contrast, our study suggests a more accurate method for determining evidence of A3 activity would be to use a set of sequences. In the case of A3A, we have identified four almost equivalent substrate signature sequences, TTCA, TTCG, CTCA, and CTCG, which should be used for identifying A3A's involvement in mutagenesis. We also found a positive correlation between A3A's sequence preference of binding and enzymatic activity. Correlation not only legitimizes the use of a DNA binding assay with inactive enzyme as a reliable method for studying specificity of A3s, it also shows that affinity for substrate is a driving factor for catalysis. Thus, factors that could enhance or perturb binding, such as pH or nucleic acid structure, would result in modulation of deamination activity.

In addition to using the full A3A signature sequences, the probability of mutagenesis should not be solely based on nucleotide sequence, but should also be

weighted by the propensity of the target sequence to be within a structured loop. Secondary structure prediction software could be used to identify the consensus sequence in loop regions of structured DNA or RNA. A3A signature sequences, (T/C)TC(A/G), that we identified, not only accounts for the discrepancies in the A3A target sequences reported in the literature such as TTCA versus CTCG [98] [106], but also leads us to advocate a new paradigm for identifying A3A's involvement in mutation of endogenous or exogenous DNA.

Designing inhibitors or activators for A3s has been extremely challenging. Our results implicate a need to incorporate the structural context of the target deoxycytidine in the therapeutic design. Larger "U" shaped macrocycles may serve as more appropriate starting scaffolds in designing cancer therapies targeting A3s, which would mimic the "U" shape of the bound ssDNA. Macrocycles have recently been shown to have good drug-like properties and may be a strategy to target these critical enzymes[188].

4.5 METHODS

4.5.1 Cloning of APOBEC3A E72A overexpression construct

The pColdII His-6-SUMO-A3A(E72A) was constructed by first cloning the SUMO gene from pOPINS His-6-SUMO into pColdII His-6 vector (Takara Biosciences) using NdeI and KpnI restriction sites. Human APOBEC3A coding sequence from pColdIII GST-A3A(E72A, C171A) was then cloned into the pColdII His-6-SUMO vector with KpnI and HindIII. The C171A mutation in the A3A construct was reverted to wild type residue by site directed mutagenesis resulting in the pColdII His-6-SUMO-APOBEC3A(E72A) catalytically inactive over-expression construct used for all experiments in this study.

4.5.2 Expression and purification of APOBEC3A E72A

Escherichia coli BL21 DE3 Star (Stratagene) cells were transformed with the pColdII His-6-SUMO-APOBEC3A(E72A) vector described above. The E72A mutation was chosen to render the protein inactive. Expression occurred at 16 °C for 22 hours in lysogeny broth medium containing 0.5 mM IPTG and 100 µg/mL ampicillin. Cells were pelleted, re-suspended in purification buffer (50 mM Tris-HCl [pH 7.4], 300 mM NaCl, 1 mM DTT) and lysed with a cell disruptor. Cellular debris was separated by centrifugation (45,000 g, 30 min, 4C). The fusion protein was separated using HisPur Ni-NTA resin (Thermo Scientific). The His6-SUMO tag was removed by means of a Ulp1 protease digest overnight at 4 °C. Untagged A3A(E72A) was separated from tag and Ulp1 protease using HisPur Ni-NTA resin. Size-exclusion chromatography using a HiLoad 16/60 Superdex 75 column (GE Healthcare) was used as a final purification step. Purified recombinant A3A was determined to be free of nucleic acid prior to binding experiments by checking OD 260/280 ratios, which was at 0.54.

4.5.3 Oligo source and preparation

Labeled and unlabeled oligonucleotides used in this assay were obtained through Integrated DNA Technologies (IDT). Labeled oligonucleotides used in the fluorescence anisotropy based binding assay contain a 50-TAMRA flourophore at their 5’ end and were re-suspended in ultra-pure water at a concentration of 20 µM. Unlabeled oligonucleotides used for the competition assays were resuspended in ultra-pure water to a concentration of 4 mM.

4.5.4 Fluorescence anisotropy-based DNA binding assay

Fluorescence anisotropy-based DNA binding assay was performed as described [84] with minor alterations. A fixed concentration of 10 nM 50-TAMRA-labeled oligonucleotides was added to A3A-E72A in 50 mM MES buffer (pH 6.0), 100 mM NaCl, 0.5 mM TCEP in a total reaction volume of 150 mL per well in nonbinding 96-well plates (Greiner). For the fluorescence anisotropy-based DNA binding assay with APOBEC3B-CTD E255A was performed in 50 mM Tris buffer (pH 7.4), 100 mM NaCl, 0.5 mM TCEP. The concentration of APOBEC3 was varied in triplicate wells. Plates were incubated for overnight at room temperature.

For the pH dependence experiments the buffer reagent used for testing was pH 4.0–5.0 sodium acetate, pH 5.5-6.5 MES, pH 7.0-8.0 HEPES, pH 8.5-9.0 TRIS. Assay was performed as described above. For the competition assays, a fixed concentration of 300 nM A3A(E72A) was used and unlabeled oligonucleotide of varied concentration was added from 0–6.1uM. A3A(E72A) was pre-incubated with unlabeled oligonucleotide for an hour in assay buffer, then labeled DNA was added and incubated overnight at room temperature.

For all experiments, fluorescence anisotropy was measured using an EnVision plate reader (PerkinElmer), exciting at 531 nm and detecting polarized emission at 579 nm wavelength. For analyzing data and determining Kd values, Prism (GraphPad) was used for least-square fitting of the measured fluorescence anisotropy values (Y) at different protein concentrations (X) with a single-site binding curve with Hill slope, a nonspecific linear term, and a constant background using the equation Y=(Bmax * X^h)/(Kd^h + X^h) +NS*X + Background, where Kd is the equilibrium dissociation constant, h

is the Hill coefficient, and Bmax is the extrapolated maximum anisotropy at complete binding.

4.5.5 ^{1}H NMR-based A3 deaminase activity assay

Deaminase activity was determined for A3A protein by assaying active enzyme against linear DNA substrates and measuring the product formation using ^{1}H NMR. Active A3A protein (50 nM) was assayed against linear DNA substrates (200 μM) in buffer with 50 mM MES pH 6.0, 100 mM NaCl, 0.5 mM TCEP, and 5% D_2O. Experiments were performed on 9-mer substrates containing the target sequences AA(A/G/T)TC(A/G/T)AAA and at 40°C to prevent the DNA from oligomerizing due to high concentration. Experiments were performed using a Bruker Avance III NMR spectrometer operating at a ^{1}H Larmor frequency of 600 MHz and equipped with a cryogenic probe. Product concentration was estimated from peak integrals with Topspin 3.5 software (Bruker Biospin Corporation, Billerica, MA) using an external standard. Activity was determined from the initial rate of product formation via first-order exponential fitting of the progress curve. Rate errors were estimated by Monte Carlo simulation using 100 synthetic data sets and taking the residuals of the initial fit to the experimental data as the concentration error.

4.5.6 Molecular Modeling

The crystal structures of A3A bound to ssDNA (PDB ID: 5KEG and 5SWW) were used for molecular modeling [104, 106]. The DNA sequence was first mutated using Coot [189]. The complex structure was then prepared and minimized by ProteinPrep Wizard in Maestro (Schrödinger) at pH6.0 with other settings as default.

4.6 ACKNOWLEDGEMENTS

This work was supported by the US National Institute of Health [R01GM118474, P01 GM091743]; and T.V.S. is supported by US National Institute of Health F31 GM11993. Funding for open access charge: US National Institute of Health. For W.M. and H.M., this project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN26120080001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

5 Chapter V: Mechanism for APOBEC3G catalytic exclusion of RNA and non-substrate DNA

Chapter V is a collaborative study that has been previously published as:

Solomon WC, Myint W, **Hou S**, Kanai T, Tripathi R, Kurt Yilmaz N, Schiffer CA, Matsuo H. "Mechanism for APOBEC3G catalytic exclusion of RNA and non-substrate DNA." *Nucleic Acids Research* 47.14 (2019): 7676-7689.

5.1 ABSTRACT

The potent antiretroviral protein APOBEC3G (A3G) specifically targets and deaminates deoxycytidine nucleotides, generating deoxyuridine, in single stranded DNA (ssDNA) intermediates produced during HIV replication. A non-catalytic domain in A3G binds strongly to RNA, an interaction crucial for recruitment of A3G to the virion; yet, A3G displays no deamination activity for cytidines in viral RNA. Here, we report NMR and molecular dynamics (MD) simulation analysis for interactions between A3Gctd and multiple substrate or non-substrate DNA and RNA, in combination with deamination assays. NMR ssDNA-binding experiments revealed that the interaction with residues in helix1 and loop1 (T201-L220) distinguishes the binding mode of substrate ssDNA from non-substrate. Using 2'-deoxy-2'-fluorine substituted cytidines, we show that a 2'-endo sugar conformation of the target deoxycytidine is favored for substrate binding and deamination. Trajectories of the MD simulation indicate that a ribose 2'-hydroxyl group destabilizes the - stacking of the target cytosine and H257, resulting in dislocation of the target cytosine base from the catalytic position. Interestingly, APOBEC3A, which can deaminate ribocytidines, retains the ribocytidine in the catalytic position throughout the MD simulation. Our results indicate that A3Gctd catalytic selectivity against RNA is dictated by both the sugar conformation and 2'-hydroxyl group.

5.2 INTRODUCTION

Cytidine deaminases perform a variety of functions ranging from diversification of antibodies to defense against viral infection. Four members of the APOBEC3 (A3) family of cytidine deaminases (A3D, A3F, A3G and A3H) have varying degrees of

effectiveness in restricting HIV-1 infection[4, 22, 190-193]. Restrictive A3 proteins are encapsidated during viral replication by associating with viral and cellular RNAs, and transported in the budding virion to the target cell. During the course of viral reverse transcription, a transient singlestranded DNA (ssDNA) intermediate is formed. Restrictive A3 proteins bind to the ssDNA intermediate and deaminate cytosine bases to uracil in preferred polynucleotide contexts (5'-TC for A3D, A3F and A3H, and 5'-CC for A3G)[194]. Upon copying of the ssDNA intermediate to form the dsDNA required for successful integration of the HIV-1 genome into the host DNA, mutated uracils base pair with adenines resulting in G to A hypermutation and loss of coding integrity[170]. Interestingly, even though these restrictive A3 proteins bind tightly to RNA in the cell[195, 196], they do not catalyze cytosine deamination in the context of RNA[197, 198]. The mechanism by which these A3 proteins distinguish between relatively rare single stranded DNAs and the abundant single stranded RNA present in the cellular milieu has been a perplexing question. Without the ability to selectively exclude ribocytidines from deamination, mRNA would acquire lethal amounts of nonsense and missense mutations[e.g., [199]], and without the ability to interact with RNA, A3 proteins would not be able to exert restrictive pressure during HIV infection since encapsidation is essential for deamination of the HIV-1 genome [e.g. [193]]. Sharma and co-workers observed the deamination of RNA by A3G in natural killer cells, lymphoma cell lines and CD8-positive T cells under specific conditions, such as cellular crowding and hypoxia, but not in cells under normal conditions[116]. Since A3G strongly disfavors ribocytidine as a substrate in vitro[197, 198], the physiological function of RNA deamination by A3G remains elusive. Structures of the catalytically active subunits of A3A, A3B, A3C, A3F, A3G and

A3H have been determined in the absence of ssDNA by us[79-81, 83, 85] and others[87, 89, 91-93, 95, 96, 98, 99, 200]. We[104] and another group[106] also determined structures of A3A bound to ssDNA, which provided insights of static interactions between substrate ssDNA and protein at the catalytic site. In the A3A–ssDNA co-crystal structures, the ssDNA exists in a tightly curved conformation with three nucleotides (the target deoxycytidine and flanking nucleotides) forming hydrogen bonds and π–π stacking interactions with A3A, and the sugar of the target deoxycytidine adopting the C2'-endo conformation typically found in DNA[104, 106]. Most recently, we determined the structure of the ssDNA-bound A3G catalytic domain using a variant of A3Gctd (A3G-CTD2) that has strong affinity for ssDNA containing a hotspot sequence, 5'-TCCCA[105]. In comparison to the A3A–ssDNA co-crystal structure, the ssDNA has a more extended conformation and larger contact surface with A3G-CTD2, by interacting with five nucleotides instead of only three. Although this cocrystal structure provided atomic details of static interactions between the hotspot nucleotides and the protein, the mechanism by which A3G strongly disfavored ribocytidine as a substrate was not revealed. In particular, a 2'-OH could fit within the spatial position of the 2'-H without significant steric hindrance[105]. Previously, Nabel et al. reported that the C2'-endo sugar conformation was important for the efficiency of deoxycytidine deamination catalyzed by human activation induced deaminase (AID) and mouse APOBEC1[201]. This finding may or may not be applicable for A3G because AID and mouse APOBEC1 are substantially different from A3G in regard to physiological targets; AID deaminates deoxycytidines in particular 5'-A/T-2A/G-1C0 hotspots of the immunoglobulin genes undergoing transcription[12], whereas APOBEC1 deaminates a specific cytidine in the apolipoprotein B (Apob) pre-mRNA[13,

[202]. Importantly, APOBEC1 but not A3G, requires an additional factor for deamination site selection and activity in cells; an RNA-binding protein, namely the APOBEC1 complementation factor or A1CF[203, 204]. For AID, different studies have found that various proteins interact with AID[205, 206]. In this study, we interrogate the differences in the interaction modes of A3Gctd for substrate or non-substrate ssDNAs, and the exclusion mechanisms for ribocytidine from deamination. We show that the mode of interaction including extent, intensity and time-scale, determined by NMR titration experiments, clearly distinguish the catalytically productive binding mode for substrate ssDNA from the inactive mode for non-substrate. In addition, we reveal the importance of 2'-endo sugar conformation for catalytically productive binding using 2'-deoxy-2'-fluorine substituted cytidines as substrates. Furthermore, molecular dynamics (MD) simulations indicate that 2'-OH causes the target ribocytidine to dislocate from the catalytic position for A3Gctd but not for A3A, consistent with A3A's ability to deaminate ribocytidine.

5.3 RESULTS

5.3.1 Assigning NMR signals of A3Gctd-2K3A-E259A at pH 6.0

Wild-type A3Gctd has weak affinity for ssDNA at neutral pH, making detection difficult of significant NMR chemical shift perturbations upon ssDNA binding[79, 89]. Enzymatic kinetics analysis of A3Gctd at pH 6 suggested that A3Gctd bound ssDNA with a higher affinity[183], but wildtype A3Gctd was not stable enough to conduct lengthy NMR experiments at that pH with high protein concentration. Therefore, we used a variant A3Gctd, termed A3Gctd2K3A that contained five amino acid substitutions

(L234K, C243A, F310K, C321A and C356A) which enhance the solubility and stability of protein, without altering catalytic activity, structure, or HIV-1 restriction[79, 81, 142]. To observe interaction and compare differences between substrate and non-substrate ssDNAs without ongoing catalytic reaction, we produced a catalytically inactive variant of A3Gctd2K3A by introducing a single alanine point mutation at the catalytic glutamate (E259A), termed A3Gctd-2K3AE259A. We completed the assignment of backbone NMR signals of A3Gctd-2K3A-E259A by using standard triple resonance NMR experiments at pH 7.3, then transferred the assignments to the spectrum recorded at pH 6.0 by following peak shifts throughout pH titration from pH 7.3 to pH 6.0. We were able to assign most of the resolved NMR signals in the 15N-HSQC spectrum at pH 6.0 (**Figure 5.1**).

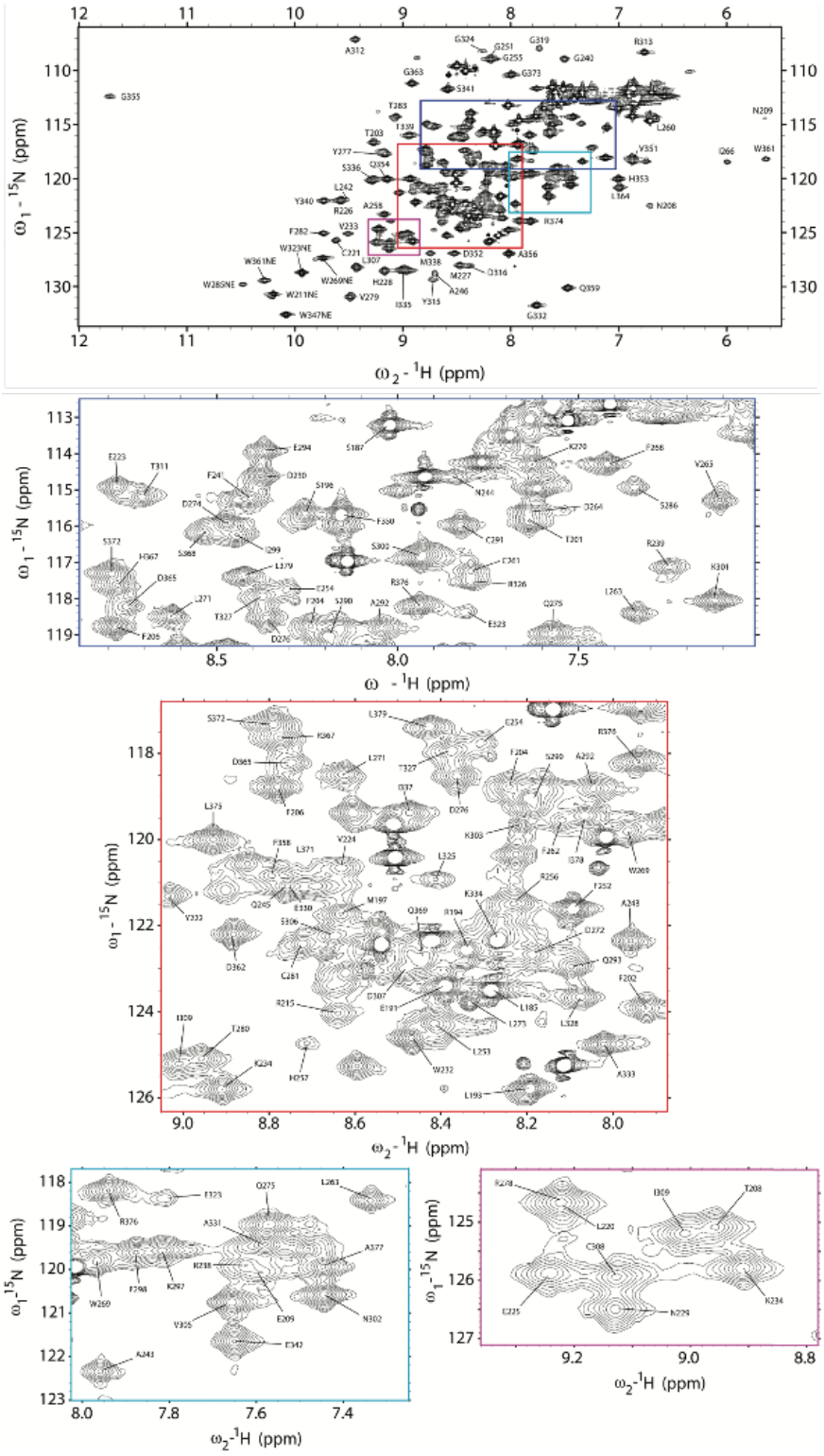

Figure 5.1: NMR signal assignments of A3Gctd-2K3A-E259A at pH 6.0.
^{1}H-^{15}N HSQC spectrum of A3Gctd-2K3A-E259A. Inset locations indicated with colored boxes correspond to expanded inset spectrum borders.

5.3.2 Identification of ssDNA-binding surfaces of A3Gctd

A3Gctd deaminates the 3 C in the 5 -CCC motif 45 times more efficiently than the middle C, and does not deaminate the 5 C in vitro[88, 183, 207, 208]. In order to determine interactions that are responsible for this disparity, we mixed ssDNA to a sample containing catalytically inactive A3Gctd-2K3A-E259A and observed interaction between the protein and ssDNA. We mixed the substrate ssDNA (5'-AATCCCAAA), the intermediate product (5'-AATCCdeoxyUAAA), the final product (5'-AATCdeoxyUdeoxyUAAA), or a ribocytidine substituted ssDNA (5'-AATCCrCAAA) to A3Gctd-2K3AE259A, and compared chemical shift perturbations (CSP) and signal intensity changes of their 15N-HSQC spectra. The 15N-HSQC spectrum of A3Gctd-2K3A-E259A showed substantial perturbations upon adding 5'-AATCCCAAA (**Figure 5.2A**). This data was quantified as described in Methods and plotted as CSP (red with right axis) and signal intensity changes (gray with left axis) in **Figure 5.2C**. Both analyses revealed three primary regions perturbed upon 5'-AATCCCAAA binding. These three regions, binding regions 1, 2 and 3 or BR1, BR2 and BR3, form a continuous surface in the 3D structure of ssDNAfree A3Gctd (PDB ID: 4ROV) (25) (**Figure 5.2D**). BR1 spans residues T201-L220, which includes residues located in helix1 (T201-N207) and loop1 (N208-T218). Especially, W211 and R215, both located in loop1, lost >70% of their signal intensity suggesting direct interactions with DNA. BR2 spans residues R238-K270, and includes β-sheet2, loop3 and helix2. Residues sequentially close to N244 and H257, both located in loop3, showed substantial CSP and intensity changes (**Figure 5.2C**). These changes are likely caused by the direct interaction of N244 and H257 with the target deoxycytidine, as observed in the co-

crystal structure of A3Gctd-ssDNA[105]. It is noteworthy that R238, located in the short loop between β-sheet2 and β-sheet2, showed substantial perturbation, although it is not located at the catalytic site. Furthermore, F262, L263, V265 and W269, all located in helix2 with their side chains directed toward the inside of the protein and forming a hydrophobic core, displayed DNA-bound as well as DNA-unbound NMR signals following substrate addition (**Figure 5.2B**), indicating slow exchange dynamics between bound and unbound states. BR3 included W285 and T311-E330, which contains loop7 (T311-G319), previously suggested to be important for recognition of the hotspot sequence[87, 172, 209]. Especially, W285 located at the catalytic pocket[79, 87] as well as loop7 residues, A312, R313, Y315 and D316, displayed substantial CSP with slow exchange dynamics (**Figure 5.2B**). These perturbations were consistent with the co-crystal structure[105] as R313, Y315 and D316 had direct interactions with ssDNA. The exchange of bound and unbound states of loop7 residues likely destabilized helix4 (E323–E330), since residues located in helix4 showed >60% reductions in signal intensity.

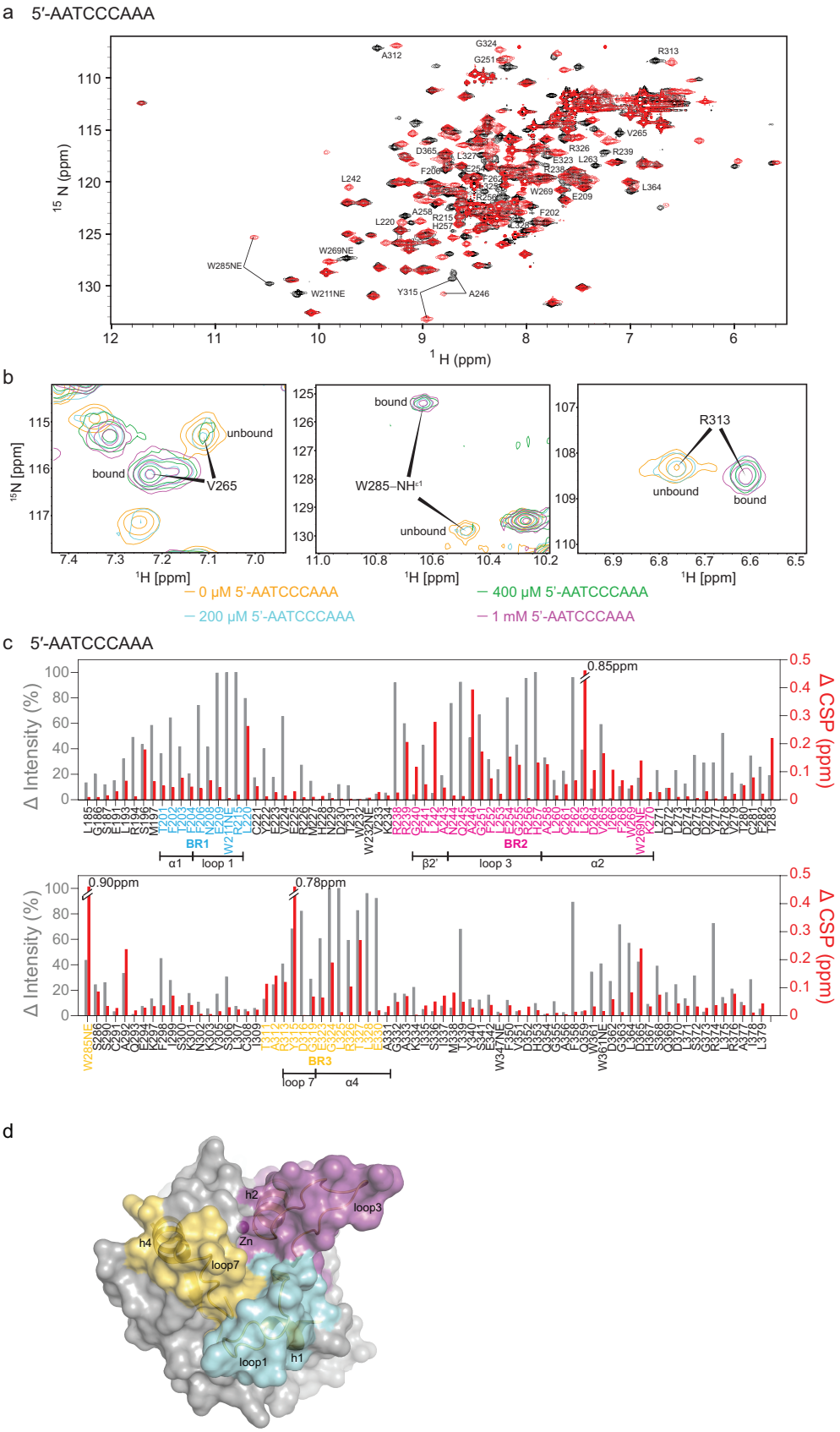
a 5′-AATCCCAAA
^{15}N (ppm)
^{1}H (ppm)
A312
G324
G251
R313
V265
D365
L327
E323
L263
R326
R239
F200
R238
L242
W269
E209
L364
A258
R215
F202
L220
H257
W269NE
W285NE
W211NE
Y315
A246
b
^{15}N [ppm]
unbound
bound
V265
W285–NH$^{\varepsilon1}$
R313
^{1}H [ppm]
0 μM 5′-AATCCCAAA
200 μM 5′-AATCCCAAA
400 μM 5′-AATCCCAAA
1 mM 5′-AATCCCAAA
c 5′-AATCCCAAA
Δ Intensity (%)
Δ CSP (ppm)
0.85ppm
BR1
α1
loop 1
BR2
β2′
loop 3
α2
0.90ppm
0.78ppm
BR3
loop 7
α4
d
h2
loop3
h4
Zn
loop7
loop1
h1

Figure 5.2: Chemical shift perturbation and signal intensity changes upon binding 5′-AATCCCAAA.

A) ^{1}H-^{15}N HSQC spectrum of 0.2 mM A3Gctd-2K3A-E259A mixed with 1 mM 5'-AATCCCAAA (red) overlaid onto 0.2 mM A3Gctd-2K3A-E259A (black). Significantly shifted peaks are labeled. B) NMR signals of residues in slow exchange regime upon titration of 5'-AAT′AAA. DNA-unbound signals are labeled unbound, whereas DNA-bound signals are labeled bound. Intensities of unbound signals decrease, while intensities of bound signals increase, upon increment of the ssDNA concentration. C) Quantification of peak intensity changes (gray bars, left axis) and chemical shifts changes (red bars, right axis). Residues in BR1, BR2 and BR3 are colored blue, magenta and yellow, respectively. Secondary structures within the binding regions are shown under the residues. D) Three ssDNA binding regions are shown on the surface of the structure of ssDNA-free wild type A3Gctd (PDB ID# 4ROV). Binding region 1 (BR1, cyan) spans residues 201-220, binding region 2 (BR2, magenta) spans residues 238-270, and binding region 3 (BR3, yellow) spans the non-consecutive residues 285, 311-330. Secondary structures of binding regions are shown in cartoon models.

We next compared CSP and intensity changes for 5'-AATCCCAAA with the intermediate product, 5'-AATCCdeoxyUAAA, by subtracting the changes for 5'-AATCCCAAA from the changes for 5'-AATCCdeoxyUAAA ('delta – delta' plot, **Figure 5.3A**). We found that 5'-AATCCdeoxyUAAA engaged all three binding regions described above for 5'-AATCCCAAA, however, the key difference was that BR1 residues displayed reduced chemical shift changes and signal intensity changes (appearing as negative red and gray bars in **Figure 5.3A**), indicating lesser interaction of BR1 with 5'-AATCCdeoxyUAAA. In addition, **Figure 5.3A** revealed that the exchange rate between bound and unbound states became faster with 5'-AATCCdeoxyUAAA than 5'-AATCCCAAA, as residues in BR2 and BR3 indicated reduced chemical shift changes (negative red bars) but increased intensity reduction (positive gray bars) caused by line-broadening due to exchange between bound and unbound states. The faster exchange rate with 5'-AATCCdeoxyUAAA was also evident in the spectrum (data not shown) since there was no residue showing two distinct bound and unbound signals, as had been displayed upon binding 5'-AATCCCAAA (**Figure 5.2B**). 5'-AATCCdeoxyUAAA contained a 5'-CC deamination motif, and the underlined C was presumably positioned at the catalytic site. The lesser interaction with 5'-CC compared to 5'-CCC was consistent with deamination efficiency since A3Gctd deaminates 5'-CCC 45-times more efficiently than 5 -CCdeoxyU[183].

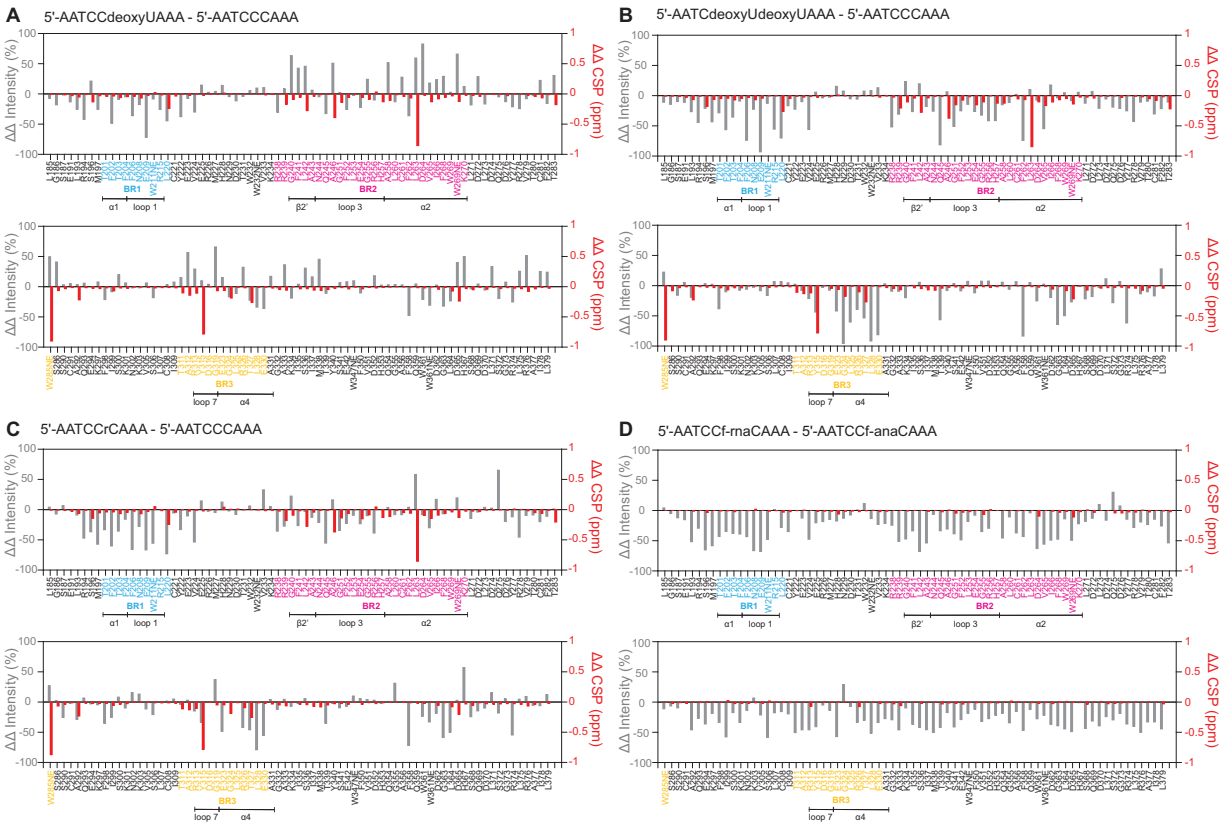

Figure 5.3: Comparison of chemical shift perturbations and intensity changes upon binding substrate and non-substrate ssDNAs.
ssDNA oligomers were mixed with A3Gctd-2K3A-E259A at 1 mM: 0.2 mM concentration ratio, and ^{1}H-^{15}N HSQC spectra were acquired. Peak intensity changes and chemical shifts changes upon addition of ssDNA were quantified for each ssDNA, then "delta – delta" plots were made by subtracting the changes for 5′-AATCCCAAA from the changes for A) 5′-AATCCdeoxyUAAA, B) 5′-AATCdeoxyUdeoxyUAAA and C) 5′-AATCCriboseCAAA. D) shows the "delta – delta" plot where the changes for 5′-AATCC(2′-F-ANA)CAAA have been subtracted from the changes for 5′-AATCC(2′-F-RNA)CAAA. The differences of chemical shift changes and signal intensity changes are shown by red bars (right axis) and gray bars (left axis), respectively.

Next, we compared a non-substrate ssDNA with the substrate by using 'delta – delta' plots, subtracting the changes for 5'-AATCCCAAA from the changes for 5'-AATCdeoxyUdeoxyUAAA (**Figure 5.3B**). 5'-AATCdeoxyUdeoxyUAAA is the final product of the deamination of 5'-CCC as A3Gctd does not deaminate the 5'-TC motif in vitro[88, 207, 210]. All three binding regions, BR1, BR2 and BR3, demonstrated greatly reduced chemical shift changes and signal intensity changes compared with 5'-AATCCCAAA (appearing as negative red and gray bars in **Figure 5.3B**), indicating interactions were lost. Although interactions with BR1 and BR3 were almost completely lost, W211NE (BR1), R215 (BR1) and D316 (BR3) retained significant reduction of signal intensities, suggesting that these residues still engage the DNA. We tested another non-substrate ssDNA containing a ribocytidine at the target position, 5'-AATCCriboseCAAA, by using 'delta–delta' plots (**Figure 5.3C**). **Figure 5.3C** displayed very similar profile to **Figure 5.3B** as all three binding regions substantially reduced both chemical shift changes and signal intensity changes compared with 5'-AATCCCAAA. Especially, BR1 residues lost interaction with the exception of W211NE and R215. BR3 was slightly more involved in the interaction with 5'-AATCCriboseCAAA than 5'-AATCdeoxyUdeoxyUAAA as BR3 residues showed smaller loss of signal intensity changes (shorter negative gray bars in **Figure 5.3C**).

The affinities of A3Gctd-2K3A-E259A for above substrate and non-substrate ssDNAs were assayed directly by using microscale thermophoresis (MST) (41). The apparent dissociation constant, *K*d , was determined for 5′ -AATCCCAAA, 5′ -AATCCdeoxyUAAA, 5′ - AATCdeoxyUdeoxyUAAA and 5′-AATCCriboseCAAA to be 1.57 ± 0.16, 2.17 ± 0.25, 2.76 ± 0.28 and 6.65 ± 0.86 mM, respectively (**Table 5.1**).

Although the differences of *K*d values among the ssDNAs were small, the direction of changes of *K*d values supported deamination activity of A3Gctd as it showed stronger affinity for the substrate (5′-AATCCCAAA) and weaker affinity for the product (5′-AATCdeoxyUdeoxyUAAA), and the intermediate product (5′ -AATCCdeoxyUAAA) showed a *K*d value between the substrate and the product. 5′-AATCCriboseCAAA displayed an affinity weaker than that of the product 5′- AATCdeoxyUdeoxyUAAA, indicating that a ribocytidine was disfavored more than deoxy-uridine for binding by A3Gctd.

Collectively, NMR and MST experiments showed that A3Gctd has multiple substrate and non-substrate ssDNA binding modes with similar affinities, but one conformation involved interaction with BR1, slightly enhancing binding, and presumably positioned the target cytosine base into the active site for the deamination to occur.

Table 5.1: Apparent Kd values of A3Gctd-2K3A-E259A for binding substrate and non-substrate ssDNAs.

ssDNA sequence	K_d [mM]
5′-AATCCCAAA	1.57 ± 0.16
5′-AATCCdUAAA	2.17 ± 0.25
5′-AATCdUdUAAA	2.76 ± 0.28
5′-AATCCrCAAA	6.65 ± 0.86
5′-AATCC(2′-F-RNA)CAAA	3.76 ± 0.30
5′-AATCC(2′-F-ANA)CAAA	1.74 ± 0.49

5.3.3 Effects of sugar conformation on ssDNA binding and deamination

Two potential mechanisms could exclude ribocytidines from catalysis by A3Gctd: the presence of the hydroxyl moiety at the sugar C2′ position of the ribocytidine or the conformation of the sugar; ribose prefers the C3'-endo conformation whereas deoxyribose prefers the C2'-endo conformation (**Figure 5.4**). To discriminate between these two possible mechanisms, we tested two fluorinated cytidine substrates, the first containing a fluorine substituted for the C2' hydroxyl of the ribose (2'-deoxy-2'-fluororibonucleic acid, 2'-F-RNA) and the second containing an arabinose sugar with the C2 hydroxyl substituted for fluorine (2'-deoxy-2'- fluoroarabonucleic acid, 2'-F-ANA) (**Figure 5.4**). The 2'-FRNA cytidine presumably had the C3'-endo conformation of the un-substituted ribose base, while the 2'-F-ANA cytidine presumably preferred the C2'-endo conformation typically seen in DNA[211]. Fluorine substitution retains an electronegative atom at the C2' position to mimic the presence of an oxygen atom with significantly weaker capability to form a hydrogen bond.

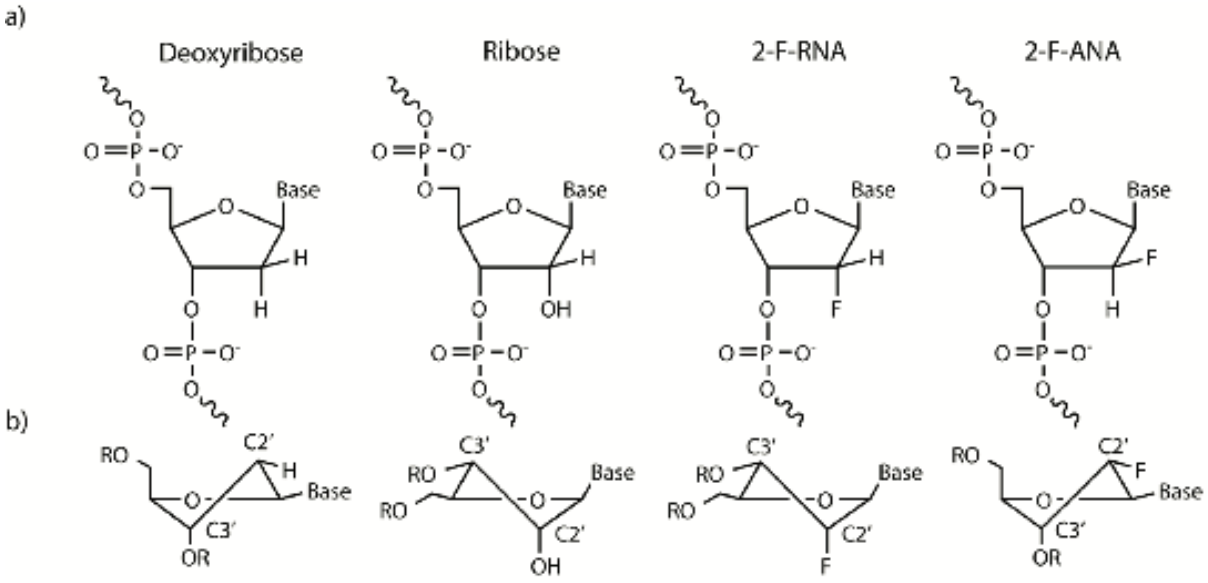

Figure 5.4: Comparison of nucleotide sugar conformation.
Varying functional group and stereochemistry at the C2′ position of the ribose ring significantly impacts the ribose ring conformation. A) Stereochemistry at the C2′ position of the nucleotides used in this study. B) Preferred conformations of the ribose ring containing specified substitutions in polynucleotide contexts. Exo-conformations indicated by vertical lines connecting functional groups at the indicated stereo-centers. Endo-conformations indicated by horizontal lines connecting functional groups.

We added 5’-ATTCC(2’-F-ANA)CAATT or 5’-ATTCC(2’-F-RNA)CAATT to a sample of A3Gctd2K3A-E259A. On the contrary, addition of 5’-ATTCC(2’-F-RNA)CAATT showed limited CSP and moderate reduction of NMR signal intensities across the protein, but did not display the intensive BR1 interaction, suggesting that 5’-ATTCC(2’- F-RNA)CAATT is not a substrate. **Figure 5.3D** shows the ‘delta-delta’ plot where the CSP and intensity changes for 5’-ATTCC(2’-F-ANA)CAATT are subtracted from the changes for 5’-ATTCC(2’-F-RNA)CAATT. 5’-ATTCC(2’-F-ANA)CAATT caused substantially increased reduction of NMR signal intensities compared with 5’-ATTCC(2’-FRNA)CAATT (negative gray bars in **Figure 5.3D**) in BR1 and BR2, but less extent in BR3.

To compare differences in affinity, apparent dissociation constant, Kd, values were determined using MST. Kd values were 1.73 ± 0.48 mM and 3.76 ± 0.30 mM for 5 -ATTCC(2’-F-ANA)CAATT and 5’-ATTCC(2’-FRNA)CAATT, respectively (**Table 5.1**). The Kd value of 5’-ATTCC(2’-F-ANA)CAATT was similar to that of substrate ssDNAs, including 5’-AATCCCAAA (Kd = 1.57 ± 0.15 mM) and 5’-AATCCdeoxyUAAA (Kd = 2.17 ± 0.25 mM), whereas the Kd value of 5’-ATTCC(2 -FRNA)CAATT was between the Kd values of two nonsubstrate ssDNAs, 5’-AATCdeoxyUdeoxyUAAA and 5’-AATCCriboseCAAA.

Since both NMR signal intensity changes and Kd values suggested that 5’-ATTCC(2’-F-ANA)CAATT might be a substrate for the deamination catalyzed by A3Gctd, we conducted 1D ^{1}H NMR deamination assays. Over the course of 8 hours, we observed the appearance of the H5 signal from the deaminated C2’-F-arabinose uracil product 5’-ATTCC(C2’-F-ANA)UAATT at 5.58 ppm, followed by the appearance of

another H5 signal at 5.68 ppm from the uracil from the product of deamination of the middle deoxycytidine, 5’-ATTCdeoxyU(C2’-F-ANA)UAATT (**Figure 5.5C**). The deamination speed for the 2’-F-ANA cyti dine was 0.06 ± 0.01 reactions/min. We also tested whether 5’-ATTCC(2’-F-RNA)CAATT could be deaminated by A3Gctd-2K3A, but over the course of 8 h, no uracil signal was observed, confirming that the 2’-F-RNA cytidine was not a substrate (**Figure 5.5D**).

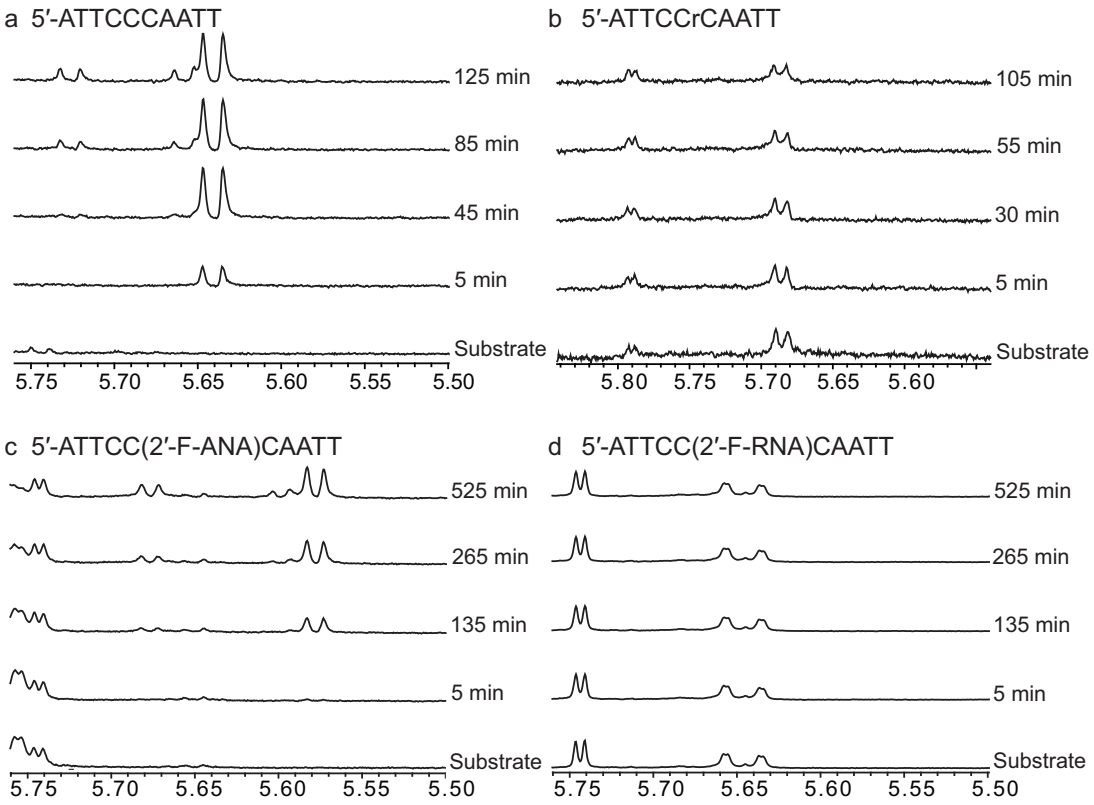

Figure 5.5: Real-time NMR deamination assays.
A) 1D ^{1}H spectra series of 150 μM 5′-ATTCCCAATT mixed with 1.5 μM A3Gctd-2K3A. A 5′-CCdU product doublet appears at 5.64 ppm, 5′-CdUdU product doublet appears at 5.73 ppm with concurrent shifting of the 3′ dU doublet to 5.66 ppm. B) 1D ^{1}H spectral series of 150 μM 5′-ATTCCrCAATT mixed with 50 μM A3Gctd-2K3A. No deamination product was observed. C) 1D ^{1}H spectra series of 150 μM 5′-ATTCC(2′-F-ANA)CAATT mixed with 30 μM A3Gctd-2K3A. A doublet signal for the (2′-F-ANA)U, which is the deamination product of (2′-F-ANA)C, was observed at 5.58 ppm. A doublet signal of the uracil resulted from deamination of the middle C, 5′-CU(2′-F-ANA)U, later appears at 5.68 ppm with concurrent shifting of the 3′ (2′-F-ANA)U to 5.60 ppm. D) 1D ^{1}H spectral series of 150 μM 5′-ATTCC(2′-F-RNA)CAATT mixed with 30 μM A3Gctd-2K3A. No deamination product was observed.

5.3.4 Molecular dynamics simulations of A3Gctd–ssDNA and A3A–ssDNA complexes

To reveal the atomic-level mechanism for how A3Gctd strongly disfavors ribocytidine (rC) as a substrate, we investigated the stability of ssDNA in the active site through molecular modeling and molecular dynamics (MD) simulations. We modeled 5'-TCCCAA and 5'-TCCrCAA with wild type A3Gctd based on the ssDNA-bound A3Gctd crystal structure (PDB ID: 6BUX) and performed MD simulations. Both MD simulations converged during the 100 ns simulation time. The deoxycytidine (dC) remained in the crystal structure conformation at the catalytic site during the simulations with 5'-TCCCAA (**Figures 5.6ABC, 5.7A, blue and B**). However, in the simulations with 5'-TCCrCAA, ssDNA still bound, but rC shifted ~3 Å away relative to the starting position within 10 ns of the MD simulation (**Figures 5.6DEF, 5.7A, red and C**). The relocation of rC was due to conformational rearrangements induced by the hydroxyl group attached to 2'C (2'-OH) in rC. H257, which is in close proximity to rC, can form a hydrogen bond with 2'-OH (5 ns; **Figure 5.6E**), which in turn destabilized the stacking interaction between the H257 imidazole ring and rC nucleobase. As a result, the critical hydrogen bonds stabilizing co-crystal structure conformation of the target ribocytidine, between the N244 sidechain and sugar, and between A258 backbone and nucleobase, were disrupted. The side chain of N244 then flipped towards rC and formed a new hydrogen bond with the rC base (10 ns; **Figure 5.6F**), and thus dislocated the rC to a position that was incompatible with the deamination reaction. rC was stable at this relocated position as it did not go back to the original catalytic position during the rest of the MD simulation (**Figure 5.7A, red**). Thus, our computational results were in agreement with

experimental data that A3G could deaminate dC but not rC. Furthermore, we performed similar modeling and MD simulations for A3A as a comparison since we observed binding[114] and deaminations of both dC and rC by A3A (**Figure 5.8**). The deamination rate for rC was two orders of magnitude slower than dC in an in vitro NMR-deamination assay (**Figure 5.8**). In agreement with experiments, the simulations showed that both dC and rC were stable in the catalytic site of A3A and maintained co-crystal structure conformation throughout the MD simulation (**Figure 5.7DEF**).

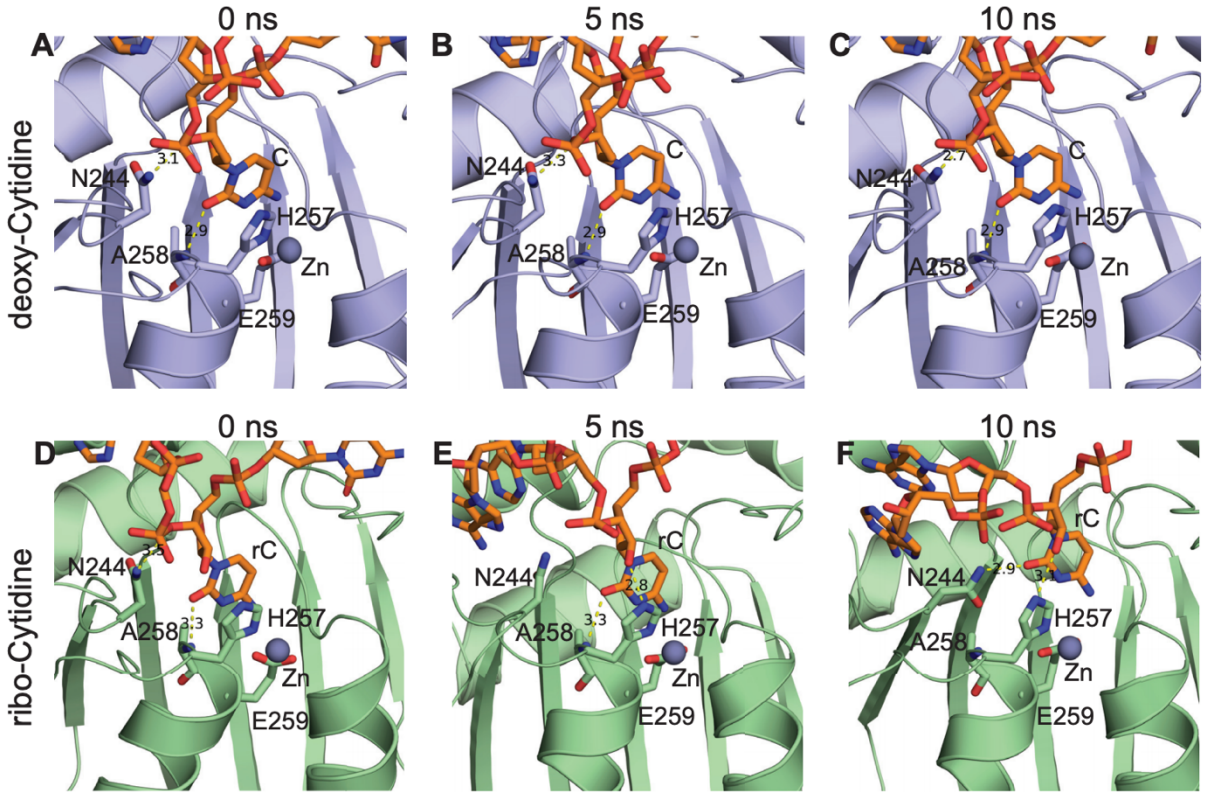

Figure 5.6: Snapshots from MD simulations with deoxy-cytidine and ribo-cytidine. All panels display the expanded view of the catalytic site of A3Gctd. Zn^{2+} is shown as a gray sphere, and yellow dashed lines indicate possible hydrogen bonding. DNAs are shown in orange stick model, and nitrogen and oxygen atoms are colored blue and red, respectively. A), B) and C) are snapshots of the 5′-TCCCAA and A3Gctd complex at 0 ns, 5 ns and 10 ns time points, respectively, while D), E) and F) are snapshots of the 5′-TCCrCAA and A3Gctd complex at 0 ns, 5 ns and 10 ns time points, respectively.

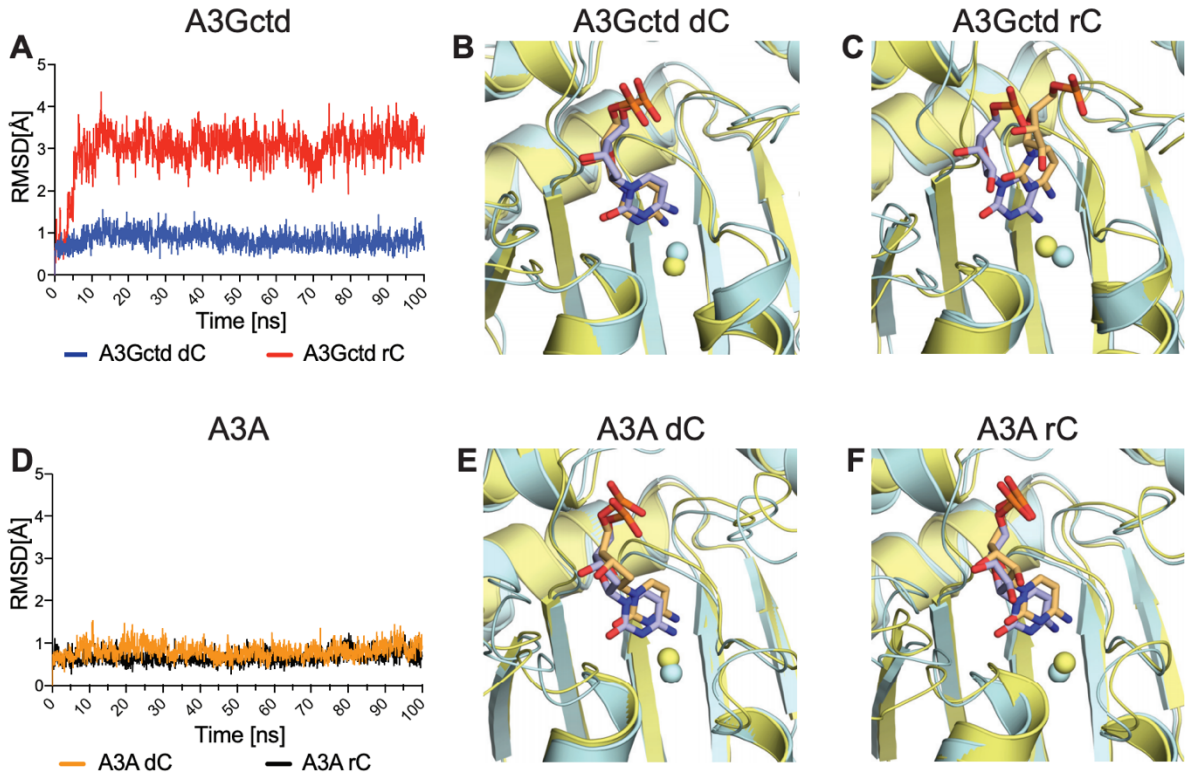

Figure 5.7: Comparison of A3Gctd and A3A in MD simulations with ssDNAs containing dC or rC.
A) A3Gctd and B) A3A root-mean-square-deviation (RMSD) of the target cytidine during MD simulation. RMSD of all heavy atoms of the target cytidine are shown for 100 ns simulation time. Deoxy-ribose C (dC) or ribose C (rC) with A3Gctd is shown in blue and red respectively in A), whereas dC or rC with A3A is shown in orange and black respectively in D). B.C) and E,F) Superposition of expanded views of the catalytic site of A3Gctd (B and C, and A3A E and F). Zn^{2+} molecules are shown as spheres. The snapshots at 0 ns are colored blue, whereas the snapshot at 100 ns is colored yellow. DNAs are shown in stick model, and nitrogen and oxygen atoms are colored blue and red, respectively. B) 5′-TCCCAA and A3Gctd complex, C) 5′-TCCrCAA and A3Gctd complex, E) 5′-AATCGAA and A3A complex, and F) 5′-AATrCGAA and A3A complex.

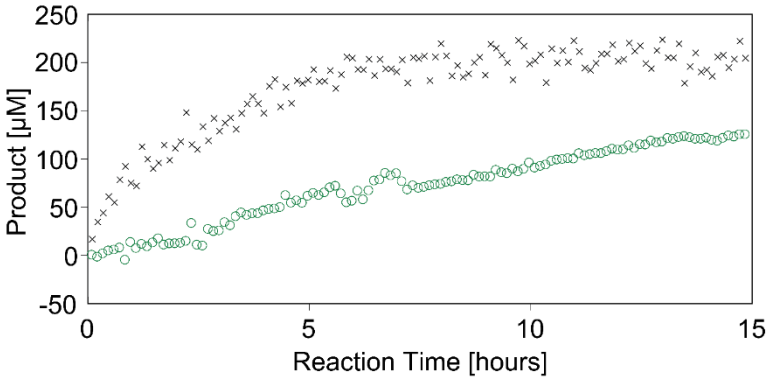

Figure 5.8: Deamination of rC and dC by A3A.

The ^{1}H NMR signal of the deamination product was tracked with respect to reaction time for 200 µM of 5′-AATTCAAAA mixed with 50 nM A3A (black crosses) and 200 µM of 5′-AATTrCAAAA mixed with 500 nM A3A (green circles). It should be noted that the concentration of A3A in the 5′-AATTrCAAAA reaction mixture was 10-fold higher than that of the 5′-AATTCAAAA reaction mixture. Spectra were measured on a Bruker Avance III 600 NMR Spectrometer at 37ºC in buffer containing 50 mM MES pH 6.0, 100 mM NaCl, and 0.5 mM TCEP. The initial rates for the 5′-AATTCAAAA and 5′-AATTrCAAAA deamination were 52 ± 2 reactions per minute and 0.3 ± 0.1 reactions per minute, respectively.

5.4 DISCUSSION

5.4.1 BR1 interaction distinguishes catalytic binding from noncatalytic binding

Characterizing the mechanism of substrate selection and non-substrate exclusion by A3Gctd is important for the development of inhibitors that can selectively modulate A3G activity as well as other A3 enzymes, including those with possible links to carcinogenesis such as A3A and A3B[150, 212]. Using a combination of experimental and computational methods, we highlight protein-ligand interactions critical for substrate binding that are absent for nonsubstrates. Our NMR data show that ssDNA binding interfaces of A3Gctd-2K3A form a continuous surface of the protein engaging loops 1, 3 and 7 (**Figure 5.2D**). Intense interactions with loop1 residues were observed only with substrate ssDNAs (**Figure 5.2C**). This finding supports the importance of loop1 residues for ssDNA binding that has been previously proposed based on the NMR and crystal structures of ssDNA-unbound A3Gctd[79, 87], and further, Carpenter et al. provided experimental evidence by swapping loop1 regions between AID and A3Gctd[209]. Our data is also consistent with our recent co-crystal structure of A3GCTD2 and ssDNA in which W211, R213 and H216, all in loop1, form direct interactions with DNA through –stacking and hydrogen bonds[105]. Ziegler and co-workers recently reported an interesting co-crystal structure of Pot1- fused-A3Gctd and ssDNA[108], and they suggested that loop1 and loop7 residues, including P210, W211, I314 and Y315, along with W285 interact with ssDNA nonspecifically during the search of deamination hotspots. Consistent with their suggestion, we found that W211, R215, D316 and W285 were involved in the interactions with substrate as well as non-substrate ssDNAs.

Allosteric regulation may be an effective strategy to develop inhibitors of A3 activity as competitive inhibitors have been elusive. The NMR signal perturbation data imply possible allosteric sites to target for A3Gctd. In particular, the slow exchange regime of helix 2 residues suggests that the reorientation of helix2 is an allosteric movement coupled with the target cytosine base positioning in Zn2+ containing active site, since helix 2 contains H257 and E259 which coordinate the Zn2+. Another possible allosteric site is 2 -strand (R238-L242) as the dynamic modulation of this strand is most likely helping N244 to stabilize the target deoxycytidine during catalysis[1, 104-106, 213, 214].

5.4.2 A3Gctd suppresses the catalytic efficiency of ribocytidine through sugar conformation and 2'-OH

Two distinct attributes of RNA could be responsible for the lack of catalytic activity; the presence of the 2'-OH directly, via steric clashing or unfavorable interaction, or indirectly, via conformational impact on the structure of the ribose ring that may prevent cytosine base access to the catalytic site. ssDNAs containing a ribocytidine at the target position have been shown to be excluded from deamination by A3G in vitro[197, 198]. Consistently, binding of 5'-AATCCriboseCAAA to A3Gctd2K3A-E259A resulted in a pattern of perturbations similar to that seen in 5'-AATCdeoxyUdeoxyUAAA, a nonsubstrate ssDNA (**Figure 5.3B and C**). We observed a slightly more extensive interaction for 5'-AATCCriboseCAAA with BR3 (residues W285, T311-E330) than seen with 5'- AATCdeoxyUdeoxyUAAA. This effect is likely due to interactions between loop7 and the 5'-CC motif present in 5'-AATCCriboseCAAA since loop7 recognizes the deoxycytidine flanking the 5 side of the target deoxycytidine in this motif. Nonetheless, 5'-AATCCriboseCAAA binding completely lacked the characteristic BR1 interactions

seen in substrate ssDNAs, suggesting that the presence of the ribose sugar prevented both the ribocytidine and the middle deoxycytidine from catalytically productive binding. Consistently, the real-time NMR deamination assay did not detect any deamination product from 5'-AATCCriboseCAAA (**Figure 5.5B**). To determine structural factors involved in DNA versus RNA differentiation, we mixed A3Gctd-2K3A-E259A with ssDNAs containing 2'-F-ANA cytidine or 2'-F-RNA cytidine at the target position. Here, fluorine serves as an isopolar and isosteric mimic of the native hydroxyl moiety in ribonucleotides, retaining similar interbond distances and similar electrostatic properties, as Pauling electronegativities are 3.44 and 3.98 for O and F, respectively[215]. Based on their chemical structures, 2'-F-ANA cytidine and 2'-F-RNA cytidine are assumed to have the C2'-endo conformation and the C3'-endo conformation, respectively[216](**Figure 5.4**). Addition of 5'-ATTCC(2'-F-RNA)CAATT resulted in moderate reduction of NMR signal intensities of A3Gctd-2K3A E259A for BR2 and BR3, yet lacked significant CSP and intensity change in BR1. Since 5'-ATTCC(2'-F-RNA)CAATT was not deaminated by A3Gctd-2K3A (**Figure 5.5D**), these results are consistent with 5'-AATCCriboseCAAA, suggesting the C3'-endo sugar conformation of the ribocytidine was disfavored for the catalytically productive binding. In contrast, the ssDNA containing a 2'-F-ANA cytidine exhibited interactions with BR1 (**Figure 5.3D**), and the 2'-FANA-cytidine was deaminated (**Figure 5.5C**), suggesting that the propensity for 2'-F-ANA to retain the C2'-endo sugar conformation of the native DNA allowed the catalytically productive binding. Furthermore, the subsequent deamination of the middle deoxycytidine in the 5'-CC(2'-F-ANA)C sequence (**Figure 5.5C**) suggested that the 2'-endo sugar conformation was also preferred for the nucleotide flanking the 3 side of the target deoxycytidine

because A3Gctd2K3A did not deaminate the middle deoxycytidine of 5'-ATTCCriboseCAATT nor 5'-ATTCC(2'-F-RNA)CAATT (**Figure 5.5B,D**). The deamination rate for the 2'-F-ANA cytidine was 0.06 ± 0.01 reactions/minute, which was over five times slower than that for deoxycytidine, suggesting that fluorine at the 2' position negatively affected the catalytic interaction with A3Gctd. Our results extended the finding of Nabel et al.[201] to A3Gctd, and provided experimental evidence as we showed that the 2'-F-ANA (presumably 2'-endo conformation) allowed the catalytically productive binding, but 2'-F-RNA (presumably 3'-endo conformation) did not. Since RNA is capable of adopting the 2'-endo sugar conformation, the remaining question is why the ribocytidine assuming the 2'-endo sugar conformation is not efficiently deaminated by A3Gctd. To answer this question, we investigated whether 2'-OH destabilizes the ribocytidine in the 2'-endo sugar conformation at the catalytic site of A3Gctd. MD simulations showed that 2'-OH triggered structural changes causing dislocation of the target base from the catalytic position. Therefore, 2'-OH may be another structural feature that negatively affects the deamination of ribocytidine by A3Gctd. On the other hand, A3A held ribocytidine at the catalytic position in the MD simulation, consistent with A3A's ability to deaminate ribocytidines albeit less efficiently compared with deoxycytidine (**Figure 5.8**). For A3A, the MD simulation showed that 2'-OH neither forms a hydrogen bond with the Zn2+-binding histidine (H70 in A3A), nor triggered subsequent structural changes of residues interacting with the target ribocytidine. The RMSD data indicated that movements of residues interacting with the ribocytidine were more restricted in A3A than A3Gctd; therefore, the target ribocytidine was stable at the catalytic position in A3A. Since A3A also deaminates 5-methyl-cytidine

as a substrate[217-219], studying A3A further by using NMR and MD simulations to understand binding modes for DNA and RNA substrates will be enlightening.

5.5 MATERIALS AND METHODS

5.5.1 Plasmid generation and protein purification

The pGEX-6P-1 vector (GE Healthcare Life Science) containing the C-terminal catalytic domain of A3G (A3Gctd), residues 191–384, with the previously reported 2K3A mutations (L234K, C243A, F310K, C321A, C356A)[79] was used as the template for Quikchange mutagenesis (Stratagene/Agilent Technologies) to introduce E259A substitution. *Escherichia coli* were transformed with the plasmid, grown to OD 0.5 at 37°C followed by a reduction in temperature to 17°C for 30 min, and protein expression was induced using a 0.1 mM final concentration of isopropyl β-D-1-thiogalactopyranoside (IPTG). The cells were lysed using sonication into buffer containing 50 mM sodium phosphate pH 7.3, 100 mM NaCl, 2 mM DTT, 0.002% Tween 20. Following high-speed centrifugation, the super- natant was bound to glutathione sepharose resin (GenScript) and washed under high salt and high detergent conditions, 400 mM NaCl and 0.06% Tween 20, followed by two washes in low salt and low detergent conditions, 30 mM NaCl and 0.002% Tween 20. The GST-tag was removed using PreScission protease (GE Healthcare Life Science) in 50 mM sodium phosphate buffer at pH 7.3 with 100 mM NaCl, 2 mM DTT and 0.002% Tween 20. Following cleavage, the protein was dialyzed into sample buffer containing 50 mM sodium phosphate pH 6.0, 100 mM NaCl, 2 mM DTT, 0.002% Tween 20 and 50 μM $ZnCl_2$.

5.5.2 NMR spectroscopy

All multi-dimensional NMR spectra were acquired on an 850 MHz Bruker Ascend spectrometer equipped with a 5 mm Z-gradient TCI cryoprobe. Samples contained a final volume of 300µL (97% H2O/3% D2O, v/v), and spectra were taken at 293 K. Backbone resonance assignments for the A3Gctd-2K3A-E259A mutant were derived using TROSY versions of a standard set of triple resonance spectra (HNCA, HN(CO)CA, HNCACB, HN(CO)CACB, HNCO, HN(CA)CO) on uniformly $^{15}N/^{13}C$ labeled protein with 85% random deuteration at pH 7.3. Assignments were transferred to pH 6.0 HSQC spectrum by titrating pH to identify relevant peak shifts. ^{15}N-HSQC with ssDNA titrations were collected on 0.2 mM 15N-labeled A3Gctd2K3A-E259A samples at pH 6.0 with unlabeled ssDNA at ratios of 1:1, 1:2 and 1:5 (A3Gctd-2K3A-E259A:ssDNA). Each titration point was collected with 128 transients and 100 real data points in the indirect ^{15}N dimension. Chemical shift and intensity changes were monitored through a series of spectra at varying relative concentration ratios. Chemical shift changes were calculated using the equation:

$$\Delta\delta_{ppm} = \sqrt{(\delta H_x - \delta H_0)^2 + \left(\frac{\delta N_x - \delta N_0}{5}\right)^2}$$

Intensity changes were calculated using the difference in peak height at the center of the ^{15}N-HSQC peak between the unbound and bound spectra divided by the unbound peak height. Real-time 1D ^{1}H NMR deaminase assays were performed on Bruker Avance III 600 MHz NMR spectrometer at 20°C in buffer containing 50mM sodium phosphate pH 6.0, 100 mM NaCl, 1 mM DTT, 10 µM ZnCl2 and 0.002% Tween-20. Oligonucleotide substrate concentrations of 150 µM were used in the assays with

enzyme concentrations ranging from 1.5 to 50 μM. Spectra were analyzed using Topspin 3.5 software package (Bruker Corporation, Billerica, MA, USA).

5.5.3 DNA oligomers

Oligonucleotides containing standard DNA and RNA bases were synthesized by Integrated DNA Technologies (IDT). Oligonucleotides containing 2'-deoxy-2'-fluororibonucleic acid (2'-F-RNA) and 2'-deoxy-2'-fluoroarabonucleic acid (2'-F-ANA) at the underlined cytosine position in the 5'-ATTCCCAATT oligonucleotide were synthesized by Boston Open Labs.

5.5.4 Microscale Thermophoresis assay (MST)

The binding affinity of purified A3Gctd-2K3A-E259A with 9nt ssDNAs (IDT), including 5'-AATCCCAAA, 5'-AATCCdeoxyUAAA, 5'-AATCdeoxyUdeoxyUAAA, 5'-AATCCriboseCAAA, 5'-ATTCC(2'-F-ANA)CAATT and 5'-ATTCC(2'-F-RNA)CAATT, were measured using Monolith NT.115 (Nano Temper Technologies)[220]. RED-tris-NTA fluorescent dye solution was prepared at 100 nM in the MST buffer (50 mM phosphate pH 6.0, 100 mM NaCl, 1 mM DTT, 0.002% Tween-20, 20 μM ZnCl2). A3Gctd-2K3A-E259A was mixed with dye at final concentration of 100 nM and incubated for 30 min at room temperature followed by centrifugation at 15 000 g for 10 min. The ssDNAs were prepared to stock concentration of 64 mM for AATCCCAAA, 5'-AATCCdeoxyUAAA, 5'-AATCdeoxyUdeoxyUAAA, 5'-AATCCriboseCAAA, or 32 mM for 5'-ATTCC(2'-F-ANA)CAATT and 5'- ATTCC(2'-F-RNA)CAATT) in the MST buffer. To determine the binding affinity, 10 μl of ssDNA solution at 16 different concentrations, ranging from 32 mM to 0.24 μM, or 16 mM to 0.12 μM for 5'-ATTCC(2'-FANA)CAATT and 5'-ATTCC(2'-F-RNA)CAATT, were prepared in LoBind centrifuge tubes (Fisher Scientific), then 10 μl

of fluorescent labelled A3Gctd-2K3A-E259A solution (100 nM) was added to each tube. The mixtures were incubated at 4°C to reach equilibrium. Each incubated solution was loaded into a Nano Temper MST premium coated capillary. The measurement was performed at room temperature using 40% LED power and 20% MST power. The experiment was repeated three times using freshly purified protein at each time, and data analysis was carried out using Nano Temper analysis software (MO affinity).

5.5.5 Molecular dynamics simulations

The structures of wild type A3Gctd with 5'-TCCCAA or 5'-TCCrCAA were modeled starting from ssDNA-bound A3Gctd crystal structure (PDB ID: 6BUX) through program Modeller 9.15 using basic modeling. The structures of wild type A3A with 5'-AATCGAA or 5'-AATrCGAA were modelled based on A3A DNA-bound crystal structure (PDB ID: 5KEG) using the same method. The phosphate groups of 5 T base in all structures were removed to prevent strong electronegative environment. All molecular dynamics simulations were performed using Desmond[164] from Schrodinger. The models were first optimized using Protein Preparation Wizard at pH 6.5. The simulation systems were then built through Desmond System Setup using OPLS3 force field[131]. Simple point charge (SPC) water model was used for solvation with cubic boundary conditions and 12 A buffer box size. The final system was neutral and had ˚ 0.15 M sodium chloride. The simulation system was first energy minimized with gradually reduced restraints (1000, 5, 0 force constant) on backbone and solute heavy atoms. A multi-stage MD simulation protocol was used. Briefly, the system was simulated using NPT ensemble with gradually increased simulation time (24, 50 and 500 ps) and decreased restraints on the solute heavy atoms to no restraints. The final production stage was performed at

300 K and 1 bar with no restraints using NPT ensemble. 100 ns MD simulations were performed for all DNA-bound structures. The analysis of MD simulations was performed separately for each trajectory. The RMSD and RMSF of protein and DNA molecule were calculated using Simulation Interactions Diagram from Schrodinger. Hydrogen bond occupancies over the trajectories and the side chain dihedral angles were calculated using program VMD. A hydrogen bond was defined as having a donor-acceptor distance of 3.6 Å and involving polar atoms nitrogen, oxygen, sulfur and fluorine. The donor-hydrogen-acceptor angle was defined as being less than 30 degrees. The trajectories from MD simulations for RMSD, distance and dihedral analysis were aligned based on whole molecules.

6 Chapter VI: Discussion and future directions

6.1 Combining molecular modeling and pMD with experimental assays to study the biology of A3s

APOBEC3s proteins (A3s), a family of cytidine deaminases, protect the host cell from endogenous retro-elements and exogenous viral infections by introducing hypermutations. However, the ability of these proteins to deaminate cytidines in ssDNA makes APOBECs a double-edged sword. When over-expressed, the resulting mis-regulated deaminase activity of A3s can contribute to genomic instability and cause cancer, as has been reported for A3A, A3B and A3H. Over the past years, several crystal and NMR structures of apo A3s and DNA/RNA-bound A3s have been determined. These structures provide the basis for understanding how A3s structurally bind to ssDNA and regulate catalytic activity, and can guide inhibitor design to target the active site of A3s to find potential anti-cancer drugs.

Why we have these functionally overlapping but distinct A3 enzymes still remains as a question. The enzymology and biological functions of A3s have been extensively studied. Despite overall structural similarity, A3 proteins have different binding affinity/deamination activity and substrate specificities. The ssDNA binding affinity of A3s could range from nM up to mM. In general, A3s prefer to deaminate the cytidine after thymidine (TC) except A3G which prefers CC. The A3 structures have suggested the importance of the loops around the active site for nucleotide specificity and binding. However, the structural mechanisms underlying A3 activity and substrate specificities require further examination.

In my thesis research I used a combination of computational modeling and pMDs with experimental verifications as a powerful method to study the A3 family. First of all,

this combined approach overcomes the challenges in determining all apo and DNA-bound A3 structures to study the structural mechanisms in A3s. The low solubility, tendency for oligomerization and low DNA affinity of certain A3 proteins have required introducing mutations to be able to structurally and biochemically characterize these proteins in vitro [79-81, 83-85, 92, 96, 99, 104, 106], or prevented such characterization especially for NTDs and full-length A3s. Molecular modeling with refinement from MD simulations enable generating reliable structures of A3 proteins (alone or with substrate oligonucleotides) and thus provide insights or propose hypothesis when crystal structures are not accessible. In addition, the parallel detailed analyses, including dynamics (RMSFs, RMSDs), intermolecular interactions (hydrogen bonds, vdW contacts) and electrostatics calculations from pMDs help reveal the underlying subtle but key differences among the highly similar A3s.

Starting with A3B (as described in Chapter II), I applied pMD with experimental binding assays to understand the structural basis for ssDNA binding, relatively lower activity, and substrate specificity in A3B compared to the highly similar but distinct A3A. The crystal structure had the critical loop 1 of A3B replaced with that of A3A, resulting in a more active chimeric enzyme. I modeled the wild-type A3B–ssDNA complex structure. I identified Arg211 in loop 1 as the gatekeeper coordinating DNA, and residues that determine nucleotide specificity at -1' position. I also found a unique auto-inhibited conformation in A3B that restricts access to the active site and may underlie lower catalytic activity. The cross-validation and agreement between computational structural analysis and experimental results in this work allowed me to apply the same method for studying other A3s as well as different specificities.

In Chapter III, I examined the structural mechanism of substrate specificities in A3s focusing on A3A, A3B and A3G as these family members have the most experimental characterization. The proposed mechanisms and observations from modeling and pMD were correlated with published experimental results. In this study, I identified an interplay between DNA binding conformation and substrate sequence specificity. I also found the potential molecular mechanisms of experimentally observed substrate specificity at -2' position for these A3s. In addition, I revealed interdependence between substrate nucleotide binding sites as well as the active site loops. Previously, we had found potential intra-DNA interactions in A3A, which correlated with its substrate sequence specificity (Chapter IV). Finally, I investigated the structural mechanism for exclusion of RNA from A3G catalytic activity using similar methods (Chapter V). Overall, the comprehensive structural analysis of A3 domains in this thesis revealed the determinants of substrate specificity and shed light into the biological function of these enzymes.

6.2 Implications of studying substrate specificities of A3 family

6.2.1 Studying the substrate specificities broadens our understanding of A3s

The A3 family is diverse with extensive polymorphisms among the family members. A3s are found only in primates. In human, A3 genes are clustered on chromosome 22. Interestingly, A3s seem to be under positive selection and thus constantly changing. For instance, A3H has at least 7 haplotypes in human. Another polymorphism in A3 is an A3B deletion allele[159]. Individuals with this deletion allele seem to be more susceptible to HIV[221] and have increased risk for cancer[222]. Besides,

A3s have can have either one or two zinc-binding domains. The pseudo-catalytic NTDs in the two-domain A3s have the same overall structural fold compared to the catalytically active CTDs. NTDs also have the conserved active site residues. The biological implications of having different polymorphisms and NTDs in the A3 family still remain largely unknown.

Having diverse A3s may help against novel retroelements. The A3 family is part of our innate immunity to protect host genome from retroviral infections and retro transportations. Considering the fast evolution of these retroelements in nature, having multiple A3s with different specificities may serve as a pool of weapons to effectively target viruses with a fast response time.

Having diverse A3s may be required for the regulation of activity to prevent cancer. Over-expressed A3s have been shown to cause heterogeneity in multiple cancers and thus help cancer evolve to escape from immune system. Moreover, study of human cancer cell lines has suggested A3s may be involved in the origination of cancer in human[167]. Therefore, the expression, localization and activity of A3s have to be regulated. For instance, cytoplasm localized A3s lost ability to deaminate RNAs possibly to avoid interfering with translation. Thus multiple A3s with different specificities may be needed for proper cellular regulation.

Therefore, studying the substrate specificities of A3 family help us better understand the biological functions of each A3, the balance between host protection versus viral evasion, how gene mutators are regulated, and potentially the evolutionary pathways leading to cancer.

6.2.2 Applying insights from substrate specificities to design specific inhibitors to target A3s

Common cytidine deaminase (CDA) inhibitors, which are usually cytidine mimics, fail to inhibit A3s despite the fact that CDA and A3s share high sequence identity and similar structure in the active site (Figure 6.1). Recent crystal structures and prior biochemical characterization suggest that unlike CDAs, which can bind a single nucleotide, A3s bind longer oligonucleotides (usually at least five nucleotides) and more extensive interactions may be needed for efficient binding. The underlying mechanisms for this disparity have remained unknown. My studies of substrate specificities may provide some insights into this difference. The interdependent interactions between upstream nucleotides and substrate target cytidine suggested that the binding event in A3s is not restricted to the active site; but instead is coordinated with loops around the active site. Loop regions are usually more susceptible to changes compared to other parts of a protein. Having binding event coordinated by these loops may allow A3s to quickly adapt to changes to accommodate different substrates and drive fast evolution to restrict novel retroelements.

Accordingly, the design of high affinity A3 inhibitors should start from chemically modified oligonucleotides instead of a single nucleotide or small molecules. Recent studies incorporating non-native bases into ssDNA have shown potential inhibition of A3s[223-225]. However, these oligos have not been optimized based on structural interactions of the ssDNA within the enzyme complex, nor for drug-like properties. My thesis work provide multiple insights to guide the inhibitor design and find inhibitors that selectively target specific A3s: 1) Optimizing the nucleotides (sequences A/T/C/G or

ribose/deoxy-ribose) at each position. My studies of protein–substrate interactions at each nucleotide position have revealed critical residues as well as preferred substrate sequence. Choosing the optimal nucleotide would not only increase binding affinity, but also specificity. For instance, inhibitors with A at the -2' position would increase affinity toward A3B but not A3A; inhibitors using ribose-based oligonucleotides may specifically target A3A but not A3G. 2) Leveraging different conformations of bound oligonucleotides. ssDNA binds A3s in different overall conformations because the variable active site loops define a unique binding groove in each A3. Hence designing oligonucleotide-based inhibitors with defined conformations would help increase specificity. For instance, designing macrocyclic or hairpin-based oligonucleotide inhibitors will specifically target A3A but not A3G. 3) Applying a similar method (modeling and pMD) for virtual screening. My studies using computational methods combined with experimental verifications have shown promising results for studying A3 substrate recognition. Inhibitor candidates based on substrate DNA sequences can also be evaluated using a similar approach. Hence, virtual screening using molecular modeling followed by MD simulations will allow effectively evaluating potential inhibitors before spending extensive efforts for synthesis and experimental characterization.

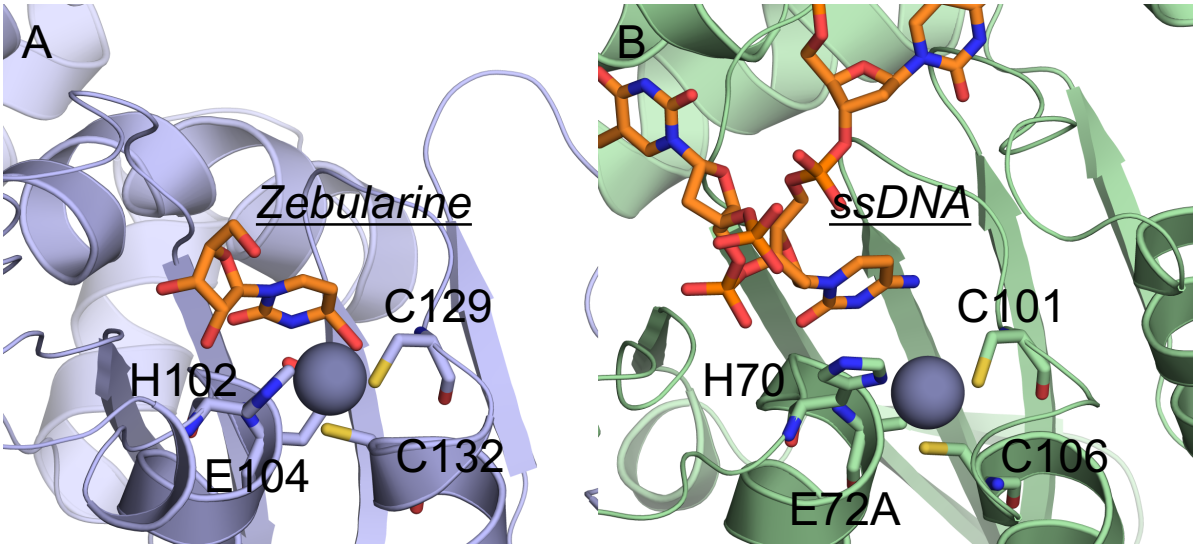

Figure 6.1: Structural comparison of the active site in human CDA and A3A. The active site comparison between human CDA (PDB: 1CTU; left) and A3A (PDB: 5KEG; right). Zinc coordinating residues, CDA inhibitor and ssDNA are shown in sticks representation.

6.2.3 Applying insights from substrate specificities to design better gene editors

Most modifications that have been implemented for improving the efficiency and fidelity of gene editing using cytosine base editors (CBEs) are not on the deaminase. Deaminase however vastly affects the performance and application of CBEs. One major limitation in using CBEs to treat genetic diseases is that the target site must naturally exist in the preferred sequence context for cytidine deaminase, which may not always happen for the desired modification. In addition, current versions of CBEs can edit off-target cytidines within the editing window due to poor substrate specificities of incorporated cytidine deaminase. CBEs incorporating APOBEC1 and A3A as the deaminase have been shown to induce transcriptome-wide off-target RNA editing[226]. One possible solution to these problems is to have libraries of context-dependent base editors as a toolkit to select from to specifically edit the desire site. My thesis studies of A3s provide additional suggestions: 1) Using A3s that lack RNA editing ability to minimize off-target RNA editing 2) Incorporating different A3s depending on the nucleotide sequence of the target sites. For instance, using A3G for 5'-CC while using A3B for 5'-TC. In addition, A3s can be used to specifically target sites with epigenetic modifications, namely A3A, A3B or A3H. 3) Engineering A3s to adapt diverse substrate sequences. In general, A3s prefer TC except A3G, which prefers CC. However, desired editing site may not always exist naturally after T or C. Hence, engineering A3s, especially introducing changes to the active site loops, may allow A3s to adapt novel sequence specificity. Besides, structural comparison between AID, which prefers A/G at -1 position, and A3s may help guide the engineering.

6.3 Applying molecular modeling and pMD to other systems

6.3.1 ssDNA binding and substrate specificities of other A3s

Many questions still remain about substrate recognition and specificities in the A3 family: How does substrate DNA bind to other A3s besides A3A, A3B and A3G? How do the diverse active site loops determine the binding, activity and substrate specificities in A3s? How do full-length two-domain A3s bind to substrate DNA, and especially how is the NTD involved in ssDNA binding? Using molecular modeling, I have modeled the wild type full-length A3G structure and performed electrostatics surface analysis. From the model, I identified a positively charged patch through the active site of CTD toward loop 1 of NTD, which suggests how ssDNA might bind to NTD simultaneously with CTD. Mutations to this positive patch have caused defects in both deamination and anti-HIV activity, supporting the proposed binding mode (Detailed results are presented in Appendix I)

In addition to substrate specificities, how A3s interact with many other proteins for different biological functions has remained unknown. Among these proteins, Vif represents the most well-known but yet experimentally challenging binding partner of A3s due to its intrinsically disordered structure. Vif binds to more than one A3 (A3G, A3F, A3D and A3H) with distinct binding surfaces. Recently, the very first Vif–A3F structure has been determined through cryo-EM. However, A3F in this structure has only CTD and multiple mutations. This A3F construct may not represent the biologically relevant interface. Using modeling and pMDs, we can generate models of wild type A3F–Vif complex and thus study the binding interface. What we learn from this

interaction may help guide the design of inhibitors that potentially disrupt Vif binding and restore innate immunity to restrict HIV.

6.3.2 Applying modeling and pMD to other proteins beyond A3s

Besides A3s, I have applied modeling and pMD to investigate multiple biological systems including influenza virus (hemagglutinin, neuraminidase and influenza antibodies), human renalase, HIV protease, and recently a neutralizing antibody against SARS-CoV-2 spike protein. The methods that I employ provide structural insights into diverse topics in biology. For example: 1. Drug design: modeling of human renalase protein revealed a potential inhibitor, which was verified in *in vivo* mouse studies; 2. Drug resistance: modeling and MD simulations of neuraminidase variants suggested the molecular mechanism for F10 antibody resistance; 3. Epitope identification: modeling of antibody binding to the viral protein identified the epitope, for influenza hemagglutinin broadly neutralizing antibody and SARS-CoV-2 spike protein antibody.

In fact, molecular modeling and pMD can be an integral part of the standard pipeline for analyzing protein structures, functions and molecular mechanisms to gain insights into biological function. This powerful combination enables characterizing and comparing distinct but related systems, such as members of an enzyme family, or a series of inhibitor modifications. Besides, experiments would also benefit from these computational analyses, for instance, for generating hypothesis to be tested or narrowing down the list of inhibitors or protein variants to be experimentally evaluated.

7 APPENDIX:

7.1 Appendix I: Crystal structure of full-length APOBEC3G bound to dinucleotide reveals domain orientation and a ssDNA-binding channel

7.1.1 PREFACE

Appendix I is a collaborative study that currently under peer review:

Atanu Maiti, Wazo Myint, Krista A. Delviks-Frankenberry, Shurong Hou**,** Christina Sierra Rodriguez, Nese Kurt Yilmaz, Vinay K. Pathak, Celia A. Schiffer, Hiroshi Matsuo. "Crystal structure of full-length APOBEC3G with a dinucleotide bound reveals domain orientation and a likely ssDNA-binding channel."

Contribution from Shurong Hou:

I performed the molecular modeling of wild type full-length A3G structure and electrostatics surface analysis of the wild type model. I made figure 4B.

7.1.2 ABSTRACT

APOBEC3G (A3G) is a single-stranded DNA (ssDNA) cytosine deaminase that can restrict HIV-1 infection by mutating the viral genome. A3G consists of a non-catalytic N-terminal domain (NTD) and a catalytic C-terminal domain (CTD). A3G-NTD helps virion incorporation and A3G catalytic function by binding ssRNA and ssDNA, respectively. Structures of both A3G domains have been solved individually, however no full-length A3G structure is available. We determined the crystal structure of a soluble variant of full-length A3G (sA3G*) with a dinucleotide bound. Furthermore, our sA3G* structure demonstrated NTD to be rotated 90° against CTD along the major axis

of the molecule, an orientation that forms a positively charged channel, consisting of NTD loop-1 and CTD loop-3. This channel connects to the dinucleotide and we propose that it could act as a ssDNA binding pathway. Structure-based mutations, *in vitro* deamination assays, and hypermutation analyses of proviruses suggest that R24 located in NTD loop-1 provides a key interaction with ssDNA that is required for efficient deamination of 5'-CC motifs in virions. Furthermore, the dinucleotide binds near the catalytic site of CTD but distant from catalytic Zn^{2+}. Hydrogen bonds and hydrophobic interactions formed between the dinucleotide and A3G differ significantly from those formed with substrate 5′-TCCCA, the latter revealed by our previous co-crystal structure of A3G-CTD with ssDNA. This new information on the non-catalytic interactions between A3G and DNA, including how nucleotides are positioned, provides a plausible mechanism by which A3G scans ssDNA for deamination target sequences.

7.1.3 INTRODUCTION

Human APOBEC3G (A3G) is a member of the seven human APOBEC3 (A3A, A3B, A3C, A3D, A3F, A3G and A3H) family of proteins which belong to the larger APOBEC super- family[4, 190-193]. All APOBEC (APOBEC1, APOBEC2, APOBEC3, APOBEC4 and activation- induced cytidine deaminase (AID)) proteins catalyze Zn-dependent deamination of deoxy- cytidine in single-stranded DNA (ssDNA) converting deoxy-cytidine to deoxy-uridine[194]. With their deaminase activity, APOBEC proteins play crucial roles in a variety of biological processes ranging from antibody diversification to defense against viral infections[12, 22, 170, 175]. Accordingly, A3G (also A3D, A3F and A3H) is capable of restricting human immunodeficiency virus type-1 (HIV-1) and other

retroviruses with its deoxy-cytidine deaminase activity[22, 145-147, 174, 175, 227, 228]. However, HIV-1 has developed a mechanism to counteract APOBEC3 proteins by one of its accessory proteins, namely viral infectivity factor or Vif, which leads APOBEC3 proteins to proteasomal degradation[228-234]. Briefly, in the absence of Vif, A3G is encapsidated into newly forming virions in association with viral and host RNAs[22, 229-231, 235-239]. Within the virion, the viral RNA is reverse transcribed into negative strand (-) DNA which acts as a template for positive strand (+) DNA synthesis. Before positive strand (+) DNA synthesis, encapsidated A3G recognizes substrate deoxy-cytidines, including hotspot sequences (5'-CCC and 5'-CC) in newly formed negative strand (-) DNA, and catalyzes the deamination of deoxy-cytidine to deoxy-uridine. Subsequently, during positive strand (+) DNA synthesis deoxy-uridine is used as a template, which results in G-to-A hypermutation in the positive strand (+) DNA. Thus, mutations exerted in the proviral DNA make the virus inactive or non-functional[23, 169, 240]. It is noteworthy that in addition to lethal hypermutation, deaminase-independent mechanisms of HIV-1 restriction by APOBEC3 proteins have been reported[29, 241-244].

Among APOBEC3 proteins, A3B, A3D, A3F and A3G consist of two homologous Zn binding domains. A3G contains catalytically inactive N-terminal (A3G-NTD) and a catalytically active C-terminal (A3G-CTD) domain[245, 246]. A3G-NTD plays an essential role in encapsidation of A3G through association with viral and host RNA[236, 247, 248]. In addition, A3G- NTD binds to ssDNA and probably supports catalysis by stabilizing the A3G-substrate ssDNA complex. A3G-NTD is also involved in Vif-mediated degradation of A3G as Vif interacts with A3G-NTD and triggers the degradation of A3G through the ubiquitin-proteasome pathway[229-232]. A3G-CTD contains a Zn^{2+} binding motif HxE-X_{23-}

$_{28}$-C-$X_{2–4}$-C and catalyzes the deamination of deoxy-cytidine to deoxy-uridine. Mechanistically, a Zn^{2+} coordinated water molecule produces a hydroxide ion which attacks the C4 atom of cytosine, causing hydrogen to be transferred to the carboxylate group of a glutamic acid coordinated to Zn^{2+} through a water molecule and ultimately to the product ammonia[1, 213, 214]. Although all APOBEC3 proteins use a similar deamination mechanism, they show varying preferences for target nucleotide sequences; A3G prefers 5'-CCC and 5'-CC, whereas other A3s prefer 5'-TC[249]. Nucleotides flanking the target motifs, and secondary structures of ssDNA can also modulate A3 binding affinity[114]. Furthermore, previous studies showed that A3G deaminates multiple target sequences processively from the 3'-end to the 5'-end of a ssDNA, although the mechanism that enables processivity remains elusive[210].

Three dimensional structures of individual domains (NTD and CTD) of A3G have been solved by us and other laboratories using NMR and X-ray crystallography [79-81, 85, 87, 89, 95, 101]. These structures are overall similar and share the same secondary structures, including six α-helices and five β-strands, and one HxE-$X_{23–28}$-C-$X_{2–4}$-C Zn^{2+} binding motif. Recently, structures of APOBEC3 domains complexed with ssDNA have emerged. Xiao et al. published a crystal structure of rhesus macaque A3G N-terminal domain complexed with poly-T ssDNA, revealing non-catalytic binding of a single thymine[99]. Co-crystal structures of A3A[104, 106] and chimeric A3B-CTD[106] with ssDNA bound have been informative with regard to the 5'-TC target sequence preference for A3A/A3B. Our recent structure of A3G-CTD co-crystalized with substrate ssDNA provided extensive information on how A3G-CTD specifically recognizes its 5'-CCC preferred target sequence[105]. Another structure of A3G-CTD with a non-preferred

adenine nucleotide bound near the catalytic pocket[108] showed possible interactions for initial DNA scanning of target sequences. Most recently, we have revealed that DNA interaction with helix-1 and loop-1 (T201-L220) of A3G-CTD distinguishes the substrate binding mode from non-substrate binding modes, and that a 2'-endo sugar conformation of the target deoxy-cytidine is important for stabilization of the substrate during catalysis[115].

Although the individual domain structures of APOBEC3 proteins are available [79-81, 83, 85, 87, 89, 91-93, 95, 96, 98, 101, 200], resolving the structures of full-length APOBEC3 proteins containing both NTD and CTD has been challenging due to their poor solubility and aggregation tendency. Without a full-length A3G structure, several important questions remain unanswered, including how the two domains are organized or oriented with respect to each other, how they interact, if this orientation changes when Vif binds, and how the two domains contribute to the search for and binding to the deamination target sequences and to RNA binding and multimerization.

To address these questions, here we present a crystal structure of double-domain A3G bound to dinucleotide at 3.0 Å resolution. To overcome the solubility and aggregation problems of wild-type full-length A3G, we generated a soluble A3G variant (soluble A3G or sA3G) amenable for structural studies. The crystal structure shows the relative positioning of the two A3G domains and suggests a channel involving both domains in ssDNA binding. In addition, the dinucleotide captured by sA3G indicates non-catalytic DNA interactions which could be used during the search for deamination target sequences.

7.1.4 MATERIALS AND METHODS

7.1.4.1 Plasmid generation and protein purification

sA3G was generated by combining a soluble NTD[85] and a soluble CTD, namely CTD2(57) with additional substitutions. Catalytically inactive variant of sA3G (sA3G*) was made by substituting E259 with alanine. pGEX6P-1 expression vector carrying sA3G* gene were transformed in BL21 (DE3) cells (Invitrogen). Cells were grown in LB media at 37°C until reaching an optical density 0.5-0.6 at 600 nm. Then temperature was changed to 17°C and cells were induced by adding 0.2 mM IPTG at optical density 0.6-0.8. Cells were further grown overnight at 17°C. All the steps for protein purification were performed at 4°C unless specified. E. coli cells were harvested by centrifugation and resuspended in lysis buffer (50 mM sodium phosphate, pH 7.3, 150 mM NaCl, 25 μM ZnCl2, 2mM DTT and 0.002% Tween-20) and protease inhibitor (Roche, Basel, Switzerland). The suspended cells were disrupted by sonication and then cell debris was separated by centrifugation at 20,000 rpm for 30 min. Supernatant containing desired protein was applied to glutathione-Sepharose (GE Healthcare Life Science) beads, preequilibrated with lysis buffer and agitated for about 2 hours. The beads were washed with PreScission Protease cleavage buffer (50 mM sodium phosphate, pH 7.5, 100 mM NaCl, 2mM DTT and 0.002% Tween-20) and incubated overnight with PreScission protease (GE Healthcare Life Science). GST-free proteins were separated from the beads by centrifugation. Supernatant having the GSTfree protein was further purified by Superdex-75 size exclusion column (GE Healthcare Life Science) in FPLC buffer (20 mM Bis-Tris, pH 6.5, 100 mM NaCl, 10 uM ZnCl2, 2mM DTT and 0.002%

Tween-20) using an AKTA FPLC system. Purity and concentration of the proteins were measured by gel electrophoresis and UV spectroscopy.

7.1.4.2 Crystal Growth and Data Collection

Purified sA3G* was concentrated to about 100 μM using Amicon Ultra-4 (Merck Millipore). Crystallization screening was performed using a commercially available crystallization screen by the sitting drop vapor-diffusion method at 4°C. Crystal drops were set up by mixing 0.3 μl sample and 0.3 μl reservoir solution in a sitting drop 2-well crystallization plate (Molecular Dimension) using a robot, Mosquito Crystal (ttp Labtech). Crystals appeared after two weeks in a condition having 0.1M Tris (pH 8.0) and 5.5% w/v PEG 4000 (MemGold B-12 screen from Molecular Dimensions). Crystals were cryoprotected using reservoir solution containing 15% v/v glycerol and flash frozen in liquid nitrogen. X-ray diffraction data were collected at the Southeast Regional Collaborative Access Team (SER-CAT) 22-ID beam line at the Advanced Photon Source, Argon National Laboratory. The collected diffraction data were indexed, integrated and scaled using the HKL2000 program[250]. The space-group of sA3G* crystals was C2.

7.1.4.3 Structure Determination and analysis

The structure was solved by molecular replacement using human A3G-CTD structure (PDB ID 6BUX, DNA was removed) and rhesus macaque A3G-NTD structure (PDB ID 5K81) as search model at 3.0 Å resolution. The molecular replacement and initial structure refinement were performed using Phaser[251] and Refmec5[252] of CCP4 program suit respectively. Model building of the protein and bound DNA were manually performed using the program Coot[189]. The final model was refined by Phenix[253, 254] to

Rwork/Rfree values of 0.26/0.28. Model was validated with PDB validation tool and Molprobity[255]. Pairwise rms deviation were calculated using Doli[256]. Structural models used for figures were generated using PyMOL.

7.1.4.4 HIV-1 infection and hypermutation assay

Plasmid construction: The plasmids in this study are designated with a "p" while the names of viruses and proviruses generated from these plasmids are not. pHCMV-G expresses the G glycoprotein of vesticular stomatitis virus (VSV-G)[257], pHDV-EGFP is an HIV-1 derived vector that expresses HIV-1 Gag-Pol and enhanced green fluorescent protein (EGFP) but does not express Env, Vif, Vpr, Vpu, or Nef[258]. pVif-HA is a codon-optimized HIV-1 Vif expressing a C-terminal HA epitope tag[259]. pFLAG-A3G expresses wild-type A3G with an Nterminal FLAG epitope tag[260]. pFLAG-A3G was subjected to site-directed mutagenesis to introduce R24A or K180A (Quick Change Lightning Site-Directed Mutagenesis Kit, Agilent Technologies) to create pFLAG-A3G-R24A and pFLAG-A3G-K180A, respectively. The structures of all final plasmids were confirmed by sequencing (Macrogen).

Tissue culture and cell lines: Human embryonic kidney 293T cells (American Type Culture Collection) and TZM-bl cells (obtained through the NIH AIDS Reagent Program [Cat. No. 8129], Division of AIDS, NIAID, NIH: TZM-bl from Dr. John C. Kappes, Dr. Xiaoyun Wu and Tranzyme Inc.[261]) were grown in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum (HyClone) and 1% penicillin-streptomycin stock (penicillin 50 U/ml and streptomycin 50 µg/ml, final concentration; Gibco). TZM-bl cells contain a HIV-1 tat-inducible luciferase reporter

gene that is expressed upon HIV-1 infection and Tat expression. All cells were maintained in humidified 37º C incubators with 5% CO2.

Transfection, virus production and single-cycle infection assays: Transfections were performed using LT1 reagent (Mirus Bio) according to manufacturer's instructions. To generate virus for infection, 4 × 105 293T cells were transfected with pHDV-EGFP (1 μg), with or without pVif-HA (2.5 μg), pHCMV-G (0.25 μg), and variable concentrations of pFLAG-A3G, pFLAGA3G-R24A or pFLAG-A3G-K180A (21, 42, 84, 170 or 340 ng). Equivalent DNA amounts in the transfection mix were maintained by adding pcDNA3.1 empty vector as needed. Forty-eight hours post-infection, virus was harvested, filtered with 0.45-μm filters, and stored at -80 °C. Capsid p24 amounts were determined using the HIV-1 p24 ELISA Kit (XpressBio) according to manufacturer's instructions. Normalized p24 was used to infect 4000 TZM-bl cells in a 96-well plate, and 48-h post-infection, luciferase activity was measured using a 96-well luminometer (LUMIstar Galaxy, BMG LABTECH). Data were plotted as the percent inhibition of luciferase activity compared to the "No APOBEC3G" control. For some experiments, portions of the viral supernatant were spun through a 20% sucrose cushion (15,000 rpm, 2 h, 4° C, in a Sorvall WX80+ ultracentrifuge), concentrated 50-fold, and used in experiments to determine virion encapsidation of FLAG-A3G, pFLAG-A3G-R24A and pFLAG-A3G-K180A by western blotting analysis.

Western blot analysis: Cell lysates were prepared using CelLytic M (Sigma) solution containing Protease Inhibitor Cocktail (Roche), followed by a 10-min, 10,000 × g spin to remove cellular debris. The cell lysates were mixed with NuPAGE LDS sample buffer (Invitrogen) containing β-mercaptoethanol and heated for 5 min at 95 °C.

Samples were analyzed on 4 –20% Tris-Glycine Gels (Invitrogen) using standard western blotting techniques. Proteins were detected with primary antibodies as follows: FLAG-A3G (mouse anti-FLAG M2 monoclonal antibody, 1:5,000 dilution, Sigma catalog #F3165); Vif-HA (mouse anti-HA monoclonal antibody, 1:5,000 dilution, Sigma catalog #H3663); glyceraldehyde 3-phosphate dehydrogenase (GAPDH); rabbit anti-GAPDH antibody, 1:10,000 dilution, Abcam catalog #ab128915). Antibody against HIV-1 p24 (monoclonal, 1:10,000 dilution) was obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: HIV-1 p24 Gag Monoclonal (#24-3) from Dr. Michael Malim (catalog #6458)[261]. An IRDye 800CW-labeled goat anti-rabbit secondary antibody (LI-COR catalog #926-32211) was used at a 1:10,000 dilution to detect rabbit primary antibodies and an IRDye 680-labeled goat anti-mouse secondary antibody (LICOR catalog #926-68070) was used at a 1:10,000 dilution to detect mouse primary antibodies. Protein bands were visualized and quantified using an Odyssey Infrared Imaging System (LICOR).

Hypermutation: Genomic DNA was isolated from infected 293T cells using the QIAamp DNA blood kit (Qiagen). An 896-nt region of reverse transcriptase from integrated proviruses was PCR-amplified with Forward primer (NL4-3 nucleotide position #3296) 5′-GGACAGCTGGACTGTCAATGACATAC-3′ and Reverse primer (NL4-3 nucleotide position #4191) 5′-CTTGTTCATTTCCTCCAATTCCTTTGTGTG-3′. The PCR products were resolved on a 1% agarose gel, the band was gel-eluted using the QIAquick Gel Extraction Kit (Qiagen), and used in the TOPO blunt cloning reaction (Invitrogen). The resulting white colonies after transformation were grown in Luria broth and plasmid DNA was extracted using the NucleoSpin 8 Plasmid Kit (Clontech).

Individual clones were sequenced (Macrogen) and analyzed for the presence of hypermutation using Hypermut (http://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html). Sequenced clones containing single G-to-A changes were not considered hypermutated as single mutations may have resulted from reverse transcriptase or RNA polymerase II errors during viral replication or during PCR and sequencing.

7.1.4.5 Real-time NMR deamination assay

We determined initial rates of deamination reaction by using 1H nuclear magnetic resonance (NMR) spectra as previously reported[89, 183]. 20nt ssDNA substrates, including 5'-AATCCCAATTTTTTTTTTTT (C is the primary target cytidine, 5'- TCCC-polyT) and 5'-AAATCCAATTTTTTTTTTTT (C is the target cytidine, 5'- TCC-polyT) (Integrated DNA Technologies), were used to determine the reaction rates. NMR spectra were acquired at 25ºC on Bruker NMR spectrometers operating at 1H Larmor frequencies of 600 MHz. To test effects of the substitution in the wild-type NTD context, sA3G-NTD was replaced by wild-type A3G-NTD in the sA3G construct called A3G-NTD-CTD2 hereafter. A3G-NTD-CTD2 was used as a template to generate substitution of R24 in NTD, called A3G-NTD-R24A-CTD2 hereafter. NMR samples contained 5% deuterium oxide with 50 nM protein, 200 μM ssDNA substrate, 100 mM NaCl, 0.002% Tween20, 1 mM DTT, 10 μM ZnCl2 and also included 50 mM sodium phosphate adjusted to pH 6.5. Concentration of deamination products were determined from integration of the H5 uracil proton peak as described previously[183]. H5 uracil proton peak was unambiguously assigned by taking 1H spectrum of expected product ssDNA which synthesized separately for each ssDNA substrate. A series of 1H spectra were

measured and the product concentrations as a function of the reaction times were used to determine the initial rate via linear regression. Reaction rates were normalized for the protein concentration and given as reactions per minute. Deamination assays were repeated 3 times independently, and errors in the initial reaction rates were taken as one standard deviation of 3 measurements.

7.1.4.6 Generation of wild-type A3G model

The structure of wild-type A3G was modelled primarily from sA3G* crystal structure (this study). The crystal structures of CTD2 with ssDNA (PDB ID 6BUX) and rhesus macaque A3G-NTD (PDB ID 5K81) were also used in modeling to provide additional structural information for CTD and NTD separately. The wild-type model was first generated through program Modeller 9.15 using basic modeling. The model was then further optimized using the Protein Preparation Wizard[262] from Schrodinger at pH 6.5 and energy minimized with gradually reduced restraints (1000, 5, 0 force constant) on backbone and solute heavy atoms using Desmond[164]. The electrostatic distribution of wild-type A3G was calculated using PDB2PQR[263] server and Pymol with the APBS plugin, and visualized with contour levels positive (+3) and negative (-3).

7.1.5 RESULTS

7.1.5.1 Generation of soluble double-domain A3G variant:

We have overcome relatively poor protein solubility of A3G by generating a soluble variant of A3G. Previously, we generated soluble variants of A3G-NTD and A3G-CTD by rational amino acid substitution, and we were able to determine their NMR and crystal structures[79, 85, 105]. Using these soluble domain variants as starting templates, we

generated a double-domain A3G variant that has improved solubility and homogeneity, which we named soluble A3G or sA3G. sA3G contains extensive substitutions of hydrophobic residues located in NTD loops that compromise encapsidation/HIV-1 restriction, as we previously reported for the soluble A3G-NTD variant[85]. For crystallization, we used a catalytically inactive (E259A) variant, referred to as sA3G* hereafter.

7.1.5.2 Co-crystal structure of sA3G* with a dinucleotide:

We solved the sA3G* crystal structure at 3.0 Å resolution (**Figure 7.1A**) in the C2 space group. We refined the final structure to Rwork/Rfree at 0.26/0.28 respectively (**Table 7.1**). A single sA3G* molecule occupied the asymmetric unit. Unexpectedly, we found a dinucleotide captured by the C-terminal domain of sA3G* (sA3G*-CTD), which we believe emanated from E. coli and remained bound to sA3G* during purification. The dinucleotide was well ordered, and a dideoxy-cytidine structure fit best to the electron density. As the full-length structure of any double-domain APOBEC3 family protein had not been available, to evaluate the sA3G* structure, we compared sA3G*-NTD and sA3G*-CTD separately with available closely related crystal structures of A3G-NTD and A3G-CTD, respectively. As expected, sA3G*-NTD has similar secondary structures (6 helices and 5 strands) as seen in A3G NTD soluble variant[85] (PDB ID: 2MZZ) and rhesus macaque A3G-NTD[99] (PDB ID: 5K81). **Figure 7.1B** shows superimposition of the sA3G*-NTD structure (this study, yellow) and rhesus macaque A3G-NTD (PDB ID: 5K81, green). It is noted that βstrand-2 and helix-2 have partial distortions which may be caused by the absence of Zn2+ ion in sA3G*-NTD probably due to the crystallization condition or amino acid substitutions. We were unable to model residues 55 to 58 of

NTD loop-2 due to poor electron density. The overall backbone structure of sA3G*-NTD is similar to that of rhesus macaque A3G-NTD (PDB ID: 5K81) as indicated by the pairwise root mean square deviation (rmsd), which is 2.2Å. Furthermore, the pairwise rmsd with A3B-NTD (PDB ID: 5TKM) is also 2.2Å indicating that individual domain structures are very well conserved among APOBEC3 proteins. We compared the sA3G*-CTD (**Figure 7.1C**, yellow) with the previously published wild-type A3G-CTD structure[95] (PDB ID: 4ROV) (**Figure 7.1C**, raspberry). Pairwise RMSD is 0.8Å, indicating that the overall backbone structure of our sA3G*-CTD remains almost unchanged(**Figure 7.1C**). Overall, the NTD and CTD in sA3G* preserve secondary and tertiary structural folds of the domain structures of APOBEC3 proteins.

Table 7.1: Crystallographic data collection and refinement statistics.

Data Collection	
Space group	C2
Cell dimensions a, b, c (Å) α, β, γ (°)	 197.42, 42.08, 60.23 90.00, 101.90, 90.00
Resolution (Å)	40.00 – 3.1 (3.21 – 3.10)*
R_{merge} (%)	18.3 (57.7)
R_{meas} (%)	22.4 (73.3)
R_{pim} (%)	12.7 (44.5)
I/σI	5.17 (1.18)
CC1/2	0.942 (0.498)
Completeness (%)	84.1 (79.2)
Redundancy	2.7 (2.1)
Refinement	
Resolution (Å)	35.23 – 3.01 (3.12 -3.01)
No. of reflections	7726
R_{work}/R_{free} (%)	26.24/28.82
No. of atoms	2870
Protein	2787
DNA	37
Ligand/ion	1 (Zn^{2+})
Water	45
B-factor	
Average B-factors (Å^2)	72.4
Protein/DNA	72.9
Ligand/ion	89.8
Water	41.7
R.m.s deviations	
Bond lengths (Å)	0.002
Bond angles (°)	0.44
*Values in parentheses are for the highest-resolution shell.	

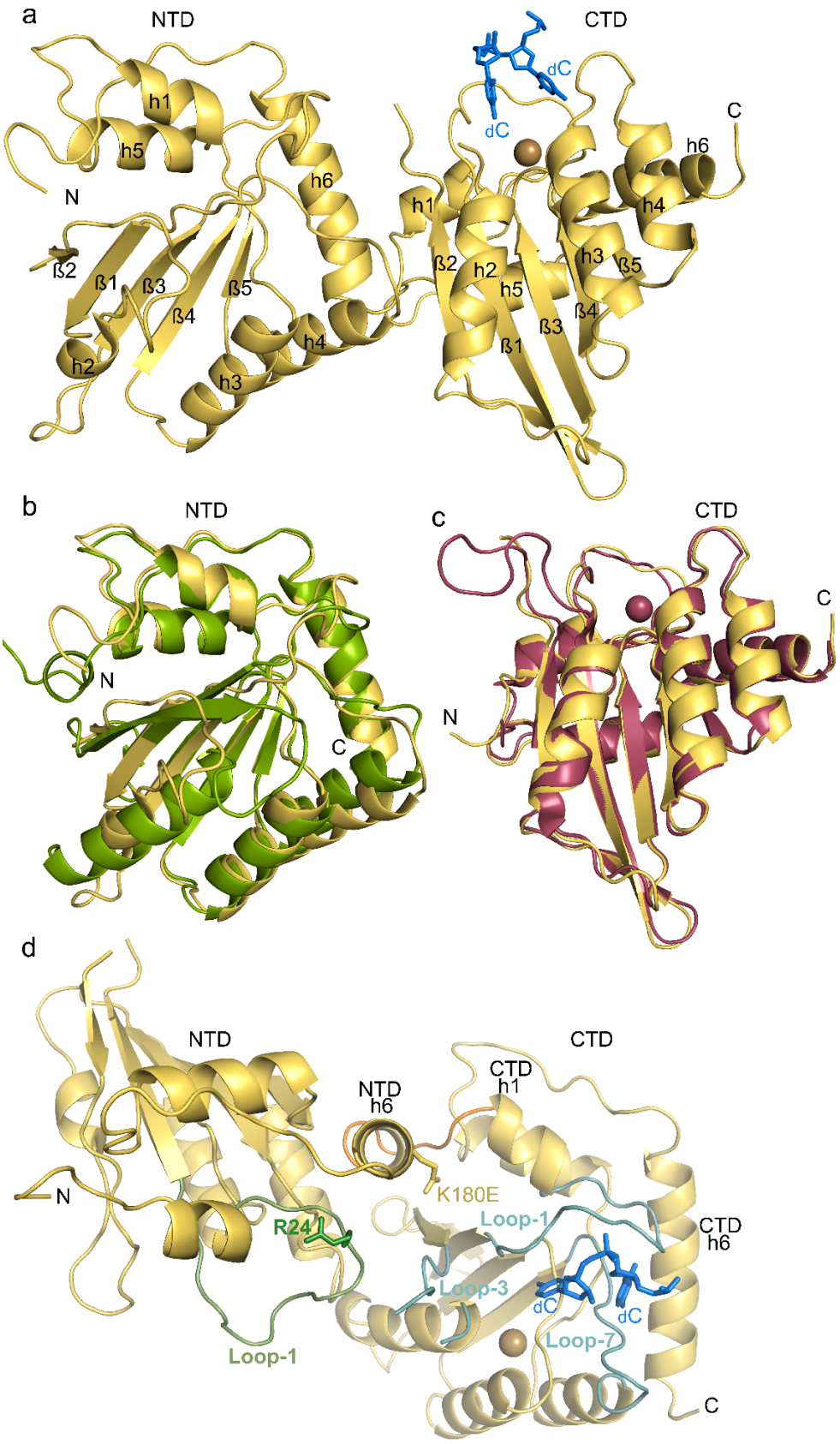
a
NTD
CTD
dC
dC
C
h1
h5
h6
N
h6
h1
h4
ß2
ß1
ß3
ß5
ß4
h4
h2
h3
ß2
h2
h5
h3
ß5
ß4
ß3
ß1
b
NTD
N
C
c
CTD
C
N
d
NTD
CTD
CTD
h1
NTD
h6
N
K180E
R241
Loop-1
CTD
h6
Loop-3
dC
dC
Loop-7
Loop-1
C

Figure 7.1: Co-crystal structure of sA3G* with a dinucleotide.
A) The asymmetric unit contains one protein (yellow) molecule comprising NTD (N-terminal domain) and CTD (C-terminal domain) with a dinucleotide (dC, blue) interacting with CTD. A sphere represents the Zn2+ ion at the catalytic site. N and C indicate the N- and C- terminal ends of the protein. B) Superimposition of the sA3G*-NTD structure (yellow, this study) with rhesus macaque A3G-NTD structure (green, PDB ID: 5K81). C) Superimposition of the sA3G*-CTD (yellow, this study) and wild-type human A3G-CTD (raspberry, PDB ID: 4ROV) structures. D) Co-crystal structure of sA3G* bound to a dinucleotide highlighting the relative orientation of NTD and CTD. Helix-6 (h6) of CTD is oriented almost perpendicular to the h6 of NTD. The linker region between the two structural domains connecting NTD-h6 and CTD-h1is colored orange. Loop-1 of NTD is colored green and potential ssDNA interacting residues R24 of NTD loop-1 and K180E of NTD helix-6 are presented in stick representation and colored green and yellow, respectively. Loops (loop-1, -3 and -7) of CTD which are involved in binding the deamination target sequence 5′-TCCCA^{105} are colored light blue.

7.1.5.3 Relative orientation of NTD and CTD:

The relationship between the relative orientation of each domain and A3G's functions has been a remaining question. The crystal structure of sA3G* shows that NTD is rotated nearly 90° relative to CTD along the major axis of the molecule, which can be seen by examining the orientation of helix-6 from each domain (**Figure 7.1D**). **Figure 7.1D** illustrates the ssDNA binding regions of CTD including loops-1, -3 and -7 (light blue) with bound dinucleotide (blue), which indicates that NTD loop-1 (green) and NTD helix-6 are positioned toward the 3' end of the dinucleotide. Two positively charged residues in NTD, R24 in loop-1 and K180 in helix-6 (K180E in sA3G*) are in good positions to interact with the 3′ side of a substrate ssDNA bound to the catalytic site of CTD (**Figure 7.1D**). The linker region (orange) connecting NTD helix-6 and CTD helix-1 is ordered, and the linker structure is well defined. The interface between NTD and CTD is 1200 Å2 with the linker (H195-D198) and 730 Å2 without the linker residues, which consists of interaction between NTD helix-6 and CTD helix-1 and between NTD loop-1 and CTD loop-3. There is no strong hydrophobic interaction involving aromatic sidechains and/or methyl groups at the inter-domain interfaces. It is noteworthy that there is a hydrogen bond between M188R of NTD and Y219 of CTD. Since M188R is a substituted residue in sA3G, as it is M188 in the wild-type sequence, this hydrogen bonding would not occur in wild-type A3G, although hydrophobic contact between the methyl group of M188 and aromatic group of Y219 is possible.

7.1.5.4 5′-CC dinucleotide captured by CTD:

Unexpectedly, we found a dinucleotide captured by sA3G*-CTD in the crystal structure. Since we did not explicitly add any ssDNA to the crystallization sample, we

believe that the dinucleotide is derived from E. coli and remained throughout purification of the protein. Two deoxy-cytidines fit best to the electron, numbered as C1 and C2 from the 5′ to 3′ direction. The dinucleotide is bound near the catalytic site, but the nucleobases are distant from Zn2+ (**Figure 7.2A**), not positioned for deamination. Therefore, the dinucleotide appears to be in a catalytically inactive position. The interactions between CTD and the dinucleotides are completely different from those observed in our previous co-crystal structure of a soluble A3G-CTD variant (A3G-CTD2) and ssDNA substrate (PDB ID: 6BUX) where the target deoxy-cytosine base was positioned deep in the catalytic pocket right next to Zn2+. **Figure 7.2B** superimposes the sA3G*-CTD:dinucleotide structure (this study, dinucleotide is colored blue) with the A3G-CTD2:ssDNA structure (6BUX, ssDNA is colored pink). Nucleobases of the dinucleotide are distant from Zn2+compared with the ssDNA substrate bound to A3G-CTD2 as the backbone phosphate of C2 from the dinucleotide is 3.7 Å away from the backbone phosphate of C-1 of ssDNA substrate (black double-arrow-headed line in **Figure 7.2B**). C1 of the dinucleotide is positioned between the C-2 and C-1 position of the ssDNA substrate (**Figure 7.2B**). It is noteworthy that although C2 is not in the catalytically active position, the Watson-Crick face of C2 points toward Zn2+ (**Figure 7.2A**). The Watson-Crick face of C1 from the dinucleotide interacts with sA3G*-CTD by two direct hydrogen bonds. The cytosine base carbonyl group makes a hydrogen bond with the mainchain amino proton of V212 from loop-1 and the cytosine base amino group makes a hydrogen bond with the mainchain amino proton of D316 of loop-7. We also observed hydrogen bonding interaction involving the C1 ribose ring O4 and NE1 atom of W211. Additionally, the C1 pyrimidine ring is stabilized by a π–π stacking

interaction with the indole ring of W211 (**Figure 7.2C**). The Watson-Crick face of C2 from the dinucleotide also interacts with sA3G*-CTD by two direct hydrogen bonds formed by the main chain carbonyl of R215 with the N3 proton and amino group of the cytosine base (**Figure 7.2C**). The cytosine base amino group also forms a hydrogen bond with side chain hydroxy-oxygen of T218 through an ordered water molecule. **Figure 7.2D** superimposes ssDNA bound A3G-CTD2 (6BUX) and sA3G*-CTD without DNAs, revealing that amino acids with direct interactions to the target DNA sequence (6BUX, **Figure 7.2D**, light purple) maintain their positions in the dinucleotide bound sA3G*-CTD (this study, **Figure 7.2D**, yellow) but with different sidechain rotamers for some residues.

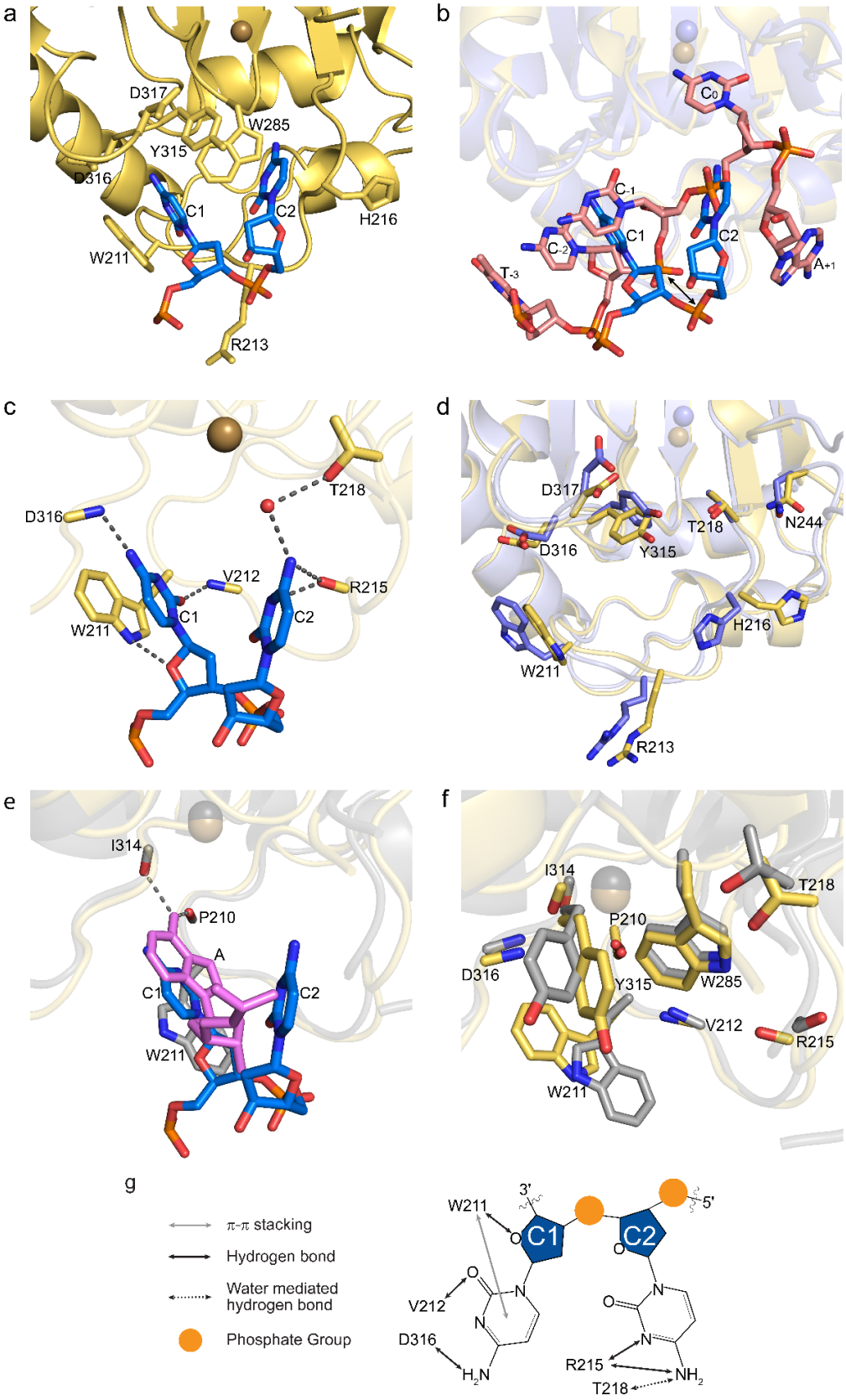
a
D317
W285
Y315
D316
C1
C2
H216
W211
R213
b
C0
C-1
C1
C2
C-2
T-3
A+1
c
T218
D316
V212
R215
C1
C2
W211
d
D317
T218
N244
D316
Y315
H216
W211
R213
e
I314
P210
A
C1
C2
W211
f
I314
T218
P210
D316
Y315
W285
V212
R215
W211
g
π-π stacking
Hydrogen bond
Water mediated hydrogen bond
Phosphate Group
W211
3'
5'
C1
C2
V212
D316
R215
T218

Figure 7.2: Non-catalytic interaction between the 5′-CC dinucleotide and sA3G*. A)The Watson-Crick faces of both cytosines (C1 and C2) position towards loop-7 and the catalytic Zn2+ ion. sA3G* is colored yellow and shows the side chains of key amino acid residues for ssDNA substrate binding as sticks. Atoms in the dinucleotide are colored blue, navy, red and orange for C, N, O and P respectively. B) Superimposition of the sA3G*:dinucleotide structure (this study, dinucleotide is colored blue) with the soluble variant A3G-CTD2:ssDNA structure (PDB ID: 6BUX, ssDNA is colored pink) comparing positions of nucleotides in the two structures. The double arrow-headed line points to the backbone phosphorous atoms of C2 of the dinucleotide and C-1 of the A3G-CTD2:ssDNA structure. C) Detailed interactions of the dinucleotide with amino acid residues of sA3G*-CTD. C, N, and O atoms are colored yellow, blue, and red, respectively, for amino acid residues of the protein. Water molecule is shown as red spheres, and Zn2+ is shown as a sand colored sphere. Dotted lines indicate hydrogen bonds. D)Superimposition of the sA3G*-dinucleotide structure (this study, sA3G* is colored yellow) with the A3GCTD2: ssDNA structure (PDB ID: 6BUX, A3G-CTD2* is colored light purple) highlighting sidechain conformational differences of key amino acid residues for ssDNA substrate binding. Side chains of the amino acid residues are shown as sticks. DNAs are not shown. E) Superimposition of the sA3G*-dinucleotide structure (this study, dinucleotide is colored blue) with the adenine bound A3G-CTD structure (PDB ID: 6BWY, adenine is colored magenta) showing overlap of the adenine base with the C1 base of dinucleotide. Dotted lines indicate hydrogen bonds between the adenine amino group and two mainchain carbonyls of A3G-CTD (colored gray). F) Superimposition of the sA3G*-dinucleotide structure (this study, sA3G* is colored yellow) with the adenine bound A3G-CTD structure (PDB ID: 6BWY, A3G-CTD is colored gray) highlighting sidechain conformational differences of amino acid residues that bind the dinucleotide and/or the adenine. Amino acid residues are shown as sticks. G) Summary of the interactions between sA3G* and the dinucleotide.

Ziegler et al. reported a crystal structure of A3G-CTD bound to a non-preferred adenine nucleotide near the catalytic site that provided insights into non-specific interactions between A3G-CTD and DNA[108] (PDB ID 6BWY). Superimposition of this adenine bound A3G-CTD (6BWY, magenta) with dinucleotide-bound sA3G*-CTD (this study, blue) shows partial overlap of the adenine base with the C1 base of the dinucleotide (**Figure 7.2E**), revealing that these two nucleotides bind to a similar position of A3G-CTD. Nonetheless, interactions between C1 and sA3G*-CTD are significantly different from that of the adenine as the amino group of adenine forms a hydrogen bond to the backbone carbonyl of P210 and I314 (**Figure 7.2E**), whereas the C1 nucleobase of the dinucleotide forms hydrogen bonds with V212 and D316 (**Figure 7.2C**). These differences in hydrogen bonding are likely a reflection of the nucleobase type difference, while both cytosine and adenine bases are stabilized by partial π–π stacking interaction with indole ring of W211 (**Figure 7.2C,E**). **Figure 7.2F** superimposes amino acid residues involved in either adenine (6BWY, gray) or dinucleotide (this study, yellow) interaction (DNAs are not shown), showing that A3G-CTD adjusts for nucleobase type differences by changing the positions of R215 and T218 and rotating the sidechains of W211 and Y315.

7.1.5.5 R24 of NTD is important for deamination of 5'-CC motifs in virions:

The sA3G*: dinucleotide co-crystal structure (this study) suggests that R24 and/or K180 may interact with DNAs located on the 3' side of a substrate ssDNA (**Figure 7.1D**). We sought to determine whether these structure-based hypotheses for A3G-NTD interaction with ssDNA are valid and influence cytidine deamination. We thus determined the influence of R24A and K180A substitutions in wild-type A3G, called

A3G-R24A and A3G-K180A hereafter, on inhibition of HIV-1 infectivity and hypermutation of provirus. To determine the effect of A3G-R24A and A3G-K180A on HIV-1 infectivity, VSVG-pseudotyped virions were prepared in the absence of A3G or in the presence of increasing amounts of wild-type A3G, A3G-R24A, or A3G-K180A expressing plasmids (**Figure 7.3A**). Western blotting analyses of the transfected cells (in the absence of Vif) indicated similar steady-state expression levels of the wild-type and mutant A3Gs. Infectivity of the virions produced from the transfected cells was determined by infection of TZM-bl cells (**Figure 7.3B**). The results indicated that expression of wild-type A3G and A3GK180A potently inhibited HIV-1 infectivity in a dose-dependent manner, but expression of A3GR24A severely reduced the A3G antiviral activity; transfection with 340 ng of wild-type A3G or A3G-R24A reduced virus infectivity to 0.5% or 47.6% of the “no A3G” control, respectively. Previous studies regarding RNA-binding of human A3G have reported that the A3G-R24A mutation reduces virion incorporation of A3G(86,87). We compared the virion incorporation of wild-type A3G in the presence of increasing amounts of A3G plasmid to the virion incorporation of A3G-R24A and A3G-K180A in the presence of 340 ng of the plasmids (**Figure 7.3C**). The results confirmed that the A3G-R24A mutant was highly defective in virion incorporation. The amount of A3G-R24A that was packaged into virions when 340 ng of plasmid was transfected (5% of 340 ng of wild-type A3G) was similar to the amount packaged when 42 ng or 84 ng of wild-type A3G was transfected (2.6% and 6.3% of 340 ng of wild-type A3G, respectively). In contrast, the amount of A3G-K180A mutant that was packaged into virions when 340 ng of plasmid was transfected was only reduced about two-fold (44% of 340 ng of wild-type A3G). It was noted that when

infectivity was compared in conditions where similar amounts of A3Gs were in virions (84 ng of wild-type A3G vs 340 ng of A3G-R24A, and 170 ng of wild-type A3G vs 340 mg of A3G-K180A), A3G-R24A showed more infectivity than wild-type A3G (47.6% for A3G-R24A, and 9.2% for wild-type A3G), while A3G-180A showed similar or less infectivity (0.6% for A3G-K180A, and 2.5% for wild-type A3G). Therefore, these infectivity assays indicate that the R24A substitution compromises its restriction function against HIV-1 infection. The A3G-R24A and A3G-K180A mutants remained sensitive to Vif-mediated degradation (**Figure 7.4A**) and in the presence of Vif, virus infectivity was restored to wild-type levels (**Figure 7.4B**).

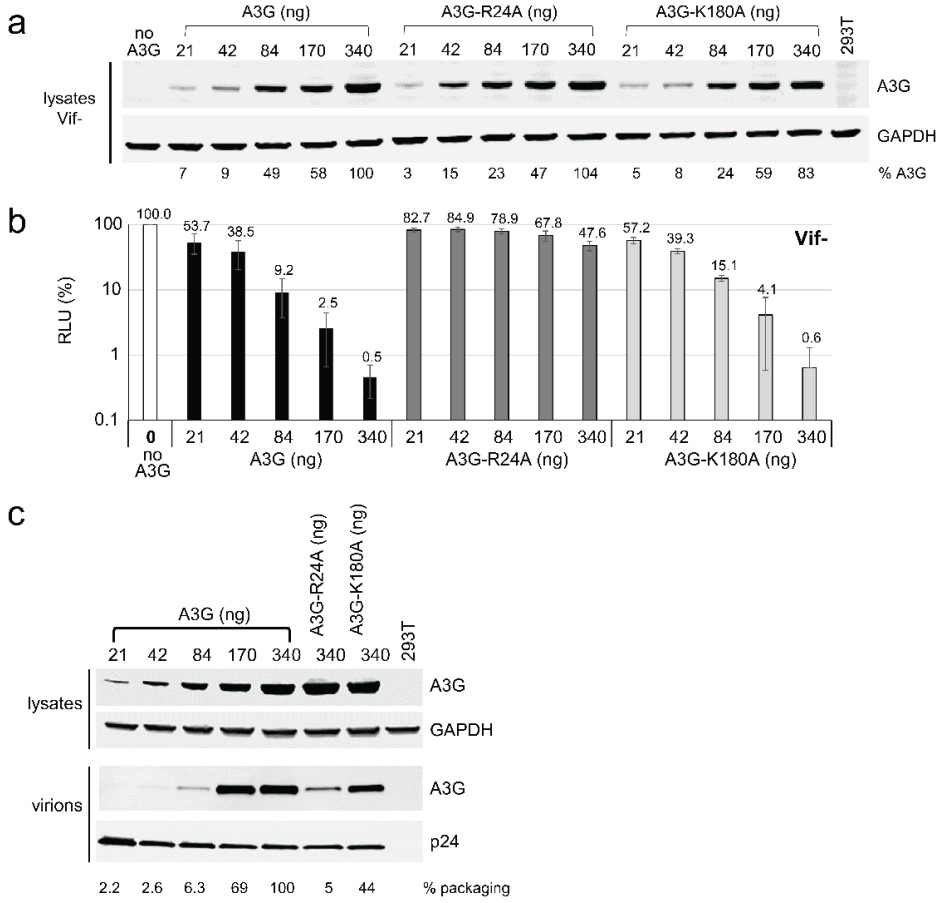

Figure 7.3: Antiviral activity and encapsidation of A3G-R24A and A3G-K180A. A) Representative western blot showing 293T cells co-transfected with increasing amounts of Nterminal FLAG-tagged wild-type A3G, A3G-R24A or A3G-K180A (21, 42, 84, 170, 340 ng), HDV-EGFP, and VSV-G in the absence of Vif. Average A3G expression from 4 independent experiments relative to 340 ng A3G (set to 100%) is shown after adjusting for GAPDH band intensities. B) Single-cycle infectivity of normalized p24 capsid amounts harvested from transfected cells in part (a) were assayed in TZM-bl target cells. Infectivity is proportional to relative light units (RLU) produced by induction of luciferase expression (normalized to the no A3G control). Error bars represent the standard deviation from three independent experiments. C) Representative western blot of 293T cell lysates and virions produced from 293T cells cotransfected with increasing amounts of N-terminal FLAG-tagged wild-type A3G (21, 42, 84, 170, 340 ng), A3G-R24A (340ng), or A3G-K180A (340 ng), HDV-EGFP, and VSV-G in the absence of Vif. Relative A3G expression was normalized to GAPDH levels in cell lysates and to capsid p24 levels in virions. The average packaging efficiency of A3G, A3G-R24A and A3GK180A was determined by dividing the amount of A3G encapsidated in the virions by the amount of A3G expressed in the lysates, and further normalized to A3G 340 ng (100%) from two independent experiments.

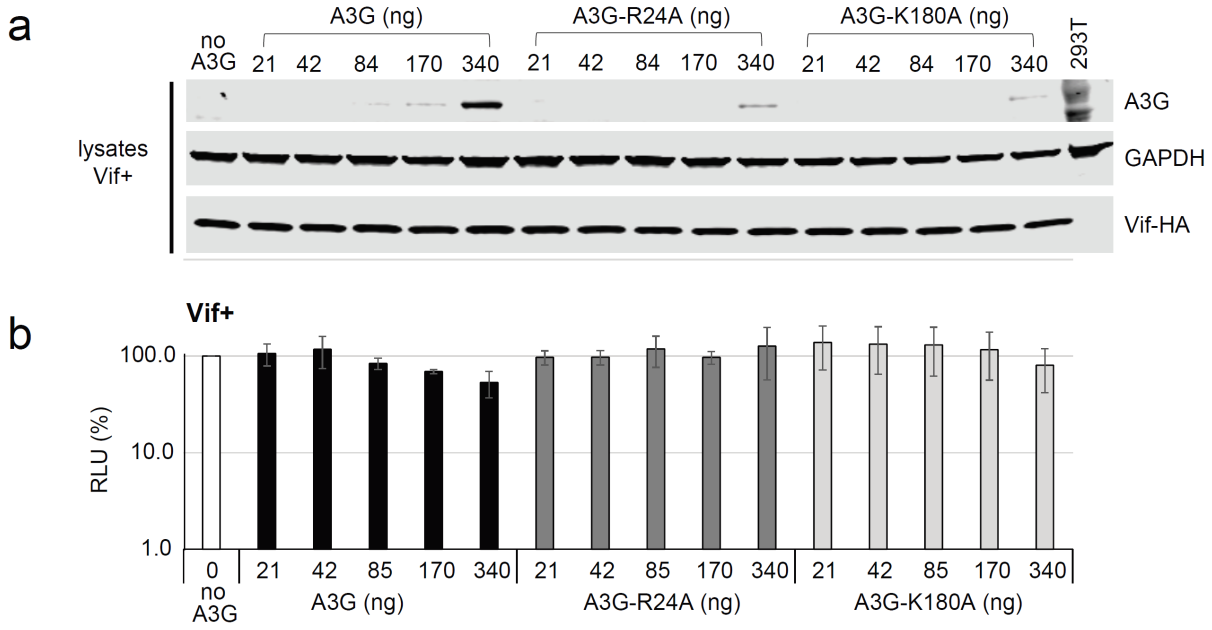

Figure 7.4: Antiviral activity of A3G-R24A and A3G-K180A in the presence of Vif. A) Representative western blot showing 293T cells co-transfected with increasing amounts of N-terminal FLAG-tagged wild-type A3G or A3G-R24A or A3G-K180A (21, 42, 84,170, 340 ng), HDV-EGFP, and VSV-G in the presence of Vif-HA. Absence of A3G signal (FLAG signal) in the Vif+ lanes indicates that mutants A3G-24A and A3G-180A were efficiently degraded by Vif. B) Single-cycle infectivity of normalized p24 capsid reflects the average relative light units (RLU) normalized to the no A3G control. Error bars represent the standard deviation from three independent experiments.

Next, we compared the effects of the A3G-R24A and A3G-K180A mutations on selection of substrate 5′-CC sites in the minus strand of the viral DNA by quantifying the G-to-A mutations in the plus strand of the viral DNA at 5′-GG sites in an 896-nt sequence of HIV-1 reverse transcriptase. Hypermutated proviral DNA sequences were obtained from cells infected with virions produced in the presence of 42, 84 or 340 ng of A3G-WT, 340 ng of A3G-R24A, or 340 ng of A3G-K180A. A comparison of the frequency distribution of mutations at 5′-GGn sites in the plus strand of viral DNA (n can be any deoxy-nucleotide, GGn hereafter) in hypermutated clones obtained from virions produced in the presence of 42, 84, and 340 ng of A3G-WT indicated that the frequencies observed with 340 ng of A3G-WT were different from those observed with 42 ng and 84 ng of A3G-WT (chi square P-values <0.00001), indicating that the mutation frequency distribution is highly dependent on the amount of A3G packaged into virions. We reasoned that virions that contained high levels of A3G would efficiently deaminate and deplete the preferred target sequences such as GGg and increase deamination of the less preferred target sites such as GGa and GGt. As a result, hypermutated clones with high levels of G-to-A mutations would display different frequencies of mutations relative to the +2 nucleotide (the “n” nucleotide in “GGn” sites within the positive strand of provirus). In order to avoid excessive deamination and depletion of a target sites, sequences with high levels of G-to-A mutations were removed from the sequence analysis of cells infected with 84 ng of A3G and 340 ng of A3G-R24A or A3G-K180A until the average number of G-to-A mutations per clone matched that observed for cells infected with 42 ng of A3G (12.5 mutations at GG sites/clone). A comparison of the proportion of G-to-A mutations at GGn positions in the

plus strand of viral DNA, when the virions were produced in the presence of 42 ng of A3G and 84 ng of A3G, showed that the frequencies of mutations at the GGn positions were not different (**Table 7.2**; chi-square analysis using 2 x 4 contingency table; P = 0.8131). A3G-K180A showed increased proportion for GGg and decreased proportion for GGt, but overall differences in the frequencies of mutations at the GGn positions for A3G-K180A were statistically not significant as P value is 0.1055 (**Table 7.2**). However, the frequencies of mutations at the GGn positions were significantly different when the virions were produced in the presence of A3G-R24A (P = 0.0016). Specifically, the frequency of mutations were 18% lower at GGa sites ([0.28/0.34] x 100% = 82%) and 13% higher at GGg sites ([0.65/0.57] x 100% = 113%) for virions produced in the presence of A3G-R24A compared to virions produced in the presence of 42 ng or 84 ng of A3G-WT (Fisher's exact test; P < 0.001; **Table 7.2**). Thus, the R24A mutation in the N-terminal domain of A3G altered the efficiency of cytidine deamination in virions, by increasing the relative efficiency of mutations at the GGg sites and decreasing the relative efficiency of mutations at the GGa sites.

Table 7.2: Comparison of deamination frequencies at GG sites with different +2 nucleotides.

A3G	No. Hypermut. Seq. Selected/ Total[a]	Avg. GG mut./ seq.[b]	Total GG Mut.[c]	GGa Mut. (freq.)[d]	GGc Mut. (freq.)	GGg Mut. (freq.)	GGt Mut. (freq.)	Chi Square *P*-value[e] vs. WT (42 ng)
WT (42 ng)	121/121	12.5	1510	511 (0.34)	11 (0.01)	867 (0.57)	121 (0.08)	
WT (84 ng)	101/154	12.5	1261	431 (0.34)	13 (0.01)	722 (0.57)	95 (0.08)	0.8131
R24A (340 ng)	94/110	12.5	1171	328 (0.28)	6 (0.01)	759 (0.65)	78 (0.07)	0.001631
K180A (340 ng)	23/92	12.6	289	92 (0.32)	2 (0.01)	182 (0.63)	13 (0.04)	0.1055

[a] Number of hypermutated sequences that were selected from the total number of hypermutated sequences.

[b] Average number of mutations at GG sites per clone; the sequence analyzed has 60 GG sites.

[c] Total number of G-to-A mutations at GG sites.

[d] Total number of G-to-A mutations at the indicated GGn sites and their frequency.

[e] Chi square statistic and *P*-value was calculated using a 2 x 4 contingency table.

Since the A3G R24A substitution significantly changed the proportion of G-to-A mutations in different GGn contexts in our mutation assays, we further tested whether the R24A substitution affects the catalytic efficiency in vitro using an NMR based deamination assay. To test effects of the substitution in the wild-type NTD context, sA3G-NTD was replaced by wildtype A3G-NTD in the sA3G construct called A3G-NTD-CTD2 hereafter as CTD was the soluble variant CTD2 that had been used for co-crystal structure with ssDNA substrate(57) and the sA3G* structure (this study). We used 20 nt ssDNA, including 5'-AATCCCAATTTTTTTTTTTT (5'-TCCC-polyT, C indicates the primary deamination site) and 5'-AAATCCAATTTTTTTTTTTT (5'- TCC-polyT) as substrates for the assay. **Table 7.3** summarizes initial speed of reaction of A3G-NTD-CTD2 and A3G-NTD-R24A-CTD2 for each substrate (deamination data is provided in **Figure 7.5**). The A3G R24A mutation reduced reaction speed by 35% for 5'- TCCC-polyT and 50% for 5'- TCC-polyT, which supports our hypermutation results, as we observed greater reduction of mutation frequency for GGa in the plus strand of viral DNA (equivalent to 5'-TCC in the minus strand of viral DNA that is the physical substrate for the deamination catalyzed by A3G) by A3G-R24A compared to wild-type A3G. Together, in vivo hypermutation assay and in vitro deamination assay suggested that R24 may play an important role in ssDNA substrate binding.

Table 7.3: Comparison of deamination speeds.
Deamination rates for the 3′-cytidine of the 5′-TCCC-polyT and 5′-TCC-polyT substrates are given as reactions/minute for A3G-NTD- CTD2 (WT) and A3G-NTD-R24A-CTD2 (R24A).

	5′-TCCC-polyT	5′-TCC-polyT
WT	4.6 ± 0.2	2.2 ± 0.2
R24A	3.0 ± 0.2	1.1 ± 0.2

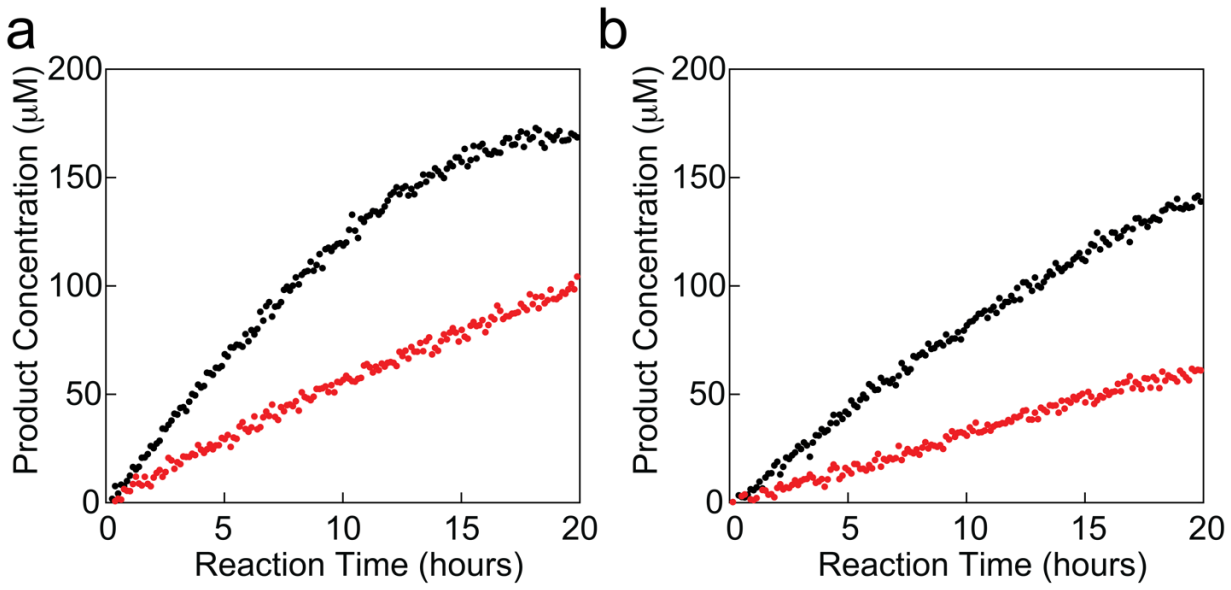

Figure 7.5: NMR based deamination assays.
^{1}H NMR detects the formation of the 3′-cytidine deamination products of 5′-TCCC-polyT (black) and 5′-TCC-polyT (red) as a function of time for a) A3G-NTD-CTD2 and b) A3G-NTD-R24A-CTD2.

7.1.5.6 Wild-type A3G structural model

To gain further structural insights in the wild-type A3G context, we generated a structural model of wild-type human A3G (wtA3G) based on the sA3G*-dinucleotide co-crystal structure (this study). The wtA3G structure shows a possible ssDNA-binding channel formed by NTD loop-1 and CTD loop-3 (**Figure 7.6A**, orange dotted line). Using this channel, NTD would interacts with DNAs located on the 3' side of a deamination target sequence, while CTD interacts with the target cytidine and adjacent nucleotides. A3G likely interacts with ssDNA in this orientation during the search for target sequences because this orientation of ssDNA is required for specific binding of the target sequence and catalysis[105]. Our deamination and hypermutation data (this study) suggest that ssDNA may interact with R24 located in the proposed ssDNA-binding channel (**Figure 7.6A**, green stick). This channel is deep and positively charged as shown in **Figure 7.6B**, which may attract the negatively charged phosphate backbone of ssDNA.

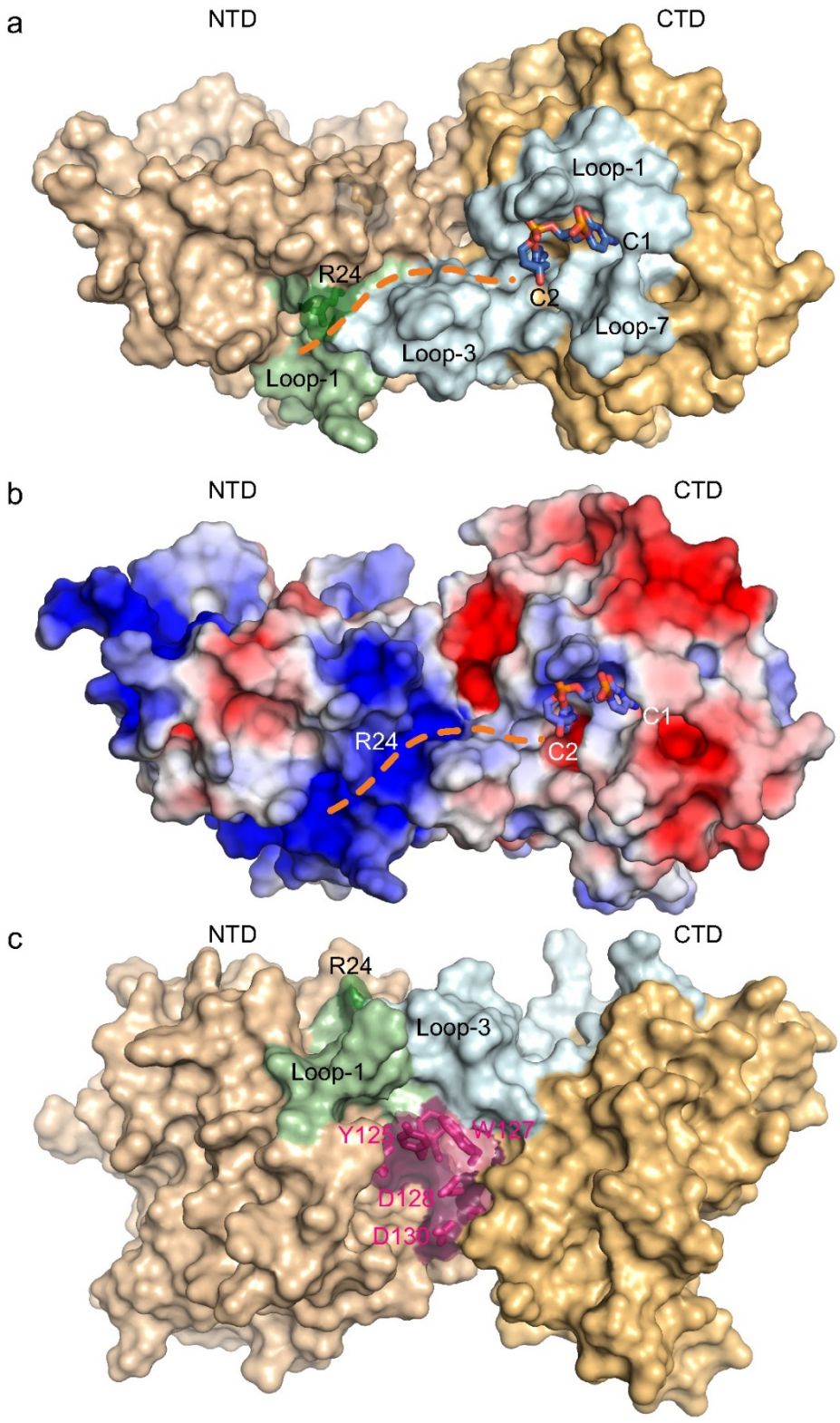

Figure 7.6: Surface representation of wild-type full-length A3G with a binding pathway for ssDNA modeled.

A) The dinucleotide bound to CTD is shown as sticks, and C, N, O, and P atoms are colored blue, dark blue, red, and orange, respectively. Loops (loop-1,-3 and -7) of CTD are colored light blue. NTD loop-1 is colored light green and potential ssDNA interacting residue R24 in loop-1 is presented in stick representation within a transparent surface and colored deep green. An orange dashed line indicates a possible ssDNA binding channel. B) Electrostatic surface distribution of the wild-type A3G structure to illustrate positive (+3, blue) and negative (-3, red) regions. The orientation of the molecule is the same as a, and an orange dashed line indicates the possible ssDNA binding channel. C) Potential HIV-1 Vif binding surface of A3G. The view of C is almost 90° rotated along the major axis of molecule from the view of A/B. Residues 124-YYFWDPD-130 of NTD are presented as sticks under a transparent surface and colored pink.

7.1.6 DISCUSSION

Our crystal structure of double-domain A3G revealed insights into the domain arrangement and enabled us to propose a structural mechanism for the involvement of catalytically inactive NTD in A3G function. The dinucleotide bound to sA3G* also provided additional insights regarding non-catalytic DNA interaction by A3G. A 5′-CC dinucleotide is positioned near the catalytic Zn2+, but lacks specific interactions required for catalysis. Interestingly, this dinucleotide and the adenine found in the previously published co-crystal structure of A3G-CTD and ssDNA[108] occupy a similar position in the A3G-CTD. As proposed by Ziegler et al., their adenine and our dinucleotide may be showing protein-DNA interactions when A3G is in search for deamination target sequences. During the search for target sequences, residues in loop-1 and loop-7 of A3G-CTD may interact with DNAs in variable ways, which enables identification of nucleobase types. This hypothesis is well supported by our most recent results using NMR[115] that suggested loops-1, -3 and -7 of A3G-CTD to be dynamically involved in non-specific DNA interactions until a deamination target sequence is found, which then causes W211 and H216 of loop-1 to latch on tightly to the target sequence providing the target cytidine stability during catalysis.

The crystal structure and wild-type structural model revealed a positively-charged channel that may be involved in ssDNA interactions. The hypermutation results of A3G R24 substitution provided an interesting observation that the proportion of mutations at GGa in the plus strand of viral DNA (equivalent to 5′-TCC in substrate ssDNA) was decreased while the proportion of mutations at GGg (equivalent to 5′-CCC in substrate ssDNA) was increased (**Table7.1.2**). Very few mutations occurred in GGt and GGc

motifs (equivalent to 5′-ACC and 5′-GCC in substrate ssDNA, respectively), and their proportions were not significantly different from wtA3G (**Table7.1.2**). It is plausible that the R24-DNA interaction becomes more significant for the 5′-TCC, 5′-ACC and 5′-GCC substrates because these substrates are missing specific hydrogen bonds involving the additional 5′ C existing in the 5′-CCC substrate[105]. R24 has lured attention since previous studies found that substitution of R24 reduces encapsidation of A3G suggesting its involvement in RNA interaction and oligomerization of A3G[264-266], although Y124, F126 and W127 were found to play central roles for those events[90, 248, 267]. Because these hydrophobic residues and oligomerization are not essential for the deamination activity of A3G[90, 268], the RNA-binding region(s) of A3G appear to be different from the ssDNA-binding channel suggested by this study (**Figure 7.1.6A**). Indeed, Huthoff et al.[264] and Xiao et al.[99] have proposed A3G homodimer models in which a single-stranded RNA binds at dimer interfaces including R24, R30, Y124, W127 and R136.

Antiviral activity of human A3G can be neutralized by HIV-1 Vif, but not by Vif from Simian Immunodeficiency Virus which infects the African green monkey (SIVagm), generating a barrier for cross-species transmission[269-273]. D128, P129 and D130 have been identified as key residues for Vif-induced degradation of human A3G as substitutions of these residues resulted in abrogated degradation[269-273]. Our wtA3G structural model shows that loop 7 of NTD, including 124-YYFWDPD-130, is exposed on a surface rotated by nearly 90° from the ssDNA binding surface (**compare Figure 7.6A,C**), and that these residues are accessible for Vif. Nevertheless, a structure of the A3G:Vif complex is required to reveal the domain orientation and atomic-level

interactions between A3G and Vif because A3G may or may not keep the same domain orientation found in our sA3G*:dinucleotide co-crystal structure when it interacts with Vif.

7.2 Appendix II: To find the first-in-class inhibitors against A3s

7.2.1 PREFACE

The following work was performed to find the first-in-class inhibitors against A3s to benefit anti-cancer therapeutics. I have performed virtual screening of small molecules/fragments libraries against A3s. I have optimized the novel fluorescence-based product release assay for measuring A3 deamination activity with the assistance of Ellen A. Nalivaika and Paul Thompson. I also performed initial crystallization trials for potential hits from virtual screening. However, as A3s requires a minimum of 5-mer DNA oligo for deamination[183, 201], finding small molecules that inhibit A3 is extremely challenging. Hence recently, our lab started the design of oligonucleotide-based inhibitors (OBI) for A3s. I performed the molecular modeling and molecular dynamics simulations for OBI-bound A3 structures.

7.2.2 METHODS AND RESULTS

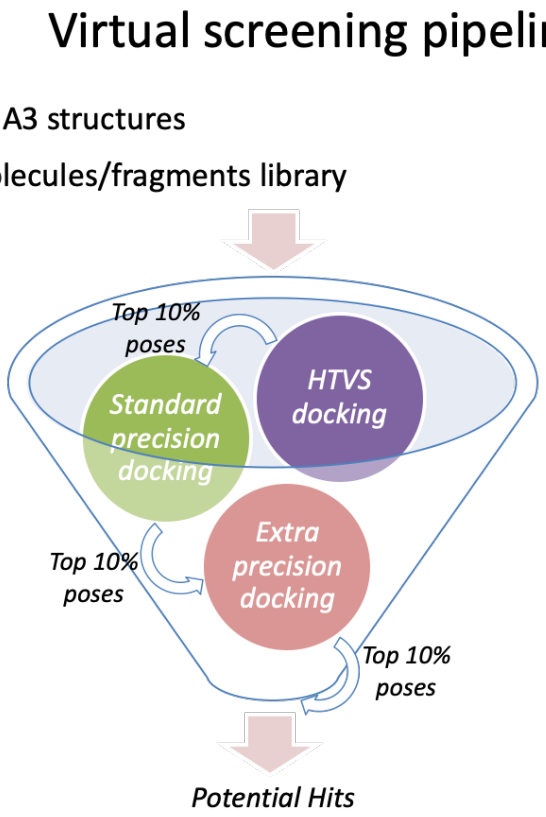

Figure 7.7: Virtual screening of small molecules/fragments against A3A/B/G. NIH Molecular Libraries Small Molecule Repository (MLSMR), AnalytiCon The FRGx library and 2 Diversity sets of small molecules from UMass Small Molecule Screening Facility (SMSF) have been used as virtual screening libraries against A3A, A3B and A3G. The screenings were performed follow the pipeline in **Figure7.2.2.1** using Glide from Schrodinger.

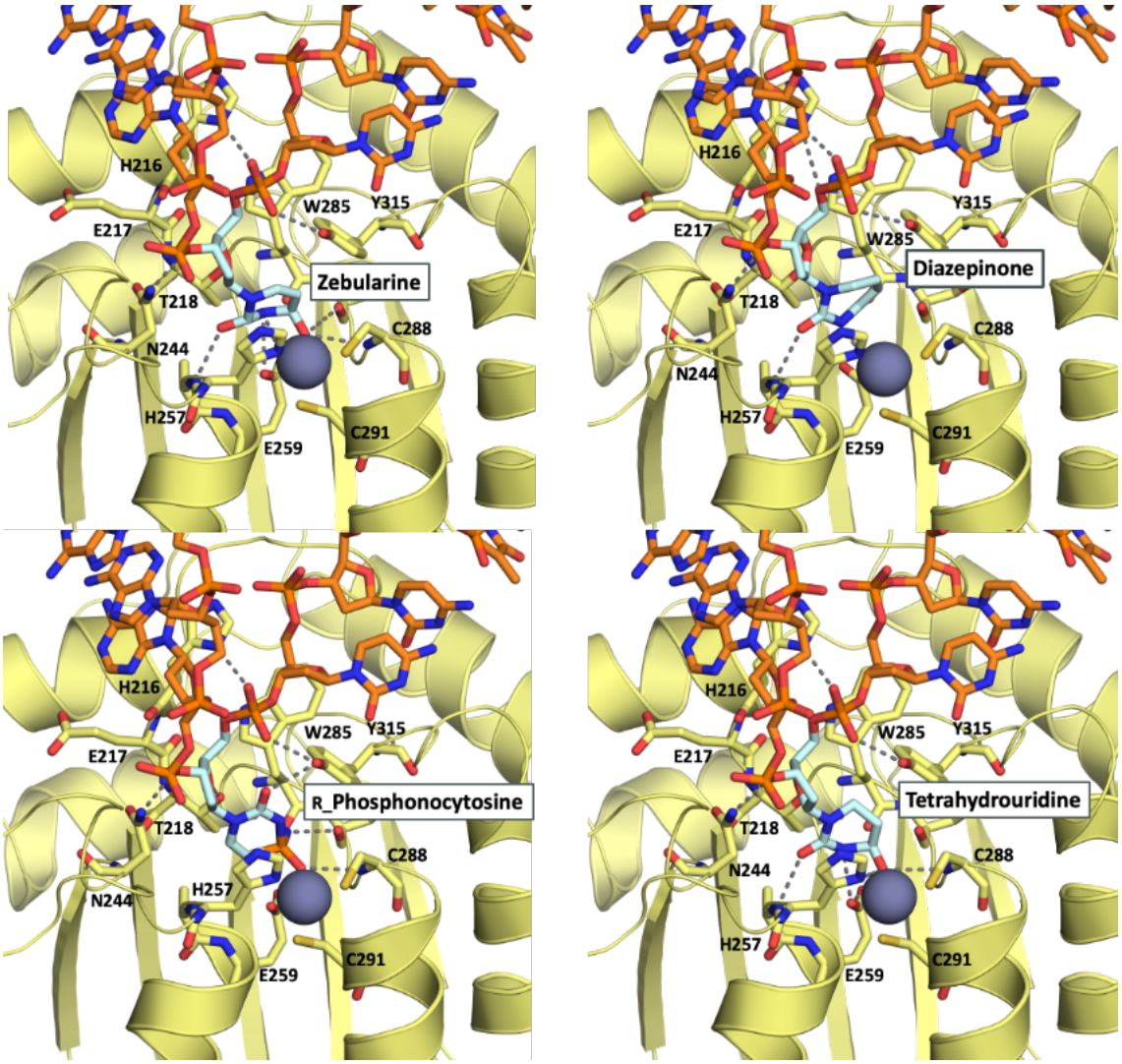

Figure 7.8: Molecular modeling for OBI-bound A3 structures.

7.3 Appendix III: Structure-based Vif fitness study

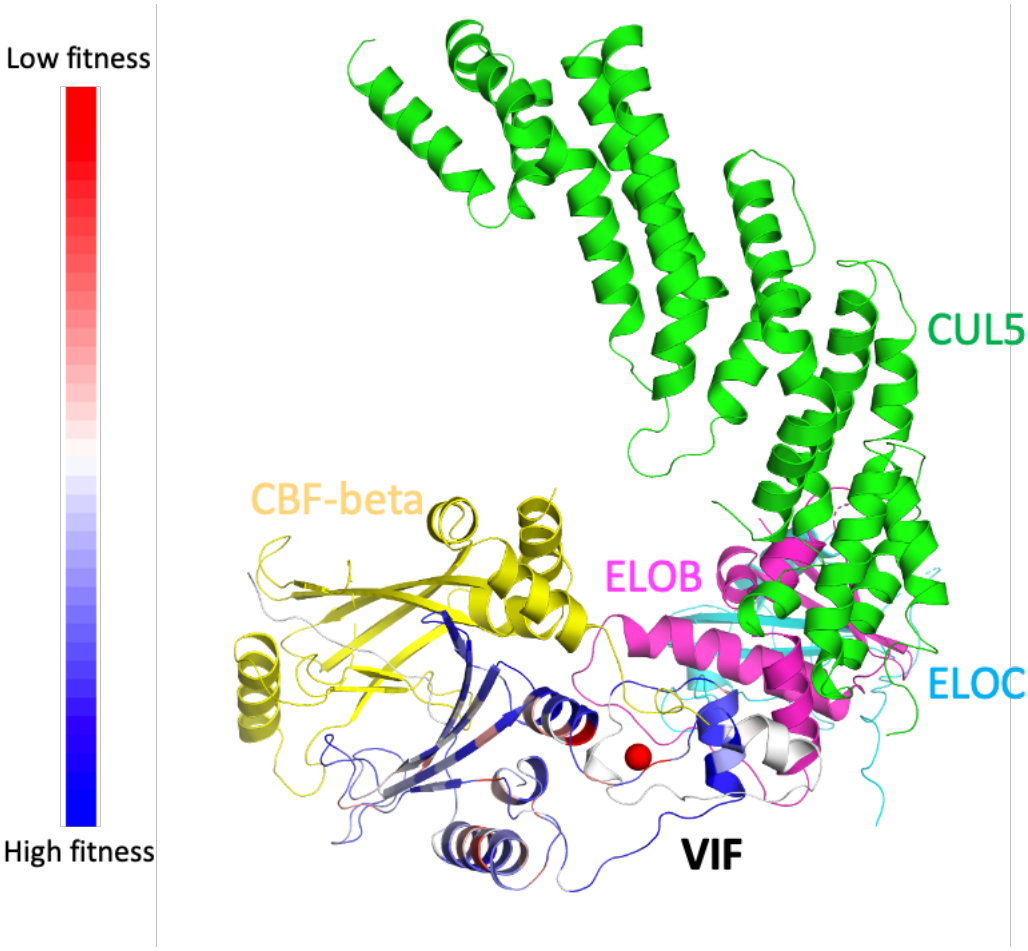

Figure 7.9: The viral fitness date plotted on Vif structure.

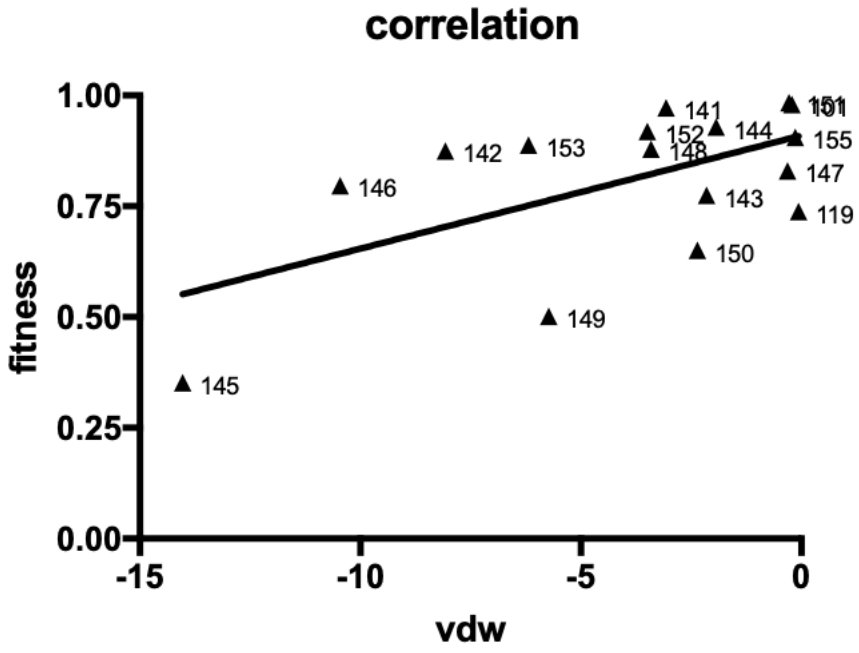

Figure 7.10: The correlation plot between vdw interactions and fitness at Vif-ELOB interface.
There is correlation between vdw interactions and fitness data at Vif-ELOB interface. The interface residues are 119A, 120I, 141K, 142V, 143G, 144S, 145L, 146Q, 148L, 149A, 150L, 152A and 153L.

8 REFERENCES

1. Betts, L.; Xiang, S.; Short, S. A.; Wolfenden, R.; Carter, C. W., Jr., Cytidine deaminase. The 2.3 A crystal structure of an enzyme: transition-state analog complex. *J. Mol. Biol.* **1994,** *235* (2), 635-56.
2. Jarmuz, A.; Chester, A.; Bayliss, J.; Gisbourne, J.; Dunham, I.; Scott, J.; Navaratnam, N., An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics* **2002,** *79* (3), 285-96.
3. Wedekind, J. E.; Dance, G. S.; Sowden, M. P.; Smith, H. C., Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends Genet.* **2003,** *19* (4), 207-16.
4. Conticello, S. G.; Thomas, C. J.; Petersen-Mahrt, S. K.; Neuberger, M. S., Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Mol Biol Evol* **2005,** *22* (2), 367-77.
5. LaRue, R. S.; Andresdottir, V.; Blanchard, Y.; Conticello, S. G.; Derse, D.; Emerman, M.; Greene, W. C.; Jonsson, S. R.; Landau, N. R.; Lochelt, M.; Malik, H. S.; Malim, M. H.; Munk, C.; O'Brien, S. J.; Pathak, V. K.; Strebel, K.; Wain-Hobson, S.; Yu, X. F.; Yuhki, N.; Harris, R. S., Guidelines for naming nonprimate APOBEC3 genes and proteins. *J. Virol.* **2009,** *83* (2), 494-7.
6. Green, A. M.; Weitzman, M. D., The spectrum of APOBEC3 activity: From anti-viral agents to anti-cancer opportunities. *DNA Repair (Amst)* **2019,** *83*, 102700.
7. Rogozin, I. B.; Basu, M. K.; Jordan, I. K.; Pavlov, Y. I.; Koonin, E. V., APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. *Cell Cycle* **2005,** *4* (9), 1281-5.
8. Johansson, E.; Mejlhede, N.; Neuhard, J.; Larsen, S., Crystal structure of the tetrameric cytidine deaminase from Bacillus subtilis at 2.0 A resolution. *Biochemistry* **2002,** *41* (8), 2563-70.
9. Ireton, G. C.; Black, M. E.; Stoddard, B. L., The 1.14 A crystal structure of yeast cytosine deaminase: evolution of nucleotide salvage enzymes and implications for genetic chemotherapy. *Structure* **2003,** *11* (8), 961-72.
10. Ko, T. P.; Lin, J. J.; Hu, C. Y.; Hsu, Y. H.; Wang, A. H.; Liaw, S. H., Crystal structure of yeast cytosine deaminase. Insights into enzyme mechanism and evolution. *J. Biol. Chem.* **2003,** *278* (21), 19111-7.
11. Xie, K.; Sowden, M. P.; Dance, G. S.; Torelli, A. T.; Smith, H. C.; Wedekind, J. E., The structure of a yeast RNA-editing deaminase provides insight into the fold and function of activation-induced deaminase and APOBEC-1. *Proc. Natl. Acad. Sci. U. S. A.* **2004,** *101* (21), 8114-9.
12. Muramatsu, M.; Kinoshita, K.; Fagarasan, S.; Yamada, S.; Shinkai, Y.; Honjo, T., Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **2000,** *102* (5), 553-63.
13. Teng, B.; Burant, C. F.; Davidson, N. O., Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science* **1993,** *260* (5115), 1816-9.
14. Kane, J. P.; Hardman, D. A.; Paulus, H. E., Heterogeneity of apolipoprotein B: isolation of a new species from human chylomicrons. *Proc. Natl. Acad. Sci. U. S. A.* **1980,** *77* (5), 2465-9.

15. Innerarity, T. L.; Boren, J.; Yamanaka, S.; Olofsson, S. O., Biosynthesis of apolipoprotein B48-containing lipoproteins. Regulation by novel post-transcriptional mechanisms. *J. Biol. Chem.* **1996,** *271* (5), 2353-6.
16. Guo, J. U.; Su, Y.; Zhong, C.; Ming, G. L.; Song, H., Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* **2011,** *145* (3), 423-34.
17. Morgan, H. D.; Dean, W.; Coker, H. A.; Reik, W.; Petersen-Mahrt, S. K., Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *J. Biol. Chem.* **2004,** *279* (50), 52353-60.
18. Petit, V.; Guetard, D.; Renard, M.; Keriel, A.; Sitbon, M.; Wain-Hobson, S.; Vartanian, J. P., Murine APOBEC1 is a powerful mutator of retroviral and cellular RNA in vitro and in vivo. *J. Mol. Biol.* **2009,** *385* (1), 65-78.
19. Gonzalez, M. C.; Suspene, R.; Henry, M.; Guetard, D.; Wain-Hobson, S.; Vartanian, J. P., Human APOBEC1 cytidine deaminase edits HBV DNA. *Retrovirology* **2009,** *6*, 96.
20. Gee, P.; Ando, Y.; Kitayama, H.; Yamamoto, S. P.; Kanemura, Y.; Ebina, H.; Kawaguchi, Y.; Koyanagi, Y., APOBEC1-mediated editing and attenuation of herpes simplex virus 1 DNA indicate that neurons have an antiviral role during herpes simplex encephalitis. *J. Virol.* **2011,** *85* (19), 9726-36.
21. Liao, W.; Hong, S. H.; Chan, B. H.; Rudolph, F. B.; Clark, S. C.; Chan, L., APOBEC-2, a cardiac- and skeletal muscle-specific member of the cytidine deaminase supergene family. *Biochem Biophys Res Commun* **1999,** *260* (2), 398-404.
22. Sheehy, A. M.; Gaddis, N. C.; Choi, J. D.; Malim, M. H., Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **2002,** *418* (6898), 646-50.
23. Mangeat, B.; Turelli, P.; Caron, G.; Friedli, M.; Perrin, L.; Trono, D., Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* **2003,** *424* (6944), 99-103.
24. Hultquist, J. F.; Lengyel, J. A.; Refsland, E. W.; LaRue, R. S.; Lackey, L.; Brown, W. L.; Harris, R. S., Human and rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H demonstrate a conserved capacity to restrict Vif-deficient HIV-1. *J Virol* **2011,** *85* (21), 11220-34.
25. Ooms, M.; Brayton, B.; Letko, M.; Maio, S. M.; Pilcher, C. D.; Hecht, F. M.; Barbour, J. D.; Simon, V., HIV-1 Vif adaptation to human APOBEC3H haplotypes. *Cell Host Microbe* **2013,** *14* (4), 411-21.
26. Belanger, K.; Savoie, M.; Rosales Gerpe, M. C.; Couture, J. F.; Langlois, M. A., Binding of RNA by APOBEC3G controls deamination-independent restriction of retroviruses. *Nucleic Acids Res.* **2013,** *41* (15), 7438-52.
27. Adolph, M. B.; Love, R. P.; Chelico, L., Biochemical Basis of APOBEC3 Deoxycytidine Deaminase Activity on Diverse DNA Substrates. *ACS Infect Dis* **2018,** *4* (3), 224-238.
28. Adolph, M. B.; Webb, J.; Chelico, L., Retroviral restriction factor APOBEC3G delays the initiation of DNA synthesis by HIV-1 reverse transcriptase. *PLoS One* **2013,** *8* (5), e64196.

29. Iwatani, Y.; Chan, D. S.; Wang, F.; Maynard, K. S.; Sugiura, W.; Gronenborn, A. M.; Rouzina, I.; Williams, M. C.; Musier-Forsyth, K.; Levin, J. G., Deaminase-independent inhibition of HIV-1 reverse transcription by APOBEC3G. *Nucleic Acids Res.* **2007,** *35* (21), 7096-108.
30. Chen, H.; Lilley, C. E.; Yu, Q.; Lee, D. V.; Chou, J.; Narvaiza, I.; Landau, N. R.; Weitzman, M. D., APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. *Curr Biol* **2006,** *16* (5), 480-5.
31. Janahi, E. M.; McGarvey, M. J., The inhibition of hepatitis B virus by APOBEC cytidine deaminases. *J Viral Hepat* **2013,** *20* (12), 821-8.
32. Vieira, V. C.; Leonard, B.; White, E. A.; Starrett, G. J.; Temiz, N. A.; Lorenz, L. D.; Lee, D.; Soares, M. A.; Lambert, P. F.; Howley, P. M.; Harris, R. S., Human papillomavirus E6 triggers upregulation of the antiviral and cancer genomic DNA deaminase APOBEC3B. *MBio* **2014,** *5* (6).
33. Wang, Z.; Wakae, K.; Kitamura, K.; Aoyama, S.; Liu, G.; Koura, M.; Monjurul, A. M.; Kukimoto, I.; Muramatsu, M., APOBEC3 deaminases induce hypermutation in human papillomavirus 16 DNA upon beta interferon stimulation. *J Virol* **2014,** *88* (2), 1308-17.
34. Suspene, R.; Aynaud, M. M.; Koch, S.; Pasdeloup, D.; Labetoulle, M.; Gaertner, B.; Vartanian, J. P.; Meyerhans, A.; Wain-Hobson, S., Genetic editing of herpes simplex virus 1 and Epstein-Barr herpesvirus genomes by human APOBEC3 cytidine deaminases in culture and in vivo. *J Virol* **2011,** *85* (15), 7594-602.
35. Seplyarskiy, V. B.; Soldatov, R. A.; Popadin, K. Y.; Antonarakis, S. E.; Bazykin, G. A.; Nikolaev, S. I., APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res.* **2016,** *26* (2), 174-82.
36. Bhagwat, A. S.; Hao, W.; Townes, J. P.; Lee, H.; Tang, H.; Foster, P. L., Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **2016,** *113* (8), 2176-81.
37. Haradhvala, N. J.; Polak, P.; Stojanov, P.; Covington, K. R.; Shinbrot, E.; Hess, J. M.; Rheinbay, E.; Kim, J.; Maruvka, Y. E.; Braunstein, L. Z.; Kamburov, A.; Hanawalt, P. C.; Wheeler, D. A.; Koren, A.; Lawrence, M. S.; Getz, G., Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **2016,** *164* (3), 538-49.
38. Hoopes, J. I.; Cortez, L. M.; Mertz, T. M.; Malc, E. P.; Mieczkowski, P. A.; Roberts, S. A., APOBEC3A and APOBEC3B Preferentially Deaminate the Lagging Strand Template during DNA Replication. *Cell Rep* **2016,** *14* (6), 1273-1282.
39. Roberts, S. A.; Sterling, J.; Thompson, C.; Harris, S.; Mav, D.; Shah, R.; Klimczak, L. J.; Kryukov, G. V.; Malc, E.; Mieczkowski, P. A.; Resnick, M. A.; Gordenin, D. A., Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **2012,** *46* (4), 424-35.
40. Nik-Zainal, S.; Alexandrov, L. B.; Wedge, D. C.; Van Loo, P.; Greenman, C. D.; Raine, K.; Jones, D.; Hinton, J.; Marshall, J.; Stebbings, L. A.; Menzies, A.; Martin, S.; Leung, K.; Chen, L.; Leroy, C.; Ramakrishna, M.; Rance, R.; Lau, K. W.; Mudie, L. J.; Varela, I.; McBride, D. J.; Bignell, G. R.; Cooke, S. L.; Shlien, A.; Gamble, J.; Whitmore, I.; Maddison, M.; Tarpey, P. S.; Davies, H. R.; Papaemmanuil, E.; Stephens, P. J.; McLaren, S.; Butler, A. P.; Teague, J. W.;

Jonsson, G.; Garber, J. E.; Silver, D.; Miron, P.; Fatima, A.; Boyault, S.; Langerod, A.; Tutt, A.; Martens, J. W.; Aparicio, S. A.; Borg, A.; Salomon, A. V.; Thomas, G.; Borresen-Dale, A. L.; Richardson, A. L.; Neuberger, M. S.; Futreal, P. A.; Campbell, P. J.; Stratton, M. R.; Breast Cancer Working Group of the International Cancer Genome, C., Mutational processes molding the genomes of 21 breast cancers. *Cell* **2012,** *149* (5), 979-93.
41. Taylor, B. J.; Nik-Zainal, S.; Wu, Y. L.; Stebbings, L. A.; Raine, K.; Campbell, P. J.; Rada, C.; Stratton, M. R.; Neuberger, M. S., DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife* **2013,** *2*, e00534.
42. Saini, N.; Roberts, S. A.; Sterling, J. F.; Malc, E. P.; Mieczkowski, P. A.; Gordenin, D. A., APOBEC3B cytidine deaminase targets the non-transcribed strand of tRNA genes in yeast. *DNA Repair (Amst)* **2017,** *53*, 4-14.
43. Alexandrov, L. B.; Nik-Zainal, S.; Wedge, D. C.; Aparicio, S. A.; Behjati, S.; Biankin, A. V.; Bignell, G. R.; Bolli, N.; Borg, A.; Borresen-Dale, A. L.; Boyault, S.; Burkhardt, B.; Butler, A. P.; Caldas, C.; Davies, H. R.; Desmedt, C.; Eils, R.; Eyfjord, J. E.; Foekens, J. A.; Greaves, M.; Hosoda, F.; Hutter, B.; Ilicic, T.; Imbeaud, S.; Imielinski, M.; Jager, N.; Jones, D. T.; Jones, D.; Knappskog, S.; Kool, M.; Lakhani, S. R.; Lopez-Otin, C.; Martin, S.; Munshi, N. C.; Nakamura, H.; Northcott, P. A.; Pajic, M.; Papaemmanuil, E.; Paradiso, A.; Pearson, J. V.; Puente, X. S.; Raine, K.; Ramakrishna, M.; Richardson, A. L.; Richter, J.; Rosenstiel, P.; Schlesner, M.; Schumacher, T. N.; Span, P. N.; Teague, J. W.; Totoki, Y.; Tutt, A. N.; Valdes-Mas, R.; van Buuren, M. M.; van 't Veer, L.; Vincent-Salomon, A.; Waddell, N.; Yates, L. R.; Australian Pancreatic Cancer Genome, I.; Consortium, I. B. C.; Consortium, I. M.-S.; PedBrain, I.; Zucman-Rossi, J.; Futreal, P. A.; McDermott, U.; Lichter, P.; Meyerson, M.; Grimmond, S. M.; Siebert, R.; Campo, E.; Shibata, T.; Pfister, S. M.; Campbell, P. J.; Stratton, M. R., Signatures of mutational processes in human cancer. *Nature* **2013,** *500* (7463), 415-21.
44. Burns, M. B.; Temiz, N. A.; Harris, R. S., Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat Genet* **2013,** *45* (9), 977-83.
45. Roberts, S. A.; Lawrence, M. S.; Klimczak, L. J.; Grimm, S. A.; Fargo, D.; Stojanov, P.; Kiezun, A.; Kryukov, G. V.; Carter, S. L.; Saksena, G.; Harris, S.; Shah, R. R.; Resnick, M. A.; Getz, G.; Gordenin, D. A., An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **2013,** *45* (9), 970-6.
46. Olson, M. E.; Harris, R. S.; Harki, D. A., APOBEC Enzymes as Targets for Virus and Cancer Therapy. *Cell Chem Biol* **2018,** *25* (1), 36-49.
47. Starrett, G. J.; Luengas, E. M.; McCann, J. L.; Ebrahimi, D.; Temiz, N. A.; Love, R. P.; Feng, Y.; Adolph, M. B.; Chelico, L.; Law, E. K.; Carpenter, M. A.; Harris, R. S., The DNA cytosine deaminase APOBEC3H haplotype I likely contributes to breast and lung cancer mutagenesis. *Nat Commun* **2016,** *7*, 12918.
48. Cortez, L. M.; Brown, A. L.; Dennis, M. A.; Collins, C. D.; Brown, A. J.; Mitchell, D.; Mertz, T. M.; Roberts, S. A., APOBEC3A is a prominent cytidine deaminase in breast cancer. *PLoS Genet* **2019,** *15* (12), e1008545.

49. Lackey, L.; Demorest, Z. L.; Land, A. M.; Hultquist, J. F.; Brown, W. L.; Harris, R. S., APOBEC3B and AID have similar nuclear import mechanisms. *J. Mol. Biol.* **2012,** *419* (5), 301-14.
50. Lackey, L.; Law, E. K.; Brown, W. L.; Harris, R. S., Subcellular localization of the APOBEC3 proteins during mitosis and implications for genomic DNA deamination. *Cell Cycle* **2013,** *12* (5), 762-72.
51. Rees, H. A.; Liu, D. R., Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **2018,** *19* (12), 770-788.
52. Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J. A.; Charpentier, E., A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **2012,** *337* (6096), 816-21.
53. Jansen, R.; Embden, J. D.; Gaastra, W.; Schouls, L. M., Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **2002,** *43* (6), 1565-75.
54. Garneau, J. E.; Dupuis, M. E.; Villion, M.; Romero, D. A.; Barrangou, R.; Boyaval, P.; Fremaux, C.; Horvath, P.; Magadan, A. H.; Moineau, S., The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **2010,** *468* (7320), 67-71.
55. Cho, S. W.; Kim, S.; Kim, J. M.; Kim, J. S., Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **2013,** *31* (3), 230-2.
56. Cong, L.; Ran, F. A.; Cox, D.; Lin, S.; Barretto, R.; Habib, N.; Hsu, P. D.; Wu, X.; Jiang, W.; Marraffini, L. A.; Zhang, F., Multiplex genome engineering using CRISPR/Cas systems. *Science* **2013,** *339* (6121), 819-23.
57. Jinek, M.; East, A.; Cheng, A.; Lin, S.; Ma, E.; Doudna, J., RNA-programmed genome editing in human cells. *Elife* **2013,** *2*, e00471.
58. Mali, P.; Yang, L.; Esvelt, K. M.; Aach, J.; Guell, M.; DiCarlo, J. E.; Norville, J. E.; Church, G. M., RNA-guided human genome engineering via Cas9. *Science* **2013,** *339* (6121), 823-6.
59. Jeggo, P. A., DNA breakage and repair. *Adv. Genet.* **1998,** *38*, 185-218.
60. Rouet, P.; Smih, F.; Jasin, M., Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol Cell Biol* **1994,** *14* (12), 8096-106.
61. Lukacsovich, T.; Yang, D.; Waldman, A. S., Repair of a specific double-strand break generated within a mammalian chromosome by yeast endonuclease I-SceI. *Nucleic Acids Res.* **1994,** *22* (25), 5649-57.
62. Rudin, N.; Sugarman, E.; Haber, J. E., Genetic and physical analysis of double-strand break repair and recombination in Saccharomyces cerevisiae. *Genetics* **1989,** *122* (3), 519-34.
63. Rouet, P.; Smih, F.; Jasin, M., Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **1994,** *91* (13), 6064-8.
64. Ran, F. A.; Hsu, P. D.; Wright, J.; Agarwala, V.; Scott, D. A.; Zhang, F., Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **2013,** *8* (11), 2281-2308.

65. Paquet, D.; Kwart, D.; Chen, A.; Sproul, A.; Jacob, S.; Teo, S.; Olsen, K. M.; Gregg, A.; Noggle, S.; Tessier-Lavigne, M., Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **2016,** *533* (7601), 125-9.
66. Lin, S.; Staahl, B. T.; Alla, R. K.; Doudna, J. A., Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife* **2014,** *3*, e04766.
67. 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* **2016,** *533* (7603), 420-4.
68. Nishida, K.; Arazoe, T.; Yachie, N.; Banno, S.; Kakimoto, M.; Tabata, M.; Mochizuki, M.; Miyabe, A.; Araki, M.; Hara, K. Y.; Shimatani, Z.; Kondo, A., Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **2016,** *353* (6305).
69. Gaudelli, N. M.; Komor, A. C.; Rees, H. A.; Packer, M. S.; Badran, A. H.; Bryson, D. I.; Liu, D. R., Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* **2017,** *551* (7681), 464-471.
70. Mol, C. D.; Arvai, A. S.; Sanderson, R. J.; Slupphaug, G.; Kavli, B.; Krokan, H. E.; Mosbaugh, D. W.; Tainer, J. A., Crystal structure of human uracil-DNA glycosylase in complex with a protein inhibitor: protein mimicry of DNA. *Cell* **1995,** *82* (5), 701-8.
71. Nishimasu, H.; Ran, F. A.; Hsu, P. D.; Konermann, S.; Shehata, S. I.; Dohmae, N.; Ishitani, R.; Zhang, F.; Nureki, O., Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **2014,** *156* (5), 935-49.
72. Komor, A. C.; Zhao, K. T.; Packer, M. S.; Gaudelli, N. M.; Waterbury, A. L.; Koblan, L. W.; Kim, Y. B.; Badran, A. H.; Liu, D. R., Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Sci Adv* **2017,** *3* (8), eaao4774.
73. Kleinstiver, B. P.; Pattanayak, V.; Prew, M. S.; Tsai, S. Q.; Nguyen, N. T.; Zheng, Z.; Joung, J. K., High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **2016,** *529* (7587), 490-5.
74. Ran, F. A.; Cong, L.; Yan, W. X.; Scott, D. A.; Gootenberg, J. S.; Kriz, A. J.; Zetsche, B.; Shalem, O.; Wu, X.; Makarova, K. S.; Koonin, E. V.; Sharp, P. A.; Zhang, F., In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **2015,** *520* (7546), 186-91.
75. Zetsche, B.; Gootenberg, J. S.; Abudayyeh, O. O.; Slaymaker, I. M.; Makarova, K. S.; Essletzbichler, P.; Volz, S. E.; Joung, J.; van der Oost, J.; Regev, A.; Koonin, E. V.; Zhang, F., Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **2015,** *163* (3), 759-71.
76. Kim, E.; Koo, T.; Park, S. W.; Kim, D.; Kim, K.; Cho, H. Y.; Song, D. W.; Lee, K. J.; Jung, M. H.; Kim, S.; Kim, J. H.; Kim, J. H.; Kim, J. S., In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nat Commun* **2017,** *8*, 14500.
77. Kleinstiver, B. P.; Prew, M. S.; Tsai, S. Q.; Topkar, V. V.; Nguyen, N. T.; Zheng, Z.; Gonzales, A. P.; Li, Z.; Peterson, R. T.; Yeh, J. R.; Aryee, M. J.; Joung, J. K., Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **2015,** *523* (7561), 481-5.

78. Lee, C. M.; Cradick, T. J.; Bao, G., The Neisseria meningitidis CRISPR-Cas9 System Enables Specific Genome Editing in Mammalian Cells. *Mol. Ther.* **2016,** *24* (3), 645-54.
79. Chen, K. M.; Harjes, E.; Gross, P. J.; Fahmy, A.; Lu, Y.; Shindo, K.; Harris, R. S.; Matsuo, H., Structure of the DNA deaminase domain of the HIV-1 restriction factor APOBEC3G. *Nature* **2008,** *452* (7183), 116-9.
80. Harjes, E.; Gross, P. J.; Chen, K. M.; Lu, Y.; Shindo, K.; Nowarski, R.; Gross, J. D.; Kotler, M.; Harris, R. S.; Matsuo, H., An extended structure of the APOBEC3G catalytic domain suggests a unique holoenzyme model. *J. Mol. Biol.* **2009,** *389* (5), 819-32.
81. Shandilya, S. M.; Nalam, M. N.; Nalivaika, E. A.; Gross, P. J.; Valesano, J. C.; Shindo, K.; Li, M.; Munson, M.; Royer, W. E.; Harjes, E.; Kono, T.; Matsuo, H.; Harris, R. S.; Somasundaran, M.; Schiffer, C. A., Crystal structure of the APOBEC3G catalytic domain reveals potential oligomerization interfaces. *Structure* **2010,** *18* (1), 28-38.
82. Li, M.; Shandilya, S. M.; Carpenter, M. A.; Rathore, A.; Brown, W. L.; Perkins, A. L.; Harki, D. A.; Solberg, J.; Hook, D. J.; Pandey, K. K.; Parniak, M. A.; Johnson, J. R.; Krogan, N. J.; Somasundaran, M.; Ali, A.; Schiffer, C. A.; Harris, R. S., First-in-class small molecule inhibitors of the single-strand DNA cytosine deaminase APOBEC3G. *ACS Chem. Biol.* **2012,** *7* (3), 506-17.
83. Bohn, M. F.; Shandilya, S. M.; Albin, J. S.; Kouno, T.; Anderson, B. D.; McDougle, R. M.; Carpenter, M. A.; Rathore, A.; Evans, L.; Davis, A. N.; Zhang, J.; Lu, Y.; Somasundaran, M.; Matsuo, H.; Harris, R. S.; Schiffer, C. A., Crystal structure of the DNA cytosine deaminase APOBEC3F: the catalytically active and HIV-1 Vif-binding domain. *Structure* **2013,** *21* (6), 1042-50.
84. Bohn, M. F.; Shandilya, S. M.; Silvas, T. V.; Nalivaika, E. A.; Kouno, T.; Kelch, B. A.; Ryder, S. P.; Kurt-Yilmaz, N.; Somasundaran, M.; Schiffer, C. A., The ssDNA Mutator APOBEC3A Is Regulated by Cooperative Dimerization. *Structure* **2015,** *23* (5), 903-11.
85. Kouno, T.; Luengas, E. M.; Shigematsu, M.; Shandilya, S. M.; Zhang, J.; Chen, L.; Hara, M.; Schiffer, C. A.; Harris, R. S.; Matsuo, H., Structure of the Vif-binding domain of the antiviral enzyme APOBEC3G. *Nat. Struct. Mol. Biol.* **2015,** *22* (6), 485-91.
86. Chelico, L.; Pham, P.; Calabrese, P.; Goodman, M. F., APOBEC3G DNA deaminase acts processively 3' --> 5' on single-stranded DNA. *Nat. Struct. Mol. Biol.* **2006,** *13* (5), 392-9.
87. Holden, L. G.; Prochnow, C.; Chang, Y. P.; Bransteitter, R.; Chelico, L.; Sen, U.; Stevens, R. C.; Goodman, M. F.; Chen, X. S., Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications. *Nature* **2008,** *456* (7218), 121-4.
88. Chelico, L.; Sacho, E. J.; Erie, D. A.; Goodman, M. F., A model for oligomeric regulation of APOBEC3G cytosine deaminase-dependent restriction of HIV. *J. Biol. Chem.* **2008,** *283* (20), 13780-91.
89. Furukawa, A.; Nagata, T.; Matsugami, A.; Habu, Y.; Sugiyama, R.; Hayashi, F.; Kobayashi, N.; Yokoyama, S.; Takaku, H.; Katahira, M., Structure, interaction and

real-time monitoring of the enzymatic reaction of wild-type APOBEC3G. *EMBO J.* **2009,** *28* (4), 440-51.
90. Chelico, L.; Prochnow, C.; Erie, D. A.; Chen, X. S.; Goodman, M. F., Structural model for deoxycytidine deamination mechanisms of the HIV-1 inactivation enzyme APOBEC3G. *J. Biol. Chem.* **2010,** *285* (21), 16195-205.
91. Kitamura, S.; Ode, H.; Nakashima, M.; Imahashi, M.; Naganawa, Y.; Kurosawa, T.; Yokomaku, Y.; Yamane, T.; Watanabe, N.; Suzuki, A.; Sugiura, W.; Iwatani, Y., The APOBEC3C crystal structure and the interface for HIV-1 Vif binding. *Nat. Struct. Mol. Biol.* **2012,** *19* (10), 1005-10.
92. Siu, K. K.; Sultana, A.; Azimi, F. C.; Lee, J. E., Structural determinants of HIV-1 Vif susceptibility and DNA binding in APOBEC3F. *Nat Commun* **2013,** *4*, 2593.
93. Byeon, I. J.; Ahn, J.; Mitra, M.; Byeon, C. H.; Hercik, K.; Hritz, J.; Charlton, L. M.; Levin, J. G.; Gronenborn, A. M., NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. *Nat Commun* **2013,** *4*, 1890.
94. Mitra, M.; Hercik, K.; Byeon, I. J.; Ahn, J.; Hill, S.; Hinchee-Rodriguez, K.; Singer, D.; Byeon, C. H.; Charlton, L. M.; Nam, G.; Heidecker, G.; Gronenborn, A. M.; Levin, J. G., Structural determinants of human APOBEC3A enzymatic and nucleic acid binding properties. *Nucleic Acids Res* **2014,** *42* (2), 1095-110.
95. Lu, X.; Zhang, T.; Xu, Z.; Liu, S.; Zhao, B.; Lan, W.; Wang, C.; Ding, J.; Cao, C., Crystal Structure of DNA Cytidine Deaminase ABOBEC3G Catalytic Deamination Domain Suggests a Binding Mode of Full-length Enzyme to Single-stranded DNA. *J. Biol. Chem.* **2015,** *290* (7), 4010-21.
96. Shi, K.; Carpenter, M. A.; Kurahashi, K.; Harris, R. S.; Aihara, H., Crystal Structure of the DNA Deaminase APOBEC3B Catalytic Domain. *J. Biol. Chem.* **2015,** *290* (47), 28120-30.
97. Shaban, N. M.; Shi, K.; Li, M.; Aihara, H.; Harris, R. S., 1.92 Angstrom Zinc-Free APOBEC3F Catalytic Domain Crystal Structure. *J. Mol. Biol.* **2016,** *428* (11), 2307-16.
98. Byeon, I. J.; Byeon, C. H.; Wu, T.; Mitra, M.; Singer, D.; Levin, J. G.; Gronenborn, A. M., Nuclear Magnetic Resonance Structure of the APOBEC3B Catalytic Domain: Structural Basis for Substrate Binding and DNA Deaminase Activity. *Biochemistry* **2016,** *55* (21), 2944-59.
99. Xiao, X.; Li, S. X.; Yang, H.; Chen, X. S., Crystal structures of APOBEC3G N-domain alone and its complex with DNA. *Nat Commun* **2016,** *7*, 12193.
100. Shi, K.; Demir, O.; Carpenter, M. A.; Wagner, J.; Kurahashi, K.; Harris, R. S.; Amaro, R. E.; Aihara, H., Conformational Switch Regulates the DNA Cytosine Deaminase Activity of Human APOBEC3B. *Sci. Rep.* **2017,** *7* (1), 17415.
101. Xiao, X.; Yang, H.; Arutiunian, V.; Fang, Y.; Besse, G.; Morimoto, C.; Zirkle, B.; Chen, X. S., Structural determinants of APOBEC3B non-catalytic domain for molecular assembly and catalytic regulation. *Nucleic Acids Res.* **2017,** *45* (12), 7494-7506.
102. Ito, F.; Yang, H.; Xiao, X.; Li, S. X.; Wolfe, A.; Zirkle, B.; Arutiunian, V.; Chen, X. S., Understanding the Structure, Multimerization, Subcellular Localization and mC Selectivity of a Genomic Mutator and Anti-HIV Factor APOBEC3H. *Sci. Rep.* **2018,** *8* (1), 3763.

103. Silvas, T. V.; Schiffer, C. A., APOBEC3s: DNA-editing human cytidine deaminases. *Protein Sci.* **2019,** *28* (9), 1552-1566.
104. Kouno, T.; Silvas, T. V.; Hilbert, B. J.; Shandilya, S. M. D.; Bohn, M. F.; Kelch, B. A.; Royer, W. E.; Somasundaran, M.; Kurt Yilmaz, N.; Matsuo, H.; Schiffer, C. A., Crystal structure of APOBEC3A bound to single-stranded DNA reveals structural basis for cytidine deamination and specificity. *Nat Commun* **2017,** *8*, 15024.
105. Maiti, A.; Myint, W.; Kanai, T.; Delviks-Frankenberry, K.; Sierra Rodriguez, C.; Pathak, V. K.; Schiffer, C. A.; Matsuo, H., Crystal structure of the catalytic domain of HIV-1 restriction factor APOBEC3G in complex with ssDNA. *Nat Commun* **2018,** *9* (1), 2460.
106. Shi, K.; Carpenter, M. A.; Banerjee, S.; Shaban, N. M.; Kurahashi, K.; Salamango, D. J.; McCann, J. L.; Starrett, G. J.; Duffy, J. V.; Demir, O.; Amaro, R. E.; Harki, D. A.; Harris, R. S.; Aihara, H., Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. *Nat Struct Mol Biol* **2017,** *24* (2), 131-139.
107. Fang, Y.; Xiao, X.; Li, S. X.; Wolfe, A.; Chen, X. S., Molecular Interactions of a DNA Modifying Enzyme APOBEC3F Catalytic Domain with a Single-Stranded DNA. *J. Mol. Biol.* **2018,** *430* (1), 87-101.
108. Ziegler, S. J.; Liu, C.; Landau, M.; Buzovetsky, O.; Desimmie, B. A.; Zhao, Q.; Sasaki, T.; Burdick, R. C.; Pathak, V. K.; Anderson, K. S.; Xiong, Y., Insights into DNA substrate selection by APOBEC3G from structural, biochemical, and functional studies. *PLoS One* **2018,** *13* (3), e0195048.
109. Shaban, N. M.; Shi, K.; Lauer, K. V.; Carpenter, M. A.; Richards, C. M.; Salamango, D.; Wang, J.; Lopresti, M. W.; Banerjee, S.; Levin-Klein, R.; Brown, W. L.; Aihara, H.; Harris, R. S., The Antiviral and Cancer Genomic DNA Deaminase APOBEC3H Is Regulated by an RNA-Mediated Dimerization Mechanism. *Mol. Cell* **2018,** *69* (1), 75-86 e9.
110. Bohn, J. A.; Thummar, K.; York, A.; Raymond, A.; Brown, W. C.; Bieniasz, P. D.; Hatziioannou, T.; Smith, J. L., APOBEC3H structure reveals an unusual mechanism of interaction with duplex RNA. *Nat Commun* **2017,** *8* (1), 1021.
111. Matsuoka, T.; Nagae, T.; Ode, H.; Awazu, H.; Kurosawa, T.; Hamano, A.; Matsuoka, K.; Hachiya, A.; Imahashi, M.; Yokomaku, Y.; Watanabe, N.; Iwatani, Y., Structural basis of chimpanzee APOBEC3H dimerization stabilized by double-stranded RNA. *Nucleic Acids Res.* **2018,** *46* (19), 10368-10379.
112. Yang, H.; Ito, F.; Wolfe, A. D.; Li, S.; Mohammadzadeh, N.; Love, R. P.; Yan, M.; Zirkle, B.; Gaba, A.; Chelico, L.; Chen, X. S., Understanding the structural basis of HIV-1 restriction by the full length double-domain APOBEC3G. *Nat Commun* **2020,** *11* (1), 632.
113. Ito, F.; Fu, Y.; Kao, S. A.; Yang, H.; Chen, X. S., Family-Wide Comparative Analysis of Cytidine and Methylcytidine Deamination by Eleven Human APOBEC Proteins. *J. Mol. Biol.* **2017,** *429* (12), 1787-1799.
114. Silvas, T. V.; Hou, S.; Myint, W.; Nalivaika, E.; Somasundaran, M.; Kelch, B. A.; Matsuo, H.; Kurt Yilmaz, N.; Schiffer, C. A., Substrate sequence selectivity of APOBEC3A implicates intra-DNA interactions. *Sci. Rep.* **2018,** *8* (1), 7511.

115. Solomon, W. C.; Myint, W.; Hou, S.; Kanai, T.; Tripathi, R.; Kurt Yilmaz, N.; Schiffer, C. A.; Matsuo, H., Mechanism for APOBEC3G catalytic exclusion of RNA and non-substrate DNA. *Nucleic Acids Res* **2019,** *47* (14), 7676-7689.
116. Sharma, S.; Wang, J.; Alqassim, E.; Portwood, S.; Cortes Gomez, E.; Maguire, O.; Basse, P. H.; Wang, E. S.; Segal, B. H.; Baysal, B. E., Mitochondrial hypoxic stress induces widespread RNA editing by APOBEC3G in natural killer cells. *Genome Biol.* **2019,** *20* (1), 37.
117. Beale, R. C.; Petersen-Mahrt, S. K.; Watt, I. N.; Harris, R. S.; Rada, C.; Neuberger, M. S., Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. *J. Mol. Biol.* **2004,** *337* (3), 585-96.
118. Pham, P.; Bransteitter, R.; Petruska, J.; Goodman, M. F., Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **2003,** *424* (6944), 103-7.
119. Chan, K.; Roberts, S. A.; Klimczak, L. J.; Sterling, J. F.; Saini, N.; Malc, E. P.; Kim, J.; Kwiatkowski, D. J.; Fargo, D. C.; Mieczkowski, P. A.; Getz, G.; Gordenin, D. A., An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* **2015,** *47* (9), 1067-72.
120. Langlois, M. A.; Beale, R. C.; Conticello, S. G.; Neuberger, M. S., Mutational comparison of the single-domained APOBEC3C and double-domained APOBEC3F/G anti-retroviral cytidine deaminases provides insight into their DNA target site specificities. *Nucleic Acids Res.* **2005,** *33* (6), 1913-23.
121. Schwede, T., Protein modeling: what happened to the "protein structure gap"? *Structure* **2013,** *21* (9), 1531-40.
122. Baker, D.; Sali, A., Protein structure prediction and structural genomics. *Science* **2001,** *294* (5540), 93-6.
123. Chothia, C.; Lesk, A. M., The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986,** *5* (4), 823-6.
124. Chung, S. Y.; Subbiah, S., A structural explanation for the twilight zone of protein sequence homology. *Structure* **1996,** *4* (10), 1123-7.
125. Khor, B. Y.; Tye, G. J.; Lim, T. S.; Choong, Y. S., General overview on structure prediction of twilight-zone proteins. *Theor. Biol. Med. Model.* **2015,** *12*, 15.
126. Schwede, T.; Sali, A.; Honig, B.; Levitt, M.; Berman, H. M.; Jones, D.; Brenner, S. E.; Burley, S. K.; Das, R.; Dokholyan, N. V.; Dunbrack, R. L., Jr.; Fidelis, K.; Fiser, A.; Godzik, A.; Huang, Y. J.; Humblet, C.; Jacobson, M. P.; Joachimiak, A.; Krystek, S. R., Jr.; Kortemme, T.; Kryshtafovych, A.; Montelione, G. T.; Moult, J.; Murray, D.; Sanchez, R.; Sosnick, T. R.; Standley, D. M.; Stouch, T.; Vajda, S.; Vasquez, M.; Westbrook, J. D.; Wilson, I. A., Outcome of a workshop on applications of protein models in biomedical research. *Structure* **2009,** *17* (2), 151-9.
127. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; Mackerell, A. D., Jr., Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J. Chem. Theory Comput.* **2012,** *8* (9), 3257-3273.

128. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmuller, H.; MacKerell, A. D., Jr., CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017,** *14* (1), 71-73.
129. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules J. Am. Chem. Soc. 1995, 117, 5179− 5197. *Journal of the American Chemical Society* **1996,** *118* (9), 2309-2309.
130. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015,** *11* (8), 3696-713.
131. Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A., OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016,** *12* (1), 281-96.
132. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996,** *118* (45), 11225-11236.
133. Collier, T. A.; Piggot, T. J.; Allison, J. R., Molecular Dynamics Simulation of Proteins. In *Protein Nanotechnology*, Springer: 2020; pp 311-327.
134. Karplus, M.; Kuriyan, J., Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U. S. A.* **2005,** *102* (19), 6679-85.
135. Ozen, A.; Lin, K. H.; Kurt Yilmaz, N.; Schiffer, C. A., Structural basis and distal effects of Gag substrate coevolution in drug resistance to HIV-1 protease. *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111* (45), 15993–15998.
136. Soumana, D. I.; Kurt Yilmaz, N.; Ali, A.; Prachanronarong, K. L.; Schiffer, C. A., Molecular and dynamic mechanism underlying drug resistance in genotype 3 hepatitis C NS3/4A protease. *J. Am. Chem. Soc.* **2016,** *138* (36), 11850–11859.
137. Paulsen, J. L.; Leidner, F.; Ragland, D. A.; Kurt Yilmaz, N.; Schiffer, C. A., Interdependence of inhibitor recognition in HIV-1 protease. *J. Chem. Theory. Comput.* **2017,** *13* (5), 2300–2309.
138. Henes, M.; Lockbaum, G. J.; Kosovrasti, K.; Leidner, F.; Nachum, G. S.; Nalivaika, E. A.; Lee, S. K.; Spielvogel, E.; Zhou, S.; Swanstrom, R.; Bolon, D. N. A.; Kurt Yilmaz, N.; Schiffer, C. A., Picomolar to Micromolar: Elucidating the Role of Distal Mutations in HIV-1 Protease in Conferring Drug Resistance. *ACS Chem. Biol.* **2019**, doi: 10.1021/acschembio.9b00370.
139. Hou, S.; Silvas, T. V.; Leidner, F.; Nalivaika, E. A.; Matsuo, H.; Kurt Yilmaz, N.; Schiffer, C. A., Structural Analysis of the Active Site and DNA Binding of Human Cytidine Deaminase APOBEC3B. *J Chem Theory Comput* **2019,** *15* (1), 637-647.
140. Quaresma, A. J.; Oyama, S., Jr.; Barbosa, J. A.; Kobarg, J., The acidic domain of hnRNPQ (NSAP1) has structural similarity to Barstar and binds to Apobec1. *Biochem Biophys Res Commun* **2006,** *350* (2), 288-97.
141. Matsumoto, T.; Shirakawa, K.; Yokoyama, M.; Fukuda, H.; Sarca, A. D.; Koyabu, S.; Yamazaki, H.; Kazuma, Y.; Matsui, H.; Maruyama, W.; Nagata, K.; Tanabe, F.; Kobayashi, M.; Shindo, K.; Morishita, R.; Sato, H.; Takaori-Kondo, A.,

Protein kinase A inhibits tumor mutator APOBEC3B through phosphorylation. *Sci. Rep.* **2019,** *9* (1), 8307.
142. Autore, F.; Bergeron, J. R.; Malim, M. H.; Fraternali, F.; Huthoff, H., Rationalisation of the differences between APOBEC3G structures from crystallography and NMR studies by molecular dynamics simulations. *PLoS One* **2010,** *5* (7), e11515.
143. King, J. J.; Manuel, C. A.; Barrett, C. V.; Raber, S.; Lucas, H.; Sutter, P.; Larijani, M., Catalytic pocket inaccessibility of activation-induced cytidine deaminase is a safeguard against excessive mutagenic activity. *Structure* **2015,** *23* (4), 615-27.
144. Gajula, K. S.; Huwe, P. J.; Mo, C. Y.; Crawford, D. J.; Stivers, J. T.; Radhakrishnan, R.; Kohli, R. M., High-throughput mutagenesis reveals functional determinants for DNA targeting by activation-induced deaminase. *Nucleic Acids Res.* **2014,** *42* (15), 9964-75.
145. Zheng, Y. H.; Irwin, D.; Kurosu, T.; Tokunaga, K.; Sata, T.; Peterlin, B. M., Human APOBEC3F is another host factor that blocks human immunodeficiency virus type 1 replication. *J. Virol.* **2004,** *78* (11), 6073-6.
146. Dang, Y.; Siew, L. M.; Wang, X.; Han, Y.; Lampen, R.; Zheng, Y. H., Human cytidine deaminase APOBEC3H restricts HIV-1 replication. *J. Biol. Chem.* **2008,** *283* (17), 11606-14.
147. Dang, Y.; Wang, X.; Esselman, W. J.; Zheng, Y. H., Identification of APOBEC3DE as another antiretroviral factor from the human APOBEC family. *J Virol* **2006,** *80* (21), 10522-33.
148. Bogerd, H. P.; Wiegand, H. L.; Doehle, B. P.; Lueders, K. K.; Cullen, B. R., APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells. *Nucleic Acids Res* **2006,** *34* (1), 89-95.
149. Muckenfuss, H.; Hamdorf, M.; Held, U.; Perkovic, M.; Lower, J.; Cichutek, K.; Flory, E.; Schumann, G. G.; Munk, C., APOBEC3 proteins inhibit human LINE-1 retrotransposition. *J Biol Chem* **2006,** *281* (31), 22161-72.
150. Burns, M. B.; Lackey, L.; Carpenter, M. A.; Rathore, A.; Land, A. M.; Leonard, B.; Refsland, E. W.; Kotandeniya, D.; Tretyakova, N.; Nikas, J. B.; Yee, D.; Temiz, N. A.; Donohue, D. E.; McDougle, R. M.; Brown, W. L.; Law, E. K.; Harris, R. S., APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **2013,** *494* (7437), 366-70.
151. Wissing, S.; Montano, M.; Garcia-Perez, J. L.; Moran, J. V.; Greene, W. C., Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells. *J. Biol. Chem.* **2011,** *286* (42), 36427-37.
152. Xu, R.; Zhang, X.; Zhang, W.; Fang, Y.; Zheng, S.; Yu, X. F., Association of human APOBEC3 cytidine deaminases with the generation of hepatitis virus B x antigen mutants and hepatocellular carcinoma. *Hepatology* **2007,** *46* (6), 1810-20.
153. Bonvin, M.; Greeve, J., Effects of point mutations in the cytidine deaminase domains of APOBEC3B on replication and hypermutation of hepatitis B virus in vitro. *J. Gen. Virol.* **2007,** *88* (Pt 12), 3270-4.
154. Leonard, B.; Hart, S. N.; Burns, M. B.; Carpenter, M. A.; Temiz, N. A.; Rathore, A.; Vogel, R. I.; Nikas, J. B.; Law, E. K.; Brown, W. L.; Li, Y.; Zhang, Y.; Maurer, M. J.; Oberg, A. L.; Cunningham, J. M.; Shridhar, V.; Bell, D. A.; April, C.; Bentley, D.; Bibikova, M.; Cheetham, R. K.; Fan, J. B.; Grocock, R.; Humphray, S.; Kingsbury, Z.; Peden, J.; Chien, J.; Swisher, E. M.; Hartmann, L. C.; Kalli, K. R.;

Goode, E. L.; Sicotte, H.; Kaufmann, S. H.; Harris, R. S., APOBEC3B Upregulation and Genomic Mutation Patterns in Serous Ovarian Carcinoma. *Cancer Res.* **2013,** *73* (24), 7222-31.
155. Sieuwerts, A. M.; Willis, S.; Burns, M. B.; Look, M. P.; Meijer-Van Gelder, M. E.; Schlicker, A.; Heideman, M. R.; Jacobs, H.; Wessels, L.; Leyland-Jones, B.; Gray, K. P.; Foekens, J. A.; Harris, R. S.; Martens, J. W., Elevated APOBEC3B correlates with poor outcomes for estrogen-receptor-positive breast cancers. *Horm. Cancer* **2014,** *5* (6), 405-13.
156. Periyasamy, M.; Singh, A. K.; Gemma, C.; Kranjec, C.; Farzan, R.; Leach, D. A.; Navaratnam, N.; Palinkas, H. L.; Vertessy, B. G.; Fenton, T. R.; Doorbar, J.; Fuller-Pace, F.; Meek, D. W.; Coombes, R. C.; Buluwela, L.; Ali, S., p53 controls expression of the DNA deaminase APOBEC3B to limit its potential mutagenic activity in cancer cells. *Nucleic Acids Res.* **2017,** *45* (19), 11056-11069.
157. Glaser, A. P.; Fantini, D.; Wang, Y.; Yu, Y.; Rimar, K. J.; Podojil, J. R.; Miller, S. D.; Meeks, J. J., APOBEC-mediated mutagenesis in urothelial carcinoma is associated with improved survival, mutations in DNA damage response genes, and immune response. *Oncotarget* **2018,** *9* (4), 4537-4548.
158. Harris, R. S., Molecular mechanism and clinical impact of APOBEC3B-catalyzed mutagenesis in breast cancer. *Breast Cancer Res.* **2015,** *17*, 8.
159. Kidd, J. M.; Newman, T. L.; Tuzun, E.; Kaul, R.; Eichler, E. E., Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet.* **2007,** *3* (4), e63.
160. Fu, Y.; Ito, F.; Zhang, G.; Fernandez, B.; Yang, H.; Chen, X. S., DNA cytosine and methylcytosine deamination by APOBEC3B: enhancing methylcytosine deamination by engineering APOBEC3B. *Biochem J* **2015,** *471* (1), 25-35.
161. Caval, V.; Bouzidi, M. S.; Suspene, R.; Laude, H.; Dumargne, M. C.; Bashamboo, A.; Krey, T.; Vartanian, J. P.; Wain-Hobson, S., Molecular basis of the attenuated phenotype of human APOBEC3B DNA mutator enzyme. *Nucleic Acids Res.* **2015,** *43* (19), 9340-9.
162. Sharma, S.; Patnaik, S. K.; Taggart, R. T.; Kannisto, E. D.; Enriquez, S. M.; Gollnick, P.; Baysal, B. E., APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. *Nat Commun* **2015,** *6*, 6881.
163. Liu, M.; Mallinger, A.; Tortorici, M.; Newbatt, Y.; Richards, M.; Mirza, A.; van Montfort, R. L. M.; Burke, R.; Blagg, J.; Kaserer, T., Evaluation of APOBEC3B recognition motifs by NMR reveals preferred substrates. *ACS Chem. Biol.* **2018**.
164. Bowers, K. J., Chow, E., Xu, H., Dror, R.O., Eastwood, M.P., Gregersen, B.A., Klepeis, J.L., Kolossvary, I., Moraes, M.A., Sacerdoti, F.D., Salmon, J.K., Shan, Y. and Shaw, D.E. In *Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters*, Proceedings of the 2006 ACM/IEEE conference on Supercomputing, 2006; p 84.
165. Leidner, F.; Kurt Yilmaz, N.; Paulsen, J.; Muller, Y. A.; Schiffer, C. A., Hydration Structure and Dynamics of Inhibitor-Bound HIV-1 Protease. *J. Chem. Theory Comput.* **2018,** *14* (5), 2784-2796.
166. Flyvbjerg, H.; Petersen, H. G., Error estimates on averages of correlated data. *J. Chem. Phys.* **1989,** *91* (1), 461-466.

167. Petljak, M.; Alexandrov, L. B.; Brammeld, J. S.; Price, S.; Wedge, D. C.; Grossmann, S.; Dawson, K. J.; Ju, Y. S.; Iorio, F.; Tubio, J. M. C.; Koh, C. C.; Georgakopoulos-Soares, I.; Rodriguez-Martin, B.; Otlu, B.; O'Meara, S.; Butler, A. P.; Menzies, A.; Bhosle, S. G.; Raine, K.; Jones, D. R.; Teague, J. W.; Beal, K.; Latimer, C.; O'Neill, L.; Zamora, J.; Anderson, E.; Patel, N.; Maddison, M.; Ng, B. L.; Graham, J.; Garnett, M. J.; McDermott, U.; Nik-Zainal, S.; Campbell, P. J.; Stratton, M. R., Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **2019,** *176* (6), 1282-1294 e20.
168. Hou, S.; Silvas, T.; Leidner, F.; Nalivaika, E. A.; Matsuo, H.; Kurt Yilmaz, N.; Schiffer, C. A., Structural analysis of the active site and DNA binding of human cytidine deaminase APOBEC3B. *J. Chem. Theory Comput.* **2018**.
169. Zhang, H.; Yang, B.; Pomerantz, R. J.; Zhang, C.; Arunachalam, S. C.; Gao, L., The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. *Nature* **2003,** *424* (6944), 94-8.
170. Yu, Q.; Konig, R.; Pillai, S.; Chiles, K.; Kearney, M.; Palmer, S.; Richman, D.; Coffin, J. M.; Landau, N. R., Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. *Nat. Struct. Mol. Biol.* **2004,** *11* (5), 435-42.
171. Suspene, R.; Sommer, P.; Henry, M.; Ferris, S.; Guetard, D.; Pochet, S.; Chester, A.; Navaratnam, N.; Wain-Hobson, S.; Vartanian, J. P., APOBEC3G is a single-stranded DNA cytidine deaminase and functions independently of HIV reverse transcriptase. *Nucleic Acids Res.* **2004,** *32* (8), 2421-9.
172. Rathore, A.; Carpenter, M. A.; Demir, O.; Ikeda, T.; Li, M.; Shaban, N. M.; Law, E. K.; Anokhin, D.; Brown, W. L.; Amaro, R. E.; Harris, R. S., The local dinucleotide preference of APOBEC3G can be altered from 5'-CC to 5'-TC by a single amino acid substitution. *J. Mol. Biol.* **2013,** *425* (22), 4442-54.
173. Stenglein, M. D.; Burns, M. B.; Li, M.; Lengyel, J.; Harris, R. S., APOBEC3 proteins mediate the clearance of foreign DNA from human cells. *Nat Struct Mol Biol* **2010,** *17* (2), 222-9.
174. Liddament, M. T.; Brown, W. L.; Schumacher, A. J.; Harris, R. S., APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 in vivo. *Curr Biol* **2004,** *14* (15), 1385-91.
175. Harris, R. S.; Bishop, K. N.; Sheehy, A. M.; Craig, H. M.; Petersen-Mahrt, S. K.; Watt, I. N.; Neuberger, M. S.; Malim, M. H., DNA deamination mediates innate immunity to retroviral infection. *Cell* **2003,** *113* (6), 803-9.
176. Ara, A.; Love, R. P.; Chelico, L., Different mutagenic potential of HIV-1 restriction factors APOBEC3G and APOBEC3F is determined by distinct single-stranded DNA scanning mechanisms. *PLoS Pathog* **2014,** *10* (3), e1004024.
177. Holtz, C. M.; Sadler, H. A.; Mansky, L. M., APOBEC3G cytosine deamination hotspots are defined by both sequence context and single-stranded DNA secondary structure. *Nucleic Acids Res* **2013,** *41* (12), 6139-48.
178. Carpenter, M. A.; Li, M.; Rathore, A.; Lackey, L.; Law, E. K.; Land, A. M.; Leonard, B.; Shandilya, S. M.; Bohn, M. F.; Schiffer, C. A.; Brown, W. L.; Harris, R. S., Methylcytosine and normal cytosine deamination by the foreign DNA restriction enzyme APOBEC3A. *J Biol Chem* **2012,** *287* (41), 34801-8.

179. Bogerd, H. P.; Wiegand, H. L.; Hulme, A. E.; Garcia-Perez, J. L.; O'Shea, K. S.; Moran, J. V.; Cullen, B. R., Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. *Proc Natl Acad Sci U S A* **2006,** *103* (23), 8780-5.
180. Vartanian, J. P.; Guetard, D.; Henry, M.; Wain-Hobson, S., Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. *Science* **2008,** *320* (5873), 230-3.
181. Pham, P.; Landolph, A.; Mendez, C.; Li, N.; Goodman, M. F., A biochemical analysis linking APOBEC3A to disparate HIV-1 restriction and skin cancer. *J Biol Chem* **2013,** *288* (41), 29294-304.
182. Love, R. P.; Xu, H.; Chelico, L., Biochemical analysis of hypermutation by the deoxycytidine deaminase APOBEC3A. *J Biol Chem* **2012,** *287* (36), 30812-22.
183. Harjes, S.; Solomon, W. C.; Li, M.; Chen, K. M.; Harjes, E.; Harris, R. S.; Matsuo, H., Impact of H216 on the DNA binding and catalytic activities of the HIV restriction factor APOBEC3G. *J. Virol.* **2013,** *87* (12), 7008-14.
184. Churchill, C. D.; Wetmore, S. D., Noncovalent interactions involving histidine: the effect of charge on pi-pi stacking and T-shaped interactions with the DNA nucleobases. *J Phys Chem B* **2009,** *113* (49), 16046-58.
185. Losey, H. C.; Ruthenburg, A. J.; Verdine, G. L., Crystal structure of Staphylococcus aureus tRNA adenosine deaminase TadA in complex with RNA. *Nat Struct Mol Biol* **2006,** *13* (2), 153-9.
186. Johnson, A. T.; Wiest, O., Structure and dynamics of poly(T) single-strand DNA: implications toward CPD formation. *J Phys Chem B* **2007,** *111* (51), 14398-404.
187. Sharma, S.; Patnaik, S. K.; Taggart, R. T.; Baysal, B. E., The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. *Sci Rep* **2016,** *6*, 39100.
188. Heinis, C., Drug discovery: tools and rules for macrocycles. *Nat Chem Biol* **2014,** *10* (9), 696-8.
189. Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **2004,** *60* (Pt 12 Pt 1), 2126-32.
190. Malim, M. H., APOBEC proteins and intrinsic resistance to HIV-1 infection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2009,** *364* (1517), 675-87.
191. Chiu, Y. L.; Greene, W. C., The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu. Rev. Immunol.* **2008,** *26*, 317-53.
192. Goila-Gaur, R.; Strebel, K., HIV-1 Vif, APOBEC, and intrinsic immunity. *Retrovirology* **2008,** *5*, 51.
193. Feng, Y.; Baig, T. T.; Love, R. P.; Chelico, L., Suppression of APOBEC3-mediated restriction of HIV-1 by Vif. *Front. Microbiol.* **2014,** *5*, 450.
194. Bishop, K. N.; Holmes, R. K.; Sheehy, A. M.; Davidson, N. O.; Cho, S. J.; Malim, M. H., Cytidine deamination of retroviral DNA by diverse APOBEC proteins. *Curr. Biol.* **2004,** *14* (15), 1392-6.
195. Chiu, Y. L.; Soros, V. B.; Kreisberg, J. F.; Stopak, K.; Yonemoto, W.; Greene, W. C., Cellular APOBEC3G restricts HIV-1 infection in resting CD4+ T cells. *Nature* **2005,** *435* (7038), 108-14.

196. Khan, M. A.; Goila-Gaur, R.; Opi, S.; Miyagi, E.; Takeuchi, H.; Kao, S.; Strebel, K., Analysis of the contribution of cellular and viral RNA to the packaging of APOBEC3G into HIV-1 virions. *Retrovirology* **2007,** *4*, 48.
197. Iwatani, Y.; Takeuchi, H.; Strebel, K.; Levin, J. G., Biochemical activities of highly purified, catalytically active human APOBEC3G: correlation with antiviral effect. *J. Virol.* **2006,** *80* (12), 5992-6002.
198. Rausch, J. W.; Chelico, L.; Goodman, M. F.; Le Grice, S. F., Dissecting APOBEC3G substrate specificity by nucleoside analog interference. *J. Biol. Chem.* **2009,** *284* (11), 7047-58.
199. Roundtree, I. A.; Evans, M. E.; Pan, T.; He, C., Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **2017,** *169* (7), 1187-1200.
200. Qiao, Q.; Wang, L.; Meng, F. L.; Hwang, J. K.; Alt, F. W.; Wu, H., AID Recognizes Structured DNA for Class Switch Recombination. *Mol. Cell* **2017,** *67* (3), 361-373 e4.
201. Nabel, C. S.; Lee, J. W.; Wang, L. C.; Kohli, R. M., Nucleic acid determinants for selective deamination of DNA over RNA by activation-induced deaminase. *Proc. Natl. Acad. Sci. U. S. A.* **2013,** *110* (35), 14225-30.
202. Navaratnam, N.; Morrison, J. R.; Bhattacharya, S.; Patel, D.; Funahashi, T.; Giannoni, F.; Teng, B. B.; Davidson, N. O.; Scott, J., The p27 catalytic subunit of the apolipoprotein B mRNA editing enzyme is a cytidine deaminase. *J. Biol. Chem.* **1993,** *268* (28), 20709-12.
203. Mehta, A.; Kinter, M. T.; Sherman, N. E.; Driscoll, D. M., Molecular cloning of apobec-1 complementation factor, a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. *Mol Cell Biol* **2000,** *20* (5), 1846-54.
204. Lellek, H.; Kirsten, R.; Diehl, I.; Apostel, F.; Buck, F.; Greeve, J., Purification and molecular cloning of a novel essential component of the apolipoprotein B mRNA editing enzyme-complex. *J. Biol. Chem.* **2000,** *275* (26), 19848-56.
205. Larijani, M.; Martin, A., The biochemistry of activation-induced deaminase and its physiological functions. *Semin. Immunol.* **2012,** *24* (4), 255-63.
206. Casellas, R.; Basu, U.; Yewdell, W. T.; Chaudhuri, J.; Robbiani, D. F.; Di Noia, J. M., Mutations, kataegis and translocations in B cells: understanding AID promiscuous activity. *Nat. Rev. Immunol.* **2016,** *16* (3), 164-76.
207. Nowarski, R.; Britan-Rosich, E.; Shiloach, T.; Kotler, M., Hypermutation by intersegmental transfer of APOBEC3G cytidine deaminase. *Nat. Struct. Mol. Biol.* **2008,** *15* (10), 1059-66.
208. Chiba, J.; Kouno, T.; Aoki, S.; Sato, H.; Zhang, J.; Matsuo, H.; Inouye, M., Electrochemical direct detection of DNA deamination catalyzed by APOBEC3G. *Chem. Commun. (Camb.)* **2012,** *48* (99), 12115-7.
209. Carpenter, M. A.; Rajagurubandara, E.; Wijesinghe, P.; Bhagwat, A. S., Determinants of sequence-specificity within human AID and APOBEC3G. *DNA Repair (Amst)* **2010,** *9* (5), 579-87.
210. Chelico, L.; Pham, P.; Calabrese, P.; Goodman, M. F., APOBEC3G DNA deaminase acts processively 3' --> 5' on single-stranded DNA. *Nat. Struct. Mol. Biol.* **2006**.

211. Williams, A. A.; Darwanto, A.; Theruvathu, J. A.; Burdzy, A.; Neidigh, J. W.; Sowers, L. C., Impact of sugar pucker on base pair and mispair stability. *Biochemistry* **2009,** *48* (50), 11994-2004.
212. Wang, Y.; Schmitt, K.; Guo, K.; Santiago, M. L.; Stephens, E. B., Role of the single deaminase domain APOBEC3A in virus restriction, retrotransposition, DNA damage and cancer. *J. Gen. Virol.* **2016,** *97* (1), 1-17.
213. Xiang, S.; Short, S. A.; Wolfenden, R.; Carter, C. W., Jr., Transition-state selectivity for a single hydroxyl group during catalysis by cytidine deaminase. *Biochemistry* **1995,** *34* (14), 4516-23.
214. Xiang, S.; Short, S. A.; Wolfenden, R.; Carter, C. W., Jr., The structure of the cytidine deaminase-product complex provides evidence for efficient proton transfer and ground-state destabilization. *Biochemistry* **1997,** *36* (16), 4768-74.
215. Batsanov, S. S., Van der Waals radii of elements. *Inorg. Mater.* **2001,** *37* (9), 871-885.
216. Watts, J. K.; Choubdar, N.; Sadalapure, K.; Robert, F.; Wahba, A. S.; Pelletier, J.; Pinto, B. M.; Damha, M. J., 2 '-Fluoro-4 '-thioarabino-modified oligonucleotides: conformational switches linked to siRNA activity. *Nucleic Acids Res.* **2007,** *35* (5), 1441-1451.
217. Wijesinghe, P.; Bhagwat, A. S., Efficient deamination of 5-methylcytosines in DNA by human APOBEC3A, but not by AID or APOBEC3G. *Nucleic Acids Res.* **2012,** *40* (18), 9206-17.
218. Suspene, R.; Aynaud, M. M.; Vartanian, J. P.; Wain-Hobson, S., Efficient deamination of 5-methylcytidine and 5-substituted cytidine residues in DNA by human APOBEC3A cytidine deaminase. *PLoS One* **2013,** *8* (6), e63461.
219. Nabel, C. S.; Jia, H.; Ye, Y.; Shen, L.; Goldschmidt, H. L.; Stivers, J. T.; Zhang, Y.; Kohli, R. M., AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat. Chem. Biol.* **2012,** *8* (9), 751-8.
220. Wienken, C. J.; Baaske, P.; Rothbauer, U.; Braun, D.; Duhr, S., Protein-binding assays in biological liquids using microscale thermophoresis. *Nat Commun* **2010,** *1*, 100.
221. An, P.; Johnson, R.; Phair, J.; Kirk, G. D.; Yu, X. F.; Donfield, S.; Buchbinder, S.; Goedert, J. J.; Winkler, C. A., APOBEC3B deletion and risk of HIV-1 acquisition. *J. Infect. Dis.* **2009,** *200* (7), 1054-8.
222. Nik-Zainal, S.; Wedge, D. C.; Alexandrov, L. B.; Petljak, M.; Butler, A. P.; Bolli, N.; Davies, H. R.; Knappskog, S.; Martin, S.; Papaemmanuil, E.; Ramakrishna, M.; Shlien, A.; Simonic, I.; Xue, Y.; Tyler-Smith, C.; Campbell, P. J.; Stratton, M. R., Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet.* **2014,** *46* (5), 487-91.
223. Kvach, M. V.; Barzak, F. M.; Harjes, S.; Schares, H. A. M.; Jameson, G. B.; Ayoub, A. M.; Moorthy, R.; Aihara, H.; Harris, R. S.; Filichev, V. V.; Harki, D. A.; Harjes, E., Inhibiting APOBEC3 Activity with Single-Stranded DNA Containing 2'-Deoxyzebularine Analogues. *Biochemistry* **2019,** *58* (5), 391-400.
224. Barzak, F. M.; Harjes, S.; Kvach, M. V.; Kurup, H. M.; Jameson, G. B.; Filichev, V. V.; Harjes, E., Selective inhibition of APOBEC3 enzymes by single-stranded DNAs containing 2'-deoxyzebularine. *Org Biomol Chem* **2019,** *17* (43), 9435-9441.

225. Kvach, M. V.; Barzak, F. M.; Harjes, S.; Schares, H. A. M.; Kurup, H. M.; Jones, K. F.; Sutton, L.; Donahue, J.; D'Aquila, R. T.; Jameson, G. B.; Harki, D. A.; Krause, K. L.; Harjes, E.; Filichev, V. V., Differential Inhibition of APOBEC3 DNA-Mutator Isozymes by Fluoro- and Non-Fluoro-Substituted 2'-Deoxyzebularine Embedded in Single-Stranded DNA. *Chembiochem* **2019**.
226. Grunewald, J.; Zhou, R.; Garcia, S. P.; Iyer, S.; Lareau, C. A.; Aryee, M. J.; Joung, J. K., Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **2019,** *569* (7756), 433-437.
227. OhAinle, M.; Kerns, J. A.; Malik, H. S.; Emerman, M., Adaptive evolution and antiviral activity of the conserved mammalian cytidine deaminase APOBEC3H. *J. Virol.* **2006,** *80* (8), 3853-62.
228. Wiegand, H. L.; Doehle, B. P.; Bogerd, H. P.; Cullen, B. R., A second human antiretroviral factor, APOBEC3F, is suppressed by the HIV-1 and HIV-2 Vif proteins. *EMBO J.* **2004,** *23* (12), 2451-8.
229. Conticello, S. G.; Harris, R. S.; Neuberger, M. S., The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. *Curr. Biol.* **2003,** *13* (22), 2009-13.
230. Marin, M.; Rose, K. M.; Kozak, S. L.; Kabat, D., HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. *Nat. Med.* **2003,** *9* (11), 1398-403.
231. Sheehy, A. M.; Gaddis, N. C.; Malim, M. H., The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. *Nat. Med.* **2003,** *9* (11), 1404-7.
232. Yu, X.; Yu, Y.; Liu, B.; Luo, K.; Kong, W.; Mao, P.; Yu, X. F., Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. *Science* **2003,** *302* (5647), 1056-60.
233. Liu, B.; Sarkis, P. T.; Luo, K.; Yu, Y.; Yu, X. F., Regulation of Apobec3F and human immunodeficiency virus type 1 Vif by Vif-Cul5-ElonB/C E3 ubiquitin ligase. *J. Virol.* **2005,** *79* (15), 9579-87.
234. Shirakawa, K.; Takaori-Kondo, A.; Kobayashi, M.; Tomonaga, M.; Izumi, T.; Fukunaga, K.; Sasada, A.; Abudu, A.; Miyauchi, Y.; Akari, H.; Iwai, K.; Uchiyama, T., Ubiquitination of APOBEC3 proteins by the Vif-Cullin5-ElonginB-ElonginC complex. *Virology* **2006,** *344* (2), 263-6.
235. Svarovskaia, E. S.; Xu, H.; Mbisa, J. L.; Barr, R.; Gorelick, R. J.; Ono, A.; Freed, E. O.; Hu, W. S.; Pathak, V. K., Human apolipoprotein B mRNA-editing enzyme-catalytic polypeptide-like 3G (APOBEC3G) is incorporated into HIV-1 virions through interactions with viral and nonviral RNAs. *J. Biol. Chem.* **2004,** *279* (34), 35822-8.
236. Zennou, V.; Perez-Caballero, D.; Gottlinger, H.; Bieniasz, P. D., APOBEC3G incorporation into human immunodeficiency virus type 1 particles. *J. Virol.* **2004,** *78* (21), 12058-61.
237. Kao, S.; Khan, M. A.; Miyagi, E.; Plishka, R.; Buckler-White, A.; Strebel, K., The human immunodeficiency virus type 1 Vif protein reduces intracellular expression and inhibits packaging of APOBEC3G (CEM15), a cellular inhibitor of virus infectivity. *J. Virol.* **2003,** *77* (21), 11398-407.
238. Simon, J. H.; Miller, D. L.; Fouchier, R. A.; Soares, M. A.; Peden, K. W.; Malim, M. H., The regulation of primate immunodeficiency virus infectivity by Vif is cell species

restricted: a role for Vif in determining virus host range and cross-species transmission. *EMBO J.* **1998,** *17* (5), 1259-67.
239. Mariani, R.; Chen, D.; Schrofelbauer, B.; Navarro, F.; Konig, R.; Bollman, B.; Munk, C.; Nymark-McMahon, H.; Landau, N. R., Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. *Cell* **2003,** *114* (1), 21-31.
240. Lecossier, D.; Bouchonnet, F.; Clavel, F.; Hance, A. J., Hypermutation of HIV-1 DNA in the absence of the Vif protein. *Science* **2003,** *300* (5622), 1112.
241. Mbisa, J. L.; Barr, R.; Thomas, J. A.; Vandegraaff, N.; Dorweiler, I. J.; Svarovskaia, E. S.; Brown, W. L.; Mansky, L. M.; Gorelick, R. J.; Harris, R. S.; Engelman, A.; Pathak, V. K., Human immunodeficiency virus type 1 cDNAs produced in the presence of APOBEC3G exhibit defects in plus-strand DNA transfer and integration. *J. Virol.* **2007,** *81* (13), 7099-110.
242. Okada, A.; Iwatani, Y., APOBEC3G-Mediated G-to-A Hypermutation of the HIV-1 Genome: The Missing Link in Antiviral Molecular Mechanisms. *Front. Microbiol.* **2016,** *7*, 2027.
243. Guo, F.; Cen, S.; Niu, M.; Saadatmand, J.; Kleiman, L., Inhibition of tRNA(3)(Lys)-primed reverse transcription by human APOBEC3G during human immunodeficiency virus type 1 replication. *J. Virol.* **2006,** *80* (23), 11710-22.
244. Bishop, K. N.; Verma, M.; Kim, E. Y.; Wolinsky, S. M.; Malim, M. H., APOBEC3G inhibits elongation of HIV-1 reverse transcripts. *PLoS Pathog.* **2008,** *4* (12), e1000231.
245. Hache, G.; Liddament, M. T.; Harris, R. S., The retroviral hypermutation specificity of APOBEC3F and APOBEC3G is governed by the C-terminal DNA cytosine deaminase domain. *J. Biol. Chem.* **2005,** *280* (12), 10920-4.
246. Gooch, B. D.; Cullen, B. R., Functional domain organization of human APOBEC3G. *Virology* **2008,** *379* (1), 118-24.
247. Burnett, A.; Spearman, P., APOBEC3G multimers are recruited to the plasma membrane for packaging into human immunodeficiency virus type 1 virus-like particles in an RNA-dependent process requiring the NC basic linker. *J. Virol.* **2007,** *81* (10), 5000-13.
248. Huthoff, H.; Malim, M. H., Identification of amino acid residues in APOBEC3G required for regulation by human immunodeficiency virus type 1 Vif and Virion encapsidation. *J. Virol.* **2007,** *81* (8), 3807-15.
249. Desimmie, B. A.; Delviks-Frankenberrry, K. A.; Burdick, R. C.; Qi, D.; Izumi, T.; Pathak, V. K., Multiple APOBEC3 restriction factors for HIV-1 and one Vif to rule them all. *J. Mol. Biol.* **2014,** *426* (6), 1220-45.
250. Otwinowski, Z.; Minor, W., Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **1997,** *276*, 307-26.
251. Bunkoczi, G.; Echols, N.; McCoy, A. J.; Oeffner, R. D.; Adams, P. D.; Read, R. J., Phaser.MRage: automated molecular replacement. *Acta Crystallogr. D Biol. Crystallogr.* **2013,** *69* (Pt 11), 2276-86.
252. Murshudov, G. N.; Skubak, P.; Lebedev, A. A.; Pannu, N. S.; Steiner, R. A.; Nicholls, R. A.; Winn, M. D.; Long, F.; Vagin, A. A., REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **2011,** *67* (Pt 4), 355-67.

253. Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Echols, N.; Headd, J. J.; Hung, L. W.; Jain, S.; Kapral, G. J.; Grosse Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R. D.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H., The Phenix software for automated determination of macromolecular structures. *Methods* **2011,** *55* (1), 94-106.
254. Echols, N.; Grosse-Kunstleve, R. W.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Headd, J. J.; McCoy, A. J.; Moriarty, N. W.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Adams, P. D., Graphical tools for macromolecular crystallography in PHENIX. *J. Appl. Crystallogr.* **2012,** *45* (Pt 3), 581-586.
255. Chen, V. B.; Arendall, W. B., 3rd; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **2010,** *66* (Pt 1), 12-21.
256. Labarga, A.; Valentin, F.; Anderson, M.; Lopez, R., Web services at the European bioinformatics institute. *Nucleic Acids Res.* **2007,** *35* (Web Server issue), W6-11.
257. Yee, J. K.; Miyanohara, A.; LaPorte, P.; Bouic, K.; Burns, J. C.; Friedmann, T., A general method for the generation of high-titer, pantropic retroviral vectors: highly efficient infection of primary hepatocytes. *Proc. Natl. Acad. Sci. U. S. A.* **1994,** *91* (20), 9564-8.
258. Unutmaz, D.; KewalRamani, V. N.; Marmon, S.; Littman, D. R., Cytokine signals are sufficient for HIV-1 infection of resting human T lymphocytes. *J. Exp. Med.* **1999,** *189* (11), 1735-46.
259. Smith, J. L.; Izumi, T.; Borbet, T. C.; Hagedorn, A. N.; Pathak, V. K., HIV-1 and HIV-2 Vif interact with human APOBEC3 proteins using completely different determinants. *J. Virol.* **2014,** *88* (17), 9893-908.
260. Russell, R. A.; Pathak, V. K., Identification of two distinct human immunodeficiency virus type 1 Vif determinants critical for interactions with human APOBEC3G and APOBEC3F. *J. Virol.* **2007,** *81* (15), 8201-10.
261. Wei, X.; Decker, J. M.; Liu, H.; Zhang, Z.; Arani, R. B.; Kilby, J. M.; Saag, M. S.; Wu, X.; Shaw, G. M.; Kappes, J. C., Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy. *Antimicrob Agents Chemother* **2002,** *46* (6), 1896-905.
262. Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W., Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **2013,** *27* (3), 221-34.
263. Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A., PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004,** *32* (Web Server issue), W665-7.
264. Huthoff, H.; Autore, F.; Gallois-Montbrun, S.; Fraternali, F.; Malim, M. H., RNA-dependent oligomerization of APOBEC3G is required for restriction of HIV-1. *PLoS Pathog.* **2009,** *5* (3), e1000330.
265. Fukuda, H.; Li, S.; Sardo, L.; Smith, J. L.; Yamashita, K.; Sarca, A. D.; Shirakawa, K.; Standley, D. M.; Takaori-Kondo, A.; Izumi, T., Structural Determinants of the APOBEC3G N-Terminal Domain for HIV-1 RNA Association. *Front Cell Infect Microbiol* **2019,** *9*, 129.

266. Pollpeter, D.; Parsons, M.; Sobala, A. E.; Coxhead, S.; Lang, R. D.; Bruns, A. M.; Papaioannou, S.; McDonnell, J. M.; Apolonia, L.; Chowdhury, J. A.; Horvath, C. M.; Malim, M. H., Deep sequencing of HIV-1 reverse transcripts reveals the multifaceted antiviral functions of APOBEC3G. *Nat Microbiol* **2018,** *3* (2), 220-233.
267. Friew, Y. N.; Boyko, V.; Hu, W. S.; Pathak, V. K., Intracellular interactions between APOBEC3G, RNA, and HIV-1 Gag: APOBEC3G multimerization is dependent on its association with RNA. *Retrovirology* **2009,** *6*, 56.
268. Morse, M.; Huo, R.; Feng, Y.; Rouzina, I.; Chelico, L.; Williams, M. C., Dimerization regulates both deaminase-dependent and deaminase-independent HIV-1 restriction by APOBEC3G. *Nat Commun* **2017,** *8* (1), 597.
269. Mangeat, B.; Turelli, P.; Liao, S.; Trono, D., A single amino acid determinant governs the species-specific sensitivity of APOBEC3G to Vif action. *J. Biol. Chem.* **2004,** *279* (15), 14481-3.
270. Mehle, A.; Strack, B.; Ancuta, P.; Zhang, C.; McPike, M.; Gabuzda, D., Vif overcomes the innate antiviral activity of APOBEC3G by promoting its degradation in the ubiquitin-proteasome pathway. *J. Biol. Chem.* **2004,** *279* (9), 7792-8.
271. Schrofelbauer, B.; Chen, D.; Landau, N. R., A single amino acid of APOBEC3G controls its species-specific interaction with virion infectivity factor (Vif). *Proc. Natl. Acad. Sci. U. S. A.* **2004,** *101* (11), 3927-32.
272. Bogerd, H. P.; Doehle, B. P.; Wiegand, H. L.; Cullen, B. R., A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor. *Proc. Natl. Acad. Sci. U. S. A.* **2004,** *101* (11), 3770-4.
273. Xu, H.; Svarovskaia, E. S.; Barr, R.; Zhang, Y.; Khan, M. A.; Strebel, K.; Pathak, V. K., A single amino acid substitution in human APOBEC3G antiretroviral enzyme confers resistance to HIV-1 virion infectivity factor-induced depletion. *Proc. Natl. Acad. Sci. U. S. A.* **2004,** *101* (15), 5652-7.