1-1-2013

Nucleic Acid Determinants of Cytosine Deamination by Aid/Apobec Enzymes in Immunity and Epigenetics

Christopher Nabel
University of Pennsylvania, cnabel@mail.med.upenn.edu

Follow this and additional works at: http://repository.upenn.edu/edissertations
Part of the Biochemistry Commons, and the Molecular Biology Commons

Recommended Citation
Nabel, Christopher, "Nucleic Acid Determinants of Cytosine Deamination by Aid/Apobec Enzymes in Immunity and Epigenetics" (2013). Publicly Accessible Penn Dissertations. 903.
http://repository.upenn.edu/edissertations/903

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/edissertations/903
For more information, please contact libraryrepository@pobox.upenn.edu.
Nucleic Acid Determinants of Cytosine Deamination by Aid/Apobec Enzymes in Immunity and Epigenetics

Abstract
A multitude of functions have evolved around cytosine within DNA, endowing the base with physiological significance beyond simple information storage. This versatility arises from enzymes that chemically modify cytosine to expand the potential of the genome. Cytosine can be methylated, oxidized, and deaminated to modulate transcription and immunologic diversity. At the crossroads of these modifications sit the AID/APOBEC family deaminases, which accomplish diverse functions ranging from antibody diversification and innate immunity to mRNA editing. In addition, novel roles have been proposed in oncogenesis and DNA demethylation. Behind these established and emerging physiologic activities remain important questions about the substrate specificity of these deaminases, reflecting a broader need to elucidate how AID/APOBEC enzymes engage their substrates for deamination. The work here addresses this larger question by focusing on the molecular basis of two important aspects of AID/APOBEC specificity: selectivity for DNA over RNA, and biochemical plausibility of deamination-coupled demethylation. To address these questions, we have synthesized chimeric nucleic acid substrates and characterized their reactivity with AID and the rest of the APOBEC family. With regards to nucleic acid selectivity, modifications to the 2'-position of the target nucleotide sugar significantly alter AID's reactivity. Strikingly, within a substrate that is otherwise DNA, a single RNA-like 2'-hydroxyl substitution at the target cytosine is sufficient to compromise deamination. Alternatively, modifications that favor a DNA-like conformation (or sugar pucker) are compatible with deamination. Inversely, with unreactive 2'-fluoro-RNA substrates, AID's deaminase activity was rescued by introducing a trinucleotide DNA patch spanning the target cytosine and two upstream nucleotides. With regards to demethylation, AID has substantially reduced activity on 5-methylcytosine relative to cytosine, its canonical substrate, and no detectable deamination of 5-hydroxymethylcytosine. This finding is explained by the reactivity of a series of modified substrates, where steric bulk at the 5-position was increasingly detrimental to deamination. We found that these nucleic acid determinants, localized to the nucleotide base and sugar, are conserved across the entire AID/APOBEC family. Taken together, we consolidate these findings into a unifying, mechanistic model for substrate engagement that clarifies the established and proposed functions of the AID/APOBEC family.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Cell & Molecular Biology

First Advisor
Rahul M. Kohli

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/903
NUCLEIC ACID DETERMINANTS OF CYTOSINE DEAMINATION BY AID/APOBEC ENZYMES
IN IMMUNITY AND EPIGENETICS

Christopher S. Nabel

A DISSERTATION

in

Cell and Molecular Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

______________________________
Rahul M. Kohli
Assistant Professor of Medicine

Graduate Group Chairperson

______________________________
Daniel S. Kessler, Associate Professor of Cell and Developmental Biology

Dissertation Committee
Marisa S. Bartolomei, Professor of Cell and Developmental Biology
Craig H. Bassing, Associate Professor of Pathology and Laboratory Medicine
Mitchell A. Lazar, Sylvan H. Eisman Professor of Medicine
Gregory D. Van Duyne, Jacob Gershon-Cohen Professor of Medical Science
Matthew D. Weitzman, Associate Professor of Pathology and Laboratory Medicine
NUCLEIC ACID DETERMINANTS OF CYTOSINE DEAMINATION BY AID/APOBEC ENZYMES IN IMMUNITY AND EPIGENETICS

COPYRIGHT

2013

Christopher Sheild Nabel

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License

To view a copy of this license, visit

http://creativecommons.org/licenses/by-ny-sa/2.0/
ACKNOWLEDGMENT

I entered graduate school with an open mind, content to take things as they come and enjoy the ride. Along the way, I have been fortunate in many regards, and there is much for which I am grateful.

As a student in the MD-PhD combined degree program, I have benefitted from the unwavering support of Skip Brass, Gary Koretzky, Maggie Krall and Maureen Kirsch. They encourage the pursuit of educational opportunities and personal growth, and I am grateful to be a part of the community that they so carefully foster and support.

I am thankful for the measured advice of my thesis committee members: Marisa Bartolomei, Craig Bassing, Mitch Lazar, Matt Weitzman, and Greg Van Duyne. I was drawn to these advisors for their collective wisdom, but I have been most fortunate for their openness and enthusiasm for the scientific process.

As a member of the Kohli lab, I am grateful for the camaraderie of my colleagues. The good spirits of my fellow labmates made joining the Kohli lab in its infancy a lot easier. In particular, I am honored to have shared my time in the lab with Kiran Gajula, Charlie Mo, and Danny Crawford, who have helped to shape the lab environment during my tenure.

I am particularly thankful for the opportunity to be Rahul Kohli’s first graduate student. I have grown as a scientist and a person through his patience, encouragement, confidence, and high standards. Above all, I have greatly enjoyed the scientific inquiry that we have shared. It is sad that my time in the lab must come to an end, but I am excited to put to good use everything I have learned from Rahul.

The support of my family has meant a lot to me over the past few years. My mom and dad have quietly fostered my passion for science without being heavy-handed, and my sisters have always encouraged me to be happy. I am fortunate to have their support and love.

Lastly, I am blessed to have shared the past few years with Mary. There have been highs and there have been lows, and I couldn’t be luckier than to go through it all with her by my side.
ABSTRACT

NUCLEIC ACID DETERMINANTS OF CYTOSINE DEAMINATION BY AID/APOBEC ENZYMES IN IMMUNITY AND EPIGENETICS

Christopher S. Nabel
Rahul M. Kohli

A multitude of functions have evolved around cytosine within DNA, endowing the base with physiological significance beyond simple information storage. This versatility arises from enzymes that chemically modify cytosine to expand the potential of the genome. Cytosine can be methylated, oxidized, and deaminated to modulate transcription and immunologic diversity. At the crossroads of these modifications sit the AID/APOBEC family deaminases, which accomplish diverse functions ranging from antibody diversification and innate immunity to mRNA editing. In addition, novel roles have been proposed in oncogenesis and DNA demethylation. Behind these established and emerging physiologic activities remain important questions about the substrate specificity of these deaminases, reflecting a broader need to elucidate how AID/APOBEC enzymes engage their substrates for deamination. The work here addresses this larger question by focusing on the molecular basis of two important aspects of AID/APOBEC specificity: selectivity for DNA over RNA, and biochemical plausibility of deamination-coupled demethylation. To address these questions, we have synthesized chimeric nucleic acid substrates and characterized their reactivity with AID and the rest of the APOBEC family. With regards to nucleic acid selectivity, modifications to the 2’-position of the target nucleotide sugar significantly alter AID’s reactivity. Strikingly, within a substrate that is otherwise DNA, a single RNA-like 2’-hydroxyl substitution at the target cytosine is sufficient to compromise deamination. Alternatively, modifications that favor a DNA-like conformation (or sugar pucker) are compatible with deamination. Inversely, with unreactive 2’-fluoro-RNA substrates, AID’s deaminase activity was rescued by introducing a trinucleotide DNA patch spanning the target cytosine and two upstream nucleotides. With regards to demethylation, AID has substantially reduced activity on 5-methylcytosine relative to cytosine, its canonical substrate, and no detectable deamination of 5-
This finding is explained by the reactivity of a series of modified substrates, where steric bulk at the 5-position was increasingly detrimental to deamination. We found that these nucleic acid determinants, localized to the nucleotide base and sugar, are conserved across the entire AID/APOBEC family. Taken together, we consolidate these findings into a unifying, mechanistic model for substrate engagement that clarifies the established and proposed functions of the AID/APOBEC family.
# TABLE OF CONTENTS

**Acknowledgment** ........................................................................................................ iii
**Abstract** ......................................................................................................................... iv
**List of Tables** .................................................................................................................. vi
**List of Illustrations** ......................................................................................................... viii

**Chapter 1: Introduction** .................................................................................................. 1
  1.1: Cytosine modifications in DNA ............................................................................... 1
  1.1.1: Chemical modification of cytosine transforms genomic potential ...................... 1
  1.1.2: Chemical reactivity of cytosine ........................................................................... 2
  1.1.3: Cytosine modification activates DNA repair pathways ..................................... 4
  1.2: Mammalian cytosine modification pathways ........................................................... 6
  1.2.1: Cytosine methylation contributes to development and oncogenesis ............... 6
  1.2.2: DNA demethylation and regeneration of unmodified cytosine ....................... 8
  1.2.3: Oxidation of Methylcytosine also contributes to development and oncogenesis ............................................................................................................................ 9
  1.2.4: Cytosine deamination contributes to adaptive and innate immunity .......... 11
  1.3: AID/APOBEC family of cytidine deaminases .......................................................... 14
  1.3.1: Members of the AID/APOBEC family ................................................................. 14
  1.3.2: Evolution of the AID/APOBEC family ............................................................... 17
  1.3.3: Catalytic determinants of the AID/APOBEC family ......................................... 18
  1.3.4: Structural characteristics of the AID/APOBEC family .................................... 21
  1.4: Thesis objectives ........................................................................................................ 24

**Chapter 2: Materials and Methods** .............................................................................. 26
  2.1: Protein expression and purification ...................................................................... 26
  2.1.1: AID/APOBEC expression and purification ..................................................... 26
  2.1.2: MBP expression and purification .................................................................... 27
  2.2: Synthesis of chimeric oligonucleotides substrates .............................................. 28
  2.3: Biochemical deamination assays ........................................................................... 34
  2.3.1: AID/APOBEC deaminase incubations ............................................................. 34
  2.3.2: DNA glycosylase based deamination assays ..................................................... 34
  2.3.3: Restriction endonuclease based deamination assay ........................................ 35
  2.3.4: Reverse Transcriptase based deamination assay ............................................. 36
  2.4: Protein binding assays ............................................................................................ 37
  2.5: Cellular deamination assays ................................................................................... 37
  2.5.1: Transient expression of deaminases and other DNA modifying enzymes .... 37
  2.5.2: Mass spectrometric experiments .................................................................... 38
  2.5.3: Deaminase and base excision activity of nuclear lysates ................................ 38

**Chapter 3: Nucleic acid determinants for selective deamination of DNA over RNA by activation-induced deaminase and APOBEC1** ................................................. 40
  3.1: Introduction ............................................................................................................... 40
  3.1.1: DNA deamination model for role of activation-induced deaminase in antibody maturation ................................................................................................................................. 40
  3.1.2: Experimental approach ...................................................................................... 41
  3.2: Results ...................................................................................................................... 42
  3.2.1: 2'-substituents of target cytosine regulate deamination by AID .................... 42
3.2.2: 2’-substituents of target cytosine regulate deamination by APOBEC1
3.2.3: Minimal DNA requirements for deamination by AID
3.3: Discussion
   3.3.1: Mechanistic basis for AID’s inherent preference for DNA deamination
   3.3.2: AID’s inherent preferences support the DNA deamination model of antibody maturation
   3.3.3: APOBEC1’s inherent preference for DNA contrasts with its physiologic RNA activity
Chapter 4: AID/APOBEC deaminases discriminate against modified cytosines implicated in DNA demethylation
   4.1: Introduction
      4.1.1: Proposed mechanisms of DNA demethylation
      4.1.2: Evidence in support of deamination-dependent demethylation and remaining questions
   4.2: Results
      4.2.1: AID/APOBECs preferentially deaminate unmodified cytosine
      4.2.2: Deamination decreases with steric bulk at C5 position
      4.2.3: Deaminases do not perturb levels of modified epigenetic bases in genomic DNA
   4.3: Discussion
      4.3.1: Re-evaluation of deamination-dependent demethylation pathways
      4.3.2: Re-evaluation of potential pathways for DNA demethylation
Chapter 5: Conclusions and Future Directions
   5.1: Model for substrate recognition and deamination by AID
      5.1.1: Description of model
      5.1.2: Remaining questions regarding deamination model
   5.2: Future biochemical studies of deamination
      5.2.1: Further studies of DNA deamination with A3A
      5.2.2: Future biochemical studies with chimeric deaminases
   5.3: Future cellular studies of deamination
      5.3.1: Evaluation of proposed role of AID in reprogramming of iPS cells
      5.3.2: Evaluation of proposed role of AID/APOBECs in cancer
   5.4: Concluding remarks
Bibliography
LIST OF TABLES

Table 1: Oligonucleotides ........................................................................................................ 30
LIST OF ILLUSTRATIONS

Figure 1: Cytosine as the Genomic “Wild Card” ...................................................... 1
Figure 2: Mechanism of DNA methylation ............................................................... 3
Figure 3: Mechanism of DNA Oxidation ................................................................. 3
Figure 4: Mechanism of DNA Deamination ............................................................ 4
Figure 5: Mechanism of base excision ................................................................. 5
Figure 6: Cytosine methylation and hydroxymethylation regulate transcription .... 10
Figure 7: Cytosine deamination in antibody maturation ........................................ 12
Figure 8: AID/APOBEC family members ............................................................... 14
Figure 9: APOBEC3 locus expansion from the mouse genome to the human genome . 16
Figure 10: Catalytic determinants of AID/APOBEC deamination ......................... 19
Figure 11: Structural characteristics of AID/APOBEC deaminases ....................... 23
Figure 12: Structural basis for interaction between hotspot targeting loop and substrate nucleic acid ................................................................. 24
Figure 13: SDS-PAGE of partially purified AID/APOBEC family members ............ 27
Figure 14: SDS-PAGE of MBP control protein ....................................................... 28
Figure 15: Nucleotide sugar pucker as a potential basis for DNA selectivity .......... 42
Figure 16: Restriction-endonuclease-based assay for deamination ...................... 43
Figure 17: 2′-substitution of target cytosine is sufficient to disrupt AID deaminase activity ........................................................................................................ 45
Figure 18: Incubation of AID with chimeric substrates in alternative sequence contexts 47
Figure 19: Full-length AID demonstrates similar reactivity compared to truncated version ................................................................. 48
Figure 20: Quantitative assessment of AID deamination of chimeric DNA substrates ... 50
Figure 21: AID binding to chimeric DNA substrates ............................................... 51
Figure 22: 2′-substitution of target cytosine disrupts APOBEC1 deaminase activity ..... 53
Figure 23: Quantitative analysis of APOBEC1 deaminase activity against chimeric DNA substrates ........................................................................................................ 54
Figure 24: APOBEC1 binding to chimeric DNA substrates ....................................... 54
Figure 25: Reverse-transcriptase-based assay for deamination of chimeric 2′-F-RNA substrates ........................................................................................................ 56
Figure 26: Rescue of deamination of chimeric 2′-F-RNA oligonucleotides requires a DNA patch from positions -2 to 0.................................................................57
Figure 27: UDG-based, alternative assay for deamination of chimeric 2′-F-RNA substrates ........................................................................................................58
Figure 28: AID binding to chimeric 2′-F-RNA substrates .............................................59
Figure 29: Proposed non-canonical role for AID/APOBEC enzymes acting on modified cytosine substrates in DNA.................................................................................65
Figure 30: Substrate and assay design to screen for deamination of substrates containing 5-modified cytosine..........................................................................................68
Figure 31: Detection limit and standard curve for deamination detection ..................69
Figure 32: AID/APOBEC enzymes preferentially deaminate unmodified cytosine ........70
Figure 33: Quantification of enzyme-dependent deamination of C-, mC-, and hmC-containing substrates ........................................................................................................72
Figure 34: Substrate and assay design for cytosines containing additional 5-substitutions .........................................................................................................................73
Figure 35: Standard curves show that DNA glycosylases are sensitive for detection of deamination of unnatural 5-modified cytosine substrates ........................................75
Figure 36: DNA deamination decreases as a function of increasing steric bulk at the 5-position of cytosine ........................................................................................................76
Figure 37: Quantification of enzyme-dependent deamination of 5-modified cytosines...77
Figure 38: Protein expression of AID/APOBECs and TDG in HEK 293T cells ...............79
Figure 39: Effects of hAID expression on genomic levels of hmU, fC, caC in HEK 293T cells ..........................................................................................................................80
Figure 40: Deamination intermediates of oxidized cytosine are not detected in genomic DNA .........................................................................................................................81
Figure 41: Uracil and not hmU is rapidly excised by HEK 293T cell nuclear extracts .... 83
Figure 42: Overexpressed hAID-DC and mAPOBEC1 are active in deamination in nuclear extracts..............................................................................................................84
Figure 43: Overexpressed hAID and mAPOBEC1 are active in deamination ...............86
Figure 44: Mechanistic model for nucleic acid targeting and selectivity .................93
CHAPTER 1: INTRODUCTION

1.1 Cytosine Modifications in DNA

1.1.1 Chemical Modification of Cytosine Transforms Genomic Potential

In the conventional view, the genome is a long polymer of A, C, G, and T, which together define and differentiate organisms. We typically think of the genome as a stable, unchanging blueprint for life. However, as life demands variety and adaptability, it has become increasingly clear that diversity within an organism is governed by the dynamic changes that occur within this genomic scaffold (1). In particular, enzymes that chemically modify cytosine introduce a physiologically important layer of complexity to the genome, beyond that seen in the primary sequence [Figure 1].

![Figure 1: Cytosine as the Genomic “Wild Card”](image)

These accessory functions to the genome mediate critical physiologic processes across all walks of life. For example, modification of DNA can help organisms distinguish self DNA from foreign DNA (2). In bacterial species, DNA methyltransferases have coevolved with partner
restriction endonucleases that share the same sequence preference. Since only host DNA is methylated, this system allows for degradation of foreign DNA by the corresponding restriction enzyme. A second adaptive role for DNA modification is to mediate the expression or silencing of genes (3). While DNA modifications share this role with histone modification enzymes, all are needed in order to properly modulate transcriptional networks. Importantly, DNA-modifying enzymes also allow for the reverse process to occur, 'resetting' the genome for proper gametogenesis or reactivation of gene expression (4). Finally, the mammalian adaptive immune system demonstrates the importance of genomic malleability. The immunoglobulin (Ig) locus is a dramatic example of how the genome is preprogrammed to foster variety, through recombination and mutation that ultimately confer an adaptive advantage (5, 6). In order to examine the relevant biological pathways, we must first introduce the enzymes in nature’s toolbox for altering cytosine within DNA.

1.1.2 Chemical Reactivity of Cytosine

In duplex DNA, the C5 and C6 positions of cytosine lie in the major groove, unobstructed by Watson-Crick interactions. The electrophilic character of the C6 position makes it a key target of modifying enzymes. For example, DNA methyltransferases (DNMTs) transiently modify C6 by attack of an active site cysteine. Methylation results from the concerted addition of a methyl group derived from S-adenosylmethionine (SAM) to the C5 position [Figure 2] (7, 8). The covalent intermediate breaks down to liberate the enzyme and generate genomic 5-methylcytosine (mC). Interestingly, in the absence of SAM, DNMTs can catalyze nonclassical reactions, such as deamination at C4 (9, 10) or the addition of aldehydes to C5 (11), raising intriguing questions about the relevance of these nonclassical functions in vivo. The epigenetic impact of C5 methylation will be discussed later.
Figure 2: Mechanism of DNA methylation. DNA methyltransferases catalyze transfer of a methyl group (shown in magenta) from S-adenosylmethionine (SAM) to the C5 position of cytosine, generating 5-methylcytosine (mC) and S-adenosylhomocysteine. Previously underappreciated, oxidative modifications of mC are also possible. In mammals, oxidation of mC is carried out by the TET family enzymes, which belong to the Fe(II)/α-ketoglutarate-dependent oxygenase family that includes histone demethylases and the DNA damage repair enzyme AlkB [Figure 3] (12, 13). Rao and colleagues initially discovered the TET family based on homology to a trypanosome enzyme known to catalyze oxidation of the exocyclic methyl group of thymine. Initially, TETs were shown to oxidize mC to 5-hydroxymethylcytosine (hmC) (12). However, more recent studies have revealed that TETs can catalyze iterative oxidation of mC. The products of iterative oxidation, 5-formylcytosine (fC) and 5-carboxylcytosine (caC), are stably detectable intermediates in genomic DNA from embryonic stem (ES) cells (14, 15). In total, the TET enzymes have provided a stable of new chemical handles whose impacts on transcriptional regulation and demethylation we will examine later.

Figure 3: Mechanism of DNA Oxidation. TET oxidases catalyze Fe(II)-dependent oxidation of mC to 5-hydroxymethylcytosine (hmC), with molecular oxygen serving as donor for oxidation
(shown in magenta), with the second oxygen atom transferred to succinate.  hmC may be further oxidized to 5-formylcytosine (fC) and 5-carboxylycytosine (caC).

The C4 position of cytosine is relatively protected while engaged in Watson-Crick pairing, but in the context of single-stranded DNA, it becomes an important site for deamination by AID/APOBEC family enzymes [Figure 4] (16). The mechanism of deamination involves activation of a zinc bound water for nucleophilic attack at C4 and generation of a tetrahedral intermediate. An active site glutamate promotes deamination of C4 and the conversion of cytosine analogues into uridine analogues (17). In addition to deamination of unmodified cytosine, some studies have suggested that mC deamination can generate thymine (16, 18). However, the evidence surrounding this possibility is conflicting (19), and the full spectrum of AID/APOBEC activity against various cytosine analogues has not yet been clarified. These questions and their impact on diversity will be explored later. The distinction between genomic malleability and instability is subtle. Deamination of cytosine and 5-methylcytosine may cause transition mutations; deamination is therefore a very relevant threat to genome stability. In response, sophisticated DNA repair machinery has evolved to ensure the integrity of DNA (18), namely, base excision repair enzymes (BER) and mismatch repair (MMR) enzymes. Interestingly, many of these ‘repair’ enzymes are exploited to support cytosine’s role in generating diversity.

Figure 4: Mechanism of DNA Deamination. AID/APOBEC deaminases catalyze deamination of cytosine to uracil through nucleophilic attack of a water molecule at the C4 position. Nucleophilic water molecule is indicated in magenta.

1.1.3 Cytosine Modification Activates DNA Repair Pathways

Several BER enzymes are worthy of particular attention, with uracil DNA glycosylase (UDG) standing out for its robust ability to excise uracil from DNA [Figure 5]. Given the need to
exclude uracil, UDG conspires with deoxyuridine triphosphatase to ensure the presence of thymine over uracil in DNA (20, 21). The only naturally occurring lesion that is efficiently targeted by UDG is uracil, though unnatural lesions such as 5-fluorouracil are also processed (22). Stringent selectivity against thymine occurs by enzymatic discrimination against bulky C5 substituents, while specific hydrogen bonding to a key active site asparagine residue selects uracil over cytosine (23-25). As we will note later, in addition to its principal role in promoting DNA fidelity, UDG is exploited to generate diversity when uracil is purposefully introduced into the genome. A second key DNA repair enzyme is thymine DNA glycosylase (TDG), which targets T:G mispairs that arise from deamination of mC in CpG motifs. Spontaneous deamination of mC produces thymine, which unlike uracil is naturally occurring in DNA and therefore more challenging to recognize as a lesion (21). Furthermore, mC is an order of magnitude more prone to spontaneous deamination than cytosine (26, 27). These factors likely contribute to the increased mutation frequency at methylated CpG sequences in cancerous cells (28). A challenge lies in editing these resulting T:G mispairs: to repair this mutation without error, repair machinery much first recognize the mispair and then specifically excise thymine and not guanine. TDG and the enzyme MDB4 are both capable of this activity. Mice deficient in MBD4 do exhibit increased C to T mutations and tumorigenesis (29, 30), although the embryonic lethality of the TDG knockout, and not MBD4, suggests additional important roles for TDG (31, 32).

Figure 5: Mechanism of base excision. DNA glycosylases, including UDG and TDG, catalyze base excision of cytosine and uracil base analogs. Nucleophilic addition of water (shown in magenta) cleaves the N-glycosidic bond between base and sugar, resulting in generation of an abasic site and liberation of the free pyrimidine base.
Several features distinguish TDG from UDG. First, the enzyme actively recognizes the opposite strand G and a neighboring G, biasing activity toward T:G mismatches within CpG motifs (33). Second, the stability of the pyrimidine N-glycosidic bond, not simply the presence or absence of C5 substituents, impacts substrate preferences. In fact, TDG can cleave not only uracil-related nucleobases but also modified cytosine residues whose N-glycosidic bond is destabilized, such as 5-fluorocytosine (34). This unique mechanism of substrate recognition went largely underappreciated until the discovered that the epigenetic bases fC and caC are substrates as well. Lastly, UDG knockout mice are viable and fertile, whereas the TDG knockout mice are embryonic lethal, standing as the only known DNA glycosylase with such a phenotype (31, 32, 35).

An additional BER enzyme that may contribute to diversity is single-stranded monofunctional DNA glycosylase (SMUG). This misnomer belies the fact that SMUG preferentially acts on double-stranded DNA and that it targets several uracil-related lesions (36). A water molecule adjacent to the C5 position provides a mechanism for selectively processing uracil. Intriguingly, a C5-hydroxymethyl substituent can replace this active-site water (37), making 5-hydroxymethyluracil (hmU) a good substrate, with potential relevance to epigenetic reprogramming (38).

1.2 Mammalian Cytosine Modification Pathways

1.2.1 Cytosine Methylation Contributes to Development and Oncogenesis

Cytosine methylation is known to modulate gene expression and cellular identity. Its significance is attested by the embryonic lethal phenotype of the genetic depletion of DNA methyltransferases. Although this modification has been well studied, in the context of considering the role of cytosine in modulating genomic potential, certain aspects of this topic are worthy of reconsideration.

Cytosine methylation upstream of transcriptional start sites is a stable chemical modification associated with transcriptional repression in eukaryotic organisms (39). Cytosine methylation occurs predominantly in the context of CpG motifs. CpG motifs are
disproportionately underrepresented in the human genome, occurring four times less frequently than would be predicted by a random distribution. Further, these motifs are highly enriched in specific regions designated as CpG islands (40). The non-random distribution of potential CpG methylation sites bolsters the notion that cytosine serves an important diversity-generating function. CpG methylation alters transcriptional repression through multiple pathways, rooted in biophysical and biochemical changes that take place in the overall DNA structure (41). DNA methylation increases the melting temperature of duplex DNA, potentially decreasing promoter accessibility to RNA polymerase (42). Further, the C5 methyl group projects into the major groove of duplex DNA, providing a biochemical handle that can be interrogated by DNA binding proteins. The impact of methylation can be direct, abrogating binding of numerous transcription factors as one means to decrease gene expression (39). Alternatively, transcriptional repression can be indirectly affected, via methyl-DNA binding proteins that subsequently recruit histone modifying enzymes (43).

Functionally, cytosine methylation can restrain the inappropriate expression of genes; thus the identity and location of the modified cytosine shapes cellular function. During embryogenesis, methylation silences the transcription of lineage-specific genes to prevent aberrant protein expression that may interfere with differentiation and proper development (3). Upon differentiation, pluripotency genes are similarly methylated to ensure the adoption of a lineage-specific cell fate (4). Methylation also regulates imprinting, the parental-specific regulation of gene expression of endogenous genes as well as autosomal transgenes (44). This process ensures monoallelic gene expression, which maintains appropriate gene dosing during development. Defective imprinting results in a panoply of genetic diseases that feature profound developmental deficits.

In addition to its importance in embryogenesis, dysregulation of DNA methylation is associated with oncogenesis. Aberrant methylation may result in inappropriate silencing of tumor suppressor genes (45, 46). Globally, altered methylation patterns are seen in tumors of several organ systems, including colon, lung, breast, thyroid and kidney (47). These observations have
led to the proposal that DNA methylation maintains epigenetic stability that prevents variability in gene expression resulting in pre-cancerous and cancerous states. In tumors where the burden of mutagenesis is low, such as AML (48), the loss of epigenetic stability may be a driver of oncogenesis. As a whole, the chemical modification of cytosine, as governed by DNMTs, plays an essential role in dictating the phenotypic outcome of the genome in a given cell.

1.2.2 DNA demethylation and regeneration of unmodified cytosine

Just as cytosine methylation is critical for repression of gene expression, the reverse of this process, the removal of the methyl group, allows cells to newly express previously repressed genes or to recover their totipotent potential. Until recently, this process of cytosine demethylation was thought to be a passive process in which replication without the action of maintenance DNMTs dilutes mC from DNA. However, mounting evidence suggests that replication-independent, “active” (enzymatic) demethylation also occurs, globally in totipotent cells (49, 50) and also in a locus specific fashion within somatic cells (51-55). Active cytosine demethylation, therefore, has now been recognized as a crucial molecular process and is yet another example of the role of cytosine in modulating genomic potential. Cytosine demethylation is relevant even at the earliest stages of mammalian development. Upon penetrating the zona pelucida, the paternal pronucleus is rapidly demethylated (49). Remarkably, the maternal pronucleus sits in the same cytoplasm and is exclusively demethylated via passive demethylation; the mechanism for such asymmetric demethylation remains unclear, but it thought to be mediated by association with the factor Stella (56). Beyond the zygote and blastula stages, a subset of cells is induced to travel to the gonadal ridge and become primordial germ cells (PGCs). Although PGC genomes are widely methylated at the time they are designated, they are globally demethylated by the time they arrive at the gonadal ridge several days later (57). Given that maintenance DNMTs are expressed in PGCs, such global demethylation is assumed to require active demethylation.

Several examples of locus-specific active demethylation suggest that this process is likewise important in the normal functioning of somatic cells. Fast methylation and demethylation cycling at the estrogen receptor promoter provide a notable example of locus-specific active
demethylation (52, 53). Other studies in CD8+ T-cells illustrated that expression of IL-2 can be induced via replication-independent demethylation, suggesting a role for active demethylation in sustained immune responses (54). Finally, even neural plasticity is impacted by active demethylation as evidenced by changes at the promoter for brain-derived neurotrophic factor (38).

1.2.3 Oxidation of Methylcytosine Also Contributes to Development and Oncogenesis

An additional layer of complexity was revealed by the discovery that mC may be oxidized to hmC. This modification was first identified in bacteriophage genomes as a strategy to evade bacterial restriction endonucleases (58). The epigenetic landscape changed significantly when Rao and colleagues discovered the TET family of mC oxidase enzymes in mammals and the Heintz group concurrently reported the presence of hmC in the mouse brain (12, 59). Further studies have demonstrated that hmC is found throughout the body, albeit at a low frequency. In tissues where hmC is most enriched, the base comprises no more than 1% of all cytosines (60, 61). Much of the focus on hmC has surrounded its presence in embryonic tissues and stem cells. Indeed, several groups have described the presence of hmC in the paternal pronucleus of the fertilized egg (62, 63), and chromatin immunoprecipitation studies have shown an association between hmC and bivalent H3K4-H3K27 histone trimethylation, an epigenetic hallmark of key embryonic genes (64, 65). Though it is known that hmC levels in ES cells decrease during differentiation (59, 66-68), the modulation of hmC in adult tissues remains poorly understood. Within the genome, much like mC, hmC localizes upstream of transcription start sites, but it may also be found in intragenic bodies (65, 66).
Figure 6: Cytosine methylation and hydroxymethylation regulate transcription. Whereas methylation typically represses gene expression, the epigenetic role of hydroxymethylation is still being explored. Hydroxymethylation is thought to play both activating and inactivating roles in gene expression.

Given that the discovery of eukaryotic hmC was so recent, work is ongoing to describe its functional significance. Initial reports implicated hmC as a “poised” intermediate on the path to cytosine demethylation, a topic to be tackled in Chapter 4 (38, 69, 70). However, the current data also strongly suggest that hmC, as a stable modification of cytosine, has its own epigenetic regulatory role with respect to modulating the genome [Figure 6]. From a biophysical perspective, hmC has been shown to partially alleviate the energetic barrier for melting mC-containing duplex DNA; Tm values are similar to those of free cytosine (42, 71). However, hmC appears enriched in the promoter region of a gene, a pattern that often correlates with transcriptional repression (65). Some DNA binding proteins like MeCP2 distinguish between mC and hmC, whereas others, such as the maintenance methyltransferase factor Uhrf1, will bind both hmC and mC (72). This implies that the information encoded by hmC may dictate chromatin structure via mechanisms distinct from mC. This notion is strengthened by the observations that TET1 associates with Sin3A repressor complexes and histone deacetylases (73), and that all three TETs associate with O-linked β-N-acetylglucosamine transferase to mediate serine/threonine glycosylation of as-yet poorly described targets (74, 75). At this time, early reports indicate that
hmC may be a stable DNA modification that, like its precursor mC, causes transcriptional repression. Currently, it is unclear what impact intragenic hmC exerts; the base may disrupt methyl-binding domain interactions that remodel euchromatin to heterochromatin (76) or may activate transcription at alternative promoters (3). Clarifying these proposed epigenetic roles of hmC, in addition to its putative role in demethylation, is an important challenge ahead.

Characterization of fC and caC within the genome remain an area of emerging study. The limited abundance of these bases—approximately 10% of hmC levels, and 0.1% of overall cytosine levels—raises questions about their physiologic relevance and poses a challenge towards their accurate detection (15, 77, 78). Currently, the strongest evidence in favor of their physiologic significance is that TDG is capable of fC and caC excision (15, 33), linking two proposed players in DNA demethylation in a novel and plausible manner. This will be discussed further in chapter 4. Moreover, techniques have recently been developed that permit the sequencing of these higher oxidation products with single base resolution, leading to the profiling of fC and enhancer elements and other distal regulatory regions (79, 80).

A gathering body of evidence supports important roles for the various TET isoforms in physiological niches where DNA demethylation is thought to be relevant. Though much remains to be resolved, genetic deletion leads to perturbed demethylation of paternal pronuclei and embryonic demise in the case of TET3 (81), dysregulation of hematopoiesis in the case of TET2 (82, 83), and diminished embryonic growth and cognitive impairment of viable offspring in the case TET1 (84-86). These genetic findings couple with biochemical studies to make a case for the TET enzymes as major regulators of DNA demethylation.

1.2.4 Cytosine Deamination Contributes to Adaptive and Innate Immunity

The numerous DNA cytosine-modifying enzymes each play important physiologic roles in generating genomic variety. On its face, cytosine deamination is antagonistic to the primary function of DNA as a stable reservoir of information. However, when the process is highly targeted and controlled, purposeful deamination is used to yield beneficial mutations.

The foremost example of deamination as a means to diversity is demonstrated by the
adaptive immune system (5, 17). The mature antibody pool is a collection of heterogeneous antigen-binding molecules produced through multiple diversity-generating mechanisms. Programmed recombination of gene segments (VDJ recombination) provides the initial repertoire of B-cells, each encoding a different surface-bound IgM molecule. However, this diversity is insufficient to yield the high-affinity interactions needed for robust immune responses. In a key transformation that occurs after exposure to antigen, B cells in the germinal center are matured by two genome-altering processes: somatic hypermutation (SHM) and class switch recombination (CSR) [Figure 7]. In SHM, antibodies evolve from low-affinity to high-affinity by the introduction of mutations into their antigen-recognition loops at a rate $10^6$ times that of spontaneous mutation. In CSR, the effector domain of the heavy chain is switched from IgM to yield the alternate isotypes IgA, IgE, or IgG.

**Figure 7: Cytosine deamination in antibody maturation.** Cytosine deamination in the immunoglobulin locus generates uracil. Error-prone repair of uracil in VDJ regions results in localized mutations that increase antibody affinity in somatic hypermutation. Clustering of uracil bases in switch regions leads to double-stranded DNA breaks that are recombined, ultimately altering the antibody isotype.

The DNA-modifying enzyme activation-induced deaminase (AID) mutates key cytosines in the Ig locus to initiate the molecular events that lead to SHM or CSR (5, 17). AID expression is largely B-cell specific and restricted to germinal centers, the site of SHM and CSR (87). In SHM, AID introduces uracil into Ig locus DNA (88). The uracil lesions are then subjected to repair pathways involving UDG, mismatch repair enzymes, and low-fidelity, rather than high-fidelity, DNA polymerases, like DNA pol η (89). The DNA “repair” pathway is therefore co-opted to promote error-prone repair, resulting in hypermutation of antibody molecules. In CSR, AID targets cytosine residues that are on opposite strands in the switch regions immediately upstream.
of the various heavy chain loci encoding IgM, IgG, IgE, or IgA (90). Clustered deamination on both DNA strands leads to double-stranded DNA breaks, which are resolved by recombination to result in isotype switching.

AID’s catalytic activity is enhanced by interactions with two protein-binding partners that mediate targeting to the actively transcribed immunoglobulin genes. To facilitate AID’s interactions with regions of single-stranded DNA, RPA binds to single-stranded DNA on the non-template strand during transcription and recruits phosphorylated AID (91, 92). These interactions with the separated strands of actively transcribed DNA are further facilitated by Spt5, which promotes AID binding to RNA polymerase II and sites of active transcription (93). Both of these factors augment AID’s deaminase activity within the cell, manifest through increased rates of somatic hypermutation and class-switch recombination.

Given the fine line between genomic malleability and instability, an important factor in deamination by AID is appropriate targeting (94, 95). Hyperactive AID is associated with common oncogenic translocations as well as leukemic progression and drug resistance in chronic myeloid leukemia (96, 97). Chromosomal translocations have been documented between IgH and every other chromosome within the cell (98, 99). AID is known to act throughout the genome but preferentially acts at the Ig locus, with a balance between deamination and repair determining function (100). Troublingly, the Myc oncogene is one of the most-frequently deaminated off-target genes. The factors that predominantly restrain AID’s activity to the immunoglobulin locus remain poorly understood, but have a critical bearing for the mutagenic potential of AID.

Though AID-catalyzed SHM and CSR are exemplars of purposeful cytosine deamination, they are not the only examples. AID is closely related to APOBEC enzymes, best known for their roles in restricting retroviruses such as HIV (101). One family member, APOBEC3G (A3G), acts as a kind of Trojan horse against HIV: it can be integrated into budding HIV virions and, upon infection of a new cell, works to damage the HIV genome. A3G deaminates the negative-stranded viral cDNA generated by reverse transcription, introducing a high frequency of uracil that impairs viral integration and disrupts essential viral proteins. As a counterattack measure,
lentiviral pathogens express Vif, a small accessory protein that targets A3G for ubiquitination and degradation (102). Intriguingly, even in the presence of Vif, A3G is occasionally packaged at low levels into HIV. This observation raises the possibility that low levels of A3G mutagenesis may in fact confer a survival advantage to HIV by yielding viral variants that can escape immune pressure or antiviral challenges (103). Indeed, sublethal mutagenesis and robust acquisition of resistance to antivirals has been demonstrated when HIV was cultured in the presence of cellular A3G (104-106). Thus, just as our immune system exploits cytosine deamination to generate variety via AID, viral pathogens, though primarily antagonized by A3G, also are able to control the deaminase to access beneficial genomic variety.

1.3 AID/APOBEC Family of Cytidine Deaminases

1.3.1 Members of the AID/APOBEC Family

The actions of AID and A3G indicate the most prominent roles for cytidine deamination, but these enzymes belong to a larger family of cytidine deaminases that are responsible for all known roles of cytosine deamination in DNA. Named for the founding member Apolipoprotein B mRNA Editing Catalytic Polypeptide 1, the entire AID/APOBEC family consists of four members: AID and APOBEC 1, 2, and 3 [Figure 8]. There are preliminary reports of additional APOBEC family members—APOBEC 4 and 5—but these remain largely putative as they rely heavily on computational homology searches within the genome that have not been properly cloned, expressed, and characterized (107).

<table>
<thead>
<tr>
<th>Cytidine deaminase domain</th>
<th>Canonical Function</th>
<th>Physiologic Substrate</th>
<th>Sequence Target (Preferred)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AID</td>
<td>Antibody Maturation</td>
<td>DNA</td>
<td>WRC (AGC)</td>
</tr>
<tr>
<td>APOBEC1</td>
<td>mRNA Editing</td>
<td>RNA</td>
<td>TW₈ (UAC)</td>
</tr>
<tr>
<td>APOBEC2</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>APOBEC3</td>
<td>Retroviral Restriction</td>
<td>DNA</td>
<td>TY₈ (TCC)</td>
</tr>
</tbody>
</table>

**Figure 8: AID/APOBEC family members.** Listed are a schematic detailing number of deaminase domains, canonical function, physiologic substrates, and preferred sequence targets.
The AID/APOBEC family originated with the discovery of APOBEC1 for its role in editing the mRNA transcripts of apolipoprotein B (108, 109). Specifically, APOBEC1 deaminates cytosine to uracil at position 6666, introducing a missense mutation ultimately resulting in a truncated form of apolipoprotein B and regulating lipid metabolism. Other mRNA targets have been identified as well, including the targets in the brain and small intestine (110). APOBEC1’s RNA deaminase activity requires the assistance of a co-factor, APOBEC1 Complementarity Factor (ACF) (111, 112). ACF seems to play a critical role in anchoring APOBEC1 to AU-rich target sequences in mRNA transcripts, but precise interactions between the two remain poorly understood. In addition to APOBEC1’s role in RNA editing, studies have proposed novel roles in retroviral restriction, memory formation, and susceptibility to testicular germ cell tumors (38, 113). However, these novel functions remain preliminary and require additional experimental validation.

Within the AID/APOBEC family, APOBEC1 is unique in that it targets RNA instead of DNA (114). Given that APOBEC1 was the first discovered member of the AID/APOBEC family, it was naturally assumed that subsequent family members might be RNA deaminases as well (87). Further biochemical characterization of APOBEC1 has demonstrated that the enzyme possesses DNA deaminase in vitro, providing an indication that DNA deamination may be a hallmark of the AID/APOBEC family (115-117). Notably, a recent report utilizing comparative genomic sequence analysis has suggested that DNA deamination is APOBEC1’s ancestral function, with RNA deaminase activity emerging only recently (118).

Much less is known about the homolog APOBEC2. Although the crystal structure of APOBEC2 has been described, no known catalytic activity has been attributed to the enzyme (119). APOBEC2 expression is restricted to cardiac and skeletal muscle lineages, and its deficiency in mice appears to confer a slight deficit in muscle regeneration (120). Curiously, while there appears to be little role for APOBEC2 in mammals, a different story emerges in zebrafish (121). In this model organism, APOBEC2 not only retains catalytic DNA deaminase activity, but has also been linked to epigenetic regulation and retroviral restriction. However, given that zebrafish are a distant ancestor from mammals, they have not evolved the full family of
APOBECs (122), and it is therefore possible that zebrafish APOBEC2 has retained certain functions that have been lost in the mammalian lineage through functional redundancy with newer homologs.

APOBEC3 represents the largest subfamily amongst the AID/APOBECs. While many species only have a single APOBEC3 representative (such as mouse and rat), a series of chromosomal duplications and expansions have left certain vertebrates (primates) with up to 11 APOBEC3 family members [Figure 9] (122). Seminal studies on human APOBEC3G (A3G) demonstrated an ability to restrict HIV infection (113, 123), suggesting a possible role for the APOBEC3 family as a viral restriction factor. Subsequent studies have confirmed such a role. Within the human APOBEC3 family, additional family members are capable of HIV restriction (124-126). APOBEC3 family members from other species are similarly capable of retroviral restriction. Other classes of viruses, including paroviruses, adeno-associated viruses and retrotransposons, are subject to restriction as well (127, 128). APOBEC3 family enzymes are predominantly expressed in myeloid and lymphoid lineages and readily induced by several cytokine-signaling pathways (122).

Figure 9: APOBEC3 locus expansion from the mouse genome to the human genome. Whereas mice have a single APOBEC3 family member, humans have 8 well-characterized APOBEC3 family members.

Appropriate targeting seems to modulate APOBEC3 deaminase activity. The role of Vif in regulation of A3G packaging into HIV has already been described (102). For several family members, cellular localization plays an additional role in targeting deaminase activity. For
homologs that target viral replication, expression in the cytoplasm localizes the restriction factor to the same cellular compartment as replicating virus (129). Cytosolic localization may also be important for sequestering the DNA deaminase activity away from the nucleus, where deamination could promote mutation and oncogenesis (130). Indeed, ectopic expression of A3B has been shown to contribute to mutagenesis in a subset of breast cancers (131-133), and when overexpressed, A3A’s robust deaminase activity is a potent activator of the DNA damage response and cell cycle arrest (134).

1.3.2 Evolution of the AID/APOBEC family

The AID/APOBEC family of deaminases descends from a much larger deaminase superfamily with a broad array of targets (114). In contrast to the AID/APOBECs, which deaminate cytosine in single-stranded DNA, some subclasses are able to deaminate adenosine instead of cytosine; further, some subclasses are capable of deaminating bases, nucleosides, nucleoside monophosphates, and nucleotides. With regards to the AID/APOBEC deaminases, one family is prominent within the deaminase superfamily. Of note are the tRNA adenosine deaminases, which are widely conserved across prokaryotes and eukaryotes. These enzymes catalyze deamination of adenosine to inosine at the wobble position of tRNA anticodons so that degenerate codons are tolerated during translation. Though the adenosine deaminases have a different substrate, they share important structural similarities with the AID/APOBEC family, which will be discussed later. Both enzyme families target single stranded nucleic acids, and it is likely that the AID/APOBEC family evolved from the tRNA adenosine deaminases. The AID/APOBEC family first emerged in bony fish, with AID and APOBEC2 as the founding members. Of note, this discovery indicates that AID predates the origin of the immunoglobulin loci, its physiologic substrate in vertebrates. Estimates date the emergence of the AID/APOBEC family prior to the divergence of bony fish from the tetrapod lineage, over 450 million years ago.

As the oldest members of the AID/APOBEC family, AID and APOBEC2 can be found the genomes of many vertebrates, ranging from fish to birds to mammals (122, 135). The more recent emergence of APOBEC1 and APOBEC3 restricts these two family members to mammals.
Based on phylogenetic analysis of its synteny with AID in both mice and humans, APOBEC1 is thought to have arisen as a gene duplication of AID (135). Similar intron-exon junctions between AID and APOBEC1 further support this notion. APOBEC3 is also thought to have evolved from AID, though its evolutionary history appears more complex (135). Unlike APOBEC1, APOBEC3 has no syntenic relationship with AID. Analysis of the mouse and rat genomes indicates a single APOBEC3 gene encoding an enzyme with two deaminase domains. This contrasts with all other members of the AID/APOBEC family—AID, APOBEC1, APOBEC2—which contain only a single deaminase domain. Thus, it is thought that the origins of APOBEC3 likely involve a duplication of ancestral AID. The APOBEC3 expansion that occurred in primates likely resulted from a similar series of chromosomal expansion events, yielding an array of single-domain and double-domain deaminases.

1.3.3 Catalytic Determinants of the AID/APOBEC Family

The core catalytic determinants of the AID/APOBEC family remain conserved with the larger deaminase superfamily. AID/APOBEC deaminase domains are characterized by the deaminase consensus amino acid sequence contained within a single exon: HXE\(_{23-28}\)XPCXXC [Figure 10] (122). The histidine and cysteines coordinate the zinc atom within the active site of the deaminase domain that activates water for hydrolytic deamination of cytosine. The active site glutamate plays a critical role in deamination by serving as a proton shuttle. The glutamate first protonates the N3 position of cytosine, activating the C4 position for nucleophilic attack by the incoming water molecule. Next, the glutamate assists the zinc atom in coordination of the incoming water molecule in the enzyme active site that attacks the C4 position. The result of the nucleophilic addition of water at the C4 position is a tetrahedral intermediate that is resolved by the departure of ammonia as a leaving group, ultimately generating the uracil deamination product.
Figure 10: Catalytic determinants of AID/APOBEC deamination. Cysteine and histidine residues coordinate zinc in enzyme active site. Zinc and glutamate activate nucleophilic addition of water to C4 position, generating a tetrahedral intermediate that collapses to yield uracil. Schematic is taken from Harris et. al (122).

Considerably less is known about how cytosine is recognized for deamination. At the level of the base, it is unknown how cytosine fits into the enzyme active site and is stabilized for deamination (136). Of particular importance is the tolerance of the active site to modifications at the 5-position of the base. Methylation and hydroxymethylation of cytosine have important roles in regulating the transcriptional activation of the cell. AID and other APOBECs have been hypothesized to deaminate these modified cytosines, though their reactivity has not been properly evaluated. This will be discussed further in Chapter 4.

Beyond the level of the base, it remains unknown how the entire nucleotide must position itself for deamination. It has been hypothesized that AID may use a mechanism similar to DNA glycosylases, in which the enzyme scans along the DNA and flips target nucleotides out of stacking position in register with neighboring bases and into the enzyme active site (137). A fundamental distinction is that these models typically describe the action of enzymes on double-stranded DNA, unique from the single-stranded substrates of the AID/APOBEC family. Nevertheless, a model of DNA binding fit to the A3A structure supports such a hypothesis (138). Specifically, in this model the bases at the -2, -1 and +1 positions are stabilized by interactions with the exterior of the enzyme, while the target cytosine nucleotide is rotated ~180 degrees away from the neighboring bases and buried within the enzyme active site. This model remains speculative, as there is currently no experimental evidence to support such a mode of recognition.
Beyond the level of the target nucleotide, it is known that AID and the other APOBECs target cytosine for deamination with a sequence preference for the neighboring nucleotides. For AID, deamination occurs preferentially at WRCY hotspot motifs (W = A or T, R = A or G, Y = C or T) (139, 140). This consensus sequence is repeated throughout the switch regions in the immunoglobulin loci, and sequence specificity is thought to be an important component in properly targeting AID’s deaminase activity (90). The other members of the APOBEC family similarly demonstrate sequence specificity (APOBEC1 = AUCC; A3G = CCC; A3A/B/F = YCA) (110, 139). The sequence preferences within this family are not absolute, as deamination occurs at disfavored coldspot motifs as well, albeit at rates that are diminished by ~10-fold (140). These relaxed constraints on sequence specificity contrast with other DNA-modifying enzymes, which are significantly more selective. For example, TDG excises thymine from T:G mismatches, but only if a G is present at the neighboring position (141). If any other base is present at the +1 position, TDG’s rate of excision drops by several orders of magnitude.

The basis for hotspot sequence targeting has been partially elucidated studying AID. Structural characterization of the APOBEC homolog A3G and the more distant adenosine deaminase relative TadA have revealed the presence of a loop at the exterior face of the enzyme, in close enough proximity to the active site that it may interact with the nucleotides neighboring the target cytosine (142-144). Though the sequence of this loop is not conserved across the AID/APOBEC family, the length of the loop largely is, indicating the possibility of a conserved function (140). Multiple groups have found that engraftment of donor loops from A3G into AID (AID-3GL) generates a chimeric deaminase that not only retains catalytic activity, but additionally possesses a sequence-specific deaminase activity skewed away from AID’s spectrum and towards that of A3G (140, 145, 146). These results were consistent across biochemical studies with purified enzyme, as well as cellular studies in E. coli and DT40 B-cells. The consistency of this finding has led to the reclassification of this loop as a ‘hotspot targeting loop’. Curiously, while engraftment of donor loops from APOBEC3 family members into AID results in successful reprogramming of sequence-specific deaminase targeting, the inverse has not born true.
engraftment of the AID donor loop into several APOBEC3 family members has largely failed to reprogram sequence targeting.

Curiously, studies of the hotspot targeting loop do not account for interactions with the phosphodiester backbone, and it remains poorly understood how AID and other APOBECs interact with their single-stranded DNA substrates. It is well documented that AID has a high binding-affinity for single-stranded DNA, with dissociation constants reported as low as 1 nM (16, 147). Binding appears to occur independent of sequence context, as determined by ChIP-Seq as well as synthetic DNA substrates (148). In addition to binding single-stranded DNA, AID also binds to artificial substrates with partial single-stranded character. Specifically, binding interactions have been reported between AID and a ‘bubble’ substrate that consists of a largely double-stranded DNA substrate with a short patch of 5-9 nucleotides that are mismatched, so as to create local single-stranded character (137, 149). It is believed that this artificial substrate better mimics the transcription bubbles that AID targets in the B-cell genome as compared to single-stranded DNA. Demonstration of AID’s tighter binding to these bubble substrates leaves open the possibility that substrate binding may play an important role in AID/APOBEC deamination.

AID’s binding characteristics largely describe the behavior of the rest of the APOBEC family. Binding interactions have been well described for APOBEC1 and A3G, which also confirm low-nanomolar dissociation constants, indicative of high-affinity binding (115, 150). A3A is the only exception to this observed pattern of high binding affinity (151, 152). Whereas all other AID/APOBECs bind nucleic acids with dissociation constants in the low-nanomolar range, A3A appears to bind in the mid-micromolar range, indicating approximately 1000-fold impaired binding as compared to its APOBEC homologs. The basis for this observation is poorly understood, particularly with regards to the kinetics of deamination. Indeed, the many remaining questions regarding the catalytic mechanisms of AID and other APOBECs reflect limited insights available from structural studies on the family.

1.3.4 Structural Characteristics of the AID/APOBEC Family
The structural characteristics of the deaminase superfamily are well established, as the structures of several members have been solved by X-ray crystallography. Initial results published on the deaminase superfamily homolog Cytidine Deaminase (CDA) demonstrate a beta-sheet consisting of five strands with mixed parallel-anti parallel character (153). Both sides of the beta-sheet are flanked by alpha helices that link the beta-sheets in the primary structure of the enzyme. The zinc-coordinating active site is contained on one face of the beta sheet by an alpha-beta-alpha motif. This motif encompasses the third beta strand, embedded deeply within the core of the beta sheet. These two structural characteristics are a hallmark of the family: 5 stranded beta sheet that serves as the enzyme core and the alpha-beta-alpha motif that contains the zinc-coordinating active site.

Structural studies of APOBEC2 and APOBEC3 family members demonstrate the conserved structural characteristics of the larger deaminase superfamily (154). APOBEC2 was the first member of the subfamily to be described structurally. X-Ray crystallography demonstrated a tetramer, with one interface at the edge of the core beta pleated sheet, the other mediated by the loops that bridge the beta-pleated sheets and alpha helices at the exterior of the enzyme (119). However, the relevance of this tetrameric oligomerization state has been questioned with regards to deamination, given that APOBEC2 lacks any catalytic activity. Shortly after the publication of the APOBEC2 structure came the structural description of an APOBEC3 family member, A3G. A3G normally exists as a double-domained deaminase, with the C-terminus being catalytically active and the N-terminus catalytically inactive. Two independent reports—one using NMR, the other using X-Ray crystallography—published the structure of the C-terminus alone (142, 144). These results verified the conserved structural determinants of the larger deaminase family: the core beta-pleated sheet and localization of the active site to the alpha-beta-alpha motif at the third beta strand [Figure 11]. However, these studies failed to report definitive mechanisms of A3G interaction with its DNA substrate, leaving open the pertinent question of how cytosine is specifically recognized by the enzyme for deamination.
Figure 11: Structural characteristics of AID/APOBEC deaminases. Catalytic residues are retained within an alpha-beta-alpha repeat, shown in magenta. Zinc-coordinating residues are shown in yellow, with zinc atom shown in white. Catalytic glutamate, in blue, is partially obscured by the zinc atom. The rest of the core beta sheet is indicated in orange.

The crystal structure of the distantly related adenosine deaminase TadA helped to elucidate the important structural characteristics of A3G, with regards to hotspot recognition. The tRNA adenosine deaminases and AID/APOBEC deaminases are unique within their larger superfamily for their ability to bind single stranded nucleic acids. The structure of RNA-bound TadA demonstrates interactions between the loop that bridges the fourth beta strand and fourth alpha helix (143). Alignment with the structure of unbound A3G demonstrates the conserved structural determinants that comprise the hotspot targeting loop [Figure 12]. High resolution modeling based on the recently-solved structure of A3A further confirms the interaction between the hotspot targeting loop and DNA substrate, indicating the importance of this structure in the AID/APOBEC deamination reaction (138).
Since the publication of the A3G structure, several additional APOBEC3 structures have been described, all unbound to nucleic acids (155, 156). The absence of a bound nucleic acid target diminishes the impact of these studies, as they fail to reveal any new determinants of deamination. Particularly in the absence of novel structural discoveries within the AID/APOBEC family, it is imperative to employ alternative methods to elucidate the characteristics of deamination.

1.4 Thesis objectives

Continued research on AID has revealed a double-edged sword with regards to genomic instability. On one hand, AID’s ability to catalyze cytosine deamination fosters immunologic diversity through somatic hypermutation and class-switch recombination, and AID’s deficiency results in immunodeficiency. On the other hand, when AID’s catalytic activity is not properly restrained, aberrant cytosine deamination results in profound mutagenesis that can drive cancer development and progression. While this delicate balance has been thoroughly demonstrated for AID, a similar picture is coming into view for other members of the APOBEC family, with
physiologic roles in innate immunity counterbalanced by off-target deamination contributing to oncogenesis. With better characterization of the interface between enzyme and substrate will come further insights into the role of cytosine deamination in health and disease.

Further biochemical characterization of the AID/APOBEC deamination reaction can also provide insights into the novel, proposed roles for deaminases in DNA demethylation. It has been proposed that AID and other APOBECs may deaminate mC or hmC. These novel physiologic roles for deaminases have been implicated in embryogenesis, memory formation, and epigenetic stability. However, the reactivity of AID/APOBECs against these modified cytosines remains a poorly profiled point of controversy.

Given the importance of elucidating these many questions, the work described within this thesis aims to clarify AID/APOBEC biology using biochemical approaches to characterize the nucleic acid determinants of deamination.
CHAPTER 2: MATERIALS AND METHODS

2.1 Protein Expression and Purification

2.1.1 AID/APOBEC Expression and Purification

Human AID (amino acids 1-181 or full-length) was cloned downstream of an N-terminal maltose binding protein in a pET41 expression plasmid. Expression vectors containing the mouse APOBEC1, mouse APOBEC2, and mouse APOBEC3 genes downstream of an N-terminal maltose-binding protein (MBP) in pET41 (Novagen) were generously provided by Junjie Guo and Hongjun Song (Johns Hopkins University). BL21-DE3 E. coli (Novagen) were transformed with expression constructs, grown to OD_{600} 0.6, and protein expression was induced by addition of 1 mM IPTG (Sigma). Following induction, cells were transferred to 16 °C and incubated for 18 hours before cultures were pelleted.

Proteins were purified essentially as described previously (140). Bacterial pellets were resuspended in 150 mM NaCl, 50 mM Tris-HCl pH 7.5, 10% glycerol with EDTA-free protease inhibitors (Roche) and cells were lysed in a microfluidizer processor. Following removal of the insoluble fraction of the cellular lysates by centrifugation, the soluble fraction was added to amylose resin (New England Biolabs) and incubated at 4 °C for one hour. Resin was washed in a high salt buffer (750 mM NaCl, 50 mM Tris-HCl pH 7.5, 10% glycerol), then low salt buffer (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 10% glycerol), and proteins were eluted with maltose-containing elution buffer (10 mM maltose, 150 mM NaCl, 50 mM Tris-HCl pH 7.5, 10% glycerol). Following elution, enzyme was dialyzed overnight at 4 °C in a storage buffer containing 75 mM NaCl, 50 mM Tris-HCl pH 7.5, 10% glycerol.

For each sample, protein concentration was initially determined by collection of UV absorbance spectrum on a UV spectrophotometer (UV-1800, Shimadzu). Absorbance at 280 nm was used to determine total protein concentration of bulk sample. For each AID/APOBEC family member, a significant peak was also observed at 260 nm, indicating the presence of co-purifying nucleic acids. To assess the purity of each protein preparation, SDS-PAGE was performed [Figure 13]. The presence of multiple co-purifying bands at unexpected molecular weights
indicates partial purity of each purified enzyme sample. Co-purifying contaminants likely represent truncation products of deaminase translation, as well as endogenous, *E. coli* proteins that bind nucleic acids.

**Figure 13: SDS-PAGE of partially purified AID/APOBEC family members.** The mouse genes encoding mAPOBEC1 (mA1), mAPOBEC2 (mA2), and mAPOBEC3 (mA3) and the human genes encoding full-length human AID (hAID) and its hyperactive C-terminally truncated variant (hAID-∆C) were expressed as N-terminal fusions to maltose binding protein (MBP). Binding to amylose resin yielded partially purified enzymes. Shown is a Coomassie-stained gel of the proteins, with the major contaminants representing MBP and truncation products.

The presence of a high 260 nm absorbance peak bears particular relevance for biochemical studies on purified enzyme. Given the high binding affinity for nucleic acids, purified AID (and other APOBECs) are predominantly bound to nucleic acids—both RNA and DNA—that originate from the cellular source of protein expression. For *in vitro* reactions, these bound nucleic acids act as a competitive inhibitor of deamination, and their removal increases enzymatic deaminase activity. RNA predominates as the most abundantly bound nucleic acid species, and pre-treatment with RNase greatly improves the enzymatic activity of purified AID.

### 2.1.2 MBP Expression and Purification

MBP control was generated via cassette mutagenesis by introducing a stop codon at the first codon of AID in the MBP-AID construct [Table 1]. The construct was transformed into BL21-DE3 *E. coli* as described for AID/APOBEC purification; however, once protein expression was
induced, cultures were maintained for only two hours at 37 °C before pelleting. Protein purification proceeded as described for AID/APOBECs, with the following exceptions: cells were lysed by sonication (Sonicator 3000, Misonix), and the wash with the high-salt buffer was omitted.

Protein concentration was assessed by UV absorbance at 280 nm. As a contrast to AID/APOBECs, no absorbance at 260 nm was detected, indicating the absence of co-purifying nucleic acids. SDS-PAGE was used to determine protein purity [Figure 14]. The absence of co-purifying bands indicates the high degree of purity of the MBP control protein. These impurities contrast with the MBP-tagged deaminase enzyme preparations, indicating that both nucleic acid binding and co-purifying contaminants are likely attributed to the presence of the deaminases.

Figure 14: SDS-PAGE of MBP control protein. Shown is a Coomassie-stained gel of all fractions from purification of MBP control protein. Purified protein is shown in the two lanes furthest to the right, as both diluted and concentrated samples. The absence of co-purifying bands demonstrates purity of the protein preparation.

2.2 Synthesis of chimeric oligonucleotide substrates

All oligonucleotides were synthesized using standard phosphoramidite chemistry. The majority of substrates were synthesized using ABI 394 Synthesizer (Applied Biosystems), and the remainder were synthesized either by Integrated DNA Technologies or the University of Calgary DNA Synthesis Core Facility [Table 1]. Phosphoramidite building blocks and reagents were obtained from Glen Research or Metkinen Chemistry and used according to the manufacturer’s instructions.
recommendations. Following synthesis, oligonucleotides were deprotected and purified using Glen-Pak DMT-ON columns, according to the manufacturer’s recommendations.

To confirm complete deprotection, purified oligonucleotides were analyzed by MALDI-TOF mass spectrometry. In brief, oligonucleotides were concentrated using ZipTips containing C18 resin (EMD Millipore) and resuspended in a matrix containing 2-picolinic acid and ammonium citrate. Samples were analyzed on a Microflex mass spectrometer (Bruker), using a negative voltage polarity for detection. All measured masses were within 5 mass units of the predicted molecular weight [Table 1]. Oligonucleotides synthesized by Integrated DNA Technologies were analyzed by MALDI TOF or ESI as part of the manufacturer’s internal quality control. All substrates were further purified PAGE purified to remove residual truncation products.

The majority of oligonucleotides substrates was synthesized on 6-Fluorescein (FAM) CPG columns and therefore contain a 3'-FAM label. For the minority of substrates that were synthesized without 3'-FAM, fluorescent labeling was necessary to permit visualization. Following synthesis, these oligonucleotides were enzymatically 3'-end labeled with ddUTP-12-FAM (Enzo Life Sciences) by incubation with Terminal Transferase (New England Biolabs) and purified using QIAquick Nucleotide Removal Kit (Qiagen).
Table 1: Oligonucleotides

<table>
<thead>
<tr>
<th>Substrate Name</th>
<th>Sequence</th>
<th>Predicted</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA Chimeric Substrates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-dC</td>
<td>5'- AGA ATT TTA AGdC TAG TTA AGT TAT -3'</td>
<td>8360.5</td>
<td>8359.1</td>
</tr>
<tr>
<td>D-dU</td>
<td>5'- AGA ATT TTA AGdU TAG TTA AGT TAT -3'</td>
<td>8361.5</td>
<td>8360.5</td>
</tr>
<tr>
<td>D-rC</td>
<td>5'- AGA ATT TTA AGrC TAG TTA AGT TAT -3'</td>
<td>8376.5</td>
<td>8376.6</td>
</tr>
<tr>
<td>D-rU</td>
<td>5'- AGA ATT TTA AGrU TAG TTA AGT TAT -3'</td>
<td>8377.5</td>
<td>8377.3</td>
</tr>
<tr>
<td>D-frC</td>
<td>5'- AGA ATT TTA AGfrC TAG TTA AGT TAT -3'</td>
<td>8378.5</td>
<td>8378.9</td>
</tr>
<tr>
<td>D-frU</td>
<td>5'- AGA ATT TTA AGfrU TAG TTA AGT TAT -3'</td>
<td>8379.5</td>
<td>8378.6</td>
</tr>
<tr>
<td>D-aC</td>
<td>5'- AGA ATT TTA AGaC TAG TTA AGT TAT -3'</td>
<td>8376.5</td>
<td>8374.9</td>
</tr>
<tr>
<td>D-aU</td>
<td>5'- AGA ATT TTA AGaU TAG TTA AGT TAT -3'</td>
<td>8377.5</td>
<td>8376.3</td>
</tr>
<tr>
<td>D-faC</td>
<td>5'- AGA ATT TTA AGfaC TAG TTA AGT TAT -3'</td>
<td>8378.5</td>
<td>8377.6</td>
</tr>
<tr>
<td>D-faU</td>
<td>5'- AGA ATT TTA AGfaU TAG TTA AGT TAT -3'</td>
<td>8379.5</td>
<td>8378.0</td>
</tr>
<tr>
<td>A1-D-dC</td>
<td>5'- AGA ATT TTA TTdC TAG TTA AGT TAT -3'</td>
<td>8326.5</td>
<td>8326.0</td>
</tr>
<tr>
<td>A1-D-dU</td>
<td>5'- AGA ATT TTA TTdU TAG TTA AGT TAT -3'</td>
<td>8327.5</td>
<td>8327.1</td>
</tr>
<tr>
<td>A1-D-rC</td>
<td>5'- AGA ATT TTA TTrC TAG TTA AGT TAT -3'</td>
<td>8342.5</td>
<td>8342.5</td>
</tr>
<tr>
<td>A1-D-rU</td>
<td>5'- AGA ATT TTA TTrU TAG TTA AGT TAT -3'</td>
<td>8343.5</td>
<td>8344.1</td>
</tr>
<tr>
<td>A1-D-frC</td>
<td>5'- AGA ATT TTA TTfrC TAG TTA AGT TAT -3'</td>
<td>8344.5</td>
<td>8345.9</td>
</tr>
<tr>
<td>A1-D-frU</td>
<td>5'- AGA ATT TTA TTfrU TAG TTA AGT TAT -3'</td>
<td>8345.5</td>
<td>8347.4</td>
</tr>
<tr>
<td>D-TAdC</td>
<td>5'- AGA ATT TTA TAdC TAG TTA AGT TAT -3'</td>
<td>8335.5</td>
<td>8335.2</td>
</tr>
<tr>
<td>D-TAdU</td>
<td>5'- AGA ATT TTA TAdU TAG TTA AGT TAT -3'</td>
<td>8336.5</td>
<td>8336.9</td>
</tr>
<tr>
<td>D-TArC</td>
<td>5'- AGA ATT TTA TArC TAG TTA AGT TAT -3'</td>
<td>8351.5</td>
<td>8351.3</td>
</tr>
<tr>
<td>D-GTdT</td>
<td>5'- AGA ATT TTA GTdT TAG TTA AGT TAT -3'</td>
<td>8296.5</td>
<td>8295.5</td>
</tr>
<tr>
<td>D-GTdU</td>
<td>5'- AGA ATT TTA GTdU TAG TTA AGT TAT -3'</td>
<td>8297.5</td>
<td>8297.1</td>
</tr>
<tr>
<td>D-GTrC</td>
<td>5'- AGA ATT TTA GTrC TAG TTA AGT TAT -3'</td>
<td>8312.5</td>
<td>8312.2</td>
</tr>
<tr>
<td>Switch-dC</td>
<td>5'- ATG AGC TGG GAA TGA GCT GAG dCTA GGC TGG AAT AGG CTG GGC TGG -3'</td>
<td>14127.2</td>
<td>14128.1</td>
</tr>
<tr>
<td>Switch-dU</td>
<td>5'- ATG AGC TGG GAA TGA GCT GAG dUTA GGC TGG AAT AGG CTG GGC TGG -3'</td>
<td>14128.1</td>
<td>14127.6</td>
</tr>
</tbody>
</table>

30
<table>
<thead>
<tr>
<th>Substrates</th>
<th>Sequence</th>
<th>14143.2</th>
<th>14142.1</th>
<th>IDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switch-rC</td>
<td>5'- ATG AGC TGG GAA TGA GCT GAG rCTA GGC TGG AAT AGG CTG GGC TGG TGG -3'</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2'-F-RNA Chimeric Substrates</th>
<th>Sequence</th>
<th>10013.5</th>
<th>10014.0</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-control C</td>
<td>5'- GGA AGG AAG UUG UGC GAG UGG GGU GGU AGG (6-FAM) -3'</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-control U</td>
<td>5'- GGA AGG AAG UUG UGU GAG UGG GGU GGU AGG (6-FAM) -3'</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S30-ATX Series</th>
<th>Sequence</th>
<th>10081.6</th>
<th>10081.6</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>S30-ATC</td>
<td>5'- GGA AGG AAG TTG ATC GAG TGG GGT GGT AGG (6-FAM) -3'</td>
<td>10081.6</td>
<td>10081.6</td>
<td>Author</td>
</tr>
<tr>
<td>S30-ATU</td>
<td>5'- GGA AGG AAG TTG ATU GAG TGG GGT GGT AGG (6-FAM) -3'</td>
<td>10082.6</td>
<td>10082.5</td>
<td>Author</td>
</tr>
<tr>
<td>S30-ATmC</td>
<td>5'- GGA AGG AAG TTG ATmC GAG TGG GGT GGT AGG (6-FAM) -3'</td>
<td>10095.7</td>
<td>10095.6</td>
<td>Author</td>
</tr>
<tr>
<td>S30-ATT</td>
<td>5'- GGA AGG AAG TTG ATT GAG TGG GGT GGT AGG (6-FAM) -3'</td>
<td>10096.7</td>
<td>10096.7</td>
<td>Author</td>
</tr>
<tr>
<td>S30-AThmC</td>
<td>5'- GGA AGG AAG TTG AThmC GAG TGG GGT GGT AGG (6-FAM) -3'</td>
<td>10111.6</td>
<td>10111.5</td>
<td>Author</td>
</tr>
<tr>
<td>S30-AThmU</td>
<td>5'- GGA AGG AAG TTG AThmU GAG TGG GGT GGT AGG (6-FAM) -3'</td>
<td>10112.6</td>
<td>10112.6</td>
<td>Author</td>
</tr>
</tbody>
</table>
S30-AT(5F)C  5'- GGA AGG TTG AT(5F)C GAG TGG GGT GGT AGG (6-FAM) -3'  10099.6  10099.6  Author
S30-AT(5F)U  5'- GGA AGG TTG AT(5F)U GAG TGG GGT GGT AGG (6-FAM) -3'  10100.6  10100.4  Author
S30-AT(5OH)C  5'- GGA AGG TTG AT(5OH)C GAG TGG GGT GGT AGG (6-FAM) -3'  10097.6  10097.7  Author
S30-AT(5OH)U  5'- GGA AGG TTG AT(5OH)U GAG TGG GGT GGT AGG (6-FAM) -3'  10098.6  10098.5  Author
S30-AT(5Br)C  5'- GGA AGG TTG AT(5Br)C GAG TGG GGT GGT AGG (6-FAM) -3'  10160.5  10160.4  Author
S30-AT(5Br)U  5'- GGA AGG TTG AT(5Br)U GAG TGG GGT GGT AGG (6-FAM) -3'  10161.5  10161.3  Author

S30-TG(5F)C  5'- GGA AGG TTG TG(5F)C GAG TGG GGT GGT AGG (6-FAM) -3'  10115.6  10115.4  Author
S30-TG(5F)U  5'- GGA AGG TTG TG(5F)U GAG TGG GGT GGT AGG (6-FAM) -3'  10116.6  10116.6  Author
S30-TG(5OH)C  5'- GGA AGG TTG TG(5OH)C GAG TGG GGT GGT AGG (6-FAM) -3'  10113.6  10113.8  Author
S30-TG(5OH)U  5'- GGA AGG TTG TG(5OH)U GAG TGG GGT GGT AGG (6-FAM) -3'  10114.6  10114.5  Author
S30-TG(5Br)C  5'- GGA AGG TTG TG(5Br)C GAG TGG GGT GGT AGG (6-FAM) -3'  10176.5  10176.5  Author
S30-TG(5Br)U  5'- GGA AGG TTG TG(5Br)U GAG TGG GGT GGT AGG (6-FAM) -3'  10177.5  10177.3  Author
S30-TG(5I)C  5'- GGA AGG TTG TG(5I)C GAG TGG GGT GGT AGG (6-FAM) -3'  10223.5  10223.6  Author
S30-TG(5I)U  5'- GGA AGG TTG TG(5I)U GAG TGG GGT GGT AGG (6-FAM) -3'  10224.5  10224.5  Author
<table>
<thead>
<tr>
<th>Complement Strands</th>
<th>Primer Sequence</th>
<th>Forward / Reverse</th>
<th>IDT 1</th>
<th>IDT 2</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGC-Complement</td>
<td>5'- ATA ACT TAA CTA GCT TAA CTT AAT TCT -3'</td>
<td>8191.4</td>
<td>8196.9</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>Bubble 5'-Complement</td>
<td>5'- ATA ACT TAA CTT TGA AAA CTT AAT TCT -3'</td>
<td>8215.4</td>
<td>Not Done</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>GTC-Complement</td>
<td>5'- ATA ACT TAA CTA GAC TAA CTT AAT TCT -3'</td>
<td>8256.4</td>
<td>8261.2</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>Switch Complement</td>
<td>5'- CCA GCC CAG CCT ATT CCA GCC TAG CTC AGC TCA TTC CCA GCT CAT -3'</td>
<td>13557.8</td>
<td>13558.5</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>TAC-Complement</td>
<td>5'- ATA ACT TAA CTA GTA TAA CTT AAT TCT -3'</td>
<td>8215.4</td>
<td>8215</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>TTC-Complement</td>
<td>5'- ATA ACT TAA CTA GAA TAA CTT AAT TCT -3'</td>
<td>8224.4</td>
<td>8224.1</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>TGC-Complement</td>
<td>5'- CCT ACC ACC CCA CTC GCA CAA CTT CCT TCC -3</td>
<td>8887.8</td>
<td>8888.3</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>Cy5-Primer</td>
<td>5'- (Cy5) CAC CCC AC -3'</td>
<td>2832.2</td>
<td>2831.5</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>S30-ATC-Comp</td>
<td>CCT ACC ACC CCA CTC GAT CAA CTT CCT TCC</td>
<td>8902.8</td>
<td>Not Done</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>S30-TGC-Comp</td>
<td>CCT ACC ACC CCA CTC GCA CAA CTT CCT TCC</td>
<td>8887.8</td>
<td>Not Done</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>MBP-Control Cassette Mutagenesis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forward</td>
<td>5'- TCG ACA TGA GAT AGC CTG CTG ATG AAC CGT CGT AAA TTT CTG TAT CAG TT -3'</td>
<td>15390.0</td>
<td>15391.4</td>
<td>IDT</td>
<td></td>
</tr>
<tr>
<td>Reverse</td>
<td>5'- CTA GAA CTG ATA CAG AAA TTT ACG ACG GTT CAT CAG CAG GCT ATC TCA TG -3'</td>
<td>15377.0</td>
<td>15377.4</td>
<td>IDT</td>
<td></td>
</tr>
</tbody>
</table>

33
2.3 Biochemical Deamination Assays

2.3.1 AID/APOBEC Deaminase Incubations

Deaminase assays were performed essentially as previously described (157). Specified concentrations of oligonucleotide and deaminase enzyme were co-incubated in the presence of 1X buffer DA (20 mM Tris-HCl pH 8.0, 1 mM Dithiothreitol, 1 mM EDTA) at 30 °C for times ranging from 1 minute to 12 hours as indicated. For chimeric substrates without any RNA content, reactions were supplemented with 1 U of RNase A (NEB). For chimeric substrates with RNA content, reactions were supplemented with 5 U RNAse OUT (Invitrogen) to inhibit RNase activity. Following incubation, deaminase enzymes were inactivated by incubation at 95 °C for 20 minutes. Subsequent to incubation with deaminase enzymes, oligonucleotides were screened for deamination with one of several downstream assays.

2.3.2 DNA Glycosylase-Based Deamination Assay

The traditional biochemical assay for detection of cytosine deamination utilizes Uracil DNA Glycosylase (UDG). UDG excises any uracil bases, generating an abasic site that may be cleaved by treatment with hot alkali. Following incubation with AID, substrates were incubated with 0.3 U/µL UDG (NEB) at 37 °C for 12 hours in 1X DA buffer. UDG digestion reactions were quenched by addition of formamide and 150 mM NaOH (final concentration: 50% v/v) to promote base-mediated cleavage of the abasic sites. Samples were run on a denaturing PAGE gel and imaged on a Typhoon 9400 scanning gel reader (Amersham Biosciences). Substrate and product band intensities were quantified using QuantityOne (BioRad), and background intensities were subtracted. Total fraction of deamination was measured as the intensity of the product band divided by the sum of the intensities of both the product and substrate bands.

UDG is unreactive to many uracil analogs that contain substitutions at the 5-position of the base. Therefore, to assay deamination of modified, 5-substituted cytosines, additional DNA Glycosylase-based assays were developed with two glycosylases that are tolerant of these substitutions: TDG and SMUG. After incubation with deaminase enzymes, excess complementary oligonucleotide was added (150 nM reaction oligonucleotide to 250 nM
complement) and annealed to substrate oligonucleotide by slow cooling from 95 °C. Duplexed DNA (final concentration 100 nM) was incubated with the appropriate glycosylase. For hSMUG reactions, 140 nM hSMUG in 1X Buffer DA with 0.1 mg/mL BSA and incubated at 37 °C for 45 min. For TDG reactions, 1.6 µM TDG was used in 1X Buffer DA supplemented with 0.1 mg/mL BSA, and reactions were incubated at 16 ºC for 12 hrs. DNA glycosylase reactions were quenched in 50% formamide (v/v) and 150 mM NaOH and heated to 95 °C for 20 minutes to cleave the abasic sites generated by glycosylases. Deamination was then quantified as described above for UDG.

For all glycosylases, cleavage of respective deamination products was incomplete and, therefore, an incomplete measure of deamination. To account for incomplete excision of deaminated bases, standard curves for quantifying the fraction of deaminated product were generated using S30-ATX substrates (X = C/mC/hmC) and S30-ATY products (Y = U/T/hmU), mixed in various ratios to 100 nM total concentration and duplexed with excess complementary strand S30-ATC-Comp. Duplexes were then incubated with UDG (for C/U), SMUG (for hmC/hmU) or TDG (for mC/T). When the BER-coupled assays are performed with 100 nM duplex DNA, the glycosylases can be used to detect 0.5 nM product. Identical conditions were used to generate standard curves with the unnaturally modified S30-TGX substrates and S30-TGY products, with UDG (for (5F)C/(5F)U) and SMUG (for (5OH)C/5(OH)U, (5Br)C/(5Br)U and (5I)C/(5I)U). The calculated fraction of cleaved product was determined by the intensity of the fluorescent product over the total product plus substrate. For all deamination assays the actual fraction of deaminated product was determined with reference to the standard curve generated for the corresponding glycosylase. Additionally, these standard curves serve as a reference for the lower limits of detection for deamination of each modified cytosine species.

2.3.3 Restriction Endonuclease-Based Deamination Assay

Several of the chimeric nucleic acid substrates in this study contain cytosine analogs bearing modifications to the 2’-position of the nucleotide sugar. These modifications are not tolerated by DNA glycosylases and necessitate a novel assay to screen for deamination. We
found that restriction endonucleases were surprisingly tolerant to DNA substrates containing sugar modifications at a single nucleotide. Following incubation with deaminase enzymes, oligonucleotides were annealed to an appropriate complementary strand at a concentration of 50 nM substrate to 100 nM complement, thereby completing a palindromic duplexed substrate for restriction endonucleases. Oligonucleotides were annealed by incubation at 75 °C for 5 minutes and slow cooling to 37 °C. 20 nM of duplexed DNA was incubated with 0.2 Units/µL FspBI (Fermentas) or BfaI (NEB) restriction endonucleases at 37 °C for 3 hours. Digestion reactions were quenched and denatured by addition of formamide (final concentration: 50% v/v) and incubation at 95 °C for 20 minutes. Samples were run on a denaturing 20% acrylamide/TBE/urea polyacrylamide gel (PAGE) at 50 °C and imaged on a Typhoon 9400 scanning gel reader (Amersham Biosciences). FspBI could accurately discriminate between a C:G match or a U:G mismatch within its recognition sequence, as could the isoschizomer BfaI, suggesting that other DNA endonucleases may be able to recognize chimeric non-DNA substrates.

To determine whether FspBI was capable of completely cleaving unreacted cytosine substrates, a standard curve was generated for quantifying the fraction of deaminated product from the FspBI assay. D-rC substrates and D-rU products were mixed in various ratios and assayed with the FspBI restriction endonuclease in triplicate. Observed fraction of D-rU was graphed as a function of actual fraction of D-rU, demonstrating a strong linear correlation. The standard curve was also able to serve as an estimate of the lower limit of detection of deamination for the assay.

2.3.4 Reverse Transcriptase-based Deamination Assay

Substrates containing prolonged stretches of 2’-fluororibonucleotides were not substrates for either DNA glycosylases or restriction endonucleases, and a novel assay was required to assess their reactivity for deamination. We found that Reverse Transcriptase was able to tolerate templates containing 2’-fluororibonucleotides and chimeric templates containing both 2’-fluororibo- and 2’-deoxynucleotides. Following incubation with AID, a 5’-Cy5-labeled primer [Table 1] was annealed to oligonucleotide substrate (166 nM substrate: 83 nM primer) by
incubation at 75 °C for 5 minutes and slow cooling to 20 °C. 50 nM primer/template duplexes were extended by MuLV Reverse Transcriptase (New England Biolabs) in the presence of 1 mM ddATP and 100 µM dCTP/dGTP/dTTP at 25 °C for 4 hours in polymerase buffer (10 mM Tris HCl pH 7.9, 10 mM MgCl₂, 50 mM NaCl, 1 mM DTT). Primer extension reactions were quenched by addition of formamide (final concentration: 50% v/v) containing 1 µM TGC-complement [Table 1] to denature primer-template duplexes and prevent reannealing of the primer strand. Samples were run on a denaturing PAGE gel, imaged using the Cy5 label for detection and quantified as described above.

2.4 Protein Binding Assays

Previous experiments assessing the binding capacity of AID/APOBEC deaminases have predominantly relied on Electrophoretic Mobility Shift Assays (EMSA). This technique does not lend itself to quantitative analysis of protein binding, given the shifting equilibrium dynamics between protein and substrate that occur during electrophoresis. Therefore, we turned to Fluorescence Anisotropy. 5 nM oligonucleotide substrate was incubated with increasing concentrations of protein under the same buffer conditions used in deaminase assays at room temperature (n=3 replicates). Fluorescence polarization was measured using a Panvera Beacon 2000 Fluorescence Polarization system. Polarization values were graphed as a function of enzyme concentration in GraphPad Prism. A one-site binding, non-linear regression was fit to the data, which yielded dissociation constants (K_d) and maximum polarization values for each substrate. For each substrate, values were normalized to maximum and minimum polarization values to yield total fraction of substrate bound.

2.5 Cellular Deamination Assays

2.5.1 Transient Expression of Deaminases and other DNA-modifying Enzymes

All cell culture and transfections were performed by the laboratory of Yi Zhang. TDG, mouse APOBEC1 (mAPOBEC1) and mouse APOBEC3 (mAPOBEC2) were each cloned from pCMV-SPORT6 vectors (Open Biosystems) into a FLAG-tagged pcDNA3 vector (pcDNA3β-FLAG). Mouse APOBEC2 (mAPOBEC2) was amplified from mouse embryonic stem cell cDNA
and cloned into pcDNA3\(\beta\)-FLAG. The cloned TET2 construct, untagged human AID and the catalytic mutant E58A of hAID were previously described. A synthetic gene encoding UGI was cloned into pIRESneo3 (Clontech). 18 hrs after plating \(8 \times 10^5\) HEK293T cells, cells were transfected with pcDNA3\(\beta\)-FLAG-TET2 (2 \(\mu\)g) and pcDNA3\(\beta\)-FLAG containing the indicated enzyme (1 \(\mu\)g), using Fugene HD transfection reagents (Roche). For experiments with UGI, transfections were done with UGI-pIRESneo3 (1 \(\mu\)g) and the indicated deaminase expression vector or controls. 48 hrs after transfection, cells were harvested for analysis of protein level and genomic DNA. Equal concentrations of cells, sonicated and boiled in protein loading buffer, were assayed by western blotting using a polyclonal \(\alpha\)-AID antibody, \(\alpha\)-AID 30F12 (Cell Signaling) or monoclonal \(\alpha\)-FLAG antibody (Sigma). Loading was controlled by probing with \(\alpha\)-tubulin (Sigma).

2.5.2 Mass Spectrometric Experiments

Mass spectrometric experiments were performed by the laboratory of Yi Zhang, essentially as previously described (14). Briefly, 2.5 \(\mu\)g genomic DNA, isolated using the DNeasy Kit (Qiagen), was heat-denatured, hydrolyzed with 90 U of Nuclease S1 (Sigma) in Buffer (0.5 mM ZnSO\(_4\), 14 mM sodium acetate, pH 5.2) at 37 °C for 1 hour, followed by the addition of 5 \(\mu\)L 10X Buffer 2 (560 mM Tris-Cl, 30 mM NaCl, 10 mM MgCl\(_2\), pH 8.3), 0.5 \(\mu\)g of phosphodiesterase I (Worthington) and 2 U of Calf Intestinal Alkaline Phosphatase (New England Biolabs) for an additional 1 hour (final volume 50 \(\mu\)L). Digested DNA was then filtered with Nanosep3K (Pall Corporation) and 15 \(\mu\)L of filtered samples were subjected to LC-MS/MS analysis as described previously with an additional transition for hmU (m/z 259.0 to 125.0) and for dU (m/z 229.0 to 113.0).

2.5.3 Deaminase and Base Excision Activity of Nuclear Lysates

Nuclear lysates were prepared from HEK293T cells as previously described (22). Cells were resuspended in a buffer containing 10 mM Heps-OH [pH 7.5], 10 mM NaCl, 1.5 mM MgCl\(_2\), 0.5 mM PMSF, and 1 mM DTT and lysed by passage through a high-gauge needle. Crude nuclei were pelleted by centrifugation at 800g, and further lysed in 20 mM Heps-OH (pH 7.5), 420 mM NaCl, 25% glycerol, 1.5 mM MgCl\(_2\), 0.2 mM EDTA, 0.5 mM PMSF, and 1 mM DTT. Following
removal of insoluble debris by centrifugation, nuclear lysates were dialyzed into a buffer containing 20 mM Hepes-OH (pH 7.5), 100 mM NaCl, 2 mM EDTA, 1 mM DTT, 0.5 mM PMSF, and 5% glycerol, and stored at -80 °C.

For analysis of deaminase activity, lysates (final protein concentration 0.3 μg/μL) were incubated with single-stranded oligonucleotide substrates (2 μM) in 1X Buffer DA and 0.1 mg/μL BSA, for 30 minutes. Reactions were quenched by addition of phenol:chloroform:isoamyl alcohol (25:24:1). To the DNA isolated from the aqueous phase, formamide and NaOH were added (50% v/v, 150 mM final concentrations). Samples were heated for 20 minutes (95 °C), separated by denaturing PAGE and analyzed as described for deamination assays.

For analysis of base excision activity against uracil or hmU in nuclear lysates, lysates (final concentration 0.2 μg/μL) were pre-incubated at 37 °C for 20 minutes before addition of 1 μM duplex DNA. Reactions were quenched and analyzed as described above.
CHAPTER 3: NUCLEIC ACID DETERMINANTS FOR SELECTIVE DEAMINATION OF DNA OVER RNA BY ACTIVATION-INDUCED DEAMINASE AND APOBEC1

3.1 Introduction

In the life of the cell, enzymes must sort through a complex milieu of biomacromolecules to identify their proper substrates and perform their intended function. For enzymes that target nucleic acids, distinguishing RNA from DNA presents a notably critical challenge. Within the AID/APOBEC family, the physiologic targets have been well-identified, but the mechanisms of selectivity remain unknown. Historically, this question has been most relevant to AID. Therefore, AID’s ability to discern between the two nucleic acids serves as the focus of this chapter and as a window into the preferences of the larger AID/APOBEC family.

3.1.1 DNA Deamination Model for Role of Activation-Induced Deaminase in Antibody Maturation

When AID was initially discovered, the closest known homolog was APOBEC1, an RNA deaminase that introduces a premature stop codon in the mRNA of apolipoprotein B (87). From this observation, it was assumed that AID was also an RNA deaminase; however, several lines of evidence have subsequently challenged this initial assumption in favor of a DNA deamination model. Early studies showed that overexpressed AID is capable of mutating the E. coli genome, and mutations could be enhanced by inhibiting or eliminating uracil DNA glycosylase (UDG), the base excision repair enzyme that accounts for the majority of uracil excision within the cell (116, 158). In mice and humans, UDG and the DNA mismatch repair enzyme Msh2 are required for SHM and CSR, providing physiological evidence for the significance of AID-generated deoxyuracil in antibody maturation (159-162). The evolving model of DNA deamination has been subsequently bolstered by in vitro biochemical studies on purified AID. The enzyme was shown to carry out deamination of single-stranded DNA oligonucleotides, without observable activity on single-stranded RNA, double-stranded DNA, or RNA-DNA hybrids (16, 115, 163, 164). Most recently, RNA sequencing has failed to demonstrate any AID-dependent RNA editing in activated B-cells (165), and AID-dependent accumulation of deoxyuracil has been demonstrated within the
Ig locus (88). This direct observation of uracil in genomic DNA provides the strongest evidence to date in favor of the DNA deamination model, and this model is now widely accepted.

Despite a compelling body of evidence in favor of the DNA deamination model, several key questions remain unanswered regarding AID’s mechanism of action. AID’s promiscuous binding interactions with RNA (16, 115) demonstrate the lack of any molecular explanation for AID’s nucleic acid selectivity. In fact, purified AID requires RNase treatment before DNA deamination activity can be observed, suggesting that RNA can competitively bind in the enzyme’s nucleic acid binding site (16). A similar pattern of DNA deamination despite RNA binding is seen with APOBEC3 homologs of AID that contribute to restriction of retroviruses such as HIV. In this case, RNA binding appears critical for the appropriate packaging of these restriction factors within their retroviral targets (166-168). Similar to AID, RNA deaminase activity has never been described for any APOBEC3 family members. Thus, there remains the persistent and unanswered question of how AID can freely interact with both RNA and DNA, yet retain catalytic specificity for DNA. Given the established and emerging roles of AID/APOBEC biology, there is a pressing need to elucidate the mechanisms of AID’s nucleic acid selectivity to provide insights into how cytosine is recognized and deaminated.

3.1.2 Experimental Approach

To elucidate the answers to this question, we interrogated the molecular basis of AID’s selectivity for DNA over RNA. RNA is distinguished from DNA by the presence of the 2’-(R)-hydroxyl group in the sugar of the nucleotide building blocks [Figure 15A]. Given that this hydroxyl group introduces a steric constraint and alters the conformational preferences of the nucleotide [Figure 15B], we reasoned that this 2’-substituent of the target cytosine itself could be an important determinant of deamination selectivity. To test this hypothesis, we synthesized chimeric substrates that contain 2’-modifications in a single target cytosine embedded within an otherwise DNA backbone: DNA with a single 2’-deoxycytidine (D-dC), 2’-ribocytidine (D-rC), 2’-fluororibocytidine (D-frC), 2’-arabinocytidine (D-aC), or 2’-fluoroarabinocytidine (D-faC) [Figure 15C]. This target cytosine was placed within an AGCT hotspot motif to ensure physiologically-
relevant sequence specificity and efficient deamination. We then assayed these substrates with purified, recombinant AID to test for differential reactivity.

**Figure 15: Nucleotide sugar pucker as a potential basis for DNA selectivity.** (A) DNA and RNA are distinguished by the presence of a 2’-(R)-hydroxyl substitution, as indicated in magenta. (B) 2’-substitution affects the equilibrium of nucleotide sugar conformations. DNA prefers a C(2’)-endo conformation, commonly known as ‘south’. RNA prefers a C(3’)-endo conformation (or ‘north’). (C) Chimeric DNA substrate design. Sequence of chimeric substrate is listed, with chimeric cytosine indicated in magenta. Chimeric cytosine conformers vary in their 2’-substituents.

### 3.2 Results

#### 3.2.1 2’-Substituents of Target Cytosine Regulate Deamination by AID

Traditional biochemical assays for deamination rely upon UDG for detection of reaction products. Since UDG does not recognize several nucleotide conformers used in our chimeric
substrates, we developed a novel assay to screen for deamination and found that DNA restriction endonucleases were capable of assaying the full panel of substrates [Figure 16A]. Specifically, the restriction endonuclease FspBI was surprisingly tolerant to the presence of a chimeric non-DNA base within its target, and could accurately discriminate between a C:G match or a U:G mismatch within its recognition sequence [Figure 16B]. Of note, the isoschizomer BfaI also tolerated these same chimeric substrates [Figure 16C], suggesting that other DNA endonucleases may be able to recognize chimeric non-DNA substrates.

**Figure 16: Restriction-endonuclease-based assay for deamination.** (A) Assay design. Following AID incubation, substrates are duplexed to a complementary strand that completes a palindromic duplex substrate for endonuclease cleavage. If the target cytosine is unreacted, the
duplex is cleaved; if target cytosine is deaminated to uracil, the duplex is not cleaved. (B) Reactivity of chimeric DNA substrates against FspBI. Above each cytosine conformer, the nucleotide sugar is shown. (C) Reactivity of chimeric DNA substrates against Bfai.

As expected, one-hour incubation of AID with the D-dC substrate yielded robust deamination [Figure 17A]. This substrate is entirely DNA and represents the canonical substrate for deamination. D-rC differs from D-dC only by the presence of a 2'-hydroxyl at the target cytosine; the presence of this single molecular substitution was sufficient to compromise deamination of the otherwise entirely DNA substrate. Incubation of AID with D-frC yielded similarly negligible deamination, demonstrating that both 2'-hydroxyl- and 2'-fluoro- substitutions with (R)-stereochemistry impair deaminase activity. By contrast, AID more efficiently deaminated cytosine substrates when the 2'-hydroxyl- and 2'-fluoro- substituents were present in epimeric, or inverted, configurations in D-aC and D-faC. The trend observed with the one-hour incubation was reflected in different initial rates of turnover at early time points [Figure 17B].
Figure 17: 2’-substitution of target cytosine is sufficient to disrupt AID deaminase activity. (A) Chimeric DNA substrates incubated with AID for 1 hour and assayed for deamination with restriction-endonuclease-based assay. Target nucleotide sugar is shown above each respective substrate. (B) Chimeric DNA substrates (250 nM) incubated with AID (500 nM) for multiple time-points up to one hour. Total product formation is graphed as a function of time (n=3 replicates, standard deviation shown). White symbols indicate target cytosines that prefer C(2’)-endo conformation, while black symbols indicate a preference for C(3’)-endo conformation.

To confirm the reproducibility of sensitivity to the 2’-(R)-hydroxyl, we designed an additional series of substrates to test the impact of this substitution [Figure 18A]. Though these substrates all bear some deviation from those described in Figures 16 and 17, they remained amenable to the restriction-endonuclease-based assay for deamination. First, we evaluated the reactivity of ribocytidine within the context of a coldspot and alternative hotspot [Figure 18B].
Neither of these conditions yielded detectable deamination. Deoxycytidine controls were ready deaminated, albeit at different levels as expected. Given that, physiologically, AID is thought to target cytosine in transcription bubbles, we also evaluated differential reactivity of the two cytosine conformers within a bubble substrate. The bubble was formed by annealing a single-stranded substrate to a complementary strand bearing a mismatch of 5 nucleotides. This result confirmed the selectivity seen with the coldspot and alternative hotspot. Additionally, the same degree of selectivity was seen with a 7-nucleotide bubble substrate as well (data not shown).

Lastly, we evaluated AID’s selectivity within a physiologically relevant sequence, chosen from the consensus sequence for the mouse Sα switch region. This sequence contained multiple hotspots for AID deamination; however, only one of those hotspots overlapped with the FspBI recognition sequence, enabling a clear readout for deamination. AID selectively deaminated the substrate containing deoxycytidine and not ribocytidine, ultimately validating the sensitivity of AID to the 2′-(R)-hydroxyl substitution of cytosine across a full panel of four additional substrates (149, 169).
Figure 18: Incubation of AID with chimeric substrates in alternative sequence contexts. (A) Sequences of additional chimeric DNA substrates containing dC or rC conformers at the target position. Substrates include coldspot and alternative hotspot targeting motifs; a bubble substrate, in which substrate is annealed to a complementary strand containing 5 mismatched nucleotides surrounding the target cytosine; and a single-stranded DNA oligonucleotide sequence derived from a consensus sequence for the mouse Sa region. For consensus sequence, additional AGCT hotspot motifs are underlined. All substrates were assayed for deamination with restriction-endonuclease-based assay. (B) Additional chimeric DNA substrates (250 nM) were incubated with AID (500 nM) for 1 hour and assayed for deamination with the FspBI restriction-endonuclease-based assay. These results confirm the sensitivity to 2’-(R)-hydroxyl substitution at the target cytosine in all contexts tested, consistent with results from Figure 17.

We used a truncated form of AID lacking its terminal exon (residues 181-198) in our experiments. This version of AID has previously been shown to retain identical sequence targeting and has increased enzymatic activity compared to the full-length enzyme (140), thereby
increasing the dynamic range of the enzyme’s reactivity and permitting the study of less favored substrates. Nevertheless, to confirm that this truncation does not affect AID’s sensitivity to 2’-substitution of the target cytosine, we purified full-length AID and repeated our experiments. As expected, when we also assayed the chimeric substrates with full-length AID, the enzyme demonstrated similar preferences, although overall enzymatic activity was diminished [Figure 19].

**Figure 19: Full-length AID demonstrates similar reactivity compared to truncated version.** Chimeric DNA substrates (250 nM) incubated with full-length AID (500 nM) for 1 hour and assayed for deamination with restriction-endonuclease-based assay. Full-length AID shows a pattern of preferences for chimeric substrates that is consistent with that in Figure 17, obtained using the truncated version of AID.

To obtain a more rigorous comparison between substrates, we returned to the hyperactive, truncated version of AID and repeated our deaminase assays under extended-incubation conditions. For a quantitative comparison of substrate reactivity, we further determined product formation as a function of enzyme concentration with extended twelve-hour incubations. We focused on a range where product formation was linearly dependent upon enzyme concentration and compared the overall substrate reactivity by normalization to those values obtained with D-dC [Figure 20A]. Values were obtained for all substrates except D-rC, as its linear regression did not significantly deviate from zero. This necessitated the determination of the lower limits of detection for the FspBI restriction-endonuclease-based assay [Figure 20B]. A standard curve was created by mixing D-rC and D-rU in various ratios, assaying with FspBI, and comparing the observed and actual percentages of D-rU. The standard curve demonstrated that the FspBI provides an accurate quantification of product formation. At the lower end of the curve,
we determined the limit of detection of deamination as 0.5%, making the assay a sensitive measure of D-rC detection.

Based on the lower limits of detection for D-rC deamination, the D-dC substrate was deaminated over 500-fold more efficiently than the D-rC substrate, highlighting AID’s striking degree of sensitivity to the 2’-(R)-hydroxyl at the target cytosine [Figure 20A]. The full degree of sensitivity is like greater than this value, limited here by the inefficient DNA deaminase activity of AID and our limits of detection for D-rC deamination. When comparing epimers of the same substituent (D-rC vs D-aC for 2’-hydroxyl-; D-frC vs D-faC for 2’-fluoro-), the arabinosyl epimers D-aC and D-faC were preferred by more than an order of magnitude. For the 2’-fluoro substituent, the arabinosyl epimer was preferred nearly 40-fold. Though the estimate for the 2’-hydroxyl substituent is constrained by the lower limit of detection, the arabinosyl epimer is preferred at least 40-fold. Thus, quantitative analysis confirms AID’s sensitivity to 2’-substitution of cytosine and reveals a strong preference for arabinosyl conformers over ribosyl counterparts.
Figure 10: Quantitative assessment of AID deamination of chimeric DNA substrates. (A) Deaminase activity of AID against chimeric DNA substrates (250 nM) as a function of enzyme concentration. Product formation of deamination of chimeric substrates is shown as a function of increasing AID concentration (n = 3 replicates; standard deviation shown). Absolute and relative deaminase activities were determined by linear regression and are listed with each substrate in figure legend. Linear regression of data for D-rC does not significantly deviate from zero; therefore, values for D-rC are based on assay detection limits determined in part B. (B) Determination of lower limit of detection for D-rC deamination. A standard curve is shown consisting of D-rC and D-rU mixed in varying ratios (n=3 replicates, standard deviation shown). Inset is the magnification of the low end of the standard curve. Statistical significance from zero was determined by a student’s T-test. NS: p > 0.05, Single Asterisk: p < 0.05, Double Asterisk: p < 0.005. 0.5% D-rU in total reaction is statistically significant limit of uracil detection, yielding upper limit of D-rC deaminase activity as 0.003 nM Product/nM AID.
With a clear pattern of differential reactivity against chimeric DNA substrates, we wanted to determine whether these results might result from preferential binding. Using fluorescence polarization, we determined that AID bound with a similar affinity across the series of substrates, suggesting that altered binding was not a sufficient explanation for selectivity [Figure 21A, B]. This is consistent with previous studies that have found limited changes in binding affinity when only a few bases are modified within a larger DNA substrate. Notably, because AID was purified as a recombinant fusion protein with maltose-binding protein (MBP), we examined binding with MBP alone to ensure that substrate binding was specifically due to AID. The failure to induce any change in polarization confirms that MBP does not contribute to substrate binding [Figure 21C].

**Figure 21: AID binding to chimeric DNA substrates.** (A) Fluorescence polarization assay for determination of binding affinity. Decreased tumbling upon formation the enzyme-nucleic acid complex leads to increase in fluorescence polarization. (B) Fluorescence polarization of chimeric DNA substrates (5 nM) in the presence of increasing concentrations of AID (n=3 replicates, standard deviation shown). Y-axis depicts the fraction of bound substrate. K_d values are reported in figure legend. Nanomolar concentrations of MBP-AID fusion protein were sufficient to induce changes in fluorescence polarization, indicative of protein binding. (C) Fluorescence polarization of chimeric DNA substrates (5 nM) in the presence of increasing concentrations of MBP (n=3 replicates). Micromolar concentrations of MBP alone were unable to induce any change in fluorescence polarization, confirming that binding affinities demonstrated in part B are specifically a result of AID in the MBP-AID fusion protein.
3.2.2 2’-Substituents of Target Cytosine Regulate Deamination by APOBEC1

The notion that AID could deaminate RNA originated from its homology to APOBEC1 (87). Despite this homology, AID acts on the immunoglobulin locus DNA, while APOBEC1 is physiologically known to target apolipoprotein B mRNA. Given the similarities and contrasts between these two deaminases, we sought to determine APOBEC1’s sensitivity to 2’-substitution of the target cytosine. As with AID, we created a series of chimeric DNA substrates containing 2’-deoxycytidine, 2’-ribocytidine, and 2’-fluororibocytidine embedded within a sequence context preferred by APOBEC1 (A1-D-xC, Table 1) (110). These substrates were incubated with APOBEC1 and evaluated using the FspBI restriction-endonuclease-based assay for deamination.

We found that APOBEC1 readily deaminated the A1-D-dC substrate, in line with previous reports that have demonstrated the enzyme’s robust DNA deaminase activity [Figure 22] (115-117). Remarkably, despite its physiological targeting of RNA, APOBEC1 deaminated A1-D-dC more efficiently than the A1-D-rC and A1-D-frC substrates, a trend additionally reflected in the differential reactivity at early time points.
Figure 22: 2'-substitution of target cytosine disrupts APOBEC1 deaminase activity. (A) Restriction-enzyme based assay of chimeric substrates (250 nM) in the presence of APOBEC1 (1.5 µM) for 1 hour. (B) Chimeric substrates were incubated with APOBEC1 for various time points over the course of one hour. Total product formation is graphed as a function of time (n = 3 replicates; standard deviation shown).

For a quantitative comparison of reactivity, as done for AID, we also determined product formation for each substrate as a function of APOBEC1 concentration with extended 12-hour incubations. The values for each substrate were normalized to that obtained with A1-D-dC [Figure 23]. Whereas deamination of D-rC was undetectable in the presence of AID, APOBEC1 yielded detectable A1-D-rC deamination. APOBEC1 deaminated the A1-D-dC substrate 110-fold more efficiently than the A1-D-rC substrate and 20-fold more efficiently than the A1-D-frC.
Compared with AID, APOBEC1 demonstrates increased tolerance of the substrates containing ribosyl epimers of cytosine.

![Quantitative analysis of APOBEC1 deaminase activity against chimeric DNA substrates.](image)

**Figure 23:** Quantitative analysis of APOBEC1 deaminase activity against chimeric DNA substrates. Product formation of deamination of chimeric substrates (250 nM) is shown as a function of increasing APOBEC1 concentration for 12 hour incubations (n = 3 replicates; standard deviation shown). Absolute and relative deaminase activities are listed with each substrate in figure legend.

As with AID, we evaluated APOBEC1’s substrate binding affinities to rule out any preferential binding. APOBEC1 bound all substrates with similarly high affinity, indicating that the differences in reactivity were not a function of altered binding [Figure 24].

![APOBEC1 binding to chimeric DNA substrates.](image)

**Figure 24:** APOBEC1 binding to chimeric DNA substrates. Fluorescence polarization of chimeric DNA substrates (5 nM) in the presence of increasing concentrations of APOBEC1 (n=3 replicates, standard deviation shown). Y-axis depicts the fraction of bound substrate. $K_d$ values are reported in figure legend. Nanomolar concentrations of MBP-AID fusion protein were sufficient to induce changes in fluorescence polarization, indicative of protein binding.
3.2.3 Minimal DNA Requirements for Deamination by AID

Analysis of the chimeric DNA substrates demonstrated that RNA-like 2'-substitutions to the target cytosine are sufficient to compromise AID’s deaminase activity, underlying AID’s specificity for DNA. To gain further insight into the molecular determinants of AID’s deamination activity, we wanted to determine whether the inverse were also true: would removal of the 2’-substituent from the target cytosine of an RNA substrate rescue deamination?

To address this question, we synthesized chimeric oligonucleotides consisting of 2’-fluororibo-nucleotides (2’-F-RNA) with a varied number of DNA nucleotides embedded at the target cytosine and neighboring positions [Figure 25A]. 2’-F-RNA was selected for its stability relative to RNA. As DNA endonucleases were intolerant of these predominantly non-DNA substrates, we designed a primer-extension assay for deamination using reverse transcriptase (RT) [Figure 25B]. RT tolerated both 2’-F-RNA and our chimeric templates, properly incorporating the chain-terminator ddATP opposite the 0 position of the uracil product controls and the -2 position of cytosine substrates during primer extension [Figure 25C].
Figure 25: Reverse-transcriptase-based assay for deamination of chimeric 2'-F-RNA substrates. (A) Substrate design. In addition to chimeric substrates, entirely DNA and 2'-F-RNA substrates were designed as positive and negative controls for deamination, respectively. (B) Assay design. Primer extension in the presence of ddATP yields a 10 base pair extension product in the presence of deamination, and a 12 base pair extension product when the template strand is not deaminated. (C) Reverse transcriptase assay using AID-incubated (1.4 µM) chimeric 2'-F-RNA substrates and controls (250 nM) as templates for primer extension.

When incubated with AID for one hour, the entirely DNA control (D-control) was readily deaminated while entirely 2'-F-RNA substrate (F-control) was not, as expected given AID’s selectivity against RNA [Figure 26]. DNA at the target cytosine alone (F-d(0)) rescued deamination negligibly compared to the all-DNA control, indicating that removing additional 2'-substituents from neighboring nucleotides may be necessary to rescue deamination. Expanding beyond the target cytosine, removing the 2'-F from the -1 position (F-d(-1:0)) also yielded
negligible, though slightly increased, deamination. On the other hand, expansion of the chimeric DNA patch from the -2 to the 0 position (F-d(-2:0)) fully rescued deamination to the level seen with the all-DNA control, as did further expansion of the DNA patch to the -3 to +1 positions (F-d(-3:+1)).

Figure 26: Rescue of deamination of chimeric 2'-F-RNA oligonucleotides requires a DNA patch from positions -2 to 0. Chimeric substrates incubated with AID for one hour and evaluated via the Reverse Transcriptase-based assay. Asterisk denotes band associated with deaminated product.

To verify these results, we verified our findings using the UDG-based assay for deamination. Neither the D-control nor F-control substrates were compatible with this deamination assay, given the overabundance of deoxyuracil in D-control and the lack of UDG reactivity of 2'-fluororibouracil in F-control. Though the positive and negative controls were lacking, the deamination status of all chimeric substrates was accurately reported [Figure 27]. As before, deamination was rescued only in the F-d(-2:0) and F-d(-3:+1) substrates. This result confirms the importance of the 2'-substituent at the two nucleotides upstream of the target cytosine.
Figure 27: UDG-based, alternative assay for deamination of chimeric 2'-F-RNA substrates. (A) After incubation with AID, the 3'-end labeled oligonucleotides are incubated with UDG, which generates abasic sites at deoxyuridine residues, with no significant reactivity with 2'-fluororibouridine. Deamination of all chimeric substrates is detectable by formation of a 15 bp product after alkaline-induced cleavage of abasic sites. The presence or absence of dU at the -2 position dictates whether unreacted cytosine-containing oligonucleotides generate a 17 bp product or the 30 bp unreacted substrate, respectively. (B) Chimeric substrates assayed for deamination with UDG-based assay. As standards, controls in the absence of AID are shown with each substrate. For the samples that received one hour incubation with AID, minimal deamination is seen with F-d(0) and F-d(-1:0), while deamination is seen with F-d(-2:0) and F-d(-3:+1). This independent assay confirms the findings of the Reverse Transcriptase-based assay.

To evaluate the contribution of binding to deamination, we determined AID’s binding affinity for 2'-F-RNA chimeras and DNA/2'-F-RNA controls by fluorescence polarization measurements. AID showed a similar affinity for all the substrates, with a slight preference for the predominantly 2'-F-RNA substrates over the all-DNA control [Figure 28]. DNA binding affinity was approximately 1.5-3 fold worse than the substrates containing 2'-F-RNA. Whereas a longstanding assumption has been that tighter binding is correlated with more efficient deamination, our results indicate somewhat to the contrary: the canonical, favored DNA substrate...
is bound slightly less efficiently than the disfavored 2'-F-RNA substrate. Given the similar conformations, 2'-F-RNA binding dynamics likely recapitulate those with true RNA.

![Figure 28: AID binding to chimeric 2'-F-RNA substrates.](image)

Fluorescence polarization of chimeric 2'-F-RNA substrates was determined in the presence of increasing concentrations of MBP-AID fusion protein (n=3 replicates, standard deviation shown).

3.3 Discussion

3.3.1 Mechanistic Basis for AID's Inherent Preference for DNA Deamination

We have demonstrated that AID's deaminase activity is strongly influenced by the 2'-substituents of its target nucleotides. Introducing a single 2'-(R)-OH at the target cytosine nucleotide of an otherwise DNA substrate was sufficient to disrupt deaminase activity by at least 500-fold. Inversely, removing the 2'-fluoro substituents from the target cytosine and two upstream nucleotides rescued deamination of a 2'-F-RNA substrate. These data indicate that 2'-substitution of the nucleotide sugar is an important molecular determinant of selectivity for AID's deaminase activity.

What is the molecular basis of the deaminases' preference for DNA over RNA? Biophysically, there are two closely related, principal mechanisms by which the 2'-OH substitution may enable them to distinguish between the two nucleic acids. The first method of discrimination is steric exclusion, wherein steric interactions with the 2'-OH of RNA exclude the target.
nucleotide. A ‘steric gate’ mechanism has been observed in DNA polymerases (170) and base excision repair enzymes, including uracil DNA glycosylase (171). While the 2'-OH provides a basis for steric exclusion, it also results in conformational differences between the nucleotide sugar of RNA and DNA, known as sugar pucker. DNA prefers a C(2')-endo (south) conformation, while RNA is more restricted to an alternative C(3')-endo (north) conformation [Figure 15] (172). For enzymes that modify nucleic acids, alternative sugar puckers can impact the reactivity or positioning of the substrate in the active site by altering the angular projection of the base from the sugar-phosphate backbone, with RNA ligases serving as one example (173).

AID’s relative reactivity with the chimeric DNA substrates suggests that sugar pucker of the target nucleotide is critical for catalysis. Specifically, the cytosine conformers that prefer the C(2')-endo (south) conformation—dC, aC, and faC (174)—were readily deaminated. Amongst these three substrates, D-dC was most favored over D-faC and D-aC, correlating reactivity with smaller 2'-S-substituents. This observation suggests that steric exclusion of the 2'-substituent may provide a secondary factor for discrimination. Further highlighting the importance of these nucleic acid determinants, the conformers that prefer the C(3')-endo (north) conformation—rC and frC (175)—were disfavored for deamination, even when embedded in a nucleic acid substrate that is otherwise entirely DNA. Prior studies have shown that chimeric nucleotides embedded in DNA independently retain their sugar pucker and perturb local helix formation (176, 177), indicating that the cytosine conformers used in this study are likely to retain their expected north-south conformations in our chimeric DNA substrates. Taken together, reactivity with our chimeric substrates supports the notion that an isolated, disfavored sugar pucker at the target cytosine is sufficient to disrupt AID’s deaminase activity and underlies selectivity for DNA. More broadly, this model implicates sugar pucker as a potential discrimination mechanism by which APOBECs and other cellular enzymes may distinguish DNA from RNA at the level of individual nucleotides.

Moving beyond the target cytosine, our study shows that the 2'-substituents play a critical role at neighboring nucleotide positions as well. Removal of the 2'-fluoro substituent from the -1
and -2 positions was necessary to rescue full deamination of 2'-F-RNA chimeric substrates. Our finding regarding the significance of the -2 to 0 positions aligns well with the identification of the “hotspot recognition” loop within AID that specifically targets deaminase activity to the cytosine of WRC trinucleotide hotspots (140, 145, 146), as well as nucleoside analog interference studies with the homolog APOBEC3G (178). Taken together, the most significant determinants of nucleic acid recognition—sugar, backbone and nucleobase recognition—appear to be confined to this critical trinucleotide patch spanning the -2 to 0 positions.

3.3.2 AID’s Inherent Preferences Support the DNA Deamination Model of Antibody Maturation

The DNA deamination model for AID’s role in antibody maturation is well supported by a bevy of genetic, cellular, and bacterial experiments that affirm DNA as AID’s true target (137). The RNA deamination hypothesis has long been largely abandoned, with the exception of continued reports from the Honjo group that fail to demonstrate any biochemical reactivity between AID and ribocytidine (179). This limited biochemical plausibility of the RNA deamination hypothesis is dwarfed by an abundance of evidence in favor of the DNA deamination model, leaving little controversy regarding AID’s true substrate. Nevertheless, the model has suffered from the lack of a mechanistic basis for selectivity.

Beyond their contributions to the body of evidence supporting the DNA deamination model, our biochemical insights contribute a mechanistic basis for AID’s nucleic acid selectivity. AID’s inherent preference for DNA over RNA is consistent with its physiologic activity. This inherent preference is likely rooted in the nucleotide sugar pucker. While these preferences were not assessed in vivo, it is highly likely that our in vitro biochemical characterizations accurately recapitulate AID’s physiologic preferences, given the preservation of other characteristics such as sequence targeting. Our study additionally provides a hypothesis that will guide further characterization of the mechanistic basis for DNA selectivity, presumably through resolution of the structural characteristics of the enzyme-substrate complex.

3.3.3 APOBEC1’s Inherent Preference for DNA Contrasts with its Physiologic RNA Activity
When assayed against a similar panel of chimeric DNA substrates, APOBEC1 also demonstrated a high degree of sensitivity to the 2'-substituent of the target cytosine. Like AID, APOBEC1 favored the A1-D-dC substrate over A1-D-rC and A1-D-frC, both of which prefer the C(3')-endo conformation at the target base. The similar pattern in favor of DNA deamination suggests that the molecular determinants of the deamination reaction have been conserved since the evolutionary divergence of APOBEC1 and AID, indicating that these determinants of deamination remain conserved across the AID/APOBEC family.

On one hand, these inherent preferences align well with recent studies on ancestral homologs of APOBEC1 that suggest that the enzyme was initially a DNA deaminase and evolved physiological RNA deaminase activity only recently (118). On the other hand, the discordance between APOBEC1’s intrinsic preference for DNA and its known physiological RNA deaminase activity also demonstrates that an enzyme’s intrinsic preferences alone do not dictate its physiological activity. Thus, our study also leaves open the possibility that AID, too, could overcome its intrinsic preferences within the cell, or that APOBEC1 may possess an as-of-yet undescribed cellular function against DNA.

If both AID and APOBEC1 indeed have greater deaminase activity against DNA, why would AID physiologically target DNA and APOBEC1 target RNA? The reason for this discrepancy may be two-fold. First, AID and APOBEC1 have distinct protein-binding partners that recruit them to their respective target nucleic acids. ACF recruits APOBEC1 to AU-rich mooring sequences on mRNA transcripts (110, 111), whereas RPA and Spt5 facilitate AID’s interactions with the B-cell genome (91, 93). It is feasible that ACF induces a permissive conformation in enzyme or substrate, or that target RNA sites intrinsically adopt a sugar pucker that permits deamination. Second, while APOBEC1 appears to prefer DNA-like sugar pucker at the target site, our data suggest that APOBEC1 also has a greater inherent tolerance of the rC substrate than AID. When measured relative to the dC substrates, APOBEC1 deaminated its D-rC substrates at least 4-fold more readily than AID. As this semi-quantitative comparison relies on the lower-limit of detection for AID’s D-rC deaminase activity, APOBEC1’s tolerance is likely
greater than this conservative estimate. Taken together, our data suggest that APOBEC1 has evolved two potentially distinct mechanisms to tolerate RNA for deamination. For future study, our findings raise the intriguing question of how APOBEC1 overcomes its inherent preferences to deaminate RNA within the cell.
4.1 Introduction

Transcriptional variability and adaptability are particularly necessary in responding to the challenges of multicellular life. As part of nature’s enzymatic toolbox, methylation of cytosine at the 5-position of the base represses gene expression and shapes cellular identity. The reverse of this process, active DNA demethylation, is equally important for cleaning the genomic slate during embryogenesis or achieving rapid, locus-specific reactivation of previously silenced genes.

The mechanism of DNA methylation has been rigorously established, catalyzed by a family of DNA methyltransferases that initiate and maintain the epigenetic memory of the methyl mark. Strikingly, the mechanisms of DNA demethylation have remained enigmatic. Studies on demethylation have been marred by a long history of seemingly disparate observations that have failed to coalesce into a consistent model (180). Against such a backdrop, this chapter explores biochemical preferences of the AID/APOBEC family to demystify the role of deamination in DNA demethylation.

4.1.1 Proposed Mechanisms of DNA Demethylation

Although active demethylation is increasingly accepted as an important physiological process, its molecular basis remains controversial. Several DNA glycosylases have been described in Arabidopsis that can excise mC specifically; however, mammals appear to lack this activity (181). In the past several years, a wealth of new evidence has implicated several of the key cytosine-modifying enzymes we have reviewed, particularly the AID/APOBEC deaminases, TET oxidases, and DNA glycosylases (182). Two major types of models have emerged: a deamination-initiated pathway and several variants of an oxidation-initiated pathway [Figure 29].
Figure 29: Proposed non-canonical role for AID/APOBEC enzymes acting on modified cytosine substrates in DNA. (A) Deamination of cytosine plays known physiological roles in adaptive immunity (AID), innate immunity against retroviruses (APOBEC3 enzymes), and mRNA editing (APOBEC1). These canonical roles involve deamination of cytosine to generate uracil. (B) DNA demethylation involves the regeneration of unmodified cytosine from mC. (C) Proposed Demethylation pathways. The function of AID/APOBEC family members on modified cytosine residues remains poorly understood despite their implication in potential pathways for active DNA demethylation. Deamination of mC or hmC, the product of TET-mediated oxidation, could generate thymidine or 5-hydroxymethyluracil (hmU), respectively. Base excision repair (BER) could subsequently excise the deaminated bases and replace them with unmodified cytosine. An alternative deamination-independent pathway involves iterative oxidation, generating 5-formylcytosine (fC) or 5-carboxylcytosine (caC). BER-mediated excision of the oxidized cytosine would result in reversion to unmodified cytosine.

Two types of deamination-dependent mechanisms have been postulated. In one scenario, deamination of mC by an AID/APOBEC enzyme generates a T:G mismatch leading to subsequent repair by the BER enzyme thymidine DNA glycosylase (TDG) (183). Alternatively, hmC could be deaminated by an AID/APOBEC enzyme to generate 5-hydroxymethyluracil (hmU) which could also be reverted to cytosine by BER (32, 38). Recent studies have also demonstrated the feasibility of a deamination-independent pathway for DNA demethylation involving oxidation of mC by TET enzymes. The product of this oxidation, hmC, can itself be iteratively oxidized to yield both 5-formylcytosine (fC) and 5-carboxylcytosine (caC) (14, 15). These higher oxidation products are detectable in the genome of embryonic stem cells and are good substrates for excision by TDG, which could ultimately regenerate unmodified cytosine.
Notably, deficiency in TDG, a potential common mediator in the various proposed pathways for DNA demethylation, is associated with developmental methylation defects and embryonic lethality (31, 32).

### 4.1.2 Evidence in Support of Deamination-Dependent Demethylation and Remaining Questions

The plausibility of deamination-dependent demethylation has been difficult to establish because of the poorly characterized activities of AID/APOBEC enzymes on C5-modified cytosines and a lack of knowledge about the functional redundancy between AID/APOBEC family members (184). Although prior studies suggest that AID’s ability to deaminate 5mC is reduced relative to its ability to deaminate cytosine (16, 183), other work proposes that the enzyme lacks any 5mC deaminase activity (19). Additional ambiguity arises because the activities of other APOBEC enzymes on 5mC have not been directly investigated, and the biochemical activities of all AID/APOBECs against 5hmC remain entirely unknown.

Further, the presence of numerous AID/APOBEC family members presents a substantial challenge to sorting out their potential roles in demethylation. A role for AID in demethylation of pluripotency promoters is suggested from heterokaryon-based systems for the generation of stem cells (185), and AID deficiency has also been found to perturb the methylome of primordial germ cells (186). However, these observations are confounded by the finding that AID deficiency is viable in both mice and humans (87), suggesting that other deaminases might also serve functionally redundant roles in demethylation. In support of this proposal, APOBEC2 enzymes were postulated to play a role in zebrafish DNA demethylation (121) and APOBEC1 has been implicated in neuronal DNA demethylation (38). Thus, biochemical characterization of the similarities and differences between these deaminases would address the functional redundancy of these enzymes in DNA demethylation. The previous implications that deaminases might be involved in DNA demethylation (136, 187) make it important to examine their activity on 5-substituted cytosine bases in DNA and the plausibility of deamination-dependent DNA demethylation.
4.2 Results

4.2.1 AID/APOBECs Preferentially Deaminate Unmodified Cytosine

We wished to profile the reactivity of representative AID/APOBEC family members with modified cytosine nucleobases. We chose to investigate the mouse enzyme family, which possesses only a single gene for each family member, rather than the human family, where extensive gene duplication and specialization at the A3 locus have generated many variants. Mouse APOBEC1 (mAPOBEC1), APOBEC2 (mAPOBEC2), APOBEC3 (mAPOBEC3) and AID (mAID) were generated as N-terminal maltose binding protein fusion constructs. Although mAID was inactive under these conditions (data not shown), we had previously expressed and characterized active human AID (hAID) by co-expression of the enzyme with the chaperone Trigger Factor in *E. coli* (140). Using this expression system, the full cohort of AID/APOBEC enzymes were all soluble and were partially purified over amylose resin as described in the materials and methods section.

We designed DNA oligonucleotides containing a single cytosine residue with several criteria in mind [Figure 30A]. First, since each AID/APOBEC family member prefers to deaminate cytosine in a different trinucleotide sequence context (139), we selected a universal sequence that would be acted upon by multiple family members (S30-TGC). Next, a guanine was introduced directly downstream of the cytosine to create a CpG motif, an important consideration given that epigenetic modifications via methylation, hydroxymethylation and demethylation are highly linked to CpG sites and islands in the mammalian genome.
Figure 30: Substrate and assay design to screen for deamination of substrates containing 5-modified cytosine. (A) Fluorophore (FAM)-labeled oligonucleotides (S30) synthesized with a single internal modified cytosine (red) embedded in a CpG motif. The preceding two bases, TG, provide a hotspot for deamination that is universally targeted by all AID/APOBEC family members. (B) After incubation with deaminases, oligonucleotides were duplexed with a complementary strand generating U:G, T:G, or hmU:G mismatches with the deaminated substrates (blue). Treatment with UDG (reactive with U:G), TDG (reactive with T:G) or SMUG (reactive with hmU:G), respectively, followed by base-mediated cleavage leads to fragmentation of deaminated products to a 15-mer (P15).

AID/APOBEC family members were assayed against the cytosine-containing substrate S30-TGC using a discontinuous, uracil DNA glycosylase (UDG)-coupled assay (Figure 30B). At the end of the deamination period, the reaction product was hybridized to a complementary strand, yielding duplexed DNA containing a U:G mismatch in the deaminated product. Treatment with UDG generated abasic sites in the deaminated oligonucleotides, while leaving unreacted substrates intact. Cleavage of the abasic sites under alkaline conditions allowed for specific detection of product after separation on a denaturing gel. To next examine the deamination activities on physiologically relevant 5-modified cytosines, we synthesized S30-TGmC and S30-TGhmC. Upon deamination, these substrates would convert to T and hmU, respectively. Since the coupling enzyme UDG is not active on either of these products, we needed to identify other glycosylase enzymes that would allow for excision of these products (171). Two useful enzymes for this purpose were TDG, which excised thymidine from T:G mismatches in CpG contexts, and SMUG, which excised hmU mispaired to G. We established that these glycosylase-coupled
assays accurately detect low levels of deamination products by generation of a standard curve (as little as 0.5% product under the condition of our deamination reaction) [Figure 31].

Figure 31: Detection limit and standard curve for deamination detection. The calculated amount of substrate to product DNA is plotted against the actual amount, showing that the DNA glycosylases are useful for detection of low levels of deamination products. The dotted line
represents the theoretical maximum. The limit of detection in a 100nM reaction is 0.5 nM deamination product. The mean and standard deviation from 3-4 independent replicates are shown in the graph.

Under these conditions the mAPOBEC1, mAPOBEC3 and hAID variants all showed deaminase activity against S30-TGC [Figure 32A]. As anticipated from prior studies mA2 showed no detectable cytosine deaminase activity (119). Using the TDG-coupled assay, deamination of S30-TGmC was detectable with all active AID/APOBEC enzymes, though product formation was decreased relative to S30-TGC [Figure 32B]. By contrast, no detectable deamination activity was evident against S30-TGhmC [Figure 32C], despite robust deamination of S30-TGC under identical conditions with mAPOBEC1, mAPOBEC3 and hAID. For the APOBEC2 deaminases, which have no known catalytic activity, but have been postulated to play a role in zebrafish demethylation (121), we found that mAPOBEC2 was also inactive against S30-TGmC or S30-TGhmC in vitro.
Figure 32: AID/APOBEC enzymes preferentially deaminate unmodified cytosine. The reaction products resulting from 12 hour incubation of substrates (200 nM) with mA1, mA2, mA3 or hAID (2 µM) are shown separated on a denaturing gel. (A) S30-TGC substrates were assayed with UDG; (B) S30-TGmC substrates were assayed with TDG; (C) S30-TGhmC substrates were assayed with SMUG.

To ensure that the discrimination against 5-substituted cytosine bases was not simply limited to a single sequence context, or perhaps the single enzyme concentration used in the above studies, we performed additional investigations. First, we tested a series of substrates containing C, mC and hmC in an ATX trinucleotide context, which is a preferred sequence for mAPOBEC1. As expected, reaction of S30-ATC with mA1 led to robust deamination, while S30-ATmC was compromised and deamination of S30-AThmC was undetectable [Figure 33A]. When higher concentrations of mAPOBEC1 were employed, deamination of S30-ATC was nearly complete, while at the highest concentrations of mA1 only ~20% of the S30-ATmC was deaminated, and no deamination of S30-AThmC was ever detected (<0.5%). Deamination was observed to be linearly dependent on the amount of enzyme at lower mAPOBEC1 concentrations, allowing the relative efficiencies for deamination of these substrates to be determined. Under these conditions, mAPOBEC1 was estimated to have ~10-fold discrimination against mC and >300-fold discrimination against hmC based on our detection limits.

We utilized hAID bearing a truncation of the C-terminal region of (hAID-ΔC), which is associated with hyperactive deamination (~3-fold) without impacting sequence-dependent targeting35. We reasoned that low-level deamination might be easier to detect with this hyperactive variant and employed it to evaluate the enzyme-dependence of deamination against the S30-TGX substrate series. As with mAPOBEC1, a similar discrimination against mC (~16-fold) and significant discrimination against hmC (>150-fold) was found with hAID-ΔC [Figure 33B]. These results are not an artifact of studying the hyperactive enzyme, as a quantitatively similar pattern can be observed with full-length hAID [Figure 33C]. Finally, mAPOBEC3 also demonstrated clear discrimination against the naturally modified cytosine nucleobases [Figure 33D], despite its distinctive canonical role from APOBEC1 and AID. We conclude that this strong discrimination against 5-substituted cytosines is an intrinsic property of the entire enzyme family.
regardless of the source organism or canonical function.

Figure 33: Quantification of enzyme-dependent deamination of C, mC, and hmC-containing substrates. Total amount of deaminated substrates is plotted as a function of increasing concentrations of respective deaminases assayed against substrates (250 nM) containing a cytosine (red), mC (blue) or hmC (yellow). (A) mAPOBEC1 was incubated with the S30-ATX series of substrates for 15 minutes. (B) hAID-ΔC, (C) full-length hAID, and (D) mAPOBEC3 were incubated with the S30-TGX series of substrates for 12 hours. Substrates were assayed for deamination as described above (C:UDG; mC:TDG; hmC:SMUG). To account for incomplete cleavage of each product (U/T/hmU) by the respective glycosylase (UDG/TDG/SMUG), product formation was determined by quantification against a standard curve generated for each glycosylase [Figure 32]. Error bars represent standard deviation from at least three independent replicates. For relative comparison of substrates, the slope of each plot in the region where product formation is linear with enzyme (dashed line) is listed. The values given for hmC substrates are the lower limits of detection with these substrates.

4.2.2 Deamination Decreases with Steric Bulk at C5 Position

The molecular basis for recognition of cytosine by AID/APOBEC enzymes has been a matter of speculation given the lack of structural information on these enzymes complexed with nucleic acid substrates. To probe the molecular impact of substitution at the 5-position of cytosine, we synthesized additional substrates with unnatural 5-substituents of varied steric and
electronic character [Figure 34A] using a sequence context appropriate for focusing further on hAID-ΔC (S30-TGX) or mAPOBEC1 (S30-ATX). One potential determinant of reactivity in this series is the electron withdrawing ability of the C5-substituents. Electronegative C5 groups could potentially enhance deamination by making C4 of cytosine more electrophilic or by lowering the pK\textsubscript{a} of N3. Alternatively, hydrophobicity of the 5-position substituent could influence selectivity (188). Finally, the size of the substituent could dominate the rate effect with a ranked steric order of 5-H < 5-F < 5-OH < 5-methyl ~ 5-Br < 5-I ~ 5-hydroxymethyl.

Figure 34: Substrate and assay design for cytosines containing additional 5-substitutions. (A) Substrate design. Full complement of 5-substituted cytosines. (B) Screening of DNA glycosylases against natural and unnatural C5-modified cytosine and uracil. Shown are the reaction products for each S30-TGC/S30-TGY substrate screened against each glycosylase and the no-enzyme control. Qualitative summary of glycosylase activity on respective DNA substrates is also shown. Red indicates that a glycosylase is specific and active for the corresponding C5-modified uracil, but not the cytosine analog, making it potentially useful for
assays. The glycosylase selected for discriminating each modified cytosine bases from each modified uracil base is noted by bold outline.

In order to study deamination of these unnatural cytosines, we first analyzed the activity of the coupling enzymes UDG, TDG and SMUG against DNA substrates containing either the modified cytosine or the corresponding 5-substituted uracil deamination product. Based on the determined substrate preferences, we selected to use UDG for the 5-fluoro substrates and SMUG for the 5-hydroxy, 5-bromo and 5-iodo substrates [Figure 34B]. The data corroborate the known studies on UDG, SMUG and TDG. In particular, UDG uses a steric discrimination mechanism that allows for excision of unmodified uracil and (5F)U, with some activity against (5OH)U. There is discrimination against cytosine bases through selective recognition of the uracil N2/C3 region through hydrogen bonding. By contrast, TDG excises bases in part through a mechanism that involves destabilization of the N-glycosidic bond. Thus, some excision activity is also seen with cytosine containing bases. SMUG has been previously shown to interact with uracil bases that have hydrogen bond donors at the 5-position.

As was done for C, mC and hmC containing oligonucleotides, we constructed standard curves to allow for accurate quantification of deamination of each unnatural 5-modified substrate using our glycosylase-coupled assay [Figure 35].
Figure 35: Standard curves show that DNA glycosylases are sensitive for detection of deamination of unnatural 5-modified cytosine substrates. The calculated amount of substrate to product DNA is plotted against the actual amount. The mean and standard deviation from 3-4 independent replicates are shown. These analyses confirm that UDG and SMUG are useful for detection of low levels of deamination products and allow for quantification of the true amount of deaminated product present based on the amount detected using the appropriate glycosylase.

Each 5-substituted substrate was incubated with hAID-ΔC, then duplexed and treated with the appropriate glycosylase to assay for deamination. For hAID-ΔC, any substitution resulted in decreased efficiency of deamination relative to unmodified cytosine [Figure 36A]. To allow for relative comparison of substrates, we next calculated the product formation under conditions where deamination was linearly proportional to enzyme concentration [Figure 37A, C]. Across these series of substrates, representing a >150-fold difference in reactivity, the size of the substituent at the 5-position appeared to be an important determinant of deamination. The smallest unnatural substituent, S30-TG(5F)C, was deaminated most readily, although it remained...
half as reactive as unmodified cytosine. Bulker halogen substituents were relatively poor substrates compared to the smaller mC. In addition to the influence of sterics, the poor hydrophobic character of hmC may play an additional role in the reactivity decrease seen between (5I)C, which has detectable deamination, and hmC, which has no detectable deamination.

**Figure 26:** DNA deamination decreases as a function of increasing steric bulk at the 5-position of cytosine. (A) Incubation of hAID-ΔC with S30-TGX substrate series for 12 hours. Substrates are ordered by increasing steric bulk of the 5-substituent. C and (5F)C substrates were assayed for deamination with UDG; mC substrates were assayed with TDG; and (5OH)C, (5Br)C, (5I)C and hmC were assayed with SMUG. The selection of the glycosylase used was based upon the survey of glycosylases [Figure 34]. (B) Incubation of mAPOBEC1 with S30-ATX substrates for 15 minutes. (C) Incubation of mAPOBEC3 with S30-TGX substrates for 12 hours.
Figure 37: Quantification of enzyme-dependent deamination of 5-modified cytosines. Quantification of enzyme-dependent deamination with (A) hAID-ΔC and (B) mA1. The amount of deaminated substrates plotted was a function of increasing concentrations of enzyme for mAPOBEC1 and hAID-ΔC, assayed against C- (red), mC- (blue), hmC- (yellow), (5F)C (green), (5OH)C (purple), (5Br)C (orange), and (5I)C (purple) containing substrates. Product formation was determined by quantification against a standard curve generated for each glycosylase [Figure 35]. Associated error bars represent standard deviation from at least three independent replicates and when not visualized are smaller than the symbol denoting the mean value. The data for C-, mC- and hmC- are the same as shown Figure 33, reproduced here to allow for comparisons to be made between substrates. (C) Shown are the electrostatic potential maps of each of the modified cytosine bases as determined using the SPARTAN program (6-31G* basis set). The electrostatic potential is colored from maximal negative (red) to positive (blue). The volume (*) is determined based on linking the 5-position substituent to a single hydrogen atom and calculating the total volume. The hydrophobic substituent constant (**) is derived from partitioning studies of substituted benzenes between octanol and water, where negative values for hydroxyl and hydroxymethyl substituents represent less hydrophobicity (Hansch C et al., J Med Chem, 1973, 16:1207-16). Reported are the values for enzyme dependent product formation (nM product/µM enzyme) for each substrate examined with each active deaminases as reported in Figure 33 (†). The relative activity for each modified substrate when compared to unmodified cytosine for each deaminase is reported parenthetically, allowing for comparisons across a row with each enzyme.

Table:

<table>
<thead>
<tr>
<th>Nucleobase</th>
<th>C</th>
<th>(5F)C</th>
<th>(5OH)C</th>
<th>mC</th>
<th>(5Br)C</th>
<th>(5I)C</th>
<th>hmC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (Å³)</td>
<td>10.5</td>
<td>16.8</td>
<td>19.4</td>
<td>30.5</td>
<td>33.0</td>
<td>37.5</td>
<td>40.8</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>0.00</td>
<td>0.14</td>
<td>-0.67</td>
<td>0.56</td>
<td>0.86</td>
<td>1.12</td>
<td>-1.03</td>
</tr>
<tr>
<td>hAID-ΔC</td>
<td>35.8 (1.0)</td>
<td>-</td>
<td>-</td>
<td>4.4 (0.12)</td>
<td>-</td>
<td>-</td>
<td>&lt;1 (&lt;0.04)</td>
</tr>
<tr>
<td>hAID-ΔC†</td>
<td>99.5 (1.0)</td>
<td>50.6 (0.51)</td>
<td>15.1 (0.15)</td>
<td>5.9 (0.06)</td>
<td>9.7 (0.10)</td>
<td>7.6 (0.08)</td>
<td>&lt;0.6 (&lt;0.006)</td>
</tr>
<tr>
<td>mAPOBEC1†</td>
<td>67.8 (1.0)</td>
<td>17.8 (0.26)</td>
<td>0.3 (0.005)</td>
<td>7.6 (0.11)</td>
<td>1.0 (0.015)</td>
<td>1.0 (0.015)</td>
<td>&lt;0.2 (&lt;0.003)</td>
</tr>
<tr>
<td>mAPOBEC3†</td>
<td>37.6 (1.0)</td>
<td>-</td>
<td>-</td>
<td>6.8 (0.18)</td>
<td>-</td>
<td>-</td>
<td>&lt;0.3 (&lt;0.008)</td>
</tr>
</tbody>
</table>
To assess the generality of reactivity determinants in the deaminase family, we additionally profiled mAPOBEC1 against a series of unnatural substrates in its preferred S30-ATX context [Figure 36B, 37B, C]. Again, 5-fluorocytosine was a good substrate for deamination, though approximately only one quarter as reactive as cytosine, while bulkier substituents were increasingly poor substrates. While smaller size remained a prerequisite for efficient deamination with mAPOBEC1, hydrophobicity also appears to play an important role. This is most strikingly notable with (5OH)C, which undergoes negligible deamination, while the larger mC is readily deaminated. An additional qualitative assessment was made with mAPOBEC3 and its preferred S30-TGC substrate series [Figure 36C]. mAPOBEC3 displayed the same selectivity, with a notably increased discrimination against the (5OH)C substrate similar to mAPOBEC1. Our extensive data show that for three distinct, active APOBEC family enzymes, efficient deamination of cytosine has steric requirements at the 5-position that contribute to the lack of any detectable enzymatic activity on hmC.

4.2.3 Deaminases Do Not Perturb Levels of Modified Epigenetic Bases in Genomic DNA

Among the multiple potential pathways for DNA demethylation, the possibility of collaboration between oxidation and deamination to generate hmU from mC has been postulated (38). To complement our in vitro findings, we addressed whether deamination of hmC could be detected in genomic DNA. Accordingly, we overexpressed one isoform of the TET oxidase family, TET2, in HEK 293T cells, which the Zhang lab has previously demonstrated generates a high prevalence of genomic hmC in a genome that otherwise does not have detectable oxidized cytosine bases (14). We then evaluated the impact of expression of individual AID/APOBEC enzymes on the levels of various modified nucleobases. The additional candidate enzymes in demethylation, TDG and the individual AID/APOBEC family members, were each cloned and co-transfected into HEK 293T cells along with TET2. Similar levels of expression for each enzyme were confirmed by Western blot [Figure 38].
Figure 38: Protein expression of AID/APOBECs and TDG in HEK 293T cells. (A) Western blot showing co-expression of FLAG-tagged TET2, TDG, and AID/APOBEC1-3 in 293T cells. Top panel: FLAG-TET2 (212 kD) detected by the FLAG antibody. Middle panel: FLAG-TDG, AID, APOBEC1, APOBEC2 or APOBEC3 detected by the FLAG antibody in the same samples as the top panel. Bottom panel: α-tubulin levels as a loading control. (B) Western blot showing co-expression of FLAG-tagged TET2 and untagged wild-type and mutant (E58A) hAID in HEK 293T cells. Top panel: FLAG-TET2 detected by the FLAG antibody. Middle panel: wild-type and mutant hAID detected by the AID antibody in the same samples as the top panel. Bottom panel: α-tubulin levels serve as a loading control.

Using methodology previously employed by the Zhang group to detect the products of iterative oxidation by TET (14), the genomic DNA was isolated and digested to generate a genomic nucleoside pool. For highly sensitive detection of the modified bases, the obtained nucleosides were subjected to LC-MS/MS in multiple reaction-monitoring mode and quantified by comparison to known standards [Figure 39].
Figure 39: Effects of hAID expression on genomic levels of hmU, fC, caC in HEK 293T cells. Mass spectrometry traces are shown for hmU, fC, and caC along with 10 or 50 fmol standards. In the fC trace, peaks at 8.9 minutes and 10.9 minutes share similar mass but do not co-elute with authentic fC. In the hmU trace, 10 fmol produces a signal that is ~3 times above background. Compared with standards, the amount of hmU in 0.75 µg genomic DNA from cells overexpressing TET2 and hAID is below 10 fmol, while fC and caC are detected at levels above 50 fmol.

As has been previously observed, TET2 overexpression in isolation leads to hmC levels that are about 1/6 that of genomic mC [Figure 40, blue bars]. The products of iterative oxidation, fC and caC, can also be detected at about 1/100 the level of mC. We posited that if AID/APOBEC enzymes could deaminate hmC, three changes should be observed: hmU should be detected in genomic DNA, hmC levels should decrease, and fC/caC levels should also decrease due to the deamination of their precursor hmC. However, none of these three predicted changes were observed when TET2 was overexpressed along with either mAPOBEC1 or hAID-ΔC [Figure 40, green and black bars]. hmU was detectable in no conditions at all, despite the ready detection of an internal synthetic standard [Figure 39]. To further examine whether other family members might influence the genomic levels of modified bases, we screened all mouse AID/APOBEC enzymes and found no changes in the modified cytosine pools [Figure 40].

By contrast to our results with overexpression of AID/APOBEC family members, we reasoned that if iterative oxidation coupled to BER is a feasible pathway for demethylation, levels
of modified cytosine bases should be readily perturbed by co-expression of TET and TDG. Indeed, we found that TDG overexpression led to dramatic reductions in the highly oxidized fC and caC species [Figure 40, magenta bars]. Despite depleting fC and caC, no overall change was observed in the levels of mC and hmC when TDG was overexpressed, likely as the highly oxidized species are far less prevalent that mC and hmC. This observation additionally suggests that TDG-mediated depletion of fC and caC does not promote further oxidation and consumption of genomic hmC. Potential explanations include the possibility that some genomic hmC is sheltered from further oxidation or plays roles independent of demethylation (65). While the dynamics of iterative oxidation will require further intensive study, from our data we conclude that the prevalence of modified cytosine nucleobases in the genome can be altered by overexpression of the players in the deamination-independent pathway, but not by those in the deamination-dependent pathway.

**Figure 40: Deamination intermediates of oxidized cytosine are not detected in genomic DNA.** Genomic DNA was extracted from HEK 293T cells co-expressing TET2 and TDG, AID/APOBEC1-3, wild-type or mutant hAID. Total amounts of mC, hmC, fC, caC, and hmU are graphed in fmol, as grouped by individual base. Asterisks: \( p \leq 10^{-3} \) for fC and caC in samples with TDG in comparison to plasmid only control.
The inability to detect hmU required additional controls to confirm that hmC was not being deaminated and then rapidly excised from the genome. First, we determined whether our inability to detect hmU could be due to its rapid enzymatic excision from the genome. To test this question, we isolated nuclear extracts from 293T cells and tested the hmU excision activity of the lysates against uracil-containing DNA substrates. We observed negligible hmU excision activity under conditions where robust nuclear uracil glycosylase activity was observed [Figure 41]. The results demonstrate that hmU glycosylase activity in HEK 293T cells is limited and quantitatively at least 400-fold less than uracil excision activity. Together these data suggest that if deamination of hmC occurs, the deaminated lesion is not being rapidly excised from the genome.
Figure 41: Uracil and not hmU is rapidly excised by HEK 293T cell nuclear extracts. (A) To analyze the endogenous glycosylase activity for uracil and hmU in nuclear lysates, duplexed S30-TGY (Y = U or hmU) substrates were incubated in the presence or absence of nuclear lysates for up to 45 minutes before reactions were quenched. (B) The reaction products were visualized on denaturing polyacrylamide gel and shown is a representative denaturing gel with time points. The gel demonstrates rapid uracil excision, with relatively less excision of hmU. (C) To quantitatively analyze product formation from the reactions in part B, the reactions were performed in triplicate and quantified as described for deamination reactions. Shown is the calculated amount of uracil or 5-hydroxymethyluracil bases excised per µg of nuclear lysate plotted as a function of time. Error bars represent standard deviation. The specific uracil glycosylase activity in HEK 293T nuclear lysates was calculated as the linear slope associated with data collected at < 1 min, given its rapid excision. The linear fit across all data collected with hmU yields the specific hmU glycosylase activity.
Next, to confirm that the overexpressed deaminases were active in the nucleus, we first examined nuclear lysates for deaminase activity using oligonucleotides containing unmodified cytosine. To determine whether AID and APOBEC1 are catalytically active in the nucleus of 293T cells, deaminase assays were performed with single stranded S30-TGC (for hAID-ΔC) and S30-ATC (for mAPOBEC1). If deamination occurs, uracil is excised by UDG in the nuclear extracts, leaving abasic sites to be cleaved by base treatment. TET2-AID nuclear extracts and negative controls (no extract, nuclear extract from untransfected 293T cells, and TET2-AID E58A) were incubated with nucleic acid substrates and analyzed by denaturing PAGE. Nuclear extracts were also incubated with a uracil-containing substrate to verify that 293T nuclear extracts have robust uracil excision activity. TET2-APOBEC1 nuclear extracts and negative controls (no extract and nuclear extract from untransfected 293T cells) were incubated S30-ATC or with the product control S30-ATU. In line with our biochemical studies, deamination of cytosine was only detectable in nuclear lysates from cells overexpressing either hAID-ΔC or mAPOBEC1 [Figure 42]. Importantly, this activity is dependent on catalysis, as no deaminase activity was seen in nuclear extracts from cells overexpressing a catalytic mutant of hAID (E58A).

Figure 42: Overexpressed hAID-ΔC and mAPOBEC1 are active in deamination in nuclear extracts. Nuclear extracts from transfected 293T cells were incubated with single stranded S30-TGC (for hAID-ΔC) and S30-ATC (for mAPOBEC1) for a total of 30 minutes. Endogenous UDG within lysates cleaves uracil and generates abasic sites for base treatment. The sample demonstrating deamination of the S30-TGC in TET2-AID extracts is highlighted (red). The sample demonstrating deamination of the S30-ATC substrate in TET2-mAPOBEC1 extracts is highlighted (red).
As a second validation of deaminase activity in the cellular setting, we examined whether overexpressed deaminases have an appreciable impact on genomic deoxyuridine (dU) levels. To overcome the rapid and efficient processing of genomic dU by multiple DNA repair pathways, we aimed to inhibit the major pathway involving UDG by overexpressing the small protein uracil DNA glycosylase inhibitor (UGI) (171) along with hAID, hAID-E58A, mAPOBEC1 or an empty plasmid control [Figure 43A]. The nuclear lysates associated with UGI overexpression were unable to excise uracil from duplexed oligonucleotides, demonstrating that UGI effectively inhibits one important pathway for uracil excision [Figure 43B]. We quantified the level of genomic dU using our highly sensitive LC-MS/MS methodology, and demonstrated a consistent increase in genomic dU in hAID samples relative to hAID-E58A or the empty plasmid control [Figure 43C]. Unlike hAID, it is unknown if mAPOBEC1 can act upon genomic cytosine. Compared to an empty vector control, we saw a trend towards increased genomic dU with mAPOBEC1 overexpression. Thus, by two measures, in lysates and through analysis of genomic DNA, the deaminases appear to be active on unmodified cytosine under conditions where no deamination of hmC is detectable.
Figure 43: Overexpressed hAID and mAPOBEC1 are active in deamination. (A) UGI expression does not alter DNA deaminase expression. Cell lysates were subjected to SDS-PAGE and western blotting to analyze expression of the proteins. Top panel: hAID and hAID(E58A) detected by α-AID antibody; FLAG-APOBEC1 detected by the α-FLAG antibody. Bottom panel: α-tubulin levels as a control for equal loading. (B) To confirm that UGI is active in the conditions used for mass spectrometry experiments and inhibits uracil excision, nuclear lysates from cells...
transfected with UGI-pIRES or an empty-vector control (pIRES) were incubated with uracil substrates. Unreacted substrates are denoted by S30 and cleaved product by P15. (C) UGI and hAID or mAPOBEC1 co-expression result in an increase in genomic uracil. Genomic DNA was extracted from HEK 293T cells co-expressing pIRES control or UGI-pIRES along with an empty vector control, hAID, hAID(E58A), or mAPOBEC1. The absolute amount of dU nucleoside was normalized to the absolute amount of deoxycytosine. The mean values are plotted with error bars from two to four independent replicates shown with p-values from unpaired t-test shown.

4.3 Discussion

4.3.1 Re-evaluation of Deamination-Dependent Demethylation Pathways

We can now reconcile the enzymatic characteristics of the AID/APOBEC deaminases determined here with their proposed function in the early steps of proposed DNA demethylation pathway. This is best accomplished by appraising the multiple discrimination mechanisms utilized by these deaminases to target their DNA substrate. At the level of the nucleobase, these deaminase enzymes have all evolved an active site that is designed to deaminate unmodified cytosine preferentially. However, deamination of mC to T can occur, albeit at ~10-fold reduced rate relative to cytosine deamination. Thus, strictly from the perspective of biochemical feasibility, deamination of mC may constitute a viable pathway for demethylation in some situations, though other constraints are important to consider (see below). By contrast, we demonstrate that no deamination of hmC was detectable in vitro or in cells when relevant enzymes were overexpressed. Our results contrast with those of a prior study that used immunoblotting of DNA, a method of uncertain specificity, to report detection of genomic hmU (38). However, our findings are in good agreement with studies on embryonic stem cells, where hmU was not detectable in genomic DNA when probed with a sensitive and specific mass spectrometry methodology (60, 78). There is a recent published report of hmU, detected in both mouse and human tissues (189). However, the levels of this base were modestly affected by TET1 overexpression, unlike hmC, fC and caC, which all saw marked increases. If hmU truly is produced at low levels during TET1 expression, it is possible that it may be derived from off-target oxidation of thymine rather than hmC deamination. Together, our biochemical and cellular data provide a strong argument against the proposed collaboration between oxidation, deamination and BER as a pathway for DNA demethylation in mammalian cells.
At the next level beyond the target cytosine, the local sequence context provides an additional potential barrier to efficient deamination. Each AID/APOBEC family member is known to act at preferred trinucleotide hotspots (139), yet methylated CpG motifs can be found in all common sequence contexts, even those that may be disfavored by the individual deaminases. For AID, guanine is the least favored nucleotide downstream of the target cytosine, making the physiologically relevant CpG motif a non-ideal substrate for deamination, regardless of the 5'-substituent (140).

Since all known AID/APOBEC enzymes are specific for single-stranded DNA (16, 164), it remains unclear how methylated CpG motifs in genomic DNA might be sufficiently targeted by AID/APOBECs. Although transcription or replication could generate single-stranded DNA, active demethylation of CpG islands does not necessarily require replication or transcription (54, 55). Further, only AID and human A3A have been shown to deaminate mammalian host genomic DNA (100, 131, 134), and these mutations appear to be localized to expressed genes. Within Peyer’s patch B-cells, where AID expression is high, preferred deamination at the immunoglobulin locus occurs >25 times more frequently than the next most targeted locus, and unexpressed genes are not mutated by AID. Finally, expression of the various deaminase family members is often restricted to particular cell lineages or tissue types, suggesting that a single deaminase is unlikely to play a universal role in demethylation. For example, detectable expression of AID and APOBEC1 has been reported in stem cells, but not of APOBEC3, limiting the potential players in maintenance of pluripotency in these cells (183).

Integrating across these layers of targeting, from the nucleobase to the cellular level, enables us to assess the plausibility of the early steps in the various proposed DNA demethylation pathways. First, our findings show that deamination of mC remains a plausible demethylation pathway based on enzymatic function, although other known limitations in targeting would seem to present several barriers to its efficient function. It is possible that deamination of mC operates in non-physiological systems, such as heterokaryon-based reprogramming (185). To this point, two separate groups have reported a potential role for AID in
the reprogramming of induced pluripotent stem cells (190, 191). However, these studies rely
heavily on fibroblasts from AID knockout mice. In the mouse genome, AID is located in close
proximity to Nanog and other factors associated with pluripotency (192). Therefore, it cannot be
excluded that knocking out AID may also remove an enhancer, noncoding RNA, or some other
genomic element that may contribute to expression of nearby pluripotency genes.

In other physiological settings, it will be important to examine whether unidentified post-
translational modifications or cellular factors can facilitate deaminase targeting to CpGs in specific
methylated promoters, allowing these enzymes to play a role in particular niches. Although our
findings exclude hmC deamination as a detectable enzymatic activity, a requirement for
AID/APOBEC enzymes in conversion of hmC to C in neurons has been previously suggested
(38). Notably, in that study, AID/APOBEC enzymes were not shown to be directly capable of hmC
deamination, and a requirement for catalytically active AID in conversion of hmC to C was also
not established. If catalysis were to be required, it is feasible that AID/APOBEC-mediated
deamination of C or mC indirectly stimulates TET or DNA damage response pathways to promote
excision of genomic hmC. Our biochemical and cellular data make it unlikely that deamination of
hmC by AID/APOBEC enzymes is involved in DNA demethylation.

4.3.2 Re-evaluation of Potential Pathways for DNA Demethylation

Our biochemical results allow us to clarify the proposed pathways of DNA demethylation.
With regards to the deamination-dependent pathways, we conclude that oxidation-coupled
deamination is not a viable pathway towards the regeneration of demethylated cytosine. We
arrive at this conclusion from the lack of deaminase reactivity against hmC, assessed in multiple
contexts. Deamination-dependent demethylation remains a viable pathway when mC is the initial
substrate. However, the reduced rate of deamination when compared to unmodified cytosine
presents an intrinsic kinetic barrier that limits its demethylation potential.

In raising doubtful concerns regarding deamination-dependent demethylation, our
findings shift the focus in favor of oxidation-dependent, deamination independent DNA
demethylation. Under the very same conditions of cellular overexpression where the product of
hmC deamination (hmU) is not observed, the proposed intermediates in the iterative-oxidation pathway, fC and caC, were readily detected. There are many important biochemical questions that currently remain unanswered regarding the mechanisms of action of the TET family. First and foremost is the urgent question of the relative rates of oxidation of mC, hmC, and fC. As hmC is the most abundant oxidized cytosine base, it remains to be determined whether iterative oxidation is a rate-limiting step in DNA demethylation. Given that the kinetics of fC and caC excision by TDG have already been described, determination of the rates of oxidation will illuminate whether production of oxidized bases may outpace their removal from the genome. Further questions, such as sequence specific targeting of DNA, will inform biological studies of the TET enzymes, and exploration of the oxidation-dependent pathway will continue to demystify the mechanisms of DNA demethylation.
CHAPTER 5: Conclusions and Future Directions

The studies within this thesis have sought to illuminate aspects of the deamination reaction of AID and its fellow APOBEC homologs for the purpose of clarifying their contributions to biology. Two particular areas have been explored: the selectivity for DNA over RNA, and the tolerance of modified cytosine substrates implicated in DNA demethylation. We have found that the preferred substrate for AID is unmodified, 2’-deoxycytosine. Modification at the 5-position of the base creates steric bulk that disfavors deamination, while modification at the 2’-position of the sugar creates a disfavored sugar pucker. The implications for DNA demethylation and the DNA deamination model of antibody maturation have been discussed in previous sections and will not be revisited here. Instead, the focus of this discussion will shift to a consolidation of our biochemical observations into a unifying model of substrate recognition and open questions for further study of AID/APOBEC-mediated deamination.

5.1 Model for Substrate Recognition and Deamination by AID

5.1.1 Description of Model

Our findings of the nucleic acid determinants of deamination support the following model for substrate recognition by AID/APOBEC deaminases, consistent with current biochemical, structural, and single-molecule studies [Figure 44] (137, 140, 145, 146, 193). The model focuses on AID, as this deaminase served as the focus of the studies described within.

The process of deamination initiates when AID binds non-specifically to the substrate backbone with a high affinity. Subsequently, the enzyme scans along single-stranded patches of DNA and ceases scanning as it interrogates the local base content for the presence of hotspot targets. Permissive sugar puckers in a DNA-like conformation, dictated by the 2’-substituents at the -2, -1 and 0 positions within the hotspot, stabilize the enzyme-substrate complex through the facilitation of hydrogen bonding and base-stacking interactions. This step is consistent with our experiments on chimeric 2’-F-RNA substrates. We observed that removal of the 2’-F groups from the -2, -1 and target positions was required for rescuing AID’s deaminase activity against the otherwise disfavored substrate.
With the substrate firmly docked to the enzyme, conformational flexibility is required of the target cytosine so that it may twist out of register with neighboring nucleotides and enter the enzyme active site. This conformational flexibility is conferred by the ability of the target nucleotide sugar to interconvert between its two dominant conformational sugar puckers: C2'-endo and C3'-endo. The importance of this conformational flexibility is highlighted by our results regarding AID’s reactivity against chimeric DNA substrates. The DNA substrates that contained cytosine conformers with a strict C3'-endo sugar pucker were significantly disfavored for deamination—up to 500-fold worse than the ideal, fully DNA substrate. By contrast, the substrates containing cytosine conformers that preferred the C2'-endo conformation were readily deaminated, indicating the importance of this sugar pucker as a permissive conformation for deamination. Most notably, the most deamination was observed with the substrate containing 2'-deoxycytidine. While this conformer prefers the C2'-endo state, it is distinguished from the rest of the assayed conformers in that it shows the greatest ability to interconvert between the C2'-endo and C3'-endo states, as determined by NMR spectrometry. The optimal reactivity of substrates with 2'-deoxycytidine, contrasted with substrates containing other conformers, suggests the importance of nucleotide flexibility in AID-mediated deamination.
Figure 44: Mechanistic model for nucleic acid targeting and selectivity. Upon AID binding to its hotspot WRC target (W = A/T, R = A/G), a preferred C(2')-endo conformation from -2 to 0 facilitates productive active site interactions, which can be antagonized by a single C(3')-endo-promoting substituent at the target nucleotide.

Once cytosine enters the deaminase active site, proper fit of the base is necessary to ensure efficient deamination. Modification of cytosine at the 5-position of the base, such as methylation or hydroxymethylation—provides a steric clash with the boundaries of the active site, displacing the base and preventing is proper reactivity for deamination. Lacking any modifications at the 5-position, unmodified cytosine is able to enter the active site, free from steric interactions with the sides of the active site, and readily undergo deamination given its proper positioning. This aspect of the model is supported by our observation of AID’s reactivity with a series of substrates in which there was an inverse correlation between the size of the 5-substituent and reactivity for deamination.

Though the data in support of this model was primarily obtained with AID, additional experiments with other members of the APOBEC family give us the confidence to conclude that
the model likely applies across the family. With regards to base modifications, AID, APOBEC1, and APOBEC3 all demonstrated a preference for unmodified cytosine over 5-methylcytosine, with no detectable deamination of 5-hydroxymethylcytosine. Not only were the relative preferences for these bases conserved across the deaminases, but also all three enzymes showed a similar pattern of discrimination against the size of the modifying group at the 5-position. Regarding the importance of the nucleotide sugar pucker at the target cytosine, AID and APOBEC1 share an inherent preference for DNA over RNA. The conservation of the molecular determinants of the nucleotide sugar and the 5-position of the base suggests that the core aspects of the deamination reaction remain conserved across the entire AID/APOBEC deaminase family since their evolutionary divergence. Thus, our model for deamination is likely broadly applicable to other family members.

5.1.2 Remaining Questions Regarding Deamination Model

While our model of deamination clarifies certain aspects of the AID/APOBEC deamination reaction, there are two unexplained areas that will require further study. The first question is how sequence-specific deamination may emerge from non-specific nucleic acid binding. This question develops out of the observation that AID binds to its nucleic acid substrates with a high affinity (Kd ~10 nM) and no known sequence specificity, yet deamination occurs at ‘hotspots’ with a greater frequency that other ‘coldspots’. One natural hypothesis is that hotspot motifs have more favorable hydrogen bonding interactions with the hotspot targeting loop. Thus, during scanning, the enzyme would be more likely to pause at hotspot motifs, providing the cytosine nucleotide a greater likelihood of entering the enzyme active site for deamination.

There are several techniques that could be leveraged to answer this interesting question. The technique best suited to answering this question is single-molecule microscopy. This technique has already been employed collaboratively by the Goodman and Rueda laboratories to explore the scanning characteristics of A3G (193). Notably, this previous study demonstrated that A3G may spent up to 10 minutes bound to the same strand of DNA before releasing, while catalyzing only a handful of deaminations. This technique is largely enabled by advances in
protein labeling, namely the conjugation of fluorescent dyes to the enzyme that permit visualization of individual monomers by fluorescent microscopy. Specifically, the expression of AID as a fusion protein to a large scaffold such as MBP or GST allows the scaffold to be engineered without significantly affecting the overall protein stability. The design of substrate DNA with fluorescent probes strategically located in hotspots or coldspots would provide a method by which to investigate the scanning occupancy at the different motifs. Ultimately, this line of analysis could reveal the contributions of scanning to the overall deamination reaction.

This question of scanning opens up a larger question of tremendous significance: what is the structural basis for substrate recognition and deamination? Answering this question through structural characterization of a DNA-bound crystal structure could provide several important insights. First and foremost, this structural information would provide the high-resolution information needed to test our proposed model of deamination. We have done our best to describe the determinants of deamination using mechanistic biochemistry, but this is no match for a crystal structure that would describe determinants of the enzyme-substrate complex. Beyond the validation of our proposed model, structural information would aid in inhibitor design, with implications for exploration of the biology and kinetics of deamination.

To date, AID has been refractory to structural characterization despite efforts from several groups. While other members of the APOBEC family have been described using X-Ray crystallography and NMR spectroscopy, none of these structures have been described in the presence of a bound nucleic acid. The difficulties in achieving a successful structural characterization largely stem from the poor solubility of AID and other APOBEC family members. Additionally, the nucleic acid binding properties of these enzymes result in purifications of soluble aggregates that are quite heterogeneous. The many impurities in these solutions pose an additional problem towards achieving a structure. Thus, there are significant impediments that currently preclude a clear picture of what AID/APOBECs look like as they deaminate their substrates.
While AID and its fellow APOBECs have proved refractory to structural characterization, our results provide insights that may illuminate a potential experimental strategy. Disulfide cross-linking technology may be used to trap complexes of AID/APOBECs bound to DNA, generating an intermolecular adduct between enzyme and substrate (194). This technology, developed by Verdine and colleagues, would involve two unique modifications, one occurring on the substrate, the other occurring on the enzyme. The first modification involves the introduction of a thiol modification on the DNA substrate. The second modification involves engineering a strategic cysteine residue into the enzyme. If the engineered cysteine on the enzyme is in close proximity to the free thiol on the substrate, the two may form a disulfide cross-link that permits enrichment of a highly-specific enzyme-substrate complex.

A similar approach may be taken to trap a member of the AID/APOBEC family with a hotspot targeting motif engaged by a DNA substrate. Our results with the 2'-F RNA substrates demonstrated that removing the 2'-F substituents from the -2, -1, and 0 positions relative to the target cytosine could rescue deamination. This mechanistic observation suggests that sequence specificity of deamination may stem from hydrogen bonding and base stacking between the hotspot loop and these residues. A thiol group engineered into the -2 or -1 positions may be able to cross-link with a cysteine residue, strategically engineered into the hotspot targeting loop. Ongoing work in the Kohli lab using deep positional mutagenesis of the hotspot targeting loop has revealed amino acid positions that are amenable to cysteine substitution, preserving enzymatic activity. These results could be used to design a strategy for structural characterization of AID trapped to its nucleic acid substrate.

A potential experimental complication may arise from AID’s tight binding affinity to single-stranded DNA. With dissociation constants in the low nanomolar range for AID and most other APOBEC family members, non-specific interactions are likely to be enriched. For X-Ray crystallography, a highly pure, homogeneous crystal is required to generate a coherent diffraction pattern. In this regard, it may be necessary to turn to A3A, a unique member of the AID/APOBEC family for its weak binding affinity. Reports suggest that A3A has a dissociation constant between
10-50 µM (151, 152). This is approximately 1000 times worse than other members of the AID/APOBEC family, indicating very weak binding affinity for substrate and low likelihood of enriching non-specific binding interactions. However, the presence of an intermolecular disulfide cross-link between A3A and its nucleic acid substrates could result in enrichment of highly-specific interactions between enzyme and substrate, providing a method for obtainment of homogeneous crystals for further study.

One potential concern about the structural studies of A3A is the applicability to the larger AID/APOBEC family. A3A has established itself as somewhat of an outlier within the AID/APOBEC family in two regards. First, A3A demonstrates weak binding affinity towards its DNA substrates, as described above. This means that A3A may contact the phosphodiester backbone or individual bases in a unique manner that does not represent the mechanisms of interaction for its homologs. While such a finding would still be of interest, as it would cast light on the mechanisms of substrate recognition for A3A, but further studies would be needed to determine the applicability to AID and the rest of the APOBEC family. Additionally, A3A demonstrates a catalytic efficiency that is several orders of magnitude greater than its AID/APOBEC homologs. It is certainly possible that this uniquely increased catalytic activity is connected to decreased binding affinity, particularly if end-product release is a rate-limiting step in turnover for deamination. Nevertheless, hypotheses for future structure-function analysis of AID and the APOBEC family. While structural studies of A3A may not reveal universal aspects regarding AID/APOBEC cytosine deamination, they would provide a beneficial starting point for future studies on the rest of the enzyme family.

5.2 Future Biochemical Studies of Deamination

5.2.1 Further Studies of DNA Deamination with A3A

While our studies have demonstrated the general properties of deamination, quantitative rigor has been lacking. The basis of our quantitative measurements—the observation of a linear relationship between product formation and enzyme concentration at fixed, 12 hour deaminase incubations—does not provide traditional measures of catalytic efficiency or substrate affinity.
The robust catalytic activity of A3A makes it well-suited for proper kinetic assessment of the questions addressed within the body of this thesis. Previous biochemical characterization of the enzyme suggests that it is capable of multiple turnover events within a few minutes (151, 152). This contrasts with the conditions of our deaminase assays with other enzymes, in which single-turnover kinetics were observed with an excess of enzyme compared to substrate. A3A may support the collection of proper kinetic measures, including $k_{\text{cat}}$ and $K_M$, that will communicate a more standard reference of enzymatic activity and proper evaluation of relative preferences.

The greater catalytic efficiency of A3A may allow us to re-examine the substrates in this thesis that were disfavored for deamination. In examining the role of base or sugar modifications, AID’s poor deaminase activity served as an impediment with regards to obtaining accurate, quantitative preferences for each substrate. It was necessary to establish a lower limit of detection for the most poorly reactive species—hmC and D-rC—which was then compared to the modest deaminase activity against the ideal substrate in order to establish a relative preference. In all cases, the limited deaminase activity hampered our ability discern the true degree of discrimination against disfavored substrates, creating a need for a deaminase with increased activity. One aspect of our conclusions is the conservation of these determinants of deamination across the active members of the AID/APOBEC family, allowing us to predict with a high degree of certainty that A3A would also display the same pattern of preferences. Given A3A’s robust DNA deaminase activity, repeating our experimental analyses with this deaminase may enable us to achieve a more accurate estimate of selectivity against hmC and the D-rC substrates. Even if these substrates remain non-reactive, normalization of our lower limits of detection to an increased default rate of deamination would more accurate measures of selectivity and bolster the confidence of our findings.

5.2.2 Future Biochemical Studies with Chimeric Deaminases

A3A also provides a handle for further study of the various properties of the enzyme that confer deaminase activity and define the protein determinants of catalytic activity. This potential
is best illustrated by studies of the hotspot targeting loop that confers sequence-specific deamination. The role of this loop in deaminase targeting has been demonstrated by several groups using both cellular and biochemical studies of AID. These initial studies have shown that A3G sequence-specific deamination can be conferred to AID by engraftment of the A3G hotspot targeting loop (140, 145, 146). However, this chimeric enzyme remains poorly characterized, in large part due to the poor baseline deaminase activity of AID.

Initial studies of an AID-A3A chimera indicate a more active form of AID that may be more tractable towards biochemical characterization. With regards to the proposed roles of AID in DNA demethylation, the Bhagwat group devised a Kanamycin selection assay within *E. coli* to screen for deamination of methylated cytosines (195). They evaluated AID and A3A with this assay and found that AID was weakly reactive with mC and A3A strongly reactive. Then, to determine whether the A3A targeting hotspot loop could augment AID’s mC deaminase activity, they tested an AID-A3A Loop (AID-3AL) chimera and found increased levels of mC deamination, relative to the ordinary AID control. This result indicates the importance of the hotspot loop as a potential enzymatic determinant of selectivity for cytosine methylation.

However, much remains to be elucidated from this study, as the proper controls were not performed. Particularly, the study by Bhagwat and colleagues did not determine whether the increase in mC deaminase activity of the AID-3AL variant was due to non-selectively increased overall deaminase activity, decreased discrimination of mC relative to C, or a combination of the two. Had the authors desired to ask this question using their bacterial mutagenesis selection assays, they could have repeated their findings with a Rifampin mutagenesis assay, which interrogates levels of unmodified C deamination. This measure could be used to normalize the mC deaminase activities of each enzyme and determine whether AID and A3A share the same relative tolerance of mC and, if not, whether the AID-3AL variant behaves more like AID or A3A.

This question may now be easily assessed using the *in vitro* biochemical approaches utilized in chapter 4 of this thesis. The AID-3AL variant may now be purified, and assessment of
the discrimination profile against the full panel of 5-modified cytosines may inform the contribution of the hotspot loop to mC/hmC deamination selectivity.

A3A may also serve as a recipient for engraftment of the hotspot loop from AID and other APOBECs. The great advantage of this approach is that A3A has much greater enzymatic activity. Therefore, modifications that reduce overall deaminase may still be detected, given the significant baseline activity of A3A. The same cannot be said for AID, for which marginal enzymatic activity is only observed at a vast excess of enzyme relative to substrate. This situates A3A as an ideal enzyme for observing a larger number of potential chimeric variants.

Initial efforts in the lab have successfully yielded an A3A-A3G Loop variant (A3A-3GL) that retains robust deaminase activity (Data not shown). This variant has been screened against a panel of 5-substituted cytosines to reveal its preferences and determine whether the preferences of this variant align with the donor enzyme (A3A) or the engrafted loop (A3G). Tentative results indicate that selectivity segregates with the engrafted loop rather than the donor enzyme, suggesting that the hotspot loop may interface with part of the enzyme active site. This will be an ongoing area of investigation, providing further mechanistic insights into the determinants of the deaminase reactions in the absence of a nucleic acid-bound crystal structure.

Expanding beyond the hotspot targeting loop, an additional aspect of the AID/APOBEC family may serve as ideal targets for future chimera studies. Similar to the hotspot targeting loop, there is an additional loop that was identified as a candidate for interaction with nucleic acid substrates, heretofore referred to as the accessory loop (146). This loop was first identified based on the crystal structure of the catalytic domain of A3G. If the hotspot loop is predicted to interact with the -2 and -1 positions of target DNA, the accessory loop is predicted to interact with the +1 position. The accessory loop has not been extensively studied, but a report from the Bhagwat group demonstrated that engraftment of both the accessory loop and hotspot targeting loop from A3G into AID boosted A3G-specific deaminase targeting over chimeric containing just the hotspot targeting loop alone. Modeling studies with the recently-solved NMR structure of A3A
similarly predict an interaction between the DNA substrate and the accessory loop, and further study is merited (138).

5.3 Future Cellular Studies of Deamination

The most pressing questions regarding the AID/APOBEC family regard established and emerging roles in DNA damage and demethylation. With regards to DNA damage, the recent implication of A3B as a driver of mutagenesis in breast cancer adds to the established body of evidence regarding AID’s pro-mutagenic effects. The possibility remains that other APOBECs, such as A3A, may contribute to mutagenesis as well. With regards to DNA demethylation, it appears that AID may have a niche role in the reprogramming of induced pluripotent stem (iPS) cells (190, 191). Given these new biological roles for the AID/APOBEC family, outside the traditional scope of adaptive and innate immunity, cellular studies are needed to further evaluate the validity of these claims. Our biochemical insights can be leveraged to add to this body of knowledge and evaluate hypotheses in a cellular setting, testing the novel proposed functions of the family.

5.3.1 Evaluation of Proposed Role of AID in Reprogramming of iPS Cells

The proposed role of AID/APOBECs in DNA demethylation has been discussed in depth in chapter 4. Our results indicate an unlikely role for hmC deamination as a route to DNA demethylation given the lack of any reactivity across the family. However, the weak reactivity of AID and other APOBECs against mC indicates that this DNA demethylation pathway is possible, though subject to an intrinsic kinetic barrier. This information can be used to expand upon the observations of two independent groups that AID may play a niche role in accelerating reprogramming of iPS cells.

This biological observation does not come without due scrutiny. The conclusions of the two separate groups stem from the use of fibroblasts derived from AID knockout mice (190, 191). AID is located in a pluripotency locus in close proximity to Nanog and other genes that play an important role during reprogramming. While AID is the only gene removed in the knockout condition, one cannot exclude the possibility that important enhancers or non-coding RNAs that
regulate expression of the neighboring pluripotency genes were removed as well (192). This concern is partially addressed by rescue experiments in which exogenous delivery of a retroviral vector containing AID is able to rescue the reprogramming process, with a catalytically-inactive version of AID unable to rescue the phenotype.

As reprogramming is proposed to depend on mC deamination, we may further test this hypothesis by rescue with the AID-3AL variant. The Bhagwat group has already demonstrated that this chimera has increased mC deaminase activity relative to wild-type AID (195). Our preliminary results indicate that this is increased mC tolerance is conferred by the A3A hotspot targeting loop. One would predict this variant to achieve greater mC deaminase activity and confer a greater capacity for reprogramming. Rather than attempting to rescue with A3A alone, this variant is uniquely positioned to contribute this answer because the AID donor mediates potentially critical interactions with actively transcribed DNA via Spt5 and RPA. Rescued reprogramming with this variant would be indicative that mC deamination contributes to the DNA demethylation process and supportive of the role of AID in the niche of iPS.

There are a few additional controls that would be required if evidence pointed towards AID’s role in mC deamination in iPS cells. A fundamental difference between wild-type AID and the A3A hotspot loop chimera is the sequence specificity of deamination targeting. While AID normal prefers to deaminate C in the WR\textsubscript{C}Y hotspot, the A3L loop variant has a targeting that is skewed closer to that of A3A: Y\textsubscript{C}A. It is possible that this manipulated sequence targeting could have a profound effect on reprogramming, independent of augmented mC deaminase activity. To control for this possibility, and additional experimental condition will need to be evaluated using the AID-3GL construct. This chimera has a similar sequence targeting to that of A3A, but is predicted to have reduced tolerance of mC. This condition could serve as a control for altered sequence targeting of the AID-3AL chimera: if mC deamination truly is contributory to reprogramming, we would expect to see rescue from AID-3AL but not AID-3GL. However, it is possible that sequence-specific deamination is more important that overall mC deaminase activity, in which case wild-type AID would preferentially rescue reprogramming over both
chimeric variants.

These experiments would demonstrate how our biochemical insights could provide new tools for evaluating the novel biological roles for AID in DNA demethylation. While these studies would certainly require biochemical validation of the assumptions made above with regards to sequence targeting and mC tolerance, they provide an opportunity to use biochemical insights to investigate questions of importance to those studying AID, genomic stability and regenerative medicine.

5.3.2 Evaluation of Proposed Role of AID/APOBECs in Cancer

A pro-mutagenic role for APOBECs has been proposed for breast cancer (131, 196, 197). This originates from sequence analysis in breast cancer genomes that identified a significant burden of mutations at TCA and CCA motifs, the preferred sequence targets of many APOBEC3 family members. Further work from the Harris group has demonstrated a role for A3B in this mutagenesis through shRNA knockdown studies in primary cell lines and a combination of over-expression and shRNA knockdown studies in transformed breast cancer cell lines (131). While it does not provide the same compelling sense of certainty that comes with studies of endogenous systems, the ability to evaluate the pro-oncogenic potential of a deaminase through cellular over-expression provides a model in which chimeric variants can be used to assess the pro-oncogenic role of the deaminases.

5.4 Concluding Remarks

The work described in this thesis presents a starting point for the detailed elucidation of the AID/APOBEC deamination reaction. Here, we have demonstrated the important nucleic acid determinants of deamination. 2'-substitution to the nucleotide sugar disfavors cytosine for deamination, presumably through effects on nucleotide sugar pucker. This selectivity appears to be important at the target cytosine and neighboring nucleotides as well. 5-substitution of the cytosine base is disfavored for deamination, presumably through steric clash within the margins of the active site. Both of these nucleic acid selectivities are conserved across the entire AID/APOBEC family, indicated that the mechanism of deamination remains unchanged within the
family despite their different biological functions. These insights support a model of deamination that illuminates future mechanistic studies into the deamination reaction. Additionally, our biochemical insights may be used to complement cellular studies of AID/APOBEC function, particularly in regards to novel proposed roles outside of immunity.
BIBLIOGRAPHY


