Model-guided mutational analysis of Activation-induced cytidine deaminase (AID) to elucidate structure-function relationships

By Justin King

A thesis submitted to the School of graduate studies in partial fulfillment of the requirements for the degree of Ph.D., department of biomedical sciences, Faculty of Medicine, Memorial university of Newfoundland

Proposed graduation date: October, 2020

St. John's, Newfoundland and Labrador

Abstract

Activation-induced cytidine deaminase (AID) is a DNA mutating cytidine deaminase enzyme that is a member of the AID/ Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) family of cytidine deaminases, altogether consisting of 11 members in humans. AID deaminates deoxycytidine to deoxyuridine at immunoglobulin loci in activated B lymphocytes to mediate secondary antibody diversification. AID-mediated mutations have also been implicated in epigenetic reprogramming, as well as genome-wide DNA damage and chromosomal translocations that transform healthy B cells into lymphoma/leukemia. Since its discovery in 1999, a key element missing in the AID field has been the lack of a protein structure. The goal of this thesis was to solve this problem and determine how AID binds its substrate nucleic acid and how these interactions result in its physiological and aberrant activities. In chapter 2 of this thesis, we explored the role AID plays in epigenetic reprogramming through genome-demethylation. We provided a glimpse into the AID catalytic pocket structure and how AID from humans and other species target cytidine *vs* 5-methylcytidine. In chapter 3, we provided the first functional and breathing structure of AID using a unique computational-biochemical approach. In this work, we provided the first fine map of AID's structure, identified the structural details of its catalytic pocket and interactions with the DNA substrate including secondary catalytic residues that are key to cytidine stabilization in the pocket, delineated two novel substrate DNA binding grooves on the surface and studied how conformational changes in the AID structure regulate its function. The work culminated in our discovery and proof that dynamic catalytic pocket closure limits AID activity. This novel catalytic pocket state which we termed the "Schrödinger's CATalytic pocket" has since been confirmed by multiple other structural studies and represents a novel mode of enzyme activity regulation across all known nucleic acid modifying enzymes. Our work also provided a structural rationale for many of the biochemical properties of AID including its

unusually low catalytic rate and atypically high nanomolar range binding affinity for its substrate DNA. In Chapter 5, we utilized these structural insights and the combined computational-biochemical approach developed in Chapter 3 to screen a large library of small molecule drug-like compounds against the catalytic pocket of AID. We have identified and functionally tested the first small molecule inhibitors that block the mutagenic activity of AID. We also demonstrated that several of these small molecules can inhibit several of AID's tumorigenic APOBEC siblings. And finally in Chapter 6, we review several DNA/RNA binding grooves previously discovered on AID and explore their possible binding combinations. We explore how these plastic modes of binding DNA and RNA simultaneously may impact AID function. Finally, we examine proposed AID dimer models bound to DNA/RNA and examine how this may act as an additional regulatory bottleneck of enzymatic activity.

Acknowledgements

I would like to thank everyone who supported me during my PhD. First, I would like to thank my parents and my brother for their support over the years. Second, I would like to thank my committee members and other faculty members for their support, guidance and constructive criticism. Finally, I would like to thank my supervisor Dr. Larijani. Dr. Larijani cultivated my critical thinking skills and inspired my love for research. Beyond academics, Dr. Larijani was also my mentor and provided advice and guidance throughout the years and during my episodes of depression. Thank you all for helping me develop into the person I am today.

Table of contents

List of Figures

List of Tables

List of abbreviations

5-mC: 5-methylcytidine

A3A: APOBEC3A

A3G: APOBEC3G

AID: Activation induced cytidine deamianse

APE: Apyrimidinic endonuclease

APOBEC: Apolipoprotein B-mRNA editing enzyme catalytic polypeptide-like complex family of cytidine deaminases

BER: Base excision repair

BL: Burkitt's lymphoma

CLL: Chronic lymphocytic leukemia

CSR: Class-switch recombination

CTD: C-terminal domain

dA: Deoxyribose Adenosine

dC: Deoxyribose cytidine

DDBs: DNA-damage binding proteins

dG: Deoxyribose guanosine

DLCL: Diffuse large B-cell lymphoma

DNA: Deoxyribose nucleic acid

DNMT: DNA methyltransferase

Dr-AID: *Dani rerio* AID

DSBs: Double-stranded breaks in DNA

dT: Deoxyribose thymidine

dU: Deoxyribose uridine

FL: Follicular lymphoma

GC: Germinal center

HIGM: Hyper-IgM syndrome

HR: Homologous recombination

Hs-AID: *Homo sapien* AID

IDLs: Insertion-deletion loops

Ig: Immunoglobulin

K_d: Dissociation constant

L2: Loop 2

L4: Loop 4

L6: Loop 6

L8: Loop 8

MM: Multiple myeloma

MMR: Mismatch repair

NER: Nucleotide excision repair

NHEJ: Non-homologous end-joining

NLS: Nuclear localization sequence

NMR: Nuclear magnetic resonance

PDB: Protein databank

RAG: Recombination-activating genes

RPA: Replication protein A

SHM: Somatic hypermutation

SSBR: Single-stranded break repair

SSBs: Single-stranded breaks in DNA

ssDNA: Single-stranded DNA

ssRNA: Single-stranded RNA

TCR: T cell receptor

TDG: Thymidine DNA glycosylase

TSS: Transcription start site

UDG: Uracil DNA glycosylase

UNG: Uracil N-glycosylase

UV: Ultraviolet

Author contributions

Chapter 2: J. King and H. Abdouni designed the research proposal, performed the experiments, analysed the data and wrote the manuscript. J. King also constructed and analysed all simulation data. M. Suliman, M. Quinlan and H. Fifield assisted with experiments for the paper. M. Larijani was the principal investigator and contributed to the writing of the paper.

Chapter 3: J. King designed the research proposal, performed the experiments/simulations, analysed the data and wrote the manuscript. C. Manuel, C. Barrett, S. Raber, H. Lucas and P. Sutter all assisted with experiments for the paper. M. Larijani was the principal investigator and contributed to writing the paper.

Chapter 4: J. King designed the research proposal, performed the experiments/simulations, analysed the data and wrote the manuscript. M. Larijani was the principal investigator and contributed to the writing of the paper.

Chapter 5: J. King designed the research proposal, performed the experiments/simulations, analysed the data and wrote the manuscript. F. Forzooee, J. Im, M. Asgharpour, C. Diamond, H. Fifield and L. Berghuis all assisted with experiments for the paper. M. Larijani was the principal investigator and contributed to the writing of the paper.

Chapter 6: J. King designed the research proposal, performed the experiments/simulations, analysed the data and wrote the manuscript. M. Larijani was the principal investigator and contributed to the writing of the paper.

Thesis chapters

Chapter 2

King JJ*, Abdouni H*, Suliman M, Quinlan M, Fifield H , Larijani M (2013). Zebrafish AID is capable of deaminating methylated deoxycytidines. *Nucleic acids research*. 41(10)
*Denotes joint-first authorship.

Chapter 3

King JJ, Manuel CA, Barrett CV, Raber S, Lucas H, Sutter P and Larijani M (2015). Catalytic pocket inaccessibility of activation-induced cytidine deaminase is a safeguard against excessive mutagenic activity. *Structure*. 23(4):615-27

Chapter 4

King JJ and Larijani M (2017). A Novel Regulator of Activation-Induced Cytidine Deaminase/APOBECs in Immunity and Cancer: Schrödinger's CATalytic Pocket. *Frontiers in Immunology*. 8(26);351

Chapter 5

King JJ, Faezeh Borzooee, Junbum Im, Mahdi Asgharpour, Cody P. Diamond, Heather Fifield, Lesley Berghuis, Mani Larijani. First generation small molecule inhibitors targeting the catalytic pocket of activation-induced cytidine deaminase. *Manuscript in preparation.*

Chapter 6

King JJ and Mani Larijani. Structural plasticity of DNA/RNA binding by AID as a regulator of its activity. *Manuscript submitted to FEBS J.*

Additional peer-reviewed publications contributed to during Ph.D. dissertation

- Diamond CP, Im J, Button EA, Huebert DNG, **King JJ**, Borzooee F, Abdouni HS, Bacque L, McCarthy E, Fifield H, Berghuis LM, Larijani M (2019). AID, APOBEC3A and APOBEC3B efficiently deaminate deoxycytidines neighboring DNA damage induced by oxidation or alkylation. *Biochim Biophys Acta Gen Subj*. 63(11): 129415
- Holland SJ, Berghuis LM, **King JJ**, Iyer LM, Sikora K, Fifield H, Peter S, Quinlan EM, Sugahara F, Shingate P, Trancoso I, Iwanami N, Temereva E, Strohmeier C, Kuratani S, Venkatesh B, Evanno G, Aravind L, Schorpp M, Larijani M, Boehm T (2018). Expansions, diversification, and interindividual copy number variations of AID/APOBEC family cytidine deaminase genes in lampreys. *PNAS*. 115(14): E3211-E3220
- Yoshioka E, Chelakkot VS, Licursi M, Rutihinda SG, Som J, Derwish L, **King JJ**, Pongnopparat T, Mearow K, Larijani M, Dorward AM and Hirasawa K (2018). Enhancement of Cancer-Specific Protoporphyrin IX Fluorescence by Targeting Oncogenic Ras/MEK Pathway. *Theranostics.* 8(8): 2134-2146.
- Abdouni H, **King JJ**, Ghorbani A, Fifield H, Berghuis L and Larijani M (2018). DNA/RNA hybrid substrates modulate the catalytic activity of purified AID. *Mol Immunology*. 93:94-106.
- Quinlan E, **King JJ**, Amemiya CT, Hsu E and Larijani M (2017). Biochemical regulatory features of AID remain conserved from lamprey to humans. *Molecular and Cellular Biology*. 37(20): e00077-17.
- Kolypetri P, **King JJ**, Larijani M and Carayanniotis G (2015). Genes and environment as predisposing factors in autoimmunity: Acceleration of spontaneous thyroiditis by dietary iodine in NOD.H2(h4) mice. *Int Rev Immunol.* 34(6):542-56
- Abdelfatah N, McComiskey DA, Doucette L, Griffin A, Moore SJ, Negrijn C, Hodgkinson KA, **King JJ**, Larijani M, Houston J, Stanton SG, Young TL. (2013). Identification of a novel in-frame deletion in KCNQ4 (DFNA2A) and evidence of multiple phenocopies of unknown origin in a family with ADSNHL. *European journal of human genetics*: *EJHG*. 21(10)
- Dancyger AM, **King JJ**, Quinlan MJ, Fifield H, Tucker S, Saunders HL, Berru M, Magor BG ,Martin A, Larijani M. (2012). Differences in the enzymatic efficiency of human and bony fish AID are mediated by a single residue in the C terminus modulating single-stranded DNA binding. *FASEB J.* 26(4)

Chapter 1: introduction

Eukaryotic DNA damage and repair mechanisms

A fundamental principle within biology is the preservation of genomic material against the constant onslaught of DNA insults (1,2). Exogenous sources of DNA damage originate from multiple sources including ionizing radiation, air pollutants, heavy metal ingestion and chemotherapeutic drugs (3–11). Endogenous sources of DNA damage largely stem from DNA replication errors or the reactive oxygen species liberated during metabolism (2,12–14). From endogenous mutation rates alone, some estimates have predicted that humans can incur ~10^4-10^5 DNA lesions per cell per day (15). Thus, to effectively safeguard the genome against exogenous/endogenous DNA damage, eukaryotes have developed efficient and redundant DNA repair mechanisms (16,17).

The type of DNA damage repair depends heavily on the nature of the DNA lesion. DNA damage can range in severity from base modifications, to mutations that distort the double-helical nature of DNA, single-stranded breaks (SSBs) or double-stranded breaks (DSBs) (16,18,19). Nucleotide excision repair (NER) depends on DNA-damage binding proteins (DDBs) that sense DNA damage that distort the double helical structure of DNA (ie: cyclobutane pyrimidine dimers formed from UV radiation). Once recognized, additional NER-associated proteins are recruited for incision upstream and downstream of the lesion, repair via DNA polymerases and ligation by DNA ligases (16,20) (Figure 1). Base excision repair (BER) acts to repair base modifications primarily caused by deamination, oxidation and alkylation (ie: 8-oxo-dG caused by oxidation of dG) (16,21). BER initiates repair by recognition and excision of the

damaged base by a lesion-specific DNA glycosylase (ie: uracil DNA glycosylase), thus forming an abasic site. Unlike NER, an intermediate step in BER involves the formation of an SSB after

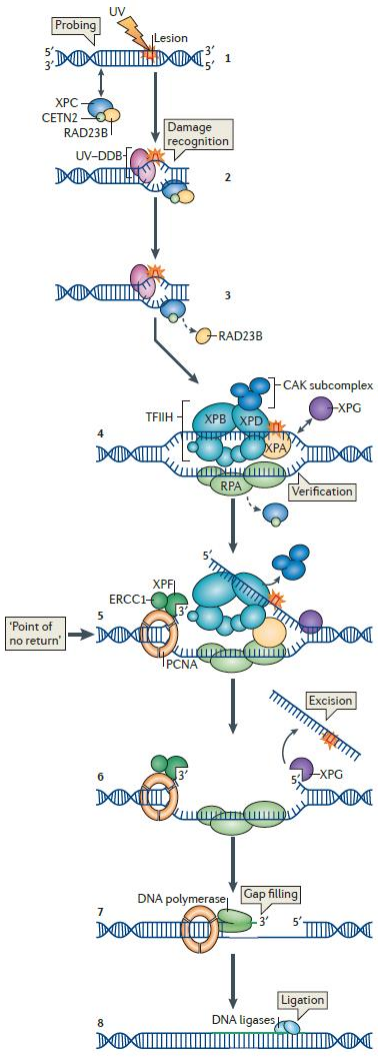


Figure 1: General overview of nucleotide excision repair (NER). In this particular case, the damaging agent is UV-radiation and XPC acts as the DNA-damage sensing factor (adapted with permission from Marteijn et al., 2014) (20).

incision by apurinic/apyrimidinic endonuclease. BER can repair the SSB with either short-patch BER, whereby only the missing nucleotide is repaired, or with long-patch BER which repairs 2-13 nucleotides surrounding the missing nucleotide. The remaining DNA gap is filled by DNA

polymerase and ligated with DNA ligase (16,21) (Figure 2). DNA SSBs represent the most common form of DNA damage with some estimates suggesting 75% of total endogenous DNA lesions (1,22). In addition to endogenous/exogenous sources of DNA damage described earlier,

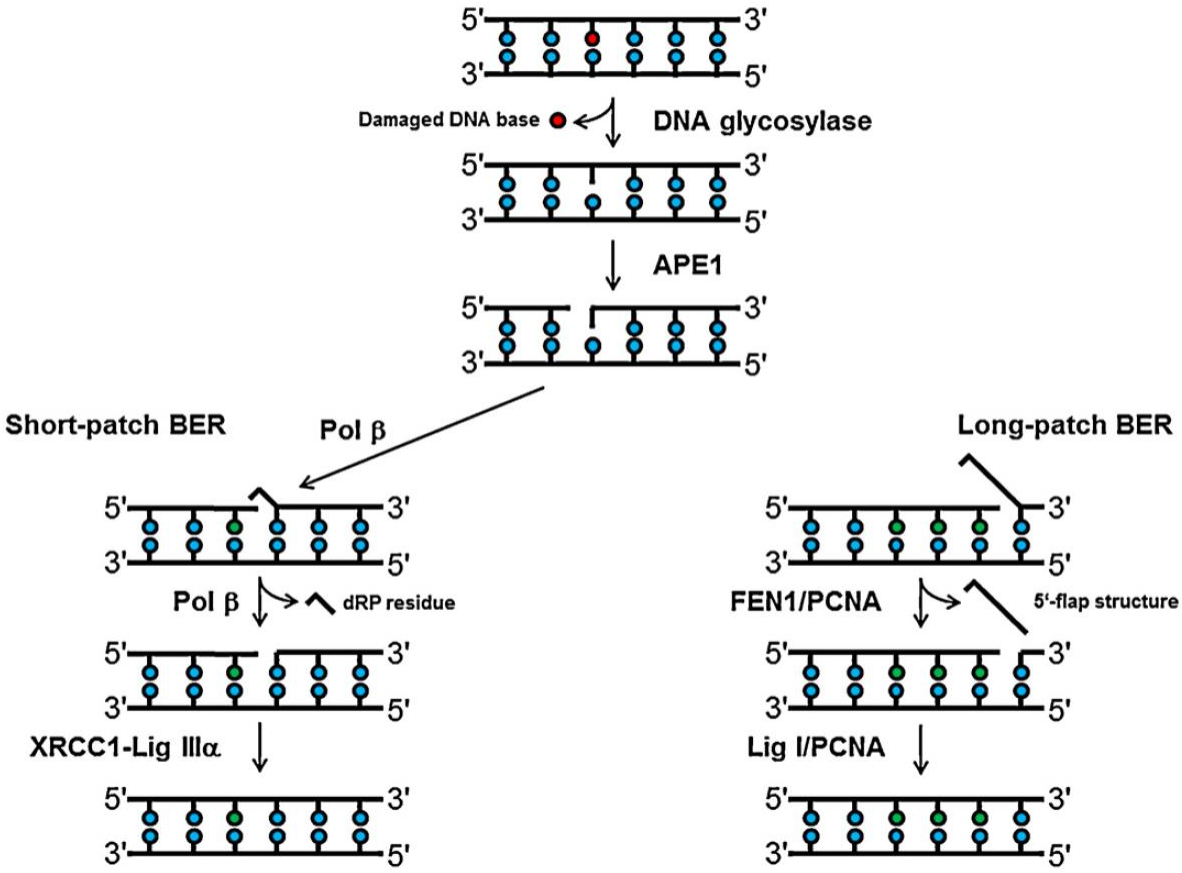


Figure 2: General overview of Base excision repair (BER). The type of DNA damage dictates the damage-specific DNA glycosylase recruited for BER (adapted with permission from Parsons et al., 2013) (21).

intermediates in certain enzymatic pathways and repair pathways such as BER involve intentional formation of SSBs (18,21). Single-stranded break repair (SSBR) efficiency depends heavily on detection by Poly ADP-ribose polymerase 1 (PARP-1). PARP-1 binds SSBs and acts to recruit downstream elements in SSBR and chromatin remodelling (18). SSB damaged ends are generally processed and repaired by DNA polymerases and associated proteins and ligated by

DNA ligase (18) (Figure 3). Mismatch repair (MMR) of DNA contributes to a 100-fold increase in DNA replication fidelity and deals primarily with miss-matched base pairs in DNA or insertion-deletion loops (IDLs) that generally arise after DNA replication (16,23). MMR in

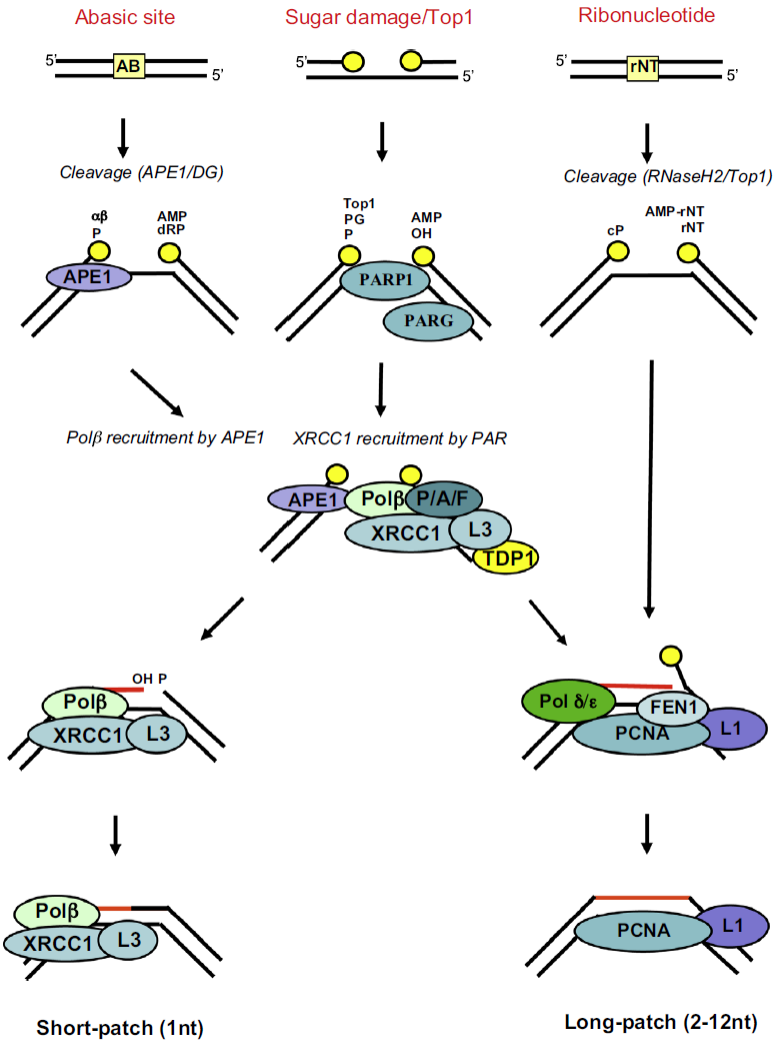


Figure 3: General overview of single-stranded break repair (SSBR). SSBs are detected depending by different SSBR repair factors, depending on the nature of the chemical moieties at DNA termini. Direct breaks (middle) are sensed via PARP1 (adapted with permission from Caldecott, 2014) (18).

humans directs repair primarily through MutSα (MSH2/MSH6 heterodimer) which recognizes base-pair mismatches and IDLs 1-2 bp in length, whereas larger IDLs of 2-10 bp long are

repaired mainly via MutSβ (MSH2/MSH3 heterodimer). In addition, MutLα (MLH1/PMS2 heterodimer) supports the repair of MutSα and MutSβ, Exo1 excises lesions 5' or 3' of the mismatches bp, replication protein A (RPA) stabilizes the gap ssDNA and PCNA in conjunction with polymerase δ/η aids in excision and repair (24) . Briefly, MSH2/MSH6 scans newly replicated DNA and once a mismatched bp is recognized MSH2/MSH6 clamps around the DNA and migrates 5'→3' from the mismatch until in reaches the PCNA DNA clamp. Exo1 then digests the daughter DNA until the mismatched bp is excised and repaired via polymerase δ or η (24) (Figure 4).

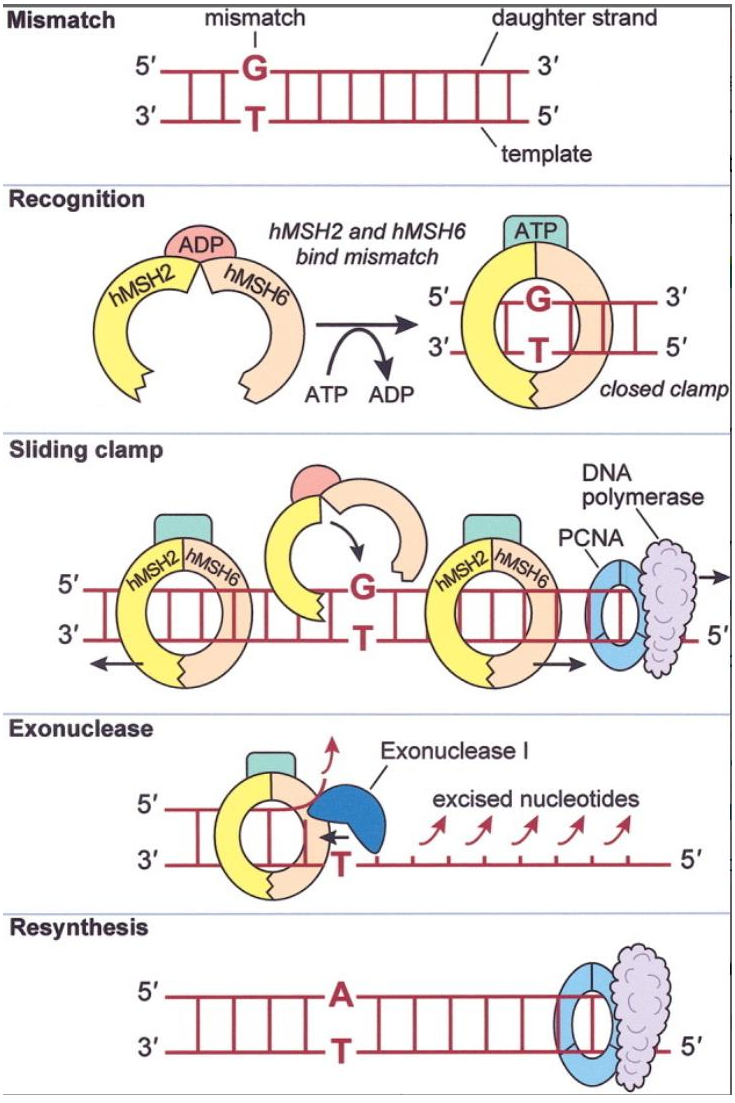


Figure 4: General overview of mismatch repair (MMR). The damage given in this example is due to polymerase error (adapted with permission from Jascur et al., 2006) (24).

Amongst all forms of DNA damage, double stranded breaks (DSBs) represent one of the most deleterious forms of damage that can elicit cell death responses and chromosomal translocations if left unrepaired (17,25). Like other forms of DNA damage, DSBs can arise from exogenous sources, such as chemotherapeutics or ionizing radiation, as well as endogenous sources such as replication or transcription (26). DSBs are repaired primarily through two pathways: Homologous recombination (HR) and Non-homologous end joining (NHEJ) (27,28) (Figure 5). HR predominantly repairs DSBs that arise during actively dividing cells and utilizes a sister chromatid as a template for repair (16,27). The MRE11-RAD50-Nijmegen (MRN) complex is recruited and binds the DSB, thereby recruiting ataxia telangiectasia mutated (ATM) and further downstream elements. The 5' ends of the DSB are resected, generating 3' overhangs of single-stranded DNA (ssDNA). The ssDNA is then bound by replication protein A (RPA), which works in conjunction with the HR machinery to replace the RPA with RAD51 and find a DNA sequence homologous to the 3' ssDNA overhang. Once a homologous DNA sequence has been identified, RAD51-coated ssDNA invades the homologous DNA in a process called strand invasion. Following strand invasion, the ssDNA is extended via DNA polymerase and ligated via DNA ligase (16,27). In contrast to HR, which operates during the S/G2 phase of the cell cycle, NHEJ is the major repair pathway in cells not undergoing active cell division and is most active during the G1 phase (16,28). NHEJ is initiated by binding of two DSBs by the Ku70/Ku80 complex, which then acts to recruit the DNA-dependent protein kinase catalytic subunits (DNA-PKcs). DNA-PKcs align the two DSB ends with one another and activate the endonuclease

Artemis, which cleaves the ssDNA overhangs until homologous DNA stretches are found. DNA polymerases fill in the gaps and the DSBs are ligated by DNA ligase (28). NHEJ is generally

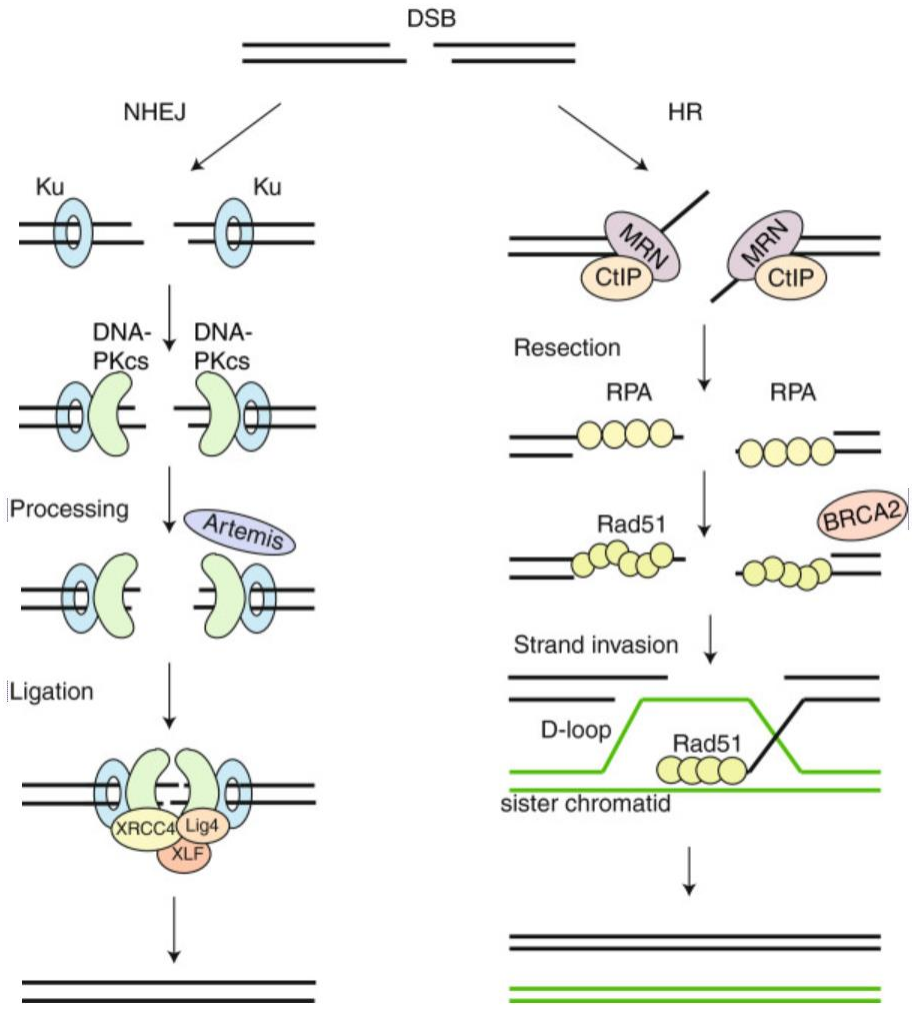


Figure 5: Non-homologous end-joining (NHEJ) and Homologous recombination (HR) mechanisms to repair DSBs in DNA (adapted with permission from Brandsma et al., 2012) (29). regarded as more error-prone than HR, due to the processing of DSBs that results in loss on genomic material at the site of the DSB (16).

Orchestrated DNA damage and repair in B and T lymphocytes: V(D)J recombination, Somatic hypermutation and class-switch recombination

Failure to handle the constant mutational burden imposed can lead to genomic instability resulting in apoptosis, senescence, oncogenesis and other diseases (2,12). Thus, genomic maintenance prevails as a top priority for the fitness of eukaryotic cells. However, in humans and other higher eukaryotes B and T lymphocytes defy this core directive and intentionally mutate and damage their own DNA to generate a diversified repertoire of antigen-specific receptors (30–32). Developing B and T lymphocytes undergo V(D)J recombination, whereby variable (V), diversity (D) and joining (J) gene segments at the Immunoglobulin (Ig) or T cell receptor (TCR) loci are rearranged to generate novel amino acid sequences of the antigen-binding portion of Igs and TCRs (30) (Figure 6). V(D)J recombination is initiated by recombination-activating genes (RAG) 1 and 2, which encode lymphocyte-specific enzymes which bind and instigate carefully orchestrated DSBs between V, D and J segments, while repair is facilitated via NHEJ (30). Thus, V(D)J recombination is responsible for development of TCRs and the initial naïve repertoire of antibodies; however, this initially developed repertoire of antibodies present on the surface of naïve B lymphocytes is generally of low affinity to any given antigen (33).

After exposure to an antigen, *Naïve* B cells expressing low affinity antibody receptors become activated (34). During B cell activation the immunoglobulin (Ig) loci are once again mutated and rearranged, followed by cellular selection for B cells with improved antibody affinities (31,32,35,36). This process known as secondary antibody diversification generates a B cell repertoire expressing higher affinity antibodies with different effector functions (36). The processes responsible for additional gene mutation and rearrangement during B cell activation are termed somatic hypermutation (SHM) and class switch recombination (CSR), respectively

(31,32,36) (Figure 7). These two reactions, both critical to adaptive immunity, are initiated by the DNA-mutating enzyme Activation-induced cytidine deaminase (AID) (37,38).

Activation induced cytidine deaminase: The sculptor of antibody memory

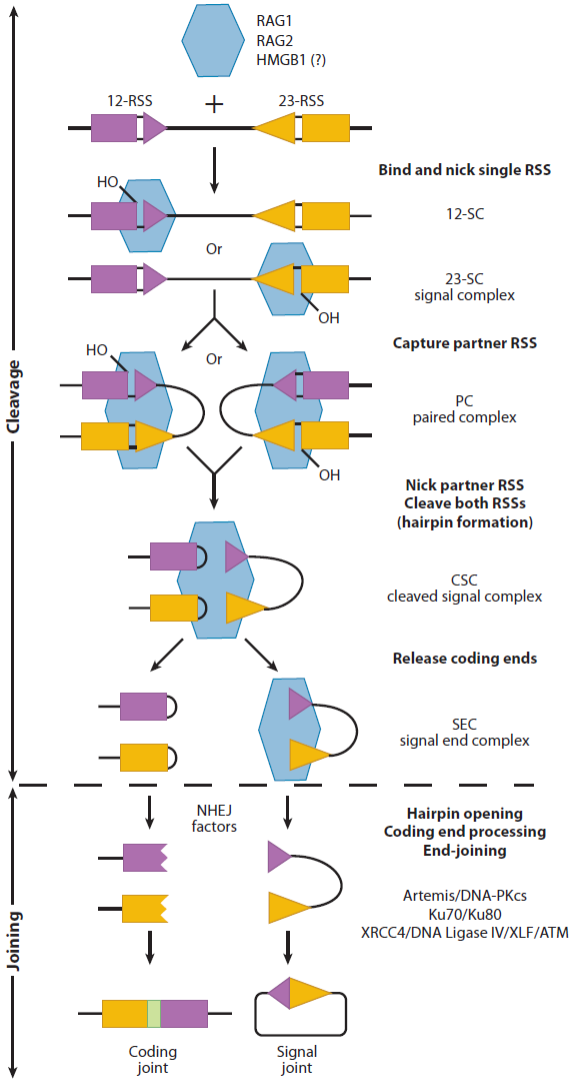


Figure 6: Overview of V(D)J recombination used to generate Ig and TCR variability (adapted with permission from Schatz et al., 2011) (30).

Human AID is a 198 amino acid (24 kDa) enzyme and is a member of the Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) family of cytidine deaminases (37). Two decades have passed since its discovery and AID's expression has been found to be largely restricted to activated B cells housed in unique structures called germinal centers in lymph nodes (37). These germinal centers are an anatomical hotspot for foreign antigen presentation (39) where initially activated B cells receive helping signals and direct cell-

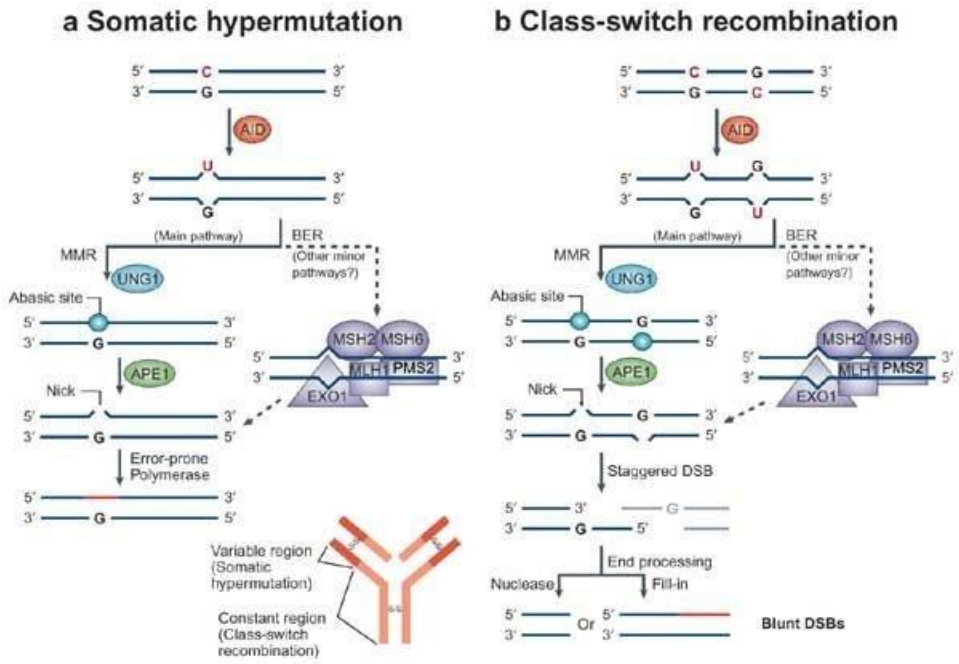


Figure 7: Overview of Somatic hypermutation (SHM) and Class-switch recombination (CSR). Both SHM and CSR are initiated by activation induced cytidine deaminase (AID) deamination (adapted with permission from Nussenzweig et al., 2004) (40).

cell interactions from other immune cells, including T lymphocytes, in order to proliferate and differentiate into Plasma and memory B cells (36,41). Plasma B cells are short-lived and secrete soluble forms of antibodies that circulate in the blood, while memory B cells can last for decades and produce membrane bound antibody on the surface of the plasma membrane (41). Successful B cell activation depends largely upon DNA mutation and rearrangement of antibody genes at immunoglobulin loci facilitated through SHM and CSR, respectively(32,38,42). These mutations

and rearrangements at the level of antibody genes change the protein makeup of the antibody expressed in that activated B cell, allowing for maturation of the early naïve low affinity antibodies into highly specific and high affinity antibodies.

Antibodies contain a variable region and a constant region that perform two different physiological tasks. The variable region binds to a specific antigen, while the constant region binds to different components of the immune system and receptors on the surface of many different immune cells to fulfill different immunological functions(43). During SHM, AID introduces point mutations into the portion of the variable region genes that encode the antigen recognition domain of the antibody molecule. Due to decreasing concentrations of antigen as the immune response proceeds, B cells expressing a lower affinity antibody than before will not receive strong enough signals to be selected and survive, whereas B cells in which these AID-mediated mutations have resulted in higher affinity antibodies will be at an advantage for selection and survival. This two part process of AID-mediated mutations followed by cellular selection of B cells is known as antibody affinity maturation(32). In addition to mutations leading to SHM, AID activity also leads to the generation of double stranded DNA breaks (DSBs) in the genes encoding the constant portion of the antibody molecule, leading to the excision of intervening regions and the replacement of an initial constant region gene segment by a downstream region. In this manner, AID mediates CSR which leads to the generation of antibodies with new effector regions such as IgG and IgA (38,44,45). Upon a second encounter with the antigen, the immune system can now mount a faster and more effective response due to the production of these higher affinity antibodies capable of different immunological function (35). Deficiencies in AID in human patients and mouse models result in immunodeficiency,

termed Hyper-IgM syndrome (HIGM), characterized by the inability to perform SHM and CSR (46,47).

Biochemical properties of AID

Past research from our laboratory and others have characterized the biochemistry and targeting specificity of AID at the molecular level. Purified AID deaminates deoxycytidine (dC) to deoxyuridine (dU) in ssDNA, with no catalytic activity detected on double-stranded DNA (dsDNA) or RNA (48–53). A wealth of biochemical studies have shown that although AID can bind any ssDNA, it appears to preferentially bind structures with topologically pronounced small ssDNA patches of 5-10 nt long, such as bubbles, stem-loops or ssDNA overhangs from G4-quadruplexes instead of relaxed linear ssDNA (49,54,55), but without regard to specific sequence. AID is thought to be catalytically active as monomers, dimers and tetramers (49,56,57). Once bound to its ssDNA target, AID can deaminate dC within the context of any ssDNA sequence, but activity is favored by 3-6 fold on WRC motifs (W=Adenine/Thymine, R=Adenine/Guanine, C= Cytidine (49). Two surface regions dictating WRC specificity have been previously identified; loop 7, termed the "hotspot specificity loop" primarily dictates specificity (58), whilst loop 1 plays a less dominant role in specificity (59). AID binds its target ssDNA with an unusually high affinity (nM-range) and once bound remains in a stable complex for 4-8 minutes (52). In contrast to its highly efficient binding of ssDNA, AID has an incredibly slow catalytic rate of 0.03 s^{-1}, mediating one dC deamination every 4 minutes, orders of magnitude slower than typical enzymes that catalyze thousands of reactions in the same time frame (60,61). We postulated that this high ssDNA binding affinity and lethargic catalytic rate had evolved to protect the genome from excessive and promiscuous AID targeting of non-Ig loci (52,55,61). In support of this notion, AID upmutants (mutants with higher catalytic activity)

showed increased genotoxicity in cells expressing them (62). AID also acts processively on its target ssDNA, in that when incubated with ssDNA many targets are unmutated but typically a few carry clustered AID mutations (50). Thus, AID remains bound to the same ssDNA substrate for several minutes and when it does dissociate it is likely to re-bind the same ssDNA again. This suggests that AID may act as a scaffold and stabilize/recruit additional repair proteins involved in SHM/CSR (63).

AID targeting in the genome

Although AID preferentially mutates the Ig loci, many studies have demonstrated that it can interact with and potentially mutate thousands of genes outside of the Ig heavy chain locus (64). Off-target AID activity can result in genomic instability and tumorigenesis (65–69). Prior to the discovery of AID, it was noticed that the initiation of both CSR and SHM was tightly associated with the initiation of active transcription through the switch and variable region genes, respectively (70–75). A "mutator factor", that would later be discovered to be AID (37), was then proposed to mediate SHM, which was suggested to mutate both transcribed and non-transcribed strands of variable region DNA in the same direction of the traversing RNA polymerase with a preference for RGYW (R =A/G, Y =C/T, W= A/T) hotspot motifs (74,74,76–78). Since the discovery of AID, the mechanism through which it is recruited to the Ig genes has been an area of interest. Since the initiation of CSR and SHM requires both transcription and AID, the major model of AID recruitment focuses on the role of transcription in bridging the gap between antibody diversity, AID and ssDNA substrate generation (79). AID recruitment is associated with genes undergoing high rates of transcription, partially explaining how AID can gain access to native Ig genes, oncogenes such as Myc and BCL6, as well as non-lymphoid loci (64,67,80). Although considerable advances have been made in understanding overall AID targeting and activity, the

precise features that lead AID to target one gene over another, leading to diversification or tumorigenesis remain elusive.

The process of transcription functions to provide AID opportunity to gain access to ssDNA substrate within an otherwise tightly organized, chromatinized, dsDNA genome. During transcription an orchestrated series of events occurs that creates a favorable environment for the targeting of a gene by AID. Nucleosome remodeling factors and histone chaperones work together to reorganize and strip nucleosomes prior to the passage of the RNA polymerase complex during elongation (81,82). The traversing RNA polymerase complex further destabilizes nucleosome complexes (83,84) and generates ssDNA patches by distorting the supercoiling of the transcribed DNA. In the absence of transcription, nucleosomes physically block AID, preventing it from reaching its target (85,86). Once transcription is initiated, AID may target both transcribed and non-transcribed strands (85,87,88). Once ssDNA is rendered accessible to AID through transcription, WRC (W =A/T, R =A/G) hotspot motifs located throughout the target gene are preferentially deaminated by AID. Although AID can bind any ssDNA, it appears to preferentially bind structures with topologically pronounced ssDNA patches, such as bubbles or stemloops, over relaxed linear ssDNA, but without regard to specific sequence. However, once bound to its ssDNA target, AID then exhibits a local sequence-dependent catalytic preference for deamination of WRC motifs (W=A/T, R=A/G) (49–53,89,90).

Another way that transcription may facilitate AID targeting is through direct association of AID with the RNA polymerase complex and/or protein co-factors. Early *in vitro* transcription studies focused on the possibility of AID targeting the non-transcribed strand within the transcription bubble while its sister DNA strand interacts with RNA (48,91). However, it is unlikely that AID would be able to physically access the non-transcribed strand within the small

~9-nt. transcription bubble, as the hefty elongation complex engrosses a larger estimated area of ~21-nt. around it (92). Additionally, the catalytic activity of AID, its binding to ssDNA structures or WRC specificity and many other regulatory features have been shown to be self-contained in AID itself rather than depend on the presence of any co-factor (49,52,89,90,93–96).

Transcription may also facilitate AID targeting through generation of secondary structures and topology that provides AID with transient ssDNA substrates. Transcription through G-rich CSR switch regions is known to promote formation of RNA-DNA hybrids known as R-loops (97–99). Chromatin remodeling and torsional strain induced by the traversing RNA polymerase complex allows the formation of secondary DNA structures such as stem-loops, cruciform (100) bubbles (101) and both negative and positive supercoiling upstream and downstream of the traversing RNAP respectively (83,84,102,103). These all have the potential to provide sufficiently stable ssDNA to attract AID, potentially aided by RNA polymerase pausing which encounters these altered DNA structures (104,105). Targeting of DNA structures and topology would also allow AID to mutate both the transcribed and non-transcribed strands, consistent with the observation that there is no strand preference for mutations in SHM (106). We and others have demonstrated that purified AID binds and exclusively mutates the ssDNA of numerous *in vitro* targets including: ssDNA, dsDNA, RNA, DNA-RNA hybrids, as well as DNA-DNA bubble structures (49–51,53,89,90,105). Moreover, AID prefers to mutate small DNA-DNA bubble structures that are 5- and 7-nt. in size over larger bubbles *in vitro*, aided by its high nanomolar affinity for ssDNA (49). Transient bubble structures 2-10-nt. in size can be transiently introduced into naked dsDNA through brief DNA fluctuations termed breathing (107), which AID may target.

Lastly, the role of sequence and structure cannot be viewed in isolation, as AID has a 3-10-fold greater preference for WRC hotspot motifs within DNA-DNA bubbles over completely ssDNA substrate (49,51). Non-optimal substrates (ie: Stemloops) containing a WRC trinucleotide motif are also mutated less efficiently than a non-WRC motif in an optimal structure (ie: 5-7nt. bubble) (49). Transcription-induced supercoiling spreads ~1.5-2kb upstream of the transcription start site (TSS) of all transcribed genes (84). During SHM AID-mediated mutations begin ~100-200 bp upstream of the V region promoter and span around 2 kb (108,109). The role of supercoiling in AID targeting and activity has been examined using an *in vitro* antibiotic resistance reversion assay in which supercoiled substrate, but not Topoisomerase I-relaxed plasmid, was efficiently targeted by AID in the absence of transcription (110). It is likely that supercoiling, and other DNA secondary structures, serve as a point of entry to an otherwise tightly organized genome.

DNA repair following AID-mediated lesions.

Although AID is a cytidine deaminase which mediates dC to dU conversion, SHM involves all possible transversion and transition mutations at any of the four nucleotides, and CSR involves physical DSBs. The link between the initial AID mediated dU lesion and the full range of these outcomes is the process of DNA repair downstream of AID. Once mutated, DNA repair pathways acting downstream of AID result in the translation of the initial dU:dG mismatch into the full spectrum of mutations of any of the 4 bases, into any of the other 4 bases (111). The MMR pathway can recognize the dU:dG mismatch and recruit downstream nucleases and error-prone polymerases, generating further mutations at surrounding bases (112,113).

Alternatively, the Uracil-N-Glycosylase (UNG) mediated mismatch repair pathway can be engaged to excise the uridine creating an abasic site, which can then lead to a transition or

transversion mutation after replication over the DNA lesion (114). Furthermore, the product produced after UNG-mediated uridine excision can act as a substrate for apyrimidinic endonuclease (APE) leading to a covalent nick in the DNA structure (16). This DNA mutation may be repaired by the BER pathway to eliminate the mutation imposed by AID; alternatively, this mutation can be further processed by NHEJ (28,114). If the initial lesion imposed by AID was in the switch region of an antibody gene, then subsequent processing by NHEJ can lead to CSR (45).

The mechanism through which AID-mediated dU repair at the Ig loci recruits more error-prone polymerases than high fidelity polymerases is poorly understood; however it is due to gap filling by error-prone polymerases that the initial dU lesion translates into the full myriad of transition and transversion mutations. Interestingly, DNA repair downstream of AID-mediated dU lesions at non-Ig loci is mediated by high fidelity polymerases (115). Repair-deficient mice and cell lines were shown to bear a much heavier genome-wide burden of AID-mediated mutations than their wildtype counterparts demonstrating that AID is constantly assaulting our genome but that the majority of these are caught and corrected by the DNA repair machinery (115).

AID mediates epigenetic modification through catalysis of 5-methylcytidine

Methylation of dC in CpG islands represents the most common epigenetic transcriptional silencing (116–122), with fully differentiated cells having nearly all CpG islands methylated. 5-methyldeoxycytidine (5-mC) can be passively demethylated through consecutive rounds of DNA replication in the absence of DNA methyltransferase (DNMT) enzymes that would otherwise maintain 5-mC in CpG islands (123). In contrast, active DNA demethylation occurs directly via

enzymatic mechanisms. Direct demethylation of 5-mC via the TET family of enzymes has been previously characterized in humans (124), although the extent to which function *in vivo* and their respective downstream pathways are poorly understood (125–132).

Alternatively, 5-mC or 5-hmC can be can be deaminated by AID, forming dT and 5-hmU, respectively, which are further processed by TDG or UDG (133–137). AID expression has been detected at low levels in pluripotent cells, while AID -/- primordial germ cells from mice have been shown to have hypermethylated genomes, analogous to terminally differentiated cells (134,135,138). The involvement of AID in mC demethylation may or may not be through direct deamination. AID may act as a scaffolding molecule, recruiting or stabilizing complexes with the ultimate function of demethylation. In support of this possibility, dozens of putative binding partners for AID have already been proposed, some of which are involved in DNA repair and transcription (55). Alternatively, several lines of evidence suggest that AID can directly deaminate mC. First, deamination of mC generates dG:dT mismatches (137). Second, MBD4, which recognizes G:T mismatches, was found to act synergistically with AID towards demethylation (137). Third, enzymatically inactive AID harboring a catalytic site point mutation failed to mediate hypomethylation (137). Fourth, AID promotes loss of DNA methylation of GC B-cells *in* vivo (139). However, AID mutation in humans and mice result in HIGM with no known developmental abnormalities (46,47). Additionally, 5-mC was shown to be a poor enzymatic target of AID in humans (51,93). Thus, genome demethylation in humans is likely orchestrated primarily by an alternate mechanism.

The role of AID in oncogenesis and tumor progression

Despite the physiological role that AID plays in secondary antibody diversification and genome demethylation, its off-target activities at non-Ig loci have been linked to many cancers. Sequencing of 118 genes expressed in germinal center (GC) B cells in wild-type and AID -/- mice revealed 25% of genes examined were mutated by AID, and this measure rose to ~50% in littermates lacking the pathways of uracil-removal or mismatch repair (140). AID also targets transgenes placed throughout the genome (141,142). AID "mistakenly" targets thousands of genes throughout the B cell genome. 95% of lymphomas diagnosed in the western world are of B cell origin, the majority of which are derived from mature B cells (143). Burkitt's lymphoma (BL), diffuse large B cell lymphoma (DLCL), follicular lymphoma (FL), multiple-myeloma (MM) and chronic lymphocytic leukemia (CLL) harbor mutations and translocations in proto-oncogenes, such as *c-myc, bcl-6, fas, pim-1, pax-5 and RhoH* (52,144–148). Strong evidence implicates AID. First, these tumors arise from centroblasts or post-centroblasts, the narrow stage of B cell proliferation where AID is expressed. Second, mutations and DNA breakpoints leading to translocations are observed at highly transcribed genes in WRC sequences, both of which are signatures of AID activity. Third, the causal translocations of these tumors (*c-myc*/IgH in BL, *bcl-2*/IgH in FL, *bcl-6*/IgH in DLCL, IgH-CCND1 in Mantle cell lymphoma) invariably involve the IgH locus with JH or Switch region breakpoints as translocation partners, locations at which AID is known to act to mediate SHM and CSR (149–151).

Direct evidence of AID involvement came with the observation that AID deficiency abrogates *c-myc*/IgH translocations in the IL6 transgenic mice. These mice are an ideal model for studying the role of AID in lymphomagenesis, since they suffer spontaneous plasmacytomas caused by a *c-myc*/IgH translocation, almost identical to the signature translocation of human Burkitt's lymphoma (152). It was shown by comparing AID deficient and sufficient littermates

that AID is necessary for the generation of DSBs at both the IgH and *c-myc* translocation sites (153,154). Similarly, in a transgenic mouse model where deregulated *bcl-6* leads to clonal splenic lymphomas, AID deficiency significantly decreased the emergence of lymphomas (155).

Multiple studies involving human lymphomas have shown that AID expression levels in CLL and DLCL correlate with poor diagnosis (156–159). This effect is presumably due to the mutagenic activities of AID causing further genetic instabilities. Indeed, this was shown to be the case in CML, where AID expression correlated strongly with progression from a "dormant" to a "blast-crisis" stage, and with resistance to Imatinib, the commonly used drug of choice for treatment. Transduction of AID into tumor cells was shown to cause hypermutations at tumor suppressor and/or DNA repair genes that are known to mediate gain of Imatinib resistance (49). Moreover, AID expression is upregulated by two common chemotherapeutic drugs (Idelalisib and Duvelisib) (160). Furthermore, AID is also upregulated by cytokines and other factors present in the inflammatory tumor micro-environment (160–163). Although healthy non-lymphoid tissues do not normally express AID, there is strong association between poor prognosis and AID expression in a growing list of non-immune cancers including breast, hepatic and gastrointestinal cancers (31-33) likely reflecting AID increasing tumor mutation rates. Therefore, it is widely agreed that inhibition of AID will be beneficial in treatment of leukemia/lymphoma and other AID-expressing tumors.

The challenges in solving the structure of AID

Because AID is a genome-wide mutator that is involved in initiation and progression of tumors, developing an inhibitor has been a high priority for the research field; however, the major hurdle to this has been the lack of a structure. In structural biology, the two main methods

for resolving atomic-level structures are X-ray crystallography and NMR, which rely on highly purified protein samples in the context of protein crystals or homogenous protein in solution, respectively. At the time when work on this thesis began in 2011, it had been 12 years since the initial discovery of AID, and despite many efforts, ongoing at the time by dozens of the world's leading X-ray crystallography and nuclear magnetic resonance (NMR) structural biology laboratories, AID's structure had remained unsolved. Hundreds of studies had elucidated AID's role in tumorigenesis, epigenetics remodeling and genome-wide mutagenesis in *in vitro* biochemical, cell-based, mouse-based and human-based systems. Despite this tremendous gain of knowledge on AID's functions in health and disease, its basic 3D structure had remained unknown. A structure would be essential to fully elucidate the basis for the unusual biochemical properties of AID. As aforementioned, AID has an unusually high nM range affinity for its substrate ssDNA and a long AID-DNA complex half-life of ~7 minutes (52), but acts ~2000 times slower than a typical enzyme (60), catalyzing one deamination reaction every 1-4 minutes (52,61). It had been hypothesized that AID's high DNA binding affinity and low catalytic rate had evolved to facilitate SHM and CSR but limit excessive genomic stress (52,55,96,164). In support of this idea, AID mutants with lower or higher catalytic rates mediated lower or higher levels of genomic damage when expressed in cells, respectively (62). Structural insights into AID would not only potentially provide insights into these highly unusual and unique enzymatic properties, but it would also be a required step for moving forward with development of potentially therapeutic inhibitors. The inability to produce an X-ray crystal or NMR structure of AID was largely due to its unique biochemical and biophysical properties. Like its other APOBEC family members, the highly charged surface of AID and extensive non-specific protein-protein and protein-nucleic acid interactions has hampered structural resolution by X-ray

crystallography and NMR (165,166). AID is notoriously challenging to isolate to high purity due to genotoxicity and extensive non-specific interactions. Indeed, this is a difficulty shared with its family member APOBEC enzymes. At the time this research was started, several APOBEC structures had been solved by X-ray or NMR; however, they were of fragmented and/or mutated versions, altered in order to enhance expression, solubility and crystallization.

In humans, AID is a small enzyme yet has a high overall positive charge of +14 at physiological pH (96,164). Due to its high density of positive charge, it binds ssDNA with nM affinity and an AID-DNA complex half-life of ~7 minutes (52). In the absence of an X-ray crystal or NMR structure of AID, homology modeling was used as an alternative to study the structure of AID. Briefly, homology modeling relies upon a previously resolved template structure with similar sequence identity/homology whereby the 3-Dimensional structure of the target protein is inferred (167). This method relies upon the well-supported assumption that proteins of similar sequence identity, ideally above 30%, share the same protein fold. Furthermore, homology models of proteins can be complemented by molecular docking of substrate to determine binding modes/binding regions on the protein surface. There are several molecular docking approaches available, but all rely on the basic premise of predicting and ranking dominant binding modes of ligand on a protein surface, using a scoring function derived from physiochemical interactions (168,169). At the time, the only available insights into AID were few reports of a truncated APOBEC2 or partial APOBEC3G-CTD structures, which had also generated crude homology models of AID based on resolved APOBEC template (170,171). However, many of these models were limited to only one conformation and lacked ab-initio structural prediction of the non-conserved C-terminal region (172). Furthermore, none of these models included any molecular docking prediction for where or how substrate ssDNA may bind

on the surface of AID (168). Thus, many molecular determinants underlying AID biochemistry remained uncharacterized. In 2015, we published the "breathing" structure of AID, whereby the entire protein structure was homology modeled based upon several APOBEC templates (96). Using this AID structure, consisting of several low-energy conformations of AID, we described the catalytic pocket, surface topology and how substrate ssDNA binds. In 2017, the X-ray crystal structure of a near-native AID was published which closely agreed with our 2015 AID structure (54).

Rationale and Scope of thesis research

In this thesis, I describe two alternative but complementary methods to solve AID's native, functional and breathing structure. These approaches integrated (1) comparative biochemistry, (2) computational biology and (3) functional enzymology. The first approach, described in chapter 2, involved comparative evolutionary biochemistry. We reasoned that divergent AID orthologs may have distinct properties and characterizing these through a combination of homology structural modeling and functional analysis of mutated and chimeric enzymes will generate structure:function insights. The Second approach, described in chapter 3, involved model-guided functional enzymology wherein I utilized 8 recently solved structures of AID's APOBEC relatives as templates to generate clusters of 10^5 AID model structures and used molecular docking to simulate the binding of each to DNA. I then rigorously examined model predictions by generating and functionally testing a library of 400 AID variants. These included human AID mutants, AID orthologs and chimeras involving regions of other deaminases exchanged into the human AID scaffold, or vice-versa. We integrated the functional data back into modeling and continually repeated this cycle. Using these two approaches, we arrived at a map of AID's functional structure including surface topology, core architecture and catalytic pocket which has since been confirmed

by published crystal structure of AID and several other studies on its related APOBEC enzymes. In Chapter 5, we identified and tested, for the first time, small molecules that inhibit the catalytic activity of AID, to be further developed as novel leukemia/lymphoma therapeutics. In Chapter 6, we review several DNA/RNA binding grooves previously discovered on AID and explore their possible binding combinations. We explore how these plastic modes of binding DNA and RNA simultaneously may impact AID function. Finally, we examine proposed AID dimer models bound to DNA/RNA and examine how this may act as an additional regulatory bottleneck of enzymatic activity.

<u>Chapter 2:</u> Zebrafish AID is capable of deaminating methylated deoxycytidines

Justin J. King[1], Hala Abdouni[1], Mussa Suliman[1], Matthew Quinlan[1], Heather Fifield[1], Mani Larijani[1]*

A version of this thesis chapter has been published previously in *Nucleic acids research* (2013). 41(10). Permission for reuse has been granted from the publisher. This chapter is a modified version of the published manuscript not the contents.

[1]Program in Immunology and Infectious Diseases, Division of Biomedical Sciences, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland A1B 3V6, Canada.

*To whom correspondence should be addressed: Tel: 011-1-709-777 2515; Fax: 011-1-709-777 8294; Email: mlarijani@mun.ca

The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors. For this chapter I contributed to the enzyme assays and carried out all the theoretical and practical structure modeling work. I contributed to data analysis and writing of the manuscript.

Funding

This work was supported by an operating grant from the CIHR to ML (MOP111132). JK is supported by a graduate fellowship from the Beatrice Hunter Cancer Research Institute. MS is supported by a Ph.D. fellowship from the Libyan-North American scholarship program. ML is supported by a CIHR New Investigator award.

Abstract:

Activation-induced cytidine deaminase (AID) deaminates deoxycytidine (dC) to deoxyuracil (dU) at immunglobulin loci in B lymphocytes to mediate secondary antibody diversification. Recently, AID has been proposed to also mediate epigenetic reprogramming by demethylating methylated cytidines (mC) possibly through deamination. AID over-expression in zebrafish embryos was shown to promote genome demethylation through G:T lesions implicating a deamination-dependent mechanism. We and others have previously shown that mC is a poor substrate for human AID. Here, we examined the ability of bony fish AID to deaminate mC. We report that zebrafish AID was unique amongst all orthologs in that it efficiently deaminates mC. Analysis of domain-swapped and mutant AID revealed that mC specificity is independent of the overall high catalytic efficiency of zebrafish AID. Structural modeling with or without bound DNA suggests that efficient deamination of mC by zebrafish AID is likely not due to a larger catalytic pocket allowing for better fit of mC, but rather due to subtle differences in the flexibility of its structure.

Introduction:

Activation-induced cytidine deaminase (AID) is a B lymphocyte specific member of the apolipoprotein B mRNA-editing catalytic component (APOBEC) family of Zn-dependent cytidine/cytosine deaminases (37,173). AID converts deoxycytidine (dC) to deoxyuridine (dU) on single-stranded DNA (ssDNA), preferentially in trinucleotide WRC (W=A/T, R=A/G) motifs (48,50–52,90,174). In activated B cells where it is primarily expressed, AID induces mutations and double-strand breaks (DSB) that initiate somatic hypermutation (SHM) and class switch recombination (CSR) (38,47,175). Inherited mutations of AID that incapacitate function result in Hyper IgM immunodeficiency syndrome (47,176). Although its activity is largely restricted to the Ig loci, AID also initiates off-target mutations and DSBs that can lead to transformations (115,143,145,156,177).

Whilst the roles of AID in immunity and lymphomagenesis are well established, a more recent body of evidence implicates AID in epigenetic reprogramming through genome demethylation (133,138). First, AID has been reported to be expressed in pluripotent tissues such as oocytes, primordial germ cells (PGC) and embryonic stem cells (ES), though at considerably lower levels than in activated B lymphocytes (134,138). Second, PGCs from AID-/- mice have a hypermethylated genome (135). Third, in a heterokaryon system of human somatic cell re-differentiation, AID is required for promoter demethylation (136). Fourth, in zebrafish embryos, AID knockdown and over-expression corresponded with promoter hyper and hypomethylation respectively and morpholino knockdowns of AID led to developmental abnormalities (137).

5-Methylcytosine (mC) in CpG motifs constitutes the most common form of genome methylation associated with transcriptional repression (116–122). In terminally differentiated

human cells, the majority of CpGs are methylated (178–181). In principle, mC can be demethylated either through DNA replication or enzymatic processes. mC can be hydroxylated to form 5hmC by TET family enzymes; though the extent of their involvement *in vivo* as well as the pathways which may act downstream to process 5hmC are poorly understood (125–132). It is also possible that either mC or 5hmC are deaminated by AID or other deaminases, leading to dT or 5hmU respectively, that can be substrates of thymine DNA glycosylase (TDG) or uracil glycoslyase (UNG; (133–137).

The involvement of AID in mC demethylation may or may not be through direct deamination. AID may act as a scaffolding molecule, recruiting or stabilizing complexes with the ultimate function of demethylation. In support of this possibility, dozens of putative binding partners for AID have already been proposed, some of which are involved in DNA repair and transcription(55). Alternatively, several lines of evidence suggest that AID can directly deaminate mC. First, deamination of mC generates G:T mismatches which were observed in the zebrafish experimental system (137). Second, MBD4 which recognizes G:T mismatches was found to act synergistically with AID towards demethylation (137). Third, enzymatically inactive AID harboring a catalytic site point mutation failed to mediate hypomethylation (137).

On the other hand, AID-deficient mice or individuals do not exhibit developmental defects (38,133,176,182), suggesting that the role of AID in genome demethylation may be significantly more subtle and/or redundant than in zebrafish. In agreement with this, we and others showed that mC is a poor substrate for human AID (51,93). In contrast, another study found that human AID efficiently deaminates mC (134), while two more recent studies have confirmed our previous finding (183,184). We have previously shown that bony fish AID exhibit distinct biochemical characteristics compared to their human counterpart (94). Here, we investigated the activity of

purified bony fish AID on mC. We found that in contrast to its human and other bony fish orthologs, zebrafish AID efficiently deaminates mC. This finding suggests that genome demethylation through direct mC deamination by AID is more likely in zebrafish than humans.

Materials and Methods:

AID expression and purification. EcoRI fragments encoding the ORF of medaka AID (Ol-AID) and tetradon AID (TnAID) were synthesized based on identified genomic sequences (http://uswest.ensembl.org/index.html) and cloned into pGEX-5x-3 (GE Healthcare, USA) to generate GST-AID expression constructs in the same manner we have described for human AID (Hs-AID), channel catfish AID (Ip-AID) and zebrafish AID (Dr-AID) (52,94). Domain swapped AID constructs were generated by synthesis of full ORF sequence (Genscript, USA) and point mutants were generated by site-directed mutagenesis. GST-AID was purified as described previously (52). In total, 3 independent preparations of Hs-AID, Ip-AID, Dr-AID, Tn-AID, Ol-AID, domain-swapped and mutant AIDs purified in parallel were used. SDS-PAGE with BSA loading standards was used to equalize the amount of GST-AID used in experiments.

Preparation of substrates. Preparation of partially single-stranded bubble substrates for AID deamination has been described (52). Briefly, 6 bubble substrates were used differing only in the dinucleotide sequence upstream of the target C nucleotide, constituting a WRC or non-WRC motif, with either dC or mC as the target of deamination (Figure 8A). 2.5 pmol of the target strand was 5' labelled with [γ-^{32}P] dATP using polynucleotide kinase (PNK; NEB, USA), followed by purification through mini-Quick spin DNA columns (Roche, USA) and annealing of 2-fold excess (5 pmol) of the complementary strand.

Alkaline cleavage deamination assay. The standard alkaline cleavage assay for measuring the activity of AID on dC has been described previously (48,49). A modified version of the assay was used to measure AID activity on substrates containing mC (Figure 8 A). Briefly, 25 fmol substrate was incubated with 0.25µg GST-AID in 100mM Phosphate buffer pH 7.3 in a volume of 10µl at the optimal temperature for each AID (Hs-AID: 37 ^{o}C, fish AID: 25 ^{o}C) for times ranging from 30 minutes to overnight. AID was deactivated at 85 ^{o}C for 30 minutes. For substrates containing dC, the volume was increased to 20 µl by adding 7.8 µl H_2O, 2 µl 10X UDG buffer and 0.2 µl (1 unit) UDG enzyme (NEB, USA), followed by 30 minutes incubation at 37 ^{o}C in order to excise the uracil and generate an alkali-labile abasic site (Figure 8A, left panel). For substrates containing mC, the volume was then increased to 20 µl by adding 8 µl H_2O, 1 µl 1 M KCl and 1 µl of 1

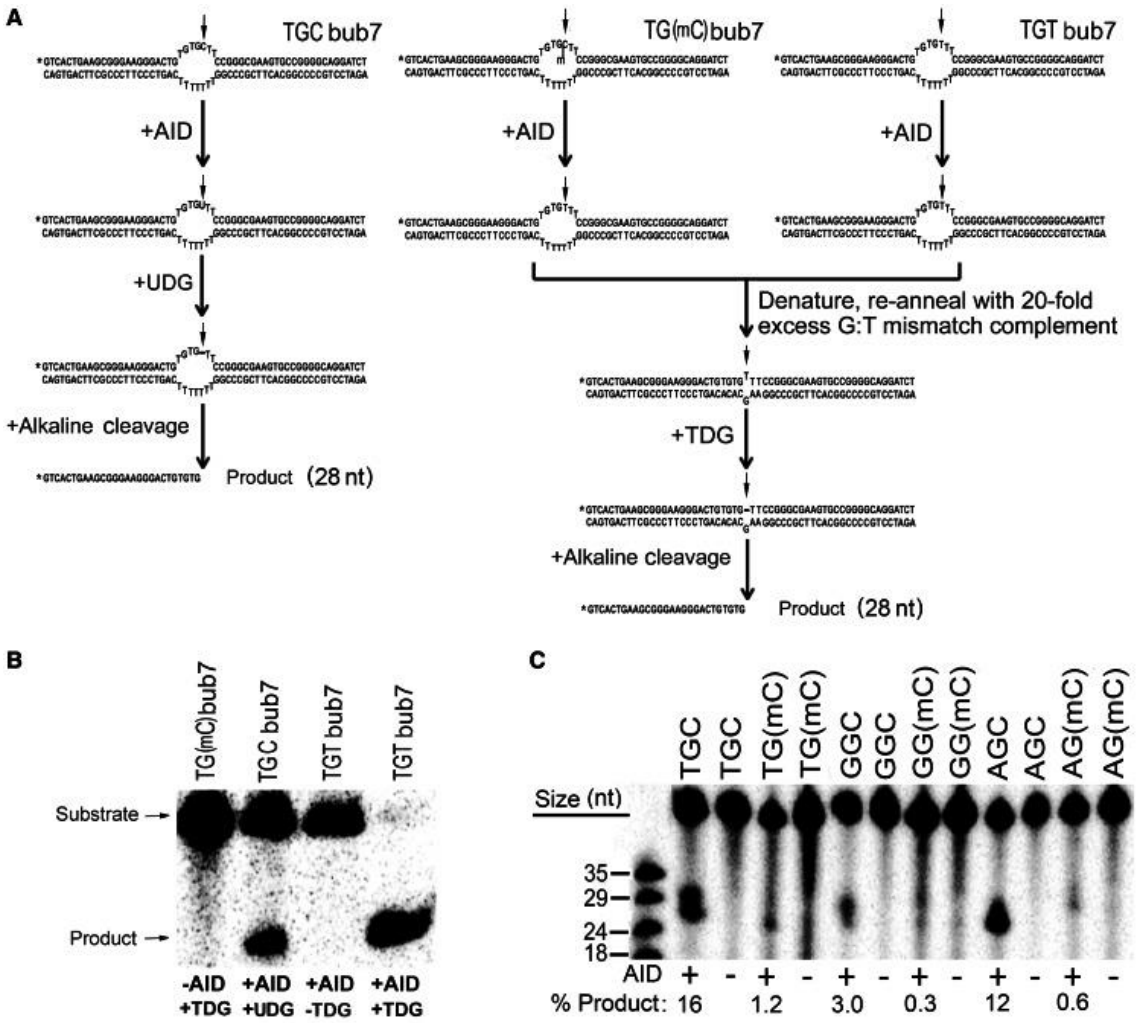


Figure 8. Experimental scheme for measurement of deamination activity on 5-methyl cytosine. (A) Typical bubble type substrates used in this study are shown. TGCbub7 denotes a substrate bearing the WRC motif TGC located in a 7 nucleotide long bubble region. Left panel shows the scheme for the UDG-based alkaline cleavage assay on the WRC bearing substrate TGCbub7. The middle panel shows the scheme for the TDG-based alkaline cleavage assay on the WR(mC) bearing substrate TG(mC)bub7. Right panel shows a control reaction for the TDG-based alkaline cleavage assay to detect the activity of AID on mC. The substrate is TGTbub7, which is the expected product of TG(mC)bub7 after deamination by AID. (B) Control reactions for the efficiency and specificity of TDG. 50 fmol TG(mC)Bub7 and TGTbub7 were incubated with AID, TDG, or both (as shown in panel A). This control was included in each subsequent experiment in order to ensure 100% TDG efficiency. (C) Comparison of Hs-AID activity on dC vs. mC located either in a WRC motif (TGC, AGC) or non-WRC motif (GGC). 50 fmol substrate was incubated with AID. A typical alkaline cleavage gel used to measure the generation of deaminated product is shown and % product formation is shown below each lane.

pmol/µl stock (40-fold excess) of a fully complementary strand. Annealing was performed to generate a G:T mismatch double-stranded substrate. Volume was increased to 30 µl by adding 6

µl H_2O, 3 µl 10x TDG buffer and 1 µl (1 unit) Thymine DNA glycosylase (TDG) enzyme (Trevigen, UK) followed by overnight incubation at 65 ^oC in order to excise the thymine (Figure 8A, middle panel). NaOH was added to 100 mM, and the sample was heated to 96 ^oC for 8 minutes to cleave the abasic site and electrophoresed on a 14% denaturing gel. Control reactions are shown in Figure 8B. The gels were exposed to a Kodak Storage Phosphor Screen GP (Bio-Rad, USA) and visualized using a PhosphorImager (Bio-Rad, USA) on Quantity One software (Bio-Rad, USA).

Data collection and quantitation. Densitometry was performed using ImageLab Analysis Software (Bio-Rad, USA). For each experiment, all AID:substrate reaction combinations were carried out in duplicate and loaded on multiple gels. Individual lanes of each gel were quantitated 3 times, thus obtaining an average value for the lane, representing a single data point (typically the triplicate measurements of each lane as well as the duplicate reaction lanes within each experiment contained a variability of < 5%). The data points are the mean of the values obtained in this manner from all experiments (4-6 experiments for each AID:substrate combination, yielding 8-12 values per data point). The data was graphed using GraphPad Prism Software (Graphpad software Inc., USA). Error bars represent standard deviations.

Structural models and substrate docking. Structure models of AID were produced by submission to the Swiss protein data bank (http://swissmodel.expasy.org/repository/) using either the crystal or NMR-determined structure of the catalytic domain of the family member APOBEC3G (PDB IDs: 3EIU, 2JYW;(185–187). The template was either chosen as best fit by the database or manually chosen and lowest energy state predictions were generated. Generation and analysis of predicted models were performed using Pymol v1.30 (http://www.pymol.org).

Stick and sphere structures of cytosine and 5-methyl cytosine were constructed in ChemDraw v12.0 (http://www.cambridgesoft.com/software/ChemDraw/) and exported into Pymol to scale. Manual orientation of docked cytosine and 5-methyl-cytosine bases was performed based on the requirement for proximity of its atoms to specific residues in the putative catalytic pocket of AID in order for deamination to proceed. Additionally, docking of 2'-deoxycytidine and 5-methyl-2'-deoxycytidine nucleosides was simulated using Autodock Vina (188), whilst allowing for flexibility of nucleoside docking into the putative catalytic pocket of AID in order to examine possible orientations in which the deamination reaction is possible.

Results:

Hs-AID is minimally active on mC on an optimal ssDNA structure

We previously reported that mC in CpG motifs were not efficiently deaminated by purified by His- or GST-tagged AID in long (~500 bp) stretches of ssDNA (93). A second study using purified GST-tagged AID found that mC located in the WRC motif AGC on a 27 nt long ssDNA substrate was deaminated ~10-fold less efficiently than dC (51). In contrast, another study found a less crass preference for dC over mC estimated at ~3 fold when considering the WRC motif AGC (134). Since AID mutates a WRC motif 6-8 fold more efficiently than a non-WRC motif (50–52,90), the results of the latter study indicate that given varied sequence contexts, mC can be deaminated by AID more efficiently than dC. Two recent studies have re-examined the efficiency of mC deamination by AID (183,184). An antibiotic gene reversion assay in bacteria revealed that AID deaminated mC at least an order of magnitude less efficiently than dC and barely above background mutations levels (183), whilst another study that measured the activity of purified MBP-tagged AID on a ssDNA substrate containing the WRC motif TGC, found AID to be 10-16 fold less active on mC as compared to dC (184). We reasoned that some of the differences amongst

studies in the degree of preference for dC over mC may lie in the structure and sequence context of the substrates, which we and others have previously shown to be an important determinant of deamination efficiency by AID (49,174). To this end, we measured the activity of purified GST-AID on a substrate in which the dC or mC target of deamination is located in the middle of a 7nt-long bubble region (Figure 8A), which we have previously shown to be an optimal structure for deamination by AID, supporting several fold higher levels of activity than fully single-stranded DNA (52,55). The activity of AID on dC is detected through alkaline cleavage of an abasic site generated by the removal of the ensuing uracil (Figure 8A, left panel). To generate abasic sites following AID activity on mC, we employed the enzyme thymine DNA glycosylase (TDG) the primary activity of which is to remove thymine bases at G:T mismatches (Figure 8A, middle panel ; (54). To ensure that this experimental approach accurately detects the deamination of mC by AID, we used a control substrate that contains a TGT motif as would result from the action of AID on a TG(mC) substrate (Figure 8A, right panel). As shown in Figure 8B, the control substrate was fully cleaved in a TDG-dependent manner. Thus, this method effectively detects the product of mC deamination by AID. In order to test whether human AID (Hs-AID) can deaminate mC in varying sequence contexts, we measured its activity on 2 WRC (TGCbub7 *vs.* TG(mC)bub7 and AGCbub7 *vs.* AG(mC)bub7) and one non-WRC (GGCbub7 *vs.* GG(mC)bub7) substrates (Figure 8C). On average, mC was deaminated 22-, 20- and 16-fold less efficiently than dC in TGC, AGC and GGC motifs, respectively. We conclude that in an optimal ssDNA structure for AID, its activity on mC is severely diminished as compared to dC.

Zebrafish AID efficiently deaminates methylated cytidines

Given that the proposed role of AID in genome demethylation is largely based on data from the zebrafish, we sought to determine whether zebrafish AID (Dr-AID) can deaminate mC. We

have previously shown that Dr-AID has a higher catalytic rate than Hs-AID (94). In order to allow for the generation of sufficient mC deamination by Hs-AID, both Hs- and Dr-AID were incubated overnight with TGC and TG(mC) since TGC is the most preferred WRC motif for both (Figure 9A). On average, Dr-AID generated 3.6-fold more deamination product on TGCbub7 as compared to Hs-AID (68 *vs.* 18%). Like Hs-AID, Dr-AID deaminated mC less efficiently than dC; however, in contrast to 10-fold less deamination by Hs-AID (1.8 ± 0.1 % *vs.* 18 ± 2%), Dr-AID generated 2-fold less product on TG(mC)bub7 (35 ± 5% *vs.* 68 ± 9%). In order to assess whether the relative lack of discrimination between dC and mC is a property shared by other bony fish AID, we examined AID from the channel catfish (Ip-AID), Japanese killifish medaka (Ol-AID) and the puffer fish tetradon (Tn-AID). We found that like Hs-AID, these deaminated mC 8-19 fold less efficiently than dC (Ip-

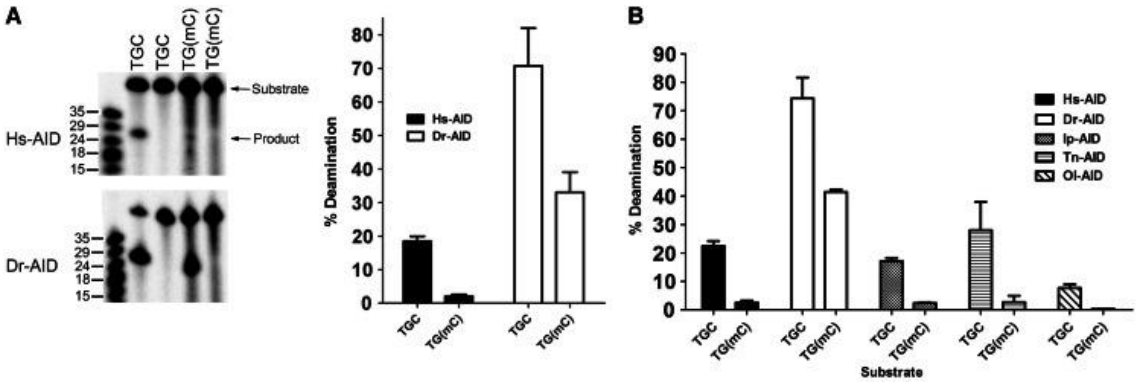


Figure 9. Dr-AID exhibits high activity on mC. (A) Left panel shows alkaline cleavage assay gels showing the deamination of TGCbub7 or TG(mC)bub7 by of Hs-AID (top panel) and Dr-AID (bottom panel). Right panel shows the graphed % deaminated product. Hs- and Dr-AID were incubated overnight with 50 fmol substrate at 37 and 25 ^{o}C respectively. (B) Deamination of TGCbub7 and TG(mC)bub7 by AID of three other bony fish (Medaka: Ol-AID, tetradon: Tn-AID and Channel catfish: Ip-AID) compared to the activities of Hs- and Dr-AID. Each AID was incubated with 50 mol substrate overnight at optimal temperature (37 and 25 ^{o}C respectively for Hs- and fish AIDs).

AID: 21 ± 2% *vs.* 2.1 ± 0.1%, Ol-AID: 7.7 ± 0.1% *vs.* 0.4%, Tn-AID: 28 ± 7% *vs.* 2.7 ± 1%; Figure 9B). To compare the initial catalytic rates of Hs- and Dr-AID on mC, we performed enzyme

kinetics. We included a second substrate TGCGbub7 that is similar to TGCbub7 but contains a dG residue following the target dC, constituting a CpG motif as may be subject to methylation *in vivo*. As we previously reported, Dr-AID was more active on both TGCbub7 and TGCGbub7 than Hs-AID (Figure 10A). The difference in product formation was apparent after a few minutes of incubation, growing to 4.5-fold at 180 minutes. As shown in Figure 10B, the activity of Dr-AID on the TG(mC)bub7 was 5-fold higher than that of Hs-AID after a 5 minute incubation, with the difference growing to 10-fold at 60 minutes and 20-fold at 180 minutes. For TG(mC)Gbub7, the activity of Dr-AID was 7-fold higher than Hs-AID after 5 minutes, with the difference growing to 12-fold at 60 minutes and 40-fold at 180 minutes. Interestingly, the activity of Dr-AID on mC was higher by approximately 2-fold when it was in the context of a typical CpG motif in TG(mG)G as compared to TG(mC)T.

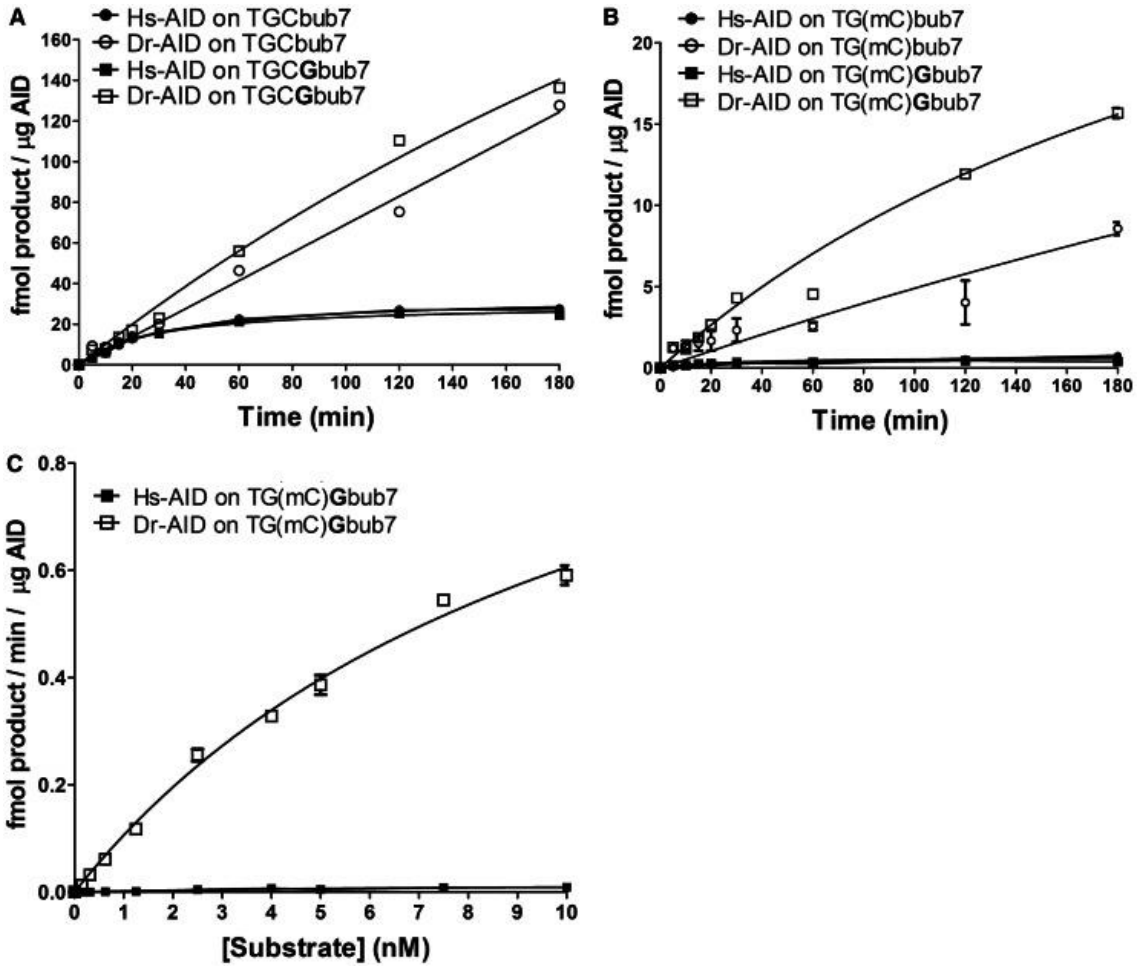


Figure 10. Deamination kinetics comparing the activities of Hs-AID and Dr-AID. (A) Comparison of activity kinetics between Hs-AID and Dr-AID on 7 nucleotide bubble substrates bearing dC in a WRC or a WRCG motif. 50 fmol substrate was incubated with AID for various times. Product formation for 1 µg AID is shown as a function of incubation time. (B) Comparison of activity kinetics between Hs-AID and Dr-AID on 7 nucleotide bubble substrates bearing mC in a WRC or a WRCG motif. (C) Deamination kinetics to compare the deamination rates of dC and mC in a WRCG motif by Hs-AID and Dr-AID. Various concentrations of the substrate TG(mC)Gbub7 ranging from 0.1 to 10 nM were incubated with AID. Velocity was calculated as the amount of deaminated product generated by a given amount of AID in a unit of time and plotted against substrate concentration.

This finding further suggests that Dr-AID has the enzymatic potential to contribute to genome demethylation in the zebrafish. We then measured the Michaelis-Menten kinetics of mC deamination by Hs- and Dr-AID (Figure 10C). On TG(mC)Gbub7 the Kcat:Km ratio for Dr-AID was estimated to be approximately 60-fold higher than that of Hs-AID. We conclude first that the

activity of Dr-AID on mC is unique amongst all bony fish orthologs tested here, second that Dr-AID can deaminate mC more efficiently than AID from other species deaminate dC, and third that the initial rates of mC deamination by Dr-AID indicate a unique specificity for mC.

Sequence context differentially influences the activity of AID orthologs on mC

We have previously shown that whilst purified Dr-AID and Ip-AID possess general WRC specificity similar to Hs-AID, they exhibit subtle alterations in their sequence preference (90,94). We sought to examine whether the relatively high level of Dr-AID activity on mC is a unique feature of the WRC motif TGC. To this end, we measured the activities of Dr-AID, Hs-AID and Ip-AID on WRC non-WRC motifs (Figure 11A). We found that Dr-AID deaminated mC 1.7-, 3- and 4.7-fold less efficiently than dC in TGC, AGC and GGC motifs respectively, compared to 8-, 7.5- and 11-fold for Ip-AID and 22-, 20- and 16-fold for Hs-AID (Figures 8C, 10B). We conclude first that Dr-AID is more active on mC than Hs-AID or Ip-AID regardless of sequence context, and second that in contrast to Hs-AID and Ip-AID which disfavor mC compared to dC almost equally regardless of sequence context, Dr-AID discriminates against mC to a lesser degree (1.7 and 3-fold *vs.* 4.7-fold) in a WRC motif.

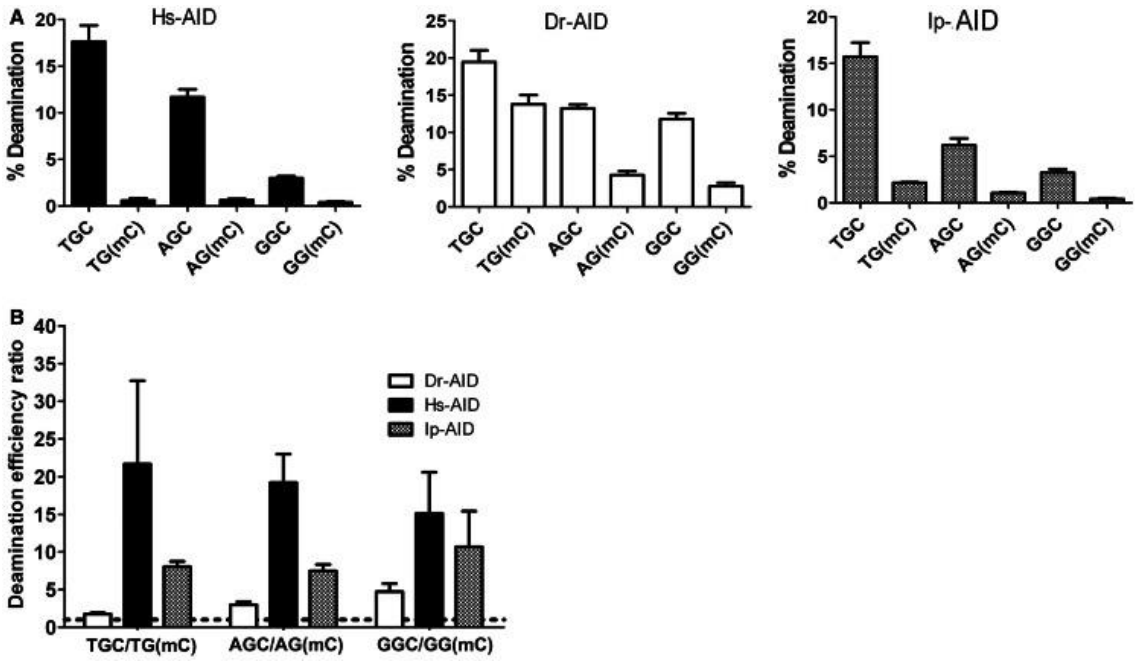


Figure 11. The influence of WRC sequence specificity of AID on mC activity. (A) Left, middle and right panels show the activities of Hs-, Dr- and Ip-AID respectively on C and mC, located in 2 WRC motifs (TGC, AGC) or a non-WRC motif (GGC). AID was incubated with 40 fmol substrate for 1 hour. % deamination product was quantitated and graphed. (B) The ratio of % deamination product for C/mC was calculated for each AID and graphed in order to show the fold preference for C over mC, in each of the WRC and non-WRC sequence motifs.

The role of putative catalytic pockets of Hs- and Dr-AID in differential discrimination against mC

One possible mechanism for the low activity of Hs-AID on mC may be a size restriction against the methyl moiety imposed by its catalytic site. In support of this, a recent study reported an inverse correlation between the "bulk" of cytidine side chain modifications and deamination efficiency by AID (184). Alternatively but not mutually exclusively, it is possible that a restriction is imposed by chemical interactions between the methyl side chain and catalytic residues of Hs-AID that impede deamination. To examine these possibilities, we sought to dock either the C or mC base or respective nucleoside versions into the putative catalytic pockets of Hs- and Dr-AID. The structure of AID has not yet been determined. Thus, we utilized models of the AID catalytic

pocket that we previously generated based on the resolved structure of the catalytic region of the family member APOBEC3G (Figure 12A; (55,94).

To manually dock C or mC into the putative catalytic pockets of Hs- Dr-AID, we took into account the mechanism of deamination by AID. This mechanism can be speculated on with a high degree of confidence, based on that of other cytidine deaminases, the resolved structure of family member APOBEC3G, and the model structure of AID itself (173,185,186,189–191). A Zn ion required for nucleophilic attack during deamination is clinched by the triad of Zn-coordinating residues (in Hs-AID: H56, C87, C90, and Dr-AID: H60, C99, C102). The carboxylate ion of a glutamic

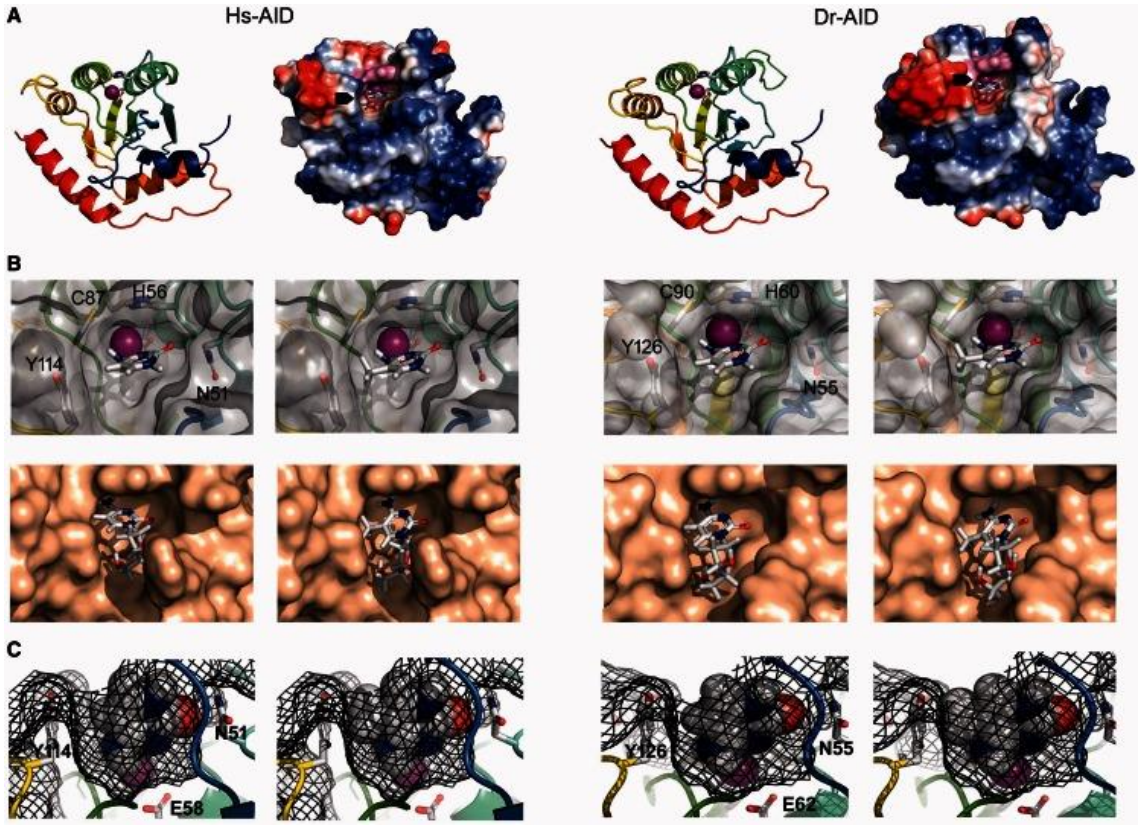


Figure 12. Models of the putative catalytic pockets of Hs- and Dr-AID with docked C or mC. (A) Models of the AID monomer were generated based on the solved structure of the APOBEC3G catalytic region. Left panel shows a ribbon and surface charge diagram of Hs-AID and right panel

shows the same for Dr-AID. N to C-terminus progression is shown in color from blue to red. The ribbon diagram shows the putative catalytic (Zn-coordinating) residues are shown in stick (Hs-AID: H56, E58, C87, C90; Dr-AID: H60, E62, C99, C102). The surface topology and charge models show positively and negatively charged residues are blue and red, respectively. The putative catalytic pockets containing the Zn-coordinating residues are shown in magenta, with a docked cytidine base inside the pocket as marked by black arrows. (B) Modeled Top down view of the surface topology of the putative catalytic pocket of Hs-AID (left panel) or Dr-AID (right panel), with either C or mC bases (top row) using manual docking, or by using Autodock software to dock dC or mC nucleoside versions (bottom row). Zn is shown as a magenta sphere. Residues that surround the catalytic pocket and affect the fit of C or mC are shown. (C) Mesh diagram of the modeled side view of the catalytic pockets of Hs-AID (left panel), or Dr-AID (right panel), showing the fit of C (left side) or mC (right side). The mesh diagram takes into account the spaces occupied by the atomic radii of both substrate and AID. Zn is shown as a magenta sphere. Residues that surround the catalytic pocket and affect the fit of C or mC are shown.

acid (E58 and E62 in Hs- and Dr-AID, respectively) coordinates a water molecule and through proton shuttling converts it into a reactive hydroxide ion. Hydrolytic deamination of C occurs through a nucleophilic attack by the Zn-OH on the C4 position of the pyrimidine ring to which the exocyclic amino group is bonded. Following proton donation, the deprotonated side chain of E58 stabilizes the N-H group of the intermediate through a hydrogen bond. Based on this reaction mechanism, we used three criteria for situating C and mC in the putative catalytic pocket: first, C4 is juxtaposed to the Zn ion. Second, the cyclic N3 and the hydrogen donor of the E58 side chain are proximal allowing for proton donation in the first step of the deamination reaction. Third, the exocyclic carbonyl group must be proximal to N51 in Hs-AID and N55 in Dr-AID. This hydrogen bond appears vital for deamination catalysis: this asparagine is proximal to the catalytic site and 100% conserved amongst orthologs. Furthermore, mutation of its equivalent residue (N244A) in APOBEC3G abolished deamination and the equivalent residue in Staphylococcus aureus tRNA adenosine deaminase (TadA) has been shown to stabilize the substrate by hydrogen bonding (185,192). Using these criteria, we positioned C and mC into the putative catalytic pockets of Hs- and Dr-AID (Figure 12B, top panel). In order to validate our manual dockings, we also utilized auto-docking software, allowing for fit of the larger nucleoside versions dC or mC. We observed

a range of rotations of the substrate molecule in which deamination was plausible, but the lowest energy state confirmed our positioning of the substrate within the putative catalytic pockets (Figure 12B, bottom panel). In order to assess fit, we modeled spaces occupied by atomic radii (Figure 12C). The dimensions of the catalytic pockets of Hs- and Dr-AID appear comparable (8.3 x 7.2 x 6.8 *vs.* 8.3 x 6.7 x 6.8 Å). The dimensions of C and mC are 7.8 x 8.3 x 4.2 and 8.8 x 8.3 x 4.2 Å, respectively, suggesting that C readily fits into the catalytic pocket whilst the fit of mC is taut (Figure 12C). Other than its size thus restricting fit, we did not observe a chemical interaction between the methyl moiety and residues constituting the "floor" or "walls" of the putative catalytic pockets that may hinder deamination catalysis. With the caveat that actual structures are not known, we conclude that the fit of C and mC into the modeled catalytic pockets of Hs- Dr-AID explain the preference for dC over mC but not the lower degree of discrimination between C and mC specific to Dr-AID.

The role of catalytic efficiency in differential activity of Dr-AID on mC

We previously showed that Dr-AID has a higher catalytic rate than Hs-AID which corresponds with higher ssDNA binding affinity (94). We sought to examine whether efficient deamination of mC by Dr-AID is a general function of its more robust activity. We reasoned that discrimination between substrates with subtly different chemical features may be less apparent in a more highly active AID enzyme, and that the size restriction of mC may be overcome by higher overall higher ssDNA affinity providing more likelihood of fit into the catalytic pocket for each AID:substrate binding event. In support of this notion, Dr-AID is also less discriminating between non-WRC and WRC motifs, exhibiting ~ 2-fold less activity on the non-WRC motif GGC as compared to the WRC motif TGC, in contrast to a ~6 fold for Hs-AID (Figure 10A; (94). We previously reported that a single mutation (D176G) in the C-terminal domain of Ip-AID results in

an enzyme nearly identical to Dr-AID with respect both to catalytic rate and ssDNA binding affinity (Kcat and Kd; (94). We found that despite its similar activity profile to Dr-AID on dC, Ip-AID D176G does not have appreciable activity on mC (2 ±0.1% for Ip-AID on mC vs. 4 ±2% for Ip-AID D176G on mC; Figure 13A). In addition, we generated two-point mutants of Hs-AID (R36A and T110A) with ~ 2-fold increased activity on dC and compared their activities on mC. Despite increased activity on dC, neither produced a

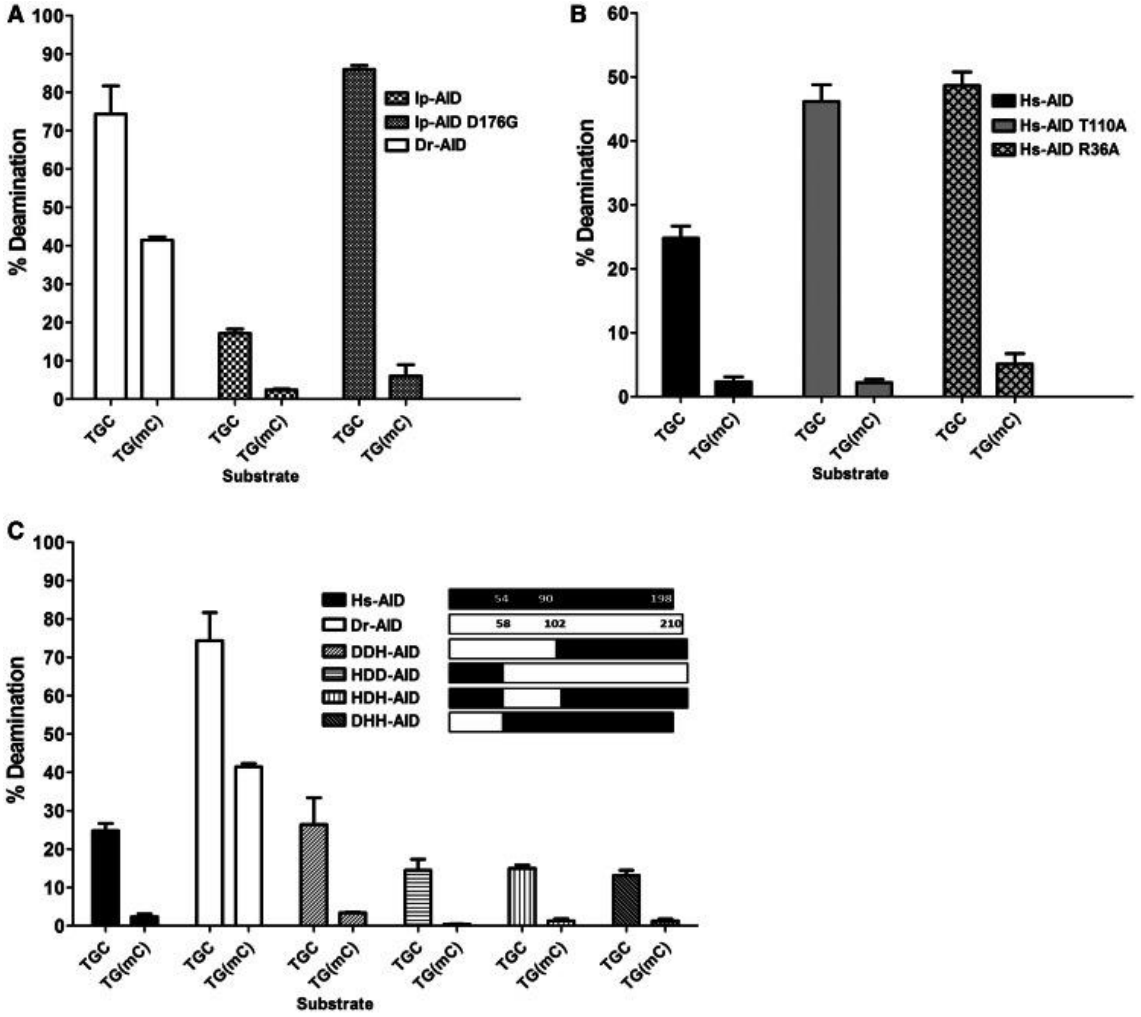


Figure 13. The activity of AID mutants and domain-swapped AIDs on mC. (A) Comparison of the deamination activities of Dr-AID, Ip-AID and Ip-AID (D176G) on TGCbub7 and TGC(mC)bub7. (B) Comparison of the deamination activities of Hs-AID, Hs-AID (R36A) and Hs-AID (T110A) on TGCbub7 and TG(mc)bub7. (C). Comparison of the deamination activities of Hs-AID, Dr-AID and hybrid enzymes with domains swapped between Hs- and Dr-AID on

TGCbub7 and TGC(mC)bub7. AID was incubated with 50 fmol substrate for 3 hours and % deamination product was graphed.

disproportionate increase of deamination at mC (2 ±0.5%, 2 ± 0.1% and 5 ± 2% for Hs-AID, Hs-AID T110A and Hs-AID R36A, respectively; Figure 13B). Thus, mutations in AID that we have previously shown to increase its catalytic activity on dC, did not result in a gain of activity on mC. We conclude that the activity of Dr-AID on mC is a biochemical property that is independent of its high catalytic efficiency on dC.

Based on the position of the catalytic residues and homology to other deaminases, AID can be divided into N-, catalytic and C-terminal domains (Figure S1; defined by arrows). To determine whether an individual domain of Dr-AID can transfer its high activity on mC, we constructed four domain-swapped enzymes between Hs-AID and Dr-AID. We found that none of these hybrid enzymes was more active on mC than Hs-AID (Figure 13C). From these results and given the collection of regions swapped in the hybrid enzymes, we conclude that this property of Dr-AID cannot be attributed to a portion of AID in the manner in which we apportioned the primary structure and is rather an attribute that likely involves multiple residues that are discontinuous in the primary structure.

Discussion

Methylation of promoter region cytidines at C5 is a marker of silenced genes, in normal differentiation as well as in oncogenesis (116–122,178–181,193–198). Several recent lines of evidence implicate AID in genome demethylation(133,138): first, over-expression, knockdown or absolute deficiencies in zebrafish and mice correlate with changes in levels of methylation in the genome and/or in introduced plasmids (136,137). Second, AID has been reported to be expressed

at low levels in germ cells where promoter demethylation is a major contributor to pluripotency (134). Third, in a zebrafish embryogenesis model, AID knockdowns exhibited gross abnormalities in neuronal development, concomitant with hypermethylation at loci encoding for transcription factors involved in neurogenesis. Evidence for the involvement of a G:T mismatch intermediate expected from mC deamination as well ablation of the hypomethylation activity of catalytically-dead AID mutants suggest that genome demethylation in zebrafish can proceed through AID-mediated deamination of mC (137).

In contrast to zebrafish, humans and mice deficient in AID bear no developmental abnormalities, suggesting that the role of AID in genome demethylation is either negligible or redundant. This is not surprising, first given that genome demethylation can proceed through a multitude of enzymatic and chemical processes, and second that other deaminases such as APOBEC-1, APOBEC-2 and APOBEC-3A have been suggested to play a role (130,134,183,199). Although the role of AID in genome demethylation in humans seems redundant, it brought forth the question of whether AID is capable of deaminating cytidines carrying a methyl group at the C5 position. To date, four reports have shown that Hs-AID deaminates mC with approximately 10-100 fold lower efficiency as dC (51,93,183,184). In contrast, one report showed that Hs-AID can deaminate mC at a comparable efficiency as dC, given a comparison of dC and mC located in different sequence contexts (134). It is formally possible that differences in the observed degree of AID's preference for dC over mC are due to differing experimental parameters such as sources of AID expression, purification methods and ssDNA structures or the local sequences in which dC or mC were located. However, the reports that show AID deaminates mC at a very low efficiency have employed a host of purification and assay methods, including prokaryotic and eukaryotic expression, and evaluating the activity of AID on mC in bacterial genomes, *in vitro* using purified

His-, GST- or MBP-tagged AID on fully or partially single-stranded substrates, and in various sequence contexts (51,93,183,184). Here, we report that even in an optimal ssDNA structure and regardless of whether the mC is positioned in a WRC or non-WRC sequence motif, it is deaminated very poorly by Hs-AID. Taking together that mC is a poor substrate for Hs-AID, that AID expression is largely restricted to activated B cells where it targets the Ig locus >90% of time (42,55,200) and that AID expression levels in germ cells are quite low, it appears unlikely that AID would be a significant contributor to genome-wide demethylation in humans. It is more likely that AID may play a minor and redundant role, the exact nature of which is yet to be determined.

We have previously shown that AID of bony fish have different biochemical properties compared to human AID (94). Thus, we hypothesized that the more prominent role of AID in genome demethylation in zebrafish may be indicative of its enzymatic capacity to act on mC. We show that this is indeed the case, and that this biochemical property appears to be unique amongst AID of other bony fish tested. The activity of Dr-AID on mC can be explained neither by an overall more accommodating modeled catalytic site, nor by its more robust enzymatic rate as compared to Hs-AID. In the absence of an experimentally determined structure, we used modeling to gain insight into the mechanism of high Dr-AID activity on mC. In modeling the structure of Hs-AID, we noted that Y114 was the closest residue to the methyl group of mC docked in the catalytic pocket (Figure S2A). We note that Y114 is conserved in AID of all species (Figure S1). Two recent studies have reported that in contrast to APOBEC3G, APOBEC3A is able to efficiently deaminate mC (183,199). Like the case in our study, this difference in mC activity could also not be attributed to an overt structural difference between the two enzymes. Based on modeling the APOBEC3A structure, it was postulated that its equivalent residue to the Y114 of Hs-AID (Y130) was positioned with its side chain further away from the catalytic pocket as compared to Hs-AID

(183). We noted in our models that this residue is located at the start of a loop that encircles the putative catalytic pocket (Figure S2A, B). Engraftment of this loop from APOBEC3A to Hs-AID transferred the ability to efficiently deaminate mC, suggesting that the composition of the loop supports a positioning of this tyrosine residue that is more conducive to docking of mC (183). However, unlike the case of APOBEC3A, we noted that this tyrosine in Dr-AID (Y126) was positioned similarly to Y114 in Hs-AID and appeared equally likely to cause steric hindrance with the fit of mC (Figure S2B). A comparison of AID models based on the NMR- *vs.* X-ray determined structure of the APOBEC3G catalytic domain revealed that the majority of residues held similar conformations; however, this tyrosine was rotated by ~180◦ and was shifted away from the catalytic pocket in Hs-AID and Dr-AID, when modeled on the NMR template (Figure S2C). We conclude that this tyrosine and the loop harboring it are likely to undergo a high degree of protein breathing, explaining their different positioning in solution and crystal forms. In addition, we noted that this loop is longer in Dr-AID than Hs-AID by one negatively charged residue (E130) that can allow for further flexibility (Figure S1). These observations suggest that due to differences in the charge make up and length of the loop harboring Y126 in Dr-AID, this tyrosine may occupy a lower energy state in the "away" conformation, as compared to Hs-AID (Figure S2 B,C), thus allowing for better fit of mC more often. For conclusive insights into the molecular basis of the difference in mC activity between Dr-AID and Hs-AID, we look forward to the resolution of their structures.

Whatever the structural basis of higher activity of Dr-AID on mC may be, our results provide a mechanistic explanation for the prominent role for AID in genome demethylation in zebrafish. It is possible that in humans APOBEC3A functions as a major contributor to genome demethylation, but in zebrafish that lacks APOBEC3A, this role is delegated to AID. On the other

hand, mice and other bony fish examined here have neither APOBEC3A, nor an AID capable of deaminating mC. This suggests strongly that even though some APOBEC family deaminases are enzymatically capable of genome demethylation, this function is mainly carried out by other pathways and that zebrafish may be an exception. Since AID is present in the earliest jawed vertebrates (ie: sharks), it is thought to be the ancestral member of the APOBEC family (201,202). On the most recent end of the APOBEC family evolutionary spectrum are the APOBEC3 conglomerate, having expanded from one to seven members in primates (173). In this light, our finding that AID from zebrafish but not three other bony fish species efficiently deaminates mC, combined with the finding that APOBEC3A but not APOBEC3G is able to efficiently deaminate mC (183,199) presents a difficulty in assigning a linear evolutionary association to activity on mC. It appears more likely that the ability to deaminate mC is a somewhat secondary byproduct of overall differences in the internal folding of the enzyme structure amongst APOBEC family members. It will be of interest to determine whether this enzymatic feature of AID corresponds with a differential role in genome demethylation amongst species.

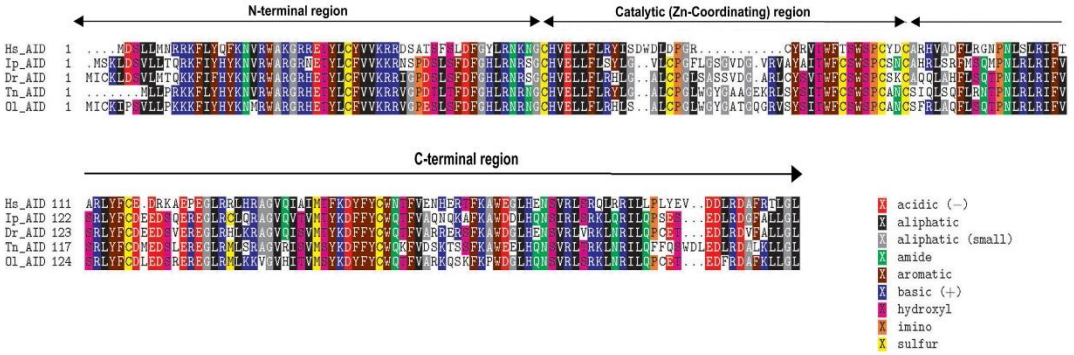


Figure S1. Comparison of the primary structure of AID from human and four bony fish. Residues are colored according to side-chain. Blue, green and red underlines indicate identical, similar and unlike residues. Arrows denote the borders of the catalytic domain.

Sequence alignment based on published sequences of AID (Homo sapiens: AAM95406.1, Danio rerio: NP001008403, Ictalurus punctatus: AAR97544) and based on unpublished sequences (Oryzias latipes and Tetradon nigroviridis).

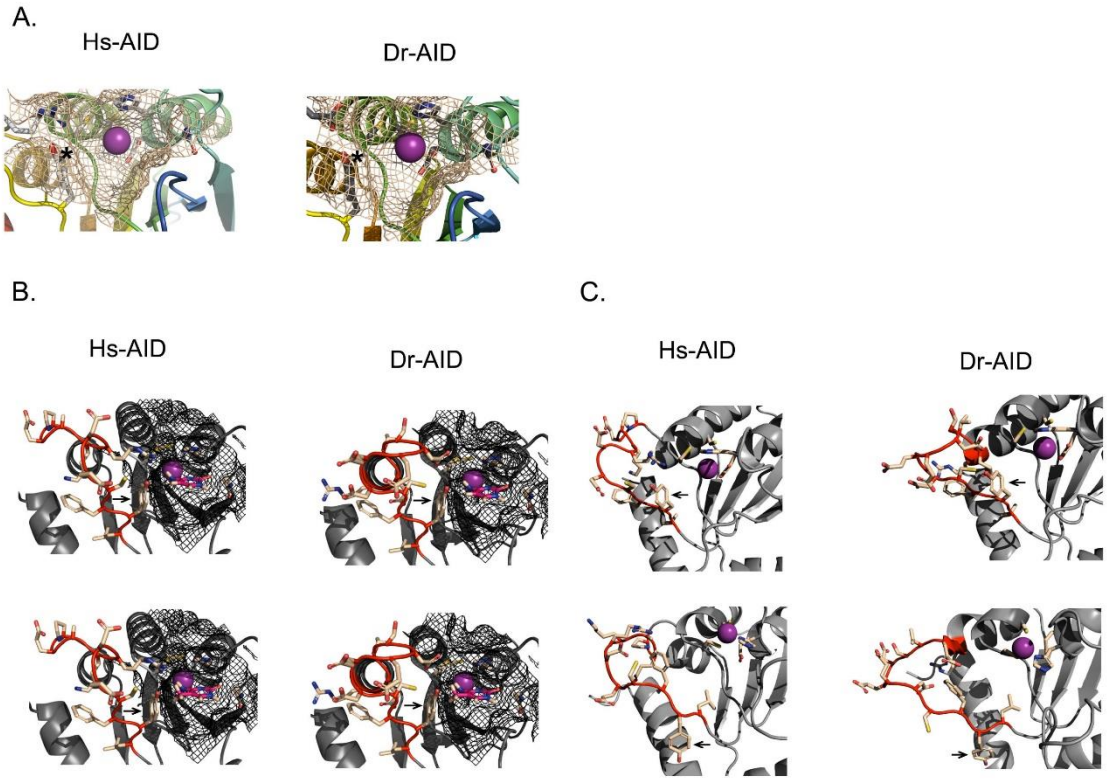


Figure S2. Models of the putative catalytic pockets of Hs- and Dr-AID with docked C or mC. (A) Combination ribbon and mesh model of Hs-AID (left panel) and Dr-AID (right panel), showing the position of Y114 in Hs-AID and its equivalent residue in Dr-AID (Y126) relative to the catalytic pocket. Positions of Y114 and 126 are indicated by *. (B). Combination mesh and ribbon diagram of Hs-AID (left panel) and Dr-AID (right panel) with C (top row) or mC (bottom row) docked into the catalytic pocket, showing proximity to Y114 (Hs-AID) and Y126 (Dr-AID). The backbone of the flexible loop (in Hs-AID: L113-P123, in Dr-AID: L125-R136) containing the tyrosine residue is colored in red. Arrows denote the position of Y114 and 126. (C). Comparison of the relative positions of Y114 in Hs-AID and Y126 in Dr-AID between models based on the X-ray determined structure of APOBEC3G catalytic unit (top row), or its NMR-determined structure (bottom row). Comparison of the top and bottom row reveals a high probability of significant breathing of Y114 and Y126. Arrows denote the position of Y114 and 126.

Chapter 3: Catalytic pocket inaccessibility of activation induced cytidine deaminase is a safeguard against excessive mutagenic activity

Justin J. King*, Courtney A. Manuel, Crystal V. Barrett, Susanne Raber, Heather Lucas, Patricia Sutter, Mani Larijani*

A version of this thesis chapter has been published previously in *Structure* (2015). 23(4):615-27. Permission for reuse has been granted from the publisher. This chapter is a modified version of the published manuscript not the contents.

Program in Immunology and Infectious Diseases Program, Division of Biomedical Sciences, Faculty of Medicine, Memorial University of Newfoundland

*To whom correspondence should be addressed:

mlarijani@mun.ca or jjk458@mun.ca

For this chapter, Dr. Larijani and I formed the hypotheses and the conceptual framework, and I conducted all the enzyme assay and modeling experiments, analyzed the data and contributed to the writing of the manuscript.

Address: Room 1809A Health Sciences Center, Memorial University of Newfoundland, 300 Prince Philip Dr, St. John's, NL, Canada A1B 3V6.

Acknowledgements

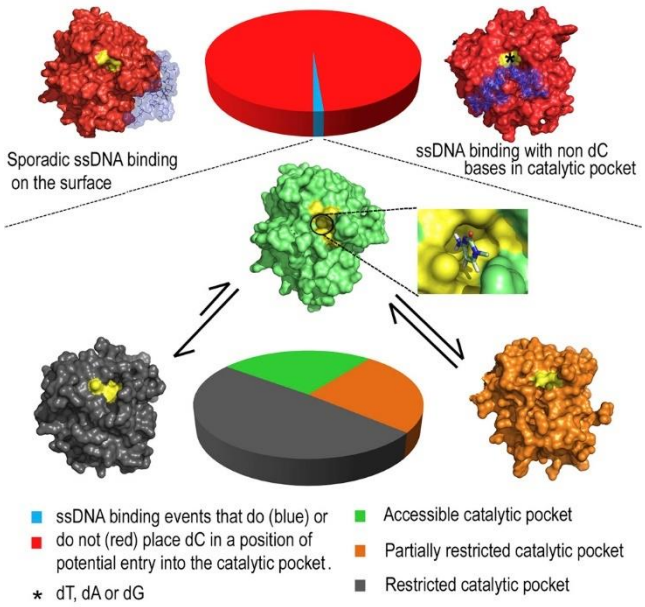
This work was supported by a Canadian Institutes of Health Research (CIHR) operating grant (MOP111132) and a Canadian Cancer Society Research Institute (CCSRI) Innovation operating grant to ML. JK was supported by a Ph.D. fellowship from the Beatrice Hunter Cancer Research Institute

(www.bhcri.ca) with funds provided by the Cancer Research Training Program as part of the Terry Fox Foundation strategic Health Research Training Program in Cancer Research at CIHR. JK is currently supported by a CIHR Frederick Banting and Charles Best Canada Doctoral Scholarship. ML is a CIHR New Investigator.

Abstract

Activation-induced cytidine deaminase (AID) mutates cytidine to uridine at immunoglobulin loci to initiate secondary antibody diversification but also causes genome-wide damage. We previously demonstrated that AID has a relatively low catalytic rate. The structure of AID has not been solved. Thus, to probe the basis for its catalytic lethargy we generated a panel of free or DNA-bound AID models based on eight recently resolved APOBEC structures. Docking revealed that the majority of AID:DNA complexes would be inactive due to substrate binding such that a cytidine is not positioned for deamination. Furthermore, we found that most AID conformations exhibit fully or partially occluded catalytic pockets. We constructed mutant and chimeric AID variants predicted to have altered catalytic pocket accessibility dynamics and observed significant correlation with catalytic rate. Data from modeling simulations and functional tests of AID variants support the notion that catalytic pocket accessibility is an inherent bottleneck for AID activity.

Graphical Abstract



Introduction

Activation-induced cytidine deaminase (AID) initiates DNA lesions at Immunoglobulin (Ig) loci that mediate antibody diversification (37,38). Inherited defects in AID result in hyper IgM immunodeficiency (38,47). AID also acts genome-wide leading to cancer driving mutations and chromosomal translocations (145,156,177,203,204). Human AID is 198 amino acids long and is a member of the apolipoprotein B mRNA editing enzyme catalytic polypeptide-like (APOBEC) family of Zn-dependent, single-stranded polynucleotide-restricted cytidine deaminases (37,173,205). It converts deoxycytidine (dC) to deoxyuridine (dU) in single-stranded DNA (ssDNA) preferentially in WRC (W=A/T, R=A/G) motifs (48–50,90). We previously showed that AID binds ssDNA with high affinity in the nM range (52). We demonstrated that AID has a significantly lower catalytic turnover rate compared to most enzymes, mediating a reaction every 1-4 minutes, a finding that has since been confirmed by others (61). We postulated that this lethargy represents an evolutionary "built in" safeguard (52,55). That AID mutants with increased catalytic rates mediate higher levels of double strand DNA breaks when transfected into cells supports this notion (62). The 3-dimensional structure

of AID has eluded resolution due to extensive non-specific interactions interfering with absolute purification (55,206). Hence, molecular determinants underlying its slow catalytic rate and other biochemical properties remain unknown.

To gain insight into its structure, we and others have presented several models of AID. In 2007, Prochnow et al. modeled AID based on the crystal structure of truncated APOBEC2 ($A2_{41-224}$), which formed a rod-shaped tetramer (170). Several hyper IgM-causing mutations were found in the predicted oligomeric interfaces, suggesting that AID forms a tetramer similar to A2. This is consistent with our previous observation that active AID is enriched in fractions corresponding to dimers or tetramers in size (52). It is noteworthy that A2 itself is inactive (173) and that the tetrameric crystal structure was obtained from a truncated human $A2_{41-224}$ whilst full length mouse A2 was later shown to be a monomer by NMR (207). Other studies showed AID, APOBEC3A (A3A) and APOBEC3G (A3G) can be active *in vitro* as monomers (57,208,209). In 2008, Patenaude et al. presented a dimeric AID model, based on $A2_{41-224}$, to illustrate that residues mediating cytoplasmic-nuclear shuttling formed a Nuclear localization signal (NLS; (171). In 2009 and 2010, two studies presented a partial model of a loop region that when transplanted between AID and A3G/A3F transfers sequence specificity of the donor (58,59). Consistent with its role in substrate specificity, in 2012, Wijesinghe et al. showed that the same loop mediates the difference between AID and A3A in the ability to deaminate 5'-methylated cytidines (5mC) (183). In 2012 and 2013, we presented several core and surface charge/topology models of bony fish and human AID, based on the resolved partial A3G structure these models had features that correlated with differences in biochemical properties amongst bony fish and human AID (55,94,95).

The aforementioned AID models had several limitations. First, each was focused on one region of interest not taking into account the impact of neighboring motifs. Second, these models were

constructed from single templates. Here we utilized 8 recently solved full or partial APOBEC structures to generate a large panel of AID models with and without bound DNA. This approach allowed us to examine the probabilities of different structural conformations as well as distinguish portions of the structure that are likely to be more rigid *vs.* flexible. We found that the majority of AID:DNA interactions do not lead to deamination and that breathing restriction of its catalytic pocket is a major bottleneck to AID activity. To test this functionally, we generated AID mutants and chimeric variants predicted to have altered catalytic pocket accessibility dynamics. Enzyme kinetic analysis of these variants revealed significant correlation between the degree of catalytic pocket accessibility and catalytic rate, supporting the model that restriction of its catalytic pocket is an inherent regulator of AID activity.

Materials and methods

APOBEC template selection, AID modeling and evaluation of model quality

Eight APOBEC structures were chosen as templates: mouse A2 NMR (PDB: 2RPZ), $A2_{41-224}$ chain A X-ray (PDB: 2NYT, chain A), $A2_{41-224}$ chain B X-ray (PDB: 2NYT, chain B), A3A NMR (PDB: 2M65), A3C X-ray (PDB: 3VOW), A3F-CTD X-ray (PDB: 4IOU), A3G-CTD NMR (PDB: 3E1U) and A3G-CTD NMR (PDB: 2KBO) (Figure S3; (185,187,207,208,210–212). All APOBEC template structures were obtained from the Protein Data Bank (http://www.rcsb.org) and visualized using PyMOL v1.3 (http://www.pymol.org/). Using the default parameters of I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) (213,214) full-length human AID (Hs-AID) and variants were modeled from the APOBEC templates to generate 39-40 models for each. In total, considering wild type human AID and AID mutants or variants, 795 models were constructed. In each model of AID, the entire protein was homology modeled except for the non-homologous 18 C-terminal amino acids, which were modeled *ab initio*. Models were compared using the superimpose function

of PyMOL, aligning structures according to the peptide backbone. Each AID model is referred to using the template that it is based on in square brackets. To evaluate overall model quality, QMEAN-6 Z-scores were calculated to compare overall model qualities. Ramachandran plots were generated for each model to check the stereochemical quality on a residue-by-residue basis using PROCHECK (215,216).

Definition of catalytic pocket and docking of DNA substrates

The putative catalytic pocket of AID was defined by Zn-coordinating and catalytic residues (H56, C87, C90 E58 in Hs-AID) as aligned with other APOBECs (55,205). We defined the catalytic pocket as the indented space containing the catalytic pocket residues. Dimensions of the catalytic pocket were measured using Pymol. DNA substrates were docked to each AID model using Swiss-Dock (http://www.swissdock.ch; (217,218). Each substrate was constructed in Marvin Sketch v.5.11.5 (http://www.chemaxon.com/products/marvin/marvinsketch/) while surface topology and docking parameters were generated using Swiss-Param (http://swissparam.ch; (219). These output files served as the ligand file in Swiss-Dock. Control non-cytosine containing (5'-TTTGTTT-3') and experimental (5'-TTTGCTT-3') ssDNA substrates were chosen since previous work has shown these to be preferred substrates of AID (48–50,52,220). AID:substrate docking simulations for each AID enzyme resulted in 5000-15000 binding modes. 8 or more recurring binding modes were clustered based on RMS values. The 32 lowest energy clusters were selected, thus representing 256 of the lowest energy individual binding events for each AID. In total, 25,440 low-energy docking clusters were analyzed for wild type human AID, mutants or variants in UCSF chimera v1.7 (https://www.cgl.ucsf.edu/chimera)(221). Deamination-conducive AID:DNA complexes were defined by the accessibility of dC NH2 substrate group to the catalytic Zn-coordinating and the glutamic acid residues (H56, C87, C90 and E58 in wild type human AID). Complexes in which dC

was bound in the catalytic pocket but improperly orientated, so as not to allow for deamination to proceed, were disregarded.

Expression, purification and enzymatic activity of AID mutants

Wild type human GST-AID was expressed and purified as previously described (52,94). Using models and DNA binding simulations as a guide, we constructed mutants by site-directed mutagenesis as previously described (94,95). To generate chimeric AID variants, EcoRI fragments encoding the AID open reading frame were synthesized (Genscript, USA) and cloned into pGEX-5X-3 (GE Healthcare, USA) to generate expression constructs. SDS-PAGE was used to equalize the amount of purified GST-AID used in experiments. We compared the catalytic activity of each AID variant in comparison to wild type using the alkaline cleavage deamination assay for enzyme kinetics as previously described (52,94,95). In all experiments, 7-nucleotide long bubble substrates containing the WRC motif TGC were used as previously described (49,52). For substrate generation, 2.5 pmol of the WRC-containing target strand was 5'-labeled with [γ-32P] dATP using polynucleotide kinase (NEB, USA), followed by purification through mini-Quick spin DNA columns (Roche, USA) and annealing of 2-fold excess (5 pmol) of the complementary strand. Purified substrate ranging in concentration from 0.15-7.5 nM was incubated with 0.40 μg AID in 100 mM phosphate buffer pH 7.3 in a volume of 10 μL for all reactions. All AID and variants were incubated at their optimal temperature of 37°C, with the exception of Dr-AID which is optimally active at 25°C (94). Enzyme assay incubation periods ranged from several minutes to 3 hours. Inactive mutants were incubated 6-30 hours to in order to distinguish from mutants or variants with minimal activity levels. The correlation plot was constructed using catalytic rate and catalytic pocket accessibilities ranked within our cohort of AID variants.

Results and Discussion

Suitability of resolved APOBEC structures to model various AID conformations

There are 11 human APOBEC enzymes: APOBEC1 (A1), APOBEC2 (A2), APOBEC3A-H (A3A-H), APOBEC4 and AID (201,205). Amongst these, eight partial or complete structures have been described to date. These include the NMR solution structure of full length mouse A2, X-ray crystal structures of either the inner (chain A) or outer (chain B) monomers of the human $A2_{41-224}$ tetramer, NMR solution structure of A3A, X-ray crystal structure of $A3C_{5-190}$, X-ray crystal structure of A3F C-terminal domain (CTD) and the X-ray crystal and NMR solution structures of A3G-CTD (Figure S3; (170,185,187,207,208,210,211). All APOBECs have a conserved core architecture akin to that of human cytidine deaminase, consisting of an α-ß-α Zn^{2+} binding motif, which is H-X-E-$X_{(23-28)}$-P-C-$X_{(2-4)}$-C (173,205,222). This catalytic pocket is characterized by its core 4 or 5-stranded ß-sheet underlying two neighboring α-helices that converge to bring the conserved Zn^{2+} coordinating residues (two cysteines and a histidine) into proximity with the glutamic acid required for deamination catalysis (Figure S3; (95,201,222,223). Amongst the APOBEC enzymes with determined structures, A2, A3A, A3C, A3G-CTD and A3F-CTD exhibit 46.9%, 56.6%, 52.5%, 55.6%, 54.3% residue homology, and 33.5%, and 42.7%, 44.6%, 46.6% and 42.9% residue identity with AID, respectively (Figure 14). All exhibit sufficient homology and spread of identical or homologous residues to serve as appropriate templates for modeling AID. However, various sections of AID's sequence are differentially homologous to each APOBEC. Therefore, considering a single template for modeling AID in its entirety may not provide an accurate picture. Even for the same APOBEC there are often conformational differences between crystal and solution structures (Figure S3; (224,225). Indeed, each structure itself may also include multiple energetically favorable conformations such as the 30

NMR conformations in the A3A NMR ensemble (208). Thus, we reasoned that generating a cohort of AID models from multiple templates may provide insights on the likelihood of different conformations. By then examining how these conformational dynamics influence the putative catalytic pocket of AID, as well as DNA binding, we could gain insights into the structural determinants of its enzymatic activity.

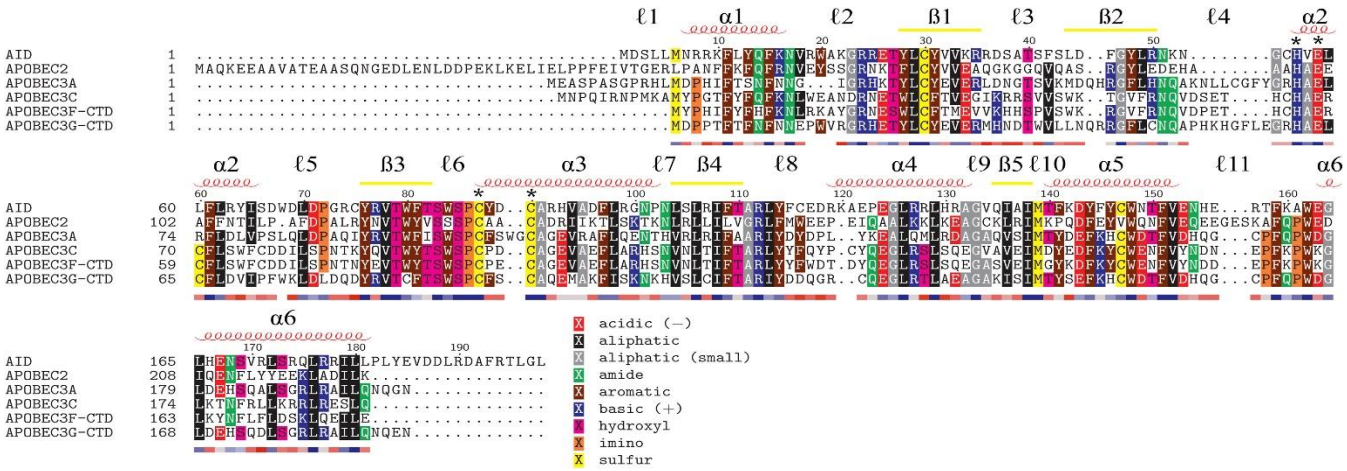


Figure 14. Primary structure alignment of AID and APOBEC modeling templates. AID was aligned with APOBEC2, 3A, 3C, 3F-CTD and 3G-CTD which served as templates for modeling AID (See also figure S1 for APOBEC template structures). The secondary structure of A3G-CTD X-ray (PDB: 3e1uA) was used to denote α-helical, ß-strand and loop regions thus providing an approximation of the secondary structure shared within the family. Every tenth residue is numbered using AID as a reference. * denotes Zn-coordinating and catalytic residues conserved amongst the family. Residues are colored according to side chain chemistry. The red to blue conservation bar under each sequence illustrates differentially homologous portions of each APOBEC, with red and blue representing non-conserved and conserved, respectively.

Quality assessment and general features of AID models

To validate our approach, we focused on the seven APOBEC family member structures that have been resolved and can be used as templates to model AID. We used four (mouse A2, human $A2_{41\text{-}224}$, A3G-CTD X-ray and NMR structures) as templates to model the other three (A3A, A3C and A3F-CTD). We generated 34 model structures and evaluated the fit of each model with its cognate resolved structure by superimposition (Figure S4). For each model:structure pair we used the global Root mean square (RMS) measure of peptide backbone as well as side chain conformation alignments, determined as the average RMS of all residues (Table S1). We found that 25/34 models yielded RMS values < 3 Å indicating acceptable fit. Thus, we concluded that our approach can produce high fidelity models.

For AID modeled based on each of the 8 templates, we examined 4-5 of the most energetically favorable conformations, totaling 39 models (Figure 15). All AID models contain a core structure comprised of a central ß-sheet with 4 or 5 ß strands (dependent upon the ß2 strand conformation) sandwiched between 6 α-helices, connected by intervening loops similar to the core APOBEC structure. However, compared to other APOBECs, AID contains an additional 18 non-homologous residues in the C-terminus (Figure 14) which we had to model *ab initio*. This C-terminal region forms an α-helix (α7) that can assume multiple conformations relative to the core. We assessed overall model quality using both qualitative and quantitative measures. We compared the features of models (Figure 16A, B) against known biochemical properties of AID (48,50,52,55,61). First, all models recapitulate the canonical position of the three Zn-coordinating residues (H56, C87 and C90) "above" a glutamic acid residue (E58), the equivalent of which in all other cytidine/cytosine deaminases serves as the catalytic proton-donating residue for deamination (Figure 16C; (226). Second, this catalytic motif forms an indented pocket on the surface with dimensions and electrostatic properties that allow docking of a

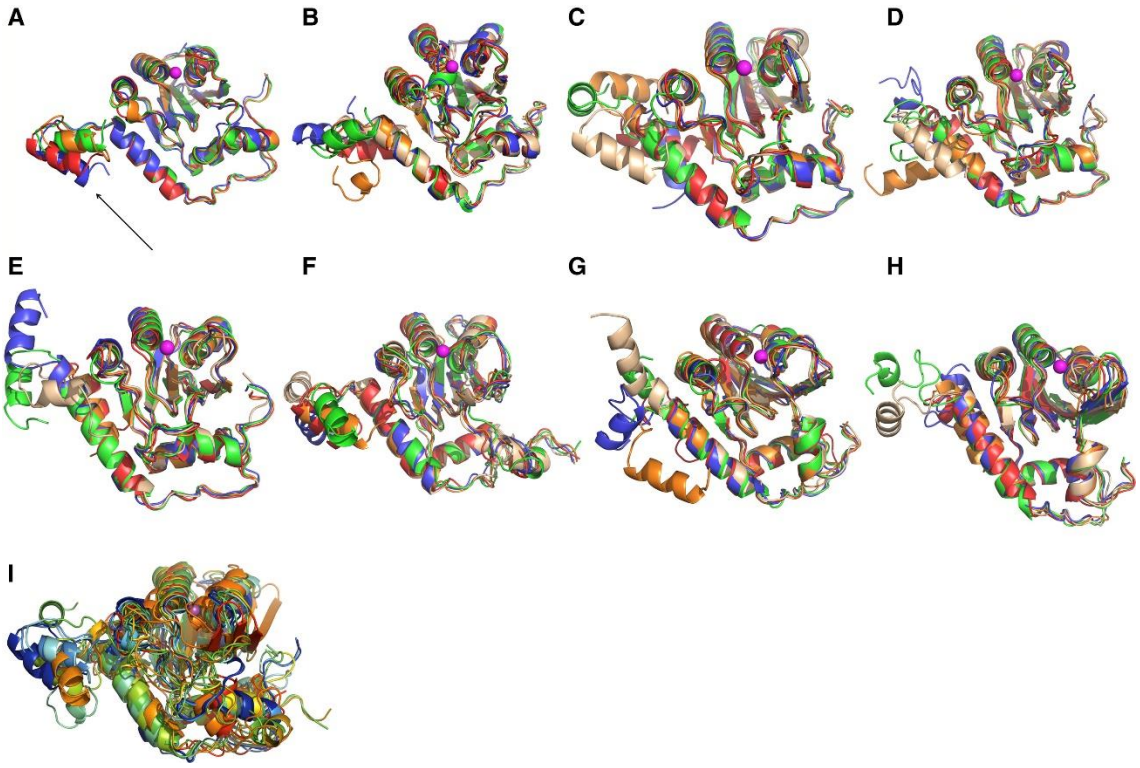


Figure 15: Superimposed AID models from different APOBEC templates highlight regions of conformational flexibility. A. AID models based on Mouse A2 NMR (PDB: 2rpzA), B. AID models based on A3A NMR (PDB: 2m65A), C. AID models based on A3C X-ray (PDB: 3vowA), D. AID models based on A3F-CTD X-ray (PDB: 4iouA), E. AID models based on A3G-CTD X-ray (PDB: 3e1uA), F. AID models based on A3G-CTD NMR (PDB: 2kboA), G. AID models based on $A2_{41\text{-}224}$ chain A (PDB: 2nytA), H. AID models based on $A2_{41\text{-}224}$ chain B (PDB: 2nytB) and I. Superimposition using a representative high probability model derived from each template (see also tables S1 and S2 and figure S4 for validation of our modeling approach involving the APOBEC templates and for description of QMEAN6 and stereochemical qualities of each AID model). Each template yielded 5 low-energy conformations of AID (except mouse A2 which yielded 4), totaling 39 models. For each of the 8 templates, we superimposed the models using a unique color for each model. The purple sphere represents the coordinated zinc in the putative catalytic pocket. The arrow in panel A points to α7 which was modeled *ab initio.* α7 has more rotational freedom and adopts multiple conformations in our models, likely because it is unrestricted on its C-terminal end. In each superimposition, the basic secondary structure is conserved while the connecting loops display increased flexibility.

dC nucleotide in deamination-feasible conformations (Figure 16; (55,94,95). That a viable catalytic pocket is formed by the 3-dimensional convergence of the 4 catalytic residues from throughout the

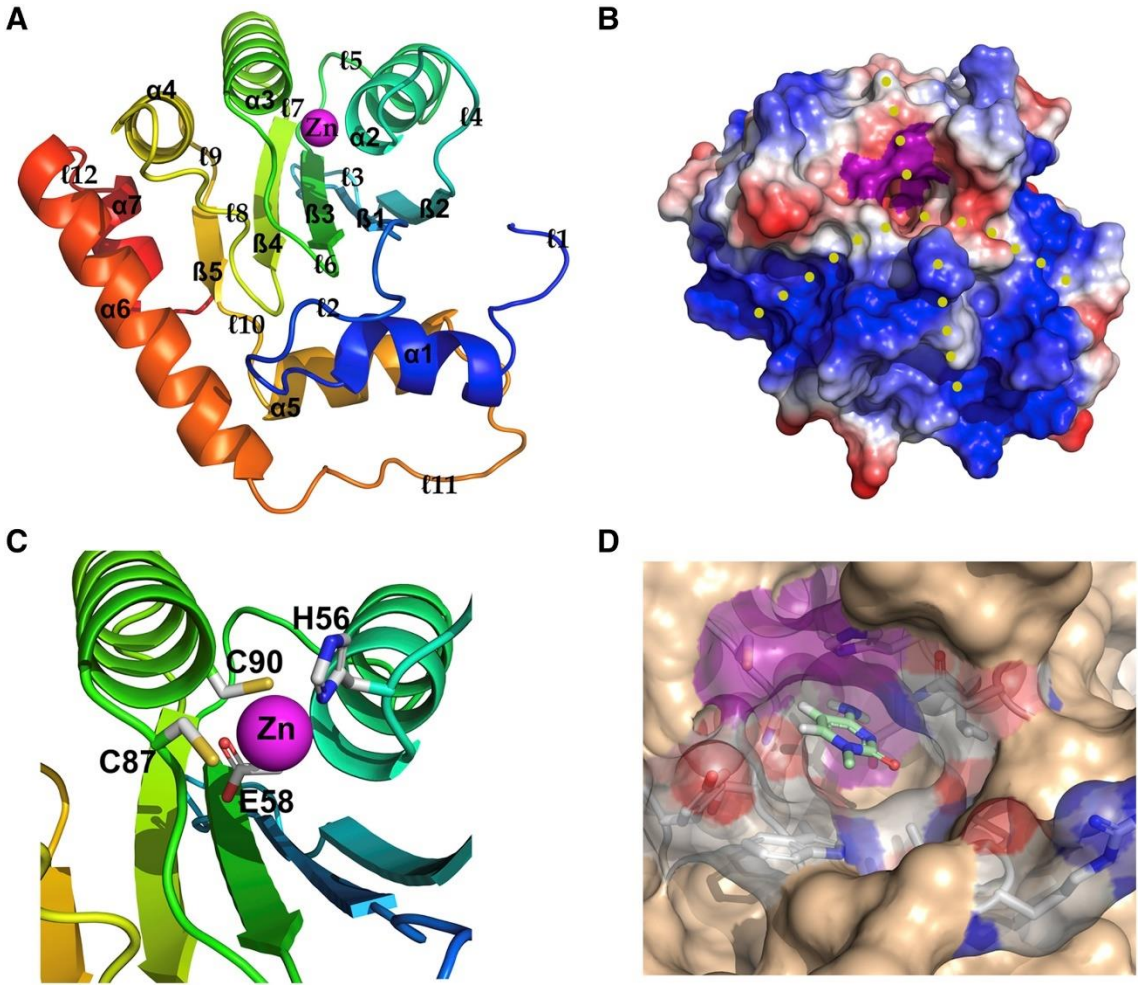


Figure 16: Highly conserved features among models support biochemical characteristics of AID. A. Representative ribbon model of AID. All AID models recapitulate the conserved APOBEC core architecture with the exception of the unique-to-AID C-terminal loop 12 and α7. N- to C-terminus progression is shown from blue to red with catalytic pocket zinc shown in purple. B. General surface topology of AID. Positive, neutral and negative residues are blue, white and red, respectively, with the putative catalytic pocket surface colored in magenta. The surface contains a high proportion of positively charged residues lining two potential ssDNA-binding grooves that pass over the putative catalytic pocket approximately denoted by yellow dots. C. The putative catalytic pocket of AID is characterized by Zn-coordinating (H56, C87 and C90) and catalytic (E58) residues. The catalytic motif is positioned at the N-termini of α2 and α3 above the central ß-sheet. D. dC docked into to a representative catalytic pocket. dC docking was carried out in the context of a 7mer ssDNA substrate, but only the dC is shown. Magenta denotes the surface topology of the putative catalytic residues that form a pocket large enough to accommodate dC and nearby residues that are likely involved in dC stabilization are shown.

model (α2, α3, loop 2, 4, 6 and 8), provides overall confidence (Figure 16A, B, C, D). Third, the surface contains a high proportion of positively charged residues, many of which fall in one of two

putative ssDNA binding grooves passing over the putative catalytic pocket (Figure 16B). These are on average ~6 Å in width, wide enough to accommodate ssDNA. These features are consistent with the ability of AID to bind exclusively ssDNA with high affinity (52,53). To quantitatively assess the models we conducted QMEAN6 Z-score and Ramachandran analysis (Table S2). All AID models showed favorable QMEAN6 Z-scores (-1.79, -1.49, -1.39, -2.27, -1.64, -1.74, -1.72 and -1.19 on average, respectively). In addition, AID models based on mouse A2, $A2_{41\text{-}224}$ chain A and B (X-ray), A3A (NMR),A3C (X-ray), A3F-CTD (X-ray), A3G-CTD (NMR) and A3G-CTD (X-ray) templates contained the majority of residues situated within favored regions of the Ramachandran plots (87.1%, 87.1%, 88.1%, 82.8%, 84.5%, 82.1%, 81.5% and 87.1% respectively) as shown in (Table S2). By thus scrutinizing the models against qualitative and quantitative criteria, we conclude that they are reliable.

Putative catalytic pocket architecture and implications for deamination catalysis

Converging at the base of the putative catalytic pocket, H56 is located at the C-terminus of loop 4, E58 at the N-terminus of α2, C87 at the N-terminus of loop 6 and C90 at the N-terminus of α3 (Figure 16A, C). This putative catalytic pocket organization supports a ordered deamination reaction chemistry that is typical of other cytidine/cytosine superfamily deaminases (173,192). After Zn^{2+} binds an activated water molecule, thus forming the Zn-OH nucleophile, AID is primed for deamination (Figure S3H) through an ordered reaction mechanism as previously suggested (95,205). Besides these Zn-coordinating and catalytic residues, we note secondary catalytic residues that compose the "walls" and "floor" of the catalytic pocket. These do not directly participate in deamination, but are predicted by the models to contact and/or stabilize the substrate dC in several conformations through hydrogen bonding, electrostatic interactions and aromatic base stacking. Therefore, these residues are likely to

play crucial roles in the catalytic activity of AID. By examining all AID:ssDNA docked complexes in which dC was in the putative catalytic pocket, we identified 21 potential secondary catalytic residues (Figure 17A). Amongst these, T27, N51 and W84 were observed in 63-75% of models. R25, V57, S85, P86 and Y114 were observed in 25-50% of models. R24, E26, K52, F115, G23, L29, N53, G54, C55, T82, D89, C116 and E122 were observed in 6-18% of the models. These residues are located predominantly on loops that form the catalytic pocket. G23, R24, R25, E26, T27 and L29 are in loop 2. N51, K52, G54, C55 and V57 are in loop 4. T82, W84, S85 and P86 reside in loop 6. Y114, F115 and E122 are in loop 8. D89 is in α3 and E122 is in α4.

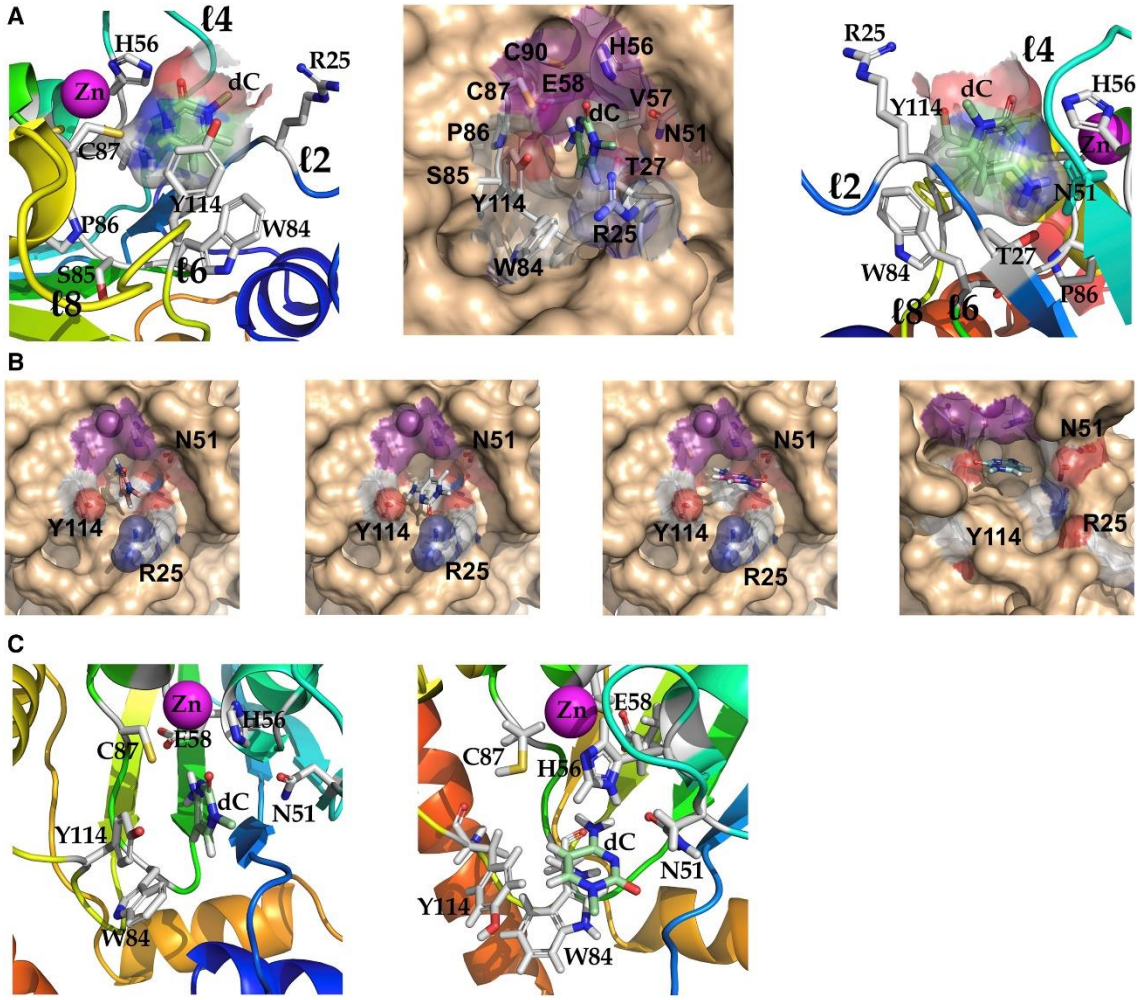


Figure 17. Secondary catalytic residues regulate the chemical microenvironment of the catalytic pocket. A. Secondary catalytic residues from loop 2 (R25, T27), loop 4 (N51, V57), loop 6 (W84, S85, P86) and loop 8 (Y114) that were found to be proximal to docked dC in the majority of AID:ssDNA complexes, at three viewing angles (left, middle and right panels). In all docks, the carbonyl of dC is hydrogen bonded by the amide group of N51 or the hydroxyl group of Y114. The docked dC carbons are colored green. B. AID:ssDNA complexes illustrating four deamination-supporting orientations of dC in the catalytic pocket (from left to right: perpendicular-1, perpendicular-2, planar-1 and planar-2). Together, perpendicular-2 and planar-2 orientations comprise the majority of dC-bound complexes. In all orientations dC carbonyl is stabilized by hydrogen bonds with one or more conserved residues lining the pocket. C. Y114 and W84 stabilize dC through base stacking and hydrophobic contacts. In the majority of conformations with dC bound in the catalytic pocket, W84 is located lying planar relative to the floor of the catalytic pocket, base stacking dC (left panel). In some conformations, Y114 adopts a side-on conformation, where it base stacks with W84 (right panel). In either case, the C_5 and C_6 of dC are in contact with W84 and/or Y114. Together, W84 and Y114 appear to stabilize the aromatic ring of dC and shield its non-polar regions from the largely polar microenvironment of the catalytic pocket.

Using the catalytic and secondary catalytic residues as reference, we examined deamination-feasible conformations of docked dC in the catalytic pocket (Figure 17B). Each complex was defined by the directionality of the dC carbonyl group: Planar 1 (carbonyl facing N51), Planar 2 (carbonyl facing away from N51), Perpendicular 1 (carbonyl facing away from the floor of the catalytic pocket) and Perpendicular 2 (carbonyl facing towards the floor of the catalytic pocket). Planar 1 and perpendicular 2 conformations represent 50% (8/16) and 25% (4/16) of clusters, respectively, whilst planar 2 and perpendicular 2 each represent 2/16 (12.5%) of docks. Thus planar 1 represents the most energetically favorable dC conformation. We found that in 63% of dC docked complexes (including all Planar 1 and perpendicular 1 conformations) the dC carbonyl group was stabilized by hydrogen bonding with the amide group of N51. We noted that first, mutation of the N51 equivalent residue (N244A in A3G, N57A in A3A) abolished deamination (185,227), second, the equivalent residue in Staphylococcus aureus tRNA adenosine deaminase TadA (N42) has been shown to stabilize its substrate in the catalytic pocket by hydrogen bonding (192) and third, N51 is 100% conserved amongst all active APOBECs (Figure 17). In the remaining 37% of conformations (planar 2 and perpendicular

2 conformations), the dC carbonyl was stabilized by a combination of polar groups in S85, C116, R25, Y114 and T27 (Figure 17A, B). In addition to polar interactions with dC carbonyl, we also noted non-polar interactions and base stacking of the pyrimidine ring. Two aromatic residues, W84 (loop 6) and Y114 (loop 8), are located on the floor and walls of the catalytic pocket, respectively (Figure 17C). In 63% and 50% of docks, the aromatic ring of dC was base stacked with the aromatic rings of W84 or Y114, respectively. In this manner, W84 and Y114 may stabilize the non-polar regions of dC and shield these from the otherwise polar microenvironment of the pocket. Recently, we and others have discussed the importance of Y114 in the context of the catalytic pocket of AID (95,183). Of the secondary catalytic residues, Y114 adopts the most diverse range of conformations in the context of its angular attitude within the catalytic pocket of the AID models. Indeed, it is partially a consequence of these variations that dC is able to dock in several planar and perpendicular conformations as described above (figure 17B). Thus, whilst the primary catalytic residues are responsible for deamination, it is likely that the makeup and arrangement of secondary residues impact deamination efficiency through dC stabilization and modulating the chemical properties of the catalytic pocket.

Catalytic pocket accessibility of AID is a key determinant of its catalytic rate

Deamination is not possible if dC cannot access the catalytic pocket. To examine the fit of dC in the catalytic pocket, we simulated a 7 nucleotide long ssDNA containing either a WRC motif (TGC) or a control non dC-containing ssDNA (TGT). We found that only 25.6% (10/39) of models bound dC in the catalytic pocket in conformations that allowed access to Zn and E58 (figure 18A). Bound conformations exhibited catalytic pockets spacious enough (average dimensions of pocket 5.6 x 4.9 x 6.7 Å) to allow dC (dimensions 6.4 x 4.3 x 6.6 Å) access to Zn and E58 (Figure 18A). 25.6% (10/39) of models formed a partially open catalytic pocket (average dimensions 4.1 x 3.6 x 7.3 Å) whilst the

remaining 49% (19/39) of models exhibited nearly or fully collapsed catalytic pockets with no apparent surface pocket (Figure 18A, B). As expected, we found dC docked only in catalytic

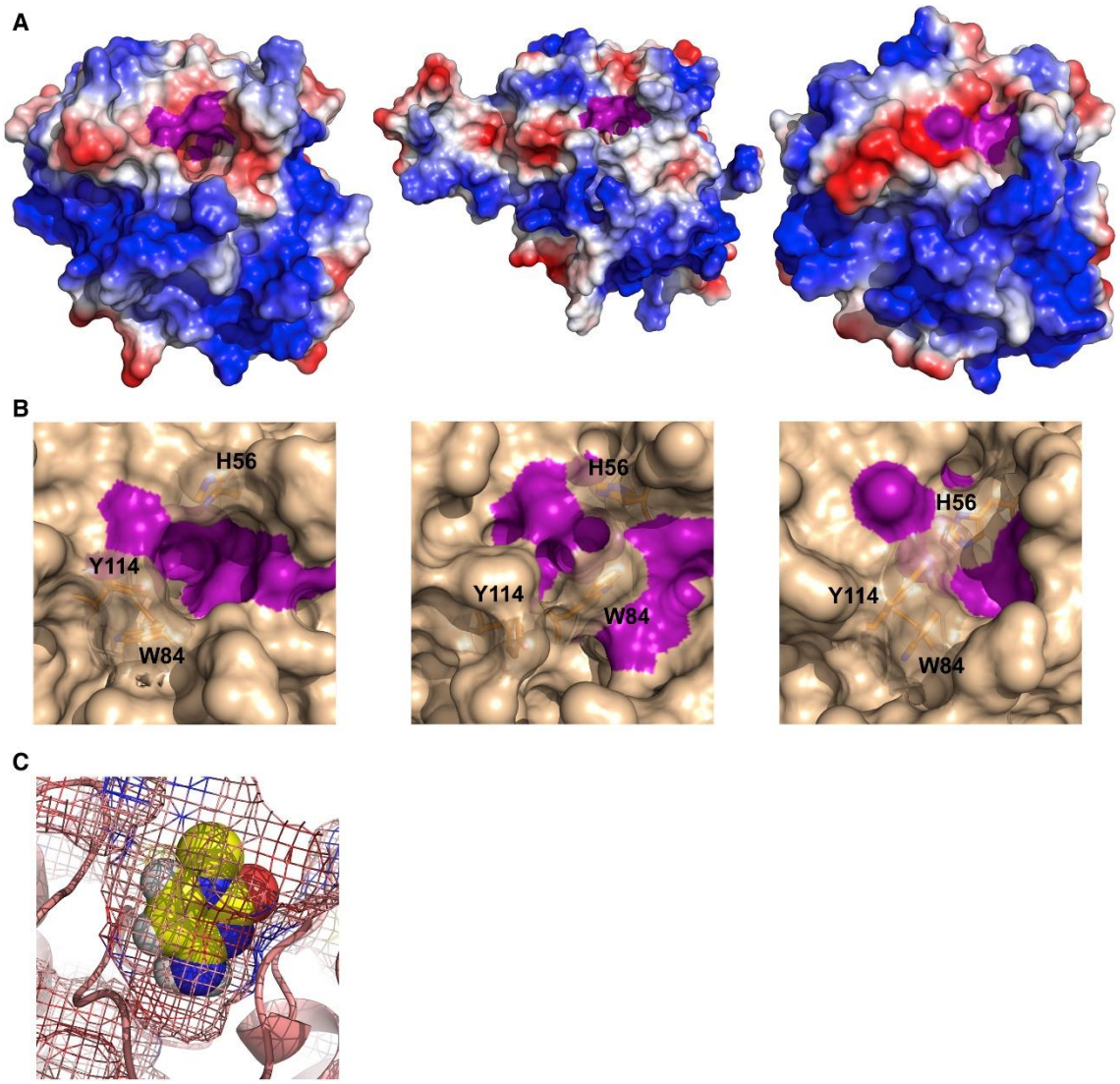


Figure 18. The catalytic pocket may assume conformations that impact accessibility. A. General surface topology of AID. Positive, neutral and negative residues are shown as blue, white and red, respectively. The surface of the putative catalytic pocket surface is shown partially in magenta. Left panel shows a fully accessible catalytic pocket able to accommodate dC and support deamination; however, in the majority of conformations, the catalytic pocket is partially (middle) or fully (right) restricted, unable to bind dC in energetically favorable conformations that allowed for deamination. B. Zoomed in surface topology of open (left), partially open (middle) and closed (right) catalytic pockets. Transparent stick format of secondary catalytic residue side chains is used to show that the catalytic pocket is collapsed due to the protrusion of one or more residue side chains, restricting dC access to catalytic residues. C. In catalytically viable conformations, dC (spheres indicate approximate Van der Waals radii) is located in the catalytic pocket forming a tight fit. The inner surface of the catalytic pocket is traced by a mesh outline.

pockets of models with accessible pockets in conformations that were energetically favorable and would allow for deamination. As evident by the average dimensions of accessible pockets and dC, even amongst these the substrate fit was taut (Figure 18C). These observations led us to hypothesize that the catalytic pocket of AID is at best marginally accessible to dC, thus explaining the low catalytic rate of AID. This hypothesis is founded on the premise that the relative proportion of time the catalytic pocket may exist in an accessible *vs.* occluded conformation is a determinant of AID's catalytic efficiency.

If this is the case, it would follow that altering accessibility states of AID's catalytic pocket ought to influence catalytic rates accordingly. Thus, we sought to design AID mutants or chimeric variants that would influence catalytic pocket accessibility. To this end, we examined the portions of AID that are likely to affect accessibility dynamics of the catalytic pocket (Figure 19A and B). Given that the basic arrangement of Zn-coordinating and catalytic residues is identical amongst APOBECs, it would stand to reason that one of the structural differences that mediates variations in catalytic robustness of APOBECs may lie in the makeup and positioning of secondary catalytic residues. As noted in the preceding section, the majority of secondary catalytic residues that make up the floor and walls of the catalytic pocket are contained in loops 2, 4, 6 and 8. Due to the extreme conservation of loop 6 (figure 14) and transplantation of loop 8 (substrate specificity loop) having no affect on the catalytic rate of AID (58,59), it is less likely that these regions alter catalytic pocket restriction. In contrast, loop 2, loop 4 and α3 all converge around the catalytic pocket and likely control access to the catalytic pocket. We noted significant variability in the length and amino acid composition of these loops amongst various APOBECs, and as a consequence, a high degree of conformational variability

between resolved APOBEC structures, and within our cohort of AID models (Figure 15I, Figure S3).

It thus appeared that differences in length and composition of these loops amongst APOBECs may

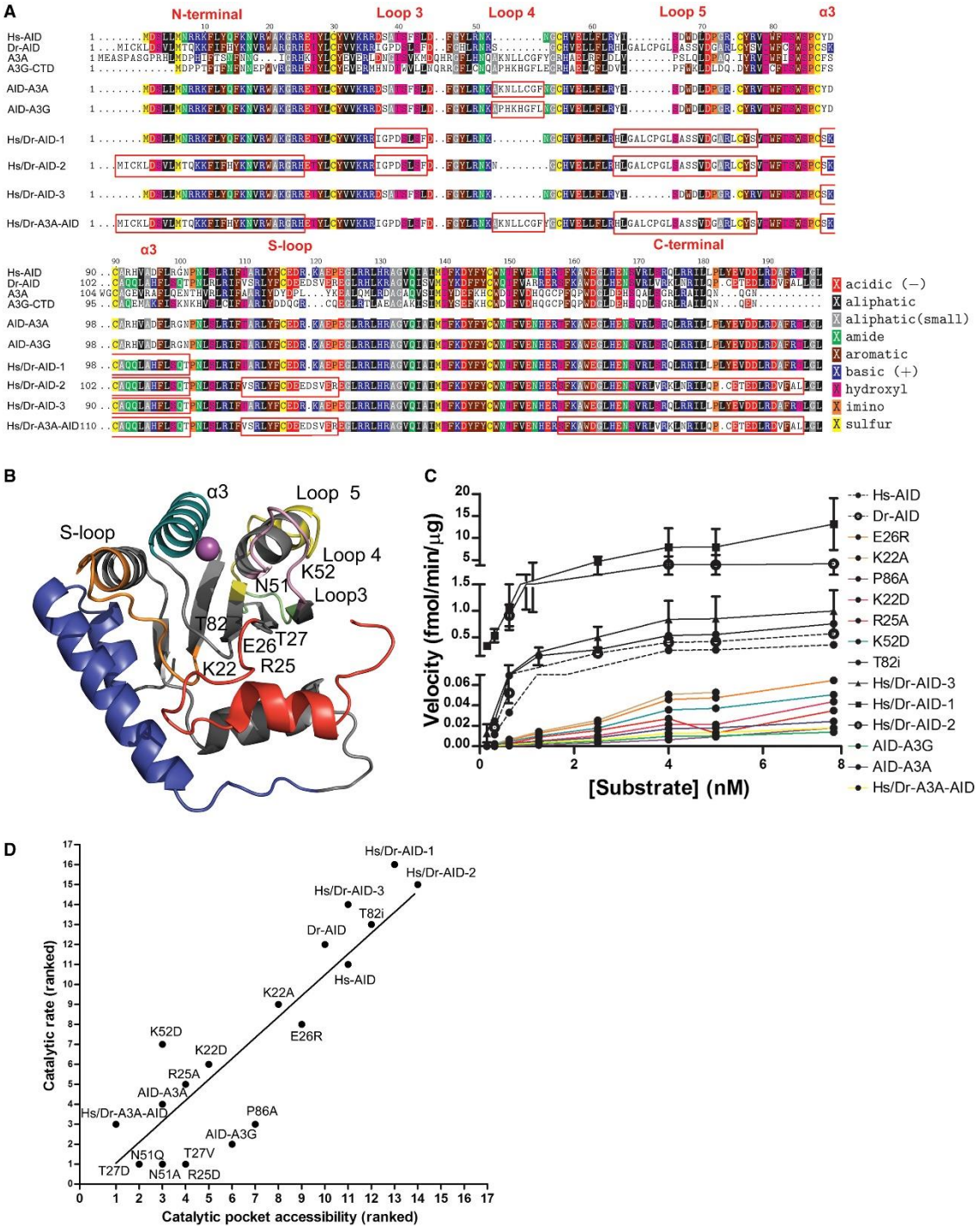


Figure 19: Modulation of catalytic pocket accessibility dynamics and consequences for catalysis. A. Sequence alignment of Hs-AID, Dr-AID, APOBEC3A, APOBEC3G-CTD and AID chimeras predicted to alter accessibility to the catalytic pocket. All AID chimeras were derived from the

human AID scaffold, with regions from Dr-AID, A3A or A3G-CTD transplanted as indicated by rectangles in each primary sequence. B. General structural regions of wild type human AID illustrating regions transplanted and mutated by color. N- and C-terminal are colored red and blue, respectively. C. Enzyme kinetic analysis of initial deamination velocities of AID variants. Substrate concentrations ranged from 0.15-7.5 nM, with incubation times ranging from several minutes to 3 hours. High activity variants are shown using solid black lines, wild type Hs- and Dr-AID are shown with dashed lines and low activity variants are shown with colored lines. Enzymatically inactive mutants (R25D, T27V, T27D, N51A, N51Q, H56D and C90F) were incubated for 6 hours or more to discriminate from low activity mutants. Lines are not shown for these inactive mutants. D. Correlation between catalytic rate and catalytic pocket accessibility of wild type AID and AID variants (see also table S3 for full quantitative analysis of catalytic pocket accessibility predictions). Frequencies of catalytic pocket accessibilities, as defined by the ability to dock dC in the catalytic pockets, were ranked within the cohort of AID variants and ordered from lowest on the left to highest on the right, on the x-axis. Maximum deamination velocities as ascertained by enzyme kinetics were ranked within the cohort and ordered from lowest on the bottom to highest on top on the y-axis. (Spearman correlation, R^2= 0.82, P<0.0001). Error bars indicate SEM.

underlie their degree of movement restriction, which would in turn impact catalytic pocket accessibility through modulating positions and/or angles of the secondary catalytic residues that make up its walls and floor.

With this rationale, we constructed 19 AID mutants or chimeric variants bearing alterations in loop 2, loop 4 and α3 (Figure 19A and B). As we had done for wild type AID, we examined the accessibility of the catalytic pockets in a cohort of models for each AID variant (756 models which generated 24,192 AID-DNA complexes, in total). We then compared the fraction of models with accessible catalytic pockets that were able to dock dC between each variant and wild type AID (Table S3). In parallel, we compared the maximal measured catalytic rate of each variant to wild type AID (Figure 19C). First, we examined mutants of residues in loops 2 and 4 which compose the catalytic pocket walls and floor and should modulate catalytic activity. We found that mutations in loop 2 (K22D, R25A, R25D, T27V and T27D) and loop 4 (N51A and N51Q and K52D) resulted in 60-80% fewer conformations able to dock dC in the catalytic pocket (Table S2). Accordingly, these mutants had lower catalytic rates compared to wild type AID, or were inactive (figure 19C). T82I was the only

mutant to have slightly more conformations able to dock dC relative to wild type AID. Accordingly, it has been shown to be slightly more active than human AID (62).

We then investigated the role of loop 4, in its entirety, in modulating catalytic rate through catalytic pocket accessibility. Relative to AID, A3A and A3G have a loop 4 that is 8 residues longer (figure 19A). Because mutation of loop 4 residues modulates catalytic rate, we wondered whether integration of this region from other APOBECs into an AID scaffold would alter the breathing dynamics of the catalytic pocket, and its activity accordingly. To this end, we constructed AID-A3A and AID-A3G incorporating loop 4 from A3A and A3G into an AID scaffold (Figure 19A). Simulations revealed that AID-A3G and AID-A3A had 50% and 70% fewer conformations able to dock dC than wild type AID, respectively (Table S3). Indeed, we found that both chimeras had severely reduced activity (figure 19C).

Next we sought to examine AID chimeras with α3 region variations. We previously reported that Zebrafish AID (Dr-AID) was more catalytically robust than human AID (Hs-AID; (94,95). We reasoned that one way to increase the catalytic rate of human AID may be to incorporate multiple regions of Dr-AID into a human AID scaffold. To this end, we constructed three chimeras, which incorporated α3 and adjacent loops from Dr-AID into a human AID scaffold (Figure 19A). We found that two chimeras (Hs/Dr-AID-1 and Hs/Dr-AID-2) exhibited higher (17-37% increase) conformations able to dock dC when compared to wild type human AID, whilst the third chimera (Hs/Dr-AID-3) which incorporates only the α3 of Dr-AID into the human AID scaffold, exhibited a more modest (7%) increase in dC accessible catalytic pockets (Table S3). Correspondingly, Hs/Dr-AID-1 and Hs/Dr-AID-2 had dramatically increased catalytic rates when compared to human AID whilst Hs/Dr-AID-3 had a more modest increase in catalytic rate (figure 19C). To further examine the role of catalytic pocket accessibility on activity, we generated chimera Hs/Dr-A3A-AID which incorporates loop 4

from A3A into Hs/Dr-AID-2 (Figure 19A). As discussed above, the transplantation of A3A loop 4 into AID resulted in diminished pocket accessibility and catalysis. This loop transplant mediated a similar effect here as evidenced by our finding that Hs/Dr/A3A-AID had 90% fewer conformations bound to dC than human AID (Table S3). Correspondingly, we found that Hs/Dr/A3A-AID had approximately 5% of the activity of human AID (Figure 19C). Beyond AID variants, a secondary prediction of our hypothesis would be that differences in catalytic pocket accessibility may also explain the different catalytic rates amongst various APOBEC enzymes. In support of this notion, we note that the experimentally determined NMR solution structure of A3A indeed includes conformations with open as well as occluded catalytic pockets (209). Moreover, in contrast to AID, a larger proportion of A3A conformations exhibit a fully open catalytic pocket (67% *vs.* 25.6%). This observation is notable given that A3A is a catalytically robust enzyme (228,229).

As summarized in Figure 19D, when ranked according to catalytic pocket accessibility *vs.* catalytic rate, we observed a positive correlation between these two parameters. Interestingly, this relationship does not appear to be linear, as relatively modest changes in pocket accessibility can result in an order of magnitude change in catalytic rate (ie: Hs/Dr-AID-1 and Hs/Dr-AID-2). Conversely, in some cases (ie: Dr-AID) relatively modest changes in catalytic rate are not associated with significant changes in catalytic pocket accessibility. In this context, it is crucial to note that our hypothesis that catalytic site accessibility is an important determinant of catalytic rate does not preclude other factors from influencing catalytic rate. For instance, N51Q introduces a bulkier side-chain that protrudes into the catalytic pocket thereby reducing the size of the catalytic pocket and abolishing activity. On the other hand, a different mutation of the same residue (N51A) does not impact the physical size of the pocket, yet it also reduces the ability to dock dC and abolishes activity (Table S3, Figure 19D). As discussed above, N51 is absolutely conserved in closely and distantly related active deaminases, and

as predicted by our models, it likely plays a crucial role in dC stabilization through hydrogen bonding of the dC carbonyl group with its amide side chain. Thus, unlike N51Q, the effect of N51A on catalysis cannot be explained by the physical occlusion of the catalytic pocket, but rather by the lack of the amide group thus perturbing the chemistry of the catalytic pocket required to stabilize dC. Taken together, these results support the model that catalytic pocket accessibility is a rate limiting step in deamination catalysis by AID. The data also further corroborate the validity of our modeling and docking approaches by demonstrating their ability to predict dynamic changes in catalytic pocket states that are in line with enzyme activity data.

The role of DNA binding in deamination catalysis

Before deamination catalysis, ssDNA ought to be bound on the surface of AID in a manner that positions dC in the catalytic pocket. As mentioned above, we observed 2 putative ssDNA binding grooves passing over the catalytic pocket (Figure 16B, 20A). In AID:ssDNA docking simulations, we identified 50 surface residues with probable contact with ssDNA when dC is bound in the catalytic pocket (Table S4). The majority of contacts consisted of negatively charged phosphodiester backbone interactions with positively charged arginines or lysines. We observed ssDNA bound on multiple surface locations in sporadic configurations. This is not surprising given the positively charged surface. Even when we restricted the docking search space to the surface area proximal to the catalytic pocket, we observed that approximately only half of the simulations (58% or 728/1248 binding clusters) resulted in binding ssDNA in the putative grooves (Figure 20B). Of these, 38% (276/728), 31% (223/728) and 31% (229/72) were ssDNA was bound entirely in groove 1 similar to A3A:ssDNA molecular docks(227), groove 2, or a combination of regions from both grooves, respectively. To assess the proportion of complexes containing dC bound in the catalytic pocket we restricted our analysis to the 320 ssDNA docking clusters involving the 10 (out of 39) models with an accessible

catalytic pocket. In this cohort, we observed that 82.8% (260/320) of complexes have a vacant catalytic pocket and only 18.8% (60/320) contained a nucleobase in the catalytic pocket (Figure 20C).

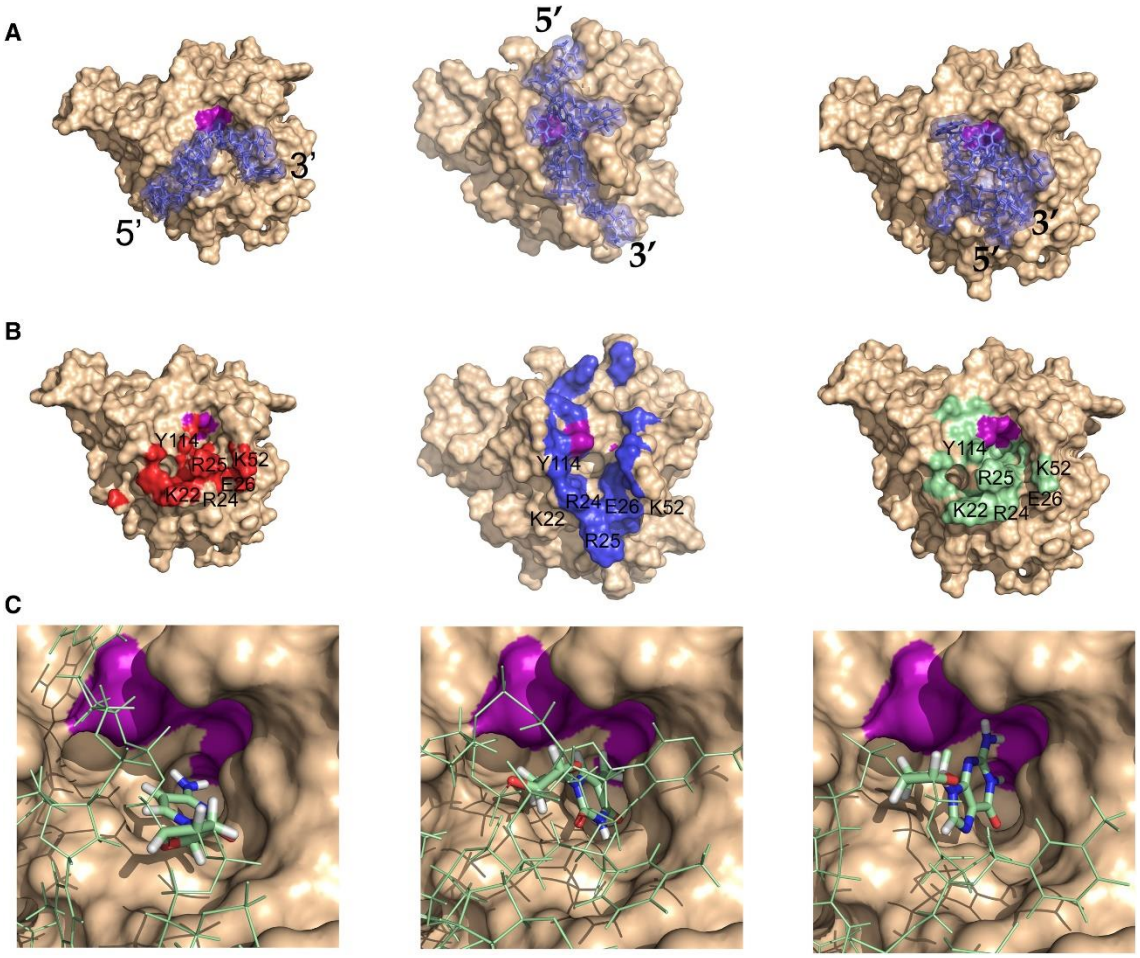


Figure 20. Docking of ssDNA on the surface of AID illustrating sporadic DNA binding and base competition for the catalytic pocket. A. Considering only conformations with accessible catalytic pockets, we docked a 7mer ssDNA containing the WRC motif TGC on the surface of AID and selected for conformations in which dC was bound to the catalytic pocket (purple). AID:ssDNA complexes illustrating ssDNA topology (blue) relative to the surface of AID bound to groove 1 (left), groove 2 (middle), or a more curved conformation involving portions of grooves 1 and 2 (right). In each case, the majority of interactions consisted of electrostatic binding between the negatively charged phosphodiester backbone of ssDNA and the positively charged residues on the surface of AID. B. Groove 1 binding residues are shown in red (Left), groove 2 binding residues in blue (middle) and a third conformation, involving binding residues from both groove 1 and 2 is shown in green (right). Residues that frequently contact ssDNA through phosphodiester backbone or base interactions are labeled (see also table S4 for a description of all ssDNA-contacting residues, contact probabilities and chemical nature of contacts). C. Besides dC in the catalytic pocket (left panel), we noted full or partial entry of other bases. We noted an equal number of T_{-2} (from TG<u>C</u>) bound (middle) and a small

proportion AID:ssDNA complexes that showed a portion of G_{-1} in the catalytic pocket (right). The ssDNA backbone is shown in green.

Amongst these, approximately a third (5%) of catalytic pockets contained dC in deamination-conducive conformations whilst the rest contained other bases. Considering all states of the catalytic pocket regardless of dC accessibility, our modeling and docking results lead us to estimate that 1.3% of all AID-DNA interactions have the potential to result in cytidine deamination. These calculations derived solely from our *in silico* observations are in good agreement with a recent study that calculated deamination-conducive ssDNA binding events by AID in solution, derived from enzyme kinetic assays(230). This study suggested that 0.7-8.0% of AID-DNA interactions result in dC deamination. Our data suggest that the actual scenario may lie closer to the lower end of this range.

In conclusion, we found several mechanisms that govern the mutagenic activity of AID. Regardless of catalytic pocket accessibility, all models were capable of binding ssDNA in or outside the putative ssDNA binding grooves. Even considering the minor fraction of AID:ssDNA complexes where ssDNA passes directly over accessible catalytic pockets, dC was not positioned in the catalytic pocket in the majority of ssDNA binding events. In the context of our suggestion here that catalytically inactive conformations likely represent the predominant breathing form of AID, it stands to reason that binding of ssDNA to inactive conformations of AID or on surface locations that do not pass over the catalytic pocket and/or do not position a dC into the pocket, limit deamination catalysis. Our data suggest that conformational breathing restriction of its catalytic pocket as well as frequent catalytically non-viable ssDNA binding are mechanisms inherent to AID that collectively limit its activity.

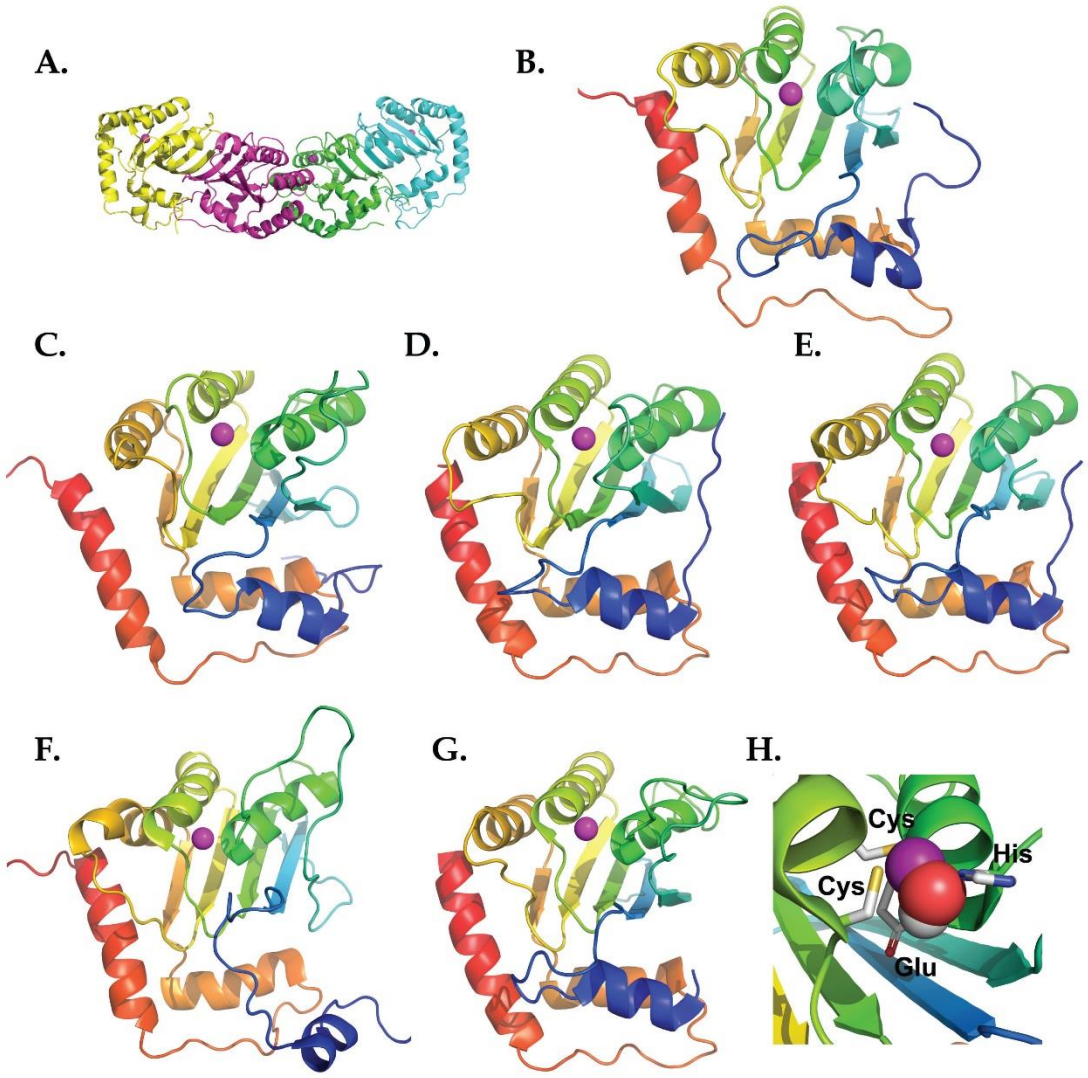


Figure S3. Ribbon structures of APOBEC enzymes that have been resolved to date. A. Tetrameric A241-224 X-ray (PDB: 2nyt), B. mouse A2 NMR (PDB:2rpzA), C. A3A NMR (PDB: 2m65A), D. A3C5-190 X-ray (PDB: 3vowA), E. A3F-CTD X-ray (PDB: 4iouA), F. A3G-CTD NMR (PDB: 2kboA) and G. A3G-CTD X-ray (PDB:3e1uA). N- to C-terminus progression is shown from blue to red, with the active site zinc denoted by the purple sphere. H. The conserved APOBEC active site, illustrating the ZnOH in an active site primed for deamination and the three Zn-coordinating residues (Cys, Cys and His) and the catalytic glutamic acid (Glu).

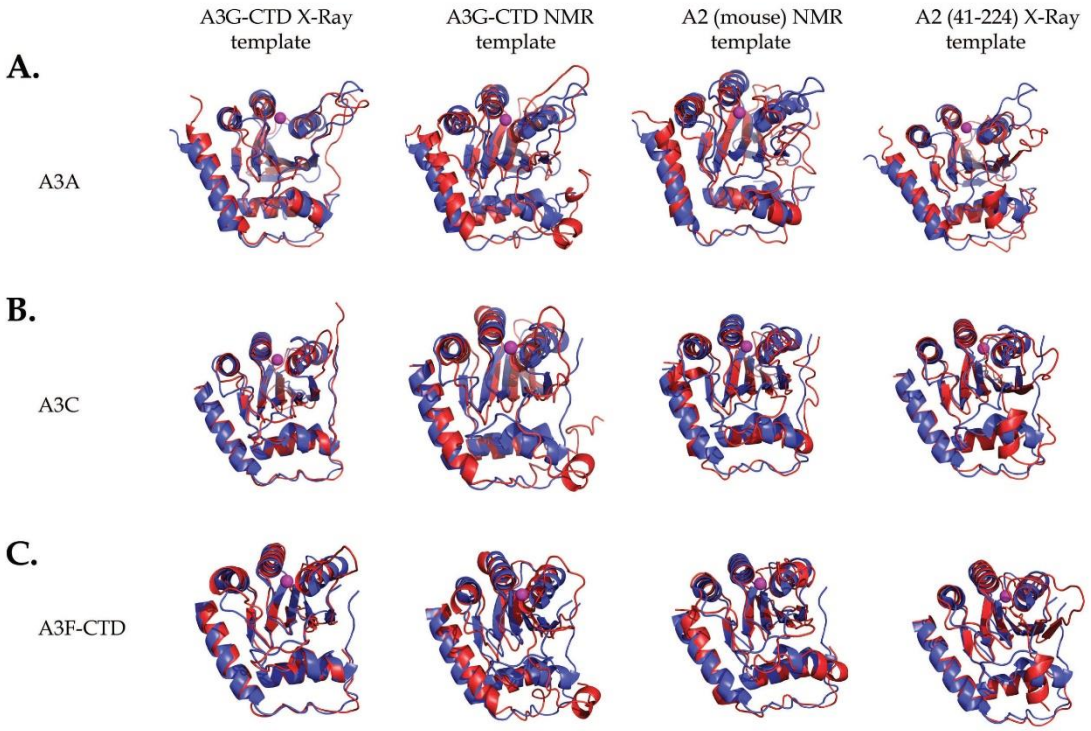


Figure S4. Validating the modeling approach by comparing models of several APOBEC enzymes derived from other APOBEC templates to their experimentally determined structures. We used five templates (A3G-CTD X-ray, A3G-CTD NMR, mouse A2 NMR and A241-224 X-ray chain A/B) to construct 1-5 models, each, of A3A, A3C and A3F-CTD, totaling 34 models. Next we superimposed each model with the experimentally determined structure and compared Root-mean square deviations of atomic positions. Each superimposition consists of the determined structure (blue) with a representative model (red) from each template. A. A3A model:structure superimpositions, B. A3C model:structure superimpositions, and C. A3F-CTD model:structure superimpositions. In all superimpositions, the regions of highest structural agreement were localized to the secondary structural regions, whilst the loop regions showed greater deviation. Despite homology differences in some of the loop regions, the conformation of 8/11 loop regions were well matched between models and structures.

Table S1: Validating the modeling approach by comparing models of several APOBEC enzymes derived from other APOBEC templates to their experimentally determined structures. We used five templates (A3G-CTD X-ray, A3G-CTD NMR, mouse A2 NMR and A241-224 X-ray chain A/B) to construct 1-5 models, each, of A3A, A3C and A3F-CTD, totaling 34 models. Next we used the global Root mean square (RMS) measure of peptide backbone as well as side chain conformation alignments, determined as the average RMS of all residues, for each model:structure pair. 74% (25/34) of model:structure pairs exhibit RMS values < 3Å, illustrating the quantitative accuracy of modeling multiple APOBEC enzymes based on other APOBECs with varying degree of homology.

A3G-CTD NMR (2kboA) template	**RMS deviation from A3A structure (Å)**	**RMS deviation from A3C structure (Å)**	**RMS deviation from A3F-CTD structure (Å)**
Model 1	3.54	2.757	2.741
Model 2	3.64	3.417	N/A
Model 3	3.52	3.34	N/A
Model 4	4.02	3.46	N/A
Model 5	3.66	3.153	N/A
$A2_{41-224}$ (2nytA) template			
Model 1	2.84	1.853	1.545
Model 2	2.38	N/A	N/A
Model 3	N/A	N/A	N/A
Model 4	N/A	N/A	N/A
Model 5	N/A	N/A	N/A
$A2_{41-224}$ (2nytB) template			
Model 1	2.71	1.76	1.65
Model 2	2.56	N/A	N/A
Model 3	2.67	N/A	N/A
Model 4	N/A	N/A	N/A
Model 5	N/A	N/A	N/A
$A2_{mouse}$ (2rpzA) template			
Model 1	2.58	2.36	2.16
Model 2	N/A	N/A	N/A
Model 3	N/A	N/A	N/A
Model 4	N/A	N/A	N/A
Model 5	N/A	N/A	N/A
A3G-CTD Xray (3e1uA) template			
Model 1	2.26	1.71	1.58
Model 2	2.26	1.35	N/A
Model 3	2.20	1.57	N/A
Model 4	2.51	2.02	N/A
Model 5	2.18	1.25	N/A

Table S2: Quantitative assessment of the overall quality of each AID model. 4-5 AID models were constructed based on 8 APOBEC templates: mouse A2 (NMR), A241-224 chain A and B (X-ray), A3A (NMR), A3C (X-ray), A3F-CTD (X-ray), A3G-CTD (NMR) and A3G-CTD (X-ray), totaling 39 models. For each model, we compared the QMEAN6 Z-score and PROCHECK analysis scores to evaluate overall model reliability and stereochemical quality of all residues, respectively. The majority of models had an average QMEAN6 Z-score > -2, indicating a high degree of general reliability. All models showed ~85% of residues in the most favored regions of the Ramachandran plot, indicating good stereochemical fit.

AID [A2 mouse NMR]	QMEAN6 Z-score	Residues in most favorable regions (%)
Model 1	-1.49	87.4
Model 2	-1.95	87.4
Model 3	-1.79	85.2
Model 4	-1.94	88.5
Model 5	N/A	N/A
Average	-1.79	87.1
AID [$A2_{41-224}$ Chain A X-ray]		
Model 1	-1.41	89.0
Model 2	-0.92	91.2
Model 3	-1.83	82.4
Model 4	-1.74	86.3
Model 5	-1.53	86.8
Average	-1.49	87.1
AID [$A2_{41-224}$ Chain B X-ray]		
Model 1	-1.70	87.9
Model 2	-1.83	87.4
Model 3	-1.69	84.1
Model 4	-1.02	90.1
Model 5	-0.73	91.2
Average	-1.39	88.1
AID [A3A NMR]		
Model 1	-1.83	83.5
Model 2	-2.45	80.2
Model 3	-2.61	82.4
Model 4	-2.36	82.4
Model 5	-2.28	85.7
Average	-2.27	82.8
AID [A3C X-ray]		
Model 1	-1.63	85.7
Model 2	-1.51	83.5
Model 3	-2.00	84.6
Model 4	-1.50	85.2
Model 5	-1.57	83.5
Average	-1.64	84.5
AID [A3F-CTD X-ray]		
Model 1	-1.59	86.3
Model 2	-2.00	80.2
Model 3	-1.88	79.7
Model 4	-1.86	81.9
Model 5	-1.36	82.4
Average	-1.74	82.1
AID [A3G-CTD NMR]		
Model 1	-1.49	83.0
Model 2	-2.70	80.8
Model 3	-1.68	82.4
Model 4	-1.46	78.0
Model 5	-1.27	83.5
Average	-1.72	81.5
AID [A3G-CTD X-ray]		
Model 1	-1.55	87.9
Model 2	-0.93	84.6
Model 3	-0.51	89.0
Model 4	-1.24	87.9
Model 5	-1.71	86.3

Table S3: Accessibility of the catalytic pocket in AID and AID variants. For each AID mutant and chimera, a single-stranded dC-containing substrate was docked with each conformation (39-40 conformations each). For each conformation, 32 low-energy AID:DNA complexes were examined totaling 1280 complexes. The docked complexes were examined, and the frequency of conformations with accessible catalytic pockets, as defined by percentage of complexes able to accommodate dC in the catalytic pocket in deamination-conducive conformations was tabulated.

AID variant	% Conformations with dC docked in catalytic pocket	% AID-DNA Complexes with dC docked in catalytic pocket
Chimera-10	35 (14/40)	2.3 (29/1280)
Chimera-8	30 (12/40)	2.0 (26/1280)
T82i	27.5 (11/40)	1.3 (17/1280)
Chimera-3	27.5 (11/40)	1.25 (16/1280)
h-AID	**25.6 (10/39)**	**1.3 (16/1248)**
Zf-AID	22.5 (9/40)	1.3 (17/1280)
E26R	20.5 (8/39)	0.8 (10/1248)
K22A	15.4 (6/39)	0.72 (9/1248)
P86A	15 (6/40)	0.5 (7/1280)
Chimera-14	12.5 (5/40)	0.78 (10/1280)
K22D	10 (4/39)	0.5 (6/1248)
T27V	10 (4/40)	0.63 (8/1280)
R25A	10 (4/40)	0.5 (7/1280)
R25D	10 (4/40)	0.5 (6/1280)
K52D	7.5 (3/40)	0.31 (4/1280)
Chimera-15	7.5 (3/40)	0.23 (3/1280)
T27D	5.0 (2/40)	0.16 (2/1280)
N51A	5.1 (2/39)	0.16 (2/1248)
N51Q	5.0 (2/40)	0.16 (2/1280)
Chimera-27	2.5 (1/40)	0.23 (3/1280)

Table S4: Putative ssDNA binding residues and their polynucleotide contacts in AID:ssDNA complexes containing a dC in the catalytic pocket. The ssDNA binding frequency describes the percentage of complexes in which each residue contacts a portion of ssDNA. For each complex bound to the corresponding residue, the phosphodiester backbone (%), Deoxyribose sugar (%) and nucleobase (%) columns describes the percentage of complexes that the residue contacts the corresponding region. As expected, we noted that positively charged residues that bound ssDNA with high or very high frequency (R25, K22, K52 and R24) contacted the phosphodiester backbone in the majority of complexes. We found that these residues electrostatically bound 2-3 phosphate groups per complex, as opposed to 1 phosphate group by positively charged residues with lower binding frequencies (ie: R50, R19, K10, R119). This suggests that high frequency positively charged binding residues act in concert to bind ssDNA at multiple sites within the DNA binding grooves.

Residue	ssDNA binding Frequency (%)	Phosphodiester backbone (%)	Deoxyribose sugar (%)	Nucleobase (%)
Very high frequency binding residues				
R25	93.8	73	67	100
K22	87.5	64	64	100
N53	81.3	54	62	38
Y114	81.3	46	46	92
E26	75	33	42	58
F115	75	50	50	67
High frequency binding residues				
K52	68.8	55	64	100
G23	62.5	30	60	90
G54	62.5	80	50	40
A21	56.3	33	33	78
C116	56.3	11	22	100
R24	50	75	75	88
W20	43.8	57	57	43
W84	43.8	43	86	100
Medium Frequency binding residues				
R50	31.3	40	60	40
Y88	31.3	60	80	40
E117	31.3	20	20	100
R19	25	75	75	50
K10	18.8	67	0	33
C55	18.8	67	100	100
D89	18.8	67	33	33
K120	18.8	33	33	67
A121	18.8	0	100	33
R171	18.8	0	0	100
Low Frequency binding residues				
Y13	12.5	100	50	100
Q14	12.5	100	50	50
L59	12.5	50	100	100
H93	12.5	0	0	100
R112	12.5	0	50	100
D118	12.5	0	0	100
R119	12.5	50	50	100
E122	12.5	0	100	50
Q175	12.5	50	50	100
K16	6.25	0	0	100
N17	6.25	100	0	0
V18	6.25	0	0	100
N51	6.25	0	0	100
L60	6.25	100	0	100
R63	6.25	0	100	100
A91	6.25	0	100	0
R92	6.25	100	100	0
D96	6.25	0	0	100
R99	6.25	0	0	100
L113	6.25	0	0	100
E124	6.25	100	0	0
R128	6.25	100	100	100
L172	6.25	0	0	100
R178	6.25	100	0	0
I179	6.25	0	0	100

<u>Chapter 4:</u> A novel intrinsic regulator of AID/APOBECs in immunity and cancer: Schrödinger's CATalytic pocket

Justin J. King, Mani Larijani*

A version of this thesis chapter has been published previously in *Frontiers in Immunology* (2017). 8(26);351. Permission for reuse has been granted from the publisher. This chapter is a modified version of the published manuscript not the contents.

Immunology and Infectious Diseases Program, Division of Biomedical Sciences, Faculty of Medicine, Memorial University of Newfoundland

*To whom correspondence should be addressed:

Mani Larijani, Division of BioMedical Sciences, Faculty of Medicine, Health Sciences Center, MUN, 300 Prince Phillip Dr., St. John's, NL, Canada A1B 3V6, mlarijani@mun.ca

For this chapter I contributed to the concept and writing of the manuscript.

Acknowledgements

The highlighted work was supported by a Canadian Cancer Society Research Institute (CCSRI) Innovation operating grant and Canadian Institutes of Health Research (CIHR) operating grant (MOP111132) to ML. JK is supported by a CIHR Frederick Banting and Charles Best Canada Doctoral Scholarship. We are grateful to our colleague Dr. Michael Grant for discussions which led to the inspiration behind the title.

Abstract

Activation-induced cytidine deaminase (AID) and its relative APOBEC3 cytidine deaminases boost immune response by mutating immune or viral genes. Because of their genome-mutating activities, AID/APOBECs are also drivers of tumorigenesis. Due to highly charged surfaces, extensive non-specific protein-protein/nucleic acid interactions, formation of polydisperse oligomers and general insolubility structure elucidation of these proteins by X-ray crystallography and NMR has been challenging. Hence, almost all available AID/APOBEC structures are of mutated and/or truncated versions. In 2015, a functional structure for AID was reported using a combined computational-biochemical approach. In so doing, a novel regulatory mechanism in human DNA/RNA-editing enzymes has come to light. This mechanism involves auto-closure of the catalytic pocket. Recent X-ray and NMR studies have shown that multiple other APOBECs also close their catalytic pockets. Here, we highlight catalytic pocket closure as an emerging and important regulatory mechanism of AID/APOBEC3s. We focus on three sub-topics: first, we propose that variable pocket closure rates across AID/APOBEC3s underlie differential activity in immunity and cancer and review the supporting evidence. Second, we discuss pocket closure as an ever-present internal regulator, in contrast to other proposed regulatory mechanisms that involve extrinsic binding partners. Third, we compare the merits of classical approaches of X-ray and NMR, with that of emerging computational-biochemical approaches, for structural elucidation specifically in the AID/APOBEC family.

Importance and challenges of solving AID/APOBEC structures

Activation induced cytidine deaminase (AID) is a 198 amino acid DNA-editing enzyme that deaminates deoxycytidine (dC) to deoxyuridine (dU) in single-stranded DNA (ssDNA; (37,48,49,51–53). It acts on immunoglobulin (Ig) loci initiating hypermutation and recombination events that lead to improved and class-switched antibodies (37,38,47,55). However, AID is loosely targeted to Ig loci and hence, it induces genome-wide mutations and double strand breaks which can lead to tumors (115,143–145,203,204,231). In addition, continued AID expression increases genetic plasticity of tumors thereby accelerating disease progression (156,159).

AID is a member of the apolipoprotein B RNA editing catalytic component (AID/APOBEC) family of cytidine deaminases, a Zn-dependent family with 11 members in humans: AID, APOBEC1, APOBEC2, the APOBEC3 sub-branch (A-H, excluding E), and APOBEC4 (55,173,205). The APOBEC3 (A3) sub-branch members are anti-retroviral/retroelement restriction factors thereby also playing an immune function (232,233); however, in the last few years, a major role in cancer initiation has also emerged for the A3 sub-branch of the family, in particular A3A,A3B (234–243) and more recently A3H haplotype I (244).

Given their intimate links to immunity and cancer, much effort has been placed on understanding the molecular structures of AID/APOBECs over the last decade. A major hurdle in this effort has been the isolation and purification of native AID/APBOEC proteins to absolute purity. To different measures for each individual family member, the challenges include cellular genotoxicity, highly charged surfaces mediating extensive non-specific protein-protein, protein-DNA/RNA interactions (55,206) and polydisperse oligomerization (165,166). Consequently, in most cases structure resolution by the traditional methodologies of X-ray crystallography and NMR have necessitated substantial alterations to stabilize protein charge and enhance solubility or crystallization. Consequently, the vast majority (22 of 24) APOBEC structures solved by X-ray or NMR to date are of truncated and/or mutated versions, necessary to enhance solubility (Table S2;

(50–52,75,76,78,84,85,124–136). Nevertheless, it has become clear that AID/APOBEC family member enzymes share the core structure of a central ß-sheet with 4 or 5 ß strands sandwiched between 6 or 7 α-helices, connected by 12-13 flexible loops of variable lengths (Figure 21A). Second, they all have highly charged DNA binding grooves, necessary to bind . negatively charged polynucleotides. The arrangement of core catalytic residues in the catalytic pocket is also conserved, consisting of a Zn-coordinating triad of two cysteines and a histidine, atop a catalytic proton-donor glutamic acid (C87, C90, H56 and E58 in AID; (55,96,173,205).

Solving the structure of AID and discovery of catalytic pocket closure in AID/APOBECs

The biochemical properties of AID have been previously described. AID has an exceptionally high affinity (nM range) for binding ssDNA and an unusually slow catalytic rate of 1 reaction in several minutes (52,188), ~2000 times slower than the typical enzyme (61). We postulated that this catalytic lethargy and high binding affinity to DNA had evolved to protect genomes from rampant AID activity (55). In direct support of this notion, mutants of AID with higher catalytic rates were shown to mediate higher levels of genome damage in cells (60). Although this body of work led to understanding AID's behavior, the molecular basis behind these properties remained an enigma because AID’s native structure eluded resolution since its discovery in 1999 until 2015, despite intense efforts. AID is notoriously challenging to isolate to absolute purity and hence its native structure had remained unsolved by X-ray and NMR. We posited that even if AID's structure were to be solved by traditional methods of X-ray or NMR it would most likely be of a truncated and/or heavily mutated version. Thus, we proposed an alternative methodology to gain insight into the functional and native structure of AID. We utilized 8 recently solved structures of AID's APOBEC relatives as templates (96) to generate thousands of AID predicted model structures followed by identification of the lowest energy clusters (Figure 21A; (96). Concurrently and guided by the model

predictions, we generated a library of 400 AID variants and carried out extensive biochemical

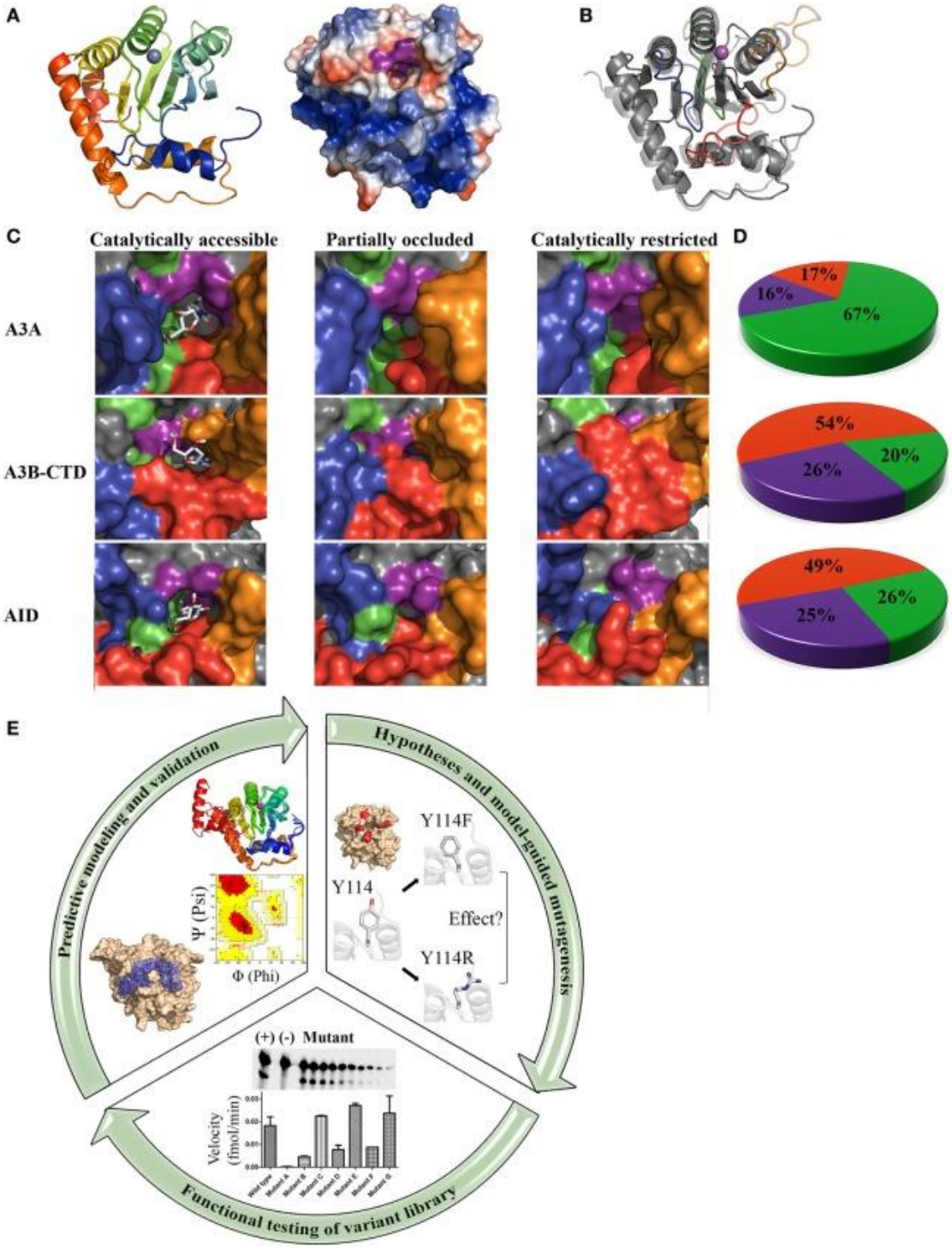


Figure 21. Core architecture, catalytic pocket occlusion and computational/biochemical approaches to solving AID/APOBEC structures. (A) Representative Ribbon structure (left) and surface topology (right) of AID modeled from APOBEC templates. In the ribbon structure, N- to C- termini progression is shown

from blue to red and the grey sphere depicts active site zinc. In the surface topology, positive, negative and neutral residues have blue, red and white surfaces, respectively. The Zn-coordinating residues and catalytic glutamic acid surface are colored purple. A distinct feature of AID amongst the APOBECs is its high positive charge at neutral pH, concentrated along two ssDNA binding grooves that pass over the catalytic pocket. (B) Ribbon structures of A3A (transparent) and AID (non-transparent) were superimposed. In each protein structure, the secondary catalytic loops 2, 4, 6 and 8 are colored red, orange, green and blue, respectively. (C) Catalytically accessible (left), partially occluded (middle) and catalytically restricted (right) conformations of A3A (top), A3B-CTD (center) and AID (bottom). The surface of secondary catalytic loop 2, 4, 6 and 8 were colored red, orange, green and blue, respectively. Catalytically accessible conformations are shown with bound dC in the catalytic pocket. Conformations were deemed catalytically accessible if they bound dC in a deamination-feasible configuration in the catalytic pocket via molecular docking (AutoDock VINA (133)). In catalytically restricted conformations, the secondary catalytic loops adopt a configuration that block the pocket. (D) Proportion of catalytically restricted (red), partially occluded (purple) and catalytically accessible (green) conformations in A3A (top), A3B-CTD (center) and AID (bottom). A3A showed a dramatically higher proportion of catalytically accessible conformations in comparison to A3B-CTD and AID. NMR conformations of A3A, (PDB: 2M65) A3B (PDB: 2NBQ) and previously reported structures of AID (58) were used. (E) Combinatorial computational/biochemical approach for solution of functional and native AID/APOBEC structures. A library of predicted structures is generated through homology modeling with a range of suitable template structures generating multiple low energy conformations. The quality of the resulting conformational ensemble is inspected (ie: construction of Ramachandran plots). Molecular docking can be used to determine the substrate binding regions in the active site and surrounding regions. Model quality and substrate-docking are checked in concordance with known biochemical properties of the enzyme. Concurrently, specific hypotheses can be formed based on the highest confidence predicted conformations and their interaction with substrate. In order to test these hypotheses and to validate the positions and relative attitudes of specific core or surface residues, a large variant library ought to be constructed and tested in functional enzyme assays. Mutational scanning ought to be used to confirm the involvement of key residues/regions in specific biochemical aspects such as substrate binding, catalysis and structural stability. For key residues involved in catalysis regulation, often several point mutations spanning the range of synonymous to severe are informative. Collectively, information from functional testing of the variant library is used to refine the understanding of the enzyme structure and its interaction with substrate.

characterization of catalytic function and DNA binding to rigorously test key predictions of models. This library included different point mutants for each residue along the length of AID, orthologous AIDs, and chimeras involving regions of other deaminases exchanged into the AID scaffold, or vice-versa. Our rationale for including AID orthologs was that divergent AID from distantly-evolved species may have distinct biochemical properties and characterizing these through a combination of homology structural modeling and functional analysis of mutated and chimeric enzymes would generate structure:function insights. Differences

amongst orthologous AIDs included catalytic rates, DNA binding affinities and thermosensitivity profiles (62,94,95). Since these differences are typically due to structural features, being reflective of catalytic motifs, surface composition and overall protein flexibility, respectively, characterizing the basis of these differences amongst orthologs proved a valuable tool to gain insight into AID's structure:function relationships. This computational-biochemical approach led to the first relatively detailed 3D maps of AID's functional structure with special focus on catalytic pocket architecture and ssDNA binding motifs (55,62,94,96).

The architecture and dynamics of an enzyme's catalytic pocket are important determinants of its activity. In addition to the core catalytic pocket composed of the aforementioned triad Zinc-coordinating residues and a Glutamic acid, we identified an additional 21 amino acids that are not directly involved in the deamination reaction, but compose the pocket's physiochemical microenvironment (96). These residues termed secondary catalytic residues form the "walls" and "floors" of the pocket and stabilize dC binding. We noted that the conformations of these secondary catalytic residues exhibited more variability than that of the primary catalytic residues amongst predictions, because these residues reside on several highly flexible connecting loops without secondary structures of their own, that surround the catalytic pocket (Figure 21B). Because of this placement, it appeared that the catalytic pocket of AID is only marginally stable such that ~75% of conformations exhibit an occluded pocket unable to accommodate dC. Thus, we hypothesized that catalytic pocket closure is a built-in mechanism that limits AID activity.

There are also several lines of indirect but strong evidence for the existence of catalytic pocket closure in AID/APOBEC3s. First, the fact that the majority of AID conformations exist in a state with closed catalytic pockets provides a mechanistic explanation for the relative catalytic lethargy of AID as discussed in the preceding section (52). Second, it is mathematically compatible with known parameters of AID: AID binds ssDNA sporadically on its surface such that most ssDNA (~95%) neither pass over AID's catalytic pocket, nor position dC for catalytic pocket entry. The proportion of catalytically-viable AID:DNA

complexes (~5%) multiplied by the ratio of open pockets (~25%) yields ~1.25%. This correlates closely with our own estimates of active AID complexes based on Michaelis-Menten parameters (52) and with other studies that carried out mathematical modeling of AID's substrate catalysis (202).

Direct proof for existence and significance of catalytic pocket closure came from two sources: one functional, and the other structure-based. First, we designed a panel of AID variants in which the secondary catalytic loops and surrounding regions were replaced with their equivalents from other APOBECs or orthologous AIDs to alter predicted pocket dynamics such that the pocket would spend either more or less time in the open conformation. We then observed that the proportion of time the pocket was predicted to assume an open conformation correlated exquisitely with catalytic rate differences amongst said AID variants some of which became up to 100 times more active than wild type AID because of a catalytic pocket that spends more time in the open conformation. This provided functional evidence that pocket closure limits activity. The second proof came from direct observation of closed pockets in several siblings of AID: in APOBEC3A by NMR(208), in APOBEC3B by X-ray crystallography (246,257) and by NMR (Figure 21 , D; (247).

Differential catalytic pocket states mediate variable biological activities amongst AID/APOBECs

From an evolutionary perspective, regulation by catalytic pocket closure provides an effective means to fine-tune variable levels of enzymatic robustness across the AID/APOBEC family, as well as impart varying types of activities amongst orthologous versions of each family member. This is because the same high degree of movement freedom in the secondary catalytic loops that lead to the fluidity of catalytic pocket dynamics in each AID/APOBEC3 enzyme, also allows for a high level of sequence and length divergence in these loops amongst the individual AID/APOBEC3s, to impart a unique range of open/closed breathing dynamics to the catalytic pocket of each member (Figure 21). To elaborate, in each APOBEC3 the catalytic pocket "walls"

and "floor" are composed of residues contributed by four secondary catalytic loops (Figure 21B).

The highest structural variation amongst the AID/APOBEC family appears in loop 2 (L2), loop 4 (L4) and loop 8 (L8) with respect to sequence homology, length and compaction relative to the core enzyme structure (Figure 21C; (96). L2 contains residues involved in ssDNA binding, catalytic pocket and dC stabilization, substrate specificity and 5-mC tolerance (59,96,199,230). L4 contains residues critical to catalysis and catalytic pocket occlusion (96). Recently, an allosteric regulatory role for L4 in A3A and A3G was identified through coordination of a secondary Zn that enhances activity (258). Interestingly, secondary Zn coordination was suggested to fine tune the position of the secondary catalytic residues, thus creating an ideal environment for cytidine deamination (258). Furthermore, secondary Zn coordination was suggested to mediate cooperative dimerization. Lastly, L8 mediates substrate sequence specificity, dC stabilization and 5-mC tolerance (94,96,230,258). Collectively, the secondary catalytic loops mediate functional differences amongst the AID/APOBECs and dictate variations in the frequency of open vs. closed catalytic pockets.

Based on these observations, we propose that differences in secondary catalytic loops mediate catalytic pocket breathing dynamics which is responsible for different enzymatic robustness amongst the AID/APOBEC enzymes. Indeed, in the last year, functional evidence in support of this novel mode of regulation has emerged. First, A3A exhibits open catalytic pockets in more conformations than AID (67% vs. 25.6%), and accordingly it is a more robust enzyme with a faster on/off rate of deaminating DNA (58,208). Second, A3B-CTD exhibits roughly one third the pockets in an open conformation compared to A3A (20% vs. 67%, respectively; Figure 21D) and this also correlates directly with a lower catalytic rate (247). It is intriguing that thus far, catalytic pocket occlusion has been observed in 3 of the most mutagenic and tumorigenic members of the AID/APOBEC family: AID, A3A and A3B. This lends credence to the idea that this is an internal protective mechanism to limit genome mutations by these enzymes. As mentioned above, the difference in catalytic activity of purified AID, A3A, A3B-CTD, correlates with the ranking of pocket

occlusion (Figure 21D). Though further study is required to clarify the relative contributions of A3A, A3B and other APOBEC3 branch enzymes such as A3H in various types of cancers, some emerging evidence indicates that there is a more dominant mutational signature observed from A3A than A3B, at least in a yeast model and in urothelial carcinoma, despite lower levels of A3A expression (229,239).

In addition to regulation of tumorigenic activity, differences in catalytic pocket dynamics also appear to correlate well with other biological functions of AID/APOBECS. As an example, zebrafish AID has a significantly higher reaction rate than human AID and is also unique amongst all AID orthologs in that it can deaminate 5-methyl-C (5-mC) in methylated CpG motifs (94). This explains a puzzling previous report that zebrafish AID plays a completely non-immune role. During embryogenesis in zebrafish, AID can mediate promotor demethylation through erasure of gene-silencing CpG methylation marks, thus orchestrating widespread gene expression required for tissue differentiation (259). This is attributable to conformational differences in the aforementioned secondary catalytic loops between human and zebrafish AID, which translate to a higher ratio of open *vs.* closed catalytic pockets. Consequently, zebrafish AID can deaminate 5mC at a high rate, as opposed to human AID whose activity on 5mC is negligible. This enzymatic difference is one factor that enables zebrafish AID to function in genome demethylation during embryonic development, an activity that is completely outside the realm of an immune function (94,96). Taken together, these lines of evidence are supportive of pocket occlusion being a key regulator of biological functions of AID/APOBECs, including their role in cancer initiation.

Catalytic pocket occlusion as internally built-in regulation

Since the discovery of AID, much effort has been directed at understanding how its activity is regulated, under the premise that a mutator so threatening must be operating under tight restrictions. To date,

almost all efforts have focused on modes of regulation that are extrinsic to the enzyme itself. This has led to the identification of over two dozen cofactors proposed to bind AID either directly or indirectly through associations with other proteins or DNA/RNA (63,87,137,171,260–281). The list of putative binding partners is rather large for a relatively small protein of 198 residues, and one must approach biological relevance with caution for several reasons: first, although some cofactors are modestly enriched at Ig loci, none can account for targeting AID to specific loci. Second, given the relatively small size of AID and the lack of clear conformational protein binding domains, the number of proposed co-factors seems high. It is rather improbable that a small 198 aa globular enzyme can fold properly to bind ssDNA, deaminate dC, maintain sequence specificity, whilst still leaving enough non-essential portions free to bind dozens of different co-factors each in a specific fashion. Indeed, a careful analysis of AID's structure reveals that most of its structure can be ascribed a function directly related to forming the core architecture essential to bind and deaminate a polynucleotide. Furthermore, a portion of the surface is likely unavailable due to forming the oligomerization surfaces, as most AID/APOBECs appear to exist as dimers or tetramers (52,170,282,283). Also, AID has a highly charged surface and a well-known propensity for high affinity non-specific interactions with other proteins (55,206). Thus, the biological significance of AID binding to many of its putative co-factors is a topic that requires further resolution. Furthermore, the very premise of searching for cofactors to explain "regulation" may be flawed in that the more AID is studied, the clearer it becomes that its activity is rather not tightly regulated: despite a modest preference for Ig loci which appears to be mediated by unique transcriptional features, (284–286) AID mutates endogenous genes and transgenes genome-wide and can do so in any prokaryotic or eukaryotic cell in which it is naturally or exogenously expressed (115,200,287–289).

Like AID, the search for regulatory mechanisms of other APBOEC3s has also focused on extrinsic binding factors, of which several have been identified including various viral proteins, and transcription factors (290–293); however, the most well-characterized APOBEC3 are Vif and cytoplasmic

ribonuclear complexes. The Virion infectivity factor (Vif) protein of HIV binds and targets A3C, A3G, A3F, A3D and A3H (to varying degrees) for degradation *via* a ubiquitin-dependent proteosomal pathway (210,294–302). Thus, when Vif is present, APOBEC3 effectiveness in viral restriction is severely diminished. Second, the activity of the anti-retroviral APOBEC3s is limited by entrapment in high-molecular-mass ribonuclear complexes (HMM) that may reach megadaltons in size, mediated by non-specific protein/DNA/RNA binding in the cytoplasm, mediated by aforementioned highly charged surfaces (303–309).

In contrast to regulation by extrinsic binding partners, be they protein or nucleic acid, catalytic pocket closure represents a novel intrinsic mode of regulation. This simple mechanism of limiting activity has several attractive features: it is ever-present, biologically reliable, mechanistically simple and structurally sound. Furthermore, as discussed in the preceding section, its variation is an evolutionary efficient mechanism for diversifying and fine-tuning the activity levels of family member enzymes, as catalytic pocket closure rates can be adjusted by minimal amino acid changes in secondary catalytic loops. It is also biologically efficient since it does not require any cellular resources, unlike the proposed complex networks of many co-factors which themselves would require regulation in different cells at different stages of differentiation or viral infection, thus requiring significant cellular resources.

Importance of determining AID/APOBEC3 structures that are native and include functional insights

X-ray crystallography and NMR have advanced the AID/APOBEC field with the full or partial structure elucidation of 7/11 APOBEC enzymes. Despite these achievements, there are pitfalls in using these traditional methods alone. First, the purification issues discussed above have necessitated working with significantly truncated and/or heavily mutated versions of AID/APOBEC proteins (Table S5). The truncations and mutations are often in functionally critical regions, such as the secondary catalytic loops. Additionally, all

double-domain APOBECs whose structure has been characterized (A3B, A3F and A3G) lack their enzymatically inactive N-terminal half which is implicated in the catalytic activity and dimerization (Table S5; (256,310,311). The N-terminal half of A3G's separate structure was recently reported; however these were also mutated and likewise lack the C-terminal half (255,256). Second, depending on crystallization or NMR conditions, even the same APOBEC structure determined by different groups can be quite distinct (185,187).These differences are likely due to differences in solution or crystallization conditions which can bias toward a specific structure or conformation (312).

In contrast, the methodology that we applied to solving AID's structure provides both a functional and native structure (96). By integrating dynamic modeling with the study of a large library of variants to functionally verify key model predictions, the emerging picture integrates the relative abundance of an enzyme's conformations with functional significance (Figure 21E). This approach is particularly advantageous in the case of AID/APOBEC3s because the functional differences amongst AID/APOBEC3 family members are dictated by subtle differences in breathing dynamics, rather than major architectural differences. It is important to note that despite being a robust methodology for determining functional and native enzyme structures, this approach is not without practical challenges: first, it is laborious and time-intensive since it requires examination of thousands of high confidence models. As modeling efforts progress, there is a continuing need to generate and test a large variant library, often necessitating several mutants of each key residue to rigorously verify its exact position, relative attitude and role(s). In addition, a sensitive enzyme assay able to detect even small differences in biochemical properties with that of wild-type ought to be in place. Practical difficulties are compounded by the fact that this approach of solving a functional and native structure is often most useful for enzymes that are challenging to purify. Second, modeling efforts depend critically on the availability of solved X-ray or NMR structures to serve as templates, with multiple templates increasing confidence. For instance, at the time of our efforts on AID, we utilized 8 available

APOBEC structures as templates. It is important to have numerous templates from different family members, so as to at least partially compensate for limitations of each template in terms of encompassing the full range of conformations. Furthermore, template structures ought to be evaluated for their suitability on the basis of extent and location of homologous/identical residues, and model quality itself ought to be rigorously scrutinized, mathematically and functionally using the variant library (Figure 21E). Third, the basic biochemical properties of the enzyme ought to have previously been determined so as to serve as a further check-point for model validity. Since we had already determined that AID has an unusually low catalytic rate and high binding affinity for ssDNA, the fact that our structure fully explained both of these properties through the abundance of closed pockets and positively charged surface residues, respectively, provided further confidence. Lastly, definite physical confirmation of findings requires observation by X-ray and NMR, as in the case of catalytic pocket closure described above. In addition to our observation of catalytic pocket closure being confirmed by direct X-ray and NMR studies as described above (246,257) and by NMR(247), other X-ray studies have also confirmed our observation of key catalytic residues as well as important DNA-binding residues of AID: following the publication of AID's functional structure, the crystal structure of an AID variant was also reported (245). As expected, it included mutations and truncations crucial to solubilize AID for X-ray crystallography (Table S5). Although this structure represents a significant achievement, it necessitated introducing mutations and truncations that removed some of the unique characteristics of AID. For instance, many mutations neutralized the positively charged surface residues lining the DNA binding groove culminating in a net charge of +4.5, as compared to AID's native charge of +14 at neutral pH. This high net positive charge of AID is a unique feature amongst AID/APOBEC3s with known structure (-2, -6, +0.5, -9 and -3.5 of A3A, A3B-CTD, A3C, A3F-CTD and A3G-CTD, respectively).

With this limitation, this structure presents a unique opportunity for a comparison of structure determination methodologies. To this end, we compared the AID variant crystal structure with the

computational-biochemical AID conformational ensemble (loop 2, 4, 6 and 8, denoted as loop 1, 3, 5 and 7, respectively, in other APOBEC publications). Overall, the AID structures shared virtually the same tertiary structure and confirmed some of the key secondary catalytic residues we posited would stabilize dC in the catalytic pocket (ie: N51 and Y114). Most of the mutations in the AID variant were localized to L1, L2 and α1, regions, while α7 was deleted. There are also several notable differences between models: First, L2 adopts a more compact conformation relative to the core structure, likely due to the deletion of three residues in L2. We and others have previously shown that L2 plays a role in catalytic activity and AID:DNA binding of AID (96,230). Second, L8 adopts a much more extended conformation in the AID variant. It was suggested this extended conformation stabilized larger purine bases upstream of the target cytidine, in contrast to other APOBECs whose shorter L8's preferred pyrimidines upstream. However, the structure of L8 is stabilized by L2, which has been shown to modulate its compaction and substrate specificity (199). Additionally, the conformation of α7 relative to the surface of AID is uncertain, although some conformations place it in direct contact with L8 (58). Therefore, although L8 was not directly altered, mutation of L1- α1-L2 together with α7 deletion may indirectly perturb its conformation in the AID variant crystal structure.

Similarly, using our computational-biochemical approach, we also highlighted two DNA binding grooves on the surface of AID, for both of which the positively charged R25 residues plays a major role in orienting the the negatively charged DNA backbone (96). Recently, DNA-bound crystal structures of mutant A3A and an A3B-CTD chimera were shown to adopt a similar DNA binding mode, whereby the DNA backbone was bound around the equivalent of R25 in AID (H29 and H212, in mutant A3A and A3B-CTD chimera, respectively (257). In this manner, X-ray and NMR structural elucidation of homologous APOBECs have provided direct physical support for notable features observed using the computational-biochemical approach.

Conclusions

In summary, we draw the parallel to the Schrödinger's Cat paradox that the catalytic pockets of AID/APOBECs appear to transition between dual states, one of which correlates with activity and the other with catalytic death, each with profound functional consequences. The second parallel between structure determination in the AID/APOBEC3 field and quantum physics is that X-ray crystallography and NMR determination of structures in the AID/APOBEC family have necessitated making alterations to structures during the very act of observation. In contrast, the computational-biochemical approach as was used to examine AID's functional structure relies on unobtrusive observation through prediction. Interventions are reserved for the functional testing phase wherein structure predictions are rigorously scrutinized by conducting enzyme assays on a large library of hundreds of mutants, orthologs, and chimeras. As described in the preceding section, it is important to note that this method is nonetheless critically dependent on the availability of multiple X-ray and NMR structure solutions, both in the beginning as templates and in the end, as independent methods to independently verify the key aspects of the structure.

In the future, as the relative contributions of each individual AID/APOBEC3 family member enzyme to immunity and cancer in different contexts become clearer, it will be important to test the hypothesis that catalytic pocket breathing differences amongst the AID/APOBEC3 family members impact their relative contributions, and to understand the extent to which this novel built-in safety switch is intertwined with other regulatory mechanisms, such as aforementioned extrinsic binding partners. Although catalytic pocket closure has been described for other enzymes (313,314) discovery of such a functionally-critical state in AID/APOBEC3s is a first across human DNA/RNA-damaging enzymes; hence, it is also important to ascertain how prevalent a regulatory mechanism dynamic catalytic pocket closure is in other DNA/RNA-editing enzymes, or whether it has evolved as a unique regulatory structural feature of the AID/APOBEC3 family.

Table S5: All X-ray and NMR solution structures of the APOBEC family

APOBEC	Experimental method	Truncations	Mutations	PDB ID
Hs-A2	X-ray	Truncated (Δ1–40)	N/A	2NYT
Mouse-A2	NMR solution	Truncated (Δ1–45)	N/A	2RPZ
Hs-A3A	NMR solution	N/A	N/A	2M65
Hs-A3A	X-ray	N/A	E72A, C171A	4XXO
Hs-A3A	X-ray	Truncated (Δ196–199)	E72A	5SWW
Hs-A3B	X-ray	Truncated (Δ1–186, Δ242–248)	F200S, W228S, L230K, Y250S, F308K	5CQK
Hs-A3B	NMR solution	Truncated (Δ1–186)	N/A	2NBQ
Hs-A3B	X-ray	Truncated (Δ1–186, Δ205–207, Δ242–249)	F200S, V205G, L209I, R210G, R212H, Q213K, W228S, L230K, Y250S, E255A, F308K	5TD5
Hs-A3C	X-ray	N/A	N/A	3VOW
Hs-A3F	X-ray	Truncated (Δ1–184)	Y196D, H247G, C248R, C259A, F302K, W310D, Y314A, Q315A, K355D, K358D, F363D	4IOU
Hs-A3F	X-ray	Truncated (Δ1–217)	N/A	4J4J
Hs-A3F	X-ray	Truncated (Δ1–186)	N/A	3WUS
Hs-A3F	X-ray	Truncated (Δ1–184)	Y196D, H247G, C248R, C259A, F302K, W310D, K355D, K358D, F363D	5HX5
Hs-A3F (Zn-depleted)	X-ray	Truncated (Δ1–184)	Y196D, H247G, C248R, C259A, F302K, W310D, K355D, K358D, F363D	5HX4
Hs-A3G	NMR solution	Truncated (Δ1–197)	L234K, C243A, F310K, C321A, C356A	2JYW
Hs-A3G	X-ray	Truncated (Δ1–196)	N/A	3E1U
Hs-A3G	NMR solution	Truncated (Δ1–192)	N/A	2KBO
Hs-A3G	NMR solution	Truncated (Δ1–190)	L234K, C243A, F310K, C321A, C356A	2KEM
Hs-A3G	X-ray	Truncated (Δ1–194)	L234K, C243A, F310K, C321A, C356A	3IR2
Hs-A3G	X-ray	Truncated (Δ1–194)	L234K, C243A, F310K, C356A	3V4K
Hs-A3G	X-ray	Truncated (Δ1–192)	D370A	4ROW
Hs-A3G	NMR	Truncated (Δ1–11, Δ78, Δ143–146, Δ197–384)	Y13D, R14P, Y22N, L62D, F71L, H72S, W73L, F74V, T101A, A109Q, D110P, P111T, K112H, F126A, C139A, K141A, R142G, M149I, R169G, E170A, L171P, E173Q, N176D, N177G, P179D, K180E, Y181H, Y182S, I183Q, L184A, H186S, I187G, M189R	2MZZ
Primate-A3G	X-ray	Truncated (Δ139–146, Δ197–384)	(C139-Q140-K141-R142-D143-G144-P145-H146) replaced with (A-E-A-G) residues	5K83
Hs-AID	X-ray	Truncated (Δ1–4, Δ20–22, Δ184–198)	N7D, R8P, R9H, K10I, L12T, Y13S, Q14N, K16N, V18G, R19I, R25H, E26K, V32E, K34E, R36L	5JJ4

Most APOBEC structures are heavily modified through mutations and/or truncations. Δ denotes the amino acids that were deleted from the structure. In cases where a study has reported several protein databank IDs of highly similar structures, a representative PDB code is listed.

<u>Chapter 5:</u> First generation small molecule inhibitors targeting the catalytic pocket of activation-induced cytidine deaminase

Justin J. King, Faezeh Borzooee, Junbum Im, Mahdi Asgharpour, Cody P. Diamond, Heather Fifield, Lesley Berghuis, Mani Larijani*

Program in immunology and Infectious Diseases, Division of Biomedical Sciences, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland A1B 3V6, Canada.

*To whom correspondence should be addressed: Tel: 011-1-709-864 6079; Fax: 011-1-709-864 6007; Email: mlarijani@mun.ca

For this chapter I formed all the hypotheses, identified all the initial hit compounds/derivatives, and conducted all the experiments and data analyses with technical help and reagents contributed by the other listed authors.

Funding

This work was supported by a Canadian Cancer Society Research Institute (CCSRI) Innovation operating grant (702145) and Canadian Institutes of Health Research (CIHR) operating grant (MOP111132) to ML. JK received support from a CIHR Frederick Banting and Charles Best Canada Doctoral Scholarship.

Abstract

Activation-induced cytidine deaminase (AID) initiates secondary antibody diversification by converting deoxycytidine to deoxyuridine at immunoglobulin loci in B lymphocytes. AID also induces genome-wide mutations and chromosomal translocations that transform normal B cells into lymphoma/leukemia. Continued AID expression in tumors accelerates tumor evolution, drug resistance and poorer prognosis. Thus, inhibiting AID has been suggested to be of therapeutic benefit. We recently used a combined computational-biochemical approach to gain insight into AID's functional structure, with emphasis on the architecture and dynamics of its catalytic pocket. Here, we describe the identification of a first-generation small molecule inhibitor that targets the catalytic pocket of AID. It inhibits purified AID, AID in whole cell extracts, and endogenous AID of lymphoma cell lines. This small molecule also inhibits AID's tumorigenic siblings APOBEC3A and APOBEC3B, though with lesser potency, but does not inhibit APOBEC3G and APOBEC3F. Analogue expansion yielded derivatives of the first-generation inhibitor, with improved inhibition potencies. Structural examination of AID-inhibitor interactions by mutational analyses highlight key stabilizing interactions between the small molecule and residues within and immediately outside of AID's catalytic pocket, representing opportunities for hit-to-lead development. Our findings are significant for therapeutic development and serve as further functional verification of AID's catalytic pocket architecture.

Introduction

The DNA-editing enzyme activation-induced cytidine deaminase (AID) is expressed in activated B lymphocytes. AID mutates deoxycytidine (dC) to deoxyuridine (dU) at immunoglobulin (Ig) genes triggering somatic hypermutation (SHM) and class switch recombination (CSR) of antibodies (37,38,47,48,55,90,173). AID deficiency results in a mild and treatable form of Hyper IgM characterized by lack of high affinity antibodies of switched isotypes (38,47,176). AID also mutates genome-wide, and some of these mutations result in double stand breaks (DSB) that mediate chromosomal translocations (115,145,203,204,315,316). Strong evidence implicates AID in initiation of leukemia/lymphomas including Burkitt's lymphoma (BL), diffuse large B cell lymphoma (DLCL), follicular lymphoma (FL), multiple-myeloma and chronic lymphocytic leukemia (144,203,317–320). These tumors arise from centroblasts or post-centroblasts, the narrow stage in a B cell's life where AID is expressed, and mutations and translocation breakpoints (c-myc/IgH in BL, bcl-2/IgH in FL, bcl-6/IgH in DLCL, IgH-CCND1 in Mantle cell lymphoma) occur at genomic sites that are frequently targeted by AID (140,141,144,321–327). The causal role of AID in lymphomagenesis was proven in IL-6 transgenic mice which develop lymphomas that mimic human BL in phenotype/genotype (c-myc/IgH). In this model, DSBs at both the IgH and translocation partner c-myc loci were demonstrated to be directly due to AID activity (143–145,177,328).

Beyond tumor initiation, AID expression can also exacerbate leukemia/lymphomas. Genome-wide AID-mediated mutation signatures are prevalent in leukemia/lymphomas and numerous studies have shown that AID levels in tumors correlate with poor diagnosis (69,146,147,149–156,203,320,329–331). In chronic myeloid leukemia (CML), AID was shown to mutate tumor suppressor and/or DNA repair genes and accelerate Imatinib resistance (156). Moreover, recent studies indicate that some therapeutic agents, such as Idelalisib and Duvelisib, can exacerbate AID-mediated genome mutations in tumors through increased AID expression and chromosomal translocation frequency between the IgH locus and off-target sites (160,161).

Thus, inhibiting AID activity has been suggested as a potentially useful approach to

treating AID-expressing malignancies or augment other therapies (159,320,331–340), but a significant hurdle over the last decade has been the lack of an AID structure. AID is a member of the apolipoprotein B mRNA editing enzyme catalytic polypeptide-like (APOBEC) family of Zn-dependent, single-stranded polynucleotide-restricted cytidine deaminases (55,173,205). AID and its related APOBEC cytidine deaminases have unique molecular properties which makes them challenging to work with: they are genotoxic, have highly charged surfaces, extensive non-specific protein-protein/DNA/RNA interactions, and are prone to formation of polydisperse oligomers in solution (55,165,166,206). These properties make it extremely difficult to purify and crystalize for X-ray and NMR studies. For these reasons, 27 of 29 related APOBEC structures solved by X-ray or NMR are of heavily mutated and/or truncated versions, necessary to enhance expression or solubility (164).

Thus, we postulated that even if AID's structure were to be solved by X-ray or NMR it would likely be of a mutated or truncated version. Instead, we utilized a combined computational-biochemical approach to glean insights into AID's native and functional structure. This approach is based on structure prediction using multiple templates, followed by rigorous functional verification of model predictions using a library of AID variants, including point/multiple mutants, orthologs and chimeric versions with portions of other deaminases exchanged into the AID scaffold, or vice-versa. We thus arrived at a map of AID's functional structure including surface topology, core architecture and catalytic pocket (96).

A detailed understanding of AID's catalytic and target DNA binding mechanism is key for inhibitor development. It was in turning our focus to its active site that we found the catalytic pocket of AID shifts dynamically between open and closed positions, and that the majority (~ 75%) of conformations at any time are predicted to exhibit a closed pocket (96). This dynamic pocket closure was the first demonstration of such an inherent regulatory mechanism in human DNA/RNA-editing enzymes, and it provided a structural explanation for the previously-demonstrated slow catalytic rate of AID of one reaction in several minutes (52,60,61). More recently, catalytic pocket closure, as we described in AID, was observed by X-ray and NMR on APOBEC3A and APOBEC3B, two siblings of AID (208,246,247,257,341). In addition, we found that the surface of AID is extremely positively charged (+14), which explains its nM-range high affinity for binding negatively-charged DNA (52,53,164). It also results in the majority (>90%) of

ssDNA binding on the surface in sporadic fashion, with only a minor fraction of bound ssDNA located in either one of two catalytically-productive ssDNA binding grooves (96). Since publication of the functional structure of AID using the computational-biochemical method, two partial crystal structures have become available (54,245) which confirmed the map of AID's catalytic pocket.

Here, as an extension of our combined computational-biochemical approach to delineating the breathing and dynamic structure of AID, we screened a library of small molecule drug-like compounds against its catalytic pocket. We identified first-generation small molecule inhibitors that specifically inhibit the mutagenic activity of purified AID, native AID in whole cell extracts, and endogenous AID of B lymphoma cells.

Materials and Methods

Virtual High-throughput screening of small molecules against the catalytic pocket of AID

The AID structure used for high throughput *in silico* identification of first generation hits is based on the functional and native AID structure described previously though a combined computational-biochemical method (96). This structure has been verified by two partial AID crystal structures (54,245). In addition, since the catalytic pocket of AID was designated as the inhibitor target for this study, this structure is advantageous because it includes several dynamic conformations of AID's catalytic pocket, verified by biochemical analysis of AID variants (96). Briefly, this structure was generated by modeling full-length AID based on eight X-ray or APOBEC structures as templates for homology modeling: A2 NMR (PDB: 2RPZ), $A2_{41–224}$ chain A and B X-ray (PDB: 2NYT:A and 2NYT:B, respectively), A3A NMR (PDB: 2M65), A3C X-ray (PDB: 3VOW), A3F-CTD X-ray (PDB: 4IOU), A3G-CTD NMR (PDB: 3E1U), and A3G-CTD NMR (PDB: 2KBO; (170,185,187,207,208,210,211). All AID/APOBEC structures were obtained from the protein databank (http://www.rcsb.org) and visualized using PyMOL v1.7.6 (http://www.pymol.org). Using the default parameters of I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) (213,214) full-length human AID (Hs-AID) and variants were modeled from APOBEC templates. The catalytic pocket is defined as the indented space containing Zn and the catalytic residues (H56, E58, C87 and C90 in Hs-AID).

Using DOCK Blaster v1.6.0 (http://blaster.docking.org/;(342) we virtually screened $4.6x10^6$ "clean-lead" small molecules from the ZINC database (http://zinc.docking.org/) against the catalytic pocket of AID. We used several AID-DNA complexes containing dC in the catalytic pocket as a template for screening (96). In total, we screened five low energy conformations of AID, representative of the range of catalytically accessible catalytic pocket conformations. As a result, small molecules were screened for their ability to bind to the catalytic pocket and surrounding DNA binding groove region across the ensemble of catalytically active AID structures. Compounds were docked and ranked based on binding energy. The 500 lowest energy compounds bound to each AID conformation were cross-referenced and compounds predicted to bind only one catalytically active AID conformation were excluded. The 40 lowest-energy compounds bound across several catalytically active AID conformations were selected for additional docking using Autodock VINA (http://vina.scripps.edu) (188) via PyRx (https://pyrx.sourceforge.io) (343) to confirm specificity to the catalytic pocket and ranked energies. The top 10 compounds (C1-C10) were then selected for testing based on binding energy ranking, as well as the chemical diversity of structure sidechains. For analogue expansion, we used the same approach against the ZINC database to identify 948 structural analogues of C4 and C8 considering a 60% similarity cutoff. Using Autodock VINA via PyRx we screened the five AID conformations and identified the top ranking structural analogues of C4 and C8, respectively.

Expression and purification of AID/APOBECs

Expression and purification of GST-AID and AID-His in bacteria and HEK 293T cells have previously been described (52,94,96). Briefly, for bacterially-expressed GST-AID, the expression and purification of human (Hs-AID), Hs-AID mutants and zebrafish (Dr-AID) GST-AID was carried out as previously described, using the pGEX5.3 expression system (94,95). Point-mutants were generated by site-directed mutagenesis using appropriate GST-AID constructs as templates, as previously described (96). Briefly, GST-AID was expressed in *E. coli* (DE3-Bl21) and purified using GST-column chromatography as per the manufacturer's guidelines. In total, 6 independent preparations of GST-AID and 2 independent preparations of each mutant/chimeric/orthologous AID were made and tested in parallel. On average, the concentration of bacterially-expressed GST-AID preparations were 300 ng/μL. For eukaryotic-expressed AID, expression and purification of

AID-His in HEK 293T cells was carried out as previously described, using the pcDNA3.1 expression system (89). The AID ORF was cloned into pcDNA3.1-V5-6xHis-Topo which was modified with the addition of 2 extra His-residues to encode an 8x C-terminal His-tag. Two versions of the expression construct were generated, with or without the AID ORF stop codon, to express either an AID-His protein, or a native untagged AID, respectively. 50 x 10 cm plates of 293T cells were transfected with 5 μg of plasmid per plate using Polyjet transfection reagent (Froggabio), incubated 48 hours and resuspended in 50 mM Phosphate Buffer pH 8.2 + 500 mM NaCl, 0.2 mM PMSF, 50 μg/ml RNAse A, and lysed using a French pressure cell press. Whole cell extracts expressing either AID-His or native untagged AID were flash frozen in liquid Nitrogen, and stored for activity analysis. Similarly, A3A/B/F/G ORFs were cloned into pcDNA3.1-V5-6xHis-Topo, modified to contain an N-terminal GST tag, and purified using GST-beads per the manufacturer's guidelines. Expression of AID/APOBECs were verified using western blotting probed with anti-GST (SantaCruz) antibodies, followed by the secondary detection by Goat anti-Rabbit IgG (SantaCruz). Western blots to verify the expression of untagged native AID were probed with rabbit polyclonal anti-AID antibody (Abcam) followed by the aforementioned secondary IgG. Relative yield and purity of each AID/APOBEC preparation were evaluated using standard SDS commassie staining. The concentration of 293T-purified AID/APOBECs ranged from 10-50 ng/μL.

Inhibition of AID/APOBEC using the alkaline cleavage assay

The standard alkaline cleavage assay for AID/APOBEC-mediated deamination was used to screen compounds for inhibition of AID, A3A, A3B, A3F and A3G. For AID reactions, we used the standard 7-nucleotide bubble substrate containing the WRC motif TGC (5'-TTTGCTT-3') as a substrate, because it has previously been optimized for the highest levels of AID activity in this enzyme assay (49,52,96). For A3A, we utilized a bubble substrate containing the preferred 5'-TC-3' dinucleotide, while A3B and A3F used a single-stranded oligo containing the preferred 5'-TC-3' dinucleotide (257,344,345) and A3G used a single-stranded 5'-CCC-3' oligo (Figure S5; (93). Substrates were labelled and purified as described previously (52,95,96). For AID reactions, purified substrate (1.7 nM) was incubated with AID enzyme (~0.9 μg of bacterially-expressed GST-AID, 1-10 ng of 293T-expressed AID-His or untagged AID) in phosphate buffer (100 mM,

pH 7.2) with H_2O, 140 mM DMSO as a vehicle control or compound (see below). All AID reactions were incubated at 37°C, except for Dr-AID, which was incubated at 25°C which was previously shown to be its optimal temperature (94,95), in a total volume of 10 μL. For A3A, A3B, A3F and A3G, alkaline cleavage reactions were conducted in the same manner (10 ng of each enzyme used per reaction), except incubated in phosphate buffer (100 mM, pH 6.0), since the A3 family enzymes are optimally active at a more acidic pH as compared to AID (346,347). Concentrated stocks of compound were sonicated at 37°C for four hours in 10% DMSO to promote dissolution. Due to solubility differences between compounds, the highest achieved concentration in 140 mM DMSO was used for initial screening. For initial C1-C10, C4.1-C4.5 and C8.1-C8.15 screening, final alkaline cleavage concentrations (in 140 mM DMSO) ranged from 500-840 μM, with the exception of C9, C8.4, C8.5 and C8.9 which had lower concentrations (310, 400, 170 and 270 μM, respectively).

AID inhibition measurement using the deamination-specific PCR assay

Daudi, Raji and Ramos cells (ATCC = CCL-213, CCL-86, CRL-1596, respectivly) were suspension cultured for 48 hours in RPMI 1640 growth media containing 10% FBS. To lyse the cells, cultures were centrifuged and cells (2.5-3 $x10^6$ cells) were washed twice with PBS. Cell were lysed using glass beads (425-600 μm) (Sigma) . The volume of glass beads used was twice the volume of the cell pellet. The mixture was then vortexed for 20-30 seconds, then incubated on ice for 30 seconds. This was repeated 2-3 more times in PBS with 0.2 mM PMSF, 50 μg/ml RNAse A. To detect AID expression in these cells, we employed quantitative real-time PCR. Total RNA was extracted from 10^6 cultured cells using TRIzol solution (Invitrogen) according to the manufacturer's instructions. The quality and quantity of the extracted RNA were estimated by spectrometry. Total RNA (1 μg) was used for cDNA synthesis with the ProtoScript First Strand cDNA Synthesis Kit (NEB, UK). Quantitative real-time PCR was performed in triplicate using SYBR Green I. The amount of GAPDH housekeeping gene transcripts was used as a reference for the level of AID gene expression. Amplification was carried out in a total volume of 20 μl containing 0.5 μg cDNA prepared as described above, 0.3 μM each of GAPDH- and AID-specific primers, 1 × reaction mixture consisting of RNase-free water and 2x QuantiTect SYBR Green PCR Master Mix (Qiagen). Thermal cycling for both genes was initiated with a denaturation step at 95°C for 10 min, followed by 40 cycles of denaturation at 95°C for 15 s, annealing at 50°C for

30 s and elongation at 72°C for 1 min. Melting curve analysis of the amplification products was performed at the end of each PCR by cooling the samples to 60°C and then increasing the temperature to 95°C at 0.2 C/s. The experiment was repeated three times for statistical analysis.

To detect AID activity in cell extracts, we utilized a deamination-specific PCR as previously described (89,90,93,348). Briefly, the plasmid used as a substrate for the deamination specific PCR assay was pcDNA3.1 V5-6xHis containing a random WRC-rich target sequence. 50 ng Supercoiled plasmid was denatured at 98 °C in 100 mM phosphate buffer pH 7.2 for 10 min followed by snap-cooling in an ice bath to generate ssDNA targets for AID to mutate. 4 µl of cell lysates was mixed either with C8 or vehicle, and added to the target plasmid and incubated for 4 h at 32 ˚C. To detect AID-mediated mutations, 1 µl of each reaction was amplified by deamination-specific nested PCR using mutation-specific primers, as previously described (89,90,93,348). The inner primers used in the second nested PCR reaction generate a 451 nt-product. PCR amplicons were subsequently TA-cloned, and 10 from each reaction were sequenced to confirm AID-mutated mutations C to T, or G to A on the sense and non-sense strands, respectively. Negative control deamination-specific reactions were conducted on the substrate alone with no added extract or AID. Positive control deamination-specific PCRs were conducted on reactions containing substrate plasmid and extracts of AID-expressing 293T cells, as described above. An additional negative control PCR reaction was carried out by adding C8 directly to the PCR reaction of the aforementioned positive control, in order to make sure that inhibition of endogenous lymphoma cell AID by C8 is not due to interference of C8 with the deamination-specific PCR step itself.

MTT Assay

A breast cancer cell line (MCF-7; ATCC = HTB22), lung cancer cell line (A549; ATCC = CCL185), embryonic kidney cell line, (293T; ATCC = CRL3216), B cell lymphoma cell line (Raji; ATCC = CCl-86), and primary Peripheral blood mononuclear cells (PBMCs) from healthy donors were used to test toxicity of C4 and C8. MCF-7, A549 and 293T cells were gifted from Dr. Kao (Memorial university of Newfoundland), while Raji cell lines were gifted from Dr. Hirasawa (Memorial university of Newfoundland). B cell lymphoma cell lines and PBMCs were grown in RPMI-1640 medium (Invitrogen, USA) supplemented with 10% fetal calf serum (Invitrogen,

USA), 200 IU/mL penicillin/streptomycin (Invitrogen, USA), 1% 1M HEPES (Invitrogen, USA), 1% L-glutamine (Invitrogen, USA) and 2.0 $x10^{-5}$ M 2-mercaptoethanol (Sigma-Aldrich, USA) whereas MCF-7, A549 and 293T cell lines were maintained in DMEM containing 100 U/mL of penicillin and 100 μg/mL of streptomycin (Invitrogen, USA), and supplemented with 10% fetal calf serum (Invitrogen, USA). PBMC from an anonymous healthy donor was obtained with approval from the Health Research Ethics Authority of Newfoundland and Labrador, Canada, carried out in accordance with the recommendations of the Canadian Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. All cells were grown at 37°C in a humidified condition containing 5% CO_2. Cells were transferred into a 96-well plate (10^4 cells per well) and treated in 8 replicates with C4 and C8 (100, 250 and 450 μM) separately, or with vehicle alone (140 mM DMSO). Untreated cells were also considered as the negative control. After 24 and 48 h incubation at 37° C, 10 μL of 12 mM MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) solution (Molecular Probes) was added to each well and incubated for 4 hours. The reaction was terminated by adding 100 μL of detergent reagent and incubation for 4 hours in the dark. Colorimetric evaluation was performed using a spectrophotometer at 490 nm. Percentage of viable cells was calculated from the absorbance values of untreated and treated cells as: $\%\text{Viable Cells} = (OD_{490}\text{treated} / OD_{490}\text{untreated}) \times 100$.

Results

Rationale for targeting the catalytic pocket of AID

AID's catalytic pocket is an ideal target for small molecule inhibition for several reasons. First, we have previously gleaned detailed insights into its architecture and conformational breathing (96). Second, the majority of structural differences between AID and related cytidine deaminases are concentrated in the catalytic pocket and proximal surface regions at the pocket opening (58,59,95,96,164,258), thus allowing maximum specificity. Third, we and others have described the pocket-adjacent main ssDNA binding groove (groove 1) in detail (54,89,96,164), thus offering a promising target for future derivatization.

We reasoned that the functional and breathing structure of AID's catalytic pocket is an advantageous template for inhibitor search for several reasons. First, the architecture of its catalytic

pocket was extensively functionally-verified by testing of mutants and chimeras(164). Second, using multiple catalytic pocket conformations is advantageous since AID/APOBEC enzymes contain highly flexible loops that compose their catalytic pocket leading to a range of different catalytic pocket conformations for each enzyme (58,59,96,164,341). Third, purely homology modeled structures, using the same homology modeling methodology used to arrive at AID's functional structure, were shown to be as reliable for generating inhibitor "hits", as crystal structures (349). Considering that the functional AID structure, especially of its inhibitor-target catalytic pocket, was backed by extensive biochemical validation of homology modeling predictions using AID mutants and chimeras (96,164), we reasoned that it ought to represent an even more high confidence template for inhibitor design. Lastly, though at the time of these experiments, there were no AID crystal structure available, the recently published crystal structures by Qiao et al. and Pham et al. (54,245), have confirmed the accuracy of the functional and native AID structure used here for inhibitor search (Figure S6; (96).

Structure-based virtual screening of small molecules against the catalytic pocket of AID

To carry out a structure-based docking search for small molecules that bind in its catalytic pocket, we included 5 different conformations of AID which are representative of the full range of conformations with accessible catalytic pockets (Figure 22A, B and C) (137). This strategy would allow for selection of compounds that bind all active AID conformations. For small molecules, we utilized the ZINC database which contains over 100 million structures of commercially available compounds (250,251). We restricted our search to the "clean lead" subset of 4.6 million compounds for several reasons. First, this set contains compounds that are lead-like; defined by pharmacological properties such as Lipinski's rule of 5 and properties generally amenable for oral intake (350,351). Second, the subset is composed of molecules with benign functionality and excludes those with potentially toxic chains such as aldehydes and thiols.

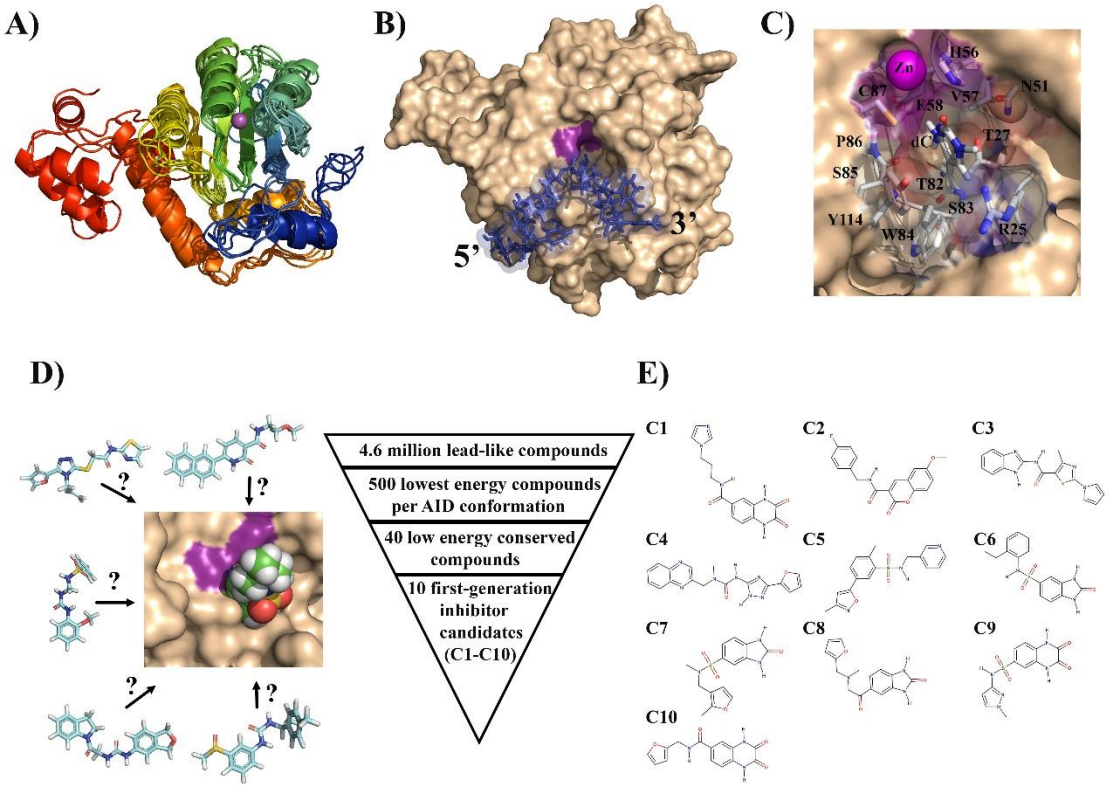


Figure 22. Virtual high-throughput screening of drug-like small molecules against the catalytic pocket of AID. A. Ensemble of AID conformations covering the range of accessible catalytic pockets dynamics used for virtual high-throughput screening. The structures exhibit conserved overall structure, with conformational changes localized to the secondary catalytic loops that compose the walls and floors of the catalytic pocket. N- to C-termini progression is shown from blue to red. The purple sphere depicts the catalytic pocket-coordinated Zn. B. Representative surface structure of a catalytically-productive AID:ssDNA complex with docked ssDNA (blue) and dC poised for deamination in the catalytic pocket (magenta). C. dC bound in a catalytically-productive configuration in the catalytic pocket highlighting the secondary catalytic residues. Adjacent ssDNA structure was omitted for clarity. This and other energetically similar conformations of the accessible catalytic pocket state of AID served as template for virtual high-throughput screening. D. The structure-based virtual high-throughput screening scheme. We screened a large library of lead-like compounds from the ZINC library against the catalytic pocket of AID. Using several independent and complementary docking protocols to the catalytic pocket alone or the entire surface of AID, we identified 10 low-energy compounds predicted to bind in the catalytic pocket. E. Structures of the 10 first-generation inhibitor candidates (C1-C10).

Using DOCK Blaster (342), we virtually screened for candidates that bind AID in a search space restricted to the catalytic pocket and proximal surface region (~10 Å; Figure 22C and D). For each AID conformation we identified the 500 lowest-energy compounds. For further

inspection, we prioritized compounds that bound several AID conformations over those that bound only to a single conformation and selected the 40 lowest binding energy compounds. To confirm specificity for the catalytic pocket, docking was repeated using the entire surface of AID, rather than just the catalytic pocket region. In addition, we employed a second independent docking algorithm, AutoDock Vina(188), to substantiate the 40 candidates identified by Dock Blaster. Even with access to AID's whole surface, 27 of 40 compounds bound preferentially in the pocket, thus validating our screening methodology whilst further refining the list of hits. The 10 inhibitor candidates with the most favorable docking energies (C1-C10) were selected for functional testing (Figure 22E).

First generation hits that inhibit purified and endogenous AID

Using the standard alkaline cleavage enzyme assay for deamination activity of purified AID/APOBEC enzymes, we measured the catalytic activity of bacterially-expressed and purified GST-AID in the presence of each compound (Figure S7). We found that 2 of 10 compounds (C4 and C8) diminished AID activity (8.3 ± 1% and 17.4 ± 1% AID catalytic activity, respectively; Figure 23A). In addition to purified GST-AID, we tested inhibition on whole cell extracts of AID-expressing 293T cells transfected with a CMV-promoter based AID-His expression vector (Figure 23B). Akin to our results with bacterially expressed GST-AID, C4 and C8 inhibited this AID as well (43.9 ± 0.3% and 40.3 ± 1.8% AID catalytic activity, respectively). To assess potency, we measured the dose-response of C4 and C8 against GST-AID and the AID-expressing 293T cell lysate (Figure 23C and D). C4 and C8 showed a similar potency in GST-AID (IC_{50}= 290 and 230 μM, respectively) as with AID-His (IC_{50}= 360 and 270 μM, respectively).

To evaluate toxicity, we incubated multiple cell lines originating from different tissues (A549, MCF-7, 293T and Raji), as well as primary peripheral blood monocytes (PBMC) with C4 or C8 and measured viability using the standard MTT assay (Figure S8). C4 caused minor toxicity towards all anchorage-dependent cells (A549, MCF-7 and 293T) but not Raji or PBMCs, while C8 was found not to be toxic. We thus focused on C8 for further development.

To confirm that C8 was a *bona fide* AID inhibitor we measured off-target inhibition of UDG, an enzyme used downstream of AID in the alkaline cleavage deamination assay, and found

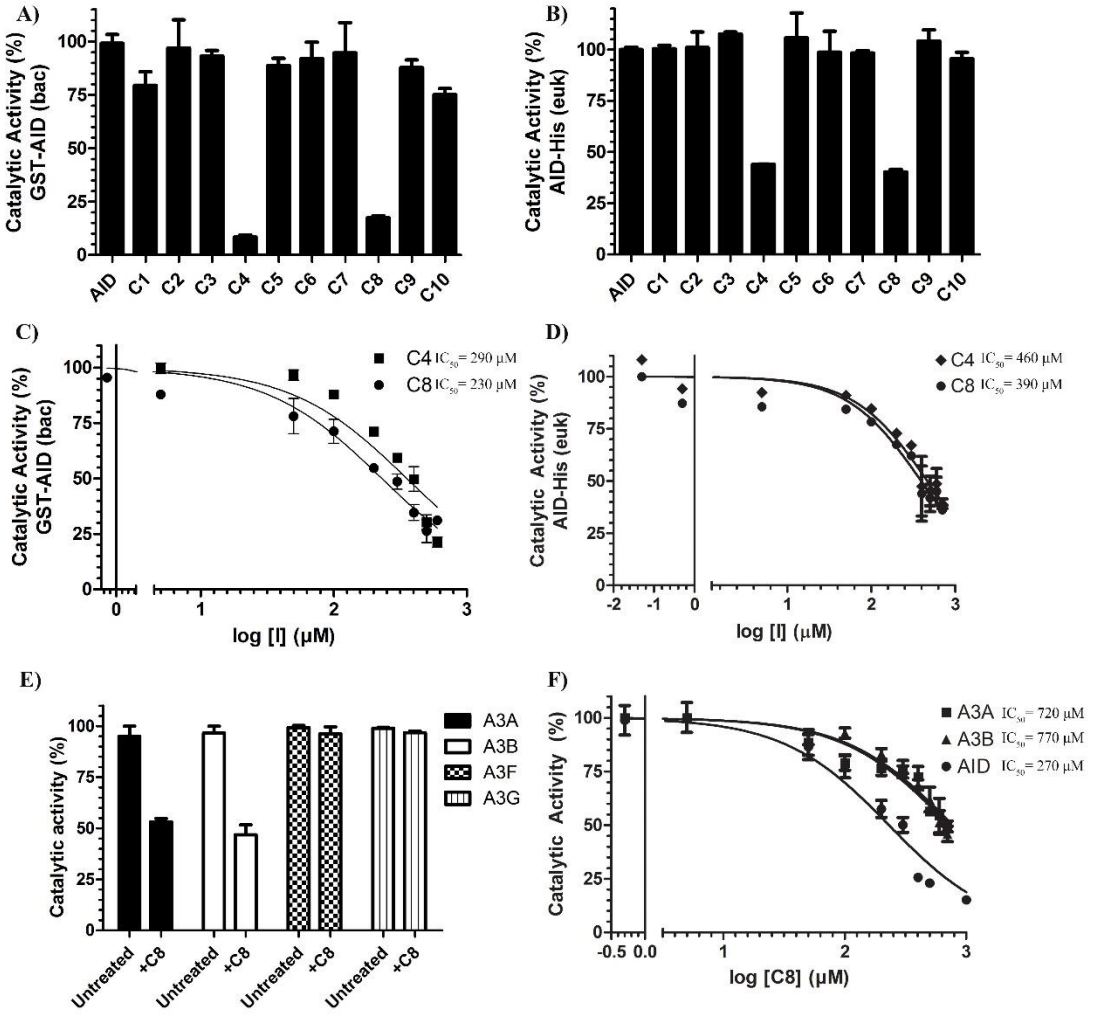


Figure 23. First-generation inhibitor candidates inhibit purified AID and AID in whole cell extracts. A. Catalytic activity of bacterially-expressed and purified GST-AID treated with C1-C10 (n=6 independent experiments conducted with 3 independently-purified preparations of GST-AID). B. Catalytic activity of eukaryotic-expressed AID in whole 293T cell lysate treated with C1-C10 (n=3 independently prepared AID-expressing whole cell extracts). C. Catalytic activity of GST-AID on C4 and C8 as a function of log inhibitor concentration. D. Catalytic activity of AID-His 293T lysate as a function of log inhibitor concentration. E. Catalytic activity of A3A, A3B, A3F and A3G treated with C8. F. Catalytic activity of A3A and A3B as a function of log C8 concentration. All experiments contained a negative control vehicle-only (140 mM DMSO) reaction which was designated as 100% AID activity. All AID reactions were performed at 37°C for 2-4 hours at pH 7.2 using 2 nM of the standard bubble oligonucleotide substrate TGCbub7 which has previously been demonstrated to be AID's most favored substrate in the alkaline cleavage assay. A3A, A3B, A3F and A3G reactions were incubated at 37°C for 2 hours in pH 6.0 using 2 nM of standard single-stranded oligonucleotide substrates containing a single target TTCA motif for A3A, A3B and A3F and a single target CCC motif for A3G (Figure S5).

C8 did not inhibit UDG (Figure S7C). To test specificity, we examined whether C8 could inhibit the catalytic activity of homologous APOBEC3A (A3A), APOBEC3B (A3B), APOBEC3F (A3F) and APOBEC3G (A3G). We found modest inhibition of A3A and A3B (52 ± 2% and 45 ±4% activity, respectively), however, neither A3F nor A3G were inhibited by C8 (Figure 23E). However, A3A and A3B were inhibited to a lesser degree than AID inhibition (IC_{50} = 720 and 770 μM, respectively vs 270 μM AID inhibition; Figure 23F). Thus, C8 acts as a specific AID inhibitor, with much lesser inhibition efficacy for the related A3A and A3B tumorigenic enzymes, and no inhibition of the anti-viral A3F and A3G family members.

C8 inhibition of multiple forms of AID (bacterially expressed GST-AID and AID-His in whole cell lysates of 293T cells) was reassuring. We then examined the ability of C8 to inhibit native (no fusion tag) full-length AID expressed in 293T cells and found that surprisingly it was even more effective at inhibiting untagged AID in whole cell lysates (IC_{50}= 10 μM) than purified AID. This result prompted us to examine the ability of C8 to inhibit the endogenous AID of AID-expressing lymphoma cell lines (Raji, Daudi and Ramos). Given that endogenous AID levels are significantly lower than that of 293T cells transfected with CMV promoter-driven AID expression vectors, we were unable to detect AID activity from cell lysates in the alkaline cleavage assay. Instead, we employed a more sensitive semi-quantitative PCR-based assay which we have previously established for measuring AID activity (Figure S9; (89,90,93,348). In this assay, a substrate plasmid is incubated with AID and subjected to PCR using deamination-specific primers that amplify DNA deaminated by AID. Using this assay on C8-treated extracts of all three lymphoma cells, we observed inhibition. (Figure 24B, S9).

C8 inhibition is sensitive to mutation of secondary catalytic and DNA binding residues

Docking revealed that C8 binds into the catalytic pocket of AID with impressive fit (Figure 25A). Akin to AID's native substrate dC (Figure 22C), C8 is stabilized through multiple interactions with secondary catalytic residues in the catalytic pocket (Figure 25B, C; (96). These secondary catalytic residues are housed on four secondary catalytic loops: Loop 2 (L2), 4 (L4), 6 (L6) and L8 (L8) (L2, L4, L6 and L8 correspond to the β1-α1 loop, β2-α2 loop, β3-α3 and the β4-α4 loop,

respectively; (96). The peptide bond of S85 and side-chains of T27, N51 and E58 hydrogen bond

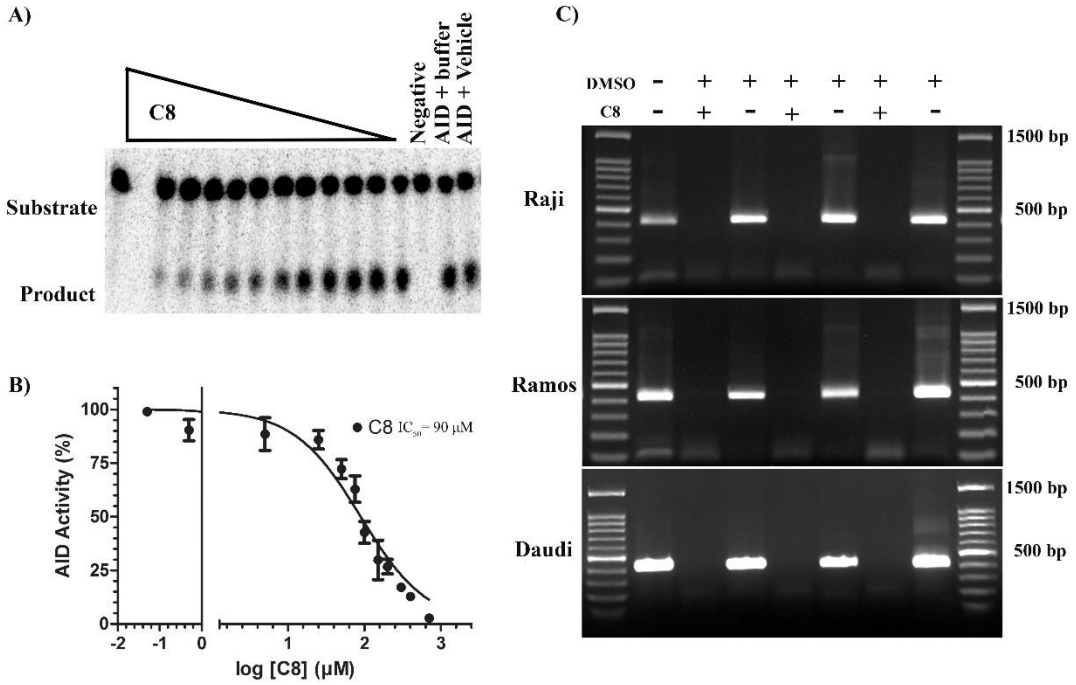


Figure 24. C8 inhibits native AID and endogenously expressed AID of B lymphoma cells. A. Representative alkaline cleavage experiment demonstrating C8 inhibition of untagged native AID in 293T whole cell lysate. B. Catalytic activity of native untagged AID in 293T lysate as a function of log inhibitor concentration. C. Deamination-specific PCR was used to detect activity of endogenous AID in extracts of AID-expressing B lymphoma cell lines, with or without addition of C8. The deamination substrate plasmid was incubated with cell extracts containing endogenously expressed AID from Raji (top panel), Ramos (middle panel) and Daudi (bottom panel) with added vehicle (140 mM DMSO) or C8. Incubation with C8, but not DMSO, abrogated any detectable PCR product.

with the benzimidazole-2-one group of C8 in the catalytic pocket. Notably, these are the same residues that interact with AID's native substrate dC for stabilization and deamination in the catalytic pocket (54,96). Additionally, W84 and Y114 partially base-stacked across several configurations of C8, analogous to dC. Beyond the benzimidazole-2-one group, the connecting carbonyl was hydrogen bonded by Y114 and R25 side-chains and the methylamino group was hydrogen bonded by the peptide bond of R25. The furan group of C8 bound to ssDNA binding groove 1 of AID with some flexibility; the majority of predicted AID-C8 complexes bound the furan group in the L2-L4 interface, nestled between the R25, E26, N51 and K52 (Figure 25C, left and middle panel) while lower-energy modes bound the furan in the L2-L8 interface between R25,

W84 and Y114 (Figure 25C, right panel). Thus, the benzimidazole-2-one group was firmly bound

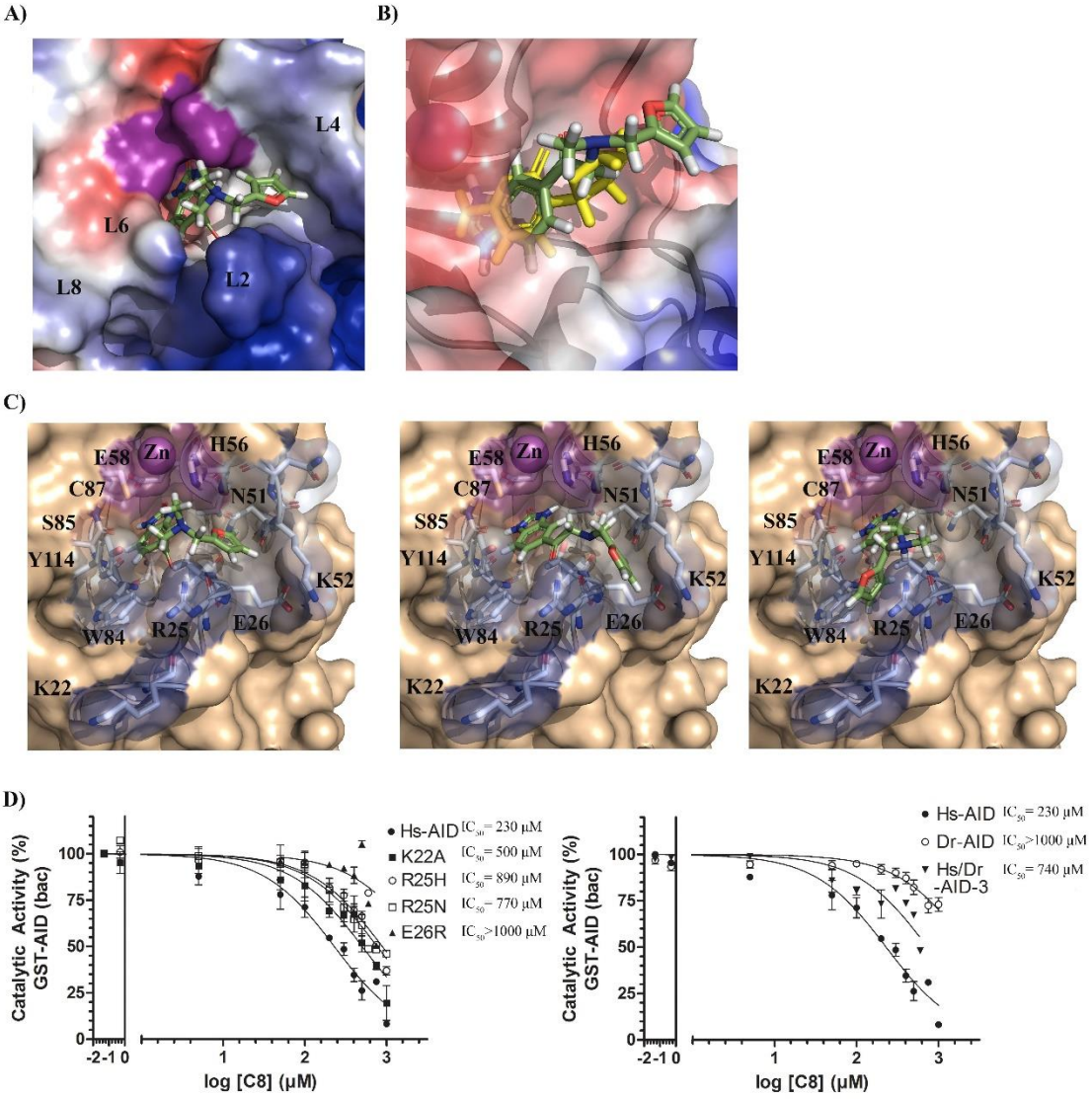


Figure 25. C8 forms multiple interactions with the catalytic pocket and ssDNA binding residues of AID. A. Representative AID-C8 docked complex illustrating a low-energy binding mode across the catalytically accessible conformations of AID. The C8 benzimidazole scaffold was bound deep in the catalytic pocket with the tail region bound outside, stabilized by several ssDNA binding residues in the L2-L4 interface. B. Superimposition of C8 (green) and ssDNA dC (yellow) bound in the catalytic pocket. For clarity, the ssDNA structure connected to dC is hidden. C. C8 binds to the catalytic pocket and surrounding ssDNA groove 1 in several configurations. Dominant AID-C8 complexes favor binding of the C8 tail in the L2-L4 interface (left and middle panel), with less dominant complexes bound in the L2-L8 interface (right panel). For C8, green represents the carbon backbone, while blue and red represents nitrogen and oxygen, respectively. D. C8 inhibition of AID mutants of residues predicted to interact with C8 by docking analysis, demonstrating that mutation of ssDNA binding groove residues that directly stabilize C8 results in

resistance to C8 inhibition. C8 inhibition of zebrafish AID (Dr-AID) and a zebrafish/human chimeric AID which contains human AID's catalytic pocket and the main ssDNA binding groove 1 with other regions contributed from Dr-AID (Hs/Dr-AID-3). C8 was ineffective against Dr-AID, while Hs/Dr-AID-3 was inhibited with close efficiency to human AID.

in the catalytic pocket, whilst the tail was predicted to adopt several configurations interacting with the main ssDNA binding groove 1.

To verify these docking predictions, we measured C8 inhibition on AID variants with mutation of residues predicted to interact with C8 (Figure 25C). We were limited in testing AID mutated at the secondary catalytic residues that are predicted to interact with C8 because even conservative mutations of these (ie: N51Q and Y114F) result in catalytically dead mutants (96); however, we tested the ability of C8 to inhibit AID mutated at each of the other contact residues located at the opening of the catalytic pocket (K22A, E26R, R25H and R25N) and found all were resistant to C8 inhibition when compared to wild type (IC_{50}= 500, 750, 820 and 820 μM, respectively). Zebrafish AID (Dr-AID) which naturally has a R25H equivalent, was also resistant to C8 inhibition (IC_{50} > 1000 μM). Hs/Dr-AID-3 is a previously described zebrafish/human AID chimera wherein part of a ssDNA binding groove 2 from Dr-AID was grafted into Hs-AID, while retaining the catalytic pocket of Hs-AID (96). Accordingly, this chimera was inhibited by C8. These data provide functional validation for the predicted AID-C8 interactions by docking and highlight the specificity of C8 for the catalytic pocket and adjacent regions of human AID. We noted that the benzimidazole-2-one group of C8 bound to the catalytic pockets of A3A and A3B-CTD in a similar fashion (Figure S10). However, we noted the tail of C8 bound differently and with less fit to the outside of A3A and A3B's catalytic pockets, due to the different surface structure of A3A and A3B-CTD in comparison to AID. This observation is consistent with C8 inhibition of A3A and A3B with lesser potency as compared to AID.

Structural analogues of C8 exhibit increased AID inhibition potency

Using the ZINC database, we identified 948 structural analogues of C8 using a 60% structural similarity cutoff. We docked each analogue with several conformations of catalytically accessible AID, using the same approach described in Figure 22. Analogous to our initial screening approach

for identifying C1-C10, we ranked C8 analogues based upon docking energy and chemical diversity and selected 15 analogues (Figure S11; (C8.1-C8.15). C8 analogues were tested for inhibition of purified AID (Figure 26A), and several were found to inhibit AID. 3 out of 15 analogues inhibited AID with modestly improved IC50 compared to C8; these included C8.5, C8.7 and C8.12 (IC_{50}= 140, 130 and 160 μM, respectively, compared to 230 μM for C8; Figure 26B). We performed a similar screen for C4 analogues (C4.1-C4.5), but none were capable of inhibiting AID (Figure S12).

Since C8, 8.5, 8.7 and 8.12 were all similarly efficient in inhibiting AID, we examined common structural motifs of these analogues (Figure 26C, top panel). Like C8, all analogues with improved potency contain the conserved benzimidazole-2-one group, while the tail structure contains three hydrogen-bond acceptor groups. C8, 8.5, 8.7 and 8.12 bound similarly to AID, with the benzimidazole-2-one group anchored in the catalytic pocket, while the flexible tail bound between the L2/L4 interface (Figure 26C, bottom panel). Depending on the bound configuration, the tail carbonyl hydrogen bonded with N51 or Y114, analogous to hydrogen bonding with the dC carbonyl group (96). The other two hydrogen bond acceptor groups in the tail were stabilized by several interactions in the L2-L4 interface. Thus, C8 and analogues form multiple specific interactions with residues that compose AID's catalytic pocket.

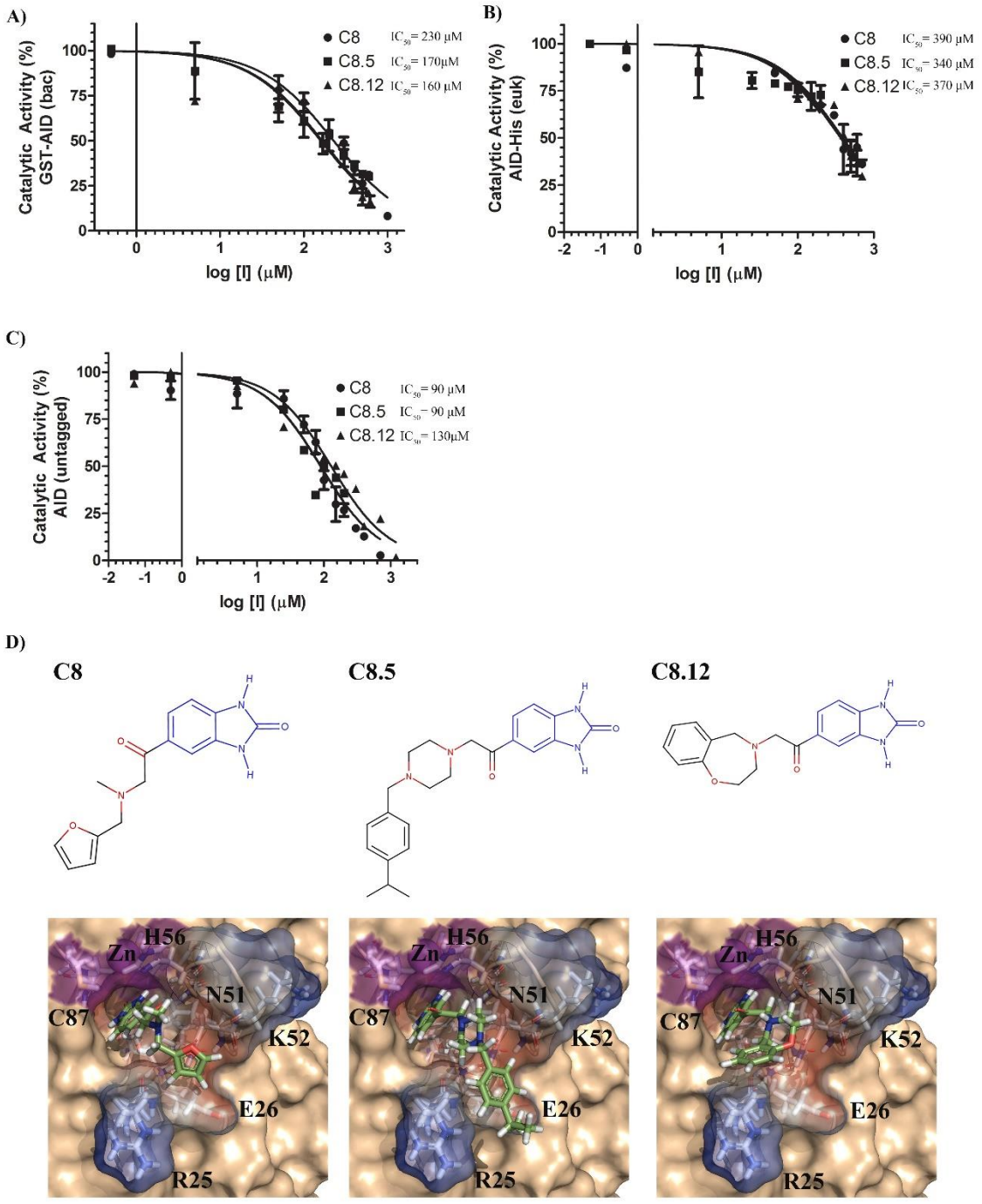


Figure 26. Structural analogues of C8 exhibit improved potency and similar binding modes in the catalytic pocket. A. The catalytic activity of bacterially-expressed purified GST-AID as a function of treatment with 15 structural analogues of C8. B. Catalytic activity of GST-AID as a function of log inhibitor concentration. In comparison to the parent C8, we found C8.5, C8.7 and C8.12 were moderately more effective. C. Top panel shows the chemical structure of C8 and analogues with similar or improved inhibition of AID. Effective inhibitors had the conserved motifs of the benzimidazole-2-one scaffold (blue) and three hydrogen-bond acceptor groups in the tail region

(red). Bottom panel shows representative AID-inhibitor docked complexes with each of the effective structural analogues compared to the parent C8. The benzimidazole-2-one group binds to the catalytic pocket in a similar configuration, with the lowest-energy complexes containing the tail motif of the inhibitor bound to the interface between L2 and L4 of AID. The carbon backbone of C8 is colored green, with nitrogen colored blue and oxygen colored red. For each inhibitor, green represents the carbon backbone, while blue and red represents nitrogen and oxygen, respectively.

Discussion

Since AID expression drives and exacerbates tumorigenesis, an AID inhibitor has been suggested to be of benefit; however, development of such an agent has not been possible since structural insights into AID were not gleaned until recently. The functional and native structure of AID described in 2015 revealed two structural features that explained the unusually low catalytic rate and high ssDNA binding affinity of AID (96). These include frequent catalytic pocket closure and sporadic ssDNA binding by a highly positively-charged surface, in positions that are not deamination-viable. This structural analysis, which has since been confirmed by two partial AID crystal structures (54,245), provided an understanding of AID's catalytic pocket conformational dynamics and ssDNA stabilization interactions proximal to the pocket (Figure 22C); thus offering an opportunity for structure-based inhibitor design.

Using these AID-DNA interactions, we scanned the ZINC database of lead-like compounds and identified 10 compounds for testing against AID (Figure 22D and 22E). We identified two compounds, C4 and C8, that inhibit the enzymatic activity of AID (Figure 23). MTT assay showed that C8 was non-cytotoxic across several cell lines and primary healthy donor cells tested (Figure S8). This suggested that C8 or derivatizations could be used successfully in future *in vivo* studies and hence we focused on C8. Although purified GST-AID was inhibited by C8 with modest IC_{50} values (230 μM), C8 could inhibit multiple forms of purified AID (both bacterially expressed GST-AID and 293T-expressed AID-His), as well as AID in whole cell lysates of 293T cells. It was particularly encouraging that native untagged AID in whole cell 293T extract, which is the best representative of endogenous cellular AID, was inhibited by C8 with the most effective $IC_{50,}$ of 10 μM, ~ 23-fold more effectively than purified AID. This increased susceptibility of C8 inhibition is likely due to the higher specific activity of native untagged AID, thus requiring a lower amount of enzyme to achieve similar catalytic rates when compared with purified GST- or His- tagged AID.

Since the topological features of the AID catalytic pocket and surrounding region are unique, we rationalized these residues ought to act as specific anchors for small molecule placement. Furthermore, inhibitor binding in this region would impart a level of structural specificity to our strategy. Akin to dC stabilization, we noted several stabilizing interactions with secondary catalytic residues and DNA binding residues in and proximal to the catalytic pocket, respectively (Figure 25A-C). The benzimdazole-2-one group was anchored in the catalytic pocket, while the tail group adopted several conformations in the DNA binding groove. To bolster our understanding of the AID-C8 complexes we probed several non-lethal surface mutants predicted to destabilize C8 binding. DNA binding groove mutants (K22A, R25H, R25N and E26R) all rendered resistance to C8 inhibition with increased IC_{50} values (Figure 25D). Dr-AID too showed resistance to C8, while Hs/Dr-AID-3 was largely susceptible, due to the retention of Hs-AID's catalytic pocket in this chimera. Interestingly, R25H and R25N represent natural variants in the AID/APOBEC family members that are more resistant to C8 inhibition (H29 in A3A, H216 in A3G, H29 in Dr-AID and N214 in A3F; Figure 23E and F). Poised at the mouth of the catalytic pocket to orchestrate ssDNA binding, R25 exemplifies a specific anchoring residue imbedded in our AID inhibition strategy; C8 binding with R25 is predicted to block deamination-feasible AID-DNA complexes, while homologous APOBECs remain largely unaffected.

Using C8 as a parent compound we examined 948 structural analogues, obtaining 15 (C8.1-C8.15) for analysis. Analogues lacking the protruding carbonyl on the benzimidazole group (C8.2, C8.3, C8.4, C8.13 and C8.14) resulted in a major loss of AID inhibition (Figure S11). Loss of the benzimidazole carbonyl likely resulted in a loss of stabilizing hydrogen bonds with secondary catalytic residues and E58. C8.5, C8.7 and C8.12 inhibited AID with a modestly improved potency relative to C8 (Figure 26B). Akin to C8, all 3 analogues contain a similar tail structure with a carbonyl, two atoms away from a Nitrogen and five atoms away from an intra-cyclic Nitrogen/Oxygen, comprising three hydrogen bond acceptor groups in the tail (Figure 26C, top panel). Similar to C8, the tail group of C8.5, C8.7 and C8.12 is predicted to bind preferentially in the L2-L4 interface, stabilized by several DNA binding residues (R25, E26, N51 and K52; Figure 26C, bottom panel). Thus, derivatization of the C8 tail structure produced analogues that were more effectively stabilized in the L2-L4 interface on the surface of AID.

Although the homologous human A3F or A3G, or the orthologous Dr-AID were not

inhibited by C8, weak inhibition of A3A and A3B suggests their catalytic pocket and surrounding region are somewhat conducive to C8 binding. Thus, it is possible that C8 could in the future be derivatized in a similar manner as described for AID, but for the surface-proximal regions of the catalytic pockets of A3A and A3B. This effort could achieve either specific inhibition of A3A or A3B, or pan inhibition of AID, A3A and A3B, three tumorigenic AID/APOBEC family member enzymes. In addition to potential for therapeutic development, the demonstration that C8 and analogues which were identified based on docking into AID's catalytic pocket structure, could indeed functionally inhibit AID, provides further verification of the pocket structure itself. In the future, these and other derived inhibitors could also be useful as biochemical probes for studying the interaction of AID with DNA/RNA, and as novel tools for studying AID biology.

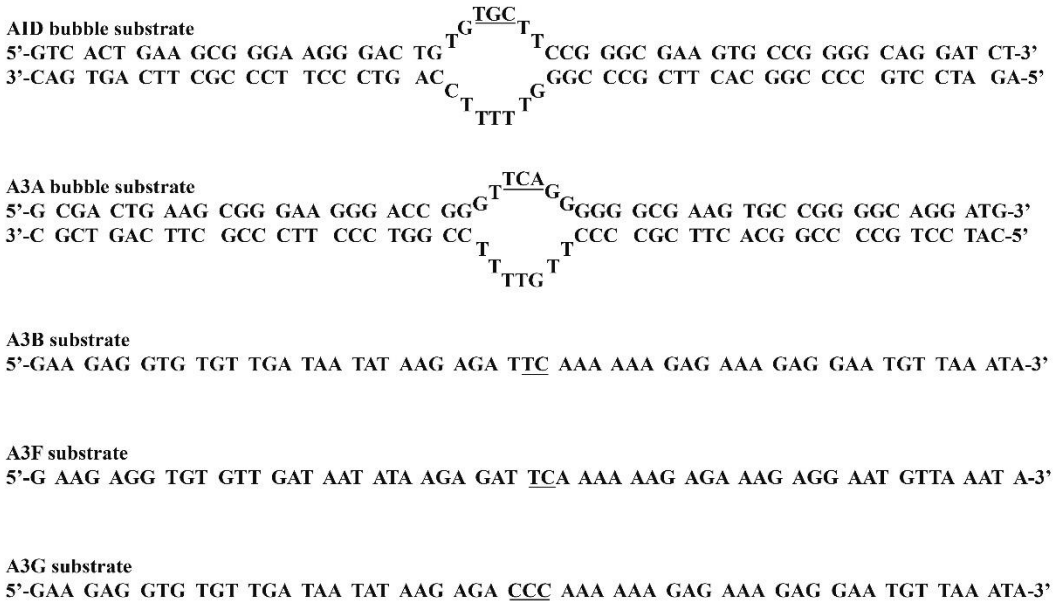


Figure S5. AID, A3A, A3B, A3F and A3G oligo substrates. The nucleotide sequence of substrate used in inhibition assays for each enzyme.

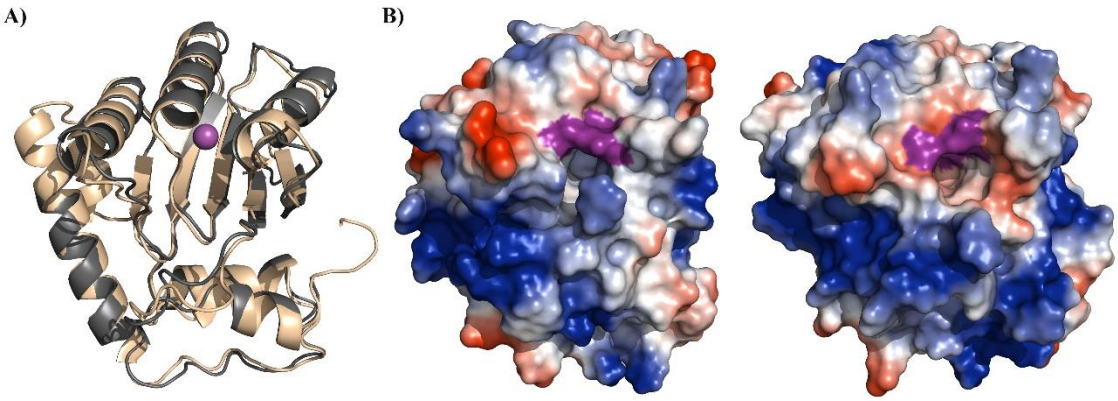


Figure S6. Comparison of the AID structure obtained by different methods. A. The crystal structure of near-native partially truncated AID (PDB: 5W1C) (gray colored) superimposed with a representative conformation of native AID published previously using our combined computational-biochemical method (beige colored). The aforementioned crystal structure lacks the α7 C-terminal domain of AID and contains a mutation (E58A) in the active site. The purple sphere depicts the coordinated Zn in the active site. B. The electrostatic surface of the crystal structure of AID (left panel) and a representative AID conformation from the AID ensemble obtained using the combined computational-biochemical method (right panel). Positive, neutral, negative and catalytic residues are represented by blue, white, red and purple surfaces, respectively.

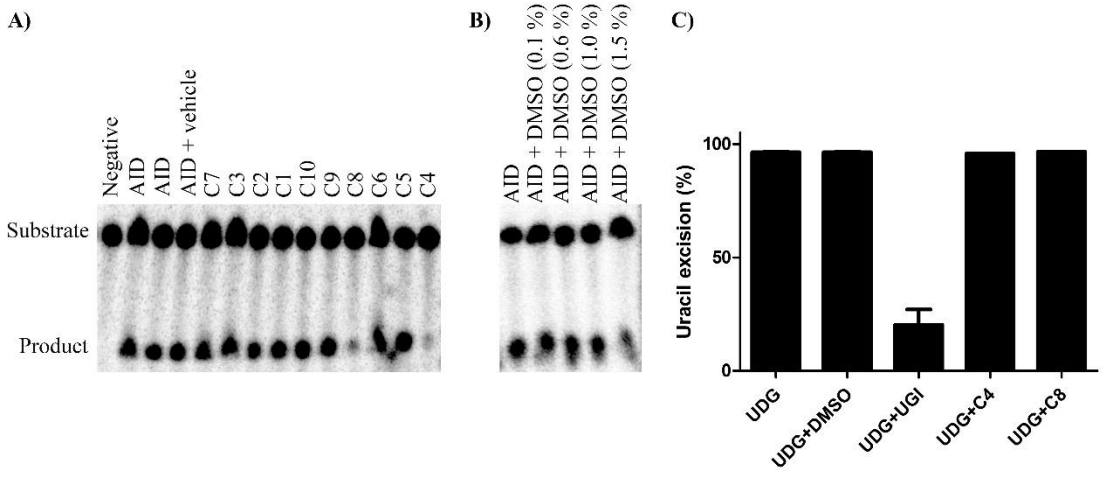


Figure S7. Representative alkaline cleavage gel of AID inhibition by C1-C10 panel and UDG inhibition assay. A. The left panel shows a representative preparation of bacterially-expressed purified GST-AID incubated with C1-10, with AID in phosphate activity buffer used as positive control, AID + vehicle (DMSO) as the 100% bench-mark reaction, and substrates alone (no AID

added) as negative control. Multiple independently purified preparations of AID (n=3-6) were used in each experiment. B. Since the candidate inhibitor compounds were dissolved in DMSO, AID activity was tested as a function of increasing DMSO concentrations. Results demonstrated that the DMSO concentrations used to solubilize the compounds do not impact AID activity negatively. C. Since UDG is used downstream of AID in the alkaline cleavage assay, we sought to ensure that C4 and C8 are not acting through inhibition of UDG. C4 and C8 did not inhibit UDG activity in comparison to a UDG + UDG inhibitor positive control (20 ± 7% UDG activity).

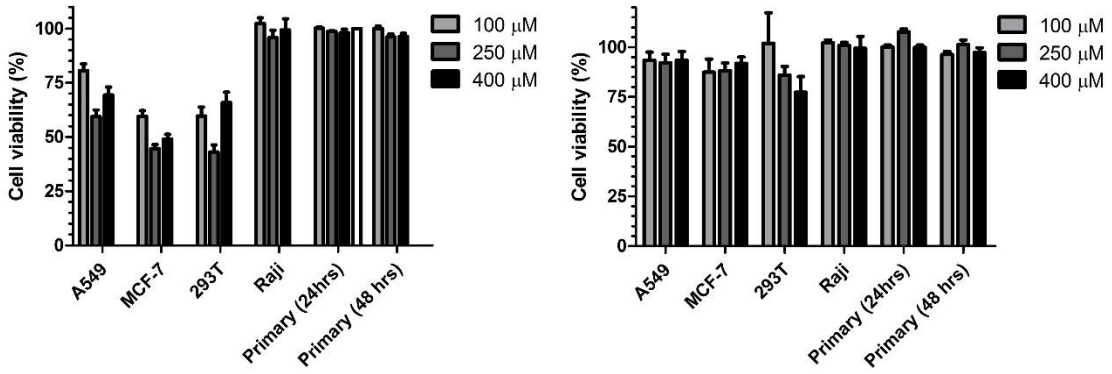


Figure S8. MTT assay to evaluate toxicity of compounds on cell lines and primary healthy cells. Four cell lines of different tissue of origin were used. These included A549 (lung), MCF-7 (breast), 293 T (human embryonic kidney) and Raji (B cell lymphoma). Cells were incubated in triplicate with varying concentrations of C4 (left panel) and C8 (right panel). Untreated cells and vehicle treated cells were considered as negative controls. After 24 hours incubation at 37 °C, 10 µl of MTT reagent [3-(4,5-dimethylthiazol-2-yl)-2,5- diphenyltetrazolium bromide, 12 mM] was added followed by incubation for 4 hours and addition of 100 µl of SDS-HCl solution to each well, and incubation for 18 hours at 37 °C. Colorimetric evaluation was carried out using a spectrophotometer at 570 nm, and percentage of cell growth was calculated from the absorbance values of untreated and treated cells as % cell growth = (OD_{570} treated / OD_{570} untreated) × 100. To test toxicity on healthy primary cells, we tested peripheral blood monocytes (PBMC) from donors; 3×10^4 cells per well were transferred into same 96-well plates and treated with same concentrations of compounds and incubated for 24 and 48 h. C8 was non-toxic to all cells, but C4 exhibited some toxicity to 3/5 cell types tested.

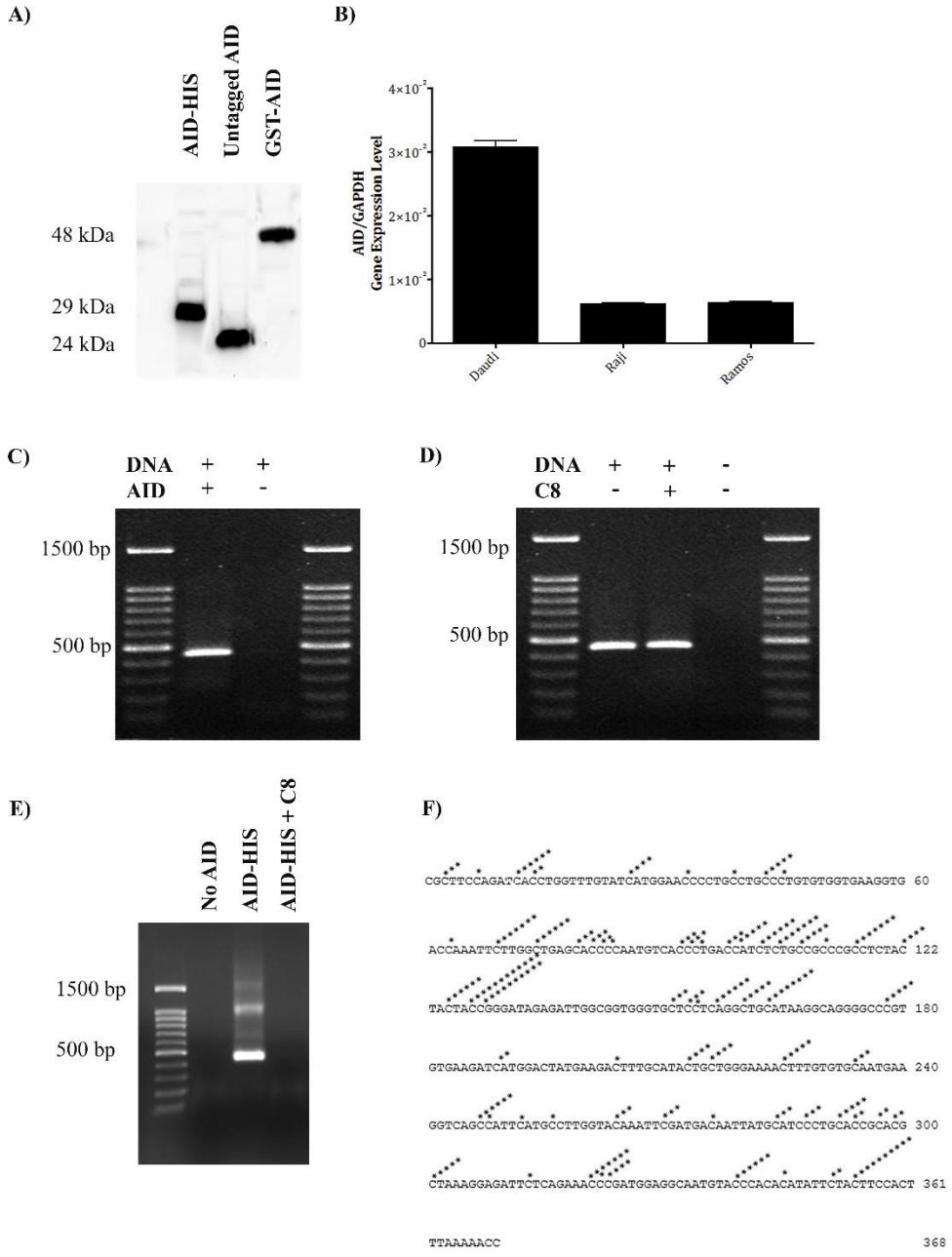


Figure S9: AID expression and deamination-specific PCR assay controls. A. Example western blot showing expression of untagged (native) AID compared with purified GST-AID and AID-His in 293T cells. B. Quantitative real-time PCR of AID expression levels relative to GAPDH expression levels in Daudi, Raji and Ramos cells indicating these cells express AID. C. Positive control reaction of AID-expressing 293T extracts incubated with substrate plasmid to demonstrate the deamination specificity of the PCR assay. D. Control PCR with C8 added after a purified AID + plasmid incubation to ensure C8 does not inhibit PCR. E. Deamination-specific PCR illustrating that like untagged AID, AID-His in 293T cell extracts is also inhibited by C8. F. Representative sequencing analysis of TA-cloned amplicons confirming the deamination specificity of the PCR assay used to detect AID activity in this experiment

(mutated dC denoted by *).

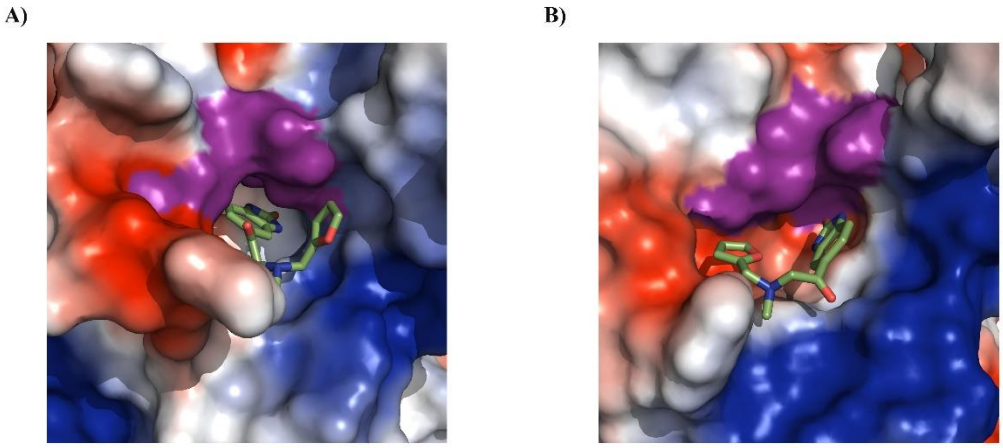


Figure S10. C8 binds to the catalytic pocket of A3A and A3B-CTD. A. C8 docked with a catalytically accessible conformation of A3A (PDB:2M65) B. C8 docked with a catalytically accessible conformation of A3B-CTD (PDB: 2NBQ). In both A3A and A3B-CTD, C8 binds to the catalytic pocket via the benzimidazole-2-one group, with the tail region binding differently to each structure dependent upon the surface structure.

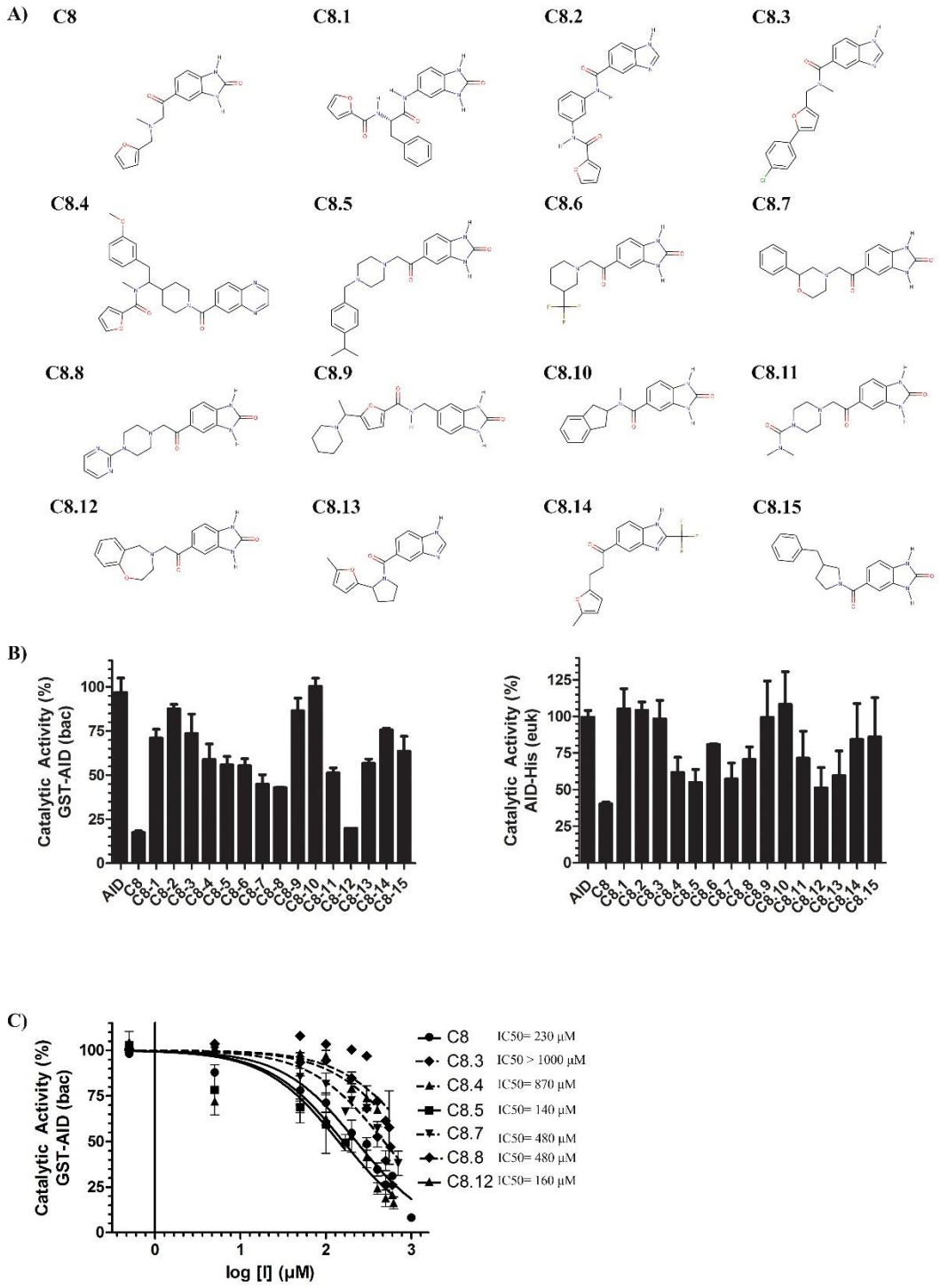


Figure S11: Structural analogues of C8 exhibit improved inhibition of AID. A. The chemical structure of C8 and C8.1-C8.15 structural analogues. B. The catalytic activity of bacterially expressed purified GST-AID as a function of treatment with 15 structural analogues of C8. C. The catalytic activity of AID-His-expressing 293T cell extracts as a function of treatment with 15 structural analogues of C8. D. Catalytic activity of GST-AID as a function of log inhibitor concentration. In comparison to the parent C8, we found C8.5 and C8.12 were moderately more

effective.

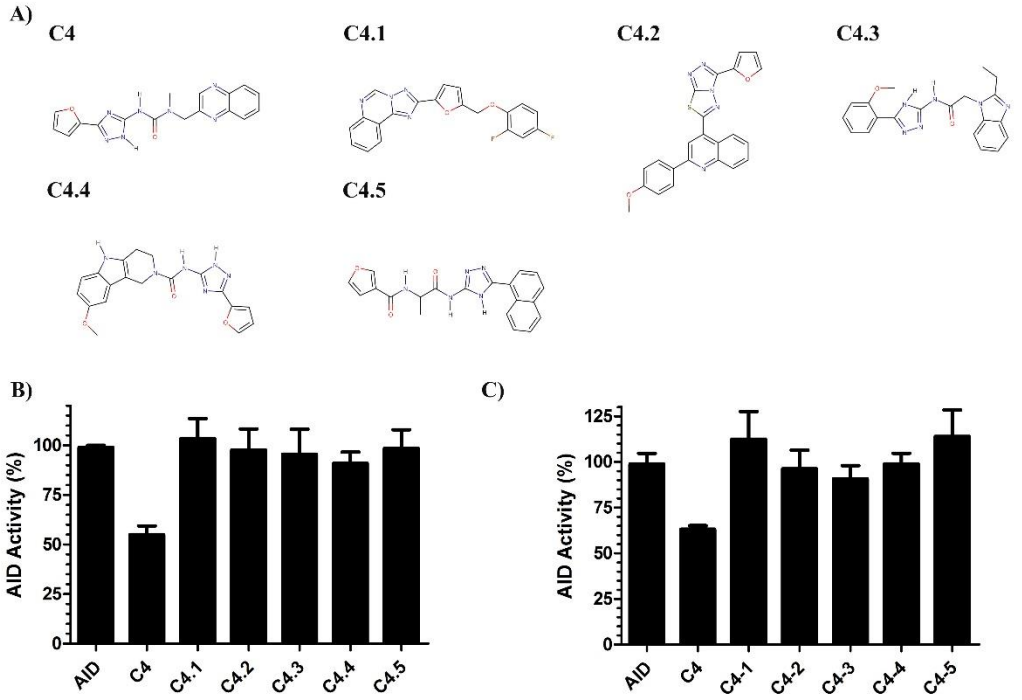


Figure S12. Structural analogues of C4 do not inhibit AID. A. Chemical structure of C4 and C4 analogues. B. The catalytic activity of bacterially expressed purified GST-AID treated with C4 and C4.1-C4.5 analogues. C. The catalytic activity of 293T-expressed AID-His treated with C4 and C4.1-C4.5 analogues.

Chapter 6: Structural plasticity of DNA/RNA binding by AID as a regulator of its activity

Justin J. King*, Mani Larijani*

Program in immunology and Infectious Diseases, Division of Biomedical Sciences, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland A1B 3V6, Canada.

*To whom correspondence should be addressed: Tel: 011-1-709-864 6079; Fax: 011-1-709-864 6007; Email: jjk458@mun.ca, mlarijani@mun.ca

Abstract

Activation induced cytidine deaminase (AID) is a DNA editing enzyme that acts as the initiator for somatic hypermutation (SHM) and class-switch recombination (CSR) in activated B lymphocytes. AID preferably targets the immunoglobulin (Ig) loci, but also acts genome-wide, causing mutations and lesions that drive tumorigenesis in lymphoid/non-lymphoid tissues. AID targets patches of single-stranded DNA (ssDNA) from structured DNA, such as DNA bubbles, stem-loops or ssDNA overhangs from G-quadruplex DNA. Initially, determining the structure of AID had proved problematic due to its highly charged surface, extensive non-specific interactions and solubility issues. In 2015, we reported a structure of native AID using a combined computational-biochemical approach, whereby the catalytic pocket, as well as a dominant and minor DNA binding groove were described. In 2017, Qiao et al. reported the crystal structure of near-native AID that closely matched the initial structure whereby they identified the same dominant groove, as well as a supportive DNA binding groove termed the assistant patch. Furthermore, an RNA binding groove, responsible for binding RNA G-quadruplexes, as well as RNA-binding proteins specific to CSR were also identified in AID. Here, we review these four DNA/RNA binding grooves in AID and explore their possible binding combinations. We explore how these plastic modes of binding DNA and RNA simultaneously may impact AID function. Finally, we examine proposed AID dimer models bound to DNA/RNA and examine how this may act as an additional regulatory bottleneck of enzymatic activity.

Solving the functional and native structure of AID

Activation-induced cytidine deaminase (AID) was discovered over 20 years ago as the essential enzyme necessary for mediating somatic hypermutation and class switch recombination of antibodies (37). In addition to its immune functions, AID and related APOBEC3 enzymes also mediate genomic mutations and lesions that cause and exacerbate tumors (234).

In the first decade after its discovery, major insights into AID targeting of genomic DNA emerged. Though AID prefers to target Immunoglobulin (Ig) loci with a modest preference (352), AID targeting is loosely regulated such that it initiates genome-wide mutations, double stand breaks (DSB) and chromosomal translocation, and these lesions have been conclusively linked to tumorigenesis in lymphoid tissues, and also proposed to do so in non-lymphoid tissues. *In vivo*, active transcription is a pre-requisite for a gene to be targeted by AID, and many highly targeted genes, such as Ig loci, have unique transcriptional properties including strong enhancers, bi-directional transcription and a high potential for R-loop formation. Biochemical studies of purified AID revealed that AID is a single-stranded DNA (ssDNA)-specific deaminase. While AID can deaminate any form of ssDNA, it favors structured ssDNA patches (ie: small 5-11 nt-long bubbles), over fully linear ssDNA without a defined secondary structure (52). Indeed, such short ssDNA patches have been shown to be enriched in AID-targeted genes. *In vitro* assays using purified AID also showed a dependence of its mutational activity on active transcription and transcription stall, and thus emerged the prevailing model that AID primarily mutates ssDNA targets *in vivo*, generated as a result of the transcription process.

In parallel, biochemical analyses of AID's enzymatic behavior revealed that AID is an unusual enzyme in that it exhibits a lethargic catalytic rate of one deamination every few minutes and

extremely high ssDNA binding affinities, in the nM range (52), such that once bound to a ssDNA target, the AID:ssDNA complex was stable for several minutes, a much longer on/off time than a typical enzyme whose on/off kinetics are in the order of seconds. These parameters, which are orders of magnitude outside usual human enzyme catalytic and substrate binding kinetics (60), are thought to represent an inherent bottleneck to AID's genome-wide activities. Namely, the high ssDNA binding affinity and low catalytic rate were proposed to mitigate AID rampantly "jumping" around the genome to carry out promiscuous genome-wide mutations. However, it also remained equally possible that AID's high ssDNA binding affinity is a required property in order to allow AID to target of transient and short-lived patches of transcription-induced ssDNA.

Over the last two decades since its discovery, studies of AID targeting of Ig loci, oncogenes and transgenes have continued to provide a wealth of information on the requirement for transcriptional and ssDNA features from a substrate point of view, however, the enzyme aspect of the equation had remained a black box until very recently. The next frontier in the field after biochemical characterization of AID was to elucidate its structure, such that insights could be gained into how AID recognizes its ssDNA targets and whether any features could explain differential targeting to various loci. However, this proved a monumental task for the many structural biology labs who attempted to tackle this issue. The insurmountable issue has been that AID is a small, highly positively charged protein (AID net charge= +14), and hence prone to extensive non-specific protein/protein and protein/nucleic acid interactions and the formation of heterogenous oligomers. Indeed, the same issue has hampered X-ray and NMR studies of AID's closely related APOBEC enzymes, such that 25 out of 27 AID/APOBEC X-ray/NMR structures published to date are of non-wildtype variants that have been mutated and/or truncated in order to facilitate crystallization or solubilization (164). As an alternative strategy for obtaining AID's

functional and native structure, in 2015 we reported a functional structure of AID using a combined computational-biochemical approach (Figure 27A, left and B; (96). This work provided two novel insights on the binding of ssDNA by AID: first, we observed an abundance of catalytically-restricted conformations of AID and sporadic ssDNA binding on the surface in positions that do not pass over an open pocket; together, explaining AID's low catalytic rate (164). Second, we observed two putative ssDNA binding grooves, the dominant "groove 1" and the less-dominant "groove 2" (figure 27C).

Comparison of the catalytic motifs of two AID structures obtained by different approaches

Since the publication of AID's structure obtained through the computational-biochemical method, two crystal X-ray structures have also been published. The first was an AID-A3A hybrid structure, termed AIDv (245) (Figure 27A, middle), with some catalytic portions from AID but a major portion of the enzyme's DNA binding surface belonging to A3A. Though AIDv confirmed the basic architecture of AID's catalytic pocket, the ssDNA binding groove located on the surface was unsurprisingly divergent from native AID (native AID net charge = +14, AIDv net charge= +4); thus, the insights provided remained limited. Recently, Qiao et al. resolved the crystal structure of a near-native AID (Figure 27A, right) complexed with cytosine, deoxycytidine (dC) or deoxycytidine monophosphate (dCMP) in the catalytic pocket (54). Comparing the functional structure of native AID obtained through a computational-biochemical method with that of the subsequent crystal structure, there is remarkable agreement in the global architecture of the enzyme (Root-mean-square-deviation between conformations of ~1 Å) and the catalytic pockets (Figure 27B), including the core catalytic residues that are responsible for Zn coordination and

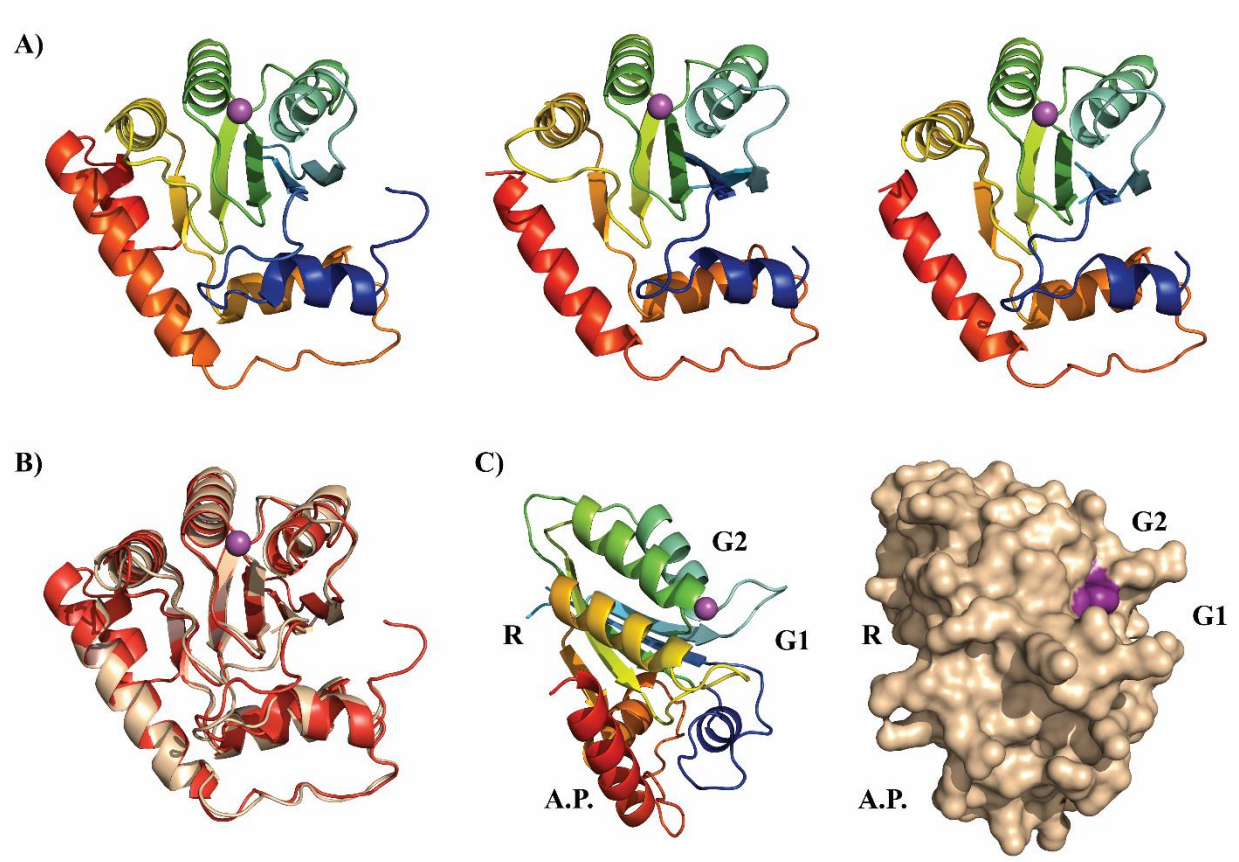


Figure 27: AID has several complementary DNA binding grooves. A. A representative conformation from the previously reported breathing model of AID structure obtained through a computational-biochemical method (left), crystal structure elucidated by Qiao et al. (middle) and the crystal structure of the AID chimera AIDv (right). The purple sphere depicts the coordinated Zn in the catalytic pocket. B. Superimposition of the Qiao et al. AID crystal structure (beige color) with a representative conformation from the breathing model of AID (red). The two structures are near-identical in secondary structural regions and display highly similar loops with an overall root-mean-square deviation (RMSD) of ~1.0 Å. C. Ribbon diagram (left) and surface structure (right) of the Qiao crystal structure, depicting the general location of the four grooves in AID (G1, G2, A.P. and R are groove 1, groove 2, assistant patch and the RNA binding groove, respectively). In the surface structure, the purple surface depicts the catalytic residues.

deamination reaction (H56, E58, C87, C90) as well as the broader set of residues that compose the architecture of the pocket and support the stability of the dC fit in the pocket, termed the "secondary catalytic residues" (96). In addition, there was also close agreement on the specific orientations of the dC within the catalytic pocket between the two structures. The computational-

biochemical structure of AID also revealed a remarkable built-in mode of activity regulation by dynamic closure and opening of the catalytic pocket. It was found that the catalytic pocket of AID, by virtue of the inherent flexibility built into the secondary catalytic loops surrounding it, could assume both closed and open confirmations, with the former prevailing by a ratio of approximately 2:1 in all simulations. This was then functionally verified by constructing AID variants with altered pocket breathing dynamics and demonstrating a direct functional correlation with enzyme activity. Although the AID crystal structure lacked this type of insight due to co-crystallization with the substrate in the pocket, the state of pocket closure, termed "Schrödinger's CATalytic pocket" (164), was later independently confirmed by NMR and crystal structures of AID's closely related APOBEC3 siblings (208,246,247,257).

Plasticity of three surface motifs responsible for ssDNA binding by AID

The first critical event that is required for AID to mutate a given ssDNA substrate is the capture and binding of ssDNA such that a dC residue binds in the catalytic pocket. Indeed, the computational-biochemical and X-ray approaches both used a similar computational docking strategy to probe ssDNA binding to the AID structure, with the difference that the enzyme structure used was either obtained using a computational-biochemical method, or crystallography. In our computational-biochemical approach, we described a prominent ssDNA binding groove, termed "groove 1", passing over the catalytic pocket that was lined with positively charged residues (R19, K22, R24, R25, R50 and K52; (Figure 27). Additionally, we observed ssDNA binding in a second surface groove, termed "groove 2" that intersects with groove 1 and traverses the catalytic pocket (Figure 27). In contrast to the high density of positively charged residues in groove 1, residues unique to groove 2 contained a mixture of positive and negative residues lining α2 and α3 (R63, D89, R92, D96 and R99; Figure 27).

In the AID crystal structure report, Qiao et al. similarly examined ssDNA binding to AID through computational modeling since AID:ssDNA co-crystals could not be obtained (54). This study also identified a main substrate binding channel, termed the "main channel", which is like the proposed dominant ssDNA binding groove 1 described in the computational-biochemical structure of AID. This study did not identify groove 2 as a ssDNA binding site, proposing instead, an additional DNA binding patch termed "the assistant patch" (Figure 27). We docked ssDNA with AID using the same methodology as employed in our computational-biochemical study of AID's structure but instead using the crystal structure obtained by Qiao et al. and analyzed predicted AID-ssDNA complexes using each binding groove (Figure 28). Unlike the main substrate channel (groove 1) and groove 2, which both facilitate catalytically-accessible ssDNA binding (Figure 28; (96), the newly identified assistant patch is housed on the C-terminal α6 (residues involved are R171, R174, R177 and R178), distal from the catalytic pocket (Figure 28; (54). Using a mutagenesis-based approach to complement the modeling, it was found that mutations of the assistant patch motif affected AID activity on branched ssDNA structures or DNA G-quadruplexes, but not on simple unbranched ssDNA. Thus, the two studies that have tackled the issue of how and where AID binds its ssDNA substrate have come to the remarkably similar overall conclusion that AID uses two apparently independent but complementary ssDNA capture grooves on its surface for binding its targets. The consistent theme is that AID binds ssDNA through multiple channels on the surface. This usage of multiple independent motifs on the surface of AID able to bind ssDNA correlates well with its observed biochemical properties, namely high nM-range affinity for ssDNA and preference for structured and branched substrates. Given that

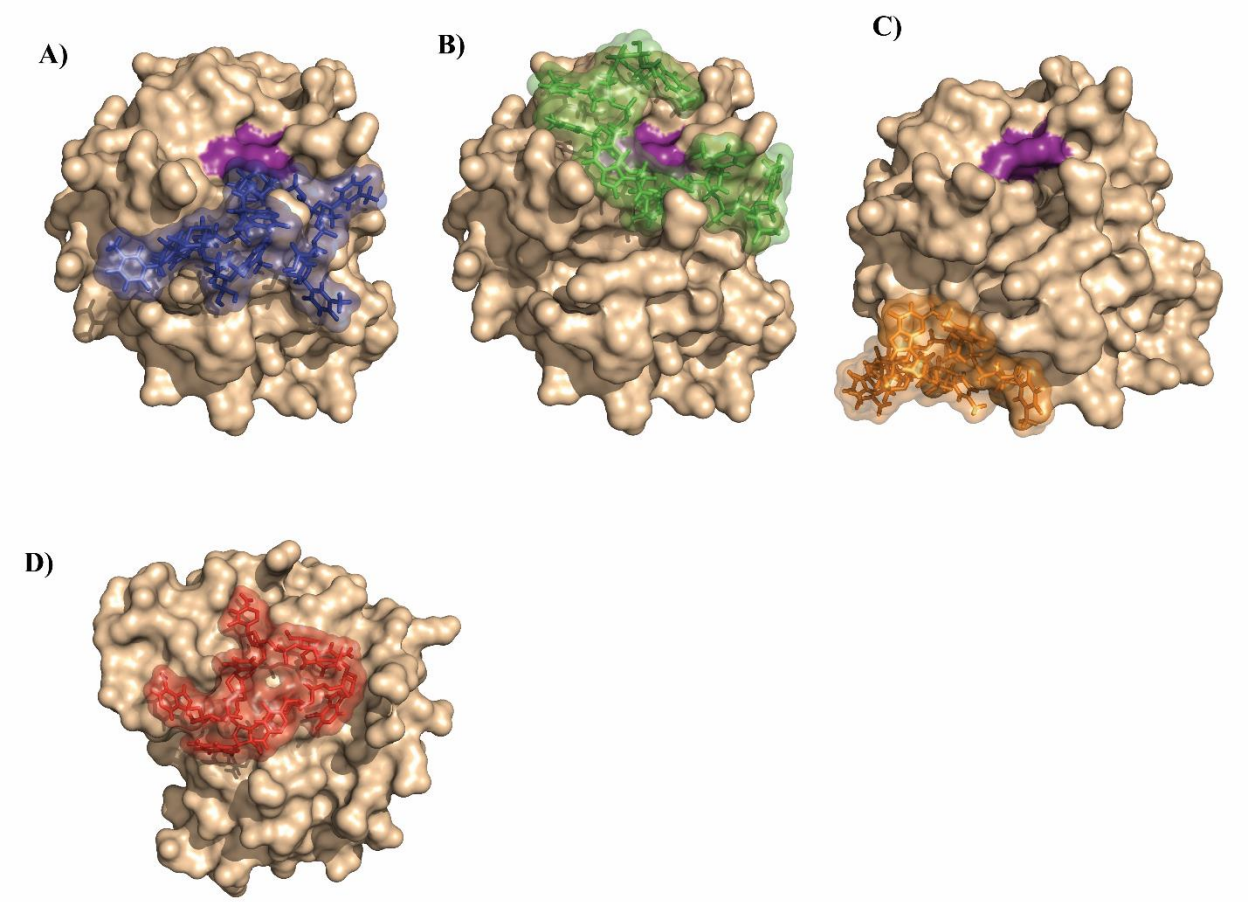


Figure 28: AID has three ssDNA binding grooves and one ssRNA binding groove. A. AID bound with ssDNA in the groove 1 (blue) with dC stabilized in the catalytic pocket. B. AID bound with ssDNA in groove 2 (green) with dC stabilized in the catalytic pocket. C. AID with ssDNA bound in the assistant patch (orange). Groove 1 and 2 bound complexes are catalytically-productive, whilst binding to the assistant patch does not directly result in deamination. D. ssRNA bound in the RNA binding groove (red), located on the AID surface opposite from groove 1, groove 2, the assistant patch and the catalytic pocket. AID surface is colored beige, with the catalytic pocket colored purple.

structured ssDNA motifs are abundant at the Ig loci and oncogenes targeted by AID, the recognition of this structural mode of ssDNA binding is a substantial step forward in understanding the role of AID's structure in genomic targeting.

On the other hand, the two studies also reached some distinct conclusions regarding ssDNA capture by AID. First, though both studies identified the same dominant primary ssDNA-binding

groove (groove 1 or the main substrate channel), they identified different secondary ssDNA binding regions (groove 2 vs. the assistant patch). Second, the computational-biochemical study determined that the majority of AID:ssDNA interactions occur altogether outside either ssDNA binding motif such that the vast majority of AID:ssDNA interactions do not result in a deamination reaction. In this model only a minor fraction of ssDNA is predicted to bind into grooves 1 or 2 and pass over the catalytic pocket, whilst the highly positively charged surface acts as a non-specific "trap", capturing ssDNA sporadically in positions that do not pass over the pocket. It was concluded that non-specific and sporadic surface ssDNA binding acts as a negative structural regulator of AID's mutagenic activities, and combined with the further bottlenecks that the catalytic pocket can accommodate non-dC bases (including purines), and that the pocket itself is closed the majority of the time, these structural properties explain the remarkably lethargic catalytic velocity of AID. When all these bottlenecks, including sporadic and catalytically non-productive ssDNA binding by AID's surface were accounted for, the resulting calculations of deamination efficiency were consistent with AID's known low catalytic rate, as well as an independent study that modeled ssDNA catalysis by AID (230). In contrast, the AID crystal structure study did not probe non-specific surface binding, suggesting instead that AID efficiently captures branched ssDNA and G-quadruplex DNA using the two surface grooves under crystallization conditions; however, the inability to co-crystalize an AID:ssDNA in this study is nevertheless consistent with sporadic and disorganized surface-wide ssDNA binding by AID (54).

AID binds RNA in an RNA binding motif distal from the catalytic pocket

In addition to the ssDNA binding modes on the surface of AID, AID has also been shown to bind RNA in a putative RNA binding groove. In 2015, Zheng et al. demonstrated AID binds G-quadruplex RNA from intronic switch-region transcripts (353). Using AID-containing extracts of

CH12 cells stimulated for CSR, AID was shown to bind G-quadruplex structures formed in sense switch RNA transcripts *in vitro* and *in vivo* using an RNA pull-down assay. In contrast, AID was unable to bind RNA transcripts lacking G-quadruplexes. Based on homology with RNA helicase associated with AU-rich element (RHAU), a G-quadruplex RNA binding protein, amino acids 130-138 in AID were identified as the putative RNA binding domain (Figure 27; (353–355). G133V, an RNA-binding domain mutation identified in HIGM patients (353), was unable to bind sense switch RNA transcripts. In contrast to wild-type AID, G133V was unable to bind S-region DNA using the ChIP assay and was unable to induce CSR in AID-deficient splenic B cells. Zheng et al. concluded that G-quadruplexes present in switch-region RNA transcripts bind the RNA binding domain of AID and guide it to switch-region DNA to facilitate CSR (353). Expanding upon these findings, in 2016 Mondal et al. showed that in contrast to wild type AID, G133V and G133P mutants were unable to associate with CSR-specific RNA-binding proteins (RBPs) and RNA critical for CSR (56). Like the assistant patch, the RNA binding groove does not pass near the catalytic pocket but instead forms a surface on the opposite face of AID (Figure 28C). In 2017, we simulated AID-RNA binding and found most complexes bound RNA to the putative RNA binding groove (89). In the RNA binding groove, 11 surface residues including G133 were identified with additional contribution from residues residing in α7 in certain conformations.

The diversity of nucleic acid binding modes of AID enables combinatorial binding

With regards to ssDNA capture on the surface of AID and implications for genome-wide AID targeting, several questions remain to be explored. First, ssDNA bubbles, DNA/RNA hybrids, stem-loops and G4 quadruplex structures have all been shown to be targeted by AID. Given the existence of several highly positively charged regions on its surface (Figure 28), it is intriguing to

speculate that AID has evolved different surface grooves/motifs to use in various combinations and in a plastic manner, not only for binding different ssDNA structures as aforementioned, but also for binding ssDNA in combination with RNA as suggested by the previous studies mentioned in the last section (Figure 29).

For instance, a model that is consistent with the collective observations of ssDNA binding by AID in both the computational-biochemical and crystallography studies is that when AID encounters simple ssDNA, deamination catalysis is inefficient due to non-specific and catalytically non-productive surface-wide ssDNA binding. In contrast, the usage of secondary motifs (groove 2 or the assistant patch) in combination with the primary groove 1 for forked or branched ssDNA substrates ensures that the binding of these is much more likely to be catalytically productive. Thus, groove 1 acts as the dominant ssDNA binding groove in AID able to accommodate linear ssDNA, as well as branched ssDNA in parallel with the assistant patch (Figure 29A. left). Groove 2 facilitates ssDNA deamination to a lesser extent than groove 1 and could also act in concert with the assistant patch (Figure 29A, right). This dominance of groove 1 over groove 2 has been inferred from ssDNA docking experiments and enzymatic assays of AID mutants targeting groove 1 or 2 (96). However, it is crucial to note that groove 1 dominance has been illustrated through the usage of dsDNA with small stretches of ssDNA bubble region (5-7 nt). Groove 2 represents a catalytically accessible DNA binding groove, located further from the assistant patch, but closer to the RNA binding groove. Thus, it is even possible that groove 2 acts in concert with the assistant patch to bind larger ssDNA structures (ie: larger ssDNA bubbles or stem-loop structures), for which the elongated groove 2 would be more efficient than groove 1 because it is located further from the assistant patch. Taken together, it is tempting to wonder how much of AID's surface is indeed a non-specific trap vs. how much of the non-catalytically viable DNA binding motifs can

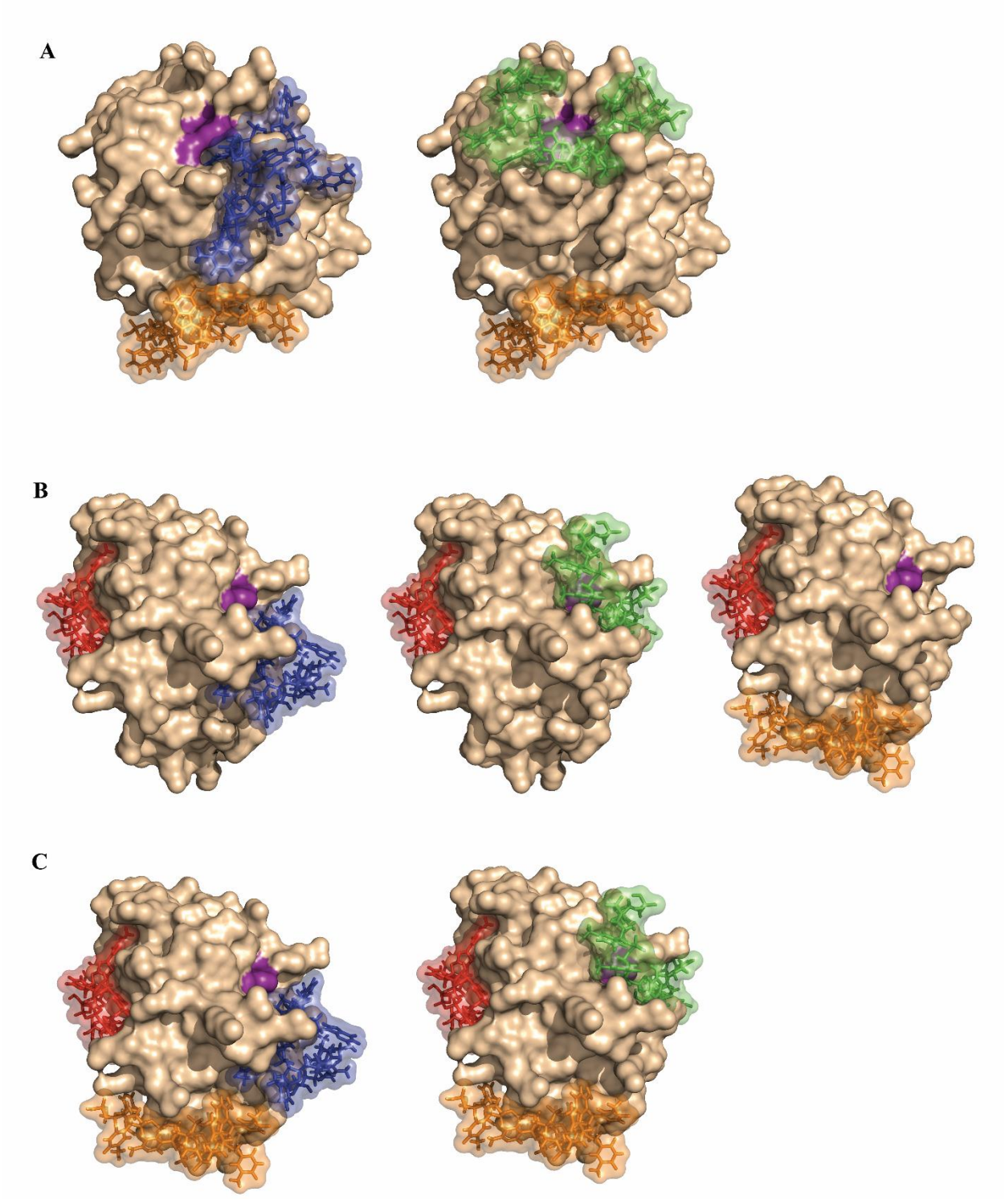


Figure 29: Combinatorial binding modes of ssDNA and ssRNA in AID. A. The most likely binding modes for branched DNA-AID with ssDNA bound to groove 1 and the assistant patch (left) or ssDNA bound to groove 2 and the assistant patch (Right). B. Binding modes in which ssRNA is bound to the RNA binding groove and ssDNA is bound in a catalytically accessible

conformation in groove 1 (left), groove 2 (middle) or in a non-productive conformation in the assistant patch (right). C. AID bound to ssRNA in the RNA binding groove, ssDNA bound in the assistant patch and ssDNA bound in either groove 1 (left) or groove 2 (right). ssDNA bound to groove 1, 2 and the assistant patch were colored blue, green and orange, respectively. ssRNA bound to the RNA binding groove was colored red, while the catalytic pocket was colored purple.

enhance catalytically accessible binding. It can be hypothesized that differences in the ability of various ssDNA structures in the relative utilization of these two modes provides a plastic mechanism through which the structure of AID underlies substrate preference.

The studies noting the role of RNA in targeting AID for class switch recombination have proposed several models of AID-DNA-RNA interactions. One model involves an RNA-to-DNA hand-off wherein G4 quadruplex RNA binds to AID and then exchanges binding for G4 quadruplex DNA (356). It has been suggested that this process could happen either in trans where AID binds free G4 quadruplex RNA first which then guides the AID-RNA complex to the switch region G4 quadruplex DNA through strand complementarity, or in cis where AID interacts with the G4 quadruplex RNA in the context of the RNA/DNA hybrid. By the same token, the discovery of multiple DNA and RNA motifs on the surface of AID strongly suggests the possibility of simultaneous DNA and RNA binding on the enzyme's surface.

In support of this notion, when structural and biochemical aspects of simultaneous DNA and RNA binding of AID were probed, it was found that the binding affinity and catalytic activity of AID on DNA/DNA and DNA/RNA bubble structures derived from random sequences vs Ig switch regions were compared (89). In randomized structures, AID was found to have higher catalytic activity and binding affinity for DNA/DNA structures instead of DNA/RNA structures. However, when binding GC-rich S-region sequences, AID preferred binding DNA/RNA hybrid structures

over conventional substrates composed entirely of DNA by 75-fold, whilst deamination was also improved several fold. AID-RNA binding was simulated, and the putative RNA-binding groove was also identified in this study as the main RNA-binding groove (89). In conjunction with DNA binding to groove 1, groove 2 and the assistant patch, several possibilities of simultaneous DNA and RNA binding may occur: Groove 1 + RNA binding groove, Groove 2 + RNA binding groove and Assistant patch + RNA binding groove co-binding by AID (Figure 29B). In the context of hybrid G4-DNA/RNA structures (where more than 2 DNA/RNA strands are available), additional concerted binding events may occur: Groove 1 + Assistant patch + RNA binding groove and Groove 2 + Assistant patch +RNA binding groove (Figure 29C).

Interestingly, the RNA binding groove of AID contained several residues that are mutated in HIGM patients (eg. H130P, G133V, I136K and M139T) (357). We and others determined that these mutations abolish activity on ssDNA substrates yet retained activity on DNA/RNA hybrids (89,353). This suggested that RNA binding to the RNA groove stabilizes DNA binding distally in the active site and that the RNA binding groove utilizes a redundant set of surface residues making it resilient towards point mutations. HIGM mutations (F15L and R24W) and other mutations in groove 1 (R24D, R25A, R25D, R25del and E26A) that disrupted/abrogated catalysis on ssDNA substrates were found to retain activity on DNA/RNA hybrids (89,96). Interestingly, mutation of groove 2 residues (D89A, D89R and D96R) resulted in lowered activity on ssDNA and DNA/RNA hybrids. Cumulatively, this suggested that when engaging ssDNA with a defective groove 1, RNA bound to the RNA binding groove can rescue the stabilization of ssDNA near the catalytic pocket. On the other hand, mutations in groove 2 hinder the ability of AID to recognize DNA/RNA hybrids suggesting that groove 2 is the dominant binding groove for binding DNA/RNA substrates, which

may be mediated by its elongated and wider architecture compared to groove 1, combined with its closer proximity to the RNA binding groove.

AID oligomerization as a regulatory mechanism for targeting branched ssDNA structures

Qiao et al. demonstrated that AID binding to G4 quadruplex structures derived from Ig switch regions induced cooperative AID oligomerization (54), in turn raising the question whether AID oligomerization can influence structured ssDNA or RNA binding preference. Using AID dimers simulated based on resolved APOBEC2 dimers which have been suggested as a good model for AID dimerization (170), we simulated AID dimer binding with ssDNA and RNA (Figure 30A). These AID dimers maintained all the same above-discussed nucleic acid binding grooves as monomeric AID, except for the emergence of a novel and extensive groove near the catalytic pocket formed by the quaternary structure at the dimeric interface. Akin to monomeric AID, we found ssDNA predominantly favored binding to groove 1 vs groove 2 (41% and 7%, respectively) in dimeric AID (Figure 30B). Unlike monomeric AID, we found a fraction of binding events bound to the dimeric interface (24%). In contrast to ssRNA binding by monomeric AID we found that only a minor fraction of RNA binding events bound to the RNA binding groove (18%), whilst the majority bound to the dimeric interface (77%; Figure 30C, right). Thus, binding simulations highlight that the dimeric interface of AID could play a significant role in ssDNA and ssRNA binding (Figure 30D).

It is intriguing to speculate whether this enhanced ssDNA and ssRNA binding at the dimeric interface explains the longer complex half-life of G4 quadruplex structures by dimeric vs monomeric AID (54). Dimeric AID was shown to bind four ssDNA overhangs from a G4

quadruplex DNA structure. In an alternative model, once AID is bound to one or more nucleic acids, monomeric AID has additional sequences overhanging that might act in concert with the

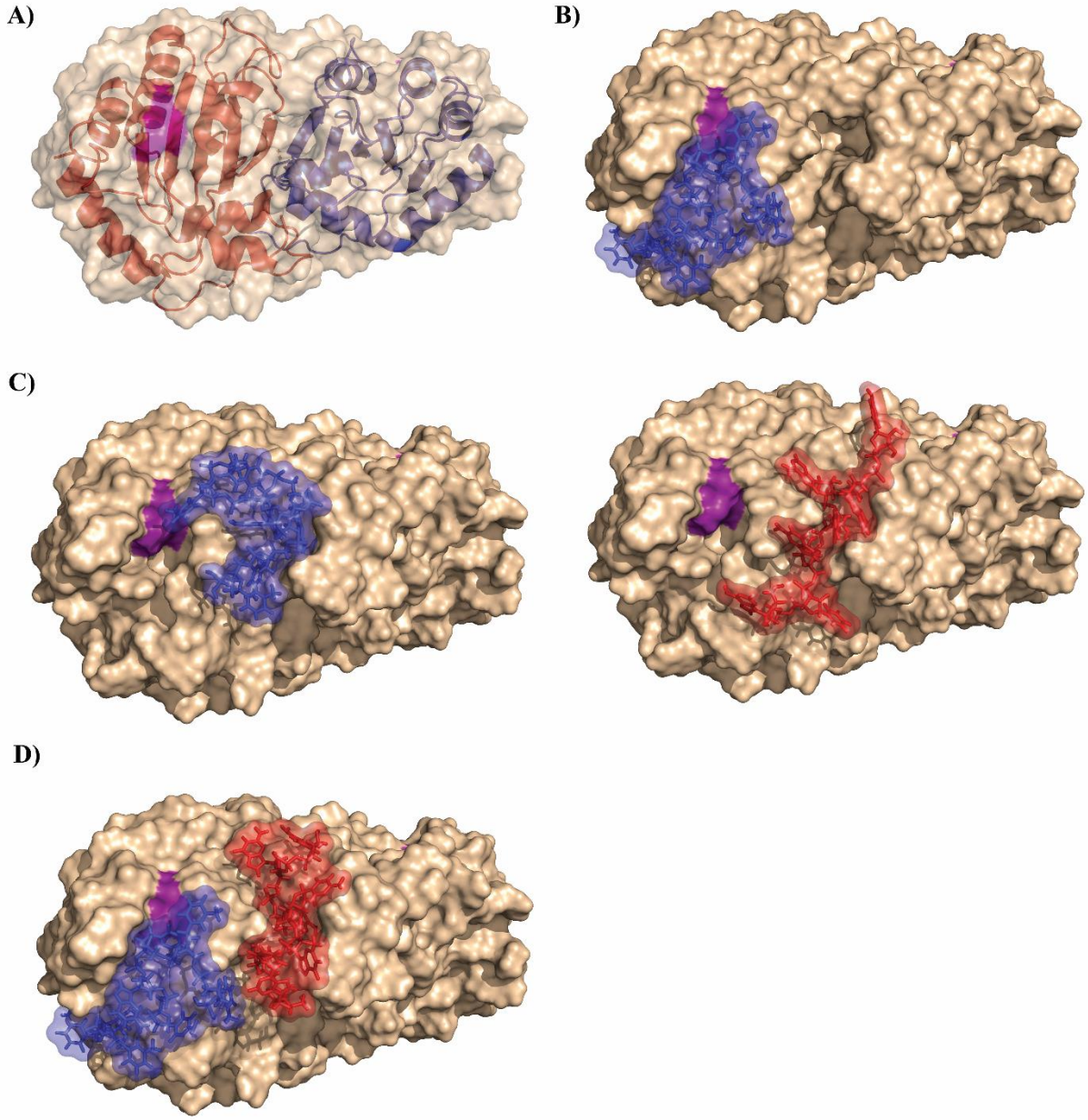


Figure 30: Dimeric AID preferentially binds ssDNA to groove 1 and ssRNA to the dimeric interface. A. Dimeric structure of AID based upon the A2-style of AID dimerization. The AID structure is transparent to show the underlying monomer chains underneath (red and blue). The catalytic pocket is colored purple. B. The most prevalent binding mode of ssDNA in dimeric AID within groove 1 C. ssDNA bound in the dimeric interface of AID (left). Unlike monomeric AID which preferentially binds ssRNA in the RNA binding groove, dimeric AID stabilizes ssRNA

primarily in the dimeric interface (right). D. Dimeric AID bound with ssDNA in groove 1 and ssRNA in the dimeric interface.

dimerization interface to stabilize and bridge an additional AID unit for the dimerization on the opposite surface. In either case, in the context of a G4 quadruplex DNA/RNA hybrid structure, it is likely that an RNA overhang binds in the dimeric interface, while each unit binds a ssDNA overhang in groove 1 for deamination (Figure 30D), with the fourth remaining overhang bound to the assistant patch or RNA binding groove (ssDNA or RNA, respectively). In the case of higher-order oligomers, additional surface groove combinations may be possible, with oligomerization interfaces stabilized by RNA or ssDNA and groove 1 acting as the main catalytically accessible ssDNA binding groove. Thus, in the context of branched ssDNA, AID could facilitate deamination of one target strand bound in groove 1/groove 2 and the assistant patch. However, AID may require oligomerization to efficiently bind and deaminate hybrid DNA/RNA branched or G4 quadruplex structures. Collectively, the biochemical data and binding simulations are highly suggestive that AID oligomerization may act as an additional regulatory mechanism for targeting more complex branched ssDNA or DNA/RNA structures.

Thesis Conclusion and impact on field

Over the last two decades since its discovery, studies of AID targeting of Ig loci, methylated cytidines within CpG islands and oncogenes have continued to provide a wealth of information on the requirement for transcriptional and ssDNA features from a substrate point of view, however, the enzyme aspect of the equation had remained a black box until very recently. In 2013, we provided an glimpse of the AID catalytic pocket in human and bony fish and how it was predicted to stabilize dC *vs* 5-mC (95). In corroboration with other groups at the time, we had confirmed that

AID from human was catalytically inefficient on 5-mC. In contrast to Hs-AID, we demonstrated evolutionarily distant Dr-AID deaminated 5-mC very efficiently. This unique property was attributed to the flexible loops regions surrounding the catalytic pocket, later termed as the secondary catalytic loops (96).

After biochemical characterization of AID, the next frontier in the field was to elucidate its structure. Questions remained regarding how AID recognizes its ssDNA targets and whether any structural features could explain differential targeting to various loci. However, this proved a monumental task for the many structural biology labs who attempted to tackle this issue. The insurmountable issue has been that AID is a small, highly positively charged protein (AID net charge= +14), and hence prone to extensive non-specific protein/protein and protein/nucleic acid interactions and the formation of heterogenous oligomers (96,164). Indeed, the same issues has hampered X-ray and NMR studies of AID's closely related APOBEC enzymes, such that the majority of AID/APOBEC X-ray/NMR structures published to date are of non-wildtype variants that have been mutated and/or truncated in order to facilitate crystallization or solubilization (164). As an alternative strategy for obtaining AID's functional and native structure, in 2015 we reported a functional structure of AID using a combined computational-biochemical approach (96). In addition to working with human AID and AID from other species, we also constructed a library of several hundred AID point mutants. This work provided two novel insights on the binding of ssDNA by AID: first, we observed an abundance of catalytically-restricted conformations of AID and sporadic ssDNA binding on the surface in positions that do not pass over an open pocket; together, explaining AID's low catalytic rate (164). This unique feature of the enzyme has since been corroborated in AID's APOBEC sibling enzymes and is now thought to be a regulatory safeguard in mutagenic APOBECs. Second, we observed two putative ssDNA binding grooves,

the dominant "groove 1" and the less-dominant "groove 2" (96,164). This finding was later supported through structural elucidation of a truncated AID X-ray crystal structure (54).

Thus, equipped with a breathing structure of AID, fine-detailed map of the catalytic pocket and AID-DNA binding modes in conjunction with a large library of AID mutants, the next logical step was to develop AID inhibitors. Due to its role in driving and exacerbating AID-expressing cancers, inhibition of AID activity has been suggested to be of clinical benefit (159,320,331–340). As an extension of our combined computational-biochemical approach to delineating the breathing and dynamic structure of AID in 2015 (96), we screened a library of small molecule drug-like compounds against its catalytic pocket. To our knowledge, this work identified the first known small-molecule inhibitors that specifically inhibit the mutagenic activity of purified AID and AID in the context of complete cellular extracts. In addition, inhibitors were also found to weakly inhibit the tumorigenic activity of APOBEC3A and 3B, thus establishing the foundation for the future development of inhibitors targeting these AID sibling enzymes. Aside their potential for therapeutic development, these inhibitors serve as additional tools to probe AID-DNA interactions further.

In conclusion, we have explored the structure-function relationship of AID. We have elucidated a comprehensive and detailed map of the AID surface topology and core architecture. Developed beyond a single conformation, we have provided a breathing and dynamic structure characterized by a series of conformations over time, thus offering the 4D structure of AID which encompassed the dimension of time as well as structure. Subsequently, the accuracy of our approach and AID structure was confirmed upon publication of the AID crystal structure two years later. Furthermore, this breathing structure of the AID catalytic pocket was also discovered in the catalytic pocket of A3A and A3B, hinting that this may be a widespread mechanism

conserved amongst APOBEC enzymes to reduce genotoxicity. Finally, we have established the physiological relevance of this 4D structure in the context of AID's role in immunity, genome demethylation and cancer, which in a sense, considering time (both real time in the sense of conformational dynamic breathing, and evolutionary time) as the fourth and function as the fifth dimension, one can regard as the 5D structure of AID. Stemming from this breathing structure of AID, we have also generated first-generation small molecule inhibitors that block the mutagenic activity of AID *in vitro*. It is our hope that our efforts in producing this first ever 5D structure of a human cancer-causing molecule and showing its utility for inhibitor development will serve as a road map for future efforts in cancer, immunology and structural biology along the same path.

References

1. Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993 Apr 22;362(6422):709–15.

2. Aguilera A, García-Muse T. Causes of genome instability. Annu Rev Genet. 2013;47:1–32.

3. Eckardt-Schupp F, Klaus C. Radiation inducible DNA repair processes in eukaryotes. Biochimie. 1999 Feb;81(1–2):161–71.

4. Santivasi WL, Xia F. Ionizing radiation-induced DNA damage, response, and repair. Antioxid Redox Signal. 2014 Jul 10;21(2):251–9.

5. Somers CM, Yauk CL, White PA, Parfett CLJ, Quinn JS. Air pollution induces heritable DNA mutations. Proc Natl Acad Sci U S A. 2002 Dec 10;99(25):15904–7.

6. Yauk CL, Quinn JS. Multilocus DNA fingerprinting reveals high rate of heritable genetic mutation in herring gulls nesting in an industrialized urban site. Proc Natl Acad Sci U S A. 1996 Oct 29;93(22):12137–41.

7. Kim HS, Kim YJ, Seo YR. An Overview of Carcinogenic Heavy Metal: Molecular Toxicity Mechanism and Prevention. J Cancer Prev. 2015 Dec;20(4):232–40.

8. Ratnaike RN. Acute and chronic arsenic toxicity. Postgrad Med J. 2003 Jul;79(933):391–6.

9. Zhang H, Wei K, Zhang M, Liu R, Chen Y. Assessing the mechanism of DNA damage induced by lead through direct and indirect interactions. J Photochem Photobiol B. 2014 Jul 5;136:46–53.

10. Rocha CRR, Silva MM, Quinet A, Cabral-Neto JB, Menck CFM. DNA repair pathways and cisplatin resistance: an intimate relationship. Clin Sao Paulo Braz. 2018 06;73(suppl 1):e478s.

11. Woods D, Turchi JJ. Chemotherapy induced DNA damage response: convergence of drugs and pathways. Cancer Biol Ther. 2013 May;14(5):379–89.

12. Zeman MK, Cimprich KA. Causes and consequences of replication stress. Nat Cell Biol. 2014 Jan;16(1):2–9.

13. Bertram C, Hass R. Cellular responses to reactive oxygen species-induced DNA damage and aging. Biol Chem. 2008 Mar;389(3):211–20.

14. Cadet J, Wagner JR. DNA base damage by reactive oxygen species, oxidizing agents, and UV radiation. Cold Spring Harb Perspect Biol. 2013 Feb 1;5(2).

15. Lindahl T, Barnes DE. Repair of endogenous DNA damage. Cold Spring Harb Symp Quant Biol. 2000;65:127–33.

16. Iyama T, Wilson DM. DNA repair mechanisms in dividing and non-dividing cells. DNA Repair. 2013 Aug;12(8):620–36.

17. Jackson SP, Bartek J. The DNA-damage response in human biology and disease. Nature. 2009 Oct 22;461(7267):1071–8.

18. Caldecott KW. DNA single-strand break repair. Exp Cell Res. 2014 Nov 15;329(1):2–8.

19. Klement K, Goodarzi AA. DNA double strand break responses and chromatin alterations within the aging cell. Exp Cell Res. 2014 Nov 15;329(1):42–52.

20. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JHJ. Understanding nucleotide excision repair and its roles in cancer and ageing. Nat Rev Mol Cell Biol. 2014 Jul;15(7):465–81.

21. Parsons JL, Dianov GL. Co-ordination of base excision repair and genome stability. DNA Repair. 2013 May 1;12(5):326–33.

22. Thompson LH, West MG. XRCC1 keeps DNA from getting stranded. Mutat Res. 2000 Feb 16;459(1):1–18.

23. Bak ST, Sakellariou D, Pena-Diaz J. The dual nature of mismatch repair as antimutator and mutator: for better or for worse. Front Genet. 2014;5:287.

24. Jascur T, Boland CR. Structure and function of the components of the human DNA mismatch repair system. Int J Cancer. 2006 Nov 1;119(9):2030–5.

25. Bohgaki T, Bohgaki M, Hakem R. DNA double-strand break signaling and human disorders. Genome Integr. 2010 Nov 5;1(1):15.

26. Bouwman BAM, Crosetto N. Endogenous DNA Double-Strand Breaks during DNA Transactions: Emerging Insights and Methods for Genome-Wide Profiling. Genes. 2018 Dec 14;9(12).

27. Daley JM, Kwon Y, Niu H, Sung P. Investigations of homologous recombination pathways and their regulation. Yale J Biol Med. 2013 Dec 13;86(4):453–61.

28. Lieber MR, Gu J, Lu H, Shimazaki N, Tsai AG. Nonhomologous DNA end joining (NHEJ) and chromosomal translocations in humans. Subcell Biochem. 2010;50:279–96.

29. Brandsma I, Gent DC. Pathway choice in DNA double strand break repair: observations of a balancing act. Genome Integr. 2012 Nov 27;3(1):9.

30. Schatz DG, Swanson PC. V(D)J recombination: mechanisms of initiation. Annu Rev Genet. 2011;45:167–202.

31. Stavnezer J, Schrader CE. IgH chain class switch recombination: mechanism and regulation. J Immunol Baltim Md 1950. 2014 Dec 1;193(11):5370–8.

32. Maul RW, Gearhart PJ. AID and somatic hypermutation. Adv Immunol. 2010;105:159–91.

33. Kerr WG, Hendershot LM, Burrows PD. Regulation of IgM and IgD expression in human B-lineage cells. J Immunol Baltim Md 1950. 1991 May 15;146(10):3314–21.

34. Treanor B. B-cell receptor: from resting state to activate. Immunology. 2012 May;136(1):21–7.

35. Briney BS, Crowe JE. Secondary mechanisms of diversification in the human antibody repertoire. Front Immunol. 2013;4:42.

36. Yoshikawa K, Okazaki IM, Eto T, Kinoshita K, Muramatsu M, Nagaoka H, et al. AID enzyme-induced hypermutation in an actively transcribed gene in fibroblasts. Science. 2002;296(5575):2033-6.

37. Muramatsu M, Sankaranand VS, Anant S, Sugai M, Kinoshita K, Davidson NO, et al. Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. J Biol Chem. 1999 Jun 25;274(26):18470–6.

38. Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell. 2000 Sep 1;102(5):553–63.

39. Berek C, Berger A, Apel M. Maturation of the immune response in germinal centers. Cell. 1991 Dec 20;67(6):1121–9.

40. Nussenzweig MC, Alt FW. Antibody diversity: one enzyme to rule them all. Nat Med. 2004 Dec;10(12):1304–5.

41. Seifert M, Küppers R. Human memory B cells. Leukemia. 2016;30(12):2283–92.

42. Delker RK, Fugmann SD, Papavasiliou FN. A coming-of-age story: activation-induced cytidine deaminase turns 10. Nat Immunol. 2009 Nov;10(11):1147–53.

43. Schroeder HW, Cavacini L. Structure and function of immunoglobulins. J Allergy Clin Immunol. 2010 Feb;125(2 Suppl 2):S41-52.

44. de Yebenes VG, Ramiro AR. Activation-induced deaminase: light and dark sides. Trends Mol Med. 2006 Sep;12(9):432–9.

45. Nagaoka H, Muramatsu M, Yamamura N, Kinoshita K, Honjo T. Activation-induced deaminase (AID)-directed hypermutation in the immunoglobulin Smu region: implication of AID involvement in a common step of class switch recombination and somatic hypermutation. J Exp Med. 2002 Feb 18;195(4):529–34.

46. Minegishi Y, Lavoie A, Cunningham-Rundles C, Bedard PM, Hebert J, Cote L, et al. Mutations in activation-induced cytidine deaminase in patients with hyper IgM syndrome. Clin Immunol. 2000;97(3):203-10.

47. Revy P, Muto T, Levy Y, Geissmann F, Plebani A, Sanal O, et al. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). Cell. 2000 Sep 1;102(5):565–75.

48. Sohail A, Klapacz J, Samaranayake M, Ullah A, Bhagwat AS. Human activation-induced cytidine deaminase causes transcription-dependent, strand-biased C to U deaminations. Nucleic Acids Res. 2003 Jun 15;31(12):2990–4.

49. Larijani M, Martin A. Single-stranded DNA structure and positional context of the target cytidine determine the enzymatic efficiency of AID. Mol Cell Biol. 2007 Dec;27(23):8038–48.

50. Pham P, Bransteitter R, Petruska J, Goodman MF. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature. 2003 Jul 3;424(6944):103–7.

51. Bransteitter R, Pham P, Scharff MD, Goodman MF. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proc Natl Acad Sci U S A. 2003 Apr 1;100(7):4102–7.

52. Larijani M, Petrov AP, Kolenchenko O, Berru M, Krylov SN, Martin A. AID associates with single-stranded DNA with high affinity and a long complex half-life in a sequence-independent manner. Mol Cell Biol. 2007 Jan;27(1):20–30.

53. Dickerson SK, Market E, Besmer E, Papavasiliou FN. AID Mediates Hypermutation by Deaminating Single Stranded DNA. J Exp Med. 2003;197(10):1291-6.

54. Qiao Q, Wang L, Meng FL, Hwang JK, Alt FW, Wu H. AID Recognizes Structured DNA for Class Switch Recombination. Mol Cell. 2017 Aug 3;67(3):361-373 e4.

55. Larijani M, Martin A. The biochemistry of activation-induced deaminase and its physiological functions. Semin Immunol. 2012 Aug;24(4):255–63.

56. Mondal S, Begum NA, Hu W, Honjo T. Functional requirements of AID's higher order structures and their interaction with RNA-binding proteins. Proc Natl Acad Sci U S A. 2016 Mar 15;113(11):E1545-54.

57. Brar SS, Sacho EJ, Tessmer I, Croteau DL, Erie DA, Diaz M. Activation-induced deaminase, AID, is catalytically active as a monomer on single-stranded DNA. DNA Repair Amst. 2008 Jan 1;7(1):77–87.

58. Kohli RM, Abrams SR, Gajula KS, Maul RW, Gearhart PJ, Stivers JT. A portable hot spot recognition loop transfers sequence preferences from APOBEC family members to activation-induced cytidine deaminase. J Biol Chem. 2009 Aug 21;284(34):22898–904.

59. Carpenter MA, Rajagurubandara E, Wijesinghe P, Bhagwat AS. Determinants of sequence-specificity within human AID and APOBEC3G. DNA Repair Amst. 2010 May 4;9(5):579–87.

60. Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, et al. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry. 2011 May 31;50(21):4402–10.

61. Pham P, Calabrese P, Park SJ, Goodman MF. Analysis of a Single-stranded DNA-scanning Process in Which Activation-induced Deoxycytidine Deaminase (AID) Deaminates C to U Haphazardly and Inefficiently to Ensure Mutational Diversity. J Biol Chem. 2011 Jul 15;286(28):24931–42.

62. Wang M, Yang Z, Rada C, Neuberger MS. AID upmutants isolated using a high-throughput screen highlight the immunity/cancer balance limiting DNA deaminase activity. Nat Struct Mol Biol. 2009 Jul;16(7):769–76.

63. Ranjit S, Khair L, Linehan EK, Ucher AJ, Chakrabarti M, Schrader CE, et al. AID recruits UNG and Msh2 to Ig switch regions dependent upon the AID C terminus [corrected]. J Immunol Baltim Md 1950. 2011 Sep 1;187(5):2464–75.

64. Yamane A, Resch W, Kuo N, Kuchen S, Li Z, Sun H, et al. Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. Nat Immunol. 2011 Jan;12(1):62–9.

65. Teater M, Dominguez PM, Redmond D, Chen Z, Ennishi D, Scott DW, et al. AICDA drives epigenetic heterogeneity and accelerates germinal center-derived lymphomagenesis. Nat Commun. 2018 15;9(1):222.

66. Pannunzio NR, Lieber MR. AID and Reactive Oxygen Species Can Induce DNA Breaks within Human Chromosomal Translocation Fragile Zones. Mol Cell. 2017 Dec 7;68(5):901-912.e3.

67. Mu Y, McBride KM. Targeting mutagenesis in B cells: Phosphorylation goes beyond AID association. Mol Cell Oncol. 2018;5(5):e1432259.

68. Matsushita M, Iwasaki T, Nonaka D, Kuwamoto S, Nagata K, Kato M, et al. Higher Expression of Activation-induced Cytidine Deaminase Is Significantly Associated with Merkel Cell Polyomavirus-negative Merkel Cell Carcinomas. Yonago Acta Med. 2017;60(3):145–53.

69. Gelmez MY, Coskunpinar E, Saracoglu B, Deniz G, Aktan M. Investigation of AID, Dicer, and Drosha Expressions in Patients with Chronic Lymphocytic Leukemia. Immunol Invest. 2017 Jul;46(5):433–46.

70. Mandler R, Chu CC, Paul WE, Max EE, Snapper CM. Interleukin 5 induces S mu-S gamma 1 DNA rearrangement in B cells activated with dextran-anti-IgD antibodies and interleukin 4: a three component model for Ig class switching. J Exp Med. 1993 Nov 1;178(5):1577–86.

71. Xu L, Gorham B, Li SC, Bottaro A, Alt FW, Rothman P. Replacement of germ-line epsilon promoter by gene targeting alters control of immunoglobulin heavy chain class switching. Proc Natl Acad Sci U S A. 1993 Apr 15;90(8):3705–9.

72. Zhang J, Bottaro A, Li S, Stewart V, Alt FW. A selective defect in IgG2b switching as a result of targeted mutation of the I gamma 2b promoter and exon. EMBO J. 1993 Sep;12(9):3529–37.

73. Rothenfluh HS, Taylor L, Bothwell AL, Both GW, Steele EJ. Somatic hypermutation in 5' flanking regions of heavy chain antibody variable regions. Eur J Immunol. 1993 Sep;23(9):2152–9.

74. Peters A, Storb U. Somatic hypermutation of immunoglobulin genes is linked to transcription initiation. Immunity. 1996 Jan;4(1):57–65.

75. Fukita Y, Jacobs H, Rajewsky K. Somatic hypermutation in the heavy chain locus correlates with transcription. Immunity. 1998 Jul;9(1):105–14.

76. Storb U, Peters A, Kim N, Shen HM, Bozek G, Michael N, et al. Molecular aspects of somatic hypermutation of immunoglobulin genes. Cold Spring Harb Symp Quant Biol. 1999;64:227–34.

77. Dörner T, Foster SJ, Farner NL, Lipsky PE. Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. Eur J Immunol. 1998;28(10):3384–96.

78. Foster SJ, Dörner T, Lipsky PE. Somatic hypermutation of VkappaJkappa rearrangements: targeting of RGYW motifs on both DNA strands and preferential selection of mutated codons within RGYW motifs. Eur J Immunol. 1999;29(12):4011–21.

79. Chandra V, Bortnick A, Murre C. AID targeting: old mysteries and new challenges. Trends Immunol. 2015 Sep;36(9):527–35.

80. Mu Y, Zelazowska MA, McBride KM. Phosphorylation promotes activation-induced cytidine deaminase activity at the Myc oncogene. J Exp Med. 2017 Dec 4;214(12):3543–52.

81. Das C, Tyler JK, Churchill MEA. The histone shuffle: histone chaperones in an energetic dance. Trends Biochem Sci. 2010 Sep;35(9):476–89.

82. Clapier CR, Cairns BR. The biology of chromatin remodeling complexes. Annu Rev Biochem. 2009;78:273–304.

83. Sheinin MY, Li M, Soltani M, Luger K, Wang MD. Torque modulates nucleosome stability and facilitates H2A/H2B dimer loss. Nat Commun. 2013;4:2579.

84. Kouzine F, Gupta A, Baranello L, Wojtowicz D, Ben-Aissa K, Liu J, et al. Transcription-dependent dynamic supercoiling is a short-range genomic force. Nat Struct Mol Biol. 2013 Mar;20(3):396–403.

85. Shen HM, Poirier MG, Allen MJ, North J, Lal R, Widom J, et al. The activation-induced cytidine deaminase (AID) efficiently targets DNA in nucleosomes but only during transcription. J Exp Med. 2009 May 11;206(5):1057–71.

86. Kodgire P, Mukkawar P, North JA, Poirier MG, Storb U. Nucleosome stability dramatically impacts the targeting of somatic hypermutation. Mol Cell Biol. 2012 May;32(10):2030–40.

87. Basu U, Meng F-L, Keim C, Grinstein V, Pefanis E, Eccleston J, et al. The RNA exosome targets the AID cytidine deaminase to both strands of transcribed duplex DNA substrates. Cell. 2011 Feb 4;144(3):353–63.

88. Boursier L, Su W, Spencer J. Analysis of strand biased “G”.C hypermutation in human immunoglobulin V(lambda) gene segments suggests that both DNA strands are targets for deamination by activation-induced cytidine deaminase. Mol Immunol. 2004 Mar;40(17):1273–8.

89. Abdouni HS, King JJ, Ghorbani A, Fifield H, Berghuis L, Larijani M. DNA/RNA hybrid substrates modulate the catalytic activity of purified AID. Mol Immunol. 2017 Nov 18;93:94–106.

90. Larijani M, Frieder D, Basit W, Martin A. The mutation spectrum of purified AID is similar to the mutability index in Ramos cells and in ung(-/-)msh2(-/-) mice. Immunogenetics. 2005 Feb;56(11):840–5.

91. Ramiro AR, Stavropoulos P, Jankovic M, Nussenzweig MC. Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. Nat Immunol. 2003 May;4(5):452–6.

92. Huang J, Sousa R. T7 RNA polymerase elongation complex structure and movement. J Mol Biol. 2000 Oct 27;303(3):347–58.

93. Larijani M, Frieder D, Sonbuchner TM, Bransteitter R, Goodman MF, Bouhassira EE, et al. Methylation protects cytidines from AID-mediated deamination. Mol Immunol. 2005 Mar;42(5):599–604.

94. Dancyger AM, King JJ, Quinlan MJ, Fifield H, Tucker S, Saunders HL, et al. Differences in the enzymatic efficiency of human and bony fish AID are mediated by a single residue in the C terminus modulating single-stranded DNA binding. Faseb J. 2012 Dec 23;

95. Abdouni H, King JJ, Suliman M, Quinlan M, Fifield H, Larijani M. Zebrafish AID is capable of deaminating methylated deoxycytidines. Nucleic Acids Res. 2013 May 1;41(10):5457–68.

96. King JJ, Manuel CA, Barrett CV, Raber S, Lucas H, Sutter P, et al. Catalytic pocket inaccessibility of activation-induced cytidine deaminase is a safeguard against excessive mutagenic activity. Struct Lond Engl 1993. 2015 Apr 7;23(4):615–27.

97. Yu K, Chedin F, Hsieh CL, Wilson TE, Lieber MR. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. Nat Immunol. 2003;4(5):442-51.

98. Fugmann SD, Schatz DG. Rna Aids DNA. Nat Immunol. 2003 May;4(5):429–30.

99. Kim N, Jinks-Robertson S. Transcription as a source of genome instability. Nat Rev Genet. 2012 Feb 14;13(3):204–14.

100. Dayn A, Malkhosyan S, Mirkin SM. Transcriptionally driven cruciform formation in vivo. Nucleic Acids Res. 1992 Nov 25;20(22):5991–7.

101. Kuetche VK. Ab initio bubble-driven denaturation of double-stranded DNA: Self-mechanical theory. J Theor Biol. 2016 21;401:15–29.

102. Krasilnikov AS, Podtelezhnikov A, Vologodskii A, Mirkin SM. Large-scale effects of transcriptional DNA supercoiling in vivo. J Mol Biol. 1999 Oct 8;292(5):1149–60.

103. Naughton C, Avlonitis N, Corless S, Prendergast JG, Mati IK, Eijk PP, et al. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nat Struct Mol Biol. 2013 Mar;20(3):387–95.

104. Rajagopal D, Maul RW, Ghosh A, Chakraborty T, Khamlichi AA, Sen R, et al. Immunoglobulin switch mu sequence causes RNA polymerase II accumulation and reduces dA hypermutation. J Exp Med. 2009 Jun 8;206(6):1237–44.

105. Canugovi C, Samaranayake M, Bhagwat AS. Transcriptional pausing and stalling causes multiple clustered mutations by human activation-induced deaminase. Faseb J. 2009 Jan;23(1):34–44.

106. Storb U, Peters A, Klotz E, Kim N, Shen HM, Hackett J, et al. Cis-acting sequences that affect somatic hypermutation of Ig genes. Immunol Rev. 1998 Apr;162:153–60.

107. Altan-Bonnet G, Libchaber A, Krichevsky O. Bubble dynamics in double-stranded DNA. Phys Rev Lett. 2003 Apr 4;90(13):138101.

108. Storck S, Aoufouchi S, Weill J-C, Reynaud C-A. AID and partners: for better and (not) for worse. Curr Opin Immunol. 2011 Jun;23(3):337–44.

109. Longerich S, Basu U, Alt F, Storb U. AID in somatic hypermutation and class switch recombination. Curr Opin Immunol. 2006 Apr;18(2):164–74.

110. Shen HM, Ratnam S, Storb U. Targeting of the activation-induced Cytosine deaminase is strongly influenced by the sequence and structure of the targeted DNA. Mol Cell Biol. 2005 Dec;25(24):10815–21.

111. Zan H, Shima N, Xu Z, Al-Qahtani A, Evinger Iii AJ, Zhong Y, et al. The translesion DNA polymerase theta plays a dominant role in immunoglobulin gene somatic hypermutation. Embo J. 2005 Nov 2;24(21):3757–69.

112. Stavnezer J, Schrader CE. Mismatch repair converts AID-instigated nicks to double-strand breaks for antibody class-switch recombination. Trends Genet. 2006 Jan;22(1):23–8.

113. Casali P, Pal Z, Xu Z, Zan H. DNA repair in antibody somatic hypermutation. Trends Immunol. 2006 Jul;27(7):313–21.

114. Di Noia J, Neuberger MS. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. Nature. 2002;419(6902):43-8.

115. Liu M, Duke JL, Richter DJ, Vinuesa CG, Goodnow CC, Kleinstein SH, et al. Two levels of protection for the B cell genome during somatic hypermutation. Nature. 2008 Feb 14;451(7180):841–5.

116. Imamura M, Miura K, Iwabuchi K, Ichisaka T, Nakagawa M, Lee J, et al. Transcriptional repression and DNA hypermethylation of a small set of ES cell marker genes in male germline stem cells. BMC Dev Biol. 2006 Jul 21;6:34.

117. Brenet F, Moh M, Funk P, Feierstein E, Viale AJ, Socci ND, et al. DNA methylation of the first exon is tightly linked to transcriptional silencing. PloS One. 2011 Jan 18;6(1):e14524.

118. Song F, Mahmood S, Ghosh S, Liang P, Smiraglia DJ, Nagase H, et al. Tissue specific differentially methylated regions (TDMR): Changes in DNA methylation during development. Genomics. 2009 Feb;93(2):130–9.

119. Chen C, Yang MC, Yang TP. Evidence that silencing of the HPRT promoter by DNA methylation is mediated by critical CpG sites. J Biol Chem. 2001 Jan 5;276(1):320–8.

120. Kroft TL, Jethanandani P, McLean DJ, Goldberg E. Methylation of CpG dinucleotides alters binding and silences testis-specific transcription directed by the mouse lactate dehydrogenase C promoter. Biol Reprod. 2001 Nov;65(5):1522–7.

121. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002 Jan 1;16(1):6–21.

122. Zhang F, Pomerantz JH, Sen G, Palermo AT, Blau HM. Active tissue-specific DNA demethylation conferred by somatic cell nuclei in stable heterokaryons. Proc Natl Acad Sci U A. 2007 Mar 13;104(11):4395–400.

123. Goll MG, Bestor TH. Eukaryotic cytosine methyltransferases. Annu Rev Biochem. 2005;74:481–514.

124. Lio C-WJ, Rao A. TET Enzymes and 5hmC in Adaptive and Innate Immune Systems. Front Immunol. 2019;10:210.

125. Hajkova P, Jeffries SJ, Lee C, Miller N, Jackson SP, Surani MA. Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. Science. 2010 Jul 2;329(5987):78–82.

126. Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature. 2010 Aug 26;466(7310):1129–33.

127. Iqbal K, Jin SG, Pfeifer GP, Szabo PE. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proc Natl Acad Sci U A. 2011 Mar 1;108(9):3642–7.

128. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009 May 15;324(5929):930–5.

129. Guo JU, Su Y, Zhong C, Ming GL, Song H. Emerging roles of TET proteins and 5-hydroxymethylcytosines in active DNA demethylation and beyond. Cell Cycle. 2011 Aug 15;10(16):2662–8.

130. Guo JU, Su Y, Zhong C, Ming GL, Song H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell. 2011 Apr 29;145(3):423–34.

131. Wossidlo M, Arand J, Sebastiano V, Lepikhov K, Boiani M, Reinhardt R, et al. Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. Embo J. 2010 Jun 2;29(11):1877–88.

132. Koh KP, Yabuuchi A, Rao S, Huang Y, Cunniff K, Nardone J, et al. Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell. 2011 Feb 4;8(2):200–13.

133. Fritz EL, Papavasiliou FN. Cytidine deaminases: AIDing DNA demethylation? Genes Dev. 2010 Oct 1;24(19):2107–14.

134. Morgan HD, Dean W, Coker HA, Reik W, Petersen-Mahrt SK. Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. J Biol Chem. 2004 Dec 10;279(50):52353–60.

135. Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, et al. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature. 2010 Feb 25;463(7284):1101–5.

136. Bhutani N, Brady JJ, Damian M, Sacco A, Corbel SY, Blau HM. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature. 2010 Feb 25;463(7284):1042–7.

137. Rai K, Huggins IJ, James SR, Karpf AR, Jones DA, Cairns BR. DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. Cell. 2008 Dec 26;135(7):1201–12.

138. Teperek-Tkacz M, Pasque V, Gentsch G, Ferguson-Smith AC. Epigenetic reprogramming: is deamination key to active DNA demethylation? Reprod Camb Engl. 2011 Nov;142(5):621–32.

139. Dominguez PM, Teater M, Chambwe N, Kormaksson M, Redmond D, Ishii J, et al. DNA Methylation Dynamics of Germinal Center B Cells Are Mediated by AID. Cell Rep. 2015 Sep 29;12(12):2086–98.

140. Pasqualucci L, Bereschenko O, Niu H, Klein U, Basso K, Guglielmino R, et al. Molecular pathogenesis of non-Hodgkin's lymphoma: the role of Bcl-6. Leuk Lymphoma. 2003;44 Suppl 3:S5-12.

141. Muschen M, Re D, Jungnickel B, Diehl V, Rajewsky K, Kuppers R. Somatic mutation of the CD95 gene in human B cells as a side-effect of the germinal center reaction. J Exp Med. 2000;192(12):1833-40.

142. Pasqualucci L, Neri A, Baldini L, Dalla-Favera R, Migliazza A. BCL-6 mutations are associated with immunoglobulin variable heavy chain mutations in B-cell chronic lymphocytic leukemia. Cancer Res. 2000;60(20):5644-8.

143. Robbiani DF, Bunting S, Feldhahn N, Bothmer A, Camps J, Deroubaix S, et al. AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. Mol Cell. 2009 Nov 25;36(4):631–41.

144. Robbiani DF, Bothmer A, Callen E, Reina-San-Martin B, Dorsett Y, Difilippantonio S, et al. AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. Cell. 2008 Dec 12;135(6):1028–38.

145. Pasqualucci L, Bhagat G, Jankovic M, Compagno M, Smith P, Muramatsu M, et al. AID is required for germinal center-derived lymphomagenesis. Nat Genet. 2008 Jan;40(1):108–12.

146. Lossos IS, Levy R, Alizadeh AA. AID is expressed in germinal center B-cell-like and activated B-cell-like diffuse large-cell lymphomas and is not correlated with intraclonal heterogeneity. Leukemia. 2004 Nov;18(11):1775–9.

147. Kuppers R, Dalla-Favera R. Mechanisms of chromosomal translocations in B cell lymphomas. Oncogene. 2001;20(40):5580-94.

148. Kuppers R. Somatic hypermutation and B cell receptor selection in normal and transformed human B cells. Ann N Acad Sci. 2003 Apr;987:173–9.

149. Willis TG, Dyer MJ. The role of immunoglobulin translocations in the pathogenesis of B-cell malignancies. Blood. 2000 Aug 1;96(3):808–22.

150. Marantidou F, Dagklis A, Stalika E, Korkolopoulou P, Saetta A, Anagnostopoulos A, et al. Activation-induced cytidine deaminase splicing patterns in chronic lymphocytic leukemia. Blood Cells Mol Dis. 2010 Apr 15;44(4):262–7.

151. Aguilera NS, Auerbach A, Barekman CL, Lichy J, Abbondanzo SL. Activation-induced cytidine deaminase expression in diffuse large B-cell lymphoma with a paracortical growth pattern: a lymphoma of possible interfollicular large B-cell origin. Arch Pathol Lab Med. 2010 Mar;134(3):449–56.

152. Palacios F, Moreno P, Morande P, Abreu C, Correa A, Porro V, et al. High expression of AID and active class switch recombination might account for a more aggressive disease in unmutated CLL patients: link with an activated microenvironment in CLL disease. Blood. 2010 Jun 3;115(22):4488–96.

153. Hancer VS, Kose M, Diz-Kucukkaya R, Yavuz AS, Aktan M. Activation-induced cytidine deaminase mRNA levels in chronic lymphocytic leukemia. Leuk Lymphoma. 2011 Jan;52(1):79–84.

154. Kano C, Wang JY. High levels of AID cause strand bias of mutations at A versus T in Burkitt's lymphoma cells. Mol Immunol. 2013 Jul;54(3–4):397–402.

155. Huemer M, Rebhandl S, Zaborsky N, Gassner FJ, Hainzl S, Weiss L, et al. AID induces intraclonal diversity and genomic damage in CD86(+) chronic lymphocytic leukemia cells. Eur J Immunol. 2014 Dec;44(12):3747–57.

156. Klemm L, Duy C, Iacobucci I, Kuchen S, von Levetzow G, Feldhahn N, et al. The B cell mutator AID promotes B lymphoid blast crisis and drug resistance in chronic myeloid leukemia. Cancer Cell. 2009 Sep 8;16(3):232–45.

157. Kou T, Marusawa H, Kinoshita K, Endo Y, Okazaki IM, Ueda Y, et al. Expression of activation-induced cytidine deaminase in human hepatocytes during hepatocarcinogenesis. Int J Cancer. 2007 Feb 1;120(3):469–76.

158. Marusawa H, Takai A, Chiba T. Role of activation-induced cytidine deaminase in inflammation-associated cancer development. Adv Immunol. 2011;111:109–41.

159. Munoz DP, Lee EL, Takayama S, Coppe JP, Heo SJ, Boffelli D, et al. Activation-induced cytidine deaminase (AID) is necessary for the epithelial-mesenchymal transition in mammary epithelial cells. Proc Natl Acad Sci U S A. 2013 Aug 6;110(32):E2977-86.

160. Compagno M, Wang Q, Pighi C, Cheong TC, Meng FL, Poggio T, et al. Phosphatidylinositol 3-kinase delta blockade increases genomic instability in B cells. Nature. 2017 Feb 23;542(7642):489–93.

161. Fruman DA, O'Brien S. Cancer: A targeted treatment with off-target risks. Nature. 2017 Feb 23;542(7642):424–5.

162. Mechtcheriakova D, Sobanov Y, Holtappels G, Bajna E, Svoboda M, Jaritz M, et al. Activation-induced cytidine deaminase (AID)-associated multigene signature to assess impact of AID in etiology of diseases with inflammatory component. PloS One. 2011;6(10):e25611.

163. Sapoznik S, Bahar-Shany K, Brand H, Pinto Y, Gabay O, Glick-Saar E, et al. Activation-Induced Cytidine Deaminase Links Ovulation-Induced Inflammation and Serous Carcinogenesis. Neoplasia N Y N. 2016 Feb;18(2):90–9.

164. King JJ, Larijani M. A Novel Regulator of Activation-Induced Cytidine Deaminase/APOBECs in Immunity and Cancer: Schrodinger's CATalytic Pocket. Front Immunol. 2017;8:351.

165. Chelico L, Sacho EJ, Erie DA, Goodman MF. A model for oligomeric regulation of APOBEC3G cytosine deaminase-dependent restriction of HIV. J Biol Chem. 2008 May 16;283(20):13780–91.

166. Shlyakhtenko LS, Lushnikov AY, Miyagi A, Li M, Harris RS, Lyubchenko YL. Atomic force microscopy studies of APOBEC3G oligomerization and dynamics. J Struct Biol. 2013 Nov;184(2):217–25.

167. Fiser A. Template-based protein structure modeling. Methods Mol Biol Clifton NJ. 2010;673:73–94.

168. Meng X-Y, Zhang H-X, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des. 2011 Jun;7(2):146–57.

169. Morris GM, Lim-Wilby M. Molecular docking. Methods Mol Biol Clifton NJ. 2008;443:365–82.

170. Prochnow C, Bransteitter R, Klein MG, Goodman MF, Chen XS. The APOBEC-2 crystal structure and functional implications for the deaminase AID. Nature. 2007 Jan 25;445(7126):447–51.

171. Patenaude AM, Orthwein A, Hu Y, Campo VA, Kavli B, Buschiazzo A, et al. Active nuclear import and cytoplasmic retention of activation-induced deaminase. Nat Struct Mol Biol. 2009 May;16(5):517–27.

172. Hardin C, Pogorelov TV, Luthey-Schulten Z. Ab initio protein structure prediction. Curr Opin Struct Biol. 2002 Apr;12(2):176–81.

173. Conticello SG. The AID/APOBEC family of nucleic acid mutators. Genome Biol. 2008;9(6):229.

174. Yu K, Huang F-T, Lieber MR. DNA substrate length and surrounding sequence affect the activation-induced deaminase activity at cytidine. J Biol Chem. 2004 Feb 20;279(8):6496–500.

175. Chaudhuri J, Alt FW. Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. Nat Rev Immunol. 2004 Jul;4(7):541–52.

176. Durandy A, Peron S, Fischer A. Hyper-IgM syndromes. Curr Opin Rheumatol. 2006 Jul;18(4):369–76.

177. Ramiro AR, Jankovic M, Eisenreich T, Difilippantonio S, Chen-Kiang S, Muramatsu M, et al. AID is required for c-myc/IgH chromosome translocations in vivo. Cell. 2004 Aug 20;118(4):431–8.

178. Bird AP. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 1980 Apr 11;8(7):1499–504.

179. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009 Nov 19;462(7271):315–22.

180. Doerfler W. In pursuit of the first recognized epigenetic signal--DNA methylation: a 1976 to 2008 synopsis. Epigenetics. 2008 Jun;3(3):125–33.

181. Antequera F, Bird A. Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A. 1993 Dec 15;90(24):11995–9.

182. Durandy A, Revy P, Fischer A. Hyper-immunoglobulin-M syndromes caused by an intrinsic B cell defect. Curr Opin Allergy Clin Immunol. 2003 Dec;3(6):421–5.

183. Wijesinghe P, Bhagwat AS. Efficient deamination of 5-methylcytosines in DNA by human APOBEC3A, but not by AID or APOBEC3G. Nucleic Acids Res. 2012 Oct 1;40(18):9206–17.

184. Nabel CS, Jia H, Ye Y, Shen L, Goldschmidt HL, Stivers JT, et al. AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. Nat Chem Biol. 2012 Sep;8(9):751–8.

185. Holden LG, Prochnow C, Chang YP, Bransteitter R, Chelico L, Sen U, et al. Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications. Nature. 2008 Nov 6;456(7218):121–4.

186. Shandilya SM, Nalam MN, Nalivaika EA, Gross PJ, Valesano JC, Shindo K, et al. Crystal structure of the APOBEC3G catalytic domain reveals potential oligomerization interfaces. Structure. 2010 Jan 13;18(1):28–38.

187. Furukawa A, Nagata T, Matsugami A, Habu Y, Sugiyama R, Hayashi F, et al. Structure, interaction and real-time monitoring of the enzymatic reaction of wild-type APOBEC3G. Embo J. 2009 Feb 18;28(4):440–51.

188. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010 Jan 30;31(2):455–61.

189. Betts L, Xiang S, Short SA, Wolfenden R, Carter CW. Cytidine deaminase. The 2.3 A crystal structure of an enzyme: transition-state analog complex. J Mol Biol. 1994 Jan 14;235(2):635–56.

190. Mejlhede N, Neuhard J. The role of zinc in Bacillus subtilis cytidine deaminase. Biochemistry. 2000 Jul 11;39(27):7984–9.

191. Ko TP, Lin JJ, Hu CY, Hsu YH, Wang AH, Liaw SH. Crystal structure of yeast cytosine deaminase. Insights into enzyme mechanism and evolution. J Biol Chem. 2003 May 23;278(21):19111–7.

192. Losey HC, Ruthenburg AJ, Verdine GL. Crystal structure of Staphylococcus aureus tRNA adenosine deaminase TadA in complex with RNA. Nat Struct Mol Biol. 2006 Feb;13(2):153–9.

193. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011 May 15;25(10):1010–22.

194. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, et al. Dynamic changes in the human methylome during differentiation. Genome Res. 2010 Mar;20(3):320–31.

195. Shiota K. DNA methylation profiles of CpG islands for cellular differentiation and development in mammals. Cytogenet Genome Res. 2004;105(2–4):325–34.

196. Sato S, Yagi S, Arai Y, Hirabayashi K, Hattori N, Iwatani M, et al. Genome-wide DNA methylation profile of tissue-dependent and differentially methylated regions (T-DMRs) residing in mouse pluripotent stem cells. Genes Cells. 2010 Jun;15(6):607–18.

197. van Vlodrop IJ, Niessen HE, Derks S, Baldewijns MM, van Criekinge W, Herman JG, et al. Analysis of promoter CpG island hypermethylation in cancer: location, location, location! Clin Cancer Res. 2011 Jul 1;17(13):4225–31.

198. Das PM, Singal R. DNA methylation and cancer. J Clin Oncol. 2004 Nov 15;22(22):4632–42.

199. Carpenter MA, Li M, Rathore A, Lackey L, Law EK, Land AM, et al. Methylcytosine and Normal Cytosine Deamination by the Foreign DNA Restriction Enzyme APOBEC3A. J Biol Chem. 2012 Oct 5;287(41):34801–8.

200. Parsa JY, Basit W, Wang CL, Gommerman JL, Carlyle JR, Martin A. AID mutates a non-immunoglobulin transgene independent of chromosomal position. Mol Immunol. 2007 Jan;44(4):567–75.

201. Conticello SG, Thomas CJ, Petersen-Mahrt SK, Neuberger MS. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Mol Biol Evol. 2005 Feb;22(2):367–77.

202. Barreto VM, Magor BG. Activation-induced cytidine deaminase structure and functions: A species comparative view. Dev Comp Immunol. 2011 Sep;35(9):991–1007.

203. McCarthy H, Wierda WG, Barron LL, Cromwell CC, Wang J, Coombes KR, et al. High expression of activation-induced cytidine deaminase (AID) and splice variants is a distinctive feature of poor-prognosis chronic lymphocytic leukemia. Blood. 2003 Jun 15;101(12):4903–8.

204. Okazaki IM, Hiai H, Kakazu N, Yamada S, Muramatsu M, Kinoshita K, et al. Constitutive expression of AID leads to tumorigenesis. J Exp Med. 2003 May 5;197(9):1173–81.

205. Conticello SG, Langlois MA, Yang Z, Neuberger MS. DNA deamination in immunity: AID in the context of its APOBEC relatives. Adv Immunol. 2007;94:37–73.

206. Coker HA, Morgan HD, Petersen-Mahrt SK. Genetic and in vitro assays of DNA deamination. Methods Enzym. 2006;408:156–70.

207. Krzysiak TC, Jung J, Thompson J, Baker D, Gronenborn AM. APOBEC2 is a monomer in solution: implications for APOBEC3G models. Biochemistry. 2012 Mar 6;51(9):2008–17.

208. Byeon IJ, Ahn J, Mitra M, Byeon CH, Hercik K, Hritz J, et al. NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. Nat Commun. 2013;4:1890.

209. Opi S, Takeuchi H, Kao S, Khan MA, Miyagi E, Goila-Gaur R, et al. Monomeric APOBEC3G is catalytically active and has antiviral activity. J Virol. 2006 May;80(10):4673–82.

210. Kitamura S, Ode H, Nakashima M, Imahashi M, Naganawa Y, Kurosawa T, et al. The APOBEC3C crystal structure and the interface for HIV-1 Vif binding. Nat Struct Mol Biol. 2012 Oct;19(10):1005–10.

211. Bohn MF, Shandilya SM, Albin JS, Kouno T, Anderson BD, McDougle RM, et al. Crystal structure of the DNA cytosine deaminase APOBEC3F: the catalytically active and HIV-1 Vif-binding domain. Structure. 2013 Jun 4;21(6):1042–50.

212. Prochnow C, Bransteitter R, Chen XS. APOBEC deaminases-mutases with defensive roles for immunity. Sci China C Life Sci. 2009 Oct;52(10):893–902.

213. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9:40.

214. Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W471-7.

215. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011 Feb 1;27(3):343–50.

216. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR. 1996 Dec;8(4):477–86.

217. Grosdidier A, Zoete V, Michielin O. Fast docking using the CHARMM force field with EADock DSS. J Comput Chem. 2011 May 3;

218. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W270-7.

219. Zoete V, Cuendet MA, Grosdidier A, Michielin O. SwissParam: a fast force field generation tool for small organic molecules. J Comput Chem. 2011 Aug;32(11):2359–68.

220. Xie K, Sowden MP, Dance GS, Torelli AT, Smith HC, Wedekind JE. The structure of a yeast RNA-editing deaminase provides insight into the fold and function of activation-induced deaminase and APOBEC-1. Proc Natl Acad Sci U A. 2004 May 25;101(21):8114–9.

221. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25(13):1605–12.

222. Smith HC, Bennett RP, Kizilyer A, McDougall WM, Prohaska KM. Functions and regulation of the APOBEC family of proteins. Semin Cell Dev Biol. 2012 May;23(3):258–68.

223. Munk C, Willemsen A, Bravo IG. An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. BMC Evol Biol. 2012;12:71.

224. Schneider M, Fu X, Keating AE. X-ray vs. NMR structures as templates for computational protein design. Proteins. 2009 Oct;77(1):97–110.

225. Igumenova TI, Frederick KK, Wand AJ. Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. Chem Rev. 2006 May;106(5):1672–99.

226. Conticello SG, Langlois MA, Neuberger MS. Insights into DNA deaminases. Nat Struct Mol Biol. 2007 Jan;14(1):7–9.

227. Bulliard Y, Narvaiza I, Bertero A, Peddi S, Rohrig UF, Ortiz M, et al. Structure-function analyses point to a polynucleotide-accommodating groove essential for APOBEC3A restriction activities. J Virol. 2011 Feb;85(4):1765–76.

228. Mitra M, Hercík K, Byeon I-JL, Ahn J, Hill S, Hinchee-Rodriguez K, et al. Structural determinants of human APOBEC3A enzymatic and nucleic acid binding properties. Nucleic Acids Res. 2014 Jan;42(2):1095–110.

229. Love RP, Xu H, Chelico L. Biochemical analysis of hypermutation by the deoxycytidine deaminase APOBEC3A. J Biol Chem. 2012 Aug 31;287(36):30812–22.

230. Mak CH, Pham P, Afif SA, Goodman MF. A mathematical model for scanning and catalysis on single-stranded DNA, illustrated with activation-induced deoxycytidine deaminase. J Biol Chem. 2013 Oct 11;288(41):29786–95.

231. Ramiro A, Reina San-Martin B, McBride K, Jankovic M, Barreto V, Nussenzweig A, et al. The role of activation-induced deaminase in antibody diversification and chromosome translocations. Adv Immunol. 2007;94:75–107.

232. Harris RS. Enhancing immunity to HIV through APOBEC. Nat Biotechnol. 2008 Oct;26(10):1089–90.

233. Stavrou S, Ross SR. APOBEC3 Proteins in Viral Immunity. J Immunol. 2015 Nov 15;195(10):4565–70.

234. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013 Sep;45(9):970–6.

235. Roberts SA, Gordenin DA. Clustered and genome-wide transient mutagenesis in human cancers: Hypermutation without permanent mutators or loss of fitness. BioEssays News Rev Mol Cell Dev Biol. 2014;36(4):382–93.

236. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. Nat Rev Cancer. 2014;14(12):786–800.

237. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug 22;500(7463):415–21.

238. Burns MB, Leonard B, Harris RS. APOBEC3B: pathological consequences of an innate immune DNA mutator. Biomed J. 2015 Apr;38(2):102–10.

239. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nat Genet. 2015 Sep;47(9):1067–72.

240. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, et al. APOBEC3B is an enzymatic source of mutation in breast cancer. Nature. 2013 Feb 21;494(7437):366–70.

241. Leonard B, Hart SN, Burns MB, Carpenter MA, Temiz NA, Rathore A, et al. APOBEC3B upregulation and genomic mutation patterns in serous ovarian carcinoma. Cancer Res. 2013 Dec 15;73(24):7222–31.

242. Knisbacher BA, Gerber D, Levanon EY. DNA Editing by APOBECs: A Genomic Preserver and Transformer. Trends Genet TIG. 2016 Jan;32(1):16–28.

243. Rebhandl S, Huemer M, Greil R, Geisberger R. AID/APOBEC deaminases and cancer. Oncoscience. 2015;2(4):320–33.

244. Starrett GJ, Luengas EM, McCann JL, Ebrahimi D, Temiz NA, Love RP, et al. The DNA cytosine deaminase APOBEC3H haplotype I likely contributes to breast and lung cancer mutagenesis. Nat Commun. 2016 21;7:12918.

245. Pham P, Afif SA, Shimoda M, Maeda K, Sakaguchi N, Pedersen LC, et al. Structural analysis of the activation-induced deoxycytidine deaminase required in immunoglobulin diversification. DNA Repair Amst. 2016 Jul;43:48–56.

246. Shi K, Carpenter MA, Kurahashi K, Harris RS, Aihara H. Crystal Structure of the DNA Deaminase APOBEC3B Catalytic Domain. J Biol Chem. 2015 Nov 20;290(47):28120–30.

247. Byeon IJ, Byeon CH, Wu T, Mitra M, Singer D, Levin JG, et al. Nuclear Magnetic Resonance Structure of the APOBEC3B Catalytic Domain: Structural Basis for Substrate Binding and DNA Deaminase Activity. Biochemistry. 2016 May 31;55(21):2944–59.

248. Siu KK, Sultana A, Azimi FC, Lee JE. Structural determinants of HIV-1 Vif susceptibility and DNA binding in APOBEC3F. Nat Commun. 2013;4:2593.

249. Nakashima M, Ode H, Kawamura T, Kitamura S, Naganawa Y, Awazu H, et al. Structural Insights into HIV-1 Vif-APOBEC3F Interaction. J Virol. 2016 Jan;90(2):1034–47.

250. Shaban NM, Shi K, Li M, Aihara H, Harris RS. 1.92 Angstrom Zinc-Free APOBEC3F Catalytic Domain Crystal Structure. J Mol Biol. 2016 Jun 5;428(11):2307–16.

251. Chen KM, Harjes E, Gross PJ, Fahmy A, Lu Y, Shindo K, et al. Structure of the DNA deaminase domain of the HIV-1 restriction factor APOBEC3G. Nature. 2008 Mar 6;452(7183):116–9.

252. Harjes E, Gross PJ, Chen KM, Lu Y, Shindo K, Nowarski R, et al. An extended structure of the APOBEC3G catalytic domain suggests a unique holoenzyme model. J Mol Biol. 2009 Jun 26;389(5):819–32.

253. Li M, Shandilya SM, Carpenter MA, Rathore A, Brown WL, Perkins AL, et al. First-in-class small molecule inhibitors of the single-strand DNA cytosine deaminase APOBEC3G. ACS Chem Biol. 2012 Mar 16;7(3):506–17.

254. Lu X, Zhang T, Xu Z, Liu S, Zhao B, Lan W, et al. Crystal structure of DNA cytidine deaminase ABOBEC3G catalytic deamination domain suggests a binding mode of full-length enzyme to single-stranded DNA. J Biol Chem. 2015 Feb 13;290(7):4010–21.

255. Kouno T, Luengas EM, Shigematsu M, Shandilya SM, Zhang J, Chen L, et al. Structure of the Vif-binding domain of the antiviral enzyme APOBEC3G. Nat Struct Mol Biol. 2015 Jun;22(6):485–91.

256. Xiao X, Li SX, Yang H, Chen XS. Crystal structures of APOBEC3G N-domain alone and its complex with DNA. Nat Commun. 2016;7:12193.

257. Shi K, Carpenter MA, Banerjee S, Shaban NM, Kurahashi K, Salamango DJ, et al. Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B. Nat Struct Mol Biol. 2017 Feb;24(2):131–9.

258. Marx A, Galilee M, Alian A. Zinc enhancement of cytidine deaminase activity highlights a potential allosteric role of loop-3 in regulating APOBEC3 enzymes. Sci Rep. 2015 Dec 18;5:18191.

259. Faltas BM, Prandi D, Tagawa ST, Molina AM, Nanus DM, Sternberg C, et al. Clonal evolution of chemotherapy-resistant urothelial carcinoma. Nat Genet. 2016 Oct 17;

260. Orthwein A, Zahn A, Methot SP, Godin D, Conticello SG, Terada K, et al. Optimal functional levels of activation-induced deaminase specifically require the Hsp40 DnaJa1. EMBO J. 2012 Feb 1;31(3):679–91.

261. Orthwein A, Patenaude AM, Affar el B, Lamarre A, Young JC, Di Noia JM. Regulation of activation-induced deaminase stability and antibody gene diversification by Hsp90. J Exp Med. 2010 Nov 22;207(12):2751–65.

262. McBride KM, Barreto V, Ramiro AR, Stavropoulos P, Nussenzweig MC. Somatic hypermutation is limited by CRM1-dependent nuclear export of activation-induced deaminase. J Exp Med. 2004 May 3;199(9):1235–44.

263. Han L, Masani S, Yu K. Cutting edge: CTNNBL1 is dispensable for Ig class switch recombination. J Immunol Baltim Md 1950. 2010 Aug 1;185(3):1379–81.

264. Conticello SG, Ganesh K, Xue K, Lu M, Rada C, Neuberger MS. Interaction between antibody-diversification enzyme AID and spliceosome-associated factor CTNNBL1. Mol Cell. 2008 Aug 22;31(4):474–84.

265. Uchimura Y, Barton LF, Rada C, Neuberger MS. REG-γ associates with and modulates the abundance of nuclear activation-induced deaminase. J Exp Med. 2011 Nov 21;208(12):2385–91.

266. Häsler J, Rada C, Neuberger MS. Cytoplasmic activation-induced cytidine deaminase (AID) exists in stoichiometric complex with translation elongation factor 1α (eEF1A). Proc Natl Acad Sci U S A. 2011 Nov 8;108(45):18366–71.

267. Pavri R, Gazumyan A, Jankovic M, Di Virgilio M, Klein I, Ansarah-Sobrinho C, et al. Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. Cell. 2010 Oct 1;143(1):122–33.

268. Chaudhuri J, Khuong C, Alt FW. Replication protein A interacts with AID to promote deamination of somatic hypermutation targets. Nature. 2004;doi:10.1038/nature02821.

269. Jeevan-Raj BP, Robert I, Heyer V, Page A, Wang JH, Cammas F, et al. Epigenetic tethering of AID to the donor switch region during immunoglobulin class switch recombination. J Exp Med. 2011 Aug 1;208(8):1649–60.

270. Okazaki I, Okawa K, Kobayashi M, Yoshikawa K, Kawamoto S, Nagaoka H, et al. Histone chaperone Spt6 is required for class switch recombination but not somatic hypermutation. Proc Natl Acad Sci U S A. 2011 May 10;108(19):7920–5.

271. Basu U, Chaudhuri J, Alpert C, Dutt S, Ranganath S, Li G, et al. The AID antibody diversification enzyme is regulated by protein kinase A phosphorylation. Nature. 2005 Nov 24;438(7067):508–11.

272. Pasqualucci L, Kitaura Y, Gu H, Dalla-Favera R. PKA-mediated phosphorylation regulates the function of activation-induced deaminase (AID) in B cells. Proc Natl Acad Sci U A. 2006 Jan 10;103(2):395–400.

273. Demorest ZL, MacDuff DA, Brown WL, Morham SG, Parise LV, Harris RS. The interaction between AID and CIB1 is nonessential for antibody gene diversification by gene conversion or class switch recombination. PloS One. 2010 Jul 20;5(7):e11660.

274. MacDuff DA, Neuberger MS, Harris RS. MDM2 can interact with the C-terminus of AID but it is inessential for antibody diversification in DT40 B cells. Mol Immunol. 2006 Mar;43(8):1099–108.

275. Nowak U, Matthews AJ, Zheng S, Chaudhuri J. The splicing regulator PTBP2 interacts with the cytidine deaminase AID and promotes binding of AID to switch-region DNA. Nat Immunol. 2011 Feb;12(2):160–6.

276. Hu W, Begum NA, Mondal S, Stanlie A, Honjo T. Identification of DNA cleavage- and recombination-specific hnRNP cofactors for activation-induced cytidine deaminase. Proc Natl Acad Sci U S A. 2015 May 5;112(18):5791–6.

277. Delker RK, Zhou Y, Strikoudis A, Stebbins CE, Papavasiliou FN. Solubility-based genetic screen identifies RING finger protein 126 as an E3 ligase for activation-induced cytidine deaminase. Proc Natl Acad Sci U S A. 2013 Jan 15;110(3):1029–34.

278. Kracker S, Di Virgilio M, Schwartzentruber J, Cuenin C, Forveille M, Deau MC, et al. An inherited immunoglobulin class-switch recombination deficiency associated with a defect in the INO80 chromatin remodeling complex. J Allergy Clin Immunol. 2015 Apr;135(4):998-1007 e6.

279. Willmann KL, Milosevic S, Pauklin S, Schmitz KM, Rangam G, Simon MT, et al. A role for the RNA pol II-associated PAF complex in AID-induced immune diversification. J Exp Med. 2012 Oct 22;209(11):2099–111.

280. Xu Z, Fulop Z, Wu G, Pone EJ, Zhang J, Mai T, et al. 14-3-3 adaptor proteins recruit AID to 5'-AGCT-3'-rich switch regions for class switch recombination. Nat Struct Mol Biol. 2010 Sep;17(9):1124–35.

281. Kanehiro Y, Todo K, Negishi M, Fukuoka J, Gan W, Hikasa T, et al. Activation-induced cytidine deaminase (AID)-dependent somatic hypermutation requires a splice isoform of the serine/arginine-rich (SR) protein SRSF1. Proc Natl Acad Sci U S A. 2012 Jan 24;109(4):1216–21.

282. Fu Y, Ito F, Zhang G, Fernandez B, Yang H, Chen XS. DNA cytosine and methylcytosine deamination by APOBEC3B: enhancing methylcytosine deamination by engineering APOBEC3B. Biochem J. 2015 Oct 1;471(1):25–35.

283. Bohn MF, Shandilya SM, Silvas TV, Nalivaika EA, Kouno T, Kelch BA, et al. The ssDNA Mutator APOBEC3A Is Regulated by Cooperative Dimerization. Structure. 2015 May 5;23(5):903–11.

284. Qian J, Wang Q, Dose M, Pruett N, Kieffer-Kwon K-R, Resch W, et al. B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. Cell. 2014 Dec 18;159(7):1524–37.

285. Alinikula J, Schatz DG. Super-enhancer transcription converges on AID. Cell. 2014 Dec 18;159(7):1490–2.

286. Meng F-L, Du Z, Federation A, Hu J, Wang Q, Kieffer-Kwon K-R, et al. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. Cell. 2014 Dec 18;159(7):1538–48.

287. Martin A, Bardwell PD, Woo CJ, Fan M, Shulman MJ, Scharff MD. Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas. Nature. 2002;415(6873):802-6.

288. Pavri R, Nussenzweig MC. AID targeting in antibody diversity. Adv Immunol. 2011;110:1–26.

289. Wang CL, Harper RA, Wabl M. Genome-wide somatic hypermutation. Proc Natl Acad Sci U A. 2004 May 11;101(19):7352–6.

290. Maeda K, Almofty SA, Singh SK, Eid MMA, Shimoda M, Ikeda T, et al. GANP interacts with APOBEC3G and facilitates its encapsidation into the virions to reduce HIV-1 infectivity. J Immunol Baltim Md 1950. 2013 Dec 15;191(12):6030–9.

291. Wang X, Ao Z, Chen L, Kobinger G, Peng J, Yao X. The cellular antiviral protein APOBEC3G interacts with HIV-1 reverse transcriptase and inhibits its function during viral replication. J Virol. 2012 Apr;86(7):3777–86.

292. Schafer A, Bogerd HP, Cullen BR. Specific packaging of APOBEC3G into HIV-1 virions is mediated by the nucleocapsid domain of the gag polyprotein precursor. Virology. 2004 Oct 25;328(2):163–8.

293. Valera M-S, de Armas-Rillo L, Barroso-González J, Ziglio S, Batisse J, Dubois N, et al. The HDAC6/APOBEC3G complex regulates HIV-1 infectiveness by inducing Vif autophagic degradation. Retrovirology. 2015 Jun 24;12:53.

294. Yu X, Yu Y, Liu B, Luo K, Kong W, Mao P, et al. Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. Science. 2003 Nov 7;302(5647):1056–60.

295. Sheehy AM, Gaddis NC, Malim MH. The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. Nat Med. 2003 Nov;9(11):1404–7.

296. Santa-Marta M, da Silva FA, Fonseca AM, Goncalves J. HIV-1 Vif can directly inhibit apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G-mediated cytidine deamination by using a single amino acid interaction and without protein degradation. J Biol Chem. 2005 Mar 11;280(10):8765–75.

297. Conticello SG, Harris RS, Neuberger MS. The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. Curr Biol. 2003 Nov 11;13(22):2009–13.

298. Santa-Marta M, Aires da Silva F, Fonseca AM, Rato S, Goncalves J. HIV-1 Vif protein blocks the cytidine deaminase activity of B-cell specific AID in E. coli by a similar mechanism of action. Mol Immunol. 2007 Jan;44(4):583–90.

299. Goncalves J, Santa-Marta M. HIV-1 Vif and APOBEC3G: multiple roads to one goal. Retrovirology. 2004;1:28.

300. Mehle A, Strack B, Ancuta P, Zhang C, McPike M, Gabuzda D. Vif overcomes the innate antiviral activity of APOBEC3G by promoting its degradation in the ubiquitin-proteasome pathway. J Biol Chem. 2004 Feb 27;279(9):7792–8.

301. Smith JL, Pathak VK. Identification of specific determinants of human APOBEC3F, APOBEC3C, and APOBEC3DE and African green monkey APOBEC3F that interact with HIV-1 Vif. J Virol. 2010 Dec;84(24):12599–608.

302. Wittkopp CJ, Adolph MB, Wu LI, Chelico L, Emerman M. A Single Nucleotide Polymorphism in Human APOBEC3C Enhances Restriction of Lentiviruses. PLoS Pathog. 2016 Oct;12(10):e1005865.

303. Wichroski MJ, Robb GB, Rana TM. Human retroviral host restriction factors APOBEC3G and APOBEC3F localize to mRNA processing bodies. PLoS Pathog. 2006 May;2(5):e41.

304. Chiu YL, Greene WC. APOBEC3G: an intracellular centurion. Philos Trans R Soc Lond B Biol Sci. 2009 Mar 12;364(1517):689–703.

305. Chiu YL. Biochemical fractionation and purification of high-molecular-mass APOBEC3G complexes. Methods Mol Biol. 2011;718:185–206.

306. Gallois-Montbrun S, Holmes RK, Swanson CM, Fernandez-Ocana M, Byers HL, Ward MA, et al. Comparison of cellular ribonucleoprotein complexes associated with the APOBEC3F and APOBEC3G antiviral proteins. J Virol. 2008 Jun;82(11):5636–42.

307. Kreisberg JF, Yonemoto W, Greene WC. Endogenous factors enhance HIV infection of tissue naive CD4 T cells by stimulating high molecular mass APOBEC3G complex formation. J Exp Med. 2006 Apr 17;203(4):865–70.

308. Kozak SL, Marin M, Rose KM, Bystrom C, Kabat D. The anti-HIV-1 editing enzyme APOBEC3G binds HIV-1 RNA and messenger RNAs that shuttle between polysomes and stress granules. J Biol Chem. 2006 Sep 29;281(39):29105–19.

309. Monajemi M, Woodworth CF, Benkaroun J, Grant M, Larijani M. Emerging complexities of APOBEC3G action on immunity and viral fitness during HIV infection and treatment. Retrovirology. 2012;9:35.

310. Bélanger K, Langlois M-A. RNA-binding residues in the N-terminus of APOBEC3G influence its DNA sequence specificity and retrovirus restriction efficiency. Virology. 2015 Sep;483:141–8.

311. Chelico L, Prochnow C, Erie DA, Chen XS, Goodman MF. Structural model for deoxycytidine deamination mechanisms of the HIV-1 inactivation enzyme APOBEC3G. J Biol Chem. 2010 May 21;285(21):16195–205.

312. Garbuzynskiy SO, Melnik BS, Lobanov MY, Finkelstein AV, Galzitskaya OV. Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? Proteins. 2005 Jul 1;60(1):139–47.

313. Malinen AM, Turtola M, Parthiban M, Vainonen L, Johnson MS, Belogurov GA. Active site opening and closure control translocation of multisubunit RNA polymerase. Nucleic Acids Res. 2012 Aug;40(15):7442–51.

314. Liu Q, Graeff R, Kriksunov IA, Lam CM, Lee HC, Hao Q. Conformational Closure of the Catalytic Site of Human CD38 Induced by Calcium. Biochemistry. 2008 Dec 30;47(52):13966–73.

315. Strout MP, Schatz DG. Imatinib resistance and progression of CML to blast crisis: somatic hypermutation AIDing the way. Cancer Cell. 2009 Sep 8;16(3):174–6.

316. Klasen M, Spillmann FJ, Marra G, Cejka P, Wabl M. Somatic hypermutation and mismatch repair in non-B cells. Eur J Immunol. 2005 Jul;35(7):2222–9.

317. Greeve J, Philipsen A, Krause K, Klapper W, Heidorn K, Castle BE, et al. Expression of activation-induced cytidine deaminase in human B-cell non-Hodgkin lymphomas. Blood. 2003 May 1;101(9):3574–80.

318. Tsukamoto T, Nakano M, Sato R, Adachi H, Kiyota M, Kawata E, et al. High-risk follicular lymphomas harbour more somatic mutations including those in the AID-motif. Sci Rep. 2017 Oct 25;7(1):14039.

319. Chesi M, Robbiani DF, Sebag M, Chng WJ, Affer M, Tiedemann R, et al. AID-dependent activation of a MYC transgene induces multiple myeloma in a conditional mouse model of post-germinal center malignancies. Cancer Cell. 2008 Feb;13(2):167–80.

320. Patten PE, Chu CC, Albesiano E, Damle RN, Yan XJ, Kim D, et al. IGHV-unmutated and IGHV-mutated chronic lymphocytic leukemia cells produce activation-induced deaminase protein with a full range of biologic functions. Blood. 2012 Dec 6;120(24):4802–11.

321. Tsai AG, Lu H, Raghavan SC, Muschen M, Hsieh CL, Lieber MR. Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. Cell. 2008 Dec 12;135(6):1130–42.

322. Lu Z, Tsai AG, Akasaka T, Ohno H, Jiang Y, Melnick AM, et al. BCL6 breaks occur at different AID sequence motifs in Ig-BCL6 and non-Ig-BCL6 rearrangements. Blood. 2013 May 30;121(22):4551–4.

323. Greisman HA, Lu Z, Tsai AG, Greiner TC, Yi HS, Lieber MR. IgH partner breakpoint sequences provide evidence that AID initiates t(11;14) and t(8;14) chromosomal breaks in mantle cell and Burkitt lymphomas. Blood. 2012 Oct 4;120(14):2864–7.

324. Gordon MS, Kanegai CM, Doerr JR, Wall R. Somatic hypermutation of the B cell receptor genes B29 (Igbeta, CD79b) and mb1 (Igalpha, CD79a). Proc Natl Acad Sci U A. 2003;100(7):4126-31.

325. Pasqualucci L, Compagno M, Houldsworth J, Monti S, Grunn A, Nandula SV, et al. Inactivation of the PRDM1/BLIMP1 gene in diffuse large B cell lymphoma. J Exp Med. 2006 Feb 20;203(2):311–7.

326. Shen HM, Peters A, Baron B, Zhu X, Storb U. Mutation of BCL-6 gene in normal B cells by the process of somatic hypermutation of Ig genes. Science. 1998 Jun 12;280(5370):1750–2.

327. Pasqualucci L, Guglielmino R, Houldsworth J, Mohr J, Aoufouchi S, Polakiewicz R, et al. Expression of the AID protein in normal and neoplastic B cells. Blood. 2004 Nov 15;104(10):3318–25.

328. Dorsett Y, Robbiani DF, Jankovic M, Reina-San-Martin B, Eisenreich TR, Nussenzweig MC. A role for AID in chromosome translocations between c-myc and the IgH variable region. J Exp Med. 2007 Sep 3;204(9):2225–32.

329. Kawamura K, Wada A, Wang JY, Li Q, Ishii A, Tsujimura H, et al. Expression of activation-induced cytidine deaminase is associated with a poor prognosis of diffuse large B cell lymphoma patients treated with CHOP-based chemotherapy. J Cancer Res Clin Oncol. 2016 Jan;142(1):27–36.

330. Kuppers R. Mechanisms of B-cell lymphoma pathogenesis. Nat Rev Cancer. 2005 Apr;5(4):251–62.

331. Arima H, Fujimoto M, Nishikori M, Kitano T, Kishimoto W, Hishizawa M, et al. Prognostic impact of activation-induced cytidine deaminase expression for patients with diffuse large B-cell lymphoma. Leuk Lymphoma. 2017 Dec 18;1–11.

332. Gruber TA, Chang MS, Sposto R, Muschen M. Activation-induced cytidine deaminase accelerates clonal evolution in BCR-ABL1-driven B-cell lineage acute lymphoblastic leukemia. Cancer Res. 2010 Oct 1;70(19):7411–20.

333. Nussenzweig A, Nussenzweig MC. Origin of chromosomal translocations in lymphoid cancer. Cell. 2010 Apr 2;141(1):27–38.

334. Montamat-Sicotte D, Litzler LC, Abreu C, Safavi S, Zahn A, Orthwein A, et al. HSP90 inhibitors decrease AID levels and activity in mice and in human cells. Eur J Immunol. 2015 Aug;45(8):2365–76.

335. Rebhandl S, Geisberger R. AIDing cancer treatment: Reducing AID activity via HSP90 inhibition. Eur J Immunol. 2015 Aug;45(8):2208–11.

336. Methot SP, Di Noia JM. Pharmacological manipulation of AID. Oncotarget. 2015 Sep 29;6(29):26550–1.

337. Tsai CT, Yang PM, Chern TR, Chuang SH, Lin JH, Klemm L, et al. AID downregulation is a novel function of the DNMT inhibitor 5-aza-deoxycytidine. Oncotarget. 2014 Jan 15;5(1):211–23.

338. McDonald G, Medina CO, Pilichowska M, Kearney JF, Shinkura R, Selsing E, et al. Accelerated Systemic Autoimmunity in the Absence of Somatic Hypermutation in 564Igi: A Mouse Model of Systemic Lupus with Knocked-In Heavy and Light Chain Genes. Front Immunol. 2017;8:1094.

339. Hsu HC, Yang P, Wu Q, Wang JH, Job G, Guentert T, et al. Inhibition of the catalytic function of activation-induced cytidine deaminase promotes apoptosis of germinal center B cells in BXD2 mice. Arthritis Rheum. 2011 Jul;63(7):2038–48.

340. Li G, Pone EJ, Tran DC, Patel PJ, Dao L, Xu Z, et al. Iron inhibits activation-induced cytidine deaminase enzymatic activity and modulates immunoglobulin class switch DNA recombination. J Biol Chem. 2012 Jun 15;287(25):21520–9.

341. Shi K, Demir O, Carpenter MA, Wagner J, Kurahashi K, Harris RS, et al. Conformational Switch Regulates the DNA Cytosine Deaminase Activity of Human APOBEC3B. Sci Rep. 2017 Dec 12;7(1):17415.

342. Irwin JJ, Shoichet BK, Mysinger MM, Huang N, Colizzi F, Wassam P, et al. Automated docking screens: a feasibility study. J Med Chem. 2009 Sep 24;52(18):5712–20.

343. Dallakyan S, Olson AJ. Small-molecule library screening by docking with PyRx. Methods Mol Biol. 2015;1263:243–50.

344. Kouno T, Silvas TV, Hilbert BJ, Shandilya SMD, Bohn MF, Kelch BA, et al. Crystal structure of APOBEC3A bound to single-stranded DNA reveals structural basis for cytidine deamination and specificity. Nat Commun. 2017 Apr 28;8:15024.

345. Fang Y, Xiao X, Li SX, Wolfe A, Chen XS. Molecular Interactions of a DNA Modifying Enzyme APOBEC3F Catalytic Domain with a Single-Stranded DNA. J Mol Biol. 2018 Jan 5;430(1):87–101.

346. Pham P, Landolph A, Mendez C, Li N, Goodman MF. A biochemical analysis linking APOBEC3A to disparate HIV-1 restriction and skin cancer. J Biol Chem. 2013 Oct 11;288(41):29294–304.

347. Harjes S, Solomon WC, Li M, Chen KM, Harjes E, Harris RS, et al. Impact of H216 on the DNA binding and catalytic activities of the HIV restriction factor APOBEC3G. J Virol. 2013 Jun;87(12):7008–14.

348. Quinlan EM, King JJ, Amemiya CT, Hsu E, Larijani M. Biochemical Regulatory Features of Activation-Induced Cytidine Deaminase Remain Conserved from Lampreys to Humans. Mol Cell Biol. 2017 Oct 15;37(20).

349. Du H, Brender JR, Zhang J, Zhang Y. Protein structure prediction provides comparable performance to crystallographic structures in docking-based virtual screening. Methods. 2015 Jan;71:77–84.

350. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001 Mar 1;46(1–3):3–26.

351. Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol. 2004 Dec;1(4):337–41.

352. Chaudhuri J, Tian M, Khuong C, Chua K, Pinaud E, Alt FW. Transcription-targeted DNA deamination by the AID antibody diversification enzyme. Nature. 2003 Apr 17;422(6933):726–30.

353. Zheng S, Vuong BQ, Vaidyanathan B, Lin JY, Huang FT, Chaudhuri J. Non-coding RNA Generated following Lariat Debranching Mediates Targeting of AID to DNA. Cell. 2015 May 7;161(4):762–73.

354. Creacy SD, Routh ED, Iwamoto F, Nagamine Y, Akman SA, Vaughn JP. G4 resolvase 1 binds both DNA and RNA tetramolecular quadruplex with high affinity and is the major source of tetramolecular quadruplex G4-DNA and G4-RNA resolving activity in HeLa cell lysates. J Biol Chem. 2008 Dec 12;283(50):34626–34.

355. Vaughn JP, Creacy SD, Routh ED, Joyner-Butt C, Jenkins GS, Pauli S, et al. The DEXH protein product of the DHX36 gene is the major source of tetramolecular quadruplex G4-DNA resolving activity in HeLa cell lysates. J Biol Chem. 2005 Nov 18;280(46):38117–20.

356. Pucella JN, Chaudhuri J. AID Invited to the G4 Summit. Mol Cell. 2017 Aug 3;67(3):355–7.

357. Durandy A, Peron S, Taubenheim N, Fischer A. Activation-induced cytidine deaminase: structure-function relationship as based on the study of mutants. Hum Mutat. 2006 Dec;27(12):1185–91.