A comparative evolutionary approach to enzyme characterization: AID and AID-like enzymes in early-evolved species

By

Emma Quinlan (BSc, BA)

A thesis submitted to the School of Graduate Studies in

partial fulfillment of the requirements for the degree of

PhD of Science in Medicine

Graduate program in Immunology and Infectious disease,

Faculty of Medicine

Memorial University of Newfoundland

June 2022

St. John’s, Newfoundland

Abstract

The immunoglobulin (Ig)-based adaptive immune system (AIS) evolved ~ 500 million years ago in jawed vertebrates, while the jawless vertebrates exhibit their own variable lymphocyte receptor (VLR)-based AIS. An evolutionarily conserved hallmark of the Ig-based AIS is the diversification of Ig-class antibodies through somatic hypermutation and/or class switch recombination of Ig genes, initiated by the DNA-mutating enzyme activation-induced cytidine deaminase (AID). The VLR-based AIS may also involve activities of AID-like enzymes, but their specific activities have remained unclear. Each clade of vertebrate fishes possesses genetically diverse AID and AID-like enzymes, whose biochemical characteristics, and in some cases functionality, were not known. To elucidate structure-function relationships and to biochemically characterize these distinct proteins, I examined AID homologs from the sea and freshwater lampreys, nurse shark, zebrafish, tetraodon, coelacanth, and human. First, I discovered that the nurse shark and coelacanth, both species representative of significant evolutionary junctures of vertebrate speciation, posses an active AID enzyme. Second, I found the highly unusual human AID biochemical properties of lethargic activity and high affinity binding to its single-stranded DNA substrate were conserved across all vertebrates, while other characteristics, like optimal temperature and sequence preference, diverged. Interestingly, while most vertebrates have only one AID gene, the freshwater lampreys were found to have multiple AID-like genes, grouped into cytidine deaminase 1 (CDA1), CDA1-like, and CDA2 genes, based on the CDA1 and CDA2 genes found previously in the sea lamprey. Furthermore, each individual expresses different combinations of the AID-like genes. The CDA1 and CDA1-like enzymes exhibited varying pH sensitivities and enzyme activity levels, similar to the differences between human AID and its APOBEC family member enzymes, thus suggesting unique roles for the CDA1 and CDA1-like

enzymes in the lamprey AIS. Faint activity was detected for the sea lamprey CDA2, similar to ongoing, yet unpublished, efforts by other research groups. This is the first study to biochemically and structurally characterize AID and AID/APOBEC-like enzymes covering the span of fish evolution. This multidimensional approach of multi-species *in silico* structural analysis and *in vitro* biochemical characterization exemplifies a comprehensive method for gaining deep insight into enzyme activity, function, and evolution.

General Summary

The immune system can be divided into two categories: the innate and adaptive immune systems. Innate immunity involves physical barriers and cells that act within seconds to defeat disease-causing pathogens, but is non-specific, so there is danger of pathogens evading this defense. Adaptive immunity provides specific and long-term, albeit slower, defense, and is centered around T and B cells. T and B cell surface receptors undergo rearrangement and mutation during development, such that each new cell contains a unique receptor with the potential to recognize an invading pathogen; if it does, the associated T and B cells are activated and send out signals to kill the invader. At this point, about a week after infection, B cell receptors are further mutated to better recognize the pathogen, thus producing a stronger kill signal. A particular protein, called activation-induced cytidine deaminase (AID), plays a vital role in the secondary mutation of B cell receptors, without which adaptive immunity is not as strong and can result multiple recurring infections. On the other hand, if AID is dysregulated it can mutate DNA in other cells and cause cancer. Therefore, learning about AID is vital to furthering our understanding of adaptive immunity. Adaptive immunity, as described above, is only found in mammals; the majority of animals have innate immunity, with the first hints of mammalian-type adaptive immunity appearing in the lamprey. Studying AID evolution in the lamprey and other fish can provide insight into immunity evolution and could help us find other means of fighting pathogens that escape human immunity. Most protein research focuses on human or mice proteins, using methodology centered around whole-cell or animal study. Methods that focus on the protein outside of the cell or those that use computer technology to analyze the protein's structure are not as valued. This thesis describes the first research to combine multiple methods of studying AID in

non-human and non-cellular contexts, which reveals deep insight into AID and adaptive immune system evolution, and thus argues for a more comprehensive method of protein study.

Acknowledgements

Writing a PhD thesis while 1) your lab, colleagues, and supervisor move to the other side of the country, 2) you get married and move yourself halfway across the country from family and friends, and 3) during the worst global pandemic in a century, is definitely a daunting task, but in no way did I do it alone. I am forever grateful to my supervisor, Dr. Mani Larijani, who took a chance on me (a Bachelor's student with limited lab experience), who instilled in me a passion for experimenting, who taught me sometimes science involves a little "voodoo" i.e. luck, who showed me what makes me unique will make me successful, who recognized and encouraged my teaching and illustration persuasions, and who pushed my lab mates and I to help and support each other. Thank you.

To my Research Assistants, Lesley, Heather, and Krista, thank you for showing me the ropes, for helping me publish my first papers, for answering my 1000 questions, and keeping the lab (my 2nd home) running smoothly. To my lab mates, Cody, Sarah, Erynn, Faezeh, Hala, Atefeh, Justin, Mahdi, Kevin, Brittany, Zoe, Jan, Patricia, and all the others, thank you for the trips to Timmies for coffee, for the troubleshooting advice, for the laughs, and for all your support. To my department, thank you for the great lunch-time discussions in the 1st floor common room, for the fun softball games, for introducing me to academic service, for keeping my lab coat superhero drawing on the lunchroom whiteboard, for the death-defying hikes, the karaoke nights, and the fond memories. A special thank you to Dr. Ann Dorward for her mentorship and support, and for being a role model of mine in academia. Of course, thank you to my committee, Dr. Michael Grant and Dr. Robert Brown, for mentoring me throughout this whole process. Another special thank you to Hala who started as a lab mate, but became my best friend, whom I will always cherish for

her grace and companionship. Thank you to another best friend, Nancy, who knew me from high school and taught me perseverance and the beauty of quiet strength. My high school mentor, Mr. Park, deserves special mention for teaching me a joy of science and who saw what I could become before I did.

I am forever grateful to my parents and brother, who throughout my academic journey supported me in everyway humanly possible. Thank you to my father, who taught me ambition and the value of education, to my mother, who taught me responsibility and hardiness, and to my brother, who taught me about growing and the value of family. I will be forever grateful for your emotional, financial, and instrumental help, and I love you.

Finally, thank you to my best friend, my confidant, my thesis “drill sergeant”, my companion, my debate advisory, my ball and chain, the Ying to my Yang, Jason. Thank you for your patience, your support, for pushing me when I felt I could not do something, and for forcing me to take breaks when I needed it. Thank you for joining me on this journey.

I am also extremely grateful to Memorial University of Newfoundland, which has been my second home since my undergraduate years, for the funding they provided, including travel funds and the William H. and Bertha Baird Memorial Scholarship. My gratitude also goes to the Natural Sciences and Engineering Research Council of Canada for granting me the Postgraduate Scholarship-Doctoral. These funds allowed me to put my total focus into my studies and I feel very fortunate to have been given that privilege.

To my friends and family whom I have not named, know I value your love and support. It has been a long road, with the last year being the weirdest and longest by far. However, I have

enjoyed every moment of my graduate degree and am satisfied in the completion of this chapter in my career. I look forward to the future and all the possibilities it brings.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A1, 2, 3, 4	APOBEC1, 2, 3, 4
AA	amino acid
ACF	APOBEC1 complementation factor
ADCC	antibody-dependent cell-mediated cytotoxicity
AID	activation-induced cytidine deaminase
AIS	adaptive immune system
AMP	adenosine monophosphate
Ap-1	activator protein 1
APC	antigen presenting cell
APOBEC	apolipoprotein B mRNA-editing catalytic component
ADAR1/2/3	adenosine deaminase acting on RNA 1/2/3
ADAT1/2/3	adenosine deaminase acting on tRNA 1/2/3
Alt-EJ	alternative end joining
APS	ammonium persulfate solution
ATP	adenosine triphosphate
BbRAG1L	Amphioxus RAG-like gene 1
BCL6	B-cell lymphoma 6
BCR	B cell receptor
BER	base excision repair
bp	base pair
BSA	bovine serum albumin

CDA1/2	cytidine deaminase like 1/2
CDR	complementarity-determining region
C/EBP	CCAAT-enhancer-binding proteins
CP	connecting peptide
CSR	class switch recombination
CT	C-terminal
CTD	C-terminal domain
CTT	C-terminal tail
dC	deoxycytidine
DMEM	Dulbecco's Modified Eagle Medium
Dr	*Danio rerio*
DSB	double-stranded breaks
DSCAM	Down syndrome cell adhesion molecule
dsDNA	double-stranded deoxyribonucleic acid
dT	deoxythymidine
DTT	dithiothreitol
dU	deoxyuridine
EDTA	ethylenediaminetetraacetic acid
EMSA	electrophoretic mobility shift assay
EMT	epithelial to mesenchymal transition
EtBr	ethidium bromide
FACT	facilitates chromatin transcription
FAS	fatty acid synthesis

GC	gene conversion
Gc	*Ginglymostoma cirratum*
GPI	glycosylphosphatidylinositol
GST	glutathione S-transferase
GVHD	graft versus host disease
HBV	hepatitis B virus
HCV	hepatitis C virus
HEK	human embryonic kidney
HEPES	(4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid)
His	histidine
HIV	human immunodeficiency virus
HMG	high mobility gene
HP	hydrophobic peptide
HP1	heterochromatin protein 1
Hs	*Homo sapiens*
HTLV-1	human T-cell leukemia virus type 1
ID	inhibitor of DNA
Ig	immunoglobulin
INT	intensity
Ip	*Ictalurus punctatus*
IPTG	isopropyl-beta-D-thiogalactopyranoside
KAP1	Krüppel-associated box-associated protein 1

Kd	dissociation constant
kDa	kilodaltons
KSHV	Kaposi’s sarcoma-associated herpesvirus
LB	lysogeny broth
Lc	*Latimeria chalumnae*
Lp	*Lampetra planeri*
LPS	lipopolysaccharide
LRR	leucine-rich repeats
MAC	membrane attack complex
MAMPs	microbial-associated molecular patterns
MBL	mannose binding lectin
MHC	major histocompatibility complex
Mm	*Mus musculus*
MMC	melanomacrophage centre
MML	multiple mini loci
mRNA	messenger ribonucleic acid
MRSA	methicillin-resistant *Staphylococcus aureus*
mya	million years ago
MZ	marginal zone
NAD1/2	novel AID/APOBEC-like deaminases 1/2
NBD	nonamer binding domain
NC	noncanonical
NF-κB	nuclear factor kappa B

NHEJ	non-homologous end joining
NK	natural killer (cell)
NLS	nuclear localization signal
NMR	nuclear magnetic resonance
nt/nts	nucleotide/nucleotides
NT	N-terminal
ORF	open reading frame
PAMPs	pathogen associated molecular patterns
PBMC	peripheral blood mononuclear cells
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PGR	programmed genome rearrangement
pI	isoelectric point
Pm	*Petromyzon marinus*
PMSF	phenylmethylsulphonyl fluoride
PNK	polynucleotide 5’-hydroxyl-kinase
RAG	recombination activating gene products
RMS	root mean square
RNP	ribonucleoprotein
rpm	rotations per minute
RSS	recombination signal sequence
SD	standard deviation

SDS-PAGE	sodium dodecyl sulfate–polyacrylamide gel electrophoresis
SEM	standard errors of the mean
SHM	somatic hypermutation
Smad3/4	mothers against decapentaplegic homolog 3/4
SNAD1/2/3	secreted novel AID/APOBEC-like deaminases 1/2/3
SOX13	sex determining region Y-related high mobility gene 13
SP	signal peptide
SRY	sex determining region Y
ssDNA	single-stranded deoxyribonucleic acid
STAT6	signal transducer and activator of transcription 6
TBE	tris-borate-EDTA
TBST	tris-buffered saline with Tween
TCF1	T cell factor 1
TCR	T cell receptors
TdT	terminal deoxynucleotidyl transferase
TEMED	tetramethylethylenediamine
TFH	follicular T helper
TGF	tumour growth factor
TIR	terminal inverted repeats

TLR	toll-like receptor
Tn	*Tetraodon nigroviridis*
TNFα	tumour necrosis factor alpha
UC	ulcerative colitis
UDG	uracil DNA glycosylase
Unt	untransfected
UTR	untranslated region
V	variable LRR cassette
Ve	end LRR cassette
Vif	viral infectivity factor
VLR	variable lymphocyte receptor
WGD	whole genome duplication
XRC	X-ray crystallography

Appendices

Chapter 1 Introduction

1.1 Overview of innate and adaptive immunity

The immune system is an interconnected system of physical barriers, cells, and cellular pathways that defend against the onslaught of disease-causing pathogens that organisms encounter every day. This system is often divided into two categories: innate and adaptive immunity. The former is defined by a quick (hours to days), unspecific defense against pathogens and includes physical barriers such as the skin and nostril hair; chemical barriers such as stomach acid and mucous; and the extracellular secretion of antimicrobial proteins such as lysozyme and surfactant proteins, which often work to destroy the cell wall of bacteria and other pathogens. Innate immunity also defines quick-acting immune cells such as phagocytic macrophages, which are white blood cells (leukocytes) that recognize ubiquitous pathogen-associated molecular patterns (PAMPs) on extracellular pathogens via their membrane-bound pattern recognition receptors (PRRs) and "consume" them, thus destroying them or rendering them or their toxins inert.[1,2]

Aside from phagocytosis, PRRs can induce other innate immune responses. Toll-like receptors (TLRs), made up of leucine-rich repeats (LRRs), recognize multiple pathogens and induce expression of transcription factors such as nuclear factor (NF)-κB, which triggers the production of 1) antimicrobial peptides; 2) chemokines, which recruit cells (e.g., neutrophils) to the site of infection; and 3) cytokines, such as interleukin (IL)-1, IL-6, and tumor necrosis factor (TNF)-α, which increase vascular permeability, recruit cells, and induce fever. TLRs can also induce the production of interferon (IFN)α and ß, which exert autocrine and paracrine antiviral effects. C-type lectin receptor (CLR) is another PRR which recognizes specific sugar moieties

such as mannose, fucose, and glucans. CLR can induce phagocytosis, as well as trigger activation of nuclear factor kappa B (NF-κB) and activator protein 1 (AP-1), the latter of which induces expression of inflammatory cytokines and IFNß; some CLRs can also trigger anti-inflammatory signals to reduce tissue damage.[1]

Soluble pattern recognition proteins, like their membrane-bound counterparts, recognize PAMPs and initiate various immune responses. Mannose-binding lectin (MBL), C-reactive protein (CRP), and complement proteins bind to extracellular pathogens and target them for destruction via the complement pathway, which destroys the pathogen's cellular membrane, or phagocytosis (opsonization). Both the complement pathway and opsonization can be enhanced by involvement of antibodies, which are a vital part of the adaptive immune system.[1]

The adaptive immune system (AIS) is the slower acting, more specific arm of the immune system, which encompasses the T and B lymphocytes and their activities. T cells facilitate cellular adaptive immunity: their T cell receptor (TCR) recognizes foreign or tumour-associated antigens (proteins or protein pieces recognized by the immune system) presented on major histocompatibility complexes (MHC) I and II, which are expressed on all nucleated cells and professional antigen presenting cells (APC), respectively. Activated T cells that recognize the MHC-antigen complex will release cytokines that either destroy virally-infected cells or recruit and activate other immune cells. B cells recognize free antigen via their B cell receptor (BCR).[1,2] The three terms of BCR, immunoglobulin (Ig), and antibody all refer to the same molecule, with the distinction that BCR includes both the membrane-bound immunoglobulin antigen receptor and its receptor-associated molecules (CD79α/β or Igα/β) responsible for intracellular signalling for activation of B cells once its cognate antigen is encountered.[1]

1.2 Antibody function and structure

Once activated, B cells will transform into plasma cells, secreting their Ig antigen receptor in the form of antibodies that recognize and bind to either extracellular or cell surface antigens.[2] When bound to extracellular antigens, the antibodies "tag" them for phagocytosis and/or elimination from the body (agglutination or precipitation), and/or rend the external pathogen or toxin inert (neutralization).[3] When bound to surface antigens on pathogens or infected cells, the antibodies mark them for antibody-dependent cell-mediated cytotoxicity (ADCC) by natural killer cells (NK cells, a type of lymphocyte), where the NK cells release cytolytic granules that cause the infected or pathogenic cell to lyse or burst.[4] Antibodies bound to surface antigens can also activate complement which are circulating precursor proteins that, when processed by activated proteases, eventually form a membrane attack complex (MAC), a channel through the cellular membrane, causing the cell to lyse.[1,5]

Antibodies are composed of heavy and light Ig protein chains, each of which can be further divided into variable (Fab) and constant (Fc) regions (see **Figure 1.1**). The variable region is the portion of the receptor that directly contacts the antigen: this portion is unique to each antibody, thus the term *variable.* The constant region is consistent amongst all receptors that belong to the same class (also referred to as isotype) of antibody; this region is also called the effector region. When excreted as an antibody, this region contacts complement or cellular receptors, affecting the type of effect the antibodies facilitate.[1,6]

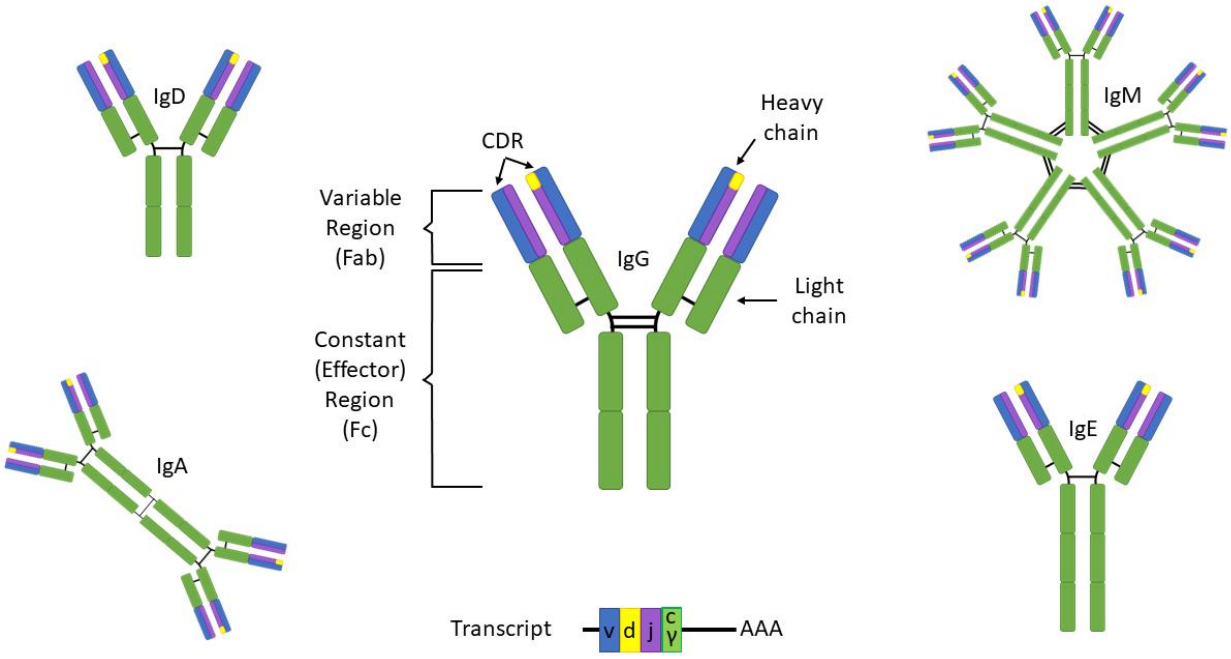


Figure 1.1 Antibody structure and immunoglobulin classes

Adapted from Owen et al., *2013*[1] *as non-commercial user-generated content.*

Centre, the structure of an antibody, class IgG. Centre bottom, the mature messenger RNA (mRNA) transcript that codes for the IgG protein. The different colours in the protein structures correspond to the different coloured genes in the mRNA transcript. Left two structures, IgD (top) and IgA (bottom). Right two structures, IgM (top) and IgE (bottom). V, variable; D, diversity; J, junction; C, constant; CDR, complementary determining region.

There are five different isotypes or classes of antibodies in mammals. The first classes that are produced early in the adaptive immune response are IgM and IgD. IgM can activate complement, leading to the opsonization of pathogens; IgM can also cross mucosal barriers acting in primary pathogen entry sites such as the gut. IgD is secreted in minute amounts compared to IgM and its function remains unknown. To initiate a more specific response to the pathogen, B cells receive differentiating signals from T cells to undergo class-switch recombination (CSR), which switches the antibodies from IgM/IgD to IgA, IgG, or IgE (see **Figure 1.1**). IgA is mainly found in mucosa and secretions, such as in the gastrointestinal and respiratory tracts, and mostly acts by inhibiting binding of pathogens to epithelial cells, preventing their transport from inhaled air or ingested food or water into the body. IgE is involved in defense against parasites, but also mediates type I hypersensitivity (i.e., allergic reactions), facilitating an immune response to otherwise harmless antigens. IgG is the most abundant antibody in serum and the most important effector of pathogen clearance, because it is able to robustly activate complement, agglutinate and opsonize antigens, and initiate ADCC.[1,6]

1.3 V(D)J Recombination: the initiator of the diverse antibody repertoire

The adaptive immune system is capable of producing millions of unique antibodies that are specific to each antigen the AIS encounters, yet there are not millions of unique antibody-encoding genes. Instead, the antibody gene loci are comprised of multiple gene segments that are somatically rearranged and edited during B cell maturation in the bone marrow, resulting in an Ig gene unique to that particular B cell. Both the single heavy and two light (kappa, κ, and lambda, λ) Ig loci are composed of multiple V (variable), J (joining), and C (constant) segments, with the heavy chain locus also containing multiple D (diversity) segments located in between the V and J segments (see **Figure 1.2**).[1,7,8]

Through a process called V(D)J recombination, the intervening gene segments between two target segments are removed via the V(D)J recombinase consisting of the recombination activating gene products 1 (RAG1) and 2 (RAG2), which recognize and cleave the DNA at the border of the recombination signal sequences (RSS) at either end of each segment. The intervening DNA between the two segments is removed and the two coding gene segments joined via DNA repair mechanisms of the non-homologous end joining (NHEJ) DNA repair pathway.[1,7]

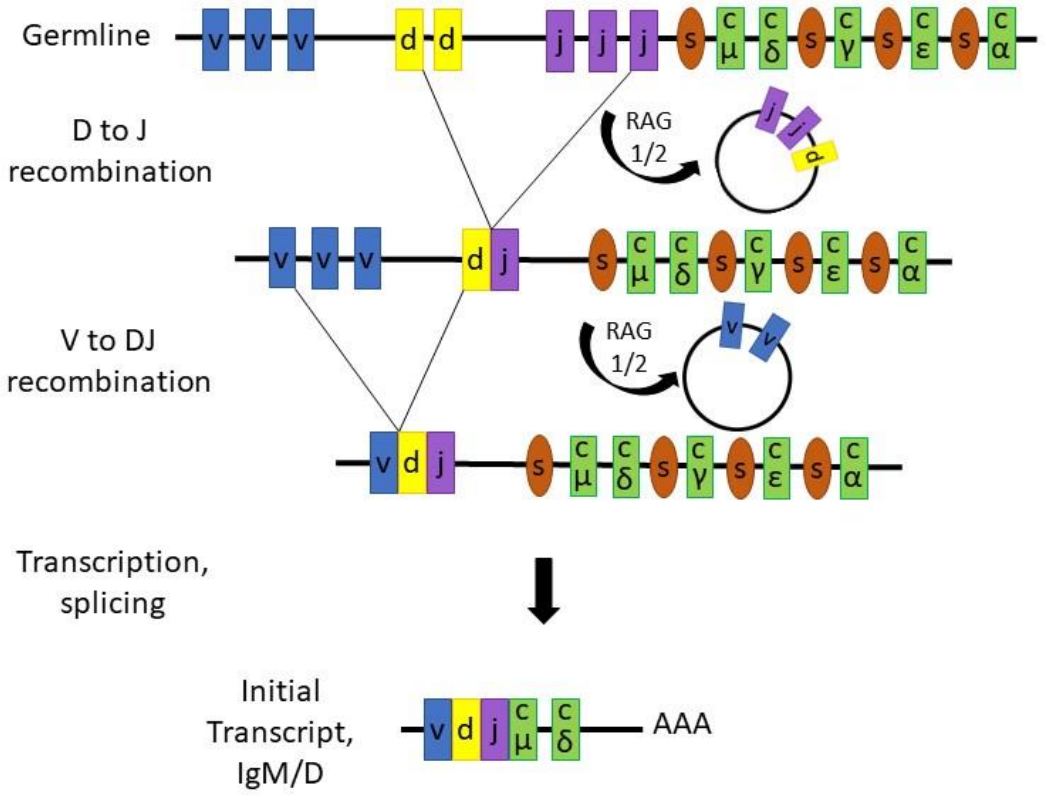


Figure 1.2 Primary antibody diversification by RAG-mediated V(D)J recombination

Adapted with permission from Little et al., *2015.*[8]

Both TCR and antibody gene loci are comprised of variable (V), joining (J) and sometimes diversity (D), segments. Each segment is in close proximity to 1–2 conserved recombination signal sequence (RSS); the V and J are immediately upstream and downstream, respectively, of their single RSS sequence, while D is book-ended by two RSS sequences. Recombination activating gene (RAG) products 1/2 cleave between one RSS and a coding sequence, forming blunt double-stranded breaks (DSB) that are joined by non-homologous end joining (NHEJ) repair mechanisms, resulting in the joining of two segments and removal of the intermediate DNA. Transcription and translation of the V(D)J product results in IgM and/or IgD antibody products (the latter through alternative splicing mechanisms).

The different combinations of various V(D)J segments constitute the first layer of diversification for Ig genes, known as combinatorial diversity, while the second layer of diversification involves junctional diversity. To join the two disparate gene segments, imperfect DNA repair mechanisms that add or remove nucleotides are employed. Deletional cleavage of intervening DNA leaves the remaining segments with hairpin-closed coding joints which can be opened asymmetrically rather than symmetrically, leaving P nucleotides. Terminal deoxynucleotidyl transferase (TdT), a member of the V(D)J recombinase, adds N nucleotides to the coding joints, while exonucleases remove them. All, none, or some of these processes can be utilized at any time, thus further diversifying the coding joints.

Once V(D)J recombination is complete for both light and heavy chains and a structurally-viable combination of light and heavy chains is produced, the developing B cells are further selected for a functional and self-tolerant BCR followed by migration of these immature B cells to the spleen. Here, the B cells are subjected to further self-antigen tolerance testing, which if passed, causes the B cells to mature into follicular (B2) or marginal zone (MZ) B cells.[1,9] The innate immune B cells, B1 B cells, are derived from a separate lineage, originating from the fetal liver.[10,11] Each B cell expresses a unique, recombined Ig receptor as either IgM or IgD to various degrees (depending on the B cell subtype) due to alternative messenger RNA (mRNA) splicing of the two most 5' constant regions in the heavy Ig chain.[1,9]

MZ and B1 B cells can play regulatory roles, supportive roles in T-dependent activation, and active roles in T-independent activation. MZ B cells specifically are vital in T-independent B-cell activation, where ubiquitous antigens on blood-borne pathogens activate these "innate" B cells into becoming short-lived plasma cells, in the absence of T cells. T-dependent B cell activation

occurs mainly to B2 B cells (hereafter simply referred to as "B cells"), which, after maturing in the spleen, migrate to lymph nodes where they come into contact with free, complement-bound, or receptor-bound antigen. The antigen-bound BCRs are then internalized, and the B cell migrates toward helper T cells that express their own, unique V(D)J-recombined T cell receptors. If these helper T cells recognize antigens expressed on lymph node-residing APCs, they become activated and activate nearby naïve B cells via the CD40/CD40L axis. The T helper cells also release stimulating cytokines to the nearby antigen-stimulated B cells, which become fully activated and begin to divide. Some of the daughter activated B cells become antibody-secreting plasma cells, which secrete the initial IgM antibodies of the adaptive immune response. Other daughter cells establish germinal centres in lymph node follicles and undergo secondary antibody diversification processes, resulting in higher affinity antibodies with specific effector responses.[1,9,12]

One such antibody diversification process is somatic hypermutation (SHM). The DNA sequences in the variable regions of both the light and heavy chains of the antigen-specific, activated B cell receptors are mutated. The resulting mutated BCRs are "tested" in the germinal centres: those with higher affinity to the antigen receive survival signals from their interaction with follicular T helper cells (T_{FH}), those with lower affinities are eliminated via T_{FH} neglect. This is also achieved through the decreasing concentration of Ag as the immune response proceeds, thereby selecting for higher and higher affinity BCRs to give those B cells the survival signal. T cells (and other cells) also direct activated B cells to undergo CSR to change the class of Ig from IgM to an Ig class customized to the type of pathogen (see section 1.2 above). Regions of DNA called switch regions are located both up- and downstream of the constant regions on the Ig heavy chain; mutations in these regions result in double-stranded breaks which are repaired in such a way

to bring two switch regions together, excising the intermediate DNA (see **Figure 1.2**). This results in a different constant region being immediately downstream of the rearranged VDJ segment, thus coding for a BCR with the same variable V(D)J antigen-recognizing motif, but with a heavy chain Fc region of a different isotype (IgG, IgE, or IgA).[1,9,12]

Both SHM and CSR are essential mechanisms of antibody diversification. Without SHM, the immune system is relegated to producing low-affinity antibodies that inefficiently bind their corresponding antigen.[13] Without CSR, the only antibody isotype the body can produce are IgM and IgD, both of which are not nearly as efficient as IgG, IgA, or IgE in effecting their respective immune responses. On the other hand, somatic mutation is something that multi-cellular organisms generally try to avoid or correct: errant mutation of essential genes such as those responsible for cellular life cycle and metabolism often leads to cell death, oncogenesis, or other disease.[14] One enzyme at the heart of this delicate balance of purposeful and errant mutation is Activation-Induced Cytidine Deaminase (AID).

1.4 Activation-Induced Cytidine Deaminase

AID is a member of the zinc deaminase apolipoprotein B mRNA-editing catalytic component (APOBEC) family. It is a small, ~ 200 amino acid enzyme mainly expressed in B cells upon recognition of antigen by their recombined, surface-expressed Ig protein and reception of secondary signalling by T cells or TLRs.[15] AID deaminates cytidine to uridine in single-stranded DNA (ssDNA). In AID's catalytic site, a zinc ion is bound to a histidine, two cysteines, and a water molecule. The Zn ion charges the water molecule, which protonates a nearby glutamic acid, which protonates N3 on the substrate. The activated water initiates a nucleophilic attack on C4, resulting in a tetrahedral intermediate. The glutamic acid then deprotonates the hydroxyl group; the oxyanion collapses, creating a double-bonded oxygen. The amino group deprotonates the glutamic acid once more, eliminating ammonia (see **Figure 1.3A**).[16] AID is provided access to ssDNA by transcription of select genes during initiation of secondary antibody diversification in B cells.[1]

During SHM, AID preferentially mutates genetic "hotspots" (WRC motifs in most vertebrates, where W = T or A and R = G or A) at complementarity-determining regions (CDR)s. The resulting uridines are recognized by uracil DNA glycosylase (UDG), which flips out and removes the uracil base. This abasic site could be repaired by error-free mechanisms, however, in activated B cells error-prone mechanisms are employed, such as nonconventional base excision repair (BER), which edits either just the bases or a stretch of one DNA strand using error-prone DNA polymerases. Uracil itself could also be recognized, and a stretch of ssDNA removed and repaired via nonconventional mismatch repair (MMR; see **Figure 1.3**). Mutations that result in

higher affinity antibodies permit the B cell that houses them to proliferate, while unproductive mutations force the B cell to die by neglect.[17]

During CSR, AID mutates cytidines on both strands in AGC-rich switch regions. The uracils are recognized and excised via UDG and MMR machinery; exonucleases remove a stretch of DNA on both strands, leading to double-stranded breaks (DSB). DSB are recognized and repaired by NHEJ and alternative-end joining (Alt-EJ) repair mechanisms, which combine two previously-unconnected sections of the Ig gene, resulting in excision of the intervening DNA and a new constant region immediately upstream of the same V(D)J region (see **Figure 1.2** and **Figure 1.3**).[17]

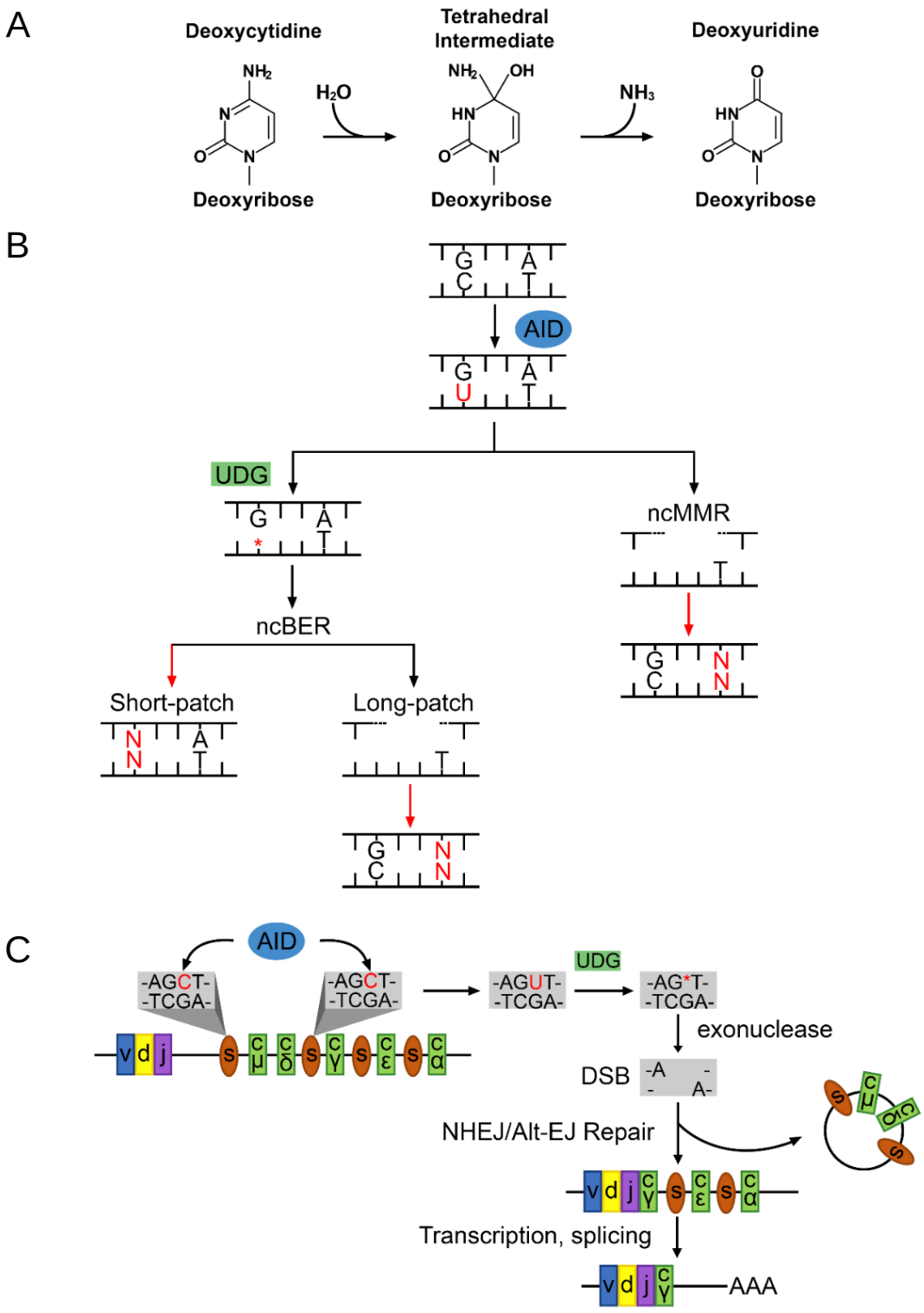


Figure 1.3 Secondary antibody diversification by AID: SHM and CSR

A is adapted from The Mechanism and Catalytic Site Atlas[16] *under a Creative Commons License. B is adapted with permission from Feng* et al., *2020.*[17]

Legend: Figure 1.3

A Deoxycytidine deamination involves the addition of water and elimination of ammonia, with a tetrahedral intermediate. **B** During somatic hypermutation (SHM), cytidines are mutated to uridines via AID deamination; these uracils are either excised by uracil DNA glycosylase (UDG) or recognized by nonconventional mismatch repair (ncMMR) machinery, the latter of which removes and repairs stretches of DNA on either side of the G:U mismatch, resulting in transition and transversion at A:T base pairs. Excised bases by UDG are repaired via short- or long-patch nonconventional base excision repair (ncBER). Short-patch ncBER removes the G opposite the abasic site and synthesizes new DNA, resulting in transition and transversion at G:C base pairs. Long-patch ncBER, like ncMMR, involves the excision of stretches of DNA around the abasic site, which is repaired in a similar fashion. Red arrows denote translesion DNA synthesis, red stars denote abasic sites. **C** During class switch recombination (CSR), AID mutates cytidines at multiple switch (S) regions; each uridine is excised via UDG and exonucleases remove stretches of DNA around the abasic site, resulting in double-stranded breaks (DSB). These DSBs are recognized and repaired via non-homologous end joining (NHEJ) or alternative-end joining (Alt-EJ) repair mechanisms; the intervening DNA between the two mutated S regions is removed, and a new constant region (green C) is joined to the previously recombined VDJ segment. Transcription and splicing mechanisms produce the final antibody transcript.

1.5 Regulation of AID

Due to AID's mutagenic nature, it is thought to be highly regulated by a multitude of mechanisms to protect genomic integrity. First, expression of AID in B cells is induced by T cell-dependent and -independent signalling. The AID locus is enriched in binding sites for transcription factors propagated through various signalling pathways (e.g., NF-kB, signal transducer and activator of transcription 6 (STAT6), CCAAT-enhancer-binding proteins (C/EBP), mothers against decapentaplegic homolog 3/4 (Smad3/4), HoxC4, Pax 5, and E protein).[18–22] Interaction of the T cell CD40 ligand with the B cell CD40 receptor gives a primary signal to B cells for AID expression; secondary signals include cytokines released by T cells activated by CD40:CD40L interaction along with recognition of antigen by the T cell's TCR. T cell independent signals include recognition of microbial-associated molecular patterns (MAMPs, e.g., lipopolysaccharide, LPS) by TLRs on the B cell.[15] Inhibitor of DNA (ID) proteins and the micro RNAs miR-155, miR-181B, and miR-93 may be negative regulators of AID expression after successful CSR and SHM.[18,23–25] There are multiple splice variants of AID that can be inactive or display higher hypermutation activity; such isoforms can be found in both healthy and transformed cells.[26–28] To downregulate AID activity once CSR and SHM is complete, splicing mechanisms to produce the inactive forms of AID may be activated.[26] Research into the higher activity AID splice variants is ongoing, with conflicting implications in tumourigenesis.[1–3]

Once AID is expressed, it is localized mainly in the cytosol with its target DNA guarded in the nucleus. To prevent passive transportation, AID may be retained in the cytoplasm by forming complexes with the proteins Hsp90-DNAja1 and/or eEF1A1.[29–31] AID is thought to be actively transported via recognition of its nuclear localization signal (NLS) by karyopherins and perhaps

trapped in the nucleus when the nuclear envelope reassembles at the end of mitosis, or it may stochastically accumulate in the nucleus in short pulses.[29,32–35] Once AID enters the nucleus, its stay is short-lived; the half life of AID is three times shorter in the nucleus than in the cytosol, as it is readily polyubiquitinated and degraded.[29,36] The C terminal of AID also contains a nuclear export signal (NES), which signals its active removal from the nucleus.[35,37–39]

To initiate SHM and CSR, transportation of AID into the nucleus must coincide with active transcription of immunoglobulin genes, which consequently unravels the genes from their chromatin structure, opens the DNA double-helix, and generates ssDNA for AID mutation.[33,40–44] Histone modification (H3K9me3 and H3K9Ac), Krüppel-associated box-associated protein 1 (KAP1), heterochromatin protein 1 (HP1), and the facilitates chromatin transcription (FACT) complex may tether AID to switch regions of Ig genes and facilitate their transcription.[45–52] Various splicing factors and *cis*-acting elements surrounding Ig genes may also target AID to Ig genes.[53–59] It is hypothesized that stalled transcription machinery may recruit AID, providing access to ssDNA for a prolonged period of time.[52,60–63] The process of transcription causes multiple formations and topologies of DNA (DNA-RNA hybrids, supercoiled, relaxed linear, breathing double-stranded DNA, etc.), all of which are possible targets for AID mutation.[43,64–67] In fact, DNA-RNA hybrids of switch regions may target AID to Ig genes for CSR.[68]

It is likely that a synchronistic cumulation of events, which includes AID expression, transportation into the nucleus, and Ig transcription, must occur to target AID to Ig genes. A further regulatory layer of error-free (non-Ig genes) and error-prone (Ig genes) DNA repair mechanisms is likely present to maintain genomic integrity.[69,70] Despite this orchestra of mechanisms to regulate AID mutation, sometimes a bad note is played and the mutations have catastrophic consequences.

1.6 The role of AID in human disease and cancer

Aside from its role in Ig diversification, the role of AID-mediated mutations in human diseases, specifically cancer, is the most researched aspect of the enzyme. In spite of the aforementioned regulatory mechanisms, AID can sometimes off-target and mutate non-Ig genes. Non-Ig genes may be targeted by AID due to them being co-transcribed alongside Ig genes, like B-cell lymphoma 6 (*BCL6*), a transcriptional repressor required for germinal centre formation, thus providing ssDNA for AID targeting.[71–73] Other genes may have *cis*-acting elements in common with Ig genes, thus attracting AID when they are transcribed.[74] Remarkably, AID has been found to mutate 25% of all genes in B cells, some of which are found in both normal and transformed B cells.[69,75] Simply mutating these genes may result in silent mutations or could render their gene products unusable, thus influencing other cellular mechanisms.[33,75] AID mutations may also be recognized by UDG and MMR proteins resulting in double-stranded breaks which, similar to CSR, are joined with unrelated genes, causing chromosomal translocations.[75] Chromosomal translocations in B cells could create novel oncogenic genes such as combining the *BCR* and *ABL* genes to create a constitutively-active tyrosine kinase.[76,77] Such translocations could also bring the oncogene *MYC* under the control of the regulatory elements of the highly-active Ig genes.[73,78,79] AID-initiated *MYC* translocations have been linked to Burkitt's lymphoma, which may be influenced by Epstein-Barr virus infection of B cells activating AID expression.[80–87] AID mutations have also been linked to chronic lymphocyte leukaemia, non-Hodgkin lymphomas, B cell lymphomas, T cell lymphomas, and multiple myeloma.[80,82,88–97]

Along with the above blood cancers, AID activity has been associated with other types of cancers. AID has been implicated in the epithelial to mesenchymal transition (EMT) for both breast

cancer and renal cell carcinoma; in this state, cells lose their physical connections to their surrounding tissues and potentially metastasize to other areas of the body.[98,99] On the surface of some of these transformed breast cancer cells are stably-expressed IgG proteins, whose transcripts have somatic mutations, further suggesting AID activity.[100,101] The overexpression of AID has also been demonstrated in multiple gastric cancers such as bile duct and liver cancers; AID is often induced by inflammatory signals such as TNFα and mutates several tumour-related genes such as *TP53* and *MYC*.[102–104] Ulcerative colitis (UC) involves inflammation throughout the intestinal tract, which often leads to colorectal cancer; while AID deficiency in mice with UC did increase morbidity (weight loss, pain, and distress), a lack of AID can decrease the risk of inflammation-associated colorectal cancer.[105]

Similar to its APOBEC relative APOBEC3G (A3G), AID may also have antiviral properties. In cirrhotic patients infected with hepatitis B and C virus (HBV/HCV), AID was present in 50% of liver samples.[106] HCV infection induced error-prone DNA polymerases and AID in liver cells *in vitro*, which subsequently resulted in mutations in these cells.[107] AID was co-immunoprecipitated with HBV nucleocapsid proteins, suggesting AID may form a ribonucleoprotein (RNP) complex with these proteins and HBV RNA, similar to how A3G is packaged into human immunodeficiency virus (HIV) virions[108]; once encapsulated, it is hypothesized that AID mutates the HBV genome, playing a protective, innate antiviral role.[109,110] Along with, or instead of, mutating the viral genome, AID may recruit RNA exosome into the HBV-RNP complex, resulting in the degradation of HBV RNA.[111]

AID may also have antiviral effects without being packaged in viral capsids. Rem protein produced by mouse mammary tumour virus may downregulate AID, similar to how the HIV

protein viral infectivity factor (Vif) targets A3G for degradation.[108] However, in this case AID was not encapsulated.[112] In response to Abelson murine leukemia virus infections, AID expression is induced, resulting in AID-associated mutations of the infected cell genome; this genotoxic activity results in checkpoint kinase-1 phosphorylation and subsequent restriction of proliferation of the infected cell. NKG2D ligand was also upregulated in infected cells, marking these cells for destruction by NK cells.[113] Similarly, AID triggers the DNA damage pathway and upregulation of NKG2D ligands in Kaposi's sarcoma-associated herpesvirus (KSHV)-infected cells. AID expression in KSHV-infected cells also reduced lytic transcripts, resulting in a ~ three-fold reduction in virus output, with the produced viruses exhibiting decreased infectivity.[114] As countermoves, KSHV encodes miR-K12-11 and miR-K12-5, which downregulate AID, and has evolved fewer WRC motifs to protect them from AID mutation.[114,115]

Like A3G, AID activity with relation to infection can be a double-edged sword. For example, HIV infection of peripheral blood mononuclear cells (PBMC) results in CD40 ligand expression, which induces AID expression in B cells that express CD40[1,85]; this could explain how HIV directly stimulates B cells, leading to accumulation of molecular lesions in non-Hodgkin's lymphoma.[85] Human T-cell leukemia virus type 1 (HTLV-1)-infected T-cell lines expressed high levels of AID compared to uninfected controls, which may initiate oncogenesis via AID mutation of the infected cell's genome.[116] *Helicobacter pylori*, a bacteria species famous in public discourse for causing common gut problems (e.g., ulcers), has been found to induce aberrant expression of AID via the NF-κB activation pathway, resulting in *TP53* tumour suppressor gene mutations in gastric cells, possibly leading to oncogenesis.[117]

1.7 Evolution of AID

Compared to studies examining the role and regulation of AID in the health-related fields of immunity and cancer, the evolution of AID and the APOBEC family is an understudied topic. All zinc deaminases are thought to originate from tRNA adenosine deaminase (Tad)/ adenosine deaminase acting on tRNA (ADAT2), the latter of which forms a heterodimer with ADAT3 to deaminate adenosine (A) to inosine (I) at position 34 on tRNA. These mutated tRNAs can then recognize multiple mRNA codons, as I pairs with U, C, or A in the wobble (3rd) position on the codon.[55,118,119] Interestingly, ADAT2 may be able to deaminate cytidine in DNA as well,[120] indicating the substrate promiscuity of the AID/APOBEC family (discussed later) may have evolved before the APOBEC family divided into the multiple family members. Other enzymes related to Tad/Adat2, but not to the AID/APOBEC family, include Tad1p/Adat1, which deaminates tRNA at position 37, perhaps in response to stress and/or to stabilize codon-anticodon interations[121–123]; adenosine deaminases acting on pre-mRNA (ADARs 1, 2, and 3) which are involved in post-translational modifications of RNA[119,123]; and cytosine deaminase, cytidine deaminase, and deoxycytidine monophosphate deaminase (dCMP), members of the pyrimidine salvage pathway which recycles nucleotides.[123] These enzymes are found throughout the metazoan phylum.[123] Members of the classical AID/APOBEC family (APOBECs 1, 2, 3, and 4) and their newly discovered sister clades and members are discussed below, in the order in which they likely evolved.

The Secreted Novel AID/APOBEC-like (SNAD) enzymes belong to a sister clade to the classical AID/APOBEC family. SNAD4 was identified in the first animals to diverge from fungi (sponges), whereas SNAD1 appears throughout the vertebrate phylum (SNAD1). SNAD2 and 3,

found only in the ray-finned fishes, are likely the result of whole genome duplication events and/or subsequent expansion of this class. SNAD enzymes are the only AID-like enzymes in multicellular eukaryotes to be secreted, potentially for delivery to virus-infected cells or extracellular parasites; however, their catalytic activity and other biochemical characteristics remain unknown. They may have originated from bacterial toxin proteins.[124]

APOBEC4 (A4), a member of the classical AID/APOBEC family, was likely next to evolve, first appearing in the cnidarians (corals), which diverged after sponges.[124] A4 lacks introns, indicating it may be the result of early retrotranspositional events.[55] A4 is present in the first vertebrates, the jawless fish (agnathans); the lobe-finned fish (sarcopterygians); and tetrapods, but is lost in sharks and often lost in ray-finned fishes (actinopterygians).[124] It is expressed in human testes, but its biological role is uknown.[125] The catalytic activity of A4 also remains unknown; unlike the other members of the classic AID/APOBEC family which are known to deaminate polynucleotides, critical amino acids required for polynucleotide deamination (SWS and F in the middle of the deaminase motif $HxE(x)_nPCxxC$) are missing from A4, indicating it may act on other substrates.[123]

Some of the next AID-like proteins to evolve include cytidine deaminase-like 1 (CDA1); CDA1-like (CDA1L)1, 2, 3, and 4; and CDA2, all of which are found in the agnathans.[124] Lamprey, one of two groups of extant jawless fish, lack many canonical “pillars” of the adaptive immune system, such as RAGs and MHC; however, they do possess antibody-like proteins (variable lymphocyte receptors, VLRs) that are diversified somatically, which led to the discovery of CDA1 and 2 in the sea lamprey, and the CDA1-like enzymes in the freshwater lampreys.[126,127] These enzymes will be discussed in detail in Chapters 3–5.

APOBEC2 (A2), along with AID, are considered the ancestral family members of the classic APOBEC family, being present in most jawed vertebrates tested to date.[124,128,129] A2 may be the result of early retrotranspositional events, which used AID as a scaffold.[55] Human A2 does not appear to be able to edit RNA, DNA, or free cytidine *in vitro*, however, its orthologs may exhibit deaminase activity in *Danio rerio* (zebrafish, ray-finned fish) retina and mouse muscle, but this activity has not yet been confirmed.[123,130–136] A2 also appears to inhibit Transforming Growth Factor (TGF)-β in *Xenopus* (frog, amphibian) and zebrafish.[128,137]

Novel AID/APOBEC-like Deaminases 1 and 2 (NAD1/2), while not being original members of the classical AID/APOBEC family, are closer in sequence to other classical members of the family (A1, A2, A3) compared to A4. NAD1 is found in ray-finned fishes, coelacanth (sarcopterygian), amphibians, lizards, and marsupials; NAD2 is found only in amphibians.[124] Neither NAD has been biochemically characterized and their biological relevance remains unknown.

APOBEC1 (A1) is the founding member of the AID/APOBEC family. It was originally thought to be first evolved in mammals due to an inverted duplication of AID; however, this duplication likely occurred in or before the lungfish, the closest fish ancestor to tetrapods.[55,124,138] A1 deaminates the cytosine at position 6666 on Apolipoprotein B mRNA, creating a premature stop codon at this position, altering ApoB100 to ApoB48, which is essential for secretion of triglyceride-rich chylomicrons.[123,139] It was later discovered that A1 is quite promiscuous, acting on retroviral substrates and ssDNA.[132,140–142] As A1 is among the later AID/APOBEC family members to evolve, it is possible that the RNA-editing capabilities seen in other members of this family is a late-evolved characteristic. On the other hand, due to the progenitors of the

AID/APOBEC family acting on RNA and, in some cases, both RNA and DNA, substrate promiscuity may be an original characteristic of the family that many members are capable of, whose activity has just not yet been fully elucidated.

APOBEC3 (A3) is the last group of AID/APOBEC enzymes to evolve, likely the result of gene duplication events of AID. In rodents, pigs, and cattle, the two AID-type genes fused to form a single gene; in horses, bats, and felines, one of the two genes repeatedly duplicated leading to an expansion of A3 genes. This expansion is greatly pronounced in primates, with seven unique A3 genes (A3A, A3B, A3C, A3DE, A3F, A3G, and A3H).[55] The duplication and expansion of these enzymes are likely a result of an arms race between mammals and viruses, the targets of A3.[55,143,144] The origin of A3 is currently contested: the initiating duplication event was thought to take place in the first placental mammal,[55] however, the sequenced lungfish genome appeared to contain a putative A3C gene.[145] Another group could not find any A3 orthologs in animals that diverged before placental mammals.[124] As this second group delved extensively into AID/APOBEC family orthologs, it is likely the A3C found in the lungfish was a novel APOBEC-like gene.

1.8 AID and its target (Ig) in non-human species

As described above, AID is an early vertebrate enzyme, playing a vital role in adaptive immunity. Adaptive immunity is characterized, among other things, by somatic diversification (SHM and CSR) of Ig genes. Somatic diversification of Ig, and the Ig loci themselves, vary greatly between evolutionarily distant species, while still accomplishing the same goal of increasing affinity for antigen. In this section, the evolution of AID, its target substrate, and its role in antibody diversification will be discussed in the context of model organisms in chronological, evolutionary order.

Pre-vertebrate animals, such as protochordates, lack AID but do appear to have AID-like enzymes, such as the aforementioned SNADs.[124] While also lacking B cell receptors, these animals do have immune receptors belonging to the immunoglobulin superfamily.[146–148] Interestingly, the freshwater snail also experiences somatic diversification of its fibrinogen-related genes, but it remains unknown whether the SNADs initiate this process.[149]

It is believed that a type of proto-AID (or AID ortholog) was present in the first vertebrate ancestor, which then diverged to CDA2 in the lamprey and to AID in the first jawed vertebrates, the shark.[126,150] Similarly, it is hypothesized that the targets of this proto-AID (somatically diversified lymphocyte receptors) diverged into three unique receptors with three unique lymphocyte cell lineages: a secreted form (VLRB in the lamprey and BCR in jawed vertebrates in B cell-like cells) and two membrane-bound receptors (VLRA/C in the lamprey and TCRαβ/γδ in jawed vertebrates in T cell-like cells; discussed further in Chapter 5).[150,151] Due to CDA1/1L genes lacking introns, it has been posited that CDA2 was the original enzyme in all three lamprey

lineages, with the ability to somatically diversify all three VLRs, and that CDA1/1L genes were the result of retrotransposon events after which CDA2 was subsequentially silenced in CDA1/1L$^+$ cell lineages.[150] This idea is supported by the fact that AID from shark initiates somatic hypermutation of both B and T cell receptors[152,153]; SHM of T cell receptors may have been lost in subsequent vertebrate lineages, appearing in limited later-diverged species, such as the Ballan wrasse (ray-finned fish)[154] and in camels.[155] Lamprey CDAs have been relatively understudied after their discovery, with their VLR targets holding the most attention[126]; their biochemical characterization is a main topic of this dissertation, which will be discussed extensively in Chapters 3–6.

The first immunoglobulin loci to evolve are thought to be in the elasmobranchs (sharks and skates) and are organized quite differently from the most-studied mammalian Ig loci (human and mice). Shark Ig loci are organized into multiple mini loci (MML),[156,157] with a mini locus or "cluster" equating to one V region placed next to one or more D regions, followed by one J segment and a single constant region $(V\text{-}DDD\text{-}J\text{-}C)_n$.[158,159] Some MML are rearranged in the germline, while most are rearranged via the RAG recombinase.[156,157] Shark Ig undergo SHM, with long, tandem substitutions unique to these species as a result of AID-initiated mutations.[128,158,160–162] It was originally believed that shark Ig did not undergo CSR[163]; however, the recombined VDJ of one cluster can be "switched" with that of another, leading to a different constant region attached to the recombined VDJ region, possibly initiated by AID.[164,165] Shark have three types of Ig: IgM, present in almost all vertebrates, IgW (may be a counterpart to IgD), and IgNAR, which is unique to sharks, being made up of only heavy chains.[128,148,158,166] At time of writing, shark AID has not been biochemically characterized outside that described in Chapter 3 of this dissertation.

Outside of humans and mice, AID, SHM, and CSR have been the most extensively studied in ray-finned fish. Poikilotherms (cold-blooded animals unable to regulate their temperature) such as ray-finned fish have modest changes in antibody affinity, which was found to be initiated by SHM.[128,159,167–172] This is likely due to inefficiencies caused by a lack of organized germinal centres; instead, ray-finned fish appear to have germinal centre-like clusters of melanomacrophages (MMC) with AID-producing cells in the centre.[169,171,173] Teleost fish (ray- and lobe-finned fish) appear to have Ig loci made up of both MML and translocon-type organizations, the latter of which is how most tetrapod Ig loci are arranged. In ray-finned fish, the V, D, and J segments are arranged as in mammalian Ig loci, with the IgM and IgD constant regions at the 3' end, one after the other. However, the teleost-unique IgZ/T constant region is located further upstream, separated from the IgM and IgD constant regions by D segments (V_n-D_n-J_n-C_Z-D_n-$C_\mu C_\delta$).[128,157,174–176] Lungfish also have IgW and the lungfish-specific IgN and IgQ.[177] However, bony fish Ig loci do not undergo CSR; the IgM and IgD constant regions are "switched" via alternative splicing, while IgZ/T is expressed after alternative V(D)J rearrangement.[174]

The first AID ortholog was identified in the channel catfish.[128,170] All teleost fish AID have a unique insert in the catalytic domain, but the biochemical consequences of this insert are as of yet unknown.[129,159] Catfish and fugu AID appear to have nuclear export and localization domains conserved with other non-mammalian vertebrate domains with expectant results upon their mutation.[38,159,178,179] There are no reports of AID transcriptional regulators outside of mammals, so the regulation of bony fish AID is currently unknown; however, it stands to reason that these enzymes are also highly regulated like their mammalian counterparts, due to the fidelity of the genomes of this animals and the apparent tight regulation of their AID expression.[159,169,180] Despite

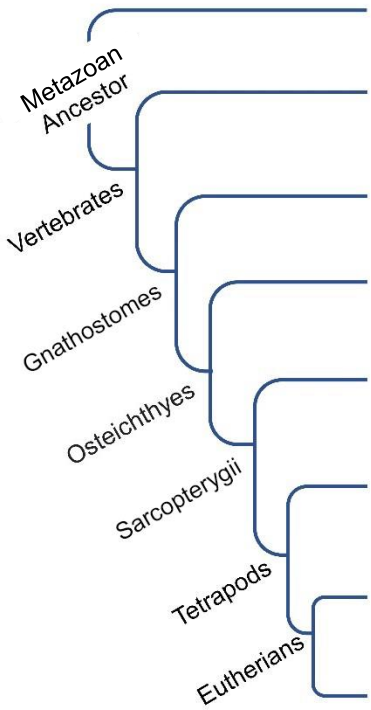
the lack of CSR in bony fish, AID from zebrafish, catfish, and fugu were able to drive CSR in activated B cells from AID-deficient mice, though their relative efficiencies varied.[159,178,179,181] It is likely that CSR developed as a consequence of full translocon organization of the loci along with evolution of regulatory components, and not due to AID mutations.[128,182] Interestingly, lungfish lack AID and *may* not experience SHM, but these findings have yet to be confirmed by other works.[145]

The temperature sensitivity of the bony fish AID enzymes were apparent in the aforementioned CSR experiments and confirmed by subsequent works: the fish AIDs were catalytically more active at cooler temperatures, while human AID was most active around 37 °C.[129,159,179,181,183] Purified zebrafish AID was found to be more catalytically active than catfish and human AID when all were tested at their optimal temperatures.[183] Zebrafish AID was also found to be more efficient than human AID at deaminating methylated cytidines, likely due to its developmental role in this species.[184–186] As seen by these referenced data, most experiments with teleost AID are performed in a cellular context; AID from the relatively recently-diverged pufferfish tetraodon is biochemically characterized outside the cell in Chapter 3 of this thesis.

As the first tetrapods, amphibian Ig loci resemble those of other recently-diverged vertebrates, organized as translocons.[128] Xenopus antibodies undergo SHM and CSR, however, the switch regions in Xenopus are interestingly AT-rich compared to the GC-rich switch regions of mammals, which may affect switching efficiency.[159,187,188] Instead of MMCs or germinal centres, amphibians may have undefined clusters of dendritic cells with both conventional and follicular dendritic cell characteristics,[159,167,168,180,189,190] demonstrating a midpoint between MMC and GC evolution. Xenopus AID has been shown to demonstrate CSR activity,[159,179] and is

expressed in hematopoietic tissues, hinting at a role in ontogeny.[159,180] Neither xenopus AID nor other amphibian AID have been extensively biochemically characterized at time of writing.

Avian (bird) Ig loci, at least the ones that have been sequenced (duck, chicken, and ostrich), are unique among the higher vertebrates in that there is a single functional germline locus (V-D_n-J or V-J) that is recombined via V(D)J recombination; further diversification occurs via gene conversion (similar to how VLRs are recombined, discussed in Chapter 4), initiated by avian AID.[165,191–193] Aside from experiments demonstrating that bovine AID can demethylate DNA via deaminase activity,[194] no other non-human, non-mouse AID has been characterized in the higher vertebrates, and its targets (Ig) and activity (SHM and CSR) in other non-human animals remain unstudied. A summary of AID/APOBEC enzymes, somatic diversification processes, and Ig loci organization can be found in **Figure 1.4**.



Clade	Enzyme	Ig/VLR	Locus	CSR	SHM	VDJ Rec
Sponges	SNAD4	/	/	/	?	/
Agnathans	CDA1 CDA1L1-4 CDA2, A4	VLRA, B, C	Cassettes	Unk	Yes	No, GC
Cartilaginous fish	AID A2	IgM, IgW, IgNAR	MML	NC	Yes	Yes
Ray-finned fish	AID A2 SNAD 1-3. NAD1	IgM, IgD, IgZ/T	MML (IgL) & TL (IgH)	NC	Yes	Yes
Lungfish	AID (?) A1, A2, A3C*, A4 SNAD1	IgM, IgW, IgN, IgQ	MML-TL (IgH), and TL (IgM)	NC (?)	Yes	Yes
Amphibians	AID A2, A4, NAD1, NAD2, SNAD1	IgY, IgM IgX/A, IgF, IgD	TL	Yes	Yes	Yes
Birds	AID A2, A4	IgM, IgY, IgA	TL	Yes	GC	Yes, some GC
Primates	AID A1, A2, A3 A4	IgM, IgD, IgG, IgA, IgE	TL	Yes	Yes	Yes

Figure 1.4 AID/APOBEC family members and Ig loci and diversification in animals

*, A3C in the lungfish is likely another APOBEC ortholog and not a true A3 gene. NC, noncanonical, i.e., the mechanism is different from what is known to take place in most jawed vertebrate species; /, not found in the clade; MML, multiple mini loci; TL, translocon; ?, questionable or unconfirmed; Unk, unknown; GC, gene conversion.

The information gathered thus far is that AID plays a vital role in adaptive immunity with potential for roles in innate immunity and development. While it is highly regulated by multi-layered, redundant systems, its dysregulation does occur and can result in catastrophic genomic consequences. Furthermore, AID from different species likely have unique biochemical characteristics due to the species-specific roles of AID in immunity and development, the varying body conditions (e.g., temperature) of these different species, and its ever-changing substrate target (Ig). However, while its roles in human disease and cancers have been extensively studied, its regulation and biochemical characterization has been less investigated—its evolution and characterization of its orthologs even less so. The relatively few biochemical studies that have been performed have been done in cellular contexts; the study of purified AID enzymes is delegated to only a few investigations.

The remaining sections of this introduction detail a theory as to why that is, based on analysis of the literature and first-hand experience conducting research in this field. The advantages and disadvantages of this weighted study are described, and a new framework is proposed, wherein these different methodologies work together to improve our understanding of not only AID, but any other enzyme.

1.9 Of cells, mice, and men

In the study of human immunology and disease, most of the field's focus is on *in vivo* model organisms, such as mice and rats, or in mammalian cell culture experiments.[195] Mice, being genetically and evolutionarily closely related to humans, provide a time-efficient, cost-effective, and ethically-favourable whole-organism alternative to studying humans directly.[196] Alternatively, cellular studies provide a microscopic view of immune system mechanisms and disease pathogenesis, allowing manipulation of the extra- and intracellular environment. The value in these methods is based on the perception that replicating nature as close as possible produces data that more accurately mirror what occurs in nature.[196–198] As cellular work is outside of the whole organism, it is often qualified as being less accurate to reality than *in vivo* work.[197,199,200] Biochemical studies done outside the confines of the cell (hereafter named *in vitro* work) are thus further devalued, being even further separated from conditions within an organism. Lastly, *in silico* computational renderings of proteins and biological processes are sometimes perceived as purely theoretical or mathematical, lacking biological expertise.[201] Each method is assigned inherent value based on its replicability or "closeness" to reality; however, this linear meritocracy is based on incomplete notions about replicability.

An alternative view to consider is that all forms of experimentation have varying correlations with reality or "accuracy". Although more genetically like humans than most other mammals, mice are insufficient models of various diseases. Mice are 2500 times smaller than humans, resulting in a much higher metabolic rate and shorter life span, which affect metabolic- and age-related diseases, such as Parkinsons'.[197,202] The inconsistency of studying the immune response to cancer in immunologically-deficient mice grafted with human tumours is clear, yet

these types of studies persist.[196] Other human diseases which are not accurately replicated in mice include Lesch-Nyhan syndrome, Lowe syndrome, X-linked andrenoleukodystrophy, and Fabry disease, among others.[203] Though other species prove to provide more accurate representation of human disease, such as the rabbit model of tuberculosis, mice work in tuberculosis and other diseases continues.[204] Attempts are made to "humanize" mice to better represent human health and disease, for example, by injecting immunocompromised mice with human peripheral blood mononuclear cells. Humanized mice have increased our understanding of multiple diseases, such as breast cancer, and development in this methodology has shown great potential in the study of personalized cancer treatment.[205–207] However, "humanized" mice often have short life spans due to graft-versus-host disease (GVHD) where the human immune cells attack the mouse cells; in fact, mouse models to study GVHD are often created using humanization methods.[207,208] Furthermore, regardless of how "human" you make a mouse, there remain stark differences between the mouse model and human response to disease, such as augmented regulatory T cell responses.[207,209] Regardless of being humanized or not, all preclinical animal models have shown to be poor predictors of human responses to drugs, with estimated failure rates as high as 90%.[197]

Considering the lack of replicability of human disease in mice, over 95% of animal studies are performed using this species due to their many advantages, such as a sequenced genome, availability of thousands of strains, and ease of care and genetic manipulation—knowledge acquired through thousands of years of experience working with these animals.[196,210] These advantages supersede the inaccuracies of mice work, something the field accepts often without qualification. Similarly, cellular work is often accepted as a second-best study model, after *in vivo* methodologies, due to it being one step removed from the whole animal. The implementation of

cellular work despite its second-best status is excused due to its inherent favourable qualities, such as the ability to closely observe and control cellular mechanisms which cannot be done *in vivo,* and its ease of care, quicker protocols, and low expense compared to most animal models. However, cellular work has many disadvantages and inaccuracies, such as the process of immortalization, which alters the growth and genetic patterns of the original cell type and an artificial external environment that does not accurately imitate that seen *in vivo*, although advances regarding the latter issue are being made (e.g., organoids and 3D biomimetic cultures).[211–213] Despite these disadvantages, cellular work is a common experimental methodology that is accepted by most scientists in the fields of biomedical sciences. It is thus recognized that the purpose of each method is unique: *in vivo* experiments study whole system interactions, while cell culture studies more specific mechanisms, and both must be performed to gain a better and complete understanding of biological processes.

Even though the field of biomedical sciences enforces a linear meritocracy of *in vivo* being more accurate than cellular work, the advantages, disadvantages, and inaccuracies of both are clear and generally accepted. However, it may be advantageous to expand this analysis and acceptance to *all* forms of biomedical study, including *in vitro* biochemical studies and *in silico* computational modeling. The argument takes two forms: 1) like their *in vivo* and cellular counterparts, both *in vitro* and *in silico* studies have advantages, disadvantages, and relative accuracy, and 2) each has an inherent value based on what they can reveal about the natural world, that other methods cannot. I suggest that instead of rating one method over the other, we appreciate each for what it can do, and use each method when it is most advantageous to gain a more thorough, accurate, and complete understanding of the immune system and its mechanisms. The methods used in this dissertation

are entirely cell-free and are often disregarded at first glance by many scientists who are most familiar with *in vivo* and cellular work regarding human biology; the aim of these remaining introductory sections is to extol the many benefits of *in vitro*, *in silico*, and non-human experimentation, to allow a higher appreciation of the work described in the following chapters.

1.10 The "anthropocentric" view of the biological sciences

It is assumed when discussing an enzyme, receptor, or biological system, that it is the human version being discussed unless otherwise stated, as though the human version is the baseline for all biological, biomedical, and biochemical research. To an extent, this is true, if one subscribes to the philosophy that the purpose of most, if not all, science is for the benefit of humanity: studying fish health preserves food stocks, studying animal toxins supports drug and product discovery, and studying cancers in mice is used to cure cancers in humans, not rodents. However, comparative science is not only beneficial for its direct implications for human health and survival, but also a crucial investigatory tool for examining the bigger question of natural life—how we got here.

Studying enzymes and biological systems of other species has often revealed important aspects of human biology or offered solutions to medical problems. In 2002, Hochachka and colleagues published a hypothesis paper that suggested an explanation for how prostate cancer cells deal with an increased oxygen demand due to malignancy, when normal prostate cells are oxygen-limited in both supply and usage.[214] Oxygen is required for cellular (aerobic) respiration, which produces Adenosine Triphosphate (ATP), the energy fuel in cells. In low-oxygen states, anaerobic metabolism or fermentation occurs instead. Anaerobic metabolism provides energy in low oxygen states, but does not produce end products that serve the other purpose of oxygen in cells, which is to balance metabolically-vital redox reactions.[214] In their previous works, Hochachka *et al.* studied how various animals, including turtles and fish, handle redox balance in low-oxygen situations, such as diving for air-breathing turtles or ice-covered water environments for fish.[215,216] The solution is the fatty acid synthesis (FAS) pathway, which produces oxidized nicotinamide adenine dinucleotide phosphate (NAPDH), which can then be used in other

metabolic pathways to balance redox.[214] This explanation, derived from comparative biology and applied to oncology, has been cited over 100 times[214] and is a great example of how studying biological mechanisms in non-human species can provide or inspire solutions to human problems.

A more direct example of comparative immunology helping the field of medicine can be found in the earthworm. Not only has the earthworm been a consequential model of tissue grafting and self versus non-self immunological study, but studies of its antimicrobial immunity have provided the basis for alternative antibiotics.[217,218] The wide use of antibiotics in the past century has decreased human suffering and death, however, multiple disease-causing bacteria species have fought back, evolving antibiotic resistance. While limiting the use of antibiotics can slow down this process and prevent resistance in the future, the search for alternative antibiotics is quickly becoming the only means of combatting these diseases.[219,220] Such weapons can be found in the most unlikely of places, such as invertebrates, who have thrived for millennia on innate immunity and limited adaptive immunity.[221] In 2003, Tincu and associates published their work on plicatamide, an antimicrobial octapeptide from a tunicate. Plicatamide was shown to kill MRSA (methicillin-resistant *Staphylococcus aureus*), the bacteria responsible for difficult-to-treat infectious commonly attained in hospitals and nursing homes.[222] Their work in the *Journal of Biological Chemistry* was cited in 101 other publications and in two patents for antimicrobial moieties, which can be used to develop alternative antibiotics.[222] Early-diverged animals have a wealth of valuable information to provide regarding human disease and the treatment thereof.

1.11 Evolving our current understanding – the value of comparative immunology

Outside of the anthropocentric view of comparative science, studying our more distantly evolved cousins, such as fishes and invertebrates, has expanded our current understanding of enzymes, mechanisms, and evolution in a variety of fields, including immunology.

Although we understand the basic mechanism of V(D)J recombination, there is still much to be elucidated about RAGs 1 and 2, such as the structure-function relationships at play in their enzymatic activity and their role in the evolution of lymphocyte receptor diversification mechanisms. While studies on human RAGs 1/2 are still providing valuable insight into these enzymes, more and more scientists are looking beyond human and mammalian RAGs to explore earlier-diverged species with non-canonical adaptive and innate immune systems. The RAG 1 and 2 construct is believed to have originated from a transposon that, after a gene duplication event duplicated the transposon, inserted itself into a proto-lymphocyte receptor gene, then exited the gene, leaving behind its terminal inverted repeats (TIRs). These TIRs became the RSSs that are required for V(D)J recombination: each sequence is divided into a heptamer and nonamer, separated by either 12 or 23 base pairs (bp), known as the 12/23 bp rule. Sometime during or after this process, the duplicated transposons, now lacking their TIRs, lost their transposase activity and became recombinases. This theory is based on the organization of the RAG construct (two inward-facing genes transcribed convergently), the similarity of recombinase and transposon activity, and the genetic similarity of RSSs to TIRs.[223,223] The discovery of proto-RAG genes surrounded by TIRs in invertebrate species lacking T and B cells (sea urchin and lancelet), corroborates this theory.[224,225]

In their 2019 *Nature* paper, Schatz and colleagues studied the structure and activity of purified human RAG and amphioxus (lancelet) RAG-like proteins, using *in vitro* structural and biochemical analysis, followed by activity tests in cells and *in vivo*.[223] The authors found that RAG1's coupled cleavage and dependency on the 12/23 rule was dependent on the presence of its N-terminal Nonamer Binding Domain (NBD), while BbRAG1L (the amphioxus RAG-like gene 1) was dependent on its C-terminal tail (CTT) for the same. By creating chimera proteins with either the RAG or BbRAG1L cores combined with their respective or the other's N and C terminals, the authors concluded that it is the presence of the RAG NBD that transfers dependency from the CTT to the N-terminal domain, thus describing an evolutionary transformation of DNA binding domains and sequence preference.[223]

Amino acid (AA) sequence comparisons between the RAG and RAG-like proteins revealed interesting structure-function relationships: by mutating Arg848 in RAG1 to Met (the AA in the equivalent position in BbRAG1L), RAG1 increased transposition activity. The reverse mutation in BbRAG1L decreased its transposition activity. BbRAG2L in amphioxus lacks an acidic hinge; removing the acidic hinge from RAG2 suppressed RAG transposition, but only in cells, thus suggesting RAG2 has a role in the post-cleavage steps of transposition. Taken together, these results suggest specific structural changes that evolved over time to suppress the transposition activity of RAGs on two fronts, thus protecting the integrity of the genome.[223] These revelations could not have been possible without species-comparative work, which highlighted specific amino acids and protein structures that, when further explored, unveiled relevant evolutionary changes and structure-function relationships of an immunologically important enzyme.

Earlier-diverged vertebrates such as bony fish have different classes of immunoglobulin, such as IgM and a fish-exclusive IgZ, yet these antibodies do not undergo CSR. However, their antibodies *are* mutated through AID-initiated SHM. This situation begot the obvious question: why do fish Ig not undergo CSR? In an attempt to answer this question, as described earlier, multiple groups in 2005 and 2006 tested fish AID in a CSR model and found that certain fish AID had the ability to initiate class switching in mouse immunoglobulin loci when introduced into AID-deficient mouse B cells.[178,179,181] These results suggested that classic CSR, which appeared in amphibians at the epoch of tetrapods, was due to the changes in AID target sequences or other cellular processes required for CSR, not AID itself. This hypothesis is further supported by our findings that all AID enzymes from varying species simply require a stretch of single-stranded DNA in order to deaminate cytidine to uridine, albeit with varying substrate preferences.[183,226–228] Recently, Hsu and colleagues discovered that the nurse shark immunoglobulin, which was previously thought to not undergo CSR, was able to switch between IgM and IgW/D, even though their Ig loci are organized into MML and lack tetrapod switch regions.[148,229] This shark-specific type of class switching was attributed to AID activity, the biochemical characteristics of which are described further in Chapter 3.[148,229] These data further support the idea of class switching being a characteristic of DNA repair mechanisms preferring translocations (rather than BER for example) at Ig loci to repair AID-initiated mutations, rather than being a unique attribute of AID. In support of this model, Zarrin and colleagues demonstrated that total replacement of AID by a restriction enzyme that cleaves at switch region DNA is sufficient to mediate CSR in mice.[188]

These comparative studies in non-mammalian species have expanded our current understanding of antibody diversification and AID. Not only is CSR a function of immunoglobulin

loci organization, but there are multiple methods of CSR that are not solely dependent on switch regions. There are two valuable outcomes of these findings: 1) these revelations have further elucidated CSR as a mechanism, which could help better understand dysregulation of human CSR; and 2) better understanding of the role AID serves in various forms of gene diversification has inspired the study of the alternative adaptive immunity in the lamprey, which I expand upon below.

With the technology of DNA sequencing becoming more accurate and available, more and more genomes from different organisms are being sequenced, with norm-shattering discoveries, such as that in the 2004 *Nature* paper by Pancer *et al.,* where the jawless vertebrate lamprey was found to have VLRs in the place of Ig antibodies.[230] Further research from Rogozin and colleagues published in 2007 verified that the adaptive immune system in these early-diverged species does not depend on Ig, nor RAGs, nor MHC, which were previously thought to be absolutely required for adaptive immunity. Although lacking these major AIS factors, AID-like transcripts were found, named CDA1 and CDA2, with CDA1 showing catalytic activity on cytidines *in vitro*.[126] This discovery revolutionized how we think about AIS evolution and the development or evolution of an AIS overall.

Adaptive immunity was previously thought to be restricted to the jawed vertebrates, beginning with the evolution of cartilaginous fishes (sharks), who demonstrate graft rejection (self vs non-self distinction), infection memory, and somatic hypermutation of antibody genes.[162,231,232] The discovery of a totally new and distinct adaptive immune system in the extant jawless vertebrates reveals an older, "parallel" adaptive immunity that can be exploited for human gain, such as unique VLR antibodies for research or health applications.[233] This parallel AIS also contains a vast array of research questions to tease apart for their own sake, such as the apparent

convergent evolution of cell- and antibody-mediated immunity, and the divergent evolution of cytidine deaminases. I, along with my colleagues, began exploring the activity of these AID-like enzymes and their function in this parallel AIS. I recently published data that describe, for the first time, each individual of a species having a unique set of *multiple* AID-like genes, the reason for which is not yet known[127]; these data are discussed in more detail in Chapter 4 of this thesis. This is one example of what can be gleaned from comparative species study and proves that there is still much value and knowledge to be gained from this pursuit.

1.12 Benefits of cellular studies

Comparative immunological investigations often involve identification and analysis of immunologically relevant enzymes and their targets. This can be conducted in two ways: within or without the cell. There are two main benefits of cellular assays over biochemical assays regarding proteins. First, within a cell the protein is translated, folded, and localized in an environment hypothetically resembling *in vivo* conditions, especially if using the same cells that produce said protein, including post-translational modifications and protein-folding chaperones. The second and most relevant benefit of cellular assays is the interaction between the enzyme and its substrate in a more "natural" environment, including intracellular conditions like pH and salt concentration, interplay between localization of enzyme and substrate, and effects of other cellular proteins and processes. The purpose of cellular assays are to replicate *in vivo* conditions as closely as possible to gain an "accurate" view of enzyme activity, while allowing more detailed manipulation and observation of intracellular processes.[199] In fact, there is a phenomenon of "moonlighting" proteins, which act completely differently in the test tube compared to inside the cell due to said protein exhibiting vastly different characteristics within different cellular compartments.[234] Yet, there is still a case to be made for more stripped-down and reductionist *in vitro* assays that focus on the biochemistry of enzyme-substrate interactions.

1.13 *In vitro* study – what an enzyme "does"

There are two veins of exploration when researching enzymes. The first is elucidation of the enzyme's normal activity and its effect in cancer, disease, and homeostasis. The second is the thorough investigation of what the enzyme is capable of. Both veins of research can be informed by biochemical assays. First, cellular assays have the potential to dampen or hide the activity of the enzyme; for example, AID-initiated mutations are often corrected by DNA-repair mechanisms present in the cell. Thus, in order to identify AID activity in the cell, uracil N glycosylase (UNG)- or mismatch repair-deficient cell lines must be used to leave the dC (deoxycytidine) → dU (deoxyuridine) mutations intact.[181] *In vitro* assays that only include the enzyme and substrate eliminate such corrective effects. Furthermore, often cell assays are "reporter" assays, with specific targets in mind for the enzyme that give broad or simple indications of activity; for example, AID activity in a cell grown on rifampicin plates can be indicated if AID mutates the *rpoB* gene, thus conferring rifampicin resistance to the cell, which grows into a colony.[179,235] Evidence of class switching can be elucidated by transforming $AID^{-/-}$ mammalian B cell lines with retroviruses containing AID mRNA, then inducing an immune response via LPS and IL-4; the efficiency of AID-initiated CSR is determined by the relative number of IgG-expressing cells.[181] Presuming a target sequence, such as in the reporter assay example, is quite limiting, especially when examining previously-untested proteins: any negative result could be because a non-optimal substrate was used. By keeping the assay simple, as in *in vitro* methods, one can introduce hundreds of variable substrates to the enzyme in a cost- and time-efficient manner, resulting in a more thorough test of possible activity. Furthermore, cellular reporter assays are often difficult to quantify, providing a simple "yes" or "no" answer with limited ability for gradations. With *in vitro* assays like alkaline

cleavage (described in Chapter 2) we can obtain a gradient of activity which enables calculation of enzyme kinetics, thus allowing objective and delicate comparison between orthologous proteins.

In the search for an enzyme's normal activity *in vivo,* cellular assays can also be detrimental due to their oft-praised complexity. If it is hypothesized that an enzyme requires other co-factors to work, one cannot ascertain what the purpose of said co-factors are until they are stripped from the enzyme; these potential co-factors can then be added on one-by-one to logically determine their order and purpose. It has been proposed that AID has multiple co-factors, but *in vitro* biochemical assays have determined that AID alone is sufficient to initiate DNA mutation both in small, partially single-stranded oligonucleotides, and in single-stranded and breathing double-stranded plasmids.[43,226–228] These potential co-factors could simply be attracted to AID due to charge differentials and not for any enzymatic purpose. Enzymatic assays in cells can also be difficult due to the fragility of the cells themselves. Some enzymes can be toxic to cells if expressed at the wrong time in the life cycle, or over/under-expressed, which can be required when exploring an enzyme's activity in the cell. For example, AID and other APOBECs are DNA-mutating enzymes, which can be genotoxic if over-expressed in cells (see Section 1.6 above).[236]

1.14 *In vitro* study – what an enzyme "can do"

Within the second vein of enzyme investigation, searching for what an enzyme is capable of not only gives us a more accurate view of structure-function relationships, it can also reveal insights into regulatory mechanisms employed by the organism. To determine the relationship between certain structural characteristics and function of the enzyme, enzymologists often create mutations at strategic points in the enzyme to determine what effect those specific amino acids have on catalytic activity or binding substrate. For example, in Drs. King and Larijani's 2015 *Structure* paper, they altered asparagine (Asn) at position 51 in human AID, which is a secondary catalytic residue conserved in all APOBECs, to either a glutamine (Gln) or alanine (Ala). Both mutations resulted in decreased activity in our alkaline cleavage assays, which was hypothesized to be due to the bulkier side-chain of Gln vs Asn, and the lack of amine side chain in Ala.[63] These delicate differences could have been obstructed in a more complicated environment like the cell, where mutations to the protein could have indirect influences on activity, such as cellular localization or co-factor/chaperone binding. These indirect effects are important to explore, but not before nor at the same time as delineating structure-function relationships.

The complexity and fragility of cells also prevent extensive research of optimal conditions for the enzyme, such as temperature, pH, salt concentration, etc., as any deviation from the cells' environmental requirements could be lethal. Deviation from these norms could also alter the cells' life cycle and normal activities, thus negating the main benefit of cellular work, in that it is closer to *in vivo* conditions than biochemical work. Not only do cells not tolerate a large range of conditions such as pH or temperature, but even within the narrow tolerated range, it would be extremely difficult to separate the direct impact of these conditions on the enzyme from that on

the many other cellular processes which could, in turn, impact the outcome of the cell-based enzyme function assays. Testing these optimal conditions, despite them being outside the realm of natural probability, can reveal interesting aspects of enzyme regulation. For example, Chester and associates in 2004 discovered an amazing characteristic of A1: using purified Glutathione S-transferase (GST)-tagged A1 in biochemical *in vitro* assays, Chester *et al*. found that the optimal temperature of A1 alone, without its cofactor APOBEC-1 complementation factor (ACF), is closer to 45 °C rather than the expected human physiological temperature of 37 °C.[237] They explained this is likely due to its substrate RNA spontaneously forming a more A1-favourable conformation at high temperatures, and that ACF works to fold RNA into these conformations at physiological temperature.[237] This raises a lot of interesting questions, such as why did A1 evolve to require RNA to be in an unnatural state? Is it because the RNA-mutating A1 evolved from the DNA-mutating AID,[123,124,129] and this unnatural state of RNA is a compromise between the two different substrates? How did ACF become recruited for this purpose? Aside from these interesting philosophical and evolutionary questions, this work exposed the thermostability of A1 and the true purpose of ACF: aspects that can be manipulated through drug interactions in a disease situation. This important work would have been impossible in a cellular assay and demonstrates the capacity of *in vitro* biochemical assays for illuminating *in vivo* characteristics.

1.15 The relationship of a reductionist *in vitro* enzyme analysis to physiological *in vivo* functions

Previous *in vitro* analyses of AID biochemical characteristics have already yielded important discoveries regarding its *in vivo* activity. In 1992, Rogozin and associates determined that SHM occurs more frequently in cytidines preceded by an A or T (W) in the –2 position and a A or G (R) in the –1 position, denoted as WRC motifs.[238] Subsequent substrate specificity experiments on purified AID proved that this preference was due to characteristics inherent in AID itself, and did not require any co-factors or extraneous conditions.[185,239–241] *In vitro* analysis of the preferred topologies of DNA substrate for purified AID revealed that AID acts on single-stranded DNA,[240–245] leading to the conclusion that AID requires transcription of Ig loci *in vivo* to provide single-stranded targets[33,40–44,243,246–248]; this discovery eventually led to further revelations on AID regulation, such as AID transport into the nucleus needing to be synchronous with transcription of Ig loci (see Section 1.5).

Transcription of DNA produces ssDNA in various conformations: within the transcription bubble as the strands are separated; adjacent to the DNA/RNA hybrid strands as the DNA is transcribed; in unwound, double-stranded DNA as it breathes; and in breathing supercoiled DNA. It is difficult to ascertain the relationship of AID activity to these different DNA topologies in *in vivo* and even cellular assays. *In vitro* analysis of purified AID activity on DNA/RNA hybrids revealed that such substrates are only preferred if they contain switch region sequences, which may be another layer of regulation to prevent AID from targeting simultaneously transcribed non-switch region DNA.[68] Exploration of AID activity without transcription on supercoiled, single-stranded, and double-stranded DNA revealed that AID can deaminate cytidines on both

supercoiled and single-stranded DNA at similar rates, while deamination of double-stranded DNA, while still present, was much reduced.[43] While many initial theories suggested AID required different protein co-factors related to transcription for its activity similar to its A1 relative,[242,246,249] the addition of transcription only modestly enhanced AID activity on supercoiled DNA. Thus AID may only require transcription to create sections of single-stranded DNA, rather than provide any direct protein-protein interaction to increase AID enzymatic activity.[43]

Other *in vitro* assays that have helped elucidate AID's *in vivo* role include that which determined AID's processivity. In 2003, Pham *et al.* found that purified AID incubated with ssDNA mutated only 2% of the available ssDNA, with 10–70 dC→dT (deoxythymidine) transitions per clone. Furthermore, many of these mutations included five consecutive deaminated dCs, which negates the possibility of multiple, immobile AID proteins being responsible. Therefore, AID likely is a processive enzyme, and the mutations seen are due to multiple mutations once bound, rather than multiple binding events.[242] This finding was affirmed in a separate experiment in which AID was incubated with ssDNA, double-stranded DNA (dsDNA), and supercoiled DNA: there was a four to nine-fold difference between the proportion of linear dsDNA and supercoiled dsDNA targeted by AID, but a 10–100-fold difference in AID-mediated mutation frequency, suggesting the preference of AID for supercoiled dsDNA lay in its processivity, not strand promiscuity.[43] This characteristic of AID could have evolved to take advantage of the fleeting availability of unobstructed ssDNA during transcription *in vivo*.

More *in vitro* work explained AID's processivity biochemically. Electrophoretic mobility shift assays (EMSA) of purified AID incubated with serial dilutions of ssDNA revealed that AID

has an unusually high affinity for ssDNA with a long protein-substrate binding half-life, which is likely due to the large net positive surface charge of most AID enzymes.[43,127,229,240]

The ability of *in vitro* experimentation to strip down a biological activity to its bare necessities, without cells or cellular processes, helped us to elucidate that AID likely evolved a high affinity for and long complex half-life with ssDNA to be more processive, in order to take advantage of transient ssDNA during transcription. These *in vitro* findings of AID's biomolecular characteristics have explained AID's *in vivo* activity, which is mutation of Ig genes during B cell development. Thus, *in vitro* experimentation is not only capable of explaining *in vivo* phenomenon but is often the only manner in which we can appropriately study certain processes.

1.16 Computational biology: the new frontier

Aside from AID, the APOBEC3s (A3s) are the most studied APOBEC family member, being implicated both positively and negatively in viral and cancer immunity.[108] Unfortunately, AID and the A3s are difficult to isolate and purify due to their insolubility, highly-charged surfaces, formation of polydisperse oligomers, and extensive non-specific protein-protein/DNA/RNA interactions. Consequently, to increase stability of the protein, all available AID and A3 structures produced using X-ray and Nuclear Magnetic Resonance crystallography (XRC and NMR, respectively) are highly altered, so much so that they lose their ability to perform known functions such as CSR and SHM for human AID.[63,250–252] That being said, the multiple, albeit mutated, structures have afforded a core template that King and associates used to generate computational models of AID based on its amino acid sequence.[63] When they focused on the lowest energy clusters, most were in a closed-pocket state; this discovery explained the low catalytic rate of AID compared to other APOBECs, which, when modelled, were more often found to be in an open-pocket formation.[63] These data were confirmed shortly thereafter with mutated AID and APOBEC X-ray structures, thus showing the validity of the computational-biochemical approach wherein computational models are explored and their closeness to the actual structure is scrutinized through testing of specific model-based hypothesis in functional enzyme assays using mutants or chimeras.[66,253]

Aside from the benefit of analyzing difficult enzyme structures, computational analysis is also more time- and cost-efficient compared to XRC and NMR crystallography, allowing for the creation of a statistically relevant cohort of structures in different conformations, thus enabling the pocket-state observation. Due to a relevant lack of time restraints and resources compared to those

required for XRC and NMR, King *et al.* could also include intentional mutations and chimeras in their modelled cohort to explore structure-function relationships.[63] Due to the mutations necessary for protein purification, this exploration would be inappropriate using XRC and NMR crystallography, as any "purifying" mutation could affect the "intentional" mutations in unexpected ways. Furthermore, even in easy-to-purify proteins, intentionally altering the enzyme could have unintended consequences on purification, thus increasing the difficulty of this exploratory process.

Due to current restrictions of technology and computational analysis, the *in silico* rendering of an enzyme's structure cannot be accurately completed without an *in vitro* crystal structure. On the other hand, especially in the cases of difficult-to-purify enzymes, the mutated crystal structures currently available cannot and should not be solely relied-upon to illustrate the enzyme's true structure. It is through careful analysis of both the *in vitro* and *in silico* structures that we can form the most accurate picture of enzyme structure and function.

1.17 Summary

It is thus apparent that *in vitro* and *in silico* methods have some advantages as well as disadvantages over their cellular and *in vivo* counterparts, and as technology progresses more will be capable using these methodologies, such as *in silico* modeling of an entire cell and all its internal interacting mechanisms.[254–256] Cellular studies are also progressing to be more like *in vivo* organ systems, such as with 3D tissue culture.[200,257–259] Genetic alterations of *in vivo* animal models are also becoming more sophisticated, allowing more fidelity in animal-human comparison.[196] Therefore, this pivotal moment in biomedical sciences where all methods are technologically advancing provides an excellent opportunity to re-evaluate our field's preconceived notions on the inherent value or truthfulness of a methodology, and instead recognize that all methods have advantages and disadvantages. Based on the examples provided in the above sections, my contention is that only by combining data from multiple types of experiments (*in vitro, in silico, in vivo*), in multiple animal models (human, mammals, fishes), from varying points of view (what a protein does *vs.* what it can do), can we fully elucidate the intricacies of life's biological processes.

1.18 Thesis synopsis

The work described in this thesis exemplifies the argument for a comprehensive, multi-dimensional approach to enzyme characterization. In each chapter, I use *in vitro* methods to elucidate the enzymatic activity of AID orthologs from various species, complemented with *in silico* modelling and observations from *in vivo* animal behaviour and immunological responses. **I hypothesize that AID and CDA orthologs from multiple species exhibit unique characteristics, the elucidation of which can offer insights into the immune mechanisms of each species, as well as the evolution of the AIS as a whole**. In Chapter 2, I describe the materials and methods used in each subsequent chapter. In Chapter 3, I describe my work comparing AID orthologs from five vertebrate species representative of key divergence junctures during vertebrate evolution, identifying evolutionarily conserved similarities and adaptive differences between them. In Chapter 4, I characterize multiple AID-like enzymes in the freshwater lampreys and suggest novel activities in this parallel AIS for their expanded AID repertoire. Finally, in Chapter 5, I illustrate my extensive search for CDA2 activity, revealing the possibility that this may be an enzymatically sluggish cytidine deaminase acting on DNA, while also suggesting further methods of exploration.

Chapter 2 Materials and Methods

2.1 AID expression and purification

AID- and CDA-encoding open reading frames (ORFs) from Chapters 3 and 5 and the unspliced CDA variant and APOBEC3G ORFs in Chapter 4 were synthesized in the pBluescript II plasmid (Genscript, US) with EcoRI fragments at both 5' and 3' ends. The remaining CDA variants in Chapter 4 were synthesized into either pGem-T_easy (CDA1L1_2 and CDA1L1_4) or pBluescript II (CDA1L1_1 and CDA1L1_3). The ORF sequences of Pm-CDA1 and Pm-CDA2 (Sea Lamprey, *Petromyzon marinus*, CDA1 and CDA2) were based on the previously published sequences.[126] The Gc-AID (Nurse Shark, *Ginglymostoma cirratum*, AID) ORF was based on a cDNA sequence obtained from a nurse shark spleen cDNA library, amplified from the highly conserved central portion, which was then used as a probe for phage display library. The ORF of Lc-AID (coelacanth, *Latimeria chalumnae*, AID) was assembled from the published coelacanth genome.[260] Human (*Homo sapiens,* Hs), zebrafish (*Danio rerio,* Dr), and tetraodon (*Tetraodon nigroviridis*, Tn) AID and human APOBEC3G ORFs were based on published sequences.[261–264] ORFs of the CDA variants in Chapter 4 were sequenced from the freshwater lamprey, *Lampetra planeri* (Lp) specimens as described.[127]

GST-tagged, purified proteins: The ORFs of Hs-, Lc-, Tn-, and Gc-AID, and Pm-CDA1 were excised from their respective plasmids via EcoRI restriction digestion (1 mg DNA, ~ 10 units of EcoRI (New England Biolabs Inc. [NEB]), 1x reaction buffer (NEB), incubated at 37 °C in excess of 1 hour). The total amount of digested DNA was diluted with 1x loading dye, from the 6x loading dye that was made in-house (60% glycerol [Fisher Scientific], 10 mM Tris-HCl, pH 7.6, 60 mM ethylenediaminetetraacetic acid [EDTA], 0.03% Bromophenol Blue, and 0.03%

Xylene Cyanol FF). The digested DNA was separated on a 1% agarose gel (1% agarose [ThermoFisher], milliQ water, 1% Ethidium Bromide [EtBr, Sigma-Aldrich], 1x Tris-Borate-EDTA [TBE]) and the ORF band (~ 600–700 bp) was visualized under ultraviolet light. The ORF band was excised from the gel using a sterilized scalpel and purified using the Qiaquick Gel Extraction Kit (Qiagen): note, 3 M sodium acetate, pH 5.0, was added to most gel extractions to maintain appropriate pH, the optional addition of 0.5 mL Buffer QG was always performed, the elution buffer was always warmed before addition to the columns (added 30 µL), and the flowthrough was always re-added to the column and spun again.

To generate GST-AID expression constructs, the EcoRI fragments containing ORFs were ligated into pGEX-5X-3 (GE Healthcare, USA) at 5:1 and 10:1 ratios (50 ng of the vector, 400 units of T4 DNA Ligase [NEB], 1x ligase reaction buffer [NEB], 16 °C for 16 hours, then 25 °C for 1 hour). Half of the ligation reaction was transformed into 30–50 µL of XL Blue cells (incubated on ice for 25 min, heat shocked in a 42 °C water bath for 1 min, and recovered on ice for 1 min), which were then resuspended in Lysogeny Broth (LB, Miller formula; 250 mL LB, 37 °C, 1 hour, 225 rotations per minute [rpm]), and plated on LB agar + ampicillin (AMP; 100 µg/mL), grown for 16–24 hours at 37 °C, in a humid environment. Colonies were picked using sterilized toothpicks which were dropped into 5 mL of LB + AMP, grown at 37 °C, 225 rpm, for 16 hours. The DNA was then extracted using the Geneaid High-Speed Plasmid Mini kit: note, the optional addition of 400 µL of W1 Buffer was always performed, the elution buffer was always warmed before addition to the column, 30 µL of elution buffer was added, and the flowthrough was always put back through the column to maximize DNA extraction.

Plasmids that contained correct inserts were verified by restriction analysis and DNA sequencing (Macrogen, Seoul, South Korea). Two independent correct expression constructs were

carried forward for expression of each AID homolog. The plasmids were then transformed into *Escherichia coli* (*E. coli*) DE3 cells. Plasmid production was induced by addition of 1 mM IPTG (isopropyl β-D-thiogalactopyranoside) to a log-phase culture, followed by incubation at 16 °C for 16 hours. Cells were lysed in a cold (~ 4 °C) French pressure cell press (Thermospectronic) twice, the supernatant applied to a column of glutathione-Sepharose high-performance beads (Amersham; located in a 4 °C room), the flowthrough collected and applied to the column again. The beads were washed twice in 20 mL of 1x PBS (phosphate-buffered saline; Sigma), and GST-tagged protein was eluted in elution buffer (0.05 M Tris, pH 8.0, and 0.01 M L-glutathione [reduced]), in 500 µL fractions. Relative amounts of total protein were measured on a ThermoFisher NanoDrop 2000 spectrophotometer, with the highest 2–4 fractions being chosen for dialysis. Eluted protein was dialyzed for 16 hours at 4 °C using SnakeSkin 22 mm Dialysis Tubing (ThermoFisher Scientific) in 1 L of AID storage buffer (20 mM Tris-HCl, pH 7.5, 100 mM NaCl, and 1 mM dithiothreitol [DTT]). The dialysis buffer was then refreshed and the protein dialyzed again for 1–2 hours. The protein was then aliquoted and flash frozen in liquid nitrogen. For each GST-AID clone, 1–4 independent GST-AID preparations were purified.

Purity and relative concentrations of each GST-AID preparation were assessed by SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis, BioRad Protean system). For each SDS-PAGE gel, an 8% acrylamide resolving gel (8% acrylamide/bis solution, 19:1 [BioRad]; 0.375 M Tris-HCl, pH 8.8 [Resolving Gel Buffer, BioRad]; 0.1% SDS [Sigma-Aldrich]; 0.066% APS [ammonium persulfate, Sigma-Aldrich]; and 0.04% TEMED [*N*,*N*,*N*′,*N*′-Tetramethylethylenediamine, Sigma-Aldrich]) was topped with a 4% stacking gel (4% acrylamide/bis acrylamide, 19:1; 0.125 M Tris-HCl, pH 6.8 [Stacking Gel Buffer]; 0.1% SDS; 0.066% APS; and 0.08% TEMED).

To initially test proper protein induction and purity, purified protein in dialysis buffer was mixed with 2x SSB (0.1 M Tris-HCl, pH 6.6 [Stacking Gel Buffer], 4% SDS, 10% Beta-Mercaptoethanol, 20% glycerol, and bromophenol blue [for colour]) at two different concentrations (usually at ratios of 1:3 and 2:2, protein and 2x SSB). IPTG-induced and control uninduced bacterial samples (1 mL), taken immediately before the 16 °C, 16-hour incubation, were pelleted, the supernatant poured off, and resuspended in 50 µL and 100 µL of 2x SSB, respectively. All protein and bacteria samples in 2x SSB were heated in a dry hot bath for ~ 10 minutes at 98 °C, and briefly spun in a microcentrifuge. The 1.0 mm-thick wells were cleaned with 1x running buffer (2.5 mM Tris [Fisher], 19.2 mM glycine [Fisher], and 0.01% SDS, pH 8.3) and 10 µL of each sample was loaded in each well; 5–7 µL of protein ladder (Thermoscientific) was also loaded. The gels were run in 1x running buffer in the BioRad SDS-PAGE apparatus at 50 V (~ 30 minutes) until the 2x SSB dye had passed the stacking gel, then at 90–120 V until the dye line neared the bottom of the resolving gel (~ 1 hour). The gels were stained in coomassie stain (0.1% Coomassie blue R-250 [ThermoFisher], 40% methanol, and 10% glacial acetic acid) followed by destaining (20% methanol and 6% glacial acetic acid). Gels were imaged on an ImageQuant phosphorimager (GE Healthcare) to visualize the purified protein bands and induced and uninduced bacterial cultures, and dried on a vacuum-sealed heater for 4 hours for conservation (see **Figure 2.1A**).

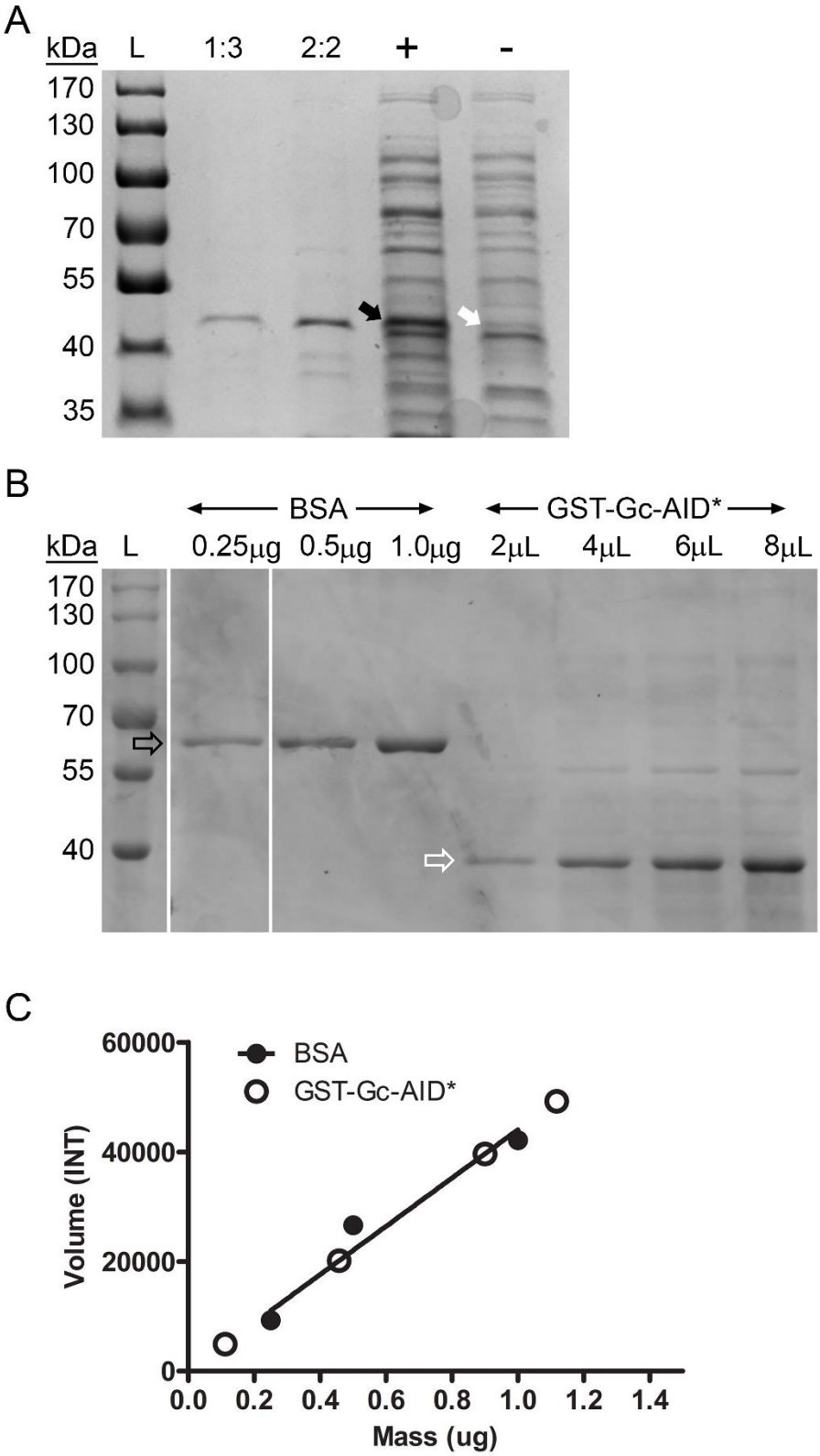


Figure 2.1 SDS-PAGE gels of GST-AID to determine relative purity and concentration of purified proteins

Legend: Figure 2.1

A SDS-PAGE gel of two dilutions of purified GST-Lc-AID (1:3 and 2:2, protein to 2x SSB), which is ~ 51 kDa in size. The black arrow is pointing to the GST-Lc-AID protein band in the induced (+) culture sample, the white arrow points to the same position in the negative control, which lacks a band. L, ladder. **B** SDS-PAGE gel of three known amounts of BSA (Bovine Serum Albumin), 0.25 µg, 0.05 µg, and 1 µg, and four dilutions of the purified dead mutant (*, H56Y, E58G) of GST-Gc-AID (2, 4, 6, and 8 µL of purified protein in 13, 11, 9, and 7 µL of 2x SSB, respectively). BSA is ~ 66 kDa (black outlined arrow) and GST-Gc-AID* is ~ 51 kDa (white outlined arrow, running slightly low on the gel). The ladder (L) was moved to the left side of the gel for clarity; the white spaces denote this edit. **C** Graph of Volume (INT, intensity) versus Mass (µg) of BSA (closed circles) and GST-Gc-AID* (open circles). The straight line of regression is analyzed from the BSA data points. The GST-Gc-AID* data points are interpolated from this line.

If there is a prominent protein band of the calculated size of the protein (GST tag + AID fusion protein) in the induced sample, if it is the same size as the purified protein bands, and if said band is not present or much reduced in the uninduced sample, it can be assumed that appropriate protein induction has occurred. The SDS-PAGE gels also indicate the purity of the purified samples, which is demonstrated by reduced or lack of excess bands compared to the bacteria culture samples. Previous optimization of this protocol has demonstrated that AID purities of 85–95% are achieved; further rinsing of the GST beads (4+ rinses of 1x PBS) resulted in inactive GST-AID, presumably reflecting the requirement of other proteins to keep GST-AID fusion protein soluble and thus able to effectively interact with substrate in solution.[226]

To determine the concentration of purified GST-AID protein, known amounts of Bovine Serum Albumin (BSA) standard were run on the same SDS-PAGE gel with different dilutions of purified GST-AID (**Figure 2.1B**). The gel was imaged and the bands quantified using ImageQuant software (GE Healthcare). The volumes (in "intensity" units) of the BSA bands were plotted against their corresponding masses and a regression line drawn (**Figure 2.1C**). This regression line was used to calculate two to four concentrations of purified GST-AID protein, which were averaged to generate the final concentration.

His-tagged proteins: KpnI and EcoRV restriction sites were added to the 5′ and 3′ ends, respectively, during synthesis of each CDA variant ORF by polymerase chain reaction (PCR), followed by cloning into the multiple cloning site of the eukaryotic expression vector pcDNA3.1/V5-8xHis-TOPO, upstream of the V5 linker and polyhistidine (His) tag (see previous section for cloning procedure). Plasmids with correct inserts were verified by restriction analysis and sequencing. For each CDA variant, two independent correct expression constructs were

chosen for transfection and protein production. Expression vectors containing Hs-, Dr-, and Gc-AID and APOBEC3G ORFs were constructed in the same manner, as positive controls.

To prepare for transfection, HEK (human embryonic kidney) 293T cells were split into 25 x 10 cm tissue culture plates and grown (at 37 °C, 5% CO_2) to 50% confluence in Dulbecco's modified Eagle medium, high glucose (DMEM) containing 1 mM sodium pyruvate, 10% supplemented calf serum, 55 μg/mL streptomycin, and 55 units/mL penicillin (all from Gibco), resulting in each plate containing 5×10^5 HEK 293T cells. The cells in each of 24 plates were transfected with 5 μg of purified ORF-containing plasmid using 2.1 μL/mL of PolyJet In Vitro DNA Transfection Reagent (SignaGen) diluted in serum-free media; one plate was left untransfected as a control for a total of 25. Cells were incubated at 37 °C, 5% CO_2 for ~ 48 hours. One plate of transfected cells was set aside for western blot, the remainder were harvested and pelleted by centrifugation and frozen.

Before purification, the harvested cells were tested for protein production by western blot. First, 1 mL of 2x SSB was added to the one plate of cells that was set aside; the cells were left at room temperature to lyse/detach from the plate, after which they were scraped off and pipetted into a 1.5 mL tube for storage at –20 °C or immediately used for western blot analysis. Of the cells + SSB, 15 μL was added to 10 μL of 2x SSB + dye, vortexed, spun down in a minifuge, boiled for 10 minutes in a dry bath at 98 °C, vortexed, and spun down again. Then, 10 μL was added to an SDS-PAGE gel prepared as previously described, run at 50 V until the dye line passed the stacking gel, and at ~ 90 V until the dye line reached the bottom of the plates (~ 1.5 hours total). Meanwhile, the Bio-Rad Western Blot gel holder cassette, two fiber pads, two blotting filter paper sheets, and a precut Hybond membrane (GE Healthcare) were soaked in 1x Transfer Buffer (0.025 M Tris, 0.192 M glycine, 20% MeOH) at 4 °C. The stacking gel was

removed from the resolving gel, the latter of which was rinsed with then soaked in 1x Transfer Buffer for 15 minutes, shaking, at 22 °C ; the buffer was refreshed three times during this period. The gel holder was then assembled as follows: black side of gel holder cassette, fiber pad, blotting filter paper sheet, gel, membrane, other blotting filter paper sheet, other fiber pad, and white/red side of cassette. The assembled gel holder was placed inside the rinsed, cold BioRad apparatus, which was then filled with cold (4 °C) 1x Transfer Buffer. The apparatus was placed in an ice bucket and run at ~ 70 V for ~ 1.5 hours. The apparatus was then disassembled and the membrane placed in ~ 15–20 mL Blotto (5% skim milk in 1x TBST [Tris-Buffered Saline with Tween; 0.02 M Tris, pH 7.6; 0.14 M NaCl; 0.1% Tween 20]) and incubated for 16 hours at 4 °C.

For the primary antibody incubation, the old Blotto was discarded and replaced with 15 mL of new Blotto plus 3 µL of polyclonal rabbit anti-V5 epitope antibody (Abcam) and incubated with the membrane at ~ 22 °C for two hours, shaking gently. The Blotto + primary antibody was discarded and the membrane washed four times with 15–20 mL 1x TBST, shaking, at 22 °C: each wash was discarded and fresh 1x TBST added each time. The last wash was discarded and 15 mL Blotto plus 3 µL of secondary antibody, goat-anti-rabbit antibody (Abcam), added to the membrane and incubated at room temperature, shaking, for 1 hour. The Blotto + secondary antibody was discarded and the membrane again washed four times, as described above. To image the membrane, the 1x TBST was discarded and 2 mL of each solution in the Millipore Immobilon Western Chemiluminescent kit added to the membrane and incubated for 5 minutes, shaking, then imaged on an ImageQuant Phosphorimager (see **Figure 2.2**). Appropriate protein expression is confirmed when a distinct band can be seen at the expected size (protein + V5 + 8x His tag; black arrow) and no band is seen in the untransfected control cell lysate.

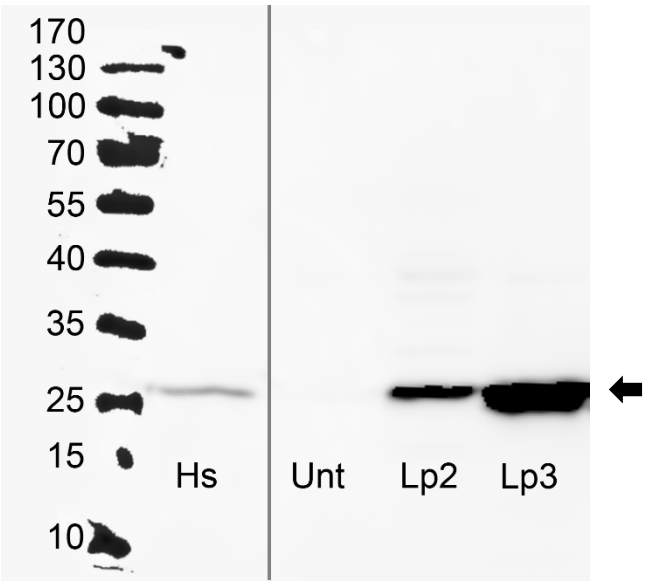


Figure 2.2 Western blot of His-tagged CDA proteins to determine proper induction and production of target protein

Chemiluminescence image of a western blot of His-tagged HsAID (Hs, ~ 26.5 kDa), LpCDA1L1_2 (Lp2, 26.9 kDa), and LpCDA1L1_3 (Lp3, 25.3 kDa) from transfected and harvested HEK 293T cells. Untransfected cell lysate (Unt) was run alongside the transfected cell lysates as a negative control. The ladder is obtained from a digitized image of the same gel, superimposed on top of the western blot image. The gray line represents the post-image cutting and removal of a superfluous lane in the gel. The black arrow indicates the three AID and AID-like protein bands. The grey line denotes the section of the imaged gel that was removed due to it containing irrelevant bands.

Cell pellets containing correctly-produced protein were resuspended in 50 mM phosphate buffer (pH 8.2), 500 mM NaCl, 50 µg/mL RNase A, and 0.2 mM PMSF (phenylmethylsulphonyl fluoride) and pre-lysed by syringing through a 19.5-gauge needle three to ten times until achieving an even consistency. The lysate was further lysed using a cold (4 °C) French press pressure cell (Thermospectronic) twice, then left at room temperature for 30 minutes to allow the RNase A to completely remove any RNA bound to the protein. The lysate was once more put through the cold French press.

For the CDA variants and the AID controls in Chapter 4, the cell lysate was centrifuged to clear cellular debris and the supernatant aliquoted and flash frozen. For the purified His-tagged A3G in Chapter 4, and the Pm-CDA2 and Hs-AID in Chapter 3, the lysed cells were centrifuged, and the supernatant was added to 1 mL of prepared nickel bead slurry (Ni Sepharose High Performance resin, GE Healthcare). The nickel beads were prepared before the French press protocol: 1–2 mL of bead slurry was centrifuged to separate the beads from the storage solution (EtOH), which was decanted. Then, 15 mL of Buffer PB (50 mM phosphate buffer, pH 8.2, + 0.5 M NaCl) was added to the beads, which were centrifuged, the supernatant poured off, and the beads resuspended again in 50 mM Buffer PB at a 1:1 ratio.

The beads were washed twice with Buffer PB plus Imidazole: 15 mL of Buffer PB + 1 mM Imidazole was added to the beads, which were then rocked gently for 20 minutes, centrifuged, and the supernatant removed; the second wash consisted of 15 mL of Buffer PB + 30 mM Imidazole, the beads were again rocked gently for 20 minutes, centrifuged, and the supernatant removed. Next, 750 µL of 500 mM Imidazole in 1x PBS was added to the beads, which were rocked at top speed for 30 minutes to elute the protein; the slurry was then centrifuged and the supernatant collected. Subsequently, 50 µg/mL BSA and 5% glycerol were added to the elute, which was then

dialyzed using SnakeSkin tubing for 16 hours at 4 °C in 1 L of dialysis buffer (0.1 M NaCl, 0.2 M Tris pH 7.5, 5% glycerol, 1 mM DTT). The dialysis buffer was refreshed and the protein dialyzed for another 1–2 hours. The protein was then aliquoted and flash frozen in liquid nitrogen. In Chapter 4, each independent clone (two per ortholog) was expressed once, therefore each ortholog had two individual preparations. In Chapter 5, three individual preparations of one clone of Pm-CDA2-His and two preparations from one clone of each Lp-CDA2 variant were also prepared (discussed further in Chapter 5).

2.2 Preparation of substrates for enzyme assays

Figure 2.3 shows all substrates and target sequences used to test the activity of AID and AID-like enzymes described in this thesis. The partially single-stranded, fully single-stranded, double-stranded, RNA/DNA hybrid, and methylated substrates were prepared as previously described.[183,184,226,227] All oligonucleotides (oligos) were synthesized by Integrated DNA Technologies (IDT); oligos intended for radioactive labelling were further purified via HPLC purification by the manufacturer. Ten units of T4 polynucleotide kinase (PNK; NEB, USA) were used to 5'-label 2.5 pmol of the target strand with 10 µCi of [γ-^{32}P] ATP in 1x PNK Buffer (NEB); the reaction was incubated at 37 °C for 1 hour, then at 65 °C for 10 minutes to deactivate the PNK enzyme. Meanwhile, a mini-Quick spin DNA column (Roche, USA) was prepared by shaking the column for ~ 30 sec to redistribute the slurry to the bottom of the column, then spun at 1000x g for 3 minutes to remove excess buffer from the column.

The total amount of the labelling reaction (10 µL) was diluted with 10 µL of TE buffer (10 mM Tris pH 8.0, 1 mM EDTA pH 8.0) and added to a prepared column, which was then centrifuged at 1000x g for 2 minutes to remove unlabeled oligo. If the substrate was meant to be single-stranded, the flowthrough was diluted to a final concentration of 50 fmol/µL with TE Buffer and stored at –20 °C. If the substrate was to be partially single-stranded or double-stranded, KCl (100 mM final concentration) and 7.5 pmol of the complementary oligo was added to the column flowthrough, which was then diluted to 50 fmol/µL with TE Buffer. This "annealing" reaction was then incubated in the annealing program: 96 °C for 5 minutes, the temperature was then decreased by 1 °C every 30 sec until it reached 4 °C. These prepared oligos were also stored at –20 °C.

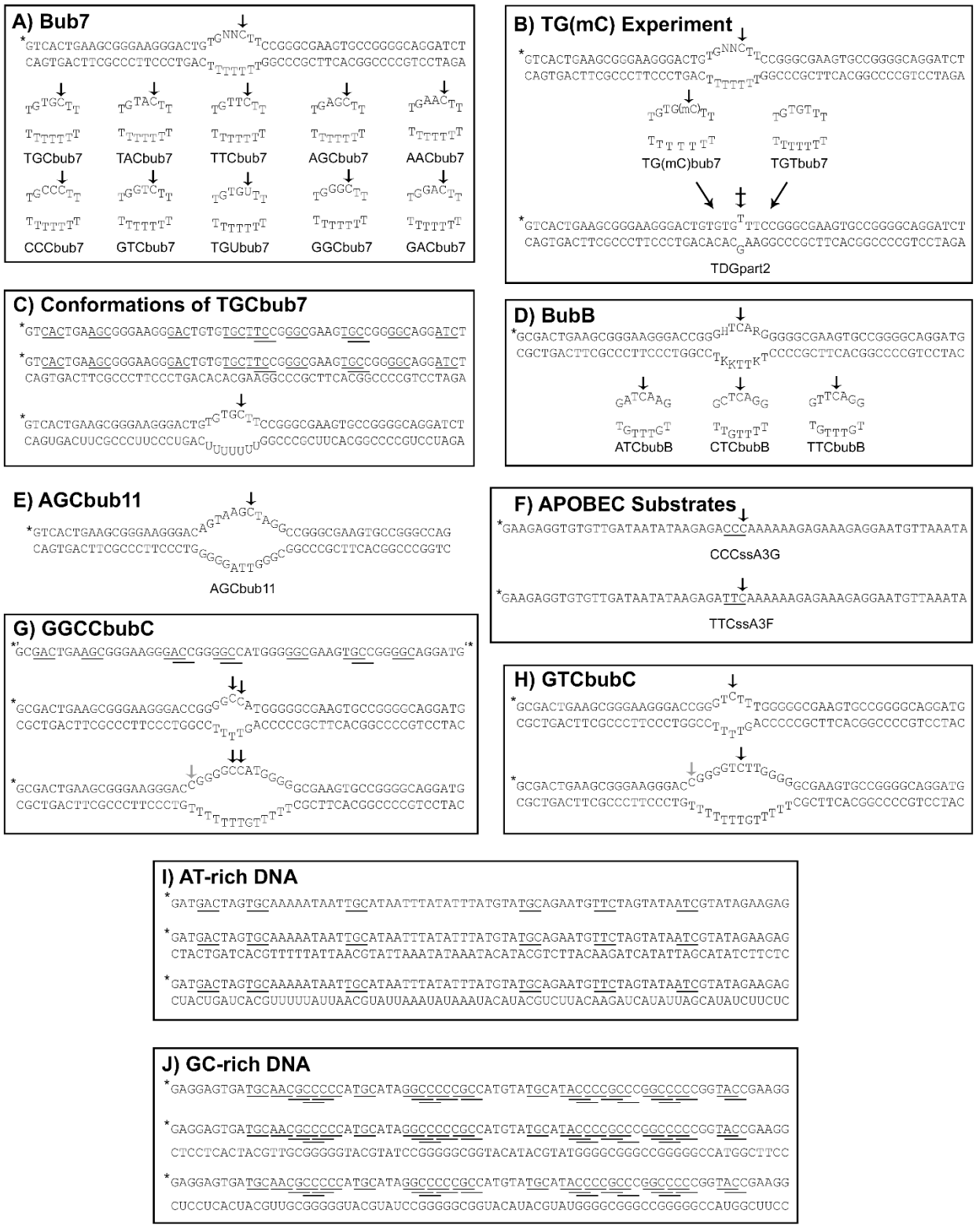


Figure 2.3 Radioactively labelled DNA substrates used in cytidine deaminase activity assays

Legend: Figure 2.3

Target cytidines (C), i.e., the only cytidine theoretically available for deamination, are indicated by black arrows, and NNC motifs by a black underline. Groups of similar substrates are enclosed in black boxes. Five prime radioactive labelling is indicated by a star. **A) Bub7**- partially double-stranded (ds) DNA oligonucleotides (oligos) with a single-stranded (ss) bubble 7 nucleotides (nts) in size. All oligos in this group have the same double-stranded arms, differing only in the NNC motif in the ss bubble. All NNC motifs used are shown. **B) TG(mC) Experiment**- both substrates in this group have the same arm sequences as those in the Bub7 group. The black arrow indicates the target methylated C (mC) in the TG(mC)bub7 substrate, while its partner (TGTbub7) lacks any cytidine in the bubble. Deamination of the methylated C results in a thymine; when the radioactively labelled top strand is reannealed to the TDGpart2 bottom strand, the newly-formed thymine is thus mismatched with the guanine in the opposite strand (TDGpart2). This mis-matched thymine is the target of Thymine-DNA Glycosylase (TDG), denoted by the black crossed arrow. **C) Conformations of TGCbub7-** ssDNA (top), dsDNA (middle), and DNA/RNA hybrid versions of TGCbub7. **D) BubB**- the three substrates in this group have the same DNA sequence in their ds arms but differ in their ss bubble sequences. **E) AGCbub11-** a unique substrate with a large (11 nt) ss bubble with a target AGC motif. **F) APOBEC substrates**- these two ss substrates have the optimal NNC motifs for A3G (CCCssA3G) and A3F (TTCssA3F). The substrates in **G)** and **H)** have the same sequences save the middle 5 nts; **G)** has a ss oligo (top), while both **G)** and **H)** have partially-ss 5 nt (middle) and 13 nt (bottom) bubble substrate versions. The gray arrow indicates a ss C that is unlikely to be targeted. **I)** and **J)** both contain ssDNA (top), dsDNA (middle), and DNA/RNA hybrid versions of **AT-rich** or **GC-rich DNA**, respectively.

2.3 Alkaline cleavage deamination assay

To measure cytidine deaminase activity, I used the standard alkaline cleavage assays for deamination, as previously described.[63,127,183–185,226,227,239] This assay was developed to be high throughput, extremely sensitive to a variety of variables (e.g., pH, temperature), and able to provide accurate comparative enzyme kinetic information. **Figure 2.4** describes the chemical mechanism of this assay, which is pervasive throughout this thesis. Briefly, 5–10 µL of purified protein or lysate (~ 1 µg of the former, concentration of the latter was unknown) was incubated with 3–100 fmol radioactively-labelled substrate in 100 mM phosphate buffer pH 5.6–8.0 or 100 mM citrate-phosphate buffer pH 3.9–7.1 (final volume 10 µL or 20 µL) at 4–40 °C, then heat-inactivated at 85 °C for 20–25 min.

The uracil resulting from cytidine deamination was excised by incubating the deamination reaction with 0.25–0.5 units of UDG (NEB, USA) in 1x UDG Buffer at 37 °C for 0.5–1 hour. NaOH was added at a final concentration of 200 mM and the total solution was heated at 98 °C to induce cleavage at the alkali-labile abasic site of the uracil and denaturation of the DNA, after which 10 µL of loading dye (95% formamide, 0.25% Bromophenol Blue) was added and the solution heated again at 98 °C for 10 minutes to ensure complete denaturation of the substrate and product (**Figure 2.4**, left).

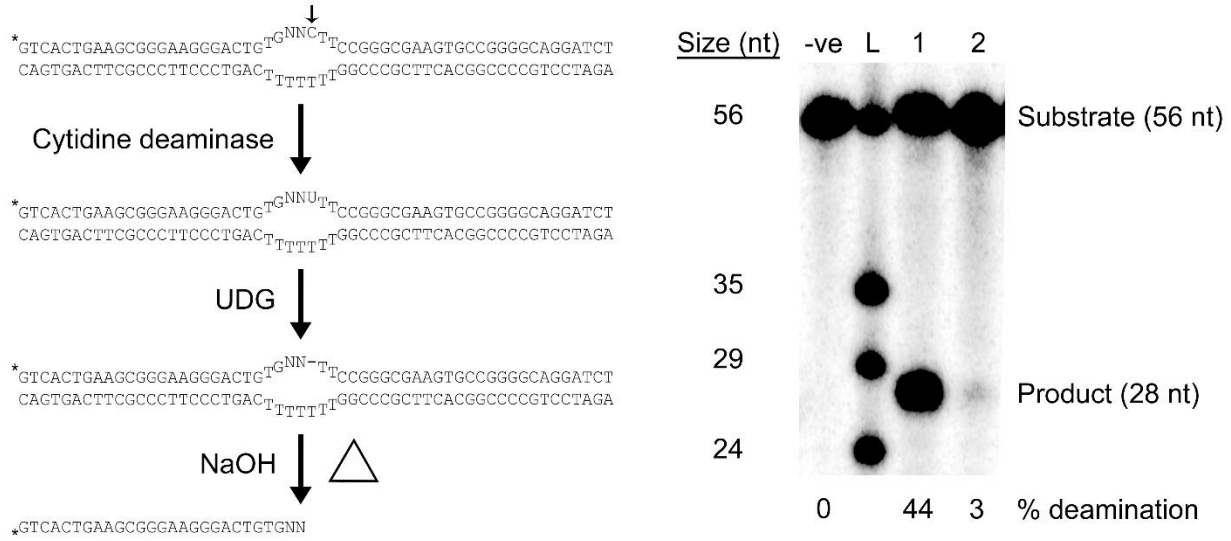


Figure 2.4 Alkaline cleavage of a radioactively labelled bubble substrate

Representative alkaline cleavage experiment using a Bub7 bubble substrate (**Figure 2.3A**). Left, incubation of the substrate with the chosen cytidine deaminase results in the target cytidine (small black arrow) being deaminated into a uridine. Incubation with UDG removes the uridine, creating an abasic site (-). Application of NaOH and heat (triangle) cleaves the substrate at the abasic site and denatures the double-stranded arms, resulting in a product that is much smaller than the un-deaminated substrate, which can be separated from each other in an acrylamide gel (right). Densitometry analysis comparing the volumes of the product and substrate bands (indicated to the right of the gel) from two different experimental incubations (1, 2) produces % deamination values (bottom of gel), which correlate to cytidine deaminase activity. L, ladder; -ve, negative (incubation of phosphate buffer and substrate, lacking protein.

Negative controls consisting of labelled oligo and buffer and/or labelled oligo, buffer, and untransfected eukaryotic cell lysate were carried out alongside the deamination reactions and treated in the same manner. Electrophoresis was performed on a 14% denaturing acrylamide gel (1x TBE, 25% formamide, 14% acrylamide:bisacrylamide, 45% urea), which was pre-run at 100 V for 0.5–2 hours. The wells of the gel were flushed with the running buffer (1x TBE) and half the volume of the final deamination reaction (~ 15 µL) was loaded into each well. A ladder consisting of 50 fmol (total) of 9 x ^{32}P-labelled oligos of various sizes (56, 35, 29, 24, 2, 18, 15, 12, and 8 nucleotides [nts]) was diluted with loading dye to a total volume of 10 µL, heated at 98 °C for 10 min, and loaded in the gel alongside the experimental reactions. The gels were exposed to a Kodak Storage Phosphor Screen GP (Bio-Rad) for 5–16 hours and visualized using a PhosphorImager on Quantity One software (Bio-Rad, USA; **Figure 2.4**, right).

2.4 Enzyme activity assays to measure thermosensitivity

Chapter 3: 50 fmol TGCbub7 (**Figure 2.3A**) was incubated with 0.5–1 µg of AID in 100 mM activity buffer (final volume 10 µL) for 0.5–5 hours at temperatures ranging from 4–40 °C. Completed reactions were immediately incubated at 85 °C for 25 minutes to ensure that the AID or AID-like enzymes were heat inactivated. Reactions were then subjected to UDG and NaOH treatment as described above. Two to six unique (i.e., AID incubation procedures performed on different days) experiments were performed for each of the two to four purified preparations of each AID ortholog. For each experiment, unique or parallel (i.e., performed on the same day), % deamination was plotted against temperature to generate a curve, with the peak revealing the optimal temperature of AID in each experiment. Optimal temperatures from six to twelve experiments were then averaged for each AID ortholog to arrive at its optimal temperature.

Chapter 4: 25–50 fmol of TGCbub7 or TTCbubB (**Figure 2.3A, D**, respectively) was incubated with lysates of His-tagged Hs-, Gc-, Dr-AID, and the CDA variants, respectively, at temperatures ranging from 4–37 °C in 100 mM activity buffer of the following pH's: CDA1L1_1 spliced and unspliced (6.18–6.40), CDA1L1_2 (6.4), CDA1L1_3 spliced and unspliced (6.20–6.40), CDA1L1_4 (7.1), Hs-AID (7.1–7.72), Dr-AID (6.7–7.1), and Gc-AID (6.7–7.1). The enzymes and substrates were incubated for the following times, followed by heat inactivation at 85 °C for 20 min: Hs-AID, Dr-AID, Gc-AID, and CDA1L1_3 spliced and unspliced (4–5 hours), CDA1L1_2 (16 hours), CDA1L1_4 (16–24 hours), and CDA1L1_1 spliced and unspliced (16–30 hours). Reactions were then subjected to UDG and NaOH treatment as described above. Two to four unique experiments were performed on each of two preparations. For each experiment, % maximum deamination was plotted against temperature to generate a curve, with the optimal

temperature at the peak. The optimal temperatures from three to eight experiments were then averaged for each AID and CDA to arrive at their respective optimal temperatures.

Chapter 5: 50 fmol of TGCbub7 or GTCbub7 (**Figure 2.3A**) was incubated with one to two preparations of purified, His-tagged Pm-CDA2 and Hs-AID for 20 or 0.5 hours, respectively, at 12–42 °C, followed by heat inactivation at 85 °C for 20 min. Reactions were then subjected to UDG and NaOH treatment as described above.

2.5 Enzyme activity assays to measure substrate sequence specificity

Chapter 3: To determine substrate sequence specificity, each AID ortholog was incubated with 50 fmol of ^{32}P-labelled bubble substrates with the following target sequence: TGC, TAC, AGC, GGC, GAC, or GTC (**Figure 2.3A**) at each AID ortholog's optimal temperature for 16 or 24 hours. Reactions were heat inactivated at 85 °C for 20–25 min. Reactions were then subjected to UDG and NaOH treatment, as described above. Three unique experiments were performed on each of two to three preparations (at least one of each clone) of each AID ortholog. Relative % deamination was used to normalize for inter-preparation variations in overall activity levels, to enable comparison between different purifications of AID.

Chapter 4: Lysate His-tagged CDA1L1 homologs (two preparations of one clone each) were incubated with 50 fmol of ^{32}P-labelled bubble substrate (**Figure 2.3A, D, E**) at 16 °C and 22 °C in 100 mM phosphate buffer pH 7.1 for 16 hours, and subsequentially heat killed at 85 °C for 20–25 min. Reactions were subjected to UDG and NaOH treatment, as described above.

Chapter 5: Purified or lysate HsAID-His and PmCDA2-His (one preparation of one clone) were incubated with 16–50 fmol of ^{32}P-labelled bubble, single-stranded, double-stranded, or DNA/RNA substrate (**Figure 2.3A–C, G–J**) at 18–37 °C in 100 mM phosphate buffer pH 6.0–8.0 for 0.5–77 hours. Reactions were heat inactivated at 85 °C for 20–25 min. All reactions except for those with TG(mC)bub7 were then subjected to UDG and NaOH treatment, as described above. For TG(mC)bub7 reactions, purified or lysate HsAID-His and PmCDA2-His were incubated with 50 fmol of ^{32}P-labelled TG(mC)bub7 or TGTbub7 at 18–37 °C in 100 mM phosphate buffer pH 7.1 for 16 hours, then heat inactivated at 85 °C for 20–25 min. Next, 1 µmol of KCl and 2.5 pmol of TGC complementary strand (bottom strand in **Figure 2.3B** TDGpart2) was added to the

deamination reaction in a final volume of 20 µL and incubated in the annealing program, as described above. One unit of Thymine DNA Glycosylase (TDG; Trevigen) was incubated with the annealed reaction and 1x TDG Buffer in a final volume of 30 µL at 65 °C for 16 hours. The samples were then treated with NaOH and heat, as described above.

2.6 Time course enzyme kinetics experiments

Chapter 3: To determine time-point kinetics, each AID ortholog was incubated with 25–50 fmol of TGCbub7 (**Figure 2.3A**) at each AID ortholog's optimal temperature for durations ranging from 5 minutes to 20 hours. Reactions were heat inactivated at 85 °C for 20–25 min. Reactions were then subjected to UDG and NaOH treatment, as described above. Three independent preparations of each AID ortholog were used, except for Lc-AID, for which six purifications were tested, all in two to seven unique experiments. Fmol deaminated substrate/µg AID was plotted against time incubated. These experiments were used to determine time of initial exponential phase of AID activity, in order to proceed to Michaelis-Menten kinetics.

Chapter 4: Purified HsAID-His and PmCDA2-His were incubated with 17–25 fmol TGCbub7 (**Figure 2.3A**) at 34 °C for 1–60 minutes and 24 °C for 1–48 hours, respectively. Reactions were heat inactivated at 85 °C for 20–25 min. Reactions were then subjected to UDG and NaOH treatment, as described above. Two unique experiments were performed on one preparation of one clone of each enzyme. Percent of deaminated substrate was plotted against time incubated. These experiments were used to estimate the initial exponential phase of AID activity.

CDA1L1_3 spliced and unspliced, Hs-AID and Dr-AID His-tagged enzymes in HEK 293T lysates were incubated with 25 fmol of TTCbubB (**Figure 2.3D**) in 100 mM phosphate buffer at pH 6.2 (CDA1L1s), pH 6.7 (Dr-AID), or pH 7.7 (Hs-AID) at 14–27 °C for 0.5–12 hours. Reactions were heat inactivated at 85 °C for 20–25 min. Reactions were then subjected to UDG and NaOH treatment, as described above. Two clones (one preparation each) of CDA1L1_3 spliced and unspliced were used, along with one preparation of one clone each of Hs-AID and Dr-AID. Percent deamination was plotted against time incubated.

2.7 Enzyme assays to determine of optimal pH for activity

To determine the optimal pH of the lysate AID and its orthologs, the enzymes were incubated with radioactively-labelled substrates in buffers of a range of pH's. To accurately determine the effective pH of the final reaction solutions, the high activity 10 µL and low activity 20 µL reaction solutions were scaled up to 50 mL solutions and the pH measured. Large-volume solutions of lysis buffer (in which the His-tagged lysates are suspended), pH buffer, and TE buffer (in which the radioactively-labelled DNA substrates are diluted) were created at the following ratios: 5:4:1 for the 10 µL reactions, and 10:9:1 for the 20 µL reactions. The pH of each solution was measured, and the resulting effective pH's of each pH buffer are listed in **Table 2.1**.

Table 2.1 Effective pH of alkaline cleavage activity buffers

	pH	Effective pH (10 µL)	Effective pH (20 µL)
Citrate-Phosphate	2.69	4.12	3.89
	3.1	4.72	4.5
	3.5	5.03	4.84
	4.1	5.7	5.5
	4.37	5.8	5.7
	4.91	6.06	5.96
	5.37	6.18	6.11
	6.41	6.56	6.53
	7.24	7.12	7.09
Phosphate	5.58	6.24	6.18
	5.93	6.3	6.24
	6.22	6.42	6.37
	6.53	6.59	6.55
	6.86	6.82	6.76
	7.2	7.09	7.05
	7.48	7.33	7.29
	7.8	7.57	7.55
	8.02	7.75	7.73

To ensure the UDG enzyme worked at the lower pH's, a test experiment was performed using 50 fmol of ^{32}P radioactively-labelled TGUbub7 (**Figure 2.3A**), which would simulate 100% AID enzyme activity. The oligo was incubated with ~ 1.0 µg GST-Hs-AID (to simulate the final, effective pH) and 4 µL of phosphate or phosphate-citrate buffer at a range of pH's at 37 °C for 5 hours. Enzymatic activity was halted by incubating at 85 °C for 20 min. Next, 0.25 units of UDG enzyme and 1x UDG buffer was added to the incubation at a final volume of 20 µL, which was incubated at 37 °C for a further 30 min. Next, 4 µM of NaOH was added, and the reaction incubated for 10 minutes at 98 °C. All reactions were run on a 14 % acrylamide gel, which was then visualized (see **Figure 2.5**). The UDG enzyme worked less efficiently at the lower pH's than at the neutral pH's (**Figure 2.5A** and **B**). Therefore, a similar experiment was run using 0.5 units of UDG enzyme and incubating at this step for 1 hour. This extended UDG step resulted in nearly 100% efficiency (**Figure 2.5C** and **D**); therefore, for all pH experiments, the extended UDG step was used to ensure accurate measurement of the AID orthologs' activity at the various pH's.

Chapter 4: The highly active His-tagged enzymes (one preparation each of Hs-AID, Dr-AID, Gc-AID; two preparations, two clones of CDA1L1_3 and CDA1L1_3spl) in cell lysate were incubated with 25 fmol of TGCbub7 (**Figure 2.3A**; Hs-, Dr-, and Gc-AID) or 25 fmol of TTCbubB (**Figure 2.3D**; CDA1L1_3 and CDA1L1_3spl) at 16–22 °C for 1–5 hours in 100 mM phosphate buffer or 100 mM citrate-phosphate buffer at effective pH's of 6.2–7.7 and 4.1–7.1, respectively. The low activity His-tagged enzymes (two preparations, two clones of CDA1L1_1, CDA1L1_1spl, CDA1L1_2, and CDA1L1_4) were incubated with 25–50 fmol TTCbubB (**Figure 2.3D**) at 16–22 °C for 24–84 hours in 100 mM phosphate buffer or 100 mM citrate-phosphate buffer at pH's of 6.2–7.7 and 3.9–7.1, respectively. The high and low activity His-tagged enzymes were tested in three to four unique experiments.

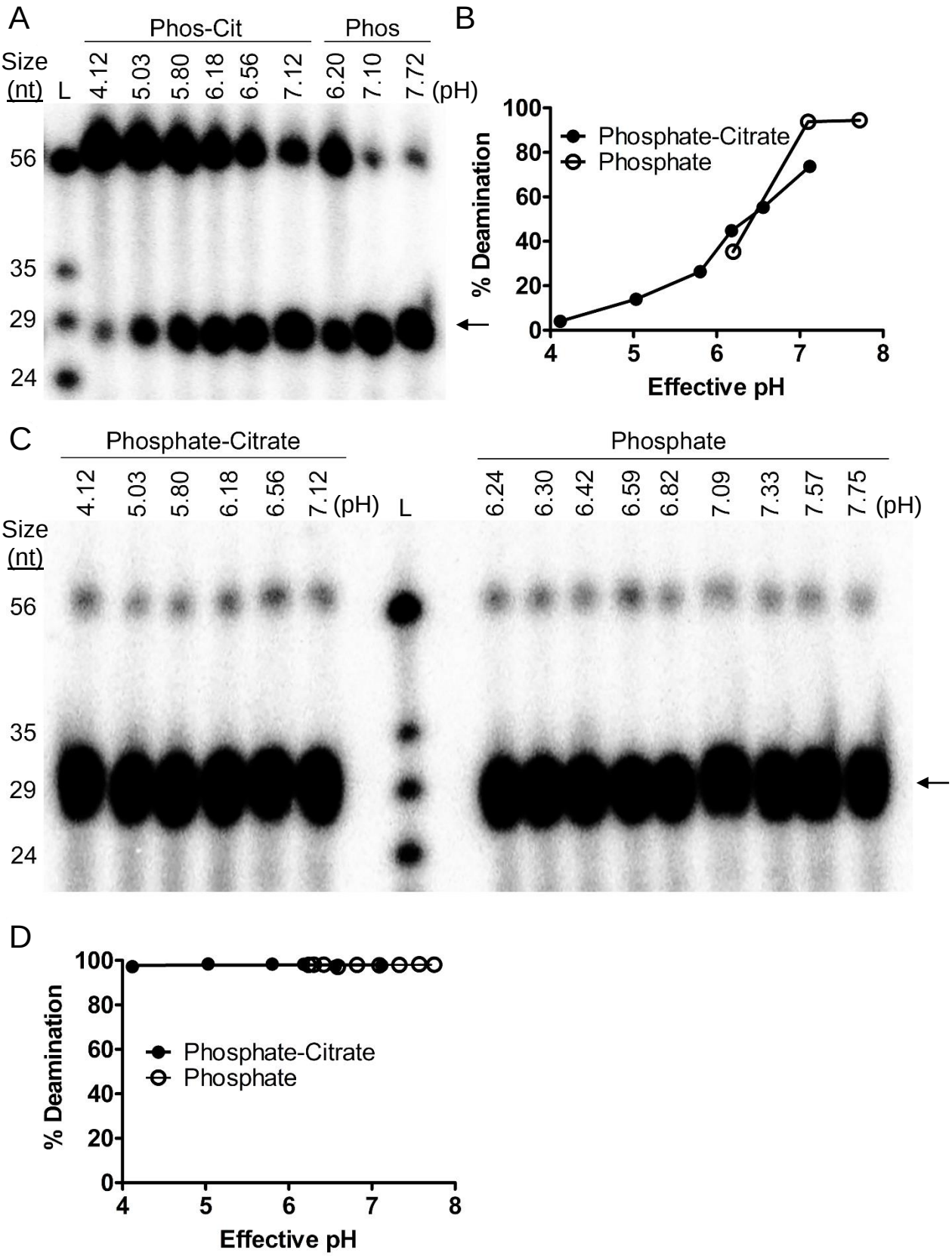


Figure 2.5 Determining conditions for optimal alkaline cleavage using lower pH buffers

Legend: Figure 2.5

A Alkaline cleavage gel of TGUbub7 incubated with GST-Hs-AID for 5 hours at 37 °C, 0.25 units of UDG for a half hour. L, ladder. **B** Graph of % deamination versus effective pH of the gel in A. **C** Alkaline cleavage gel of TGUbub7 incubated with 5 µL lysis buffer pH 8.2 for 1 hour at 37 °C, 0.5 units UDG for 1 hour. L, ladder. **D** Graph of % deamination versus effective pH of the gel in C. The black arrow indicates the product band on each alkaline cleavage gel.

One preparation of GST-Pm-CDA1 purified enzyme was incubated in two unique experiments with 8.33 fmol of TGCbub7 (**Figure 2.3A**) at 16 ℃ for 6 hours in 100 mM phosphate buffer of pH 5.9–8.2. One preparation of purified A3G-His was incubated with CCCssA3G (**Figure 2.3F**) in modified (1 mM DTT and 0.05 mg/mL RNaseA added) 100 mM phosphate buffer or 100 mM citrate-phosphate buffer at pH's of 6.2–7.7 and 4.1–7.1, respectively, at 37 ℃ for 0.5 hour. All reactions were heat inactivated at 85 ℃ for 20–25 min, then subjected to UDG and NaOH treatment, as described above.

For each experiment, % deamination was plotted against effective pH to generate a curve, with the peak representing the optimal pH of AID/CDA activity in each experiment. These pH's were then averaged for each ortholog to arrive at their overall optimal pH.

Chapter 5: Purified His-tagged Hs-AID and Pm-CDA2 were incubated with 50 fmol GTCbub7 (**Figure 2.3A**) at 37 ℃ in 100 mM phosphate buffer at pH's of 5.9–8.2 for 20 hours, then treated with UDG and NaOH, as described above. The percent deamination was plotted against pH.

2.8 Enzyme assays to examine Michaelis-Menten kinetics

Chapter 3: To determine Michaelis-Menten kinetics for comparison of initial velocities, 0.2 µg of AID was incubated with 1.5–100 fmol of TGCbub7 (**Figure 2.3A**) at each AID ortholog's optimal temperature for 2–7 hours (depending on the results of time-point kinetics, to ensure that reaction times did not exceed the initial linear phase of time dependency). All reactions were heat inactivated at 85 °C for 20–25 min, then subjected to UDG and NaOH treatment, as described above. Two to three independent preparations of each AID ortholog were tested in two to three unique experiments. Percent deamination was used to calculate velocity (fmol product/minutes incubation/µg AID), which was then plotted against substrate concentration.

Chapter 4: To determine initial velocity of PmCDA2-His and to compare it to that of Hs-AID-His, these lysate enzymes were incubated with 1.5–100 fmol of TGCbub7 (**Figure 2.3A**) at 24 °C for 48 hours (Pm-CDA2-His) and 34 °C for 90 minutes (Hs-AID-His). All reactions were heat inactivated at 85 °C for 20–25 min, then subjected to UDG and NaOH treatment, as described above. One experiment was performed on one preparation of one clone of each enzyme.

2.9 PCR-based deamination assay

Chapter 3: My colleagues and I have previously described the deamination-specific PCR assay to detect activity of purified AID on a ~ 500 bp-long stretch of denatured plasmid.[68,185,239] 100 ng of HindIII-digested linearized p219 plasmid (see **Figure 2.6**) was denatured at 99 °C for 10 minutes in 100 mM phosphate buffer (pH 7.1) followed by snap-cooling in an ice bath. ~ 1 µg of Hs- and Tn-AID, 1.5 µg of Lc-AID, 2 µg of Pm-CDA1, and 5 µg of Gc-AID was added to their respective denatured plasmid samples and the solution incubated for 1 hour at the optimal temperature for each AID ortholog as previously determined.

The plasmid substrate was again denatured and snap-cooled, followed by a second round of AID addition to a final volume of 40 µL, and incubated for another hour. Subsequently, 5 ng of plasmid (2 µL of the incubation solution) was then PCR-amplified using 7.5 pmol of deamination-specific primers in a total volume of 25 µL (94 °C for 2 min; followed by 35 cycles of 94 °C for 30 seconds, 50–64.7 °C for 1 minute, and 72 °C for 1 minute; with a 72 °C extension phase of 10 minutes). The annealing temperature of the PCR protocol was optimized for each ortholog to avoid biasing the PCR products toward highly-mutated amplicons in which all dCs were mutated by AID (see **Figure 2.7**). The optimized annealing temperature would produce a PCR band for the ortholog that was stronger than that seen in the AID-negative control at the same temperature, meaning some non-mutated strands were amplified alongside mutated strands. Nurse shark AID required extensive optimization compared to the other AID orthologs, likely due to its comparatively lower enzymatic activity.

Figure 2.6 p219 PCR-amplified sequence

The black text is the p219 plasmid with all cytidines converted to uridines, assuming 100% deamination by a cytidine deaminase. The first reverse primer (P1R, red text) anneals to the deaminated p219 plasmid, resulting in the replicated (blue text) strand, which is recognized by the first forward primer (P1F, red text). The resulting amplicons are 411 nt in length (dashed and solid underlined sequence). The nested primers (P2F and P2R, red text) result in amplicons 328 nt long (solid underlined sequence).

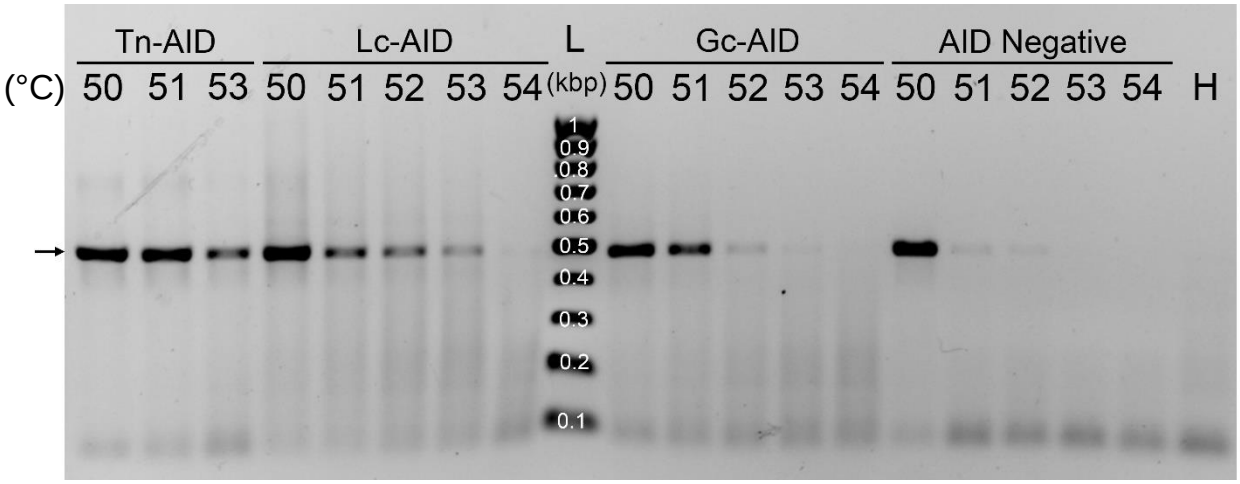


Figure 2.7 Optimization of PCR annealing temperature to amplify AID-mutated plasmids

Representative 1% agarose gel showing PCR-amplified bands of AID-mutated p219 plasmid at different annealing temperatures (inverted for better visibility). Tn-AID, Lc-AID, and Gc-AID were each incubated with p219 linearized plasmid at 20, 25, and 20 ℃, respectively. An AID-negative control (phosphate buffer) was incubated similarly at 25 ℃. 2 µL of each incubation product, plus a water negative (H) lacking plasmid, was subjected to PCR amplification with deamination-specific primers at different annealing temperatures (50–54 ℃ in this experiment). Bands can be visually detected in the AID negative controls up to 52 ℃, meaning non-mutated plasmids are amplified at these annealing temperatures; however, the bands at 51 and 52 ℃ are quite faint, thus these annealing temperatures are possible choices for a balance between PCR amplification and AID mutation. Tn-AID bands are of high intensity at all temperatures tested, whereas Lc-AID bands are slightly weaker at the higher temperatures, Gc-AID even less so. The optimal annealing temperatures for each ortholog were chosen such that the amplified bands at those temperatures were visually more intense than those of the AID-negative control at the same temperature. Thus, the following annealing temperatures were chosen: Tn-AID, 51 ℃; Lc-AID, 51 and 52 ℃; and Gc-AID, 51 ℃. L, ladder. Black arrow indicates the band of the correct size (~411 bp).

Once the primary PCR protocol was optimized, the amplified DNA (407 nucleotides in size) was then gel purified (Qiagen), and a second more stringent nested deamination-specific PCR (with the same annealing temperature as the first, unique to each ortholog) was carried out wherein 2 µL of the first PCR reaction was added to the nested PCR reaction (94 °C for 2 min; followed by 40 cycles of 94 °C for 30 seconds, 50–64.7 °C for 30 seconds, and 72 °C for 30 seconds; with a 72 °C extension phase of 10 minutes). The nested PCR protocol includes 7.5 pmol of a nested set of deamination-specific primers, resulting in amplicons of 329 nucleotides in size.[239] Both primary and secondary PCR products were TA-cloned (Invitrogen) and for each AID ortholog ~ 100 amplicons were sequenced (Macrogen, South Korea). To avoid analyzing multiple amplicons of the same original plasmid, thus biasing the WRC and other analyses, duplicate amplicons with the same mutation pattern were removed.

2.10 Electrophoretic mobility shift assay

Chapter 3: 0.4–50 fmol of TGCbub7 (**Figure 2.3A**) was incubated with 0.1–0.8 µg of each AID ortholog in binding buffer (1 mM DTT, 2 µM $MgCl_2$, 50 µM NaCl, dH_2O) in a total volume of 10 µL at 25 °C or its optimal temperature for 1 hour followed by UV-cross-linking on ice, as previously described.[227] The 0.2 mL PCR tubes were laid on their side on a bed of ice slush in a styrofoam box and cross-linked at 4 °C, 120 mJ/cm^2 , < 1 cm from the lamps, for 15 sec. The ice box was rotated 180° and samples were cross-linked for another 15 sec. Next, 5 µL of loading dye (50% glycerol, 0.25% Bromophenol Blue, 0.25% Xylene Cyanol) was added to each sample followed by electrophoresis on a pre-run (300 V for 1–2 hours) 8% native acrylamide gel (0.5x TBE, 8% acrylamide:bisacrylamide, 6% glycerol, dH_2O) and electrophoresed at 4 °C, 300 V for 3 hours. Gels were exposed to a Kodak Storage Phosphor Screen GP (Bio-Rad) for 16 hours and visualized using a PhosphorImager (Bio-Rad, Hercules, CA, USA) on Quantity One software (Bio-Rad, Hercules, CA, USA). Two to three independent preparations of each AID ortholog were tested, each in two to three unique experiments. Fmol substrate bound to AID was plotted against nM of free substrate, followed by non-saturation binding kinetic analysis (GraphPad Prism Software, GraphPad, San Diego, CA, USA) to obtain binding half-saturation Kd values.

2.11 Structure prediction of AID and AID-DNA complexes

Chapter 3 and 4: Similar to the methods described in King and Larijani 2017[252] and King *et al*. 2015[63], five resolved homologous APOBEC structures were chosen as templates for homology modeling: mouse A2 NMR (PDB: 2RPZ), A3A NMR (PDB: 2M65), A3C (PDB: 3VOW), A3F-CTD (C-terminal domain) X-ray (PDB: 4IOU) and A3G-CTD X-ray (PDB: 3E1U); the structures were obtained from 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/) full-length AID was modeled from the APOBEC templates to generate 25 models for each of the five structures, totaling 125 models. For each model, much of the protein was homology-modeled save the non-homologous 18 C-terminal amino acids, which were modeled *ab initio*. The indented space containing the Zn-coordinating and catalytic residues (H56, E58, C87 and C90 in Hs-AID) defined the catalytic pocket. The PARSE force field in PROPKA 3.0 was used to calculate the protein charge and isoelectric point (pI) for each AID structure.[265–267]

To dock the DNA substrates to each AID model, Swiss-Dock (http://www. swissdock.ch) was used.[268,269] Each substrate was constructed in Marvin Sketch v.5.11.5 (http://www.chemaxon.com/products/marvin/marvinsketch/), while docking parameters and surface topology were generated using Swiss-Param (http://swissparam.ch)[270]; the output files served as the ligand file in Swiss Dock. The ssDNA substrates for human and bony fish had a target sequence of 5'-TTTGCTT-3' ssDNA, while 5'-TTGTCTT-3' was chosen for Pm-CDA1, as both substrates were determined to be preferred by their respective enzymes in this and previous studies.[226,227,240,271] Simulations of substrate docking for each AID enzyme resulted in 5,000–15,000 binding modes, of which 8 or more were clustered based on root mean square (RMS)

values. The 32 lowest-energy clusters were selected, representing 256 of the lowest-energy individual binding events for each AID; 5600 low-energy docking clusters in total were analyzed for AID orthologs in UCSF Chimera v1.7 (https://www.cgl.ucsf.edu/chimera).[272] AID-DNA structure complexes were chosen based on accessibility of the dC-NH_2 substrate group to the catalytic Zn-coordinating and the glutamic acid residues (H56, C87, C90, and E58 in wild-type Hs-AID).

Chapter 5: For structural modeling and substrate docking, we used the same strategy as described above, which was used to arrive at the first published computational-biochemical AID structure that has since been confirmed by a partial AID crystal structure.[63,66] Using I-TASSER (https://zhanglab.ccmb.med.umich.edu/I-TASSER/), full-length orthologous AID and CDA protein structures were modeled at pH 7.0 from APOBEC templates to generate 9–25 models for each, totaling 145 models. Nonhomologous regions between the target and APOBEC template were modeled *ab initio*. The 5′-TTTGCTT-3′ ssDNA substrates were chosen, as the TGC motif is the most commonly-preferred motif among all AID enzymes tested thus far. ssDNA was docked within 30 × 30 × 30 Å (x, y, and z, respectively) from the Zn-coordinating histidine in the catalytic pocket. CDA-DNA complexes conducive to deamination were defined by the accessibility of the (d)CNH2 substrate group to the catalytic Zn-coordinating and the glutamic acid residues within the putative catalytic pockets.

2.12 Data quantification and statistics

The amino acid alignment analysis was performed using STRAP (http://www.bioinformatics.org/strap/) version 2018-08-11. The nucleotide alignment analysis was performed using PROMALS3D (http://prodata.swmed.edu/promals3d/promals3d.php).

The sequence logos for the PCR deamination assay in Chapter 3 were generated using WebLogo 3 software (weblogo.threeplusone.com). The mutation frequencies of the XXC motifs generated by each AID ortholog were translated into sequence consensus data: the frequency of mutation for a certain XXC motif is directly correlated to the number of multiples of said motif in the consensus.

For enzyme activity and binding assays, band densitometry was performed using Image Lab Analysis Software (Bio-Rad, Hercules, CA, USA) and Quantity One (Bio-Rad, Hercules, CA, USA). For each experiment, individual lanes of each gel were quantified 2–3 times; variability between each quantification was < 5%. Values from individual experiments were averaged and plotted as a single data point. For thermosensitivity and binding assays, the optimal temperatures and Kd values, respectively, were calculated using GraphPad Prism Software (GraphPad, San Diego, CA, USA) for each enzyme in each experiment and were averaged to obtain final values. For substrate specificity, time course, and Michaelis-Menten kinetics experiments, data points from all unique experiments were averaged for each AID ortholog. Therefore, each numerical characteristic (optimal pH/substrate/temperature, Michaelis-Menten kinetics) is derived from 2–3 quantifications each of up to six individual experiments each of two to four protein preparations each of one to two clones of each AID ortholog. The data were graphed and statistically analyzed using GraphPad. Error bars represent standard deviations (SD). *P* values for the substrate

specificity assay in Chapter 3 were determined by the Mann Whitney Test, which compared the relative deamination of different substrates. The *P* values for the Michaelis-Menten kinetics assay were determined by two-tailed, non-parametric paired *t* test, which compared the initial velocities of the five AID orthologs.

2.13 Accession numbers

Table 2.2 Accession numbers for AID protein amino acid sequence

Genbank (https://www.ncbi.nlm.nih.gov/genbank/)

Chapter	Species	Protein Name		Accession No
		Full	Short	
3	*Homo sapien*	activation-induced cytidine deaminase	Hs-AID	AB040430.1
	Callithrix jacchus	predicted single-stranded DNA cytosine deaminase isoform X1	Cj-AID	XP_017832076.1
	Pteropus vampyrus	predicted single-stranded DNA cytosine deaminase	Pv-AID	XP_011364076.1
	Ornithorhynchus anatinus	predicted single-stranded DNA cytosine deaminase	Oa-AID	XP_001516174.2
	Anolis carolinensis	predicted single-stranded DNA cytosine deaminase isoform X1	Ao-AID	XP_008102036.1
	Pleurodeles waltl	activation-induced cytidine deaminase	Pw-AID	CBG76578.2
	Danio rerio	activation-induced cytidine deaminase	Dr-AID	NM_001008403
	Petromyzon marinus	cytosine deaminase	Pm-CDA1	ABO15149.1
4	*Lampetra planeri*	cytidine deaminase	LpCDA1L1_1	AVN88307.1
	Lampetra planeri	cytidine deaminase	LpCDA1L1_1sp	AVN88319.1

Table 2.2 continued

Chapter	Species	Protein Name		Accession No
		Full	Short	
4	*Lampetra planeri*	cytidine deaminase	LpCDA1L1_2	AVN88308.1
	Lampetra planeri	cytidine deaminase	LpCDA1L1_3	AVN88309.1
	Lampetra planeri	cytidine deaminase	LpCDA1L1_3sp	AVN88320.1
	Lampetra planeri	cytidine deaminase	LpCDA1L1_4	AVN88310
	Homo sapiens	cytidine deaminase	APOBEC3G	NM_021822.4
5	*Petromyzon marinus*	cytosine deaminase	Pm-CDA1	ABO15150.1
	Lampetra planeri	cytidine deaminase	Pm-CDA1	AVN88323.1
	Lampetra planeri	cytosine deaminase	LpCDA2_v3.1	AVN88324.1
	Lampetra planeri	cytosine deaminase	LpCDA2_v3.2	AVN88325
Uniprot (http://www.uniprot.org/):				
3	*Tetraodon nigroviridis*	uncharacterized protein (CMP/dCMP-type deaminase)	Tn-AID	H3CRQ9

Chapter 3 Biochemical regulatory features of AID have remained conserved from lamprey to humans

3.1 Authorship statement

The data in this chapter were published in a manuscript titled "Biochemical Regulatory Features of Activation-Induced Cytidine Deaminase Remain Conserved from Lampreys to Humans" in the Molecular and Cellular Biology journal in 2017.[229] Author E Hsu provided the sequence of the nurse shark AID gene. CT Amemiya previously sequenced and published the coelacanth genome in Nature "The African coelacanth genome provides insights into tetrapod evolution" (2013), and, with Dr. Hsu, assisted in reconstructing the 5' end of the AID gene. Drs. Hsu and Amemiya provided scientific analysis and review of the project and final manuscript. JJ King performed predictive structure modelling and substrate docking of the AID proteins. M Larijani and I designed and analyzed the research and edited the manuscript. I expressed and purified all AID proteins, performed all the experimental work except for the aforementioned structural modeling, conducted the literature review, analyzed the data, and wrote the manuscript and this chapter. I also illustrated and designed all figures.

3.2 Introduction

Since their discovery, murine and human AIDs have been subject to intense research, with the latter proven to cause DNA damage resulting in tumourigenesis through targeting of non-immunoglobulin genes.[98,252,273–277] While its role in cancer and disease has been the main focus of the majority of research regarding human AID (Hs-AID), some of its biochemical characteristics have been elucidated. Hs-AID preferentially deaminates dC to dU within WRC (W is A/T; R is A/G) motifs in ssDNA.[185,239,240,242,244,248] Human AID has a relatively slow catalytic rate (one reaction every 1–4 minutes) and exceptionally high affinity (nM range) for ssDNA binding,[226,240] while typical human enzymes have catalytic and substrate on/off rates thousands of times faster.[278] We recently discovered that AID's strong affinity for negatively-charged ssDNA is due to its highly positively-charged surface (+14 in neutral pH), resulting in DNA binding non-specifically outside the catalytic pocket. Further, AID appears to conform to a catalytically unfavourable structure ~ 75% of the time, which, combined with the positively-charged surface, explains AID's lethargic activity: only ~ 1% of AID:ssDNA binding events result in mutation catalysis by AID. We have proposed that these unusual characteristics evolved to protect genomic integrity.[63,226,228,240,252] This concept was supported by the finding that AID "upmutants", which exhibited higher catalytic rates *in vitro*, inflicted more damage to the genomes of cells in which they were expressed.[279]

Non-mice AID orthologs from distantly-related vertebrates have also been identified and their functionality investigated, including AID from the channel catfish (*Ictalurus punctatus*, Ip-AID),[170] and zebrafish (*Danio rerio*, Dr-AID),[263] yet their activity relative to human AID remained elusive. This is because orthologous AIDs have been studied in the context of xenogeneic cell lines with the suppressing effects of cellular DNA repair processes, making it difficult to quantify and

compare their biochemical and structural properties. To address this issue, our previous studies comparing the enzymatic properties of Dr-AID and Ip-AID with Hs-AID involved highly purified enzymes. We found the fish AID orthologs were more active at colder temperatures compared to human AID (20–25 °C vs 30–37 °C, respectively). Furthermore, although both earlier-diverged orthologs bound ssDNA with the same high affinities as human AID, Dr-AID exhibited the fastest catalytic rate, followed by Hs-AID, while Ip-AID was the slowest of the three,[183,184] yet all catalytic activities remained well below that of other common enzymes.[278] Thus, investigation of these inter-species differences using mutant and chimeric enzymes led to significant insights into structure-function relationships of AID. However, how well-conserved these characteristics are across evolution remained unclear.

Complicating interpretation was the fact that Dr-AID may be quite unique among orthologs in that it appears to play epigenetic roles outside the immune system. Dr-AID is required for neurogenesis in developing zebrafish embryos by turning on broad gene expression cascades through widespread demethylation of promoter CpG motifs.[186] We showed that Dr-AID is unique among all tested bony fish AID orthologs in its capacity to deaminate 5-methylcytosine (mC) in CpG motifs, hence its ability to function in mC demethylation.[184]

In this chapter, I sought to study the biochemical properties of other earlier-diverged AID orthologs to evaluate the conservation or divergence of AID's unusual properties. I focused on AID from species representative of key divergence points of adaptive immune systems evolution in fishes. These include the sea lamprey (Pm-CDA1), the nurse shark (Gc-AID), the 'living fossil' fish coelacanth (Lc-AID), and a recently-diverged bony fish, tetraodon (Tn-AID). I report that these earliest-diverged AID orthologs exhibit unique substrate specificities and optimal temperature tolerances, while AID's lethargic enzymatic rate and high affinity for ssDNA are

remarkably conserved. These results shed light on the molecular evolution of AID in the context of diverging adaptive immune systems and highlight the importance of its unique biochemical properties.

3.3 Results

3.3.1 Selection of AID orthologs covering the broadest window of evolution

The earliest-diverged vertebrates are the agnathans, a.k.a. jawless fish, of which the cyclostomes are the only remaining lineages, consisting of the lamprey and hagfish (**Figure 3.1**).[280] The lamprey is of major interest to immunologists due to its unique adaptive immune system capable of memory and graft rejection, but lacking major cornerstones of canonical adaptive immunity, such as T and B cell receptors and MHC proteins. However, lamprey do have three types of variable lymphocyte receptors (VLRA, B, and C) which are expressed on lymphocyte-like cells. These VLRs have structures similar to Toll-like receptors and play similar roles to those of immunoglobulin (Ig; VLRB) and T cell receptors (VLRA and C). VLRA/C receptors are bound to the membrane and expressed in cells associated with the thymus-like thymoid tissue in the gills; VLRB receptors are both bound and secreted by lymphocytes associated with the lamprey kidney.[151,230,281–289] These VLRs are clonally-expressed and somatically recombined in response to immune stimulation via a type of gene conversion, but both jawless vertebrates lack the recombination activating gene products (RAG 1 and 2).[126,145,151,230,281–284,286–288,290–292]

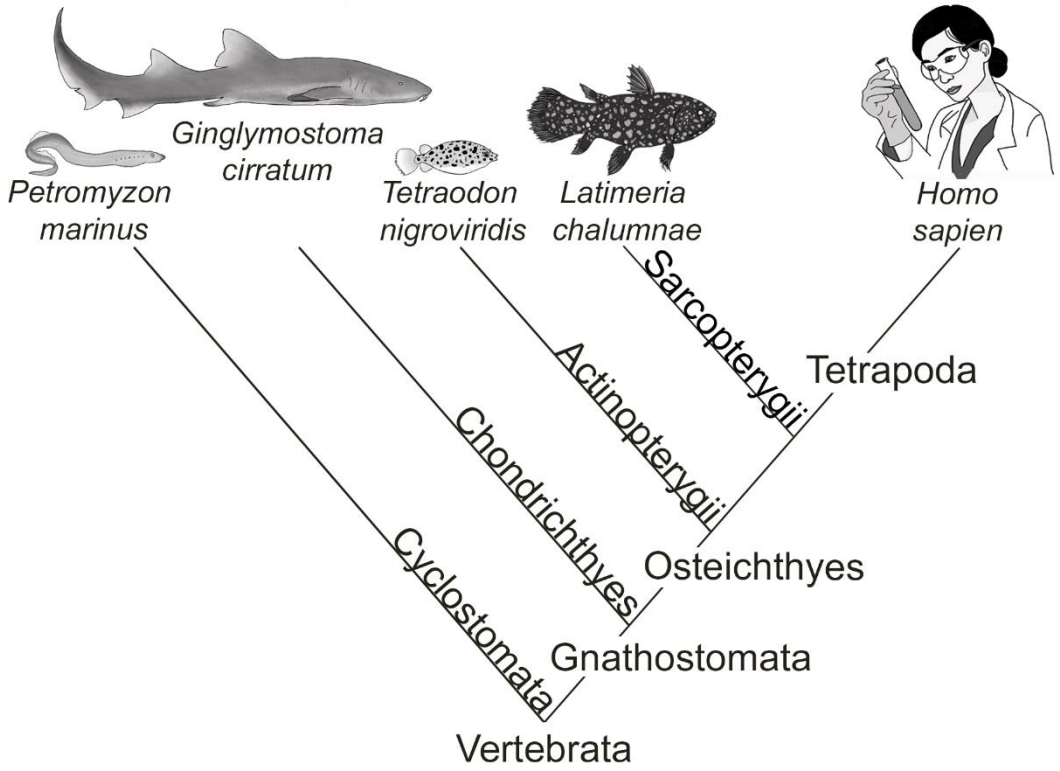


Figure 3.1 Phylogenetic tree of representative species

Adapted with permission from Figure 1 in Quinlan et al.*, 2017.*[229]

Phylogenetic tree showing the evolutionary relationships between fish species and humans: *Petromyzon marinus* (lamprey), *Ginglymostoma cirratum* (nurse shark), *Tetraodon nigrovirigis* (tetraodon), *Latimeria chalumnae* (coelacanth) and *Homo sapiens* (human). Vertebrata, subphylum of chordates with backbones; Cyclostomata, class of jawless vertebrates; Gnathostomata, superclass of jawed vertebrates; Chondrichthyes, class of cartilaginous fish; Osteichthyes, superclass of bony fish; Actinopterygii, class of ray-finned fishes; Sarcopterygii, class of lobe-finned fishes; Tetrapoda, superclass of four-limbed animals.

While attempting to identify how VLRs are assembled, Rogozin *et al*. identified CDA1 and CDA2, the earliest-diverged vertebrate AID orthologs identified to date.[126] Investigations into the expression patters of each CDA determined that CDA1 is associated with the T-cell-like cells VLRA and VLRC, while CDA2 is expressed in B-cell-like VLRB cells.[151,291,292] Both cytidine deaminases have poor sequence similarity to other AID sequences, yet CDA1 was found to be an active cytidine deaminase, inducing Ig recombination in yeast and dC to dT mutations in *E. coli*.[126] CDA1 exhibited mutagenic activity when expressed in cells, however, it had not yet been purified to be directly studied. Therefore, I chose to purify CDA1 from *Petromyzon marinus* (Pm-CDA1) to biochemically compare the earliest-diverged AID ortholog with the latest-diverged and most studied, Hs-AID.

The earliest-diverged jawed vertebrates (gnathostomes) are the cartilaginous fish (chondrichthyes). Sharks are descendants of this lineage and are among the first species to evolve a V(D)J- and Ig-based adaptive immune system.[148,160,163,231,293,294] Shark Ig genes undergo secondary diversification modification via SHM and recent studies suggest they may also undergo a rudimentary type of class-switching.[148,163,295,296] Furthermore, there is evidence of productive somatic AID mutations in the TCRs of two shark species; a rare occurrence, AID-initiated SHM of TCRs has only been found in two other gnathostomes, the dromedary camel and the Ballan wrasse fish.[153,154,297,298] At time of writing, only one cytidine deaminase from a shark species has been identified (elephant shark),[232] and no AID from any shark species has been biochemically evaluated. To this end, we identified an AID-coding gene from nurse shark (*Ginglymostoma cirratum*, Gc-AID), and expressed and purified this putative enzyme.

About 50 million years after the divergence of cartilaginous fish, the first bony fish (Osteichthyes) emerged, eventually giving rise to the ray-finned fish (Actinopterygii). These ray-

finned fish constitute the largest and most diverse vertebrate group, with unique V(D)J immunoglobulin loci organization and isotypes (see Section 1.8).[280] The AID orthologs of several earlier-diverged Actinopterygii species have been shown to be functional cytidine deaminases, mostly in cellular contexts.[164,169,170,178,179,181,183,184,186,263] For this study, I chose a lesser-studied Actinopterygii AID ortholog from tetraodon (*Tetraodon nigroviridis*).

Lastly, the closest fish ancestors to tetrapods (four-legged animals) are the lobe-finned fish (Sarcopterygii), of which coelacanth (*Latimeria chalumnae*) is an extant member whose genome, including its AID gene (Lc-AID), has been sequenced.[138] The only other extant members are lungfish species, who apparently lack an AID ortholog, making coelacanth our closest fish ancestor to have an AID-encoding gene.[145] Evidence of V(D)J recombination of coelacanth Ig has also been reported.[260,299] Thus, as shown in **Figure 3.1**, the four chosen fish AID orthologs are representative of key points in the evolution of adaptive immunity, while simultaneously covering the broadest evolutionary distance relative to each other and to Hs-AID.

3.3.2 Genetic comparison of AID orthologs

Figure 3.2A shows an alignment of the amino acid (AA) sequences of the five AID orthologs, while **Figure 3.2B** shows the degree of AA identity and similarity among the orthologs. As expected, the ortholog from the earliest-diverged lineage, Pm-CDA1, is the least conserved with Hs-AID (17% identity, 37% similarity). Also as expected, the most recently diverged fish AID ortholog, Lc-AID, exhibited the highest level of sequence similarity with Hs-AID (74% identity, 85% similarity). Tn-AID exhibited a moderate level of conservation with Hs-AID, between that of Pm-CDA1 and Lc-AID (55% identity, 68% similarity). Tn-AID has two inserts unique to ray-finned fish spanning AA positions 65 to 79 (**Figure 3.2A**).[183,184] Without these inserts, Tn-AID's AA identity and similarity to Hs-AID only increases by 2 and 3%, respectively (**Figure 3.2B**, dotted line); therefore, Tn-AID's lower sequence similarity is not due to these inserts. Gc-AID exhibited higher conservation with Hs-AID (66% identity, 78% similarity) compared to Tn-AID.

Along with the AA sequences, I compared relative conservation among each AID ortholog's cDNA sequence (**Figure 3.2C**). The nucleotide comparison shows a similar pattern of sequence identity between the AID orthologs as seen in the AA comparison. Pm-CDA1 is the least conserved (39%), while Lc-AID has the closest nucleotide conservation to Hs-AID (73%). Gc-AID is slightly more homologous to Hs-AID than Tn-AID (69% vs 63%, respectively). I conclude that sequence similarity to Hs-AID generally corresponds with evolutionary distance. This suggests that if the biochemical properties of AID are indeed divergent among species, this ought to be observable in our comparative study.

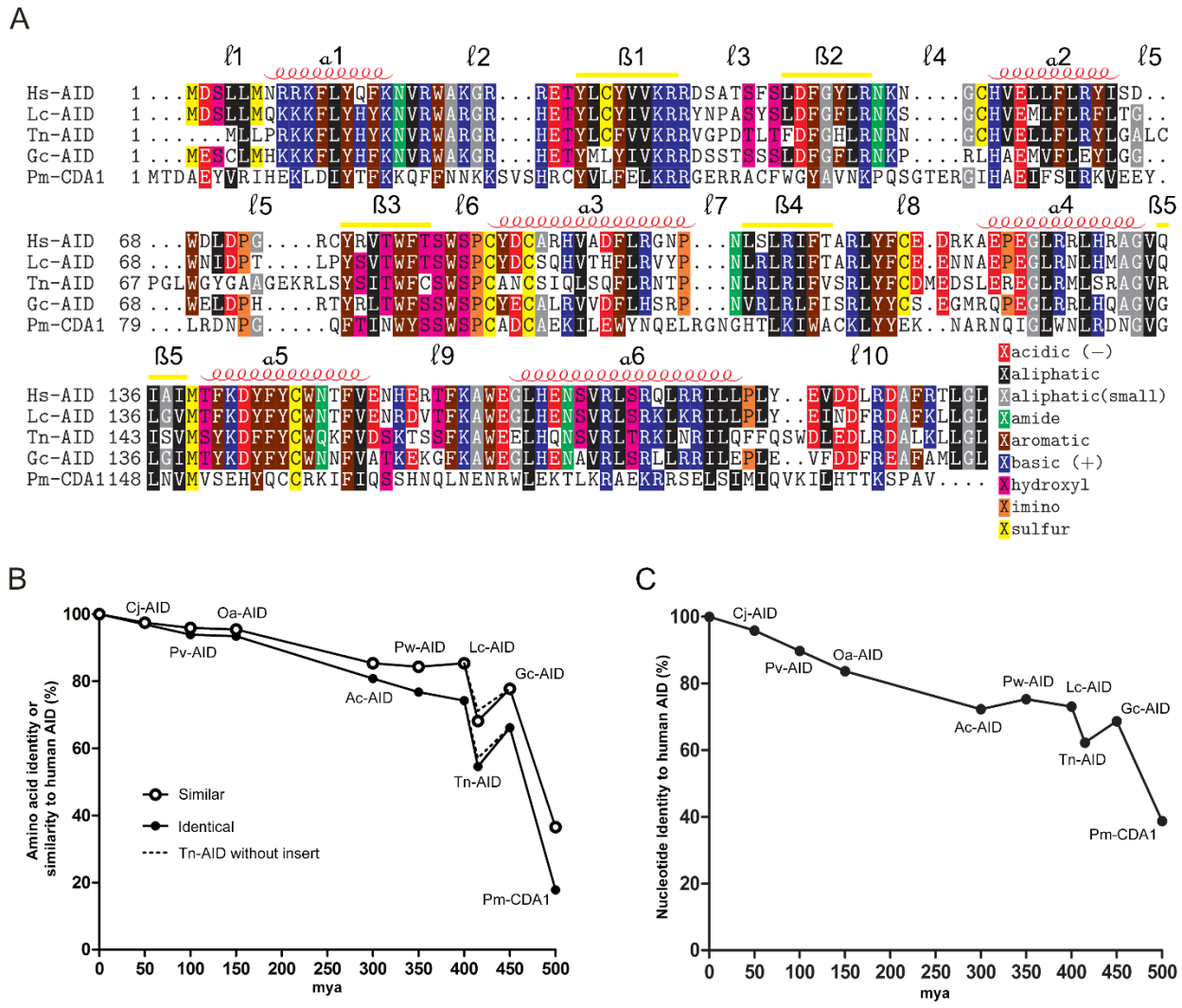


Figure 3.2 AID in the context of evolution

Adapted with permission from Figure 1 in Quinlan et al., *2017.*[229]

Legend: Figure 3.2

A Alignment of AID protein sequences from human (Hs-AID), coelacanth (Lc-AID), tetraodon (Tn-AID), nurse shark (Gc-AID), and lamprey (Pm-CDA1). L, loop; α, helix; and β, strand. **B** Percent identity and similarity were calculated by comparing identical and similar amino acids between the indicated AID orthologs and Hs-AID. To investigate the contribution of the ray-finned fish insert to sequence identity and similarity, identical and similar homologies between Tn-AID and Hs-AID were calculated without this insert. The approximate period of species evolution (mya, million years ago) is shown on the x axis. Sequence identity and similarity to AID are based on published sequences of AID; some are predicted: 0 mya, *Homo sapiens*; 50 mya, *Callithrix jacchus* (Cj-AID); 100 mya, *Pteropus vampyrus* (Pv-AID); 150 mya, *Ornithorhynchus anatinus* (Oa-AID); 300 mya, *Anolis carolinensis* (Ac-AID); 350 mya, *Pleurodeles waltl* (Pw-AID); 400 mya, *Latimeria chalumnae* (Lc-AID), *Tetraodon nigroviridis* (Tn-AID); 450 mya, *Ginglymostoma cirratum* (Gc-AID); and 500 mya, *Petromyzon marinus* (Pm-CDA1). **C** Percent sequence identity was calculated by comparing the nucleotide sequence between the AID-encoding cDNA sequences of fish AID orthologs and Hs-AID. The approximate period of species appearance (mya) is shown on the x axis. The cDNA sequences of the same non-fish species as used to provide context for amino acid conservation were again used: 0 mya, Hs-AID; 50 mya, Cj-AID; 100 mya, Pv-AID; 150 mya, Oa-AID; 300 mya, Ac-AID; 350 mya, Pw-AID; 400 mya, Lc-AID, Tn-AID; 450 mya, Gc-AID; and 500 mya, Pm-CDA1.

3.3.3 Fish AID orthologs are active cytidine deaminases

To compare some of their biochemical properties, I generated expression constructs and expressed and purified AID orthologs from the aforementioned species as glutathione S-transferase (GST)-tagged fusion proteins, as previously described.[183,184,226] First, I verified the quality and determined the concentrations of multiple independently purified preparations of each AID on denaturing SDS-PAGE gels (**Figure 3.3A**). Once appropriate purification of each AID preparation was confirmed, I tested whether they are active cytidine deaminase enzymes.

For initial activity tests, I utilized the standard alkaline cleavage assay for cytidine deamination which I and others have previously established to measure AID activity on ssDNA.[142,183–185,227,300,301] I incubated each AID at long intervals at three different temperatures (18 °C, 25 °C, and 37 °C) with TGCbub7 (**Figure 2.3A**), which is a standard bubble-type substrate that contains the WRC motif TGC (**Figure 3.3B**, top). The longer incubation at three different temperatures accounts for possible differences in absolute activity levels, as well as optimal temperatures among AID orthologs, and would allow for enzymatic activity, however faint, to be seen. As shown in **Figure 3.3B**, bottom, when a representative preparation of each AID ortholog was incubated with TGCbub7 overnight, Pm-CDA1 and Hs-AID showed moderate-to-high levels of activity (54.7% and 93.0% deamination, respectively), whereas the other AID orthologs appeared to be less active (1.8%, 5.1%, and 16.3% for Gc-, Tn-, and Lc-AID, respectively).

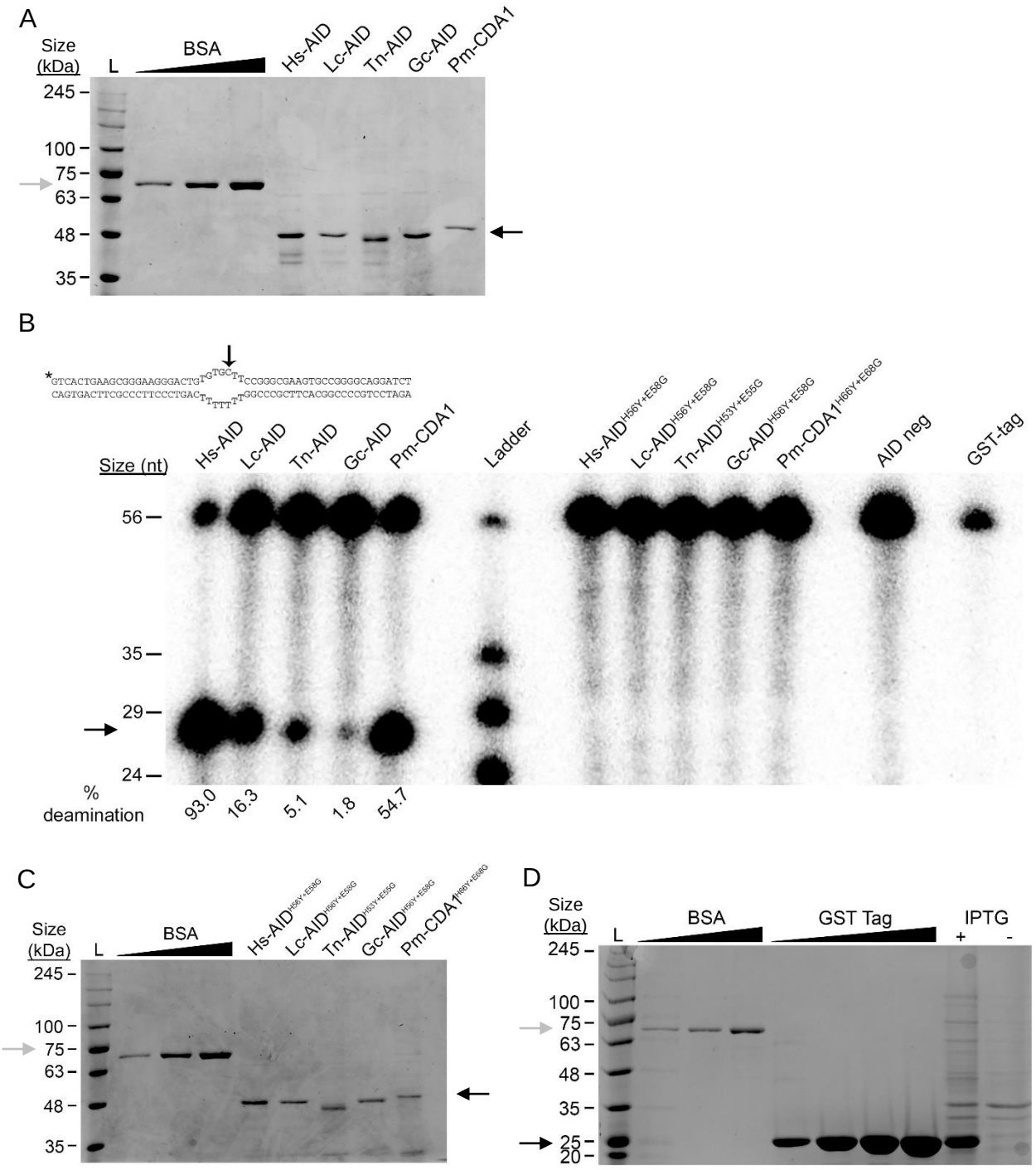


Figure 3.3 AID orthologs are active cytidine deaminase enzymes

Adapted with permission from Figure 2 and Supplementary Figure 1 in Quinlan et al., *2017.*[229]

Legend: Figure 3.3

A Representative SDS-PAGE gel of AID purifications. Multiple concentrations (0.25 µg, 0.5 µg and 1.0 µg) of BSA standard (~ 66.5 kDa, grey arrow) is used to confirm the size and amount (0.5 µg) of wild type Hs-, Lc-, Tn-, Gc-AID and Pm-CDA1 purified preparations (black arrow). All five GST-AID orthologs are ~ 50 kDa in size. **B**, **top** TGCbub7 substrate. *, ^{32}P label. The vertical black arrow shows the target cytidine in the bubble. **B, bottom** Representative alkaline cleavage gel showing activity of 0.5 µg of Pm-CDA1, Gc-, Tn-, Lc-, and Hs-AID, and their corresponding catalytically dead mutants (single-point mutations were made in two catalytic residues – histidine 56 and glutamic acid 58 in Hs-AID) on TGCbub7, incubated at each enzyme's optimal temperature (14.5 °C, 20.5 °C, 20.0 °C, 25 °C, and 31 °C, respectively – discussed further in the following section). The substrate and product (black arrow) are 56 and 28 nucleotides in length, respectively. The negative control is the substrate TGCbub7 and activity buffer with no enzyme added. **C** SDS-PAGE of purified Hs-, Lc-, Tn-, and Gc-AID and Pm-CDA1 catalytically inactive mutant preparations (0.5 µg each, black arrow) with a BSA standard (grey arrow). **D** SDS-PAGE of 7–22 µg of GST tag, expressed from the vector bearing only the GST tag with no AID insert, and purified in the same manner as the wild-type GST-AID preparations (~ 25 kDa, black arrow), with a BSA standard (grey arrow).

To confirm that the observed activity was *bona fide* AID activity through cytidine deamination, I created catalytically "dead" AID mutants, wherein one of the zinc-coordinating amino acids and the proton-donating glutamic acid in the catalytic pocket, previously shown to be necessary for cytidine deamination,[63,244,302] were mutated (H56Y and E58G for Hs-AID, Lc-AID, and Gc-AID; H53Y and E55G for Tn-AID; and H66Y and E68G for Pm-CDA1). I purified each mutant, confirming purification on SDS-PAGE (**Figure 3.3C**).

When a representative preparation of each catalytically dead AID mutant was incubated with TGCbub7 (**Figure 2.3A**) overnight and tested alongside the wild-type AID orthologs (**Figure 3.3B**), no cleavage of the substrate was observed. As a second independent negative control of the purity and specificity of the AID purifications, the construct bearing only the GST tag with no AID insert was expressed, and the GST tag was purified in the same manner as the wild-type GST-tagged AID orthologs. The purity of this preparation was determined through visualization on an SDS-PAGE gel (see **Figure 3.3D**). When the GST tag preparation was incubated overnight with TGCbub7 (**Figure 2.3A**) and visualized on an acrylamide gel alongside the wild-type and dead AID orthologs, the substrate was not cleaved (**Figure 3.3B**). Thus, any activity seen in our alkaline cleavage assay can be attributed to cytidine deaminase activity of each purified AID ortholog.

To independently confirm that the fish AID proteins were active, I also used a second type of deamination assay. In this PCR-based assay that we previously established,[185,239] a plasmid substrate, rendered single-stranded through restriction enzyme digestion and heat denaturation, is incubated with AID, followed by PCR using deamination-specific primers to selectively amplify AID-mutated DNA (**Figure 3.4A**), which is then sequenced to verify AID-mediated mutations.

Using this assay, I confirmed that all five AID orthologs catalyzed dC-to-dU mutations detectable by PCR (**Figure 3.4B**). When the p219 plasmid was incubated with the GST tag alone, purified in parallel with the AID orthologs, no amplification using the deamination-specific primers was observed (**Figure 3.4B**); therefore, any cytidine deaminase activity seen in this assay can be directly attributed to specific AID-mediated dC-to-dU conversion.

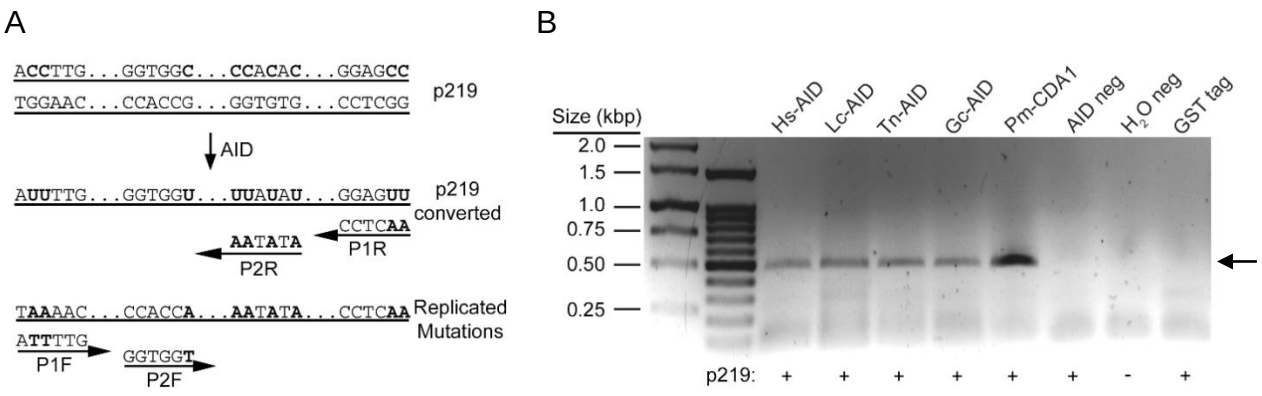


Figure 3.4 AID orthologs are active on long ssDNA

Adapted with permission from Figure 2 in Quinlan et al.*, 2017.*[229]

A Schematic of p219 showing the principle of the PCR assay for detecting AID activity on long stretches (several hundred nucleotides) of plasmid ssDNA. Outside primers: P1F (GGAAGGTATGAAAATAGGAAAAGAAAATAAATAAATTTTG) and P1R (CCCCTAACTTTTATACCCAACCCTAACTCC). Nested primers: P2F (CCCCCCGATCCGTATTTTTGGATAGTTAGGTGGT) and P2R (CCCCCCGATCCAATTAACAACCCTAAAATATAA). Reverse primers are designed to preferentially anneal with deaminated cytidines; the forward primers anneal to the replicated mutations. **B** PCR amplification of a section of p219 plasmid untreated (AID neg) or incubated with five AID orthologs or GST at their optimal temperature (25 °C for GST) twice for 1 hour each. To amplify AID-mutated sequences, p219 plasmid incubated with Hs-AID, Pm-CDA1, and GST was annealed to deamination-specific primers at 52 °C; Lc-, Tn-, and Gc-AID reactions were annealed at 51°C. PCR-amplified bands (black arrow) indicate successful dC→dU deamination.

3.3.4 Fish AID orthologs are cold adapted

Previous studies using cell lines transfected with fish AID, as well as *in vitro* enzyme assays with purified fish AID, demonstrated that these orthologs mediate higher mutation rates at colder temperatures than 37 °C, suggesting that AID orthologs could have unique thermosensitivity preferences.[129,183] To evaluate this, I measured the activity of each AID on TGCbub7 (**Figure 2.3A**) at 18 temperature increments ranging from 4 °C to 40 °C (**Figure 3.5A**). In order from earliest to most recently diverged orthologs, Pm-CDA1 is the coldest adapted out of all five species tested, with an average optimal temperature of 14.5 °C (**Figure 3.5B**). Gc-AID and Tn-AID are both relatively cold adapted, with average optimal temperatures of 20.5 °C and 20.0 °C, respectively. Lc-AID is the warmest adapted of the fish AIDs, preferring an optimal temperature of 25.0 °C. Hs-AID is the warmest adapted of all the AID orthologs tested, with an optimal temperature of 30.8 °C (**Figure 3.5B**). Based on these data, I conclude that these fish AID orthologs are active cytidine deaminase enzymes that are colder adapted, one with as much difference in optimal temperature as 15 °C compared to Hs-AID.

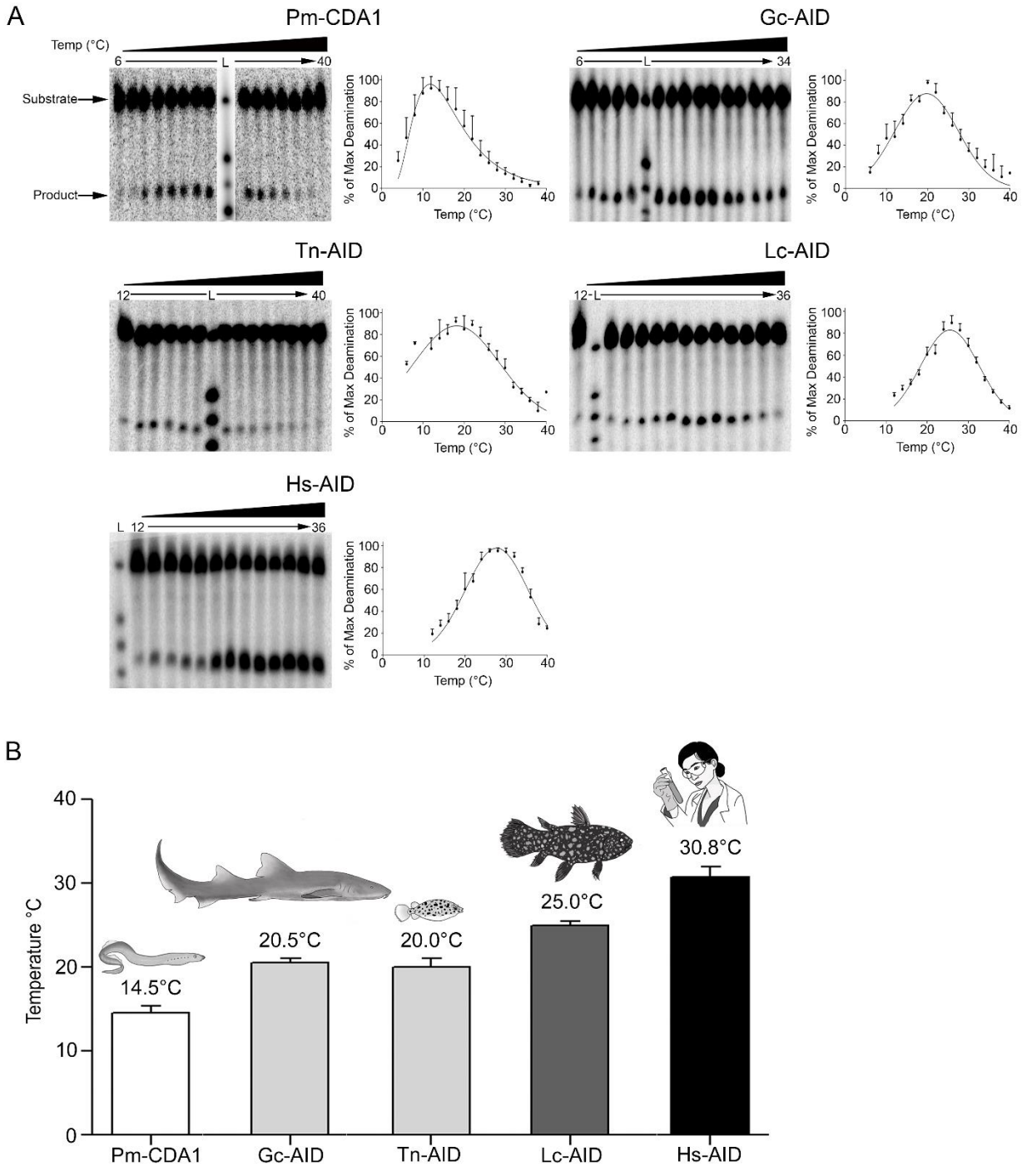


Figure 3.5 Fish AID orthologs are cold adapted compared to Hs-AID

Adapted with permission from Figure 3 in Quinlan et al.*, 2017.*[229]

Legend: Figure 3.5

A Representative AID activity alkaline cleavage gels and data from thermosensitivity assays. Gels show AID incubated with TGCbub7 for temperatures ranging from 4 °C to 40 °C. Graphs present thermosensitivity curves, the peak of which indicates the AID ortholog's optimal temperature. The y axis shows the percentage of maximum deamination. Error bars represent standard deviations (SD). L, ladder. **B** Graph of average optimal temperatures for each AID ortholog, indicated by name in the x axis, with each organism illustrated and the final average optimal temperature above each bar. Each average is determined from three to six individual experiments using two to four independently purified preparations expressed from two or three independently constructed expression vector clones for each AID ortholog, resulting in, from left to right, seven, four, five, twelve, and five unique experiments, respectively. Error bars represent the standard errors of the mean (SEM).

3.3.5 AID orthologs have unique substrate sequence specificities

Hs-AID preferentially mutates dC in signature trinucleotide WRC motifs.[227,238–240,248,303–307] To determine the sequence preference of the earlier-diverged AID orthologs, I tested them using the alkaline cleavage assay on six different substrates that are identical to TGCbub7 (**Figure 2.3A**) except for the –2 and –1 position nucleotides immediately upstream of the target dC. **Figure 3.6A** demonstrates the WRC specificity of Hs-AID. As expected, Hs-AID is three-fold more active on the WRC substrates (on TGCbub7, TACbub7, and AGCbub7: 62.0%, 47.0%, and 42.7% deamination, respectively) than non-WRC substrates (on GACbub7, GTCbub7, and GGCbub7: 10.7%, 18.0%, and 27.1% deamination, respectively). We previously demonstrated that both Hs- and Dr-AID preferred the three tested WRC motifs at similar rates, whereas Ip-AID was three-fold more active on TGCbub7 than on all other WRC and non-WRC motifs tested.[183] Here, I found that Pm-CDA1 was four-fold more active on TGCbub7 and TACbub7 than the non-WRC GGCbub7 and GACbub7; however, Pm-CDA1 also unexpectedly showed 2.5-fold activity on the non-WRC GTCbub7 compared to the WRC AGCbub7 (**Figure 3.6B**). Both Gc-AID and Tn-AID exhibited a statistically significant preference for TGCbub7 over all other motifs tested (three- and eight-fold, respectively). Lc-AID also exhibited a slightly skewed WRC preference: it showed five-fold more activity on the WRC motifs tested compared to GTCbub7 and GACbub7; however, it was also five-fold more active on the non-WRC GGCbub7 compared to the other non-WRC substrates (**Figure 3.6B**).

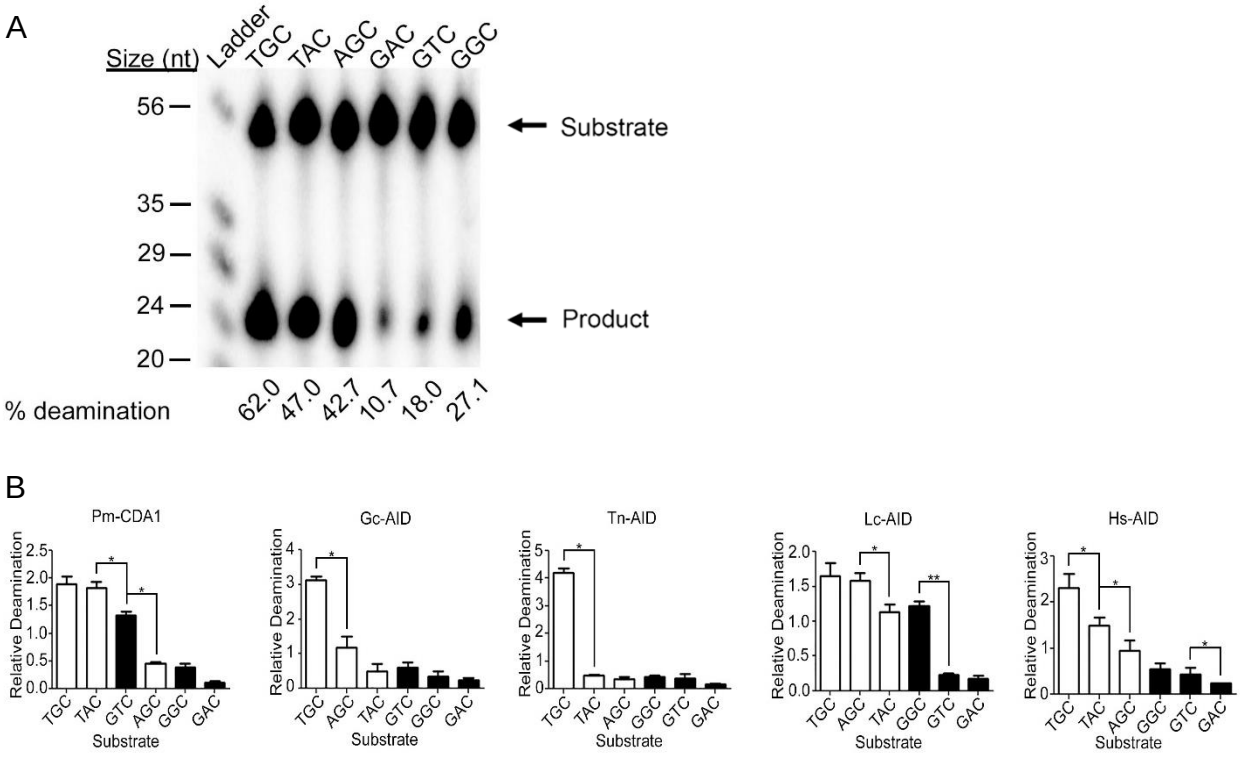


Figure 3.6 AID orthologs exhibit unique substrate specificity

Adapted with permission from Figure 4 in Quinlan et al.*, 2017.*[229]

A Representative alkaline cleavage gel of the substrate specificity assay, showing activity of Hs-AID on six bubble substrates (TGCbub7, TACbub7, AGCbub7, GACbub7, GTCbub7, and GGCbub7). **B** Bar graphs showing substrate specificity of each AID ortholog on six bubble substrates with the following target sequence: TGC, TAC, AGC, GGC, GTC, and GAC. Two or three purifications of one or two clones from each AID orthologs were tested two or three times for a total of three to five unique experiments for each species. The y axis shows relative deamination efficiency (preference relative to average of all six substrates) to enable comparison between AID orthologs whose absolute activity levels on each substrate vary. Error bars represent the SEM. *, $P \leq 0.05$; **, $P \leq 0.005$. *P* values were determined using one way ANOVA test with Bonferroni's Multiple Comparison Test.

The alkaline cleavage assay thus revealed the relative preference of each AID on three WRC and three non-WRC-bearing oligonucleotide substrates (**Figure 3.6A** and **B**); however, to examine the complete sequence specificity profile of each AID ortholog, I utilized the aforementioned deamination-specific PCR assay, as the plasmid substrate contains dC in all 16 possible trinucleotide NNC motifs. To this end, PCR amplicons representing plasmid DNA mutated by each of the five AID orthologs (**Figure 3.4B**) were sequenced, and the mutations were mapped (**Figure 3.7A**). Rates ranged from 0.1 to 6 mutations on average per 100 bp of the PCR amplicon, with all orthologs mediating either comparable or lower mutation loads than Hs-AID (**Error! Reference source not found.**). Pm-CDA1 mutated cytidines in an NYC motif (N= any nucleotide; Y= C or T) at higher-than-average rates compared to other motifs (**Figure 3.7B**). Gc-AID mutated cytidines in the AAC and ACC motifs at four- and three-fold above average, respectively, while not mutating any C in a trinucleotide motif that begins with G. Tn-AID preferred the AAC motif, followed by AGC, CGC, and TGC, with the GNC motifs among the least mutated. Lc-AID mutated CCC motifs two-fold above average, followed by AGC and CAC.

Complete agreement between the alkaline cleavage and PCR assays is not expected due to major differences in the nature of the substrates, one of which is an oligonucleotide with a single target dC and the other a plasmid with many dCs located in different regions. Nevertheless, when the substrate specificity data gleaned from the PCR deamination assay is focused on the same six motifs studied in the alkaline cleavage assay, the similarities between the two results become more apparent (**Figure 3.7C**). The WRC versus non-WRC preferences of Gc-, Tn-, and Hs-AID are maintained, with all three AID orthologs preferring TGC, TAC, and AGC over GTC, GGC, and GAC. The unique preferences of Pm-CDA1 for GTC and Lc-AID for GGC (along with their preferred WRC motifs) are also conserved.

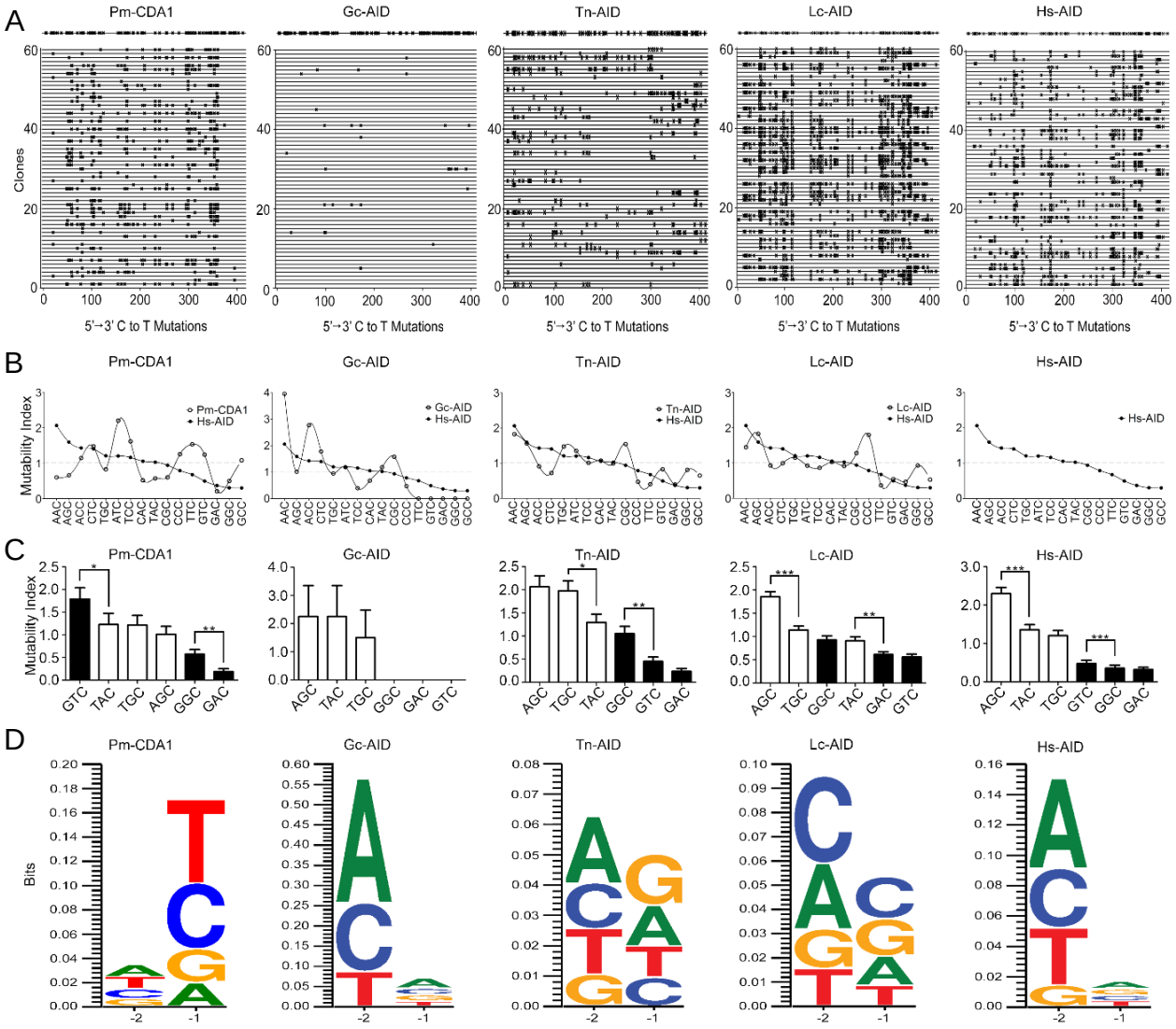


Figure 3.7 PCR assay confirms unique substrate specificity of AID orthologs

Adapted with permission from Figure 4 in Quinlan et al.*, 2017.*[229]

Legend: Figure 3.7

A Visualization of mutations. Each horizontal line is one individual sequenced amplicon of p219 that was mutated by the indicated AID ortholog, and PCR amplified; each "X" is a dC→dT mutation. The x axis maps the section of p219 that is PCR amplified, minus the primers: 407 nucleotides from the first set of primers. The line at the top of each graph shows all possible dC mutations, indicated by an X. The numbers of amplicons analyzed for each AID ortholog were as follows: Pm-CDA1, 61; Gc-AID, 82; Tn-AID, 103; Lc-AID, 112; and Hs-AID, 114. **B** Comparison of substrate specificity of Hs-AID to fish AID orthologs on p219. The y axis shows the mutability index, where 1 is the average rate of mutation (dotted line) for all 16 NNC motifs. The relative preference for each individual NNC sequence was obtained by dividing its mutation rate by the average value for all 16 NNCs. The x axis shows XXC DNA motifs, ordered from most to least mutated by Hs-AID. Error bars represent the SD. **C** Bar graphs showing the substrate specificity of each AID ortholog on six NNC motifs—TGC, TAC, AGC ("hot spots", white bars) and GGC, GTC, and GAC ("cold spots", black bars)—in a PCR-based assay. The y axis shows the mutability index. The relative preferences for the six NNC sequences were obtained by dividing the mutation rate by the average value for the six NNC motifs identified. *, $P \leq 0.05$; **, $P \leq 0.005$, ***, $P \leq 0.0001$. P values were determined by using a Mann-Whitney test (nonparametric, two-tailed). Error bars represent the SEM. **D** Sequence logos showing the relative AID specificity of each nucleotide at the –2 and –1 nucleotide positions. The height of the stack shows the consensus of nucleotides at that position, and the height of each symbol within the stack indicates the relative frequency of each nucleotide at that position.

Table 3.1 Mutation rates of AID orthologs on denatured plasmid DNA

AID ortholog	Avg dC mutations (per amplicon)	Mutation rate (per 100 nt)
Pm-CDA1	13	4
Gc-AID	0.4	0.1
Tn-AID	8	2
Lc-AID	22	6
Hs-AID	16	4

Adapted with permission from Table 1 in Quinlan et al.*, 2017.*[228]

Avg, average; dC, deoxycytidine; nt, nucleotides

The corresponding sequence logo analysis of the PCR assay results depicted in **Figure 3.7D** agrees with the initial sequence preference analysis. Pm-CDA1 is again unique among the other orthologs tested, since the –2 position in the XXC motif is not as relevant as the –1 position, showing equal weight for all four nucleotides in the –2 position while T and C in the –1 position are most preferred. In contrast, Gc-AID, Tn-AID, Lc-AID, and Hs-AID demonstrate a greater preference for certain nucleotides in the –2 position rather than the –1 position. Overall, I conclude that each AID ortholog has unique substrate sequence specificity patterns, with lowest preference for GNC trinucleotides being the only ubiquitous feature among three of the four jawed vertebrate orthologs.

3.3.6 All earlier-diverged AID orthologs have relatively low catalytic rates

I sought to assess the catalytic rates of AID orthologs by comparing initial Michaelis-Menten enzyme velocities, in the same manner previously used to demonstrate Hs-AID's lethargic catalytic rate.[227] To determine the optimal duration of incubation with substrate for each AID ortholog, which will be used later to measure initial catalytic velocities, I first performed time course enzyme kinetics comparisons. Each AID ortholog was incubated with TGCbub7 (**Figure 2.3A**) at its optimal temperature for durations between 5 and 1,200 minutes (**Figure 3.8A**). The substrate TGCbub7 was chosen because I found in this and previous works that it is the most preferred target of all AID orthologs in the alkaline cleavage assay, which is used to determine enzyme kinetics.[183,184] Following a similar pattern to the initial activity test seen in **Figure 3.3B**, I found that Hs-AID was the most active, catalyzing the maximum deamination of the second most active AID ortholog, Pm-CDA1 (42 fmol/ug at 1,200 min), more than three times as fast. Tn-AID was as active as Pm-CDA1 until 360 min, after which Tn-AID catalyzed 20% less product than Pm-CDA1 at 1,200 min. Gc-AID activity plateaued at 360 min, after which it catalyzed 0.5 fmol/µg product every hour until it reached a low maximum of 17.5 fmol/µg at 1,200 min, which is three-fold less than the activity of Hs-AID at this time. Lc-AID catalyzed a maximum of 10 fmol/µg at 1,200 min, approximately one-fifth the product catalyzed by Hs-AID at this time point.

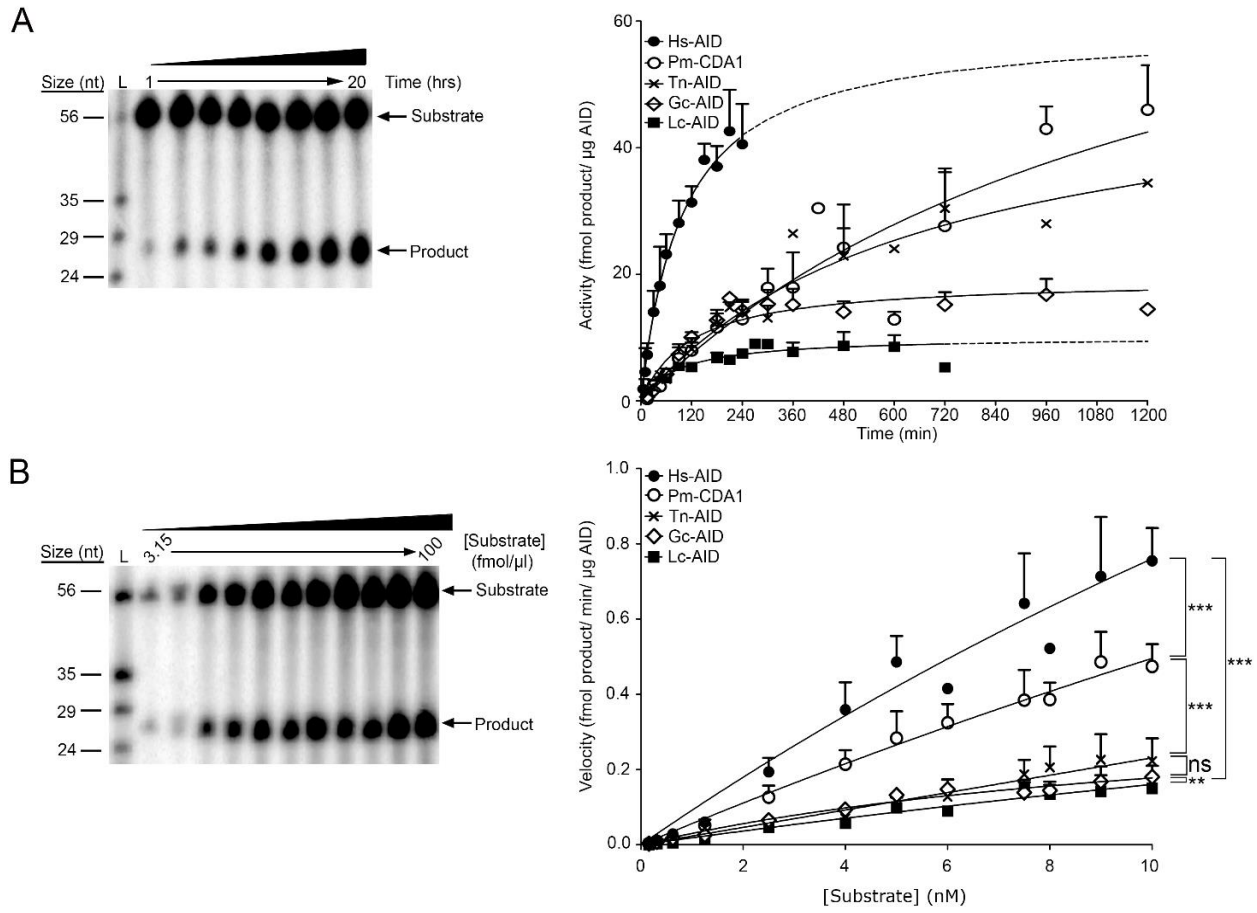


Figure 3.8 Fish AID orthologs exhibit lower enzymatic efficiency compared to Hs-AID

Adapted with permission from Figure 5 in Quinlan et al., *2017.*[229]

Legend: Figure 3.8

A, left Representative time course alkaline cleavage gel. Pm-CDA1 was incubated with TGCbub7 at optimal temperature (14 °C) for 1–20 hrs. **Right** Graph of combined time course data from two to seven experiments using three to six independently purified preparations of each AID ortholog; each point on the graph represents four to ten individual data points. Error bars represent the SD. **B, left** Representative Michaelis-Menten enzyme kinetics alkaline cleavage gel used to determine the initial enzymatic velocity of each AID ortholog. A 0.2 µg portion of Pm-CDA1 was incubated with concentrations of TGCbub7 ranging from 3.15 to 100 fmol/ul at the optimal temperature (14 °C) for 7 hours. **Right** Graph of Michaelis-Menten kinetics showing initial velocities of the different AID orthologs. Each point represents four to six individual data points from three independent experiments on three to six independently purified preparations of each AID ortholog. Error bars represent the SD. **, $P \leq 0.005$; ***, $P \leq 0.0005$, determined by two-tailed, nonparametric t test (Wilcoxon signed rank test).

The optimal length of incubation for comparison of Michaelis-Menten kinetics between AID orthologs was determined based on the duration of the initial velocity phase of time course kinetics: Pm-CDA1, 240 to 420 min; Gc-, Tn-, and Lc-AID, 150 to 300 min; and Hs-AID, 120 min. I thus measured the catalytic kinetics of each AID ortholog at its optimal time and temperature on a range of TGCbub7 concentrations (**Figure 2.3A** and **Figure 3.8B**). Unsurprisingly, the order from highest to lowest initial catalytic rate followed the same trend as time course enzyme kinetics in **Figure 3.8A**: Hs-AID exhibited the highest catalytic rate with an initial velocity of 0.75 fmol/µg/min. Pm-CDA1 is 33% less active than Hs-AID, with an initial velocity of 0.50 fmol/µg/min. Tn-AID and Gc-AID had lower and similar initial velocities, between 0.18 and 0.22 fmol/µg/min, nearly a quarter the rate of Hs-AID. Similar to the time course kinetics, Gc-AID and Lc-AID were again the slowest, having initial velocities of 0.18 and 0.15 fmol/µg/min, respectively, one-fifth that of Hs-AID. Thus, based on time course and initial catalytic rate comparisons, I conclude that all orthologs studies have a lower catalytic rate than Hs-AID. Interestingly, the earliest-diverged ortholog (Pm-CDA1) is the most active of the fish orthologs.

3.3.7 All AID orthologs have high binding affinity to their substrate

An important intrinsic variable that affects enzymatic activity is binding affinity; optimal enzymatic activity requires not only effective initial binding of the enzyme to its substrate, but also disassociation from the substrate after processing the substrate, to free the enzyme to act on other substrates. Binding affinity can be affected by temperature, salt concentration, and other variables that affect the structure and mobility of the protein and its substrate, as well as the intermolecular interactions between them.[308]

The EMSA allows for analysis of enzymatic binding affinity to its substrate; when performed on a serial dilution of substrate concentration, a dissociation constant (Kd) can be calculated. Enzymes with a small dissociation constant have high binding affinity to its substrate and may have slower enzymatic activity. Enzymes with a large dissociation constant may also demonstrate slow activity due to an inherently low binding affinity with the substrate, thus requiring higher concentrations of substrate. In any case, analysis of an enzyme's binding affinity to its optimal substrate can further elucidate relationships between the enzyme's biochemical properties and its function. In the case of AID and its orthologs, comparison of their dissociation constants and enzymatic activities can elucidate evolutionary changes, explain genetic and cellular regulatory mechanisms, and further characterize these understudied proteins.

To ascertain whether the differences in catalytic activity between the AID orthologs are due to differences in affinity of binding to the substrate ssDNA, I performed EMSAs previously optimized for measuring affinity of overall AID binding to DNA (**Figure 3.9A**, left).[183,226,227] The average Kd values (in nM) were as follows: Pm-CDA1, 0.11; Gc-AID, 0.14; Tn-AID, 0.58; Lc-AID, 0.044; and Hs-AID, 0.14 (**Figure 3.9A**, right).

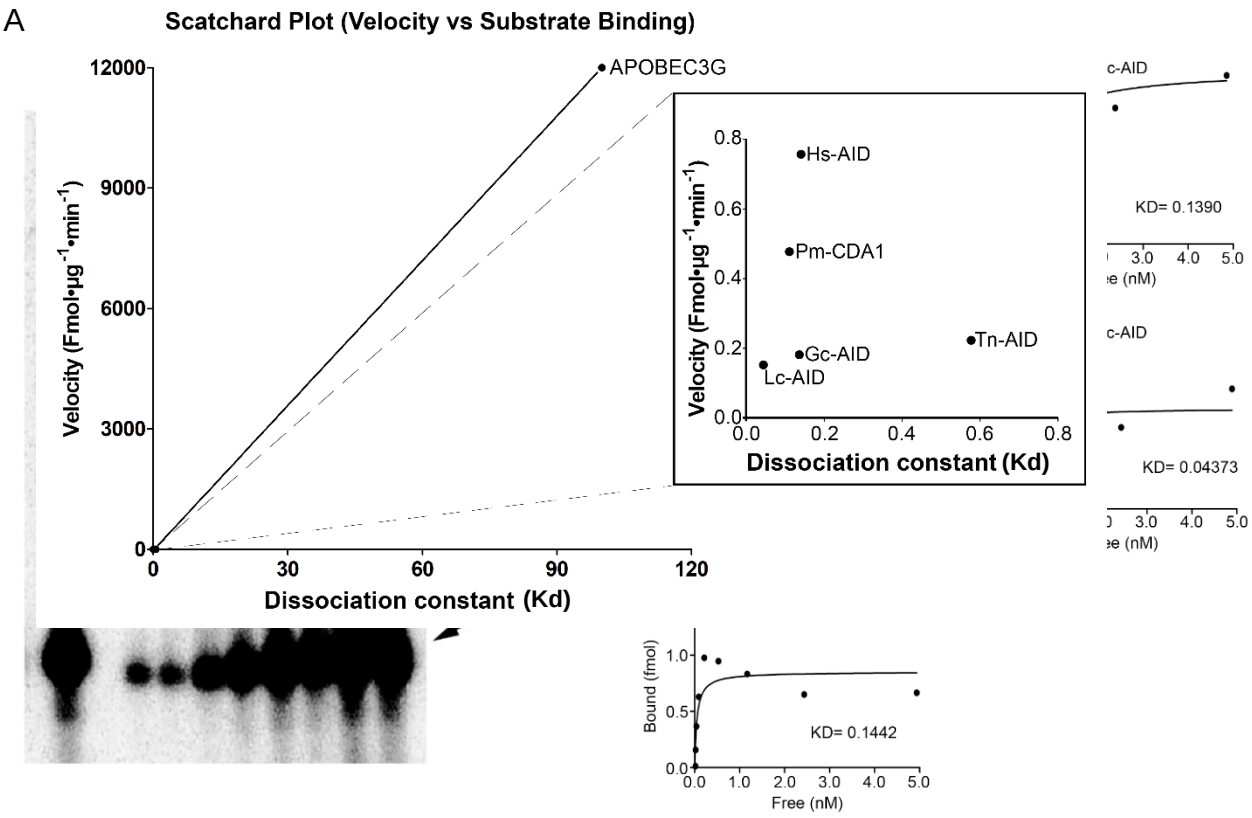


Figure 3.9 Extremely high substrate binding of AID orthologs is correlated to low enzymatic activity compared to APOBEC family member

Adapted with permission from Figure 6 and Supplementary Figure 2 in Quinlan et al.*, 2017.*[229]

Legend: Figure 3.9

A, left Representative EMSA gel of Hs-AID incubated with various concentrations of TGCbub7 at optimal temperature (31 °C) for 1 hour. AID in complex with substrate is found in the bound bands, while free substrate continues to migrate toward the bottom of the gel, as indicated. The negative control is TGCbub7 in binding buffer with no AID added. **A, right** To calculate Kd half-saturation binding affinities for each ortholog, the fraction of shifted substrate was quantified, and a bound versus free plot was generated. The average Kd value for each AID ortholog is indicated within each graph, determined from three independently purified preparations of each AID ortholog. **B** Scatchard Plot of Velocity versus Binding Affinities of the five AID orthologs studied in this section compared to those of human APOBEC3G (determined by Chelico *et al*[309,310]). The insert graph is a zoomed-in view of the position of AID orthologs on the graph.

The differences in binding affinities were not directly correlated with variation in catalytic activity (see **Figure 3.9B,** insert); however, since all ssDNA dissociation constants were extremely low (in the nM range), together they correspond to slower activity, compared to APOBEC3G, another member of the APOBEC family, that exhibits μM range ssDNA dissociation constants, and a much faster catalytic rate (**Figure 3.9B**).[309,310] Thus, I conclude that all AID orthologs examined here are able to bind ssDNA efficiently with very high affinities in the nM range, similar to Hs-AID.

3.3.8 Predicted structures of AID orthologs provide insight into biochemical differences

To gain structure-function insights of the different AID orthologs, we modeled each AID's structure using several related and solved APOBEC structures as the templates, as previously described (**Figure 3.10**, first column).[63,252] We examined the 25 lowest-energy predicted conformations for each ortholog and found that all formed a core structure akin to the conserved core architecture of the AID/APOBEC family.[63] Consistent with our finding that these orthologs are active cytidine deaminases, they all exhibited an enzymatically viable catalytic pocket at the centre of which are the characteristic triad α-β-α Zn-coordinating motif of two cysteines and a histidine atop the proton-donating catalytic glutamic acid (H66, E68, C97, and C100 for Pm-CDA1; H56, E58, C87, and C90 for Gc-AID; H53, E55, C93, and C96 for Tn-AID; H56, E58, C87, and C90 for Lc-AID; H56, E58, C87, and C90 for Hs-AID).

We noted that each predicted AID structure exhibited a high concentration of surface-exposed, positively-charged residues distributed over the entire surface in locations proximal to the catalytic pocket, as we previously described for Hs-AID.[63] Consequently, all AID orthologs exhibited similar isoelectric properties (pI values for Hs-, Lc-, Tn-, and Gc-AID and Pm-CDA1 of 10.0, 9.7, 10.1, 10.2, and 10.0, respectively) and charges at neutral pH (net charges for Hs-, Lc-, Tn-, and Gc-AID and Pm-CDA1 of +14.3, +11, +15.9, +11.7, and +12.7, respectively). This conservation is remarkable since there is significant divergence in pI and surface charges within the AID/APOBEC family (–2, –6, +0.5, –9, and –3.5 of human A3A, A3B-CTD, A3C, A3F-CTD, and A3G-CTD, respectively), thus making the net charge of approximately +14 a defining feature of AID.[252] The conservation in high positive net charge is also consistent with our observation that all AID orthologs bind ssDNA with high-nM-range affinity (**Figure 3.9**, right).

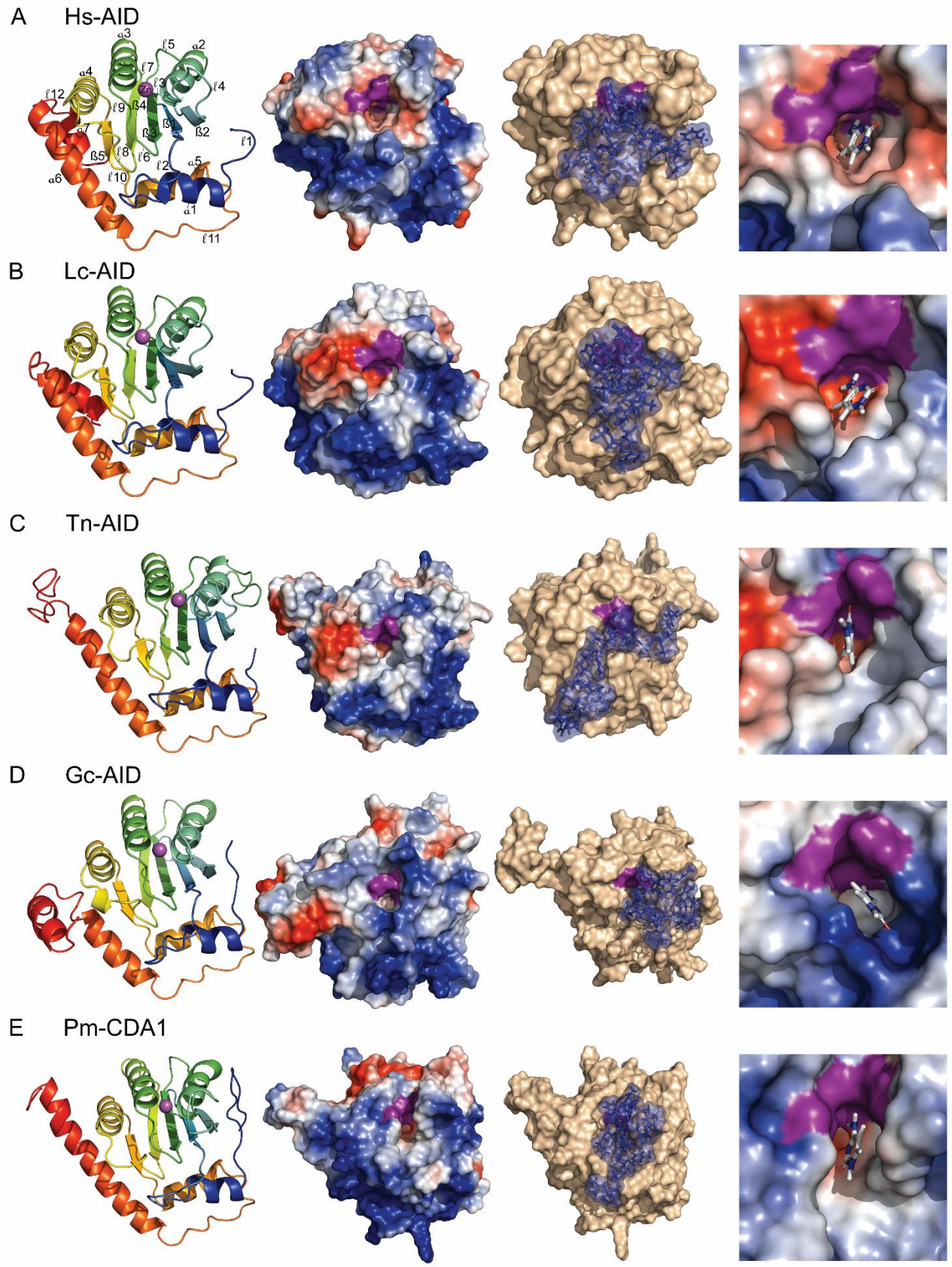


Figure 3.10 Predicted structures and substrate docking of AID orthologs

Legend: Figure 3.10

Adapted with permission from Figure 7 in Quinlan et al., *2017.*[228]

Shown are the cartoon (first column), surface topology (second column), surface model with single-stranded section of TGCbub7 in the DNA binding groove (third column), and catalytic pocket docked with dC (fourth column) of A, Hs-AID; B, Lc-AID; C, Tn-AID; D, Gc-AID; and E, Pm-CDA1. Each structure is a representative of the 25 lowest-energy predicted conformations, using multiple related APOBEC X-ray and NMR structures as the templates. N-to-C terminus progression is shown from blue to red in the cartoon structures. The ribbon diagram shows Zn coordinated in the catalytic pocket (purple sphere): ℓ is loop, α is helix, and β is strand. All models illustrate the same global architecture with notable differences found in the N-terminal tails, connecting loops (ℓ2, ℓ4, ℓ5, ℓ6, and ℓ8), particularly in the loop 5 extension that is unique to teleost fish, and in the structure/absence of the C-terminal α7. In the surface structures, positive and negative residues are blue and red, respectively. The catalytic pocket is seen as an indentation in the centre, with the catalytic residues (Zn-coordinating triad of cysteines and histidine, catalytic glutamic acid) coloured purple. A large proportion of the surface is positively charged, as reflected by the relatively high pI and net charge at neutral pH compared to other APOBEC enzymes, and similar to Hs-AID. The fourth column shows catalytically accessible conformations of each AID ortholog docked with dC in the catalytic pocket. All orthologs exhibited both open and closed catalytic pocket confirmations, similar to Hs-AID, and a representative open pocket confirmation was chosen for dC docking, since closed conformations are unable to dock dC in the pocket and thus represent an inactive conformation of AID.

Since I found that all AID orthologs examined had varying but consistently low catalytic robustness, we compared regions that govern the catalytic pocket and substrate specificity (**Figure 3.10**, first and second columns). We previously described a set of secondary catalytic residues that support the aforementioned primary catalytic residues by stabilizing dC in the catalytic pocket and contribute to the physical architecture of the catalytic pocket.[63] The secondary catalytic residues are contained in loops 2, 4, 6, and 8 in Hs-AID; relative breathing motion of these loops contributes to the fluidity of Hs-AID's catalytic pocket dynamics.[63,252] We previously demonstrated that these loops contain much of the sequence divergence within the human AID/APOBEC family and that these differences mediate variable breathing dynamics of catalytic pockets.[252] Here, we found that the AID orthologs also had significant differences in these loops. Pm-CDA1 had a three-residue insert and a replacement of two normally conserved secondary catalytic residues (R32 and C33) in loop 2. Nevertheless, the conformation of loop 2 remains similar to other AID and APOBECs. We noticed a large insert in loop 4 of Pm-CDA1 that compacts the residues neighboring the catalytic pocket. We also noted considerable sequence differences in loop 8 among all AID examined, with Pm-CDA1 having the shortest loop 8. In Tn-AID, we observed a unique sequence in loop 5, which is conserved in AID from ray-finned fish. This sequence contains a high number of helix-breaking residues and was predicted to form an extension of loop 5. Furthermore, this extension contains a high number of hydrophobic residues and was found to mediate contact with α2 and the N terminus, which is likely important for dimerization.

We previously demonstrated that the catalytic pocket of AID transitions dynamically between accessible and occluded states and that only the former can accommodate dC for deamination.[63] Therefore, we sought to compare the efficiency of dC binding in the catalytic

pocket of the AID orthologs. To this end, we docked a 7-nucleotide ssDNA containing the WRC motif TGC to the surface of AID and measured the proportions of ssDNA bound near the catalytic pocket and how many of those conformations were catalytically viable (**Figure 3.10**, third and fourth columns; **Table 3.2**). As previously demonstrated for Hs-AID,[63] blind docking simulations showed sporadic binding of the ssDNA oligonucleotide on the surface, with a minor proportion of substrate bound proximal to the catalytic pocket: 18% (59/320), 39% (112/288), 29% (56/192), 42% (161/384), and 44% (99/224) of DNA bound clusters for Hs-AID, Lc-AID, Gc-AID, Tn-AID, and Pm-CDA1, respectively.

We then restricted the docking to the catalytic pocket and surrounding region; for Hs-AID we found 58% of low-energy binding clusters resulted in DNA bound to the DNA binding grooves. For Tn-AID, Gc-AID, Lc-AID, and Pm-CDA1 we found 63% (507/800 clusters), 62% (498/800 clusters), 74% (590/800 clusters), and 72% (572/800 clusters) of low-energy ssDNA binding clusters bound to the DNA binding grooves located on the surface, respectively. For Tn- and Gc-AID, the proportions were similar to Hs-AID, but slightly higher in Lc-AID and Pm-CDA1. These higher proportions might be attributed to a slightly more favourably shaped DNA-binding groove. We also observed that like Hs-AID, each ortholog had only a minority proportion of catalytic pockets in an open and accessible conformation able to dock dC (40% [10/25], 36% [9/25], 24% [6/25], 48% [12/25], and 28% [7/25]) for Hs-, Lc-, Tn-, and Gc-AID and Pm-CDA1, respectively (**Table 3.2**)). This observation is consistent with our demonstration here that all fish AID orthologs have a low catalytic rate and suggests that catalytic pocket occlusion is a crucial aspect of AID regulation that has been preserved during its divergence.

Table 3.2 Percent substrate bound to AID near the catalytic pocket and correctly oriented within the DNA binding groove

AID ortholog	% Substrate bound near the catalytic pocket		% Catalytically viable states
	Not restricted	Restricted[α]	
Hs-AID	18 (59/320)	58 (466/800)	40 (10/25)
Lc-AID	39 (112/288)	74 (590/800)	36 (9/25)
Gc-AID	29 (56/192)	62 (498/800)	24 (6/25)
Tn-AID	42 (161/384)	63 (507/800)	48 (12/25)
Pm-CDA1	44 (99/224)	72 (572/800)	28 (7/25)

Adapted with permission from Table 2 in Quinlan et al.*, 2017.*[228]

[α]To within 30 by 30 Å of the catalytic pocket

To address the structural basis of the variable substrate sequence specificities, we simulated TGC- and GTC-containing ssDNAs binding to Pm-CDA1 and Hs-AID. Our docking simulations reflected our *in vitro* results: we found that Hs-AID preferred TGC versus GTC (2.1% versus 0.9% of DNA complexes containing dC in the catalytic pocket), while Pm-CDA1 has a greater preference for GTC (2.1% versus 1.1% of DNA complexes containing dC in the catalytic pocket). These results not only explain our *in vitro* finding that Pm-CDA1 prefers GTC, the opposite preference to that of Hs-AID, but also serve as further validation of our AID:ssDNA structural prediction strategy.

3.4 Discussion

Previous works have mainly focused on testing ray-finned fish AID in cell-based reporter assays, determining that these AID orthologs can initiate tetrapod-exclusive CSR in mammalian cells.[159,178,179,181] This study is the first to compare purified versions of the earliest-diverged known AID orthologs. I found that all four fish AID orthologs are active cytidine deaminases, supporting the theory that the downstream effects of AID mutation, like CSR and SHM, are likely more dependent on substrate availability, such as the concurrent exposure of single-stranded Ig switch-regions during transcription and transportation of AID into the nucleus of the activated B cell.

I also found that while each fish AID demonstrated various levels of activity, all AID orthologs share the exceptional biochemical properties of Hs-AID: relatively slow catalytic rates and high-nanomolar-range ssDNA binding affinities.[227,228,311] These characteristics are consistent with predicted shared structural features of frequently occluded catalytic pockets and highly positively charged surfaces, respectively. We have shown that these two structural features of Hs-AID mediate sporadic ssDNA binding on its surface and infrequent positioning of dC in the catalytic pocket that, together, limit the activity of AID. Our finding here that even the most divergent orthologs studied share these biochemical and structural features points to the evolutionary importance of this safeguard.[63]

In support of our finding that AID orthologs are lethargic *in vitro*, the relatively slow catalytic activity of wild-type AID has also been demonstrated *in vivo*: AID mutants that were more catalytically active than their wild-type Hs-AID counterpart were shown to increase antibody diversification and chromosomal translocations in B cells, suggesting that AID has evolved a suboptimal specific activity to avoid possibly tumourigenic DNA damage.[279] To confirm our

finding that AID's lethargy is consistent across species as a protective evolutionary trait against DNA damage, it will be critical to develop similar *in vivo* assays using higher-activity mutants of nonhuman AID orthologs.

The AID orthologs studied here show similarly high nanomolar range binding affinities, but also unusually low enzymatic activity. An enzyme requires a minimum binding strength to have any activity, but if the bound complex between product and enzyme is too strong, the product may remain bound too long and act like a competitor, decreasing activity. Indeed, when a related enzyme with a much higher specific activity, like APOBEC3G, is considered for comparison, an inverse relationship between specific activities and binding affinities emerges (see **Figure 3.9B**). In this context, it is remarkable that even though the amino acid sequence of Pm-CDA1 is more different from Hs-AID than other human APOBECs are from Hs-AID, Pm-CDA1 has maintained the high binding affinity and low catalytic activity common to the other AID orthologs, further supporting the evolutionary importance of these biochemical properties.

Furthermore, as suggested in previous works, [129,183,184] I found that purified fish AID are more cold-adapted than their human counterpart, indicating that AID itself has adapted to changes in body temperature. Adult lamprey inhabit waters at 20 °C,[312] nurse sharks prefer 25 °C to 30 °C,[313] tetraodon are found in tropical freshwater environments at 26 °C,[314] coelacanths prefer deep water at 13 °C to 25 °C,[315] and humans keep an internal temperature of 37 °C: these physiological temperatures are only slightly warmer than the optimal temperatures of the associated AID. The concurrence of the optimal temperature of an enzyme to the physiological temperature of the species is referred to as "corresponding states," where orthologous proteins from species living at different temperatures are in corresponding stable tertiary structures and exhibit similar thermodynamic stability at their respective optimal temperatures.[316–318] Cold-adapted proteins are

often more flexible than their warm-adapted counterparts so as to adopt a similar level of structural rigidity at lower temperatures.[318,319]

I demonstrated that Lc-, Gc-, and Tn-AID is less catalytically active than Pm-CDA1 and Hs-AID, thus we were surprised not to observe gross differences in apparent compaction of predicted AID ortholog structures. Rather, structure prediction suggests other reasons for variation in enzymatic activity. For example, in Tn-AID there is an extension of loop 5, which is unique to AID from ray-finned fish (not shown). This modified loop 5, enriched in proline/glycine residues, effectively bulges loop 5 away from the core toward the β2 face. Based upon the tetrameric structure of APOBEC2, AID has been proposed to dimerize at this interface, which involves interactions from 2 and loop 5.[320] Given the context of this insert, we suspect this loop 5 insert in Tn-AID may alter the dimer structure, which in turn may have ramifications on its other biochemical properties, such as activity. Furthermore, in Pm-CDA1 we noticed a distinct structure of the N-terminal tail in comparison to other AID structures examined. While the N-terminal in other AID structures forms a short extension, which makes minor contact with β2, the N-terminal in Pm-CDA1 forms an extended conformation that maintains significantly more contact with β2 and partial contact with α2. This extended N-terminal tail effectively shields β2, analogous to the full-length structure of APOBEC2. Unlike the N-terminally truncated APOBEC2 which forms a tetramer, the full-length APOBEC2 structure exists solely as a monomer because the extended N-terminal tail, analogous to Pm-CDA1, effectively shields β2-β2 dimerization. This suggests that the N-terminal extension in Pm-CDA1 may result in shielding of the dimerization interface, possibly leading to prevalence of monomers, which would explain our observation of its relatively higher specific activity. The predicted models also reveal differences in the substrate specificity loops which explain the unique substrate preferences of these AID orthologs.

It has been hypothesized that Pm-CDA1 is responsible for initiating gene conversion between the placeholder untranslated region of VLRA/C and the leucine-rich repeats, which form the completed transcript.[126,151,286] Unlike VLRB, which is predicted to perform the same actions in cyclostomes as Ig in gnathostomes, VLRA and VLRC are theorized to be more similar to T cell receptors. Therefore, it is possible that the unique preference for Pm-CDA1 for non-WRC motifs (NYC) is a reflection of coevolution with its target substrate (VLRA/C). Based on our work, we suggest differences present in loops 2 and 8 of Pm-CDA1, as well as the overall compacted state of the catalytic pocket and surrounding region due to a four-residue insert (G60–R63) pushing loop 4 closer toward the catalytic pocket, are responsible for its unique substrate preference. Concurrent with Ig loci appearing as the primary antigen recognition receptor in the shark family, AID seems to have evolved a preference for substrates with WRC motifs, its substrate specificity pattern becoming more like that of Hs-AID as the evolutionary distance between the species closes. Given the involvement of loop 8 in substrate recognition in Hs-AID and other APOBECs, it was not surprising to observe considerable differences of loop 8 between the AID orthologs.

This study is the first to carry out an *in vitro* comparison of specific biochemical properties in purified earliest-diverging orthologs of a tumourigenic DNA-mutating enzyme. I discovered that the AID orthologs found in these fish species maintain the uniquely low enzymatic rate and high-affinity DNA binding of their human counterpart and that, despite structural differences that lead to various optimal temperatures and DNA substrate sequence preferences, the three defining regulatory aspects of the structure are remarkably conserved across species: an abundance of high positive surface charge, catalytic pocket inaccessibility, and frequently catalytically-nonviable ssDNA binding. Together, these features inherent to AID are responsible for its lethargic catalytic activity, which I demonstrated is also conserved throughout evolution. Our results provide strong

evidence that the inherent structural features that limit AID's "dangerous" mutagenic activities are biologically significant, as they are remarkably conserved across evolution.

Chapter 4 Lamprey has an expanded and diverse repertoire of enzymatically distinct AID-like enzymes

4.1 Authorship Statement

Data described in Section 4.2 are published in "Expansions, diversification, and interindividual copy number variations of AID/APOBEC family cytidine deaminase genes in lampreys" in the Proceedings of the National Academy of Sciences of the United States of America (PNAS) in 2018.[127] Last co-authors T Boehm and M Larijani designed and supervised the research. Authors SJ Holland, LM Iyer, K Sikora, S Peter, F Sugahara, P Shingate, I Trancoso, N Iwanami, E Temereva, C Strohmeier, S Kuratani, B Venkatesh, G Evanno, L Aravind, and M Schorpp performed DNA extraction, whole genome sequencing, generation of contig assemblies, phylogenetic comparison of APOBEC genes, RT-PCR, transcriptome assembly, RNA sequence expression analysis, and the bacterial resistance assays. LM Berghuis expressed and purified the CDA variants, performed the protein activity assays, and analyzed the resulting data. JJ King performed structure prediction modeling and prediction of the biochemical characteristics of the CDA variants. SJ Holland, M Larijani, and T Boehm analyzed the data and wrote the paper. I assisted in data analysis and designing and creating figures 2, 3A–G, S3B, S8, and s9.

In the subsequent sections, LM Berghuis expressed and purified the CDA variants. JJ King performed the structure prediction and modelling of the variants and their biochemical characteristics. David Hubert predicted the biochemical characteristics of some CDA and AID variants. I designed and performed the cytidine deaminase activity assays, analyzed the data, performed the literary research, and wrote the chapter. M Larijani supervised the work and edited this chapter.

4.2 Introduction

The first report of AID-like enzymes in the lamprey identified two different orthologs: Pm-CDA1, which is a 208 amino acid (aa) protein with cytidine deaminase activity in cellular reporter assays, and Pm-CDA2, a longer 331 aa protein without any discernable cytidine deaminase activity in the reporter assays used. No other "pillars" of an adaptive immune system were found, such as MHC, RAGs, or T/B cell receptors.[126] However, lymphocyte-like cells were discovered, with receptors similar to T and B cells; these were similar to TLRs in shape, had many leucine-rich repeats, and were named Variable Lymphocyte Receptors.[230]

Another factor that unites both jawed and jawless vertebrates is somatic rearrangement of their lymphocyte receptors: mature VLR transcript sequences do not match their genomic counterparts, suggesting they are somatically rearranged. Each VLR gene contains four exons which together house the 5' untranslated region (UTR), a signal peptide, the 3' end of the LRRCT (the most 3' LRR segment), and the C terminal, which contains a glycosylphosphatidylinositol (GPI) membrane-anchorage motif, with slight variations depending on the type of VLR and the species. The LRR cassettes that are copied into the non-coding middle region of the mature VLR are found both up- and downstream of the main gene (see **Figure 4.1**).[126,230,321] It is hypothesized that CDA1 and CDA2 initiate the gene conversion events that create the VLR transcript by mutating cytidines located in the non-coding regions of the VLRA/C and VLRB genes, respectively, in a process similar to how AID mutation initiates gene conversion events at immunoglobulin loci in sheep, rabbits, and chickens.[322] In such a gene conversion process, after the DNA double strand break, DNA repair mechanisms would use the variable LRR segments as templates to replace the mutated DNA in the non-coding region through nonreciprocal homologous recombination.[321,323]

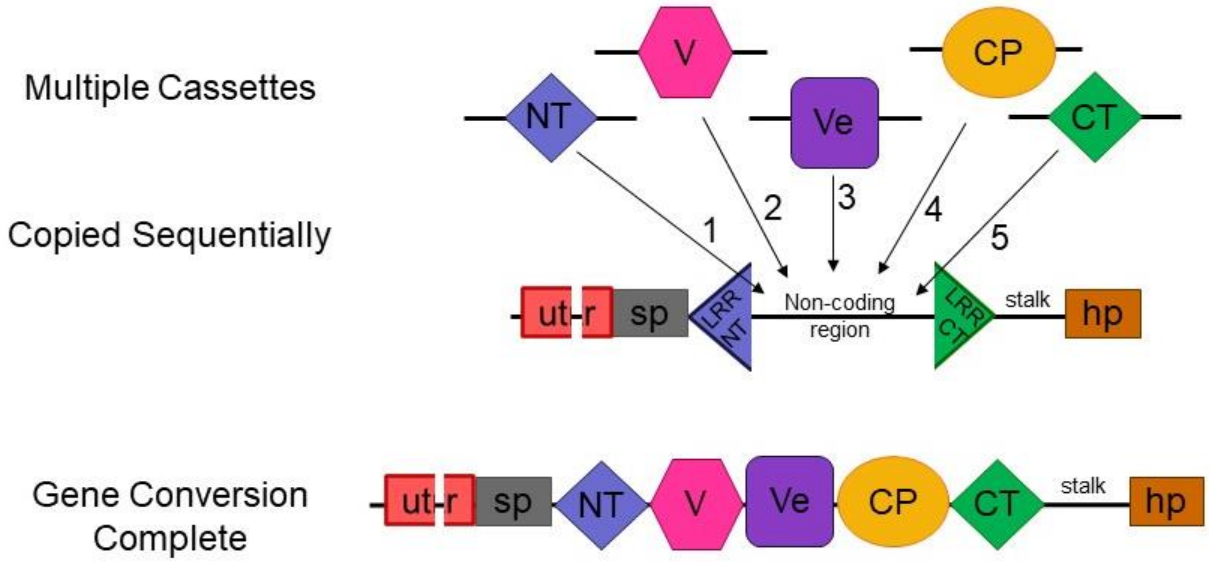


Figure 4.1 Somatic rearrangement of the lamprey variable lymphocyte receptor

Adapted from Figure 2, Box 7-2 in Owen et al., *2013*[1] *as non-commercial user-generated content.*

The Variable Lymphocyte Receptor (VLR) loci is arranged in the genome with the untranslated region (UTR), signal peptide (sp), N-terminal (NT) Leucine-rich Repeat (LRR), a non-coding region, the C-terminal (CT) LRR, a stalk, and a hydrophobic peptide (hp). It is hypothesized that CDA1/2 mutate cytidines at the either end of the non-coding region; the resulting uridine is removed and the abasic site repaired through DNA repair mechanisms. Often these mechanisms caused single- or double-stranded breaks, which signals gene conversion (GC) to begin. GC uses various cassettes found both up- and downstream of the main cassette as templates to fill in and repair the DNA break, essentially replacing the non-coding region with NT (N-terminal), V (variable), Ve (end variable), connecting peptide (CP) and CT cassettes. These cassettes are copied sequentially, one after the other, and may be copied whole or partially, creating diversity in the final transcript. Diversity is also created via different numbers of V LRR modules and different combinations of LRR modules, the latter of which is similar to V(D)J recombination.

P. marinus, discussed in Chapter 3, is one of ~ 40 species of lamprey, which, along with hagfish, are the only remaining members of the jawless fish superclass Agnatha, i.e., cyclostomes.[280] Lamprey can be divided into parasitic and non-parasitic species. The parasitic sea lamprey is arguably the most-studied species, classified as "invasive" in the North American Great Lakes. Sea lamprey spend over half their life as ammocoetyes (larvae) living on freshwater riverbeds feeding off detritus, afterwards undergoing metamorphosis to a sexually-immature parasitic form, migrating to larger bodies of fresh or salt water to feed on the blood of larger animals. Once sexually mature, the sea lamprey cease to eat, returning to their ammocoetye riverbed to spawn and die. Non-parasitic lamprey, such as the European brook and Japanese lampreys, skip the juvenile parasitic phase and metamorphosize almost directly from ammocoetyes into sexually-mature adults, gaining teeth yet losing the ability to eat, and remaining near their birth sites to spawn.[324]

During embryogenesis, lamprey lose ~ 20% of their genome from their somatic cells.[325] This process is called programmed DNA elimination or programmed genome rearrangement (PGR, see **Figure 4.2**) and occurs in ~ 10% of all vertebrates, including the cyclostomes, some shark species (holocephalans), bandicoots, spiny mice, and Passeriformes (song birds).[326–328] In lamprey, the eliminated DNA is composed of transposable elements, satellite repeats, and genes. These eliminated genes are responsible for organism development and have homologs in mice that are largely silenced by the polycomb repressive complex (PRC).[325,328] The PRC plays a protective role, silencing genes that, while instrumental during development, become detrimental once the animal has reached reproductive age (e.g., oncogenes), a phenomenon called antagonistic pleiotropy.[329,330] PGR is believed to play a similar role, eliminating genomic content that has surpassed its usefulness, protecting the soma while maintaining the integrity of the germline.[326,331]

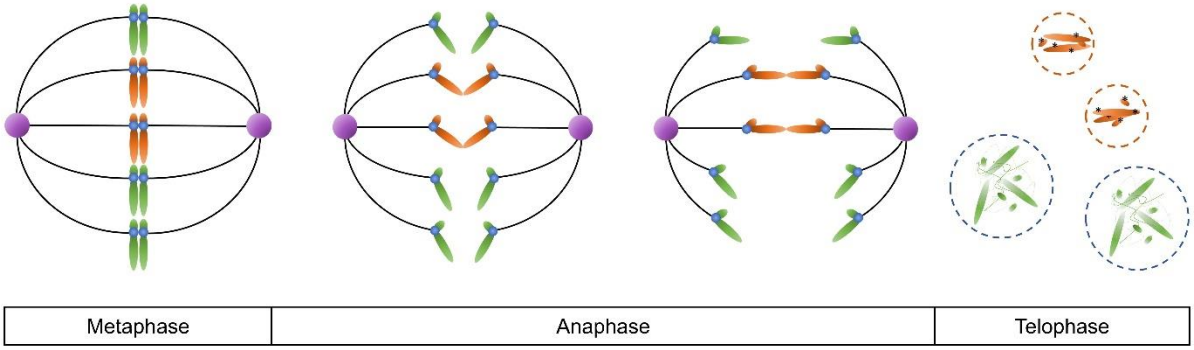


Figure 4.2 Mechanism of programmed genome rearrangement in the lamprey

Adapted with permission from Smith et al., *2021*.[326]

Diagram illustrating the proposed mechanism of programmed genome rearrangement (PGR) in lampreys. At the sixth cell division, duplicated chromosomes align centrally along the metaphase plane (left). During anaphase (centre left), as the chromosomes are pulled away from the metaphase plane via spindle fibres (black lines) towards the spindle poles (purple spheres), germline-specific chromosomes (orange) remain closer to the metaphase plane than somatic chromosomes (green). Germline-specific sister chromosomes are prevented from separating due to an interaction at their terminal ends, where repetitive DNA elements are located (*Germ2* and *Germ6*; centre right). During telophase (right), retained chromosomes form nuclei (blue dashed circles), while germline-specific chromosomes form small micronuclei (orange dashed circles). These micronuclei are enriched in trimethylated H3K9 and accumulate 5-methylcytosine (black stars), targeting the germline-specific DNA for degradation.

The ability to manage different pH, temperature, and salinity levels; their vastly different life stages; the fact their genome undergoes PGR during embryogenesis; combined with their unique place in vertebrate evolution highlight the lamprey as a fascinating model organism for studying evolutionary biology and immunology. Thus far, multiple investigations have found evidence of parallels between lamprey and gnathostome immune systems. For example, Hirano *et al*. found that the third lymphocyte lineage ($VLRC^+$) not only expressed genes similar to γδT cells (such as Sex determining Region Y [SRY]-related High Mobility Gene [HMG] 13 [*SOX13*] encoding a γδT cell fate-determining transcription factor), but was also localized to lamprey epithelial tissue where they were dendritic in shape, similar to mouse dendritic epithelial T cells.[151,332] Another example is work by Li, Wu, and associates, who studied lamprey complement component C1q, an initiator of the mammalian classical pathway, which activates C3 in lamprey. This group demonstrated that lamprey C1q mediates complement-dependent cytotoxicity via interaction with secreted VLRB *in vitro,*[333] while their *in vivo* C1q-knockdown study suggested C1q is involved in $VLRA/C^+$ T cell-like response; these interactions are analogous to the influence of gnathostome C1q over humoral and cellular immunity.[5,334] Study of lamprey innate immunity revealed further similarities: for example, a lamprey NF-κB homolog with a 3D structure and *in situ* activity similar to its human counterpart[335]; and the presence of leukocyte autophagy in response to antigen stimulation, similar to the response in mice and other vertebrates.[336]

The aforementioned work is a small portion of what has been investigated, and much is still not yet known about the immune system of this extant agnathan. My work, and that of my colleagues, described in the third chapter added to this growing field by highlighting the unique biochemical characteristics of the sea lamprey CDA1 enzyme, which includes the first case of non-WRC sequence preference shown in an AID ortholog while demonstrating the evolutionarily

conserved AID characteristics of high substrate binding and relatively low catalytic activity. Our data suggests that, as an active AID-like enzyme expressed in T cell-like cells, CDA1 could play a mutagenic-diversification role in the sea lamprey cellular adaptive immune system. This chapter describes our previously published discovery and initial analysis of new freshwater lamprey CDA genes, research that is expanded in the second half of the chapter describing our as-yet unpublished work.

4.3 Expansions, diversification, and interindividual copy number variations of AID/APOBEC family cytidine deaminase genes in lampreys

To identify AID homologs in freshwater lamprey species, genomic DNA was extracted from whole larvae and testes of European brook lamprey *Lampetra planeri* (Lp) and Japanese lamprey *Lethenteron japonicum* (Lj), respectively, these two species being closely related non-parasitic cousins to the comparatively better studied and parasitic *P. marinus* species. This DNA was then sequenced and searched for Pm-CDA1 and Pm-CDA2 orthologs. Close genetic homologs of CDA1 and CDA2 were found in both species, along with, unexpectedly, multiple CDA1-like genes that could be divided into two distinct groups (CDA1L1_1, _2, _3, _4 and CDA1L2_1, _2) (see **Figure 4.3A**). Genomic DNA from other individuals were searched for homologs of these new CDA1-like genes, which were found, along with splice variants of CDA1L1_1 and CDA1L1_3.[127]

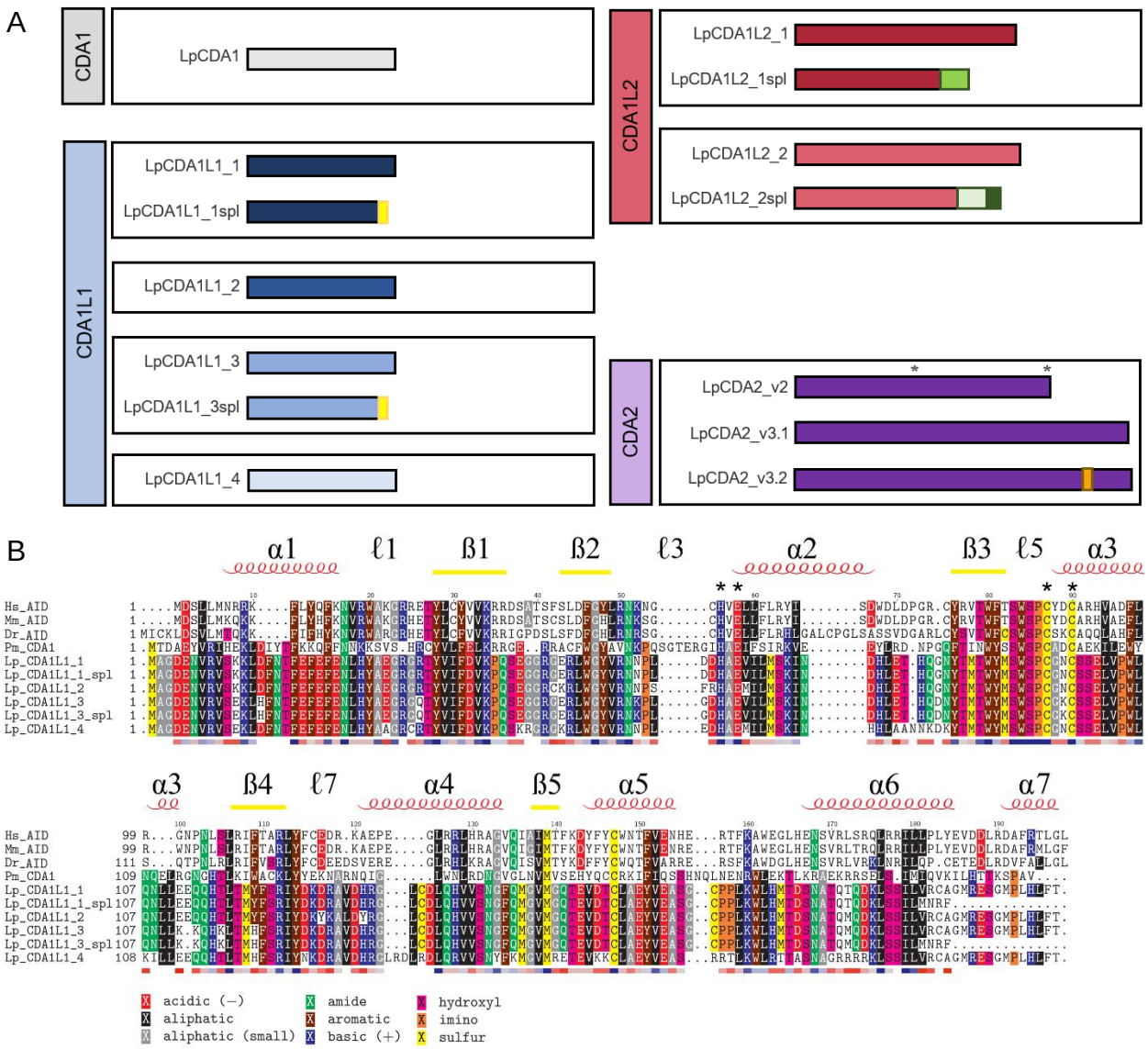


Figure 4.3. CDA-like genes in *Lampetra planeri*

Adapted with permission from Figure 2 in Holland et al.*, 2018.*[127]

Legend: Figure 4.3

A Representative schematic of all CDA-like genes found in the *L. planeri* fish. Each gene is colour-coded: different shades of the same colour denote similar genetic sequence, same shades of the same colour symbolize exact sequence: CDA1, grey; CDA1L1, blue; CDA1L2, red; and CDA2, purple. The LpCDA1L1_1spl and LpCDA1L1_3spl 3' end sequences are genetically exact (yellow). The 3' ends of LpCDA1L2_1spl and LpCDA1L2_2spl are genetically similar (different shades of green). All LpCDA2 genes are genetically exact, except for two point mutations (stars) and truncation in LpCDA2_v2, and a short insert in LpCDA2_v3.2 (orange). **B** Alignment of amino acid sequences of human, mouse, and zebrafish AID (Hs-, Mm-, and Dr-AID, respectively) and CDA-like proteins from *L. planeri* (LpCDA1L1_1, _1spl, _2, _3, _3spl, and _4). α, alpha helix; β, beta sheet; ℓ, loop; *, zinc-coordinating motif residues.

When their amino acid sequences were compared with those of other AID orthologs (**Figure 4.3B**), these novel CDA-like proteins were found to contain the conserved deaminase core catalytic motif ($HxE(x)_nPCxxC$), suggesting they could be active cytidine deaminases. *In silico* modeling of each CDA ortholog also demonstrated their possible catalytic activity, as each protein formed a putative cytidine deaminase catalytic site (**Figure 4.4**).[127]

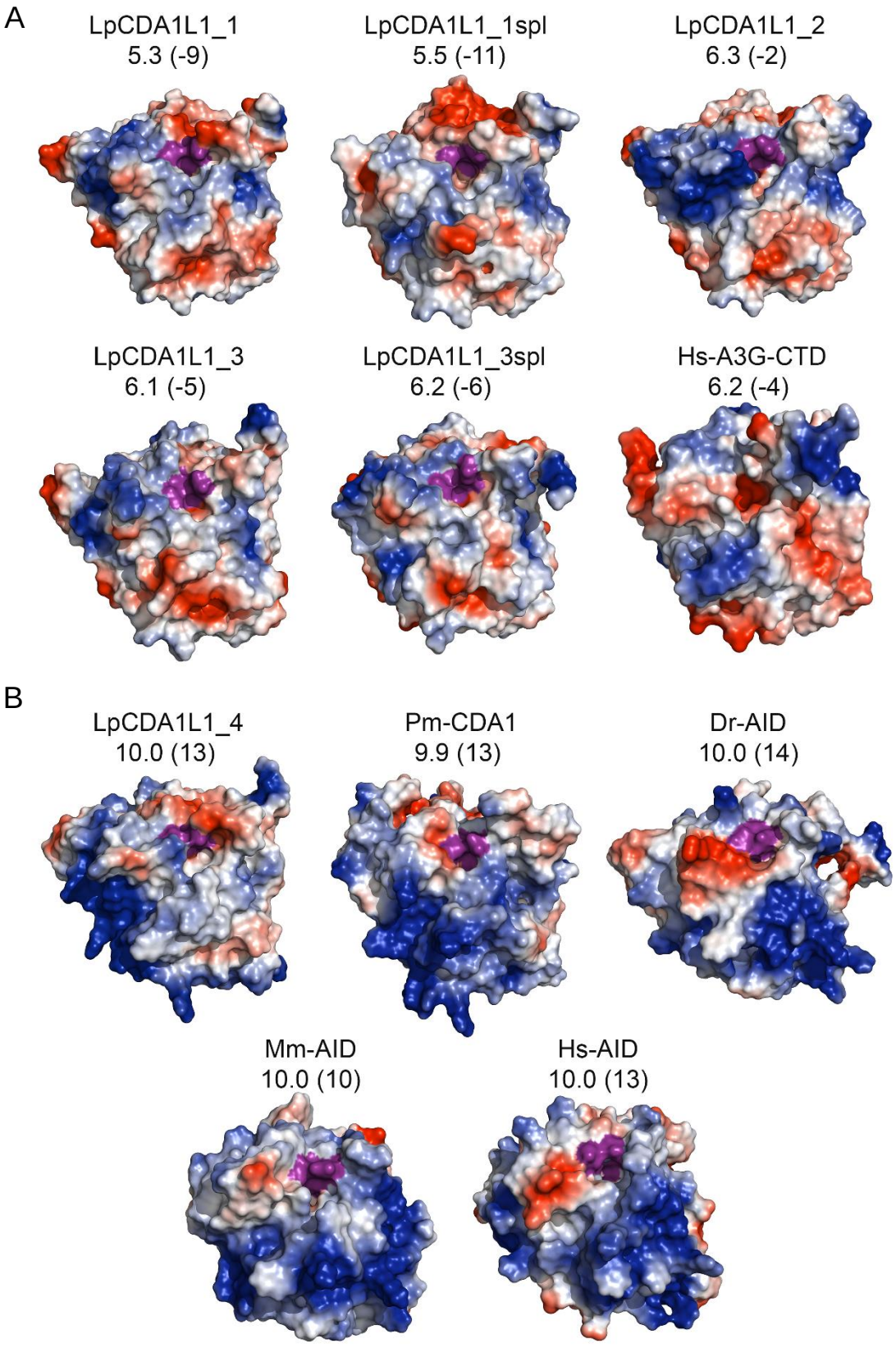


Figure 4.4. Comparison of isoelectric points and surface charges of APOBEC proteins

Legend: Figure 4.4

Adapted with permission from Figure 2 in Holland et al.*, 2018.*[127]

Representative surface topology models of the lamprey CDA1L1 proteins, compared with sea lamprey (*Petromyzon marinus*) CDA1 (Pm-CDA1), zebrafish (*Danio rerio*) AID (Dr-AID), house mouse (*Mus musculus*) AID (Mm-AID), human (*Homo sapien*) AID (Hs-AID), and the C terminal domain (CTD) of human APOBEC3G (Hs-A3G-CTD). Positively- and negatively-charged residues are coloured blue and red, respectively. The catalytic pockets are indicated as an indentation in the centre, with the Zn-coordinating triad (two cysteines and a histidine) and the catalytic glutamic acid coloured purple. Each protein's predicted isoelectric point (first number) and charge at neutral pH (second number, in brackets) is indicated above the associated surface structure. **A** Proteins with negative charge at neutral pH (mostly red surface residues). **B** Proteins with positive charge at neutral pH (mostly blue surface residues).

Another surprising finding was that each individual had a unique repertoire of CDA1L1 spliced and unspliced genes, differing in number and combination of genes (see **Table 4.1**). Due to the genetic variation between these CDA1L1 genes, the enzymatic activity of the resulting proteins could also vary; that each individual contained a unique repertoire of these genes suggests an evolutionary advantage to diversity of these genes, which could be linked to a variation in enzyme activity. Most intriguing, if one individual lacked a CDA1 gene, it had a CDA1L1_4 gene, and *vise versa*; in contrast, each individual could have all, some, or none of the remaining CDA1L1 genes. This suggests CDA1 and CDA1L1_4 play a similar, necessary role in the animal's survival, while the presence or absence of the other CDA1L1 genes is influenced by other factors, further suggesting a variation in the enzymatic activity or role of these proteins.[127]

Additionally, *in silico* structural modeling of all identified proteins found that the surface charge of CDA1L1_4 was more similar to that of Pm-CDA1 and AID from mouse, human and zebrafish (mostly positive at neutral pH), than that of the other CDA1L1 proteins, whose surface charges matched that of the APOBEC3 enzymes (negative at neutral pH; **Figure 4.4**).[127] While both are nucleotide mutators, AID and APOBEC3 play vastly different roles in immunity: the former diversifies and strengthens the antibody response, while the latter mutates viral genomes to render them ineffectual.[17,29,108] The differences between the surface charges of AID and APOBEC3 is reflective of their different subcellular localizations and enzymatic activities; the difference in the surface charges of the CDA1L1 proteins similarly suggests different enzymatic activities.[127] Taken together, these data suggested unique roles for the different CDA1L1 enzymes that could be further explored using an unbiased, simplified *in vitro* protocol, as described in the following sections.

Table 4.1 Combinatorial diversity of AID-like genes in the lamprey

	Lampetra planeri				***Lampetra fluviatillis***		***Lampetra japonica***		***Petromyzon marinus***	
	Lp #173	Lp #175	Lp #236	Lp #242	Lf #29	Lf #33	Lj #1	Lj #2	Pm #1	Pm #144
CDA1				X				X	X	
CDA1L1_1	X					X	X	?		
CDA1L1_2	X	X		X	X			?		
CDA1L1_3	X			X		X		?		
CDA1L1_4	X	X	X		X	X	X	?		X
CDA1L2_1	X	X	X	X			X	?	X	X
CDA1L2_2		X					X	?		

Adapted with permission from Figure 4 in Holland et al., *2018.*[127]

Inventory of CDA and CDA-like genes in individual fishes whose genomes were extracted, to demonstrate combinatorial diversity. Numbers in row 2 denote individual animal from which the genomic DNA was taken. CDA1 and CDA1L1_4 are highlighted in red. X, presence of gene; ?, unexplored presence of genes; Lp, *Lampetra planeri*; Lf, *Lampetra fluviatillis*; Lj, *Lampetra japonica*; Pm, *Petromyzon marinus*.

To enzymatically characterize these novel AID-like proteins in order to better understand their role in lamprey immunity, each CDA1L1 gene (LpCDA1L1_1, _1spl [spliced], _2, _3, _3spl, _4) was processed in two of our protein expression systems which use two different plasmids (pGEX-5X-3 and pcDNA 3.1-V5-HIS-TOPO) containing two different tags (GST and HIS, respectively) at different terminal ends of the protein (N and C termini, respectively) in different types of cells (prokaryotic and eukaryotic, respectively, the latter having more extensive post-translational modifications[337]). Expressing the same proteins in these two systems would certify if the enzymatic activity and characteristics were due to any external influences on the protein, such as the tag or post-translational modifications, or due to an inherent property of the enzyme. That being said, our previous works have demonstrated that, at least with human AID, the different expression systems and tags appear to have little influence on the characteristics of the enzyme.[43,68]

Contrary to the success seen with other AID orthologs, most of these new CDA1 orthologs could not be successfully cloned and purified using the prokaryotic expression system. The eukaryotic expression system was more productive, perhaps due to a more native protein production facilitated by post-translational modifications that are lacking in *E. coli* cells.[337] However, the resulting proteins were not purified as in the prokaryotic expression system. To release the C-terminally His-tagged proteins in this second system, the HEK293T cells are simply lysed; enzymatic activity declined drastically after purification, therefore, subsequent enzyme activity tests were performed on whole cell lysates containing the protein of interest. The previously studied human, zebrafish, and nurse shark AID (Hs-AID, Dr-AID, and Gc-AID, respectively) were cloned and purified in parallel with the CDA1L1 genes, acting as controls to ensure accurate protein production and to serve as positive controls in downstream activity assays. Accurate protein production was verified by western blot analysis (see **Figure 4.5**).[127]

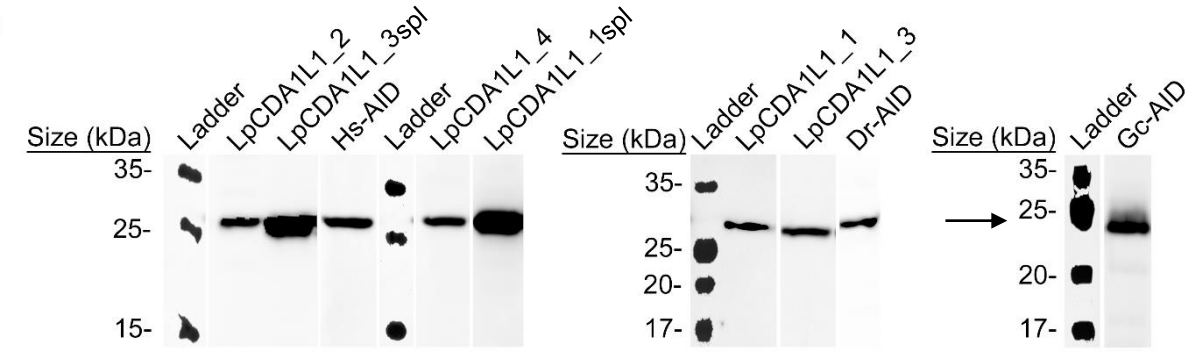


Figure 4.5. Western blots of LpCDA1L1 and control enzymes to test for accurate purification

Adapted with permission from Supplementary Figure 8 in Holland et al.*, 2018.*[127]

Western blots of His-tagged LpCDA1L1 orthologs and controls (Hs-, Dr-, and Gc-AID, human, zebrafish, and nurse shark AID, respectively) in whole cell lysate using anti-V5 antibodies. The ladders come from digitized versions of the same gels, which were inverted and superimposed onto the original gel. Empty and extraneous lanes containing irrelevant samples were removed for simplicity, denoted by the white space separating the cut gels. The sizes of each protein combined with the linker, V5 epitope, and 8xHis tag are thus, following the order each protein appears above: LpCDA1L1_2 (~ 29.0 kDa), LpCDA1L1_3spl (27.5 kDa), Hs-AID (28.6 kDa), LpCDA1L1_4 (~ 29.9 kDa), LpCDA1L1_1spl (~ 27.6 kDa), LpCDA1L1_1 (~ 28.9 kDa), LpCDA1L1_3 (~ 28.8 kDa), Dr-AID (~ 29.3 kDa), and Gc-AID (~ 28.5 kDa), the latter of which is running lower on the gel than expected, but at the appropriate size compared to a Hs-AID sample run on the same gel (not shown). The correct Gc-AID band is indicated by a black arrow to the left of the gel.

Cell lysate was tested for cytidine deaminase activity in our alkaline cleavage assay, alongside untransfected cell lysate negative controls (**Figure 4.6**). Initial activity tests included 16 hour incubations at three temperatures (14 °C, 22 °C, and 37 °C), three phosphate buffers of different pH's (5.9, 6.4, and 7.1), on 12 partially single-stranded, radioactively-labelled substrates, each bearing a different target sequence motif (TGC-, TAC-, AGC-, TTC-, GTC-, GGC-, AAC-, and GACbub7; ATC-, CTC-, and TTCbubB; and AGCbub11; see **Figure 2.3A**, **D**, **E**). This wide variety of parameters was necessary for the initial activity tests of these proteins due to their activity status and optimal conditions being unknown.

All LpCDA1L1 lysates appeared to be active cytidine deaminases on partially single-stranded substrates (**Figure 4.6A–D**), with the LpCDA1L1_2 and LpCDA1L1_4 lysate preparations demonstrating the lowest enzymatic activity (**Figure 4.6B**). The untransfected lysate and negative enzyme controls were indeed negative, therefore, any cytidine deaminase activity seen in the alkaline cleavage activity could confidently be concluded to represent *bona fide* activity of the tested LpCDA1L1 enzymes (**Figure 4.6D**).[127]

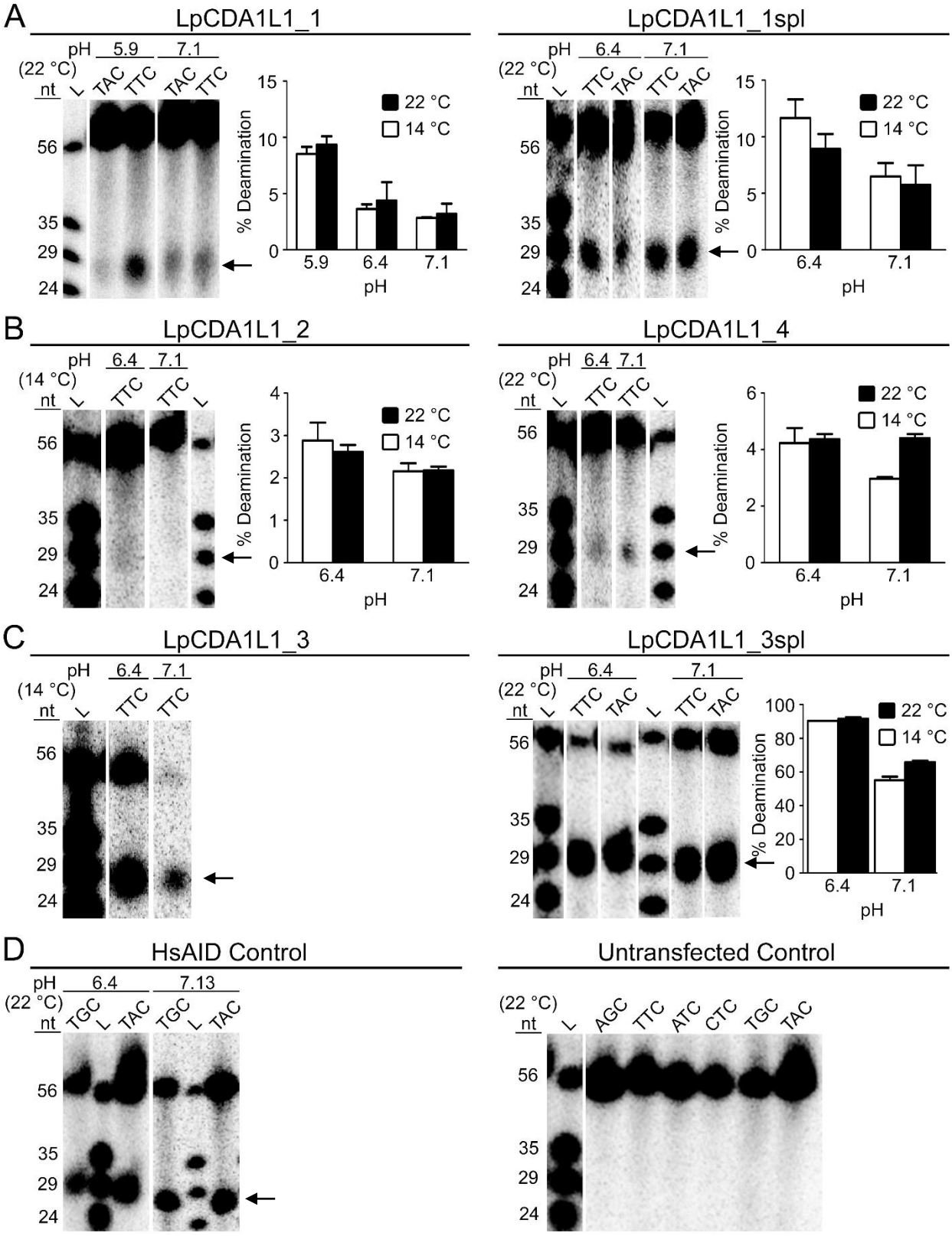


Figure 4.6. Initial activity tests of LpCDA1L1 orthologs and controls

Legend: Figure 4.6

Adapted with permission from Supplementary Figure 9 in Holland et al., *2018.*[127]

Representative alkaline cleavage gels and % deamination graphs of **A** LpCDA1L1_1 and LpCDA1L1_1spl, **B** LpCDA1L1_2 and LpCDA1L1_4, **C** LpCDA1L1_3 and LpCDA1L1_3spl, and **D** Hs-AID and untransfected control incubated 16 hours at 14 °C or 22 °C, at pH 6.4 or 7.1, with 12 substrates (TGC-, TAC-, AGC-, TTC-, GTC-, GGC-, AAC-, and GACbub7; ATC-, CTC-, and TTCbubB; and AGCbub11); only the darkest product bands are shown, indicated by a black arrow. The % deamination graphs are of the TTCbubB results of each ortholog. L, ladder. Gels C and D were unable to be quantified due to low-quality imaging of the gels. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

The optimal temperatures of the CDA1L1 enzymes ranged between 14 and 22 °C (see **Figure 4.6**). In line with the differing surface charges of each protein, in the limited range of pH's tested, the optimal pH of CDA1L1_4, along with Hs-AID and Dr-AID, was 7.1, while the other CDA1L1 enzymes showed optimal activity at pH 6.4 or 5.9 (CDA1L1_1 only; **Figure 4.6**). All CDA1L1 enzymes showed a preference for TTC (a non-WRC motif) and TAC (a WRC motif) (**Figure 4.7**). These results are similar to those seen for GST-tagged Pm-CDA1, which also demonstrated a substrate specificity of both WRC and non-WRC motifs, while most AID enzymes tested thus far have solely WRC preferences.[127]

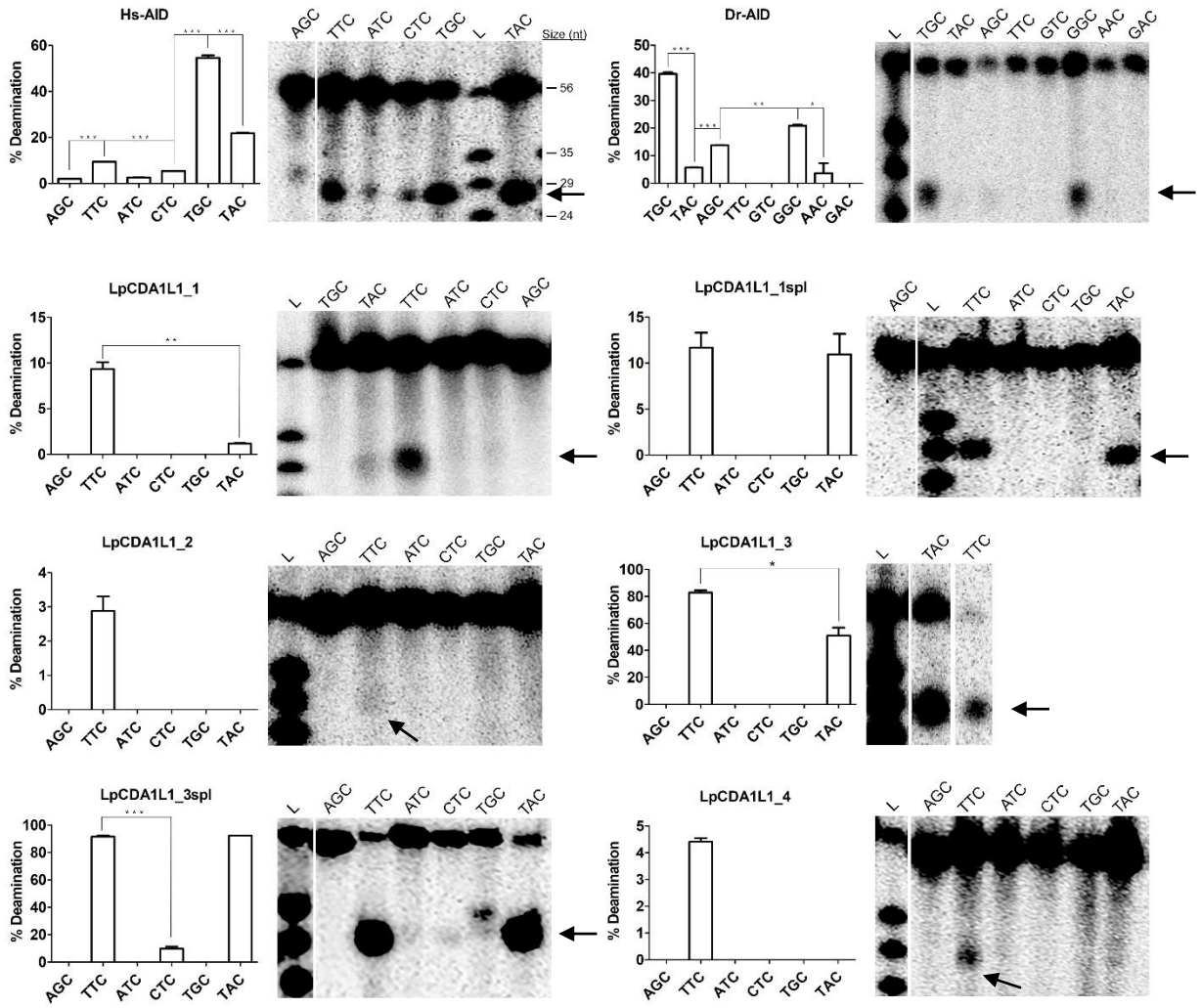


Figure 4.7. Substrate specificity profiles of LpCDA1L1 enzymes and controls

Legend: Figure 4.7

Adapted with permission from Figure 3 in Holland et al., *2018.*[127]

Combined results (graphs) of lysate, His-tagged control enzymes (Hs-AID and Dr-AID) and LpCDA1L1 orthologs, and associated representative alkaline cleavage gels. Substrates (whose products are indicated by a black arrow), in order of appearance on each gel: Hs-AID, LpCDA1L1_1spl, LpCDA1L1_2, LpCDA1L1_3spl, and LpCDA1L1_4→ AGCbub11, TTCbubB, ATCbubB, CTCbubB, TGCbub7, and TACbub7; Dr-AID→ TGC-, TAC-, AGC-, TTC-, GTG-, GGC-, AAC-, and GACbub7; LpCDA1L1_1→ TGCbub7, TACbub7, TTCbubB, ATCbubB, CTCbubB, and AGCbub11; LpCDA1L1_3→ TACbub7, TTCbubB. Error bars = SD. ***, $P < 0.005$; **, $P < 0.05$; *, $P < 0.05$. *P* values were determined using unpaired, two-tailed t test. The white lines denote areas of the imaged gel which were removed due to the lanes being empty or containing irrelevant samples.

These results suggest CDA1 enzymes from lamprey have unique substrate specificities compared to AID orthologs from jawed vertebrate species, which may be due to different genes targeted by jawless versus jawed vertebrate AID enzymes (VLR verses Ig genes, respectively). These data combined (see **Table 4.2**) suggest that these CDAs have unique biochemical characteristics worth exploring in more detail. In the remainder of this chapter, the thermosensitivity and optimal pH profiles of each CDA1L1 gene are further characterized in comparison to Hs-AID, Dr-AID, and Gc-AID, whose characteristics were detailed in the previous chapter and published in Quinlan *et al*. 2017.[229]

Table 4.2 Optimal conditions from initial activity tests of CDA-like enzymes

CDA-Variant	Max Deam. (%)	Optimal Inc. Temp. (°C)	Optimal pH of Activity Buffer	Optimal Substrate
CDA1L1_1	9.4	22	5.9	GTTCAGG
CDA1L1_1spl	11.7	14	6.4	GTTCAGG
CDA1L1_2	2.88	14	6.4	GTTCAGG
CDA1L1_3	82.79	22	6.4	GTTCAGG
CDA1L1_3spl	92.3	22	6.4	TGTACTT
CDA1L1_4	4.41	22	7.1	GTTCAGG
Human AID	53.51	22	7.1	TGTGCTT
Zebrafish AID	40.19	18	7.1	TGTGCTT

Adapted with permission from Figure 3 in Holland et al., *2018.*[127]

Deam., deamination; Inc., incubation; Temp., temperature. The optimal substrate sequences in the bubble are shown in the last column, the nucleotides in red are the target XXC sequences.

4.4 Results

4.4.1 Thermosensitivity profiles of eukaryotically-expressed, His-tagged controls

Before determining the optimal temperatures of the CDA1L1 variants, the thermosensitivity profiles of the eukaryotically-expressed His-tagged (lysate) AID controls (Hs-, Dr-, and Gc-AID) were compared with those of their prokaryotically-expressed GST-tagged (purified) counterparts to determine if there were any substantial differences in their enzymatic characteristics. Two unique preparations of each His-tagged AID control (derived from one clone each) were incubated for 4 hours with a ^{32}P radioactively labelled TGCbub7 (see **Figure 2.3A**) in phosphate buffer at pH 7.2 at temperatures ranging from 4 °C to 37 °C (**Figure 4.8A**). The optimal temperatures from each experiment were combined for each lysate and compared to previously-determined optimal temperatures from their purified counterparts (**Figure 4.8B**) — the data for prokaryotically-expressed, GST-tagged proteins can be found in Chapter 3 (Hs- and Gc-AID, **Figure 3.5**) and **Appendix 1** (Dr-AID).

The jawed vertebrate eukaryotically-expressed HIS-tagged AID proteins in lysate exhibited lower optimal temperatures than their previously reported, prokaryotically-expressed, purified, GST-tagged counterparts: Hs-AID (20.5 °C vs 30.8 °C), Dr-AID (14.0 °C vs 20.5 °C), and Gc-AID (13.5 °C vs 20.5 °C), though they follow the same general trend of fish AID being more cold-adapted than their human counterpart (**Figure 4.8**).[183,184,229]

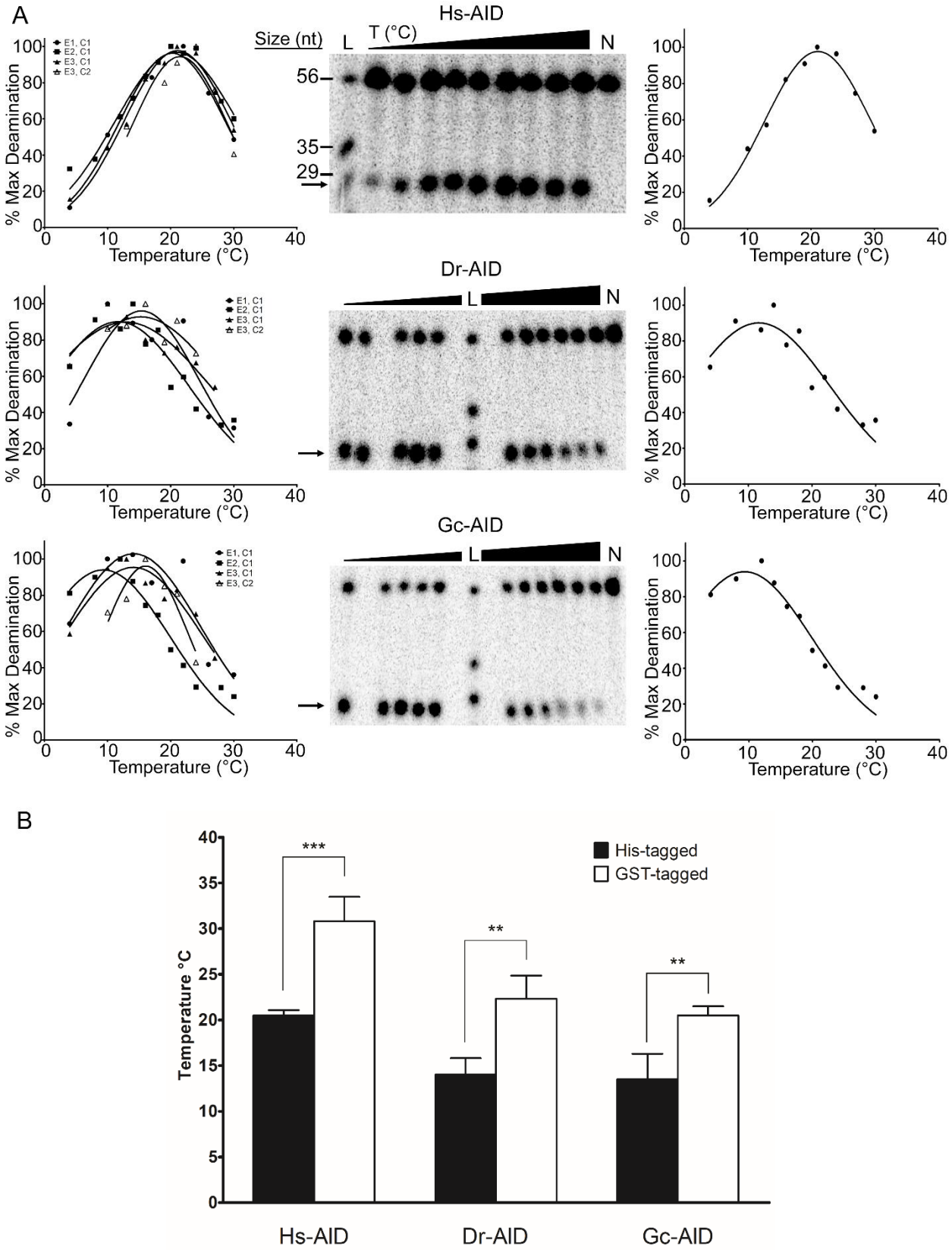


Figure 4.8. Optimal temperature profiles of His-tagged control enzymes in eukaryotic lysates are lower than their GST-tagged, prokaryotic purified counterparts

Legend: Figure 4.8

A Left Graphs present all thermosensitivity curves of each data set used to determine optimal temperature for each control AID enzyme, which was determined from the peak of each curve. The y axis show percentage of maximum deamination to correct for inter-preparation, -clone, and -experimental activity differences. The x axis show temperature range in °C from 4 °C to 30 °C. **Middle and Right** Representative gels (middle) and corresponding graphs (right) from thermosensitivity assays. Gels show AID incubated with TGCbub7 for temperatures ranging from 4 °C to 30 °C. L, ladder; N, negative control; T, temperature. The product band is indicated by a black arrow. **B** Graph of average optimal temperatures for each eukaryotically-expressed, His-tagged AID ortholog compared to that of their prokaryotically-expressed, GST-tagged counterparts. The x axis shows the name of each ortholog. Each average is determined from two to four unique experiments (E) using two independently constructed expression vector clones (C). Error bars represent SD. **, $P \leq 0.005$; ***, $P \leq 0.0001$. *P* values were determined using one way ANOVA test with Bonferroni's Multiple Comparison Test.

The aforementioned initial activity tests showed a high degree of difference between the enzymatic activity of the different CDA1L1 enzymes, therefore, the CDAs were divided into two groups: the "high" activity enzymes (CDA1L1_3 and CDA1L1_3spl with maximum overnight incubation deamination percentages of 82.8% and 92.3%, respectively) and the "low" activity enzymes (CDA1L1_1, CDA1L1_1spl, CDA1L1_2, and CDA1L1_4 with maximum overnight deamination percentages of 9.4%, 11.7%, 2.9%, and 4.4%, respectively). Multiple preparations of the high and low activity enzymes were incubated for 4 and 16 hours, respectively, with ^{32}P radioactively-labelled TTCbubB (**Figure 2.4**) at a range of temperatures (4–37 °C; **Figure 4.9**).

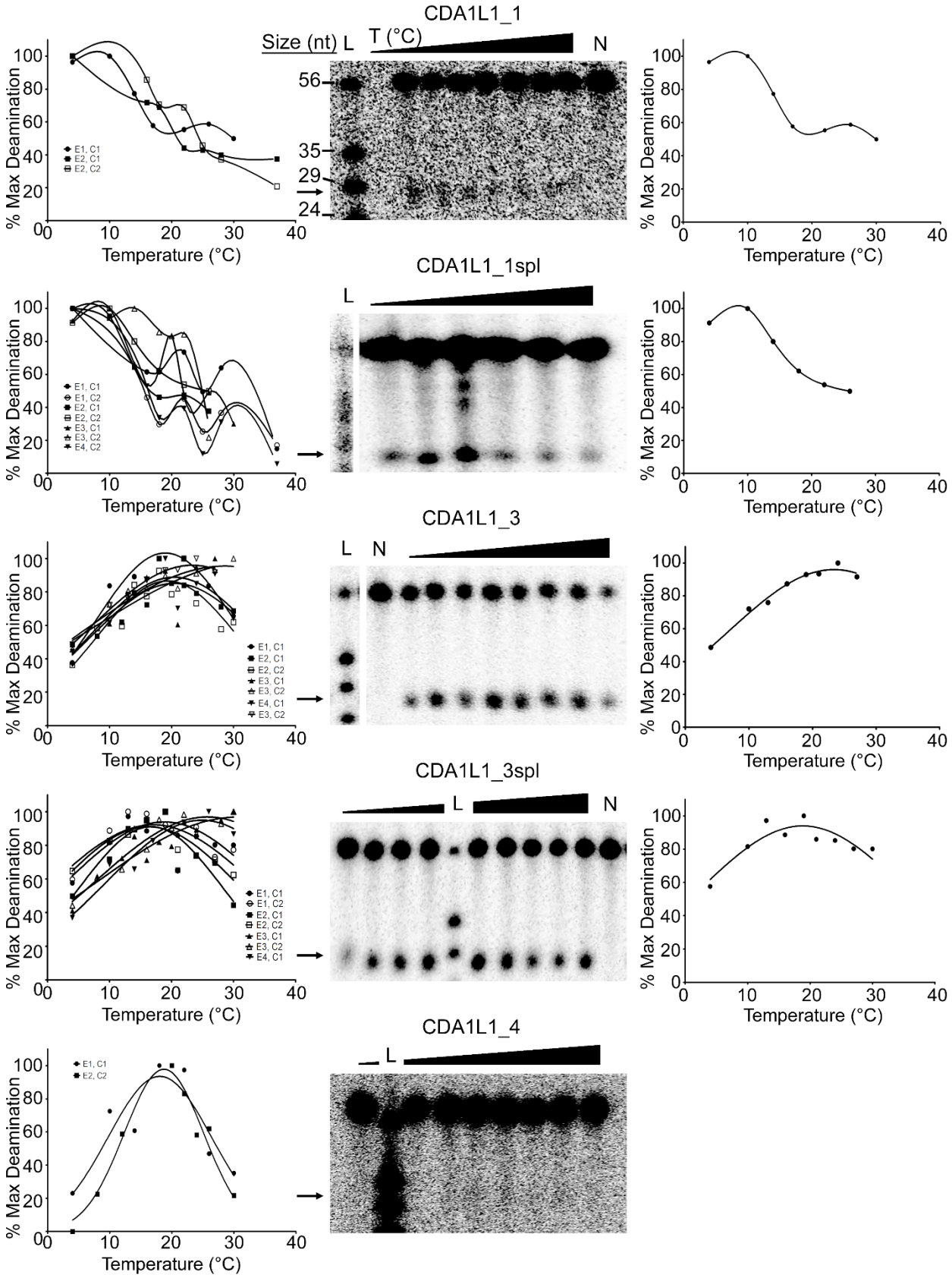


Figure 4.9. CDA1L1 variants exhibit unique thermosensitivity profiles

Legend: Figure 4.9

A Left Graphs present all thermosensitivity curves of each data set used to determine optimal temperature for each CDA ortholog, which was determined from the peak of each curve. The y axis show percentage of maximum deamination to correct for inter-preparation, -clone (C), and -experimental (E) activity differences. The x axis show temperature range in °C from 4 °C to 30 °C. **Middle and Right** Representative gels (middle) and corresponding graphs (right) from thermosensitivity assays. Gels show AID incubated with TGCbub7 for temperatures ranging from 4 °C to 30 °C. L, ladder; N, negative control; T, temperature. The product band is indicated by a black arrow. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

There was an unexpected variation between optimal temperatures of the CDA1L1 enzymes, despite being from the same species. CDA1L1_1 and CDA1L1_1spl preferred the lowest temperatures of 7.3 °C and 6.7 °C, respectively, which were even cooler than the fish AID lysate controls Dr- and Gc-AID (14.0 °C and 13.5 °C, respectively). CDA1L1_3 and CDA1L1_3spl were optimal at 22.7 °C and 21.5 °C, respectively, similar to the optimal temperature of Hs-AID lysate (20.5 °C). CDA1L1_4 had maximum activity at 18.3 °C (**Figure 4.9** and **Figure 4.10**). The activity of CDA1L1_2 in our alkaline cleavage assay was too weak to determine an optimal thermosensitivity profile.

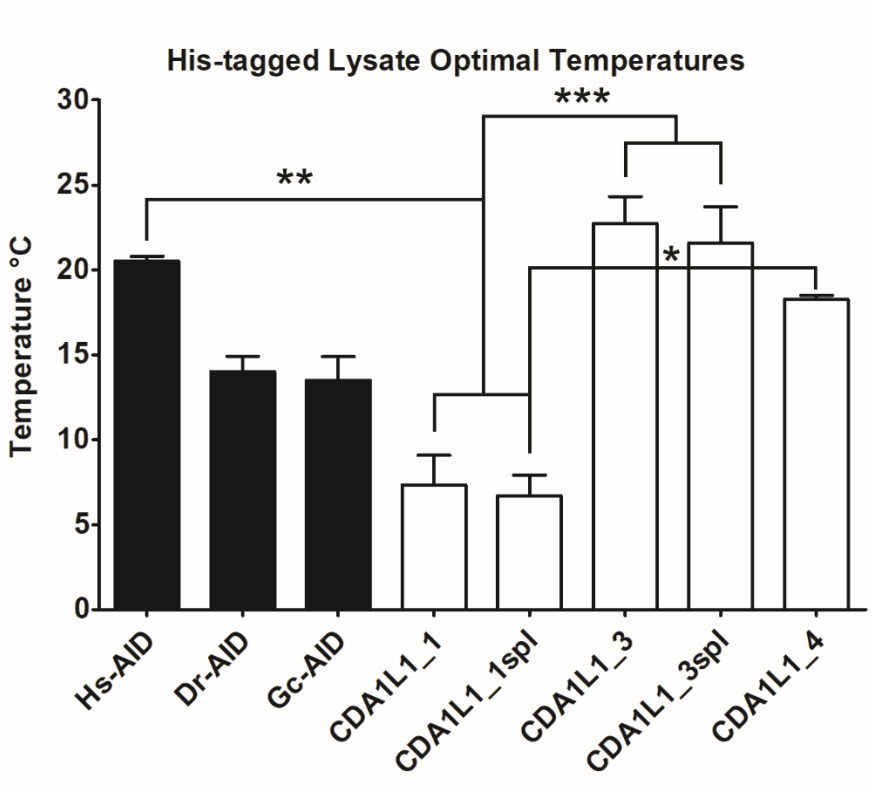


Figure 4.10. Comparison of optimal temperatures of His-tagged eukaryotic lysates

Bar chart of optimal temperatures of all eukaryotically-expressed, His-tagged lysates. Black, controls. White, CDA1L1 lysates. Error bars represent SD. Each bar is a mean of optimal temperatures derived from multiple thermosensitivity experiments. Hs-AID, Dr-AID, Gc-AID, n = 4; CDA1L1_3, CDA1L1_3spl, n = 7; CDA1L1_1spl, n = 8; CDA1L1_1, n = 3; CDA1L1_4, n = 2. *, $P \leq 0.05$; **, $P \leq 0.005$; ***, $P \leq 0.0001$. *P* values were determined using one way ANOVA test with Bonferroni's Multiple Comparison Test.

4.4.2 Optimal pH Profiles of CDA1L1 variants

Due to the different predicted surface charges and isoelectric points for each CDA variant, I surmised that they may exhibit different optimal pH profiles. Therefore, I examined the enzymatic activity of the CDA1L1 variants in our alkaline cleavage experiment, using a range of phosphate buffers at different pH's. I also examined the pH profiles of a number of previously studied AID and APOBEC orthologs, to not only ensure the accuracy of the experiments, but to compare their pH profiles with those of the CDA1L1 variants; any similarities between the different orthologs may offer insight into the enzymatic roles these CDA1L1 proteins may play in the lamprey immune system.

I examined the aforementioned AID control enzymes (Hs-, Dr-, and Gc-AID), along with Pm-CDA1 (see Chapter 3, Quinlan *et al*., 2017), a close ortholog to the CDA variants, which all have similar isoelectric points and surface charges to LpCDA1L1_4. I also tested human A3G, which shares a similar isoelectric point and surface charge to the other LpCDA1L1 variants (**Figure 4.4**). The Hs-, Dr-, and Gc-AID enzymes, along with A3G, were eukaryotically-expressed and His-tagged, while the Pm-CDA1 enzyme was prokaryotically-expressed and GST-tagged due to the former expression system not producing an active Pm-CDA1 enzyme, likely due to its unique protein-folding needs: 1) it is the most cold-adapted of the aforementioned orthologs, likely requiring protein production at cooler temperatures than what is optimal for HEK cells, 2) it is also the most evolutionarily-distant, perhaps requiring unique tRNA codons not present in HEK cells. Hs-, Dr-, and Gc-AID-His and GST-Pm-CDA1 were incubated with ^{32}P-labelled bubble substrate TGCbub7 (**Figure 2.3A**) at 16–22 °C for a maximum of 6 hours at a range of pH's (4.12–7.75); A3G-His was incubated with radioactively-labelled single-stranded CCCssA3G (**Figure 2.3F**) for 30 minutes at 37 °C at a range of pH's (3.89–7.73; see **Figure 4.11**).

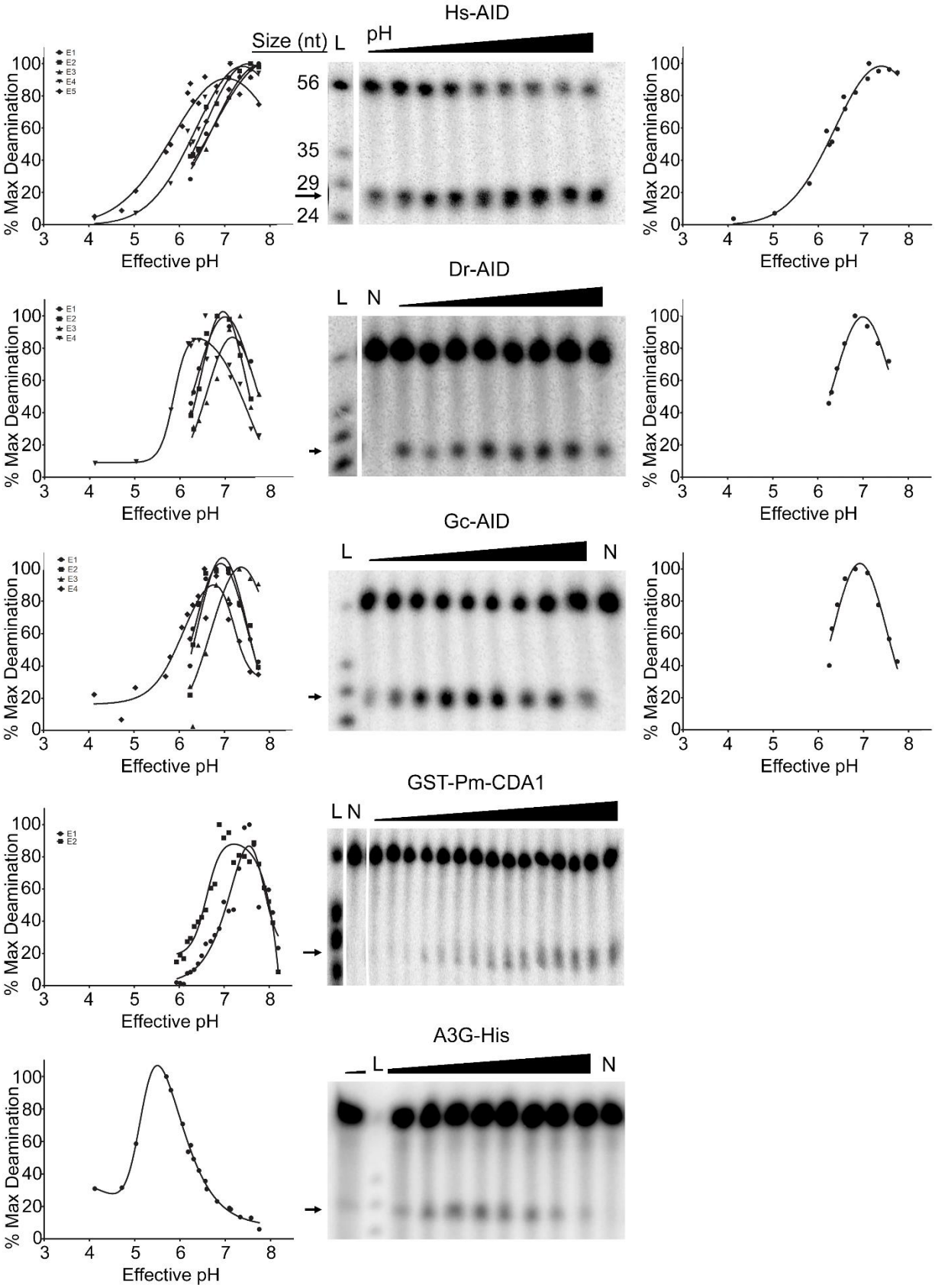


Figure 4.11. Control AID enzymes have neutral pH profiles

Legend: Figure 4.11

Left Graphs present all pH curves of each data set used to determine optimal pH for each eukaryotically-expressed His-tagged lysate (Hs-, Dr-, and Gc-AID), eukaryotically-expressed His-tagged C-terminal domain of human A3G, and for the prokaryotically-expressed, GST-tagged Pm-CDA1, which was determined from the peak of each curve. The y axis shows percentage of maximum deamination to correct for inter-preparation, -clonal, and -experimental activity differences. The x axis shows effective pH range from 4.1–7.8. **Middle and Right** Representative gels (middle) and corresponding graphs (right) from the pH assays. Gels show His-tagged, eukaryotic lysate AID incubated with TGCbub7 at effective pH's 6.2–7.8 for Hs-AID, and 6.2–7.7 for Dr-AID and Gc-AID. The GST-Pm-CDA1 gel shows GST-Pm-CDA1 purified from prokaryotic lysate incubated with TGCbub7 at effective pH's 6.3–7.1. The A3G gel and corresponding graph shows His-tagged, eukaryotic purified C-terminal portion of human A3G incubated with CCCssA3G at effective pH's 4.1–7.1. L, ladder; N, negative control; E, experiment number. The product bands are indicated by a black arrow. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

Hs-AID exhibited the highest activity at pH 7.5, while Dr-AID and Gc-AID had slightly more acidic optimal pH's of 6.9 and 7.0, respectively (**Figure 4.11**). Like the thermosensitivity profiles, the optimal pH's of these AID orthologs are similar to their prokaryotically-expressed, GST-tagged counterparts : 7.2–7.5[338,339] for Hs-AID and 7.2 for Dr- and Gc-AID (**Appendix 2**). Prokaryotically-expressed GST-Pm-CDA1 also demonstrated a somewhat neutral optimal pH of 7.6, while human A3G-His exhibited a vastly more acidic optimal pH of 5.5 (**Figure 4.12**).

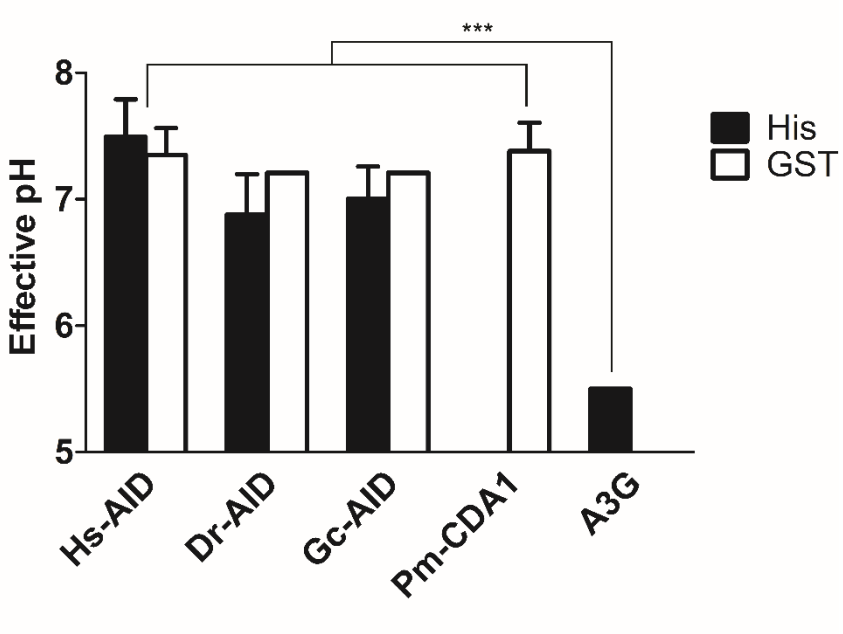


Figure 4.12 Eukaryotically-expressed, His-tagged AID orthologs have similar optimal pH's to their prokaryotically-expressed, GST-tagged counterparts

Bar chart of optimal pH's of eukaryotically-expressed, His-tagged lysates Hs-, Dr-, Gc-AID, and human A3G, compared to the optimal pH's of prokaryotically-expressed, GST-tagged purified proteins Hs-, Dr-, and Gc-AID, and Pm-CDA1. Black, His-tagged enzymes. White, GST-tagged enzymes. Error bars represent SD. The following bars are a mean of optimal temperatures derived from multiple thermosensitivity experiments: Hs-AID-His, n = 5; Dr-AID-His, n = 4; Gc-AID-His, n = 4; GST-Pm-CDA1, n = 2, GST-Gc-AID, n = 2. The bars denoting GST-Dr-AID, GST-Dr-AID, and A3G-His are representative of one experiment each. The bar for GST-Hs-AID is a mean of two optimal pH values from one previously published report[338] and one manuscript in progress. ***, $P \leq 0.0001$. *P* values were determined using one way ANOVA test with Bonferroni's Multiple Comparison Test.

To determine the optimal pH profiles of the CDA1L1 variants, the high activity His-tagged enzymes in eukaryotic cell lysate (CDA1L1_3 and CDA1L1_3spl) were incubated with ^{32}P radioactively-labelled bubble substrate TTCbubB (**Figure 2.3D**) at 16–22 °C for a maximum of 5 hours at a range of pH's (4.12–7.75). Likewise, the low activity His-tagged lysates (CDA1L1_1, CDA1L1_1spl, CDA1L1_2, CDA1L1_4) were incubated with TTCbubB at 16–22 °C for 24–84 hours at a range of pH's (3.89–7.73).

Similar to their unexpected optimal temperature profiles, which varied between proteins despite being from the same animal, the CDA1L1 variants had a range of optimal pH's. Indeed, two variants displayed optimal activity at two pH's, which has never been seen before in any AID/APOBEC enzymes studied to date. The activity curves for CDA1L1_1 showed two peaks at pH 6.1 and 7.7, while CDA1L1_1spl also showed two peaks of activity at pH 5.8 and 7.7. CDA1L1_3 and CDA1L1_3spl had one optimal pH each, at pH 6.0 and 5.5, respectively (**Figure 4.13**). CDA1L1_2 and CDA1L1_4 had too low of activity to accurately determine optimal pH profiles.

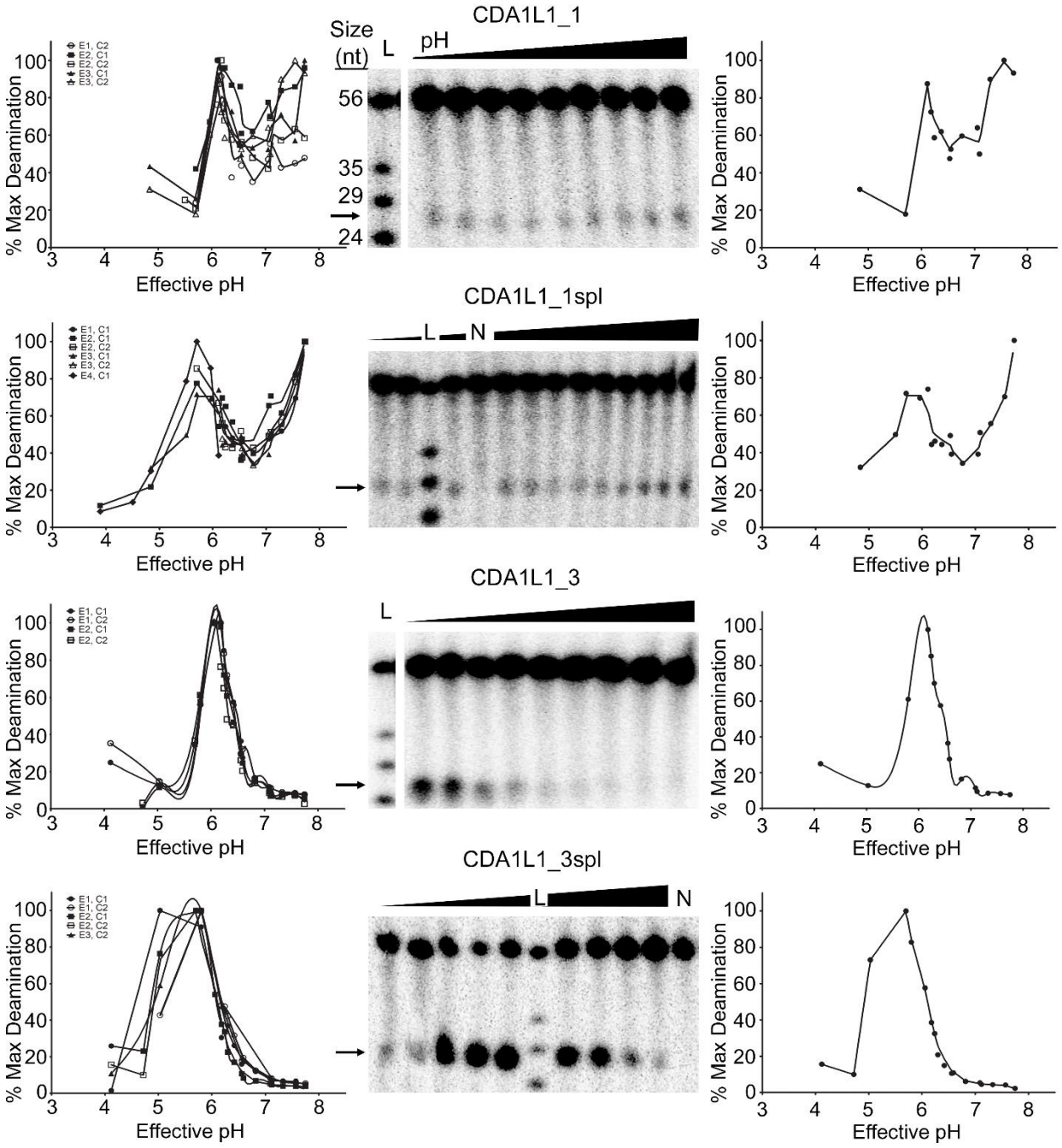


Figure 4.13. CDA1L1 variants have variable optimal pH's

Legend: Figure 4.13

Left Graphs present all pH curves of each data set used to determine optimal pH for each His-tagged CDA1L1 variant in lysate, which was determined from the peak of each curve. The y axis show percentage of maximum deamination to correct for inter-preparation, -clonal, and -experimental activity differences. The x axis shows effective pH range from 3.9–7.7 for the low-activity CDA1L1 variants (CDA1L1_1 and CDA1L1_1spl) and from 4.1–7.8 for the high-activity CDA1L1 variants (CDA1L1_3 and CDA1L1_3spl). **Middle and Right** Representative gels (middle) and corresponding graphs (right) from the pH assays. Gels show AID incubated with TTCbubB at effective pH's 6.2–7.7 for CDA1L1_1, 6.11–7.73 for CDA1L1_1spl, 6.2–7.8 for CDA1L1_3, 4.1–7.1 for CDA1L1_3spl. L, ladder; N, negative control; E, experiment number; C, clone number – note, each CDA1L1 protein preparation is derived from one unique genetic clone. The product bands are indicated by a black arrow. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

The comparison of the optimal pH values of these CDA1L1 variants with those of previously-studied AID and APOBEC orthologs has revealed an interesting pattern. While all AID orthologs studied to date have neutral to slightly-basic optimal pH's, two of the CDA1L1 variants have acidic optimal pH's (CDA1L1_3 and CDA1L1_3spl), which is more akin to that of human A3G. Two other CDA1L1 variants, CDA1L1_1 and CDA1L1_1spl, displayed two optimal pH values each, one of which being more neutral (like the AID orthologs) and one more acidic (like A3G). A graph showing the optimal effective pH's of all enzymes tested is shown in **Figure 4.14**.

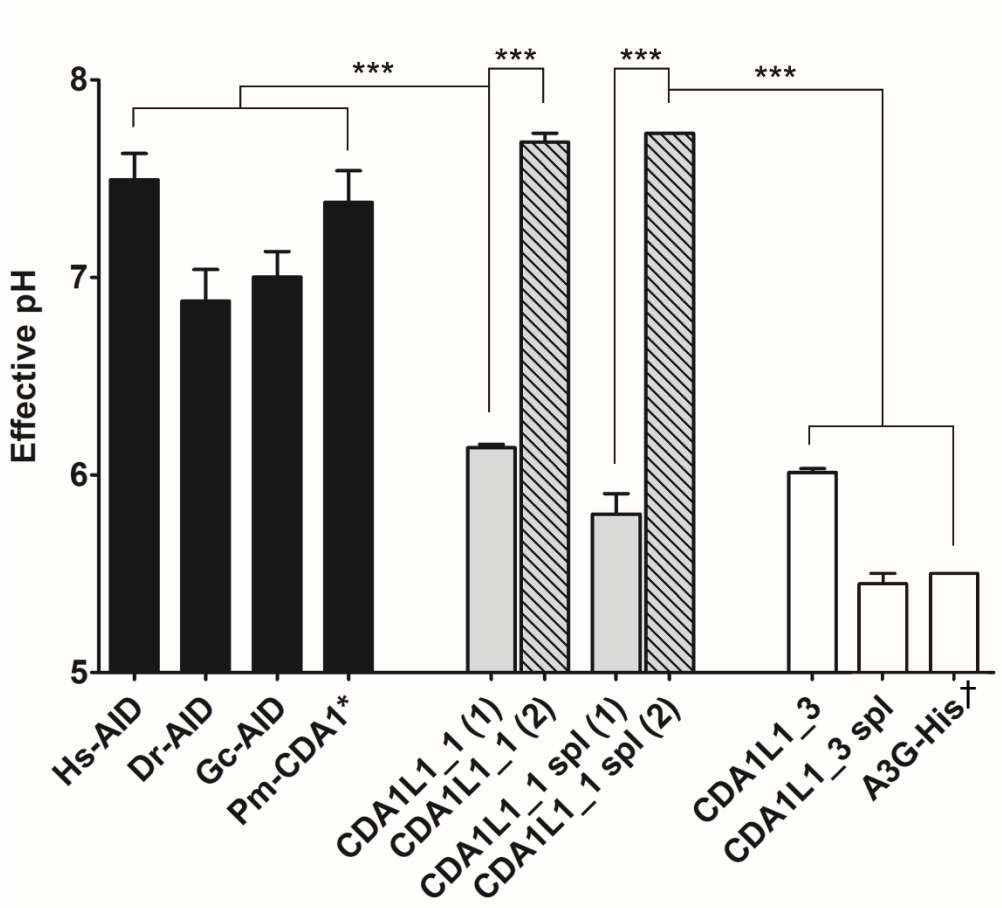


Figure 4.14. Graph of average optimal pH's for each His-tagged AID ortholog and CDA1L1 variant in lysates.

The x axis shows the name of each ortholog/variant (His-tagged, eukaryotic lysate unless otherwise stated). Each average is determined from two to four unique experiments using two independently constructed expression vector clones, except for A3G-His, which is a result of one experiment on one protein purification derived from one clone, as its low optimal pH has been previously shown, and GST-Pm-CDA1, which is a result of two experiments on one protein purification from a single clone. Error bars represent SEM. *, Pm-CDA1 is GST-tagged, purified from prokaryotic cells. †, A3G is His-tagged, purified from eukaryotic cells. Black, control AID orthologs; grey, CDA1L1 variants that display two optimal pH's; white, acidic CDA1L1 variants and A3G. ***, $P \leq 0.0001$. *P* values were determined using one way ANOVA test with Bonferroni's Multiple Comparison Test.

4.5 Discussion

I previously reported that each CDA1L1 variant had the capacity to bind ssDNA *in silico* in such a way that dC fit in the catalytic pocket in an enzymatically favourable position, thus they each had the potential to be active cytidine deaminases.[127] Our initial activity tests on His-tagged lysates of CDA1L1_1, CDA1L1_3, their splice variants, and CDA1L1_2 and CDA1L1_4 confirmed that they are active cytidine deaminases on ssDNA (**Figure 4.6** and **Figure 4.7**,). These initial tests also hinted that each enzyme had unique optimal temperatures and pH profiles (ex., CDA1L1_1, 22 °C and pH 5.9; CDA1L1_1spl, 14 °C and pH 6.4) , despite being from the same species, *Lampetra planeri*; this was unexpected, as orthologous enzymes generally evolve to accommodate the body temperature of the species in which they are expressed (see theory of "corresponding states" discussed in Chapter 3).[316–318] These different enzyme characteristic between variants suggested unique biochemical activities, and were thus examined in more detail.

To further elucidate these characteristics, I first tested the optimal thermosensitivity profiles of each enzyme. The mammalian Hs-AID-His was more active at a warmer temperature than the zebrafish and shark AID orthologs (**Figure 4.8**), as expected due to previous results from their prokaryotically-expressed, GST-tagged, purified counterparts (see Chapter 3, theory of corresponding states and Quinlan *et al*. 2017).[229] That being said, the eukaryotically-expressed, His-tagged lysate controls (Hs-, Dr-, and Gc-AID) showed cooler optimal temperatures than their GST-tagged, purified counterparts.[127,183] The difference in absolute optimal temperatures between the two types of enzyme preparation could be due to fact that the His-tagged proteins 1) are in full cell lysates rather than purified form, thus any change in temperature would also change protein-protein interactions within the lysate, and/or 2) have been subjected to post-translational

modifications in the eukaryotic cells, which likely result in different thermodynamic influences that could affect thermostability, compared to purified proteins in dialysis buffer.

I used cell lysates containing CDA1L1-His proteins as we were unable to purify the His-tagged CDA1L1 lysates in an active form; this is likely due to the post-translational modifications that take place in eukaryotic cells happening at 37 °C in our HEK293T cells, rather than at physiological temperatures for these fish proteins (5–20 °C):[340–342] these sub-optimal conditions possibly result in a majority of the produced protein being misfolded and improperly modified. Taking this into consideration, the purification process results in extreme loss of protein compared to full cell lysates, possibly reducing the percentage of properly folded and modified protein. This hypothesis is strengthened by our previous experience with human proteins such as Hs-AID and human A3G, which are able to be purified as His-tagged proteins from our HEK293T cells. It is logical that human proteins are expressed, folded, and modified more "correctly" from human cells, compared to fish proteins. The His-tagged purification process is currently being optimized, and these proteins are being cloned into pcDNA3.1/V5-TOPO GST vectors, which we hope will facilitate expression of active, GST-tagged enzymes. Another avenue of future research includes expression of these AID orthologs in fish cell lines which could be grown at a cooler temperature, hypothetically enabling the more accurate (i.e., closer to *in vivo*) production of these proteins.

Compared to the AID orthologs, the CDA1L1_1 and CDA1L1_3 proteins and their spliced variants exhibited broader thermosensitivity curves, suggesting these proteins are more thermostable across a wide range of temperatures (**Figure 4.8**–**4.10**). This is in contrast to CDA1L1_4, which exhibited more narrow thermosensitivity curves similar to that of the control proteins: these data further support our previously-published theory that CDA1L1_4 is more similar to Pm-CDA1, Hs-AID, and other AID orthologs, than the other CDA1L1 variants.[127]

Furthermore, CDA1L1_1 and CDA1L1_1spl preferred much cooler temperatures compared to the other CDA1L1 variants (~ 7 °C compared to ~ 20 °C). These unique temperature characteristics for each CDA1L1 ortholog were unexpected, as all of these proteins are from the same species. That being said, lamprey species are among the rare vertebrates to undergo varied life cycle stages that involve metamorphosis: brook lamprey change from filter-feeding, stationary larvae to mobile adults that lack parts of their digestion system.[324] While the optimal water temperatures for the European Brook Lamprey throughout all their life stages is currently unknown, spawning in British rivers occurs around 11 °C, the optimal temperatures for larvae survival in laboratory conditions was found to be 14–18 °C, and freshwater temperatures in Britain range between 5 °C and 20 °C annually.[340–342] It is possible that the expansion of CDA1L1 variants seen in the lamprey serves to provide immunological support at the different life cycles and temperature fluctuations experienced by this species. The broad thermosensitivity curves of the CDA1L1_1 and CDA1L1_3 orthologs, meaning these enzymes are active at a wide range of temperatures, also support this theory.

A second possibility for the optimal temperature variance between LpCDA1L1 orthologs is that one or more of these CDA1L1 variants require accessory proteins to function at physiological temperature. As previously stated in this manuscript, a close relative of AID, A1, has an optimal temperature of 45 °C, and requires an accessory protein, ACF, to fold its substrate (mRNA) into an enzymatically-favourable conformation at 37 °C that naturally occurs at >40 °C [237]. Future work to explore this theory could involve mRNA analysis on co-transcribed genes, or pull-down assays using a CDA1L1 variant in developing VLRA/C cells.

A third possibility is that these CDA1L1 variants evolved non-physiological optimal temperatures to maintain genomic integrity, similar to the theory that some characteristics of

Hs-AID evolved sub-optimally for the same reason.[63,226,229] That being said, CDA1L1_4, CDA1L1_3, and CDA1L1_3spl have similar optimal temperatures of ~ 20 °C, yet these enzymes have vastly different maximum overnight % deamination values (4.4%, 82.8%, and 92.3%, respectively). It is apparent that this characteristic must be further explored, using differently tagged and expressed enzymes, within and without the cell, in order to better understand these enzymes and their functions.

The other biochemical characteristic that I studied in this chapter was optimal pH. We previously determined, *in silico*, that the surface charge of CDA1L1_4 was more similar to that of other AID orthologs (human, zebrafish, sea lamprey CDA1, and mouse), than that of the other CDA1L1 variants, whose overly negative surface charges matched that of the C-terminal domain of A3G (**Figure 4.4**). This hypothesis was supported by our preliminary work that showed, out of three pH's, CDA1L1_4 preferred a neutral pH of 7.1 compared to the other CDA1L1 variants, which preferred more acidic pH's (6.4, aside from CDA1L1_1, whose optimal pH was 5.9).[127] I later performed a more thorough pH analysis of the CDA1L1 variants and human, shark, and zebrafish controls. The control AID orthologs preferred roughly neutral pH's, with human AID preferring a slightly more basic pH and zebrafish AID preferring a slightly more acidic pH (**Figure 4.11** and **Figure 4.12**). Like their thermosensitivity curves, the pH curves of these control enzymes were narrow (**Figure 4.11**). The pH curves of the CDA1L1 enzymes were also narrow, suggesting that, unlike with thermosensitivity, the CDA1L1 variants are more sensitive to changes in pH (**Figure 4.13**). However, that is where the similarities end. The CDA1L1_1 spliced and unspliced proteins show a preference for two pH's, ~ 6 and ~ 7.7 (**Figure 4.14**). This is contrasted with the CDA1L1_3 spliced and unspliced variants pH curves, which only had one peak; that being said, in some experiments it appears as though the curve begins to incline towards a second peak at

more acidic pH's, but this is not consistent across all experiments (**Figure 4.14**). In all cases, a wider range of pH's is required to get a clearer picture of the pH profiles of these enzymes.

The first optimal pH peak of ~ 6 for CDA1L1_1spl and CDA1L1_1 is expected due to their overall negative surface charge. It could be argued that the second peak that is present at higher pH's for the weaker CDA1L1_1 spliced and unspliced variants, while lacking in the more active CDA1L1_3 spliced and unspliced variants, is due to the His tag. Histidine becomes deprotonated at higher pH's, potentially causing structural changes that affect protein activity; such changes might be more apparent in weaker enzymes than in more active enzymes. However, this phenomenon is not seen in any other His-tagged protein we have tested, regardless of differences in enzymatic activity; for example, shark AID (Gc-AID) is among the lowest-activity AID orthologs we've tested,[229] yet its His-tagged version exhibits the expected single optimal pH. Furthermore, CDA1L1_1spl and CDA1L1_1 also displayed unique temperature preferences as compared to the other CDA1L1 variants; it is possible that these enzymes have vastly different biochemical characteristics compared to the other CDA1L1 variants, thus their pH profiles are not artifacts of His-tag interference, but are in fact representative of these differences.

The dual optimal pH peaks could indicate two different structural states that are enzymatically favourable, such as oligomerization. Using the results from one pH experiment for CDA1L1_1 as an example (**Figure 4.15**), the low deamination activity seen at low pH could be indicative of low stability of CDA1L1_1 as a monomer.

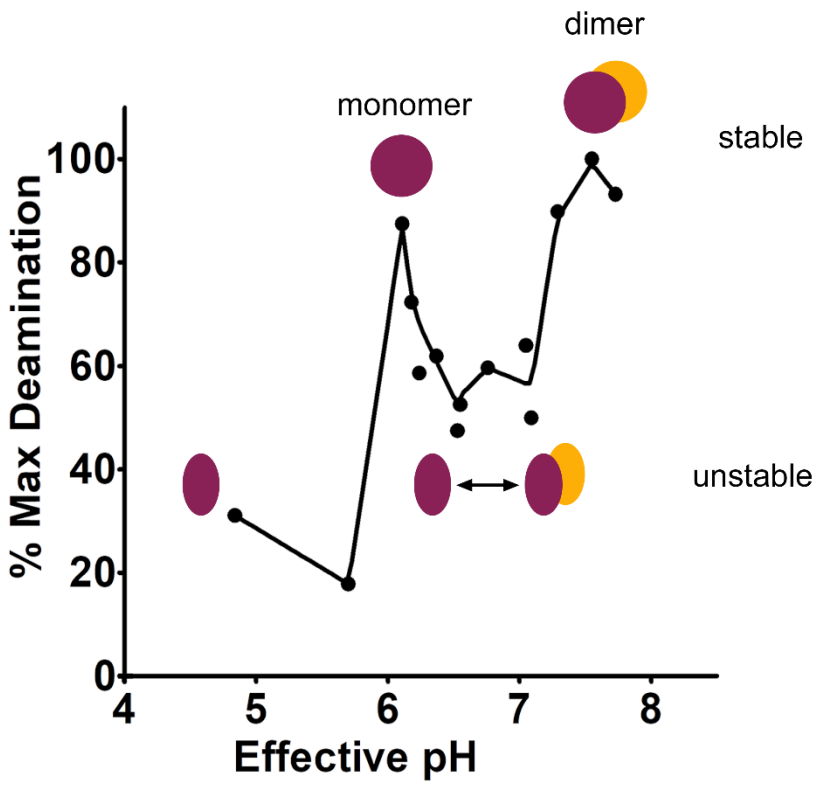


Figure 4.15. Two optimal pH peaks could be due to CDA1L1_1 oligomerization

% Maximum deamination of LpCDA1L1_1-His incubated with 50 fmol of TTCbubB at 22 °C for 84 hours in citrate-phosphate pH 3.89-7.09 and phosphate buffer pH 6.18-7.73. Stable proteins, full circles; unstable proteins, ovals; first CDA1L1_1 protein, red-purple; second CDA1L1_1 protein, gold.

As the pH of the solution increases, the enzyme reaches peak deamination activity at pH 6.1, indicative of peak stability and activity in its monomeric state. As the solution becomes more basic, some of the monomers will transition to an unstable dimeric state, while others will remain as monomers, albeit at suboptimal stability, thus the lower deamination activity at these pH levels. Then, as the pH increases further and the enzyme becomes more deprotonated, more and more of it will exist as dimers, thus achieving a second stable and enzymatically favourable state at pH 7.5. This is currently a hypothesis, as whether CDA1L1_1 dimerizes or disassociates at the lower or higher pH remains unknown; this hypothesis could be verified by running a purified sample of CDA1L1_1 on a native PAGE gel in different pH buffers.

Oligomerization of the CDA1L1 orthologs would not be unusual. It is hypothesized that 30–50% of all enzymes self-oligomerize and that any protein with free N or C termini can potentially form domain-swapped dimers.[343,344] Many members of the APOBEC family are only active as multimers (A1, A3B, A3D, A3F, A3G, and A3H),[345] a yeast ortholog of AID and A1 (CDD1) acts as a dimer,[271] and it is currently debated whether human AID acts *in vivo* as a monomer or multimer.[301,346,347] Both AID and the APOBEC3s have been found to bind to RNA and that oligomerization may be dependent on such binding.[347] AID oligomerization has been suggested to be part of the mechanism for CSR, whereas A3G oligomerization is thought to play a role in its interaction with HIV virions.[108,347] Thus, oligomerization of CDA1L1_1 and CDA1L1_1spl could be related to other unknown immune modulating mechanisms. Whether these proteins oligomerize and, if so, what roles the multimers and monomers play in the lamprey immune system, is an exciting avenue for future research.

Overall, the CDA1L1_4 optimal neutral pH matches that of human, zebrafish, shark, and Pm-CDA1, while the other CDA1L1 variants displayed optimal pH's more similar to that of

A3G-His (**Figure 4.14**). This is corroborated by the thermosensitivity profiles (**Figure 4.10**). These data lend credence to our hypothesis that CDA1L1_4 has more biochemical characteristics in common with those of the AID orthologs and Pm-CDA1, while the other CDA1L1 variants are more enzymatically similar to A3G. AID and A3G have vastly different roles in the immune system. While AID acts inside the nucleus of developing B cells to mutate immunoglobulin genes and is a relatively weak enzyme,[1,229,348] A3G works in the cytoplasm of virally-infected cells to mutate the viral genome and is a highly active enzyme.[108] Due to the arms race between primates and HIV-like viruses, the APOBEC3 sub family has expanded uniquely in primates, from one to seven different enzymes, with various population-specific allelic differences.[108] It is possible that the expansion of CDA1L1 variants we see in the lamprey, which lacks other members of the APOBEC family, is a similar evolutionary response to an infectious agent: CDA1 and CDA1L1_4 may play similar, fundamental roles in antibody assembly, while CDA1L1_1, _2, and _3 may serve anti-pathogen functions, and their variations in copy number between individuals could be explained by allelic differences due to pathogenic pressure.[127]

The variation between number and type of CDA1L1 genes between individuals could also be due to the nature of the lamprey genome: lamprey are unique in that during embryonic development, their somatic cells eliminate ~ 20% of their genome. They end up with two separate genomes: their full, intact genome in their gonadal cells, and a reduced genome in their somatic cells. It is hypothesized that the removed genes are developmental in purpose, and are disposed of when no longer needed, to save energy in replicating them during cell division and growth and to reduce the risk of these developmental genes becoming inappropriately activated.[324,325] It is possible that each individual lamprey has, at least, a CDA1L1_4 and/or CDA1 gene and most, if not all, of the other CDA1L1 variants as embryos.[127] During development, either due to embryonic

contact with environmental pathogens, or epigenetic regulation related to pathogenic interaction of the parents, the developing somatic cells reject CDA1L1 variants that would not be helpful in counteracting the infectious agents said individual is likely to encounter. By removing mutagenic gene products from the somatic genome, the lamprey would save replicative energy and maintain genomic integrity. That all being said, previous examinations of somatic and gonadal genomes of the sea lamprey have suggested CDA genes are not affected by programmed genomic rearrangement.[325] Further examination of gonadal and somatic genomes in various individuals from different ecological niches are required to answer these important questions.

If true, the hypothesis that the expansion of the lamprey CDA1 repertoire is similar to that of mammalian AID/APOBEC has interesting implications in evolution. The emergence of enzymes that somatically mutate antibody genes appears to necessitate their eventual expansion beyond antibody diversification into antiviral roles, while earlier-evolved animals lacking AID/APOBEC orthologs, such as arthropods, appear to produce effective immune responses without AID orthologs. This hypothesis also begs the question, why was this expansion of AID repertoire required in the lamprey, while lacking in its close evolutionary relatives the lungfish and shark? The answers may be found in the complex interplay of factors required to balance evolutionary resources (genetic material, animal fecundity, etc.) with the energy required to elicit an immune response, while considering the effects of life span, communal living, and other animal behaviours, on the chance of infection, along with the benefits and disadvantages of eliciting such a reponse.[349]

For example, perhaps the multiple whole genome duplication (WGD) events which occurred in the jawed vertebrate lineages provided ample and specific genetic material to expand other non-AID/APOBEC immune functions.[350–352] As a case-in-point, codfish, whose ancestors

also underwent a third WGD event followed by genetic loss,[352] have a truncated AID gene, which results in depleted AID activity,[353] no SHM nor CSR, and a lack of MHC II.[354] However, the codfish has expanded MHC I and TLR families; therefore, the codfish has compensated its reduced adaptive immunity with increased innate immune functions.[354] Another example of "compensatory" immune system evolution is seen in arthropods, which have developed the adaptive immunity-like RNAi system, which silences viral genomes using small interfering RNA fragments derived from viral mRNA, and the DSCAM (Down syndrome cell adhesion molecule) gene, which encodes an Ig superfamily protein and can undergo alternative splicing, generating tens of thousands of pathogen-specific isoforms.[221] Both codfish and arthropods lack the same adaptive immunity seen in mammalian species, but have instead developed an expanded innate immune system and an adaptive-like immune system, respectively, while benefiting from a decreased chance of errant somatic mutations and autoimmunity from a lack of AID-like enzymes.[355]

Perhaps lamprey have an expanded CDA repertoire due to the proto-CDA/AID gene being available while lacking certain genetic material afforded to the gnathostome lineage, due to the asymmetrical distribution of genetic material to the cyclostome and gnathostome lineages after the first WGD event in the vertebrate ancestor.[350,351,356–358] Therefore, single-gene or -locus duplication events to expand the lamprey CDA/AID repertoire to create antiviral mechanisms may have been more energetically favourable than expanding innate immune mechanisms, like the codfish or arthropods. As for the jawed vertebrate ancestor, perhaps the expansion of their AID repertoire was not yet needed due to the aforementioned interplay of infection risk, evolutionary material, and energy balance; for example, sharks have appeared to evolve increased wound-healing and anti-tumour capabilities beyond that of teleost fish,[359] which may have decreased the

former's need for an increased anti-viral AID-like repertoire. Thus, the AID/APOBEC repertoire remained simplified in the early gnathostome lineage, until viral risk surpassed the benefits of genetic fidelity, which appears to have occurred at various points throughout jawed vertebrate evolution (see **Figure 1.4**). In species such as the primates, with such conditions as low fecundity (increased need for individuals to survive), long-term communal living (increased viral risk), and a lack of specific genetic material to create new antiviral mechanisms, expansion of the APOBEC3 lineage may have been the most favourable solution to increased primate-specific retrovirus activity.[108,360–362] Further exploration of the lamprey genome and immune system, and those of earlier-diverged fish species, could further elucidate the similarities between the lamprey and mammalian immune systems.

Chapter 5 Lamprey CDA2 *may* be an active cytidine deaminase

5.1 Authorship Statement

The data in this chapter are not yet published. Lesley Berghuis expressed and purified the CDA2 variants, assessed protein expression using western blot experiments, and performed initial cytidine deaminase activity checks. M Larijani supervised, designed, and analyzed the research, and edited this chapter. I designed and conducted the extended cytidine deaminase activity experiments, analyzed the data, performed the literary research, and wrote the chapter.

5.2 Introduction

The first suggestion of lamprey having a type of adaptive immune system occurred in the 1960's, when it was discovered that they produce circulating "antibodies" (that were not immunoglobulins) in response to immunization with killed Brucella cells, reject repeated skin allografts at an accelerated pace, and produce a delayed type hypersensitivity reaction to tuberculin.[363,364] These "antibodies" and the cells that produced them were further investigated, revealing that lamprey have lymphocytes that behave similarly to those of their jawed vertebrate counterparts, in that they proliferate in response to antigenic stimulation and express similar transcription factors.[363,365–367] In 2004, Pancer and colleagues stimulated an immune response in lamprey larvae by intraperitoneal injections of an antigen/mitogen cocktail, resulting in increased proliferation of lamprey leukocytes. The cDNA libraries from circulating leukocytes revealed over 200 unique transcripts containing LRRs, thus named variable lymphocyte receptors; further investigation determined that each leukocyte expressed a single VLR transcript that was somatically rearranged from a number of LRR modules or cassettes (see **Figure 4.1**).[230]

A year later, the same group published results from similar experiments on hagfish, the second group of cyclostomes, and found two homologs of the lamprey VLR, named VLRA and VLRB. The immature genomic DNA of hagfish VLRA and B, located in two separate loci, were similar to that of lamprey VLR, containing an UTR, signal peptide, and N-terminal LRR, followed by a non-coding region, then a C-terminal LRR, stalk, and hydrophobic tail, all of which was surrounded by LRR modules. The hagfish VLRB and lamprey VLR were found to have a similar threonine/proline-rich region in the stalk which was not found in the hagfish VLRA. Furthermore, while all three proteins contained a GPI anchor site, hagfish VLRA lacked a cleavage signal which was found in the other two VLRs, indicating the former is membrane-bound and likely not secreted

like hagfish VLRB and lamprey VLR.[281] Subsequent searches of the sea lamprey genome by Rogozin *et al*. discovered a second VLR locus similar to the hagfish VLRA and was thus named VLRA; due to the similarities between hagfish VLRB and the initial lamprey VLR, the latter was renamed to VLRB. It was this publication that also revealed the existence of the two CDA proteins (Pm-CDA1 and Pm-CDA2) suspected to be responsible for initiating the gene conversion-like events that somatically rearranged the mature VLR gene.[126] Lamprey VLRA and B were later found to be both membrane-bound in unstimulated animals, whereas only VLRB was secreted upon immune stimulation.[291,368]

Three years after the discovery of a second VLR in the lamprey, another group identified a third VLR protein in the Japanese lamprey with a distinct 3' terminus, named VLRC. Like the other VLR proteins, the cells producing VLRC proteins on their surface produced only VLRC and not any other VLR, and each VLRC+ clone produced a genetically unique VLRC transcript. Furthermore, like the other VLRs, the germline VLRC gene contained only the 5' and 3' regions of the mature transcript (see **Figure 4.1**), lacking the middle LRR modules. However, the 5' LRRCT region of the VLRC exhibited low levels of sequence diversity, unlike the same region in the previously characterized VLRA and VLRB. Despite these differences, phylogenetic analysis determined that VLRC is more closely related to VLRA than B.[284] A third VLR was also found in the hagfish, and due to its similarity to the VLRC found in the lamprey, was also named VLRC.[287]

Therefore, both hagfish and lamprey were found to have three distinct VLR genes which are somatically rearranged and exclusively expressed in distinct leukocyte populations. It has been hypothesized that these three VLR proteins are similar to the B and T cell receptors of the jawed vertebrates, and that both types of receptors derive from a single lineage which also used three distinct somatically-rearranged receptors for adaptive immunity before the divergence of the jawed

and jawless vertebrates.[151] Recent work has detailed further similarities between these two groups of receptors, which appear to support this theory.

Aside from expressing only one mature VLR protein, VLR+ cells also exhibit unique tissue distribution. VLRB+ cells are found primarily in lamprey blood and kidneys, similar to where B cells are found in jawed vertebrates. VLRA+ cells mainly populate the skin, typhlosole (intestine inner lining), and gills, the latter of which is the thymus-like organ (thymoid) in the lamprey. VLRC+ cells are also found in the gills and are found in greater number in the typhlosole than VLRA+ cells. In the skin, VLRC+ cells were also less genetically diverse than VLRA+ cells.[151] The tissue-specificity of VLRA+ and VLRC+ cells reflects that of αβ and γδ T cells, which populate the T-cell development organ, the thymus[1]; γδ T cells are also more abundant than their αβ counterparts in the intestinal lining in jawed vertebrates, with limited diversity in the skin.[369]

Another similarity between the VLRs and B and T cell receptors is their pattern of secretion in response to immunization. VLRB+ cells differentiate into plasma cells and secrete their VLRB receptors when activated, whereas VLRA+ and VLRC+ cells do not secrete their receptors, though they do proliferate.[151] This appears to be parallel to the B and T cell receptors in jawed vertebrates where BCRs are secreted by plasma cells and neither population of T cells secrete their receptors in response to immunization.[1]

As mentioned above, lamprey lymphocyte cells appear to express homologs of genes that are often expressed exclusively or more often by jawed vertebrate lymphocytes. For example, VLRB+ cells express *CDA2* and a homolog of gnathostome *BCL6*, a transcription regulator vital to the secondary diversification processes in mature B cells[370]; VLRA+ cells express *CDA1* and a homolog of gnathostome T cell factor 1 (*TCF1*), which is essential in T cell commitment to the αβ

T cell CD4+ lineage[371]; VLRC+ cells also expressed *CDA1*, along with a homolog of mammalian *SOX13*, a γδ T cell lineage-determining transcription factor.[151,372] Interestingly, during lamprey lymphocyte development in the thymoid, both *VLRA* and *VLRC* genes were transcriptionally active in the same cell, suggesting each T cell-like cell has the potential to become either a VLRA+ or VLRC+ cell, yet only one receptor is eventually expressed.[151] This is similar to jawed vertebrate T cells, which, during their development in the thymus, begin to rearrange the β, γ, and δ receptors, and the first productive rearrangement, be it either the β chain or both the γ and δ chains, results in αβ or γδ T cell lineage determination, respectively.[1] Thus, the similar reaction to immunization, tissue distribution, and genetic expression profiles of the lamprey and gnathostome lymphocytes suggest that the lamprey VLRB+, VLRA+, and VLRC+ cells are analogous to jawed vertebrate B cells, αβ T cells, and γδ T cells, respectively.

I have previously shown that Pm-CDA1, and its orthologous variants from freshwater lamprey, are active cytidine deaminases with a similar protein structure to Hs-AID, despite being only 17% identical and 37% similar in aa sequence to human AID (see Chapters 3 and 4, and **Figure 3.2**).[127,229] The Pm-CDA2 protein structure has not yet been determined experimentally, and its structure is difficult to accurately simulate, as there are multiple stretches of aa sequence of unknown structure and poor homology to any other known structure; they do not align with Hs-AID nor Pm-CDA1 (see **Figure 5.1A**). Thus, the 3D modelling program is forced to model these sections *ab initio*, leading to low confidence sections of the model that appear intrinsically disordered. Along with its structure, the activity of Pm-CDA2 as a cytidine deaminase has yet to be demonstrated.

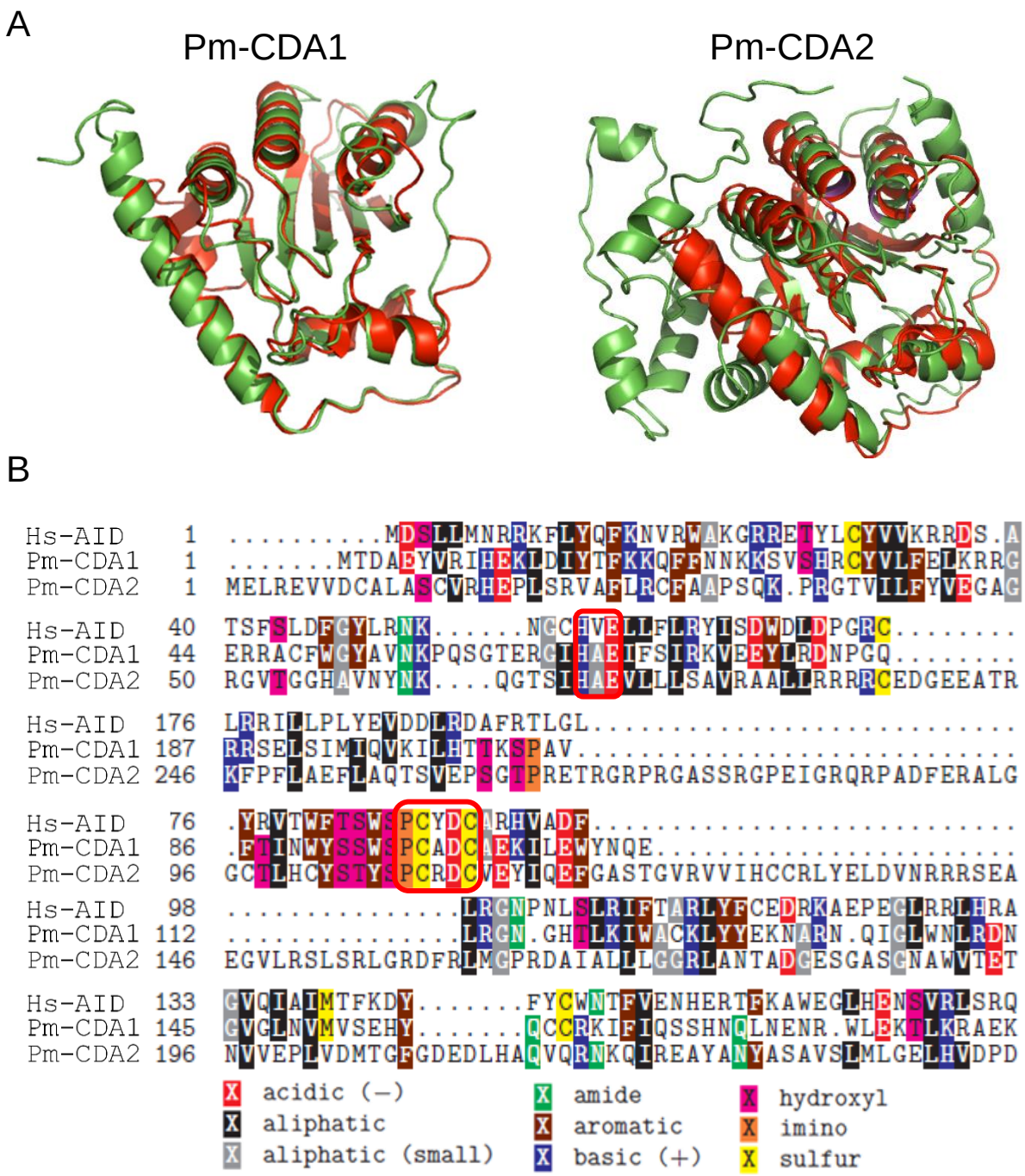


Figure 5.1. Comparison of Pm-CDA1 and Pm-CDA2 to Human AID

A Ribbon structures of Pm-CDA1 (left; green) and Pm-CDA2 (right; green) over human AID (red). **B** Alignment of Hs-AID, Pm-CDA1, and Pm-CDA2 amino acid sequence. The CDA motif is circled in red.

However, there are several reasons to believe that Pm-CDA2 is an active deaminase enzyme. First, Pm-CDA2 does contain the canonical cytidine deaminase motif ($HxE(x)_nPCxxC$) (see **Figure 5.1B**). Second, CDA2 structure modeling, low confidence as it is, does generate a viable 3D active site with the same arrangement of catalytic residues as AID and the CDA1s (**Figure 3.10** and **Figure 5.1**). Third, the CDA2 ortholog in *Lampetra planeri* has shown limited cytidine deaminase activity in cell-based reporter assays performed by our collaborators (oral communication, Boehm lab). Fourth, CRISPR/Cas9 knockout experiments eliminating CDA2 from lamprey larvae demonstrated that CDA2 is required for VLRB antibody gene assembly.[373] Therefore, there is evidence that this protein is vital for lamprey immunity and might be active in some enzymatic form. In this Chapter, I present my work to investigate if this enzyme is an active cytidine deaminase *in vitro.*

5.3 Results

5.3.1 Multiple preparations made from transfection of one clone of Pm-CDA2 are active cytidine deaminases

The Pm-CDA2 ORF was cloned into the pcDNA3.1/V5-8xHis-TOPO vector and expressed in HEK293T cells. Expression was confirmed via western blot using an anti-V5 epitope antibody before concentrating the His-tagged protein by batch-binding with nickel beads (**Figure 5.2A**). As a negative and positive control, the GST tag (forming a GST-V5-His protein) and Hs-AID, respectively, were concurrently cloned, expressed, and isolated in the same manner. Initial activity was determined by incubating multiple preparations of multiple clones of Pm-CDA2 (clones 18 and 25) with our TGCbub7 substrate (**Figure 2.3A**) overnight at 25 °C; the only preparations that demonstrated activity were from one transfection of one clone (clone 25; **Figure 5.2B**). Furthermore, the deaminase activity seen with these preparations was relatively low (~ 1 %) compared to overnight incubations with other AID orthologs (see Chapters 3 and 4) and to the Hs-AID-His control which was incubated concurrently with TGCbub7 (**Figure 2.3A**) for 1 hour at 37 °C (~ 94%). Untransfected HEK293T cell lysate was used as a negative control for any non-specific, non-CDA2, or non-AID cytidine deaminase activity that may be present in the cell lysate. A second negative control containing buffer and substrate, but lacking protein, was also used; both controls were indeed negative for cytidine deaminase activity. These results indicate that the cytidine deaminase activity seen with clone 25 of Pm-CDA2 is likely due to the enzymatic activity of Pm-CDA2, and not an unknown enzyme or chemical in the cell lysate, buffer, or labelled DNA preparation.

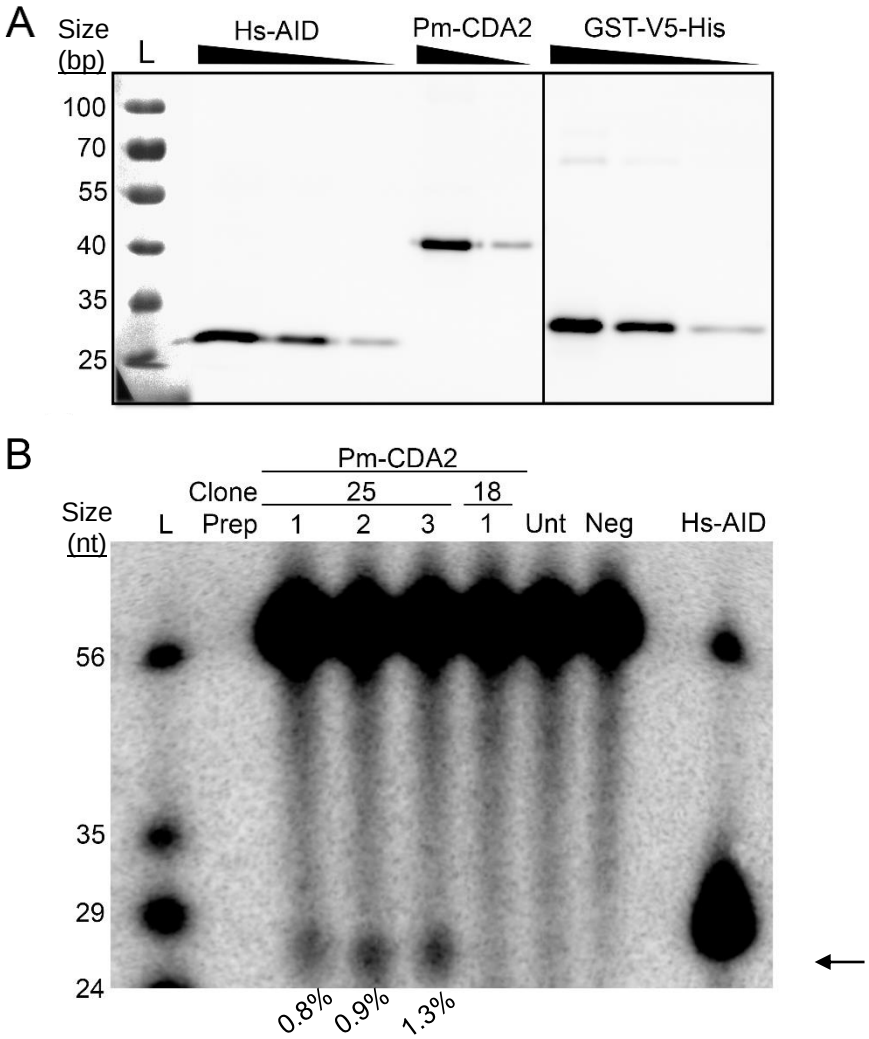


Figure 5.2. Pm-CDA2-His clone 25 is an active cytidine deaminase

A Western blot (WB) of His-tagged Hs-AID (~ 28.5 kDa); Pm-CDA2, clone 25, Preparation (Prep) 3 (~ 41 kDa); and a combined GST-V5-8xHis tag (~ 32.5 kDa). The digitized ladder is superimposed onto the WB image for size reference. **B** Alkaline cleavage gel of multiple preparations of two clones of His-tagged Pm-CDA2 (the active clone 25 [3 preparations] and the inactive clone 18 [1 prep]). Radioactively labelled TGCbub7 was incubated with 10 µL of each preparation of Pm-CDA2 and 5 µL of Hs-AID for 16 hours at 25 °C and 1 hour at 37 °C, respectively. The numbers on the left of the gel indicate the ladder sizes in nucleotides (nt). The product band is at the appropriate size (28 nt, black arrow), showing cytidine deaminase activity (percentage indicated below the gel). L, ladder; Unt, untransfected HEK292T cells; Neg, negative control (buffer and substrate).

5.3.2 Pm-CDA2-His showed temperature-sensitive substrate specificity

To determine the substrate specificity of Pm-CDA2, preparations of Pm-CDA2-His Clone 25 were incubated for 20 hours at 18, 25, and 37 °C with six radioactively-labelled bubble substrates that are identical except for the target C-containing motif in the single-stranded bubble (TGC-, TAC-, AGC-, GGC-, GTC-, and GACbub7; **Figure 2.3A**). The three different temperatures were tested because the optimal temperature for Pm-CDA2 activity had not yet been determined, and these three temperatures encompassed the range of temperatures favoured by other AID orthologs (see Chapters 3 and 4).

At 18 °C, Pm-CDA2 showed minimal activity on TGC- and TACbub7, with no perceptible activity on the other four substrates (**Figure 5.3A**). Surprisingly, at 25 °C, Pm-CDA2 not only showed activity on TGC- and TACbub7, but also AGC- and GTCbub7 (**Figure 5.3B**). The substrate specificity profile changed again at 37 °C, when Pm-CDA2 activity on the WRC motif substrates dropped drastically, while its activity on GTCbub7 increased dramatically (**Figure 5.3C**). The absolute difference between Pm-CDA2 activity on TGC- and GTCbub7 at 25 and 37 °C is apparent in **Figure 5.3D**; the substrate specificity profiles at all three temperatures is shown in **Figure 5.3E**. The untransfected and enzyme-negative controls remained negative at all temperatures, suggesting this difference in substrate specificity profiles at different temperatures is likely not influenced by lysate contents or experimental reagents (**Figure 5.3D**).

By contrast, when Hs-AID was incubated with the same substrates at the same temperatures for 16 hours (**Figure 5.3F**), the substrate specificity profiles were similar across the different temperatures, with the WRC motifs showing the highest activity compared to the non-WRC motifs (**Figure 5.3G**).

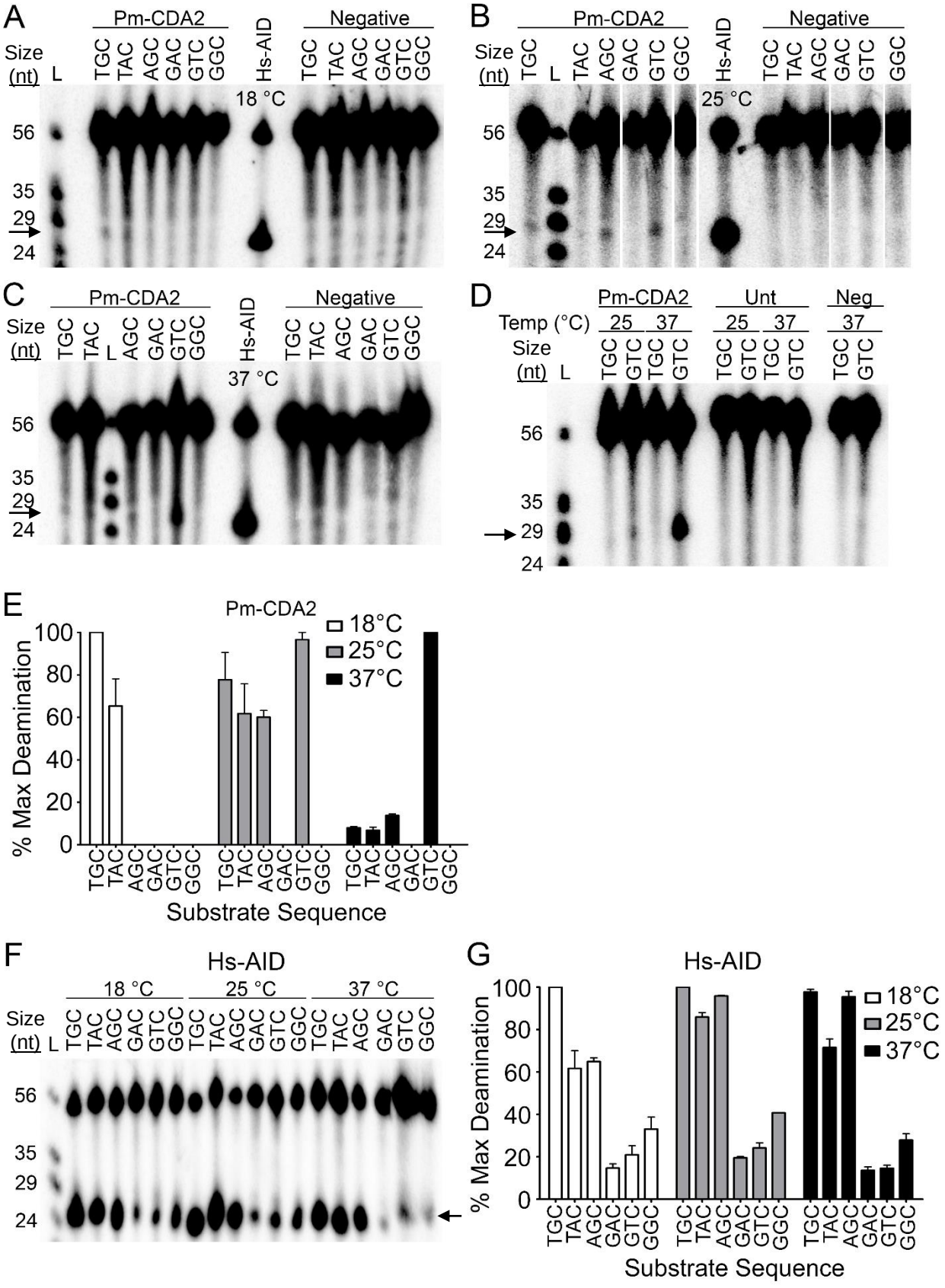


Figure 5.3. Pm-CDA2-His shows temperature-sensitive substrate specificity

Legend: Figure 5.3

A, B, C Alkaline cleavage gels showing the activity of Pm-CDA2 (preparation 1 of Clone 25) on ^{32}P-labelled substrates TGC-, TAC-, AGC-, GAC-, GTC-, and GGCbub7 incubated for 16 hours at 18 °C (**A**), 25 °C (**B**), and 37 °C (**C**). Negative samples contained just buffer and substrate, no protein. **D** Alkaline cleavage gel of Pm-CDA2 (preparation 1 of Clone 25) incubated with TGCbub7 and GTCbub7 for over 20 hours at 25 °C and 37 °C; Unt, untransfected HEK293T cells incubated with substrate for 20 hours; Neg, negative control (phosphate buffer and substrate, lacking protein), also incubated for 20 hours. **E** Bar graph showing % maximum deamination results from the gels in **A, B,** and **C**; the 25 °C bars also contain data from a second experiment, not shown. **F** Alkaline cleavage gel of Hs-AID incubated with TGC-, TAC-, AGC-, GAC-, GTC- and GGCbub7 for 16 hours at 18, 25, and 37 °C. **G** Bar graph showing % maximum deamination results from the gels in **F**, along with a second experiment, not shown. Error bars in **E** and **G** are SD. Black arrows next to the alkaline cleavage gels indicate the product band (28 nt). L, ladder. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

Due to the thermosensitive nature of Pm-CDA2 substrate specificity, more accurate thermosensitivity assays of the concentrated His-tagged Pm-CDA2 and Hs-AID proteins on TGCbub7 and GTCbub7 were performed (**Figure 5.4**). Pm-CDA2 was incubated with TGCbub7 (**Figure 2.3A and Figure 5.4A,** top) and GTCbub7 (**Figure 5.4A,** middle) for 20 hours at a range of temperatures (15–42 °C), while Hs-AID was incubated with both TGCbub7 (**Figure 5.4B,** top) and GTCbub7 (**Figure 5.4B,** middle) for 30 minutes (due to its comparatively high activity rate) at 12–38 °C. Pm-CDA2 showed optimal activity on TGCbub7 at ~ 27 °C, while showing optimal activity on GTCbub7 at the hottest temperature tested (42 °C), and still continuing to increase (**Figure 5.4A**, bottom). Hs-AID showed the expected optimal temperature of ~ 32 °C on both TGCbub7 and GTCbub7 (**Figure 5.4B**, bottom).

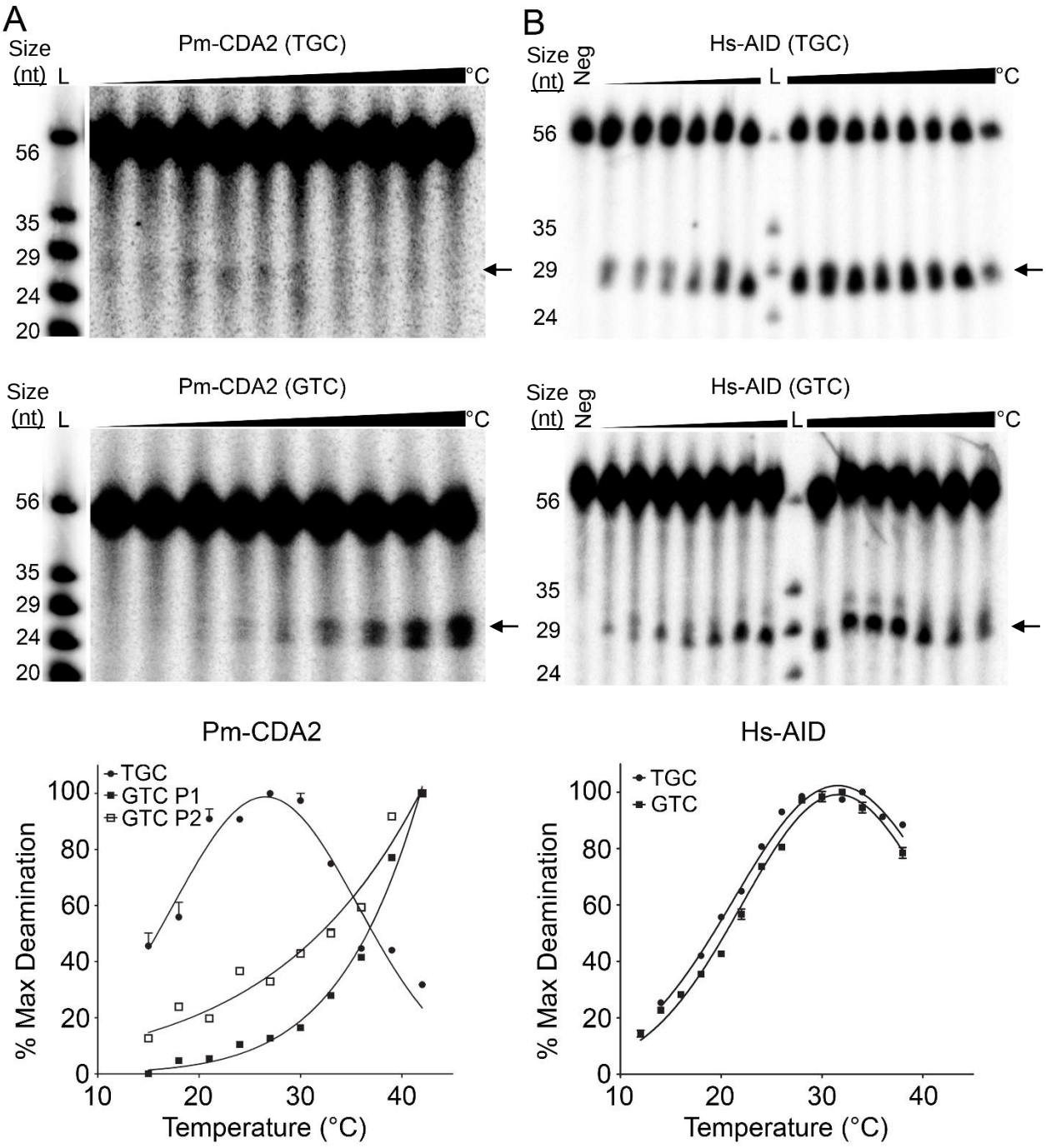


Figure 5.4. Pm-CDA2-His has two distinct thermosensitivity profiles

Legend: Figure 5.4

A Top and middle, alkaline cleavage gels of Pm-CDA2 (preparation 1 of Clone 25) incubated with TGCbub7 (top) and with GTCbub7 (middle) for 20 hours at temperatures ranging from 15 ℃ to 42 °C. Bottom, bar graph of the % maximum deamination values from the two gels in **A**, along with data from a second preparation of Clone 25 of Pm-CDA2 on GTCbub7 incubated in the same manner. P1, preparation 1; P2, preparation 2. **B** Top and middle, alkaline cleavage gels of Hs-AID incubated with TGCbub7 (top) and GTCbub7 (middle) for 30 min at temperatures ranging from 12 ℃ to 38 °C. Bottom, bar graph of the % maximum deamination values from the two gels in **B**. Black arrows next to the alkaline cleavage gels indicate the product band (28 nt). L, ladder. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

5.3.3 Pm-CDA2-His is active on canonical cold spots, ssDNA, and RNA/DNA hybrid bubbles

To test whether Pm-CDA2 is active on different conformations of DNA, preparations of Clone 25 of Pm-CDA2 and Hs-AID (as a control) were incubated at 24 °C (due to this temperature showing the most inclusive substrate specificity profile for Pm-CDA2) for 24 hours with single-stranded, double-stranded, and RNA/DNA hybrid substrates with a target TGC motif, and TGCbub7 with a methylated C (**Figure 5.5A,** see **Figure 2.3B** and **C** for substrate sequences). Pm-CDA2 was more active on the TGC motif in the single-stranded TGCbub7 substrate than in the partially single-stranded TGCbub7 and the RNA/DNA hybrid substrate. Hs-AID was more active on the partially single-stranded TGCbub7 and RNA/DNA hybrid substrates compared to the ssTGCbub7 substrate, similar to the substrate preference previously demonstrated by GST-tagged, prokaryotic-purified Hs-AID in Chapter 3 and published in Quinlan *et al*., 2017[68] (**Figure 5.5A**).

To test the activity of Pm-CDA2 on different DNA sequences and bubble sizes, Pm-CDA2 and Hs-AID were incubated at 24 °C for 24 hours with single-stranded and partially-single stranded substrates (with bubble sizes of five or 13 nucleotides) with different target sequences (GGC, GTC, CCC, GAC, AGC, ACC, TGC, and GCC; **Figure 2.3A** and **Figure 5.5B**). Pm-CDA2 was active on GGC and GAC motifs on single-stranded DNA, GTC in a 13 nt bubble, and single-stranded CCC, while Hs-AID was active on all substrates at multiple motifs, preferring the 5 nt bubble over the 13 nt bubble.

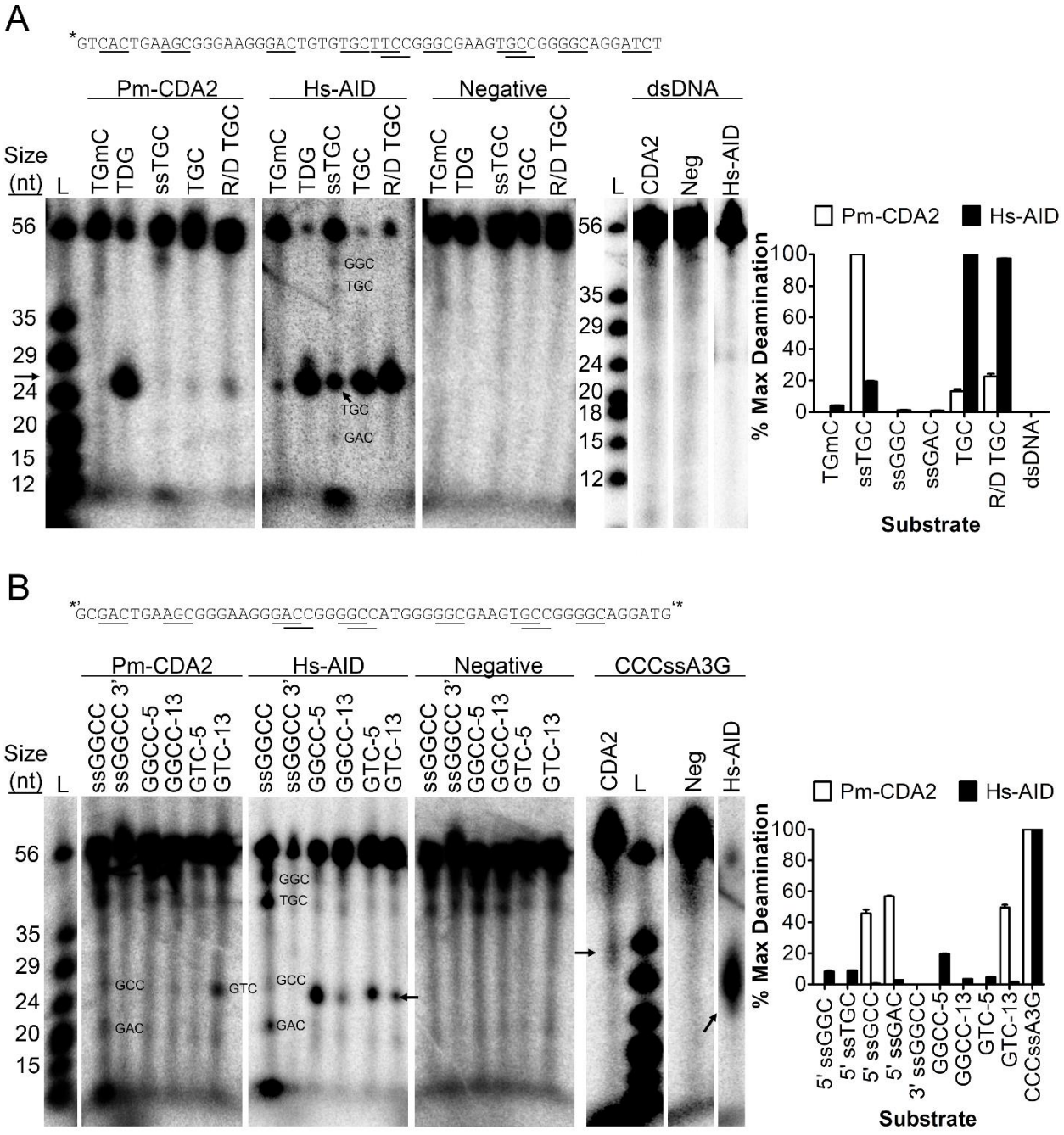


Figure 5.5. Pm-CDA2-His is active on canonical cold spots, ssDNA, and RNA/DNA hybrid bubbles

Legend: Figure 5.5

Left, alkaline cleavage gels of 10 µl Pm-CDA2 and 2 µl Hs-AID incubated at 24 °C for 24 hours with substrates of different structures and sequences. Right, graphs of % maximum deamination results from their respective graphs. **A** TGmC: TGCbub7 with a methylated target C; TDG, control for methylated C alkaline cleavage; ssTGC, single-stranded top strand of TGCbub7 (see sequence above gel image with XXC motifs underlined; *, 5' radioactive labelling); ssTGC/ssGGC/ssGAC- XXC motifs in the ssTGCbub7 substrate deaminated by either Hs-AID or Pm-CDA2; TGC, TGCbub7; R/D TGC, RNA/DNA partially single-stranded TGCbub7; dsDNA, double-stranded TGCbub7; Negative, buffer and substrate, lacking enzyme. The black arrow to the left of the gel indicates where all product bands should lie for deamination of TGC (28 nt) in the TGCbub7 substrate. Product bands in the ssTGC substrate are indicated on the gel for clarity. **B** ssGGCC, single-stranded top strand of GGCCbubC ^{32}P-labelled on the 5' end (see sequence above gel image with XXC motifs underlined; *', 5' or 3' radioactive labelling); ssGGC/ssTGC/ssGCC/ssGAC, XXC motifs in the 5'-labelled ssGGCCbubC substrate deaminated by either enzyme; ssGGCC 3', single-stranded top strand of GGCCbubC ^{32}P-labelled on the 3' end; GGCC-5, GGCCbubC with a 5 nt bubble; GGCC-13, GGCCbubC with a 13 nt bubble; GTC-5, GTCbubC with a 5 nt bubble; GTC-13, GTCbubC with a 13 nt bubble Negative/Neg- buffer and substrate, lacking enzyme; CCCssA3G- optimal substrate of A3G. The black arrow to the right of the Hs-AID gel indicates where all product bands should lie for deamination of GGCC or GTC (27 nt) in the GGCCbubC and GTCbubC substrates, respectively. Product bands in the ssGGCCbubC (motifs) and CCCssA3G (black arrows) substrate lanes are indicated on the gel for clarity. Error bars represent SEM. L, ladder. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

To test the effect of the surrounding DNA on the activity of Pm-CDA2 on a target sequence, Pm-CDA2 and Hs-AID were incubated with AT-rich and GC-rich single-stranded, double-stranded, or RNA/DNA hybrid substrates (**Figure 2.3I** and **J**, **Figure 5.6**). Pm-CDA2 was active on one motif (TGC) in the single-stranded, AT-rich DNA, while Hs-AID was also only active on single-stranded, AT-rich DNA, but on more motifs (TTC, TGC, and GAC; **Figure 5.6A**). Pm-CDA2 could not be confirmed to be active on any of the GC-rich substrates, while Hs-AID was active on the single-stranded, GC-rich DNA on the following motifs: TAC, TGC, CCC, and AAC (**Figure 5.6B**).

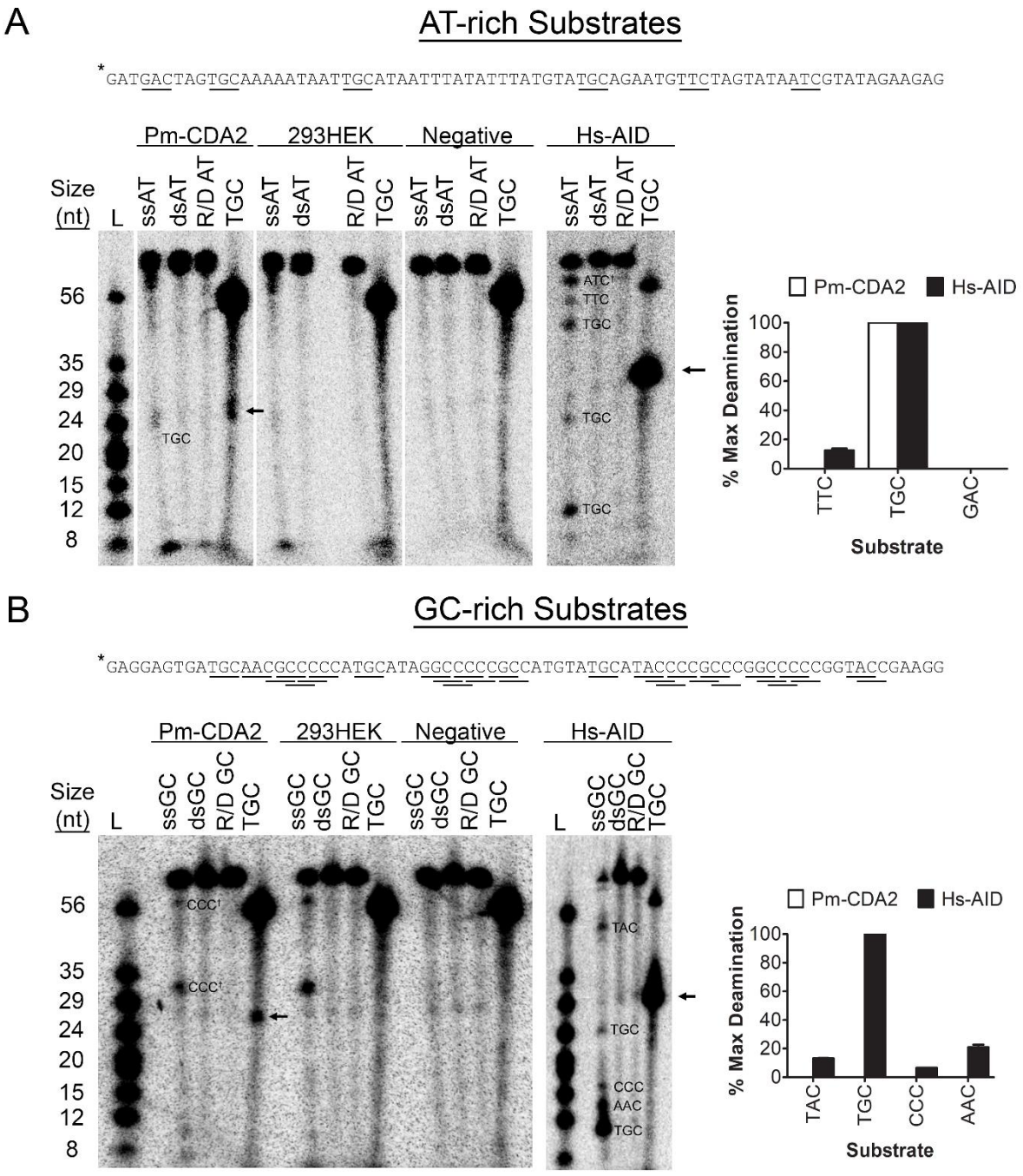


Figure 5.6 Pm-CDA2 is active on single-stranded AT-rich DNA

Legend: Figure 5.6

Left, alkaline cleavage gels of 10 µl Pm-CDA2 and 2 µl Hs-AID incubated at 24 °C for 32 hours with AT- (**A**) and GC-rich (**B**) substrates of different structures and sequences. Right, graphs of % maximum deamination results from their respective graphs; only ss substrates were measured, deamination values from bands of the same motif were combined. The ssDNA sequences of both AT- and GC-rich substrates are written above their respective gels; the star indicates 5' radioactive labelling, all possible XXC motifs are underlined. ssAT/GC, single-stranded AT- or GC-rich substrate, respectively; dsAT/GC, double-stranded AT- or GC-rich substrate, respectively; R/D AT/GC, RNA/DNA AT- or GC-rich substrate, respectively; 293HEK, untransfected 293HEKT cells; Negative, buffer anzd substrate, lacking enzyme. 28 nt TGC product bands are indicated by a black arrow; possible product bands in ssDNA substrate lanes are indicated by the target motif — motifs with † indicate a possible deamination target that could not be confirmed (thus not measured), as there is a band present of the same length in the 293HEK control lanes. Error bars represent SEM. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

5.3.4 Pm-CDA2-His is optimal at physiological pH

As the freshwater lamprey CDA1L1 variants showed unique optimal pH profiles, yet GST-tagged Pm-CDA1 showed a preference for a neutral pH (see Chapter 4),[127] I wanted to determine the optimal pH for the distinct His-tagged Pm-CDA2 enzyme. Thus, Pm-CDA2 was incubated with GTCbub7 at 37 ℃ for 20 hours at a range of pH's from 5.94–8.20, as these conditions and substrate showed the optimal activity with Pm-CDA2. Hs-AID was incubated with TGCbub7 at 37 °C for 20 hours at a range of pH's, as a control (**Figure 5.7**). Both Pm-CDA2 and Hs-AID showed optimal activity around neutral pH (7.1 and 7.6, respectively).

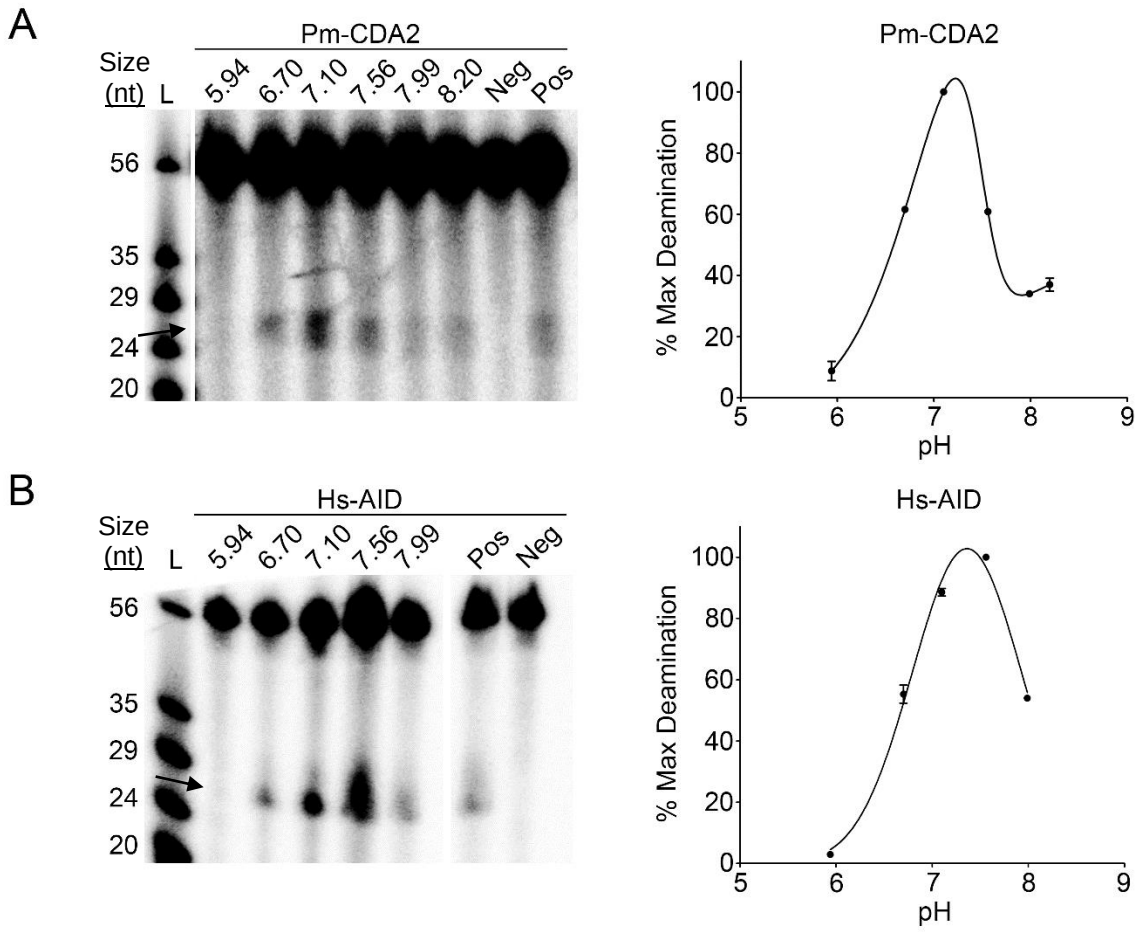


Figure 5.7. His-tagged Pm-CDA2 and Hs-AID were optimally active at neutral pH

A Left, alkaline cleavage gel of Pm-CDA2 incubated with GTCbub7 at 37 °C at a range of pH's for 20 hours (actual pH of buffers indicated). Right, graph of % maximum deamination values analyzed from the gel on the left. **B** Left, alkaline cleavage gel of Hs-AID incubated with TGCbub7 at 37 °C at a range of pH's for 20 hours (actual pH of buffers indicated). Right, graph of % maximum deamination values analyzed from the gel on the left. Pos, positive control, enzyme incubated with substrate in phosphate buffer at pH 7.3, 37 °C; Neg, negative control, just substrate and buffer (pH 7.3) incubated at 37 °C. Error bars represent SD. Product bands (28 nt) are indicated by a black arrow. The white space in **A** denotes a deleted area of the imaged gel that contained irrelevant lanes.

5.3.5 Pm-CDA2-His catalytic activity is much slower than that of Hs-AID

Due to both Pm-CDA2 and Hs-AID showing activity on TGCbub7, albeit at different temperatures, enzyme kinetic experiments were performed on both enzymes on TGCbub7 around their optimal temperatures (24 °C and 34 °C, respectively; **Figure 2.3** and **Figure 5.8**). Pm-CDA2 was not tested on GTCbub7 due to its optimal temperature on this substrate being undetermined. Absolute velocity could not be calculated due to the precise concentration of the enzyme within the cell lysates being unknown. Assuming the concentrations of each enzyme are not greatly different, the kinetics can be compared taking into consideration the relative amounts of enzyme used. Pm-CDA2 lysate was used at 5 times the volume of Hs-AID lysate, thus a factor of 5 was used for the Pm-CDA2 calculations and a factor of 1 was used for Hs-AID calculations.

For the time course kinetics, Pm-CDA2 was incubated for much longer than Hs-AID (up to 48 hours compared to 60 min, respectively) due to the much lower activity of the former observed in previous experiments. Pm-CDA2 activity began to plateau around 48 hours at 0.035 fmol/enzyme unit (**Figure 5.8A**), while Hs-AID activity plateaued ~ 1 hour at 34 fmol/enzyme unit (**Figure 5.8B**). For the substrate concentration kinetics, Pm-CDA2 and Hs-AID were incubated with a range of concentrations of TGCbub7 (**Figure 2.3**) at 24 °C for 48 hours, and 34 °C for 90 min, respectively. Neither enzyme reached saturation at these conditions, which has been seen previously for prokaryotically expressed, GST-Hs-AID[183,184,229] (see also **Figure 3.8**). At the maximum substrate concentration tested for Pm-CDA2 (5.0 fmol/µL), the maximum velocity reached was 0.015 fmol/min for Pm-CDA2 (**Figure 5.8C**) and 37 fmol/min for Hs-AID (**Figure 5.8D**).

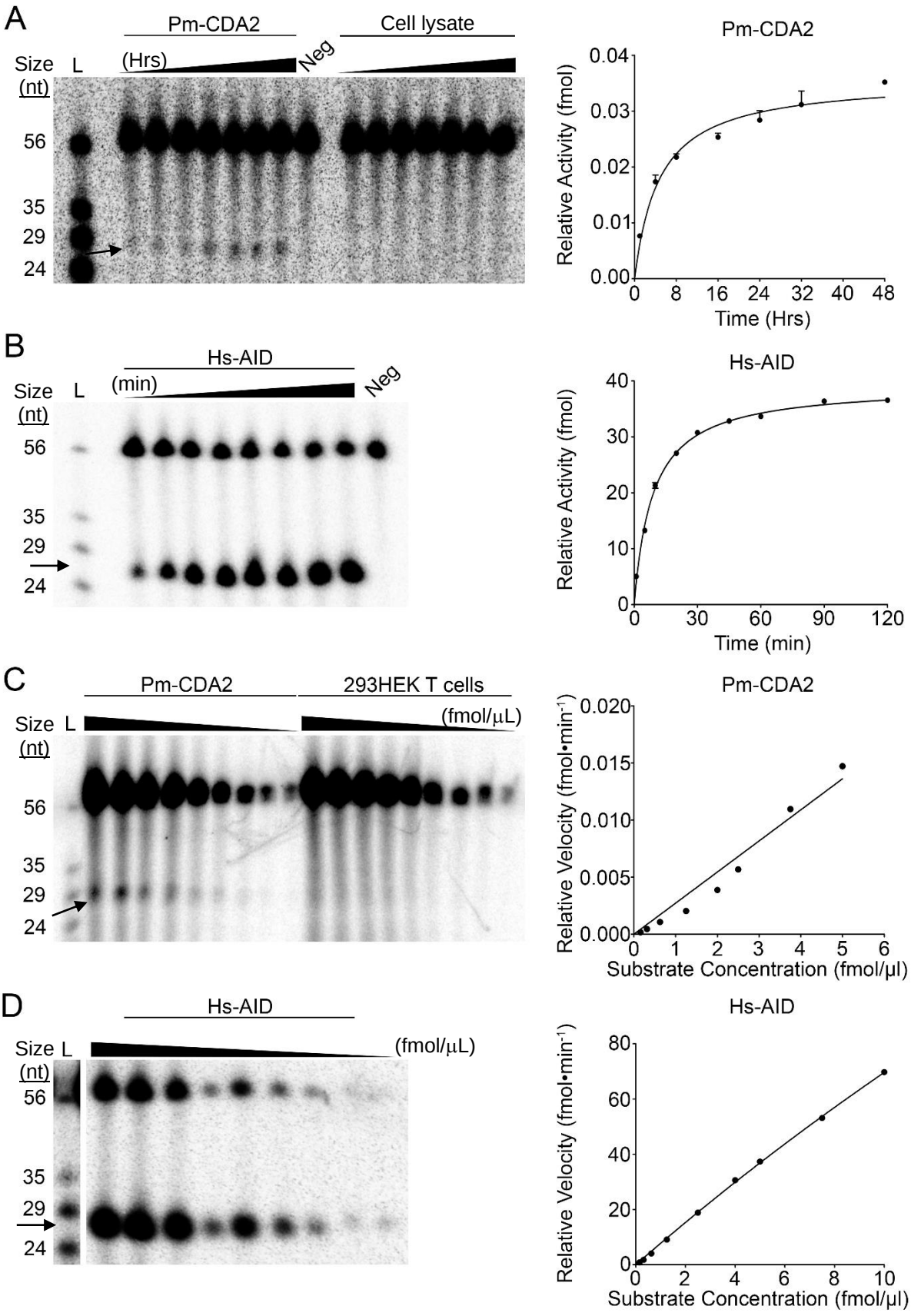


Figure 5.8. His-tagged Hs-AID is vastly more catalytically active than Pm-CDA2

Legend: Figure 5.8

A Time course experiment of Pm-CDA2 and 293HEKT cells lysate (lacking plasmid, negative control) incubated with 50 fmol/µL of TGCbub7 at 24 °C for 1 to 48 hours. Left, alkaline cleavage gel; right, corresponding Relative Velocity versus Time (hours) graph. **B** Time course experiment of Hs-AID incubated with TGCbub7 at 34 °C for 1 to 120 minutes. Left, alkaline cleavage gel; right, corresponding Relative Velocity versus time (minutes) graph. Neg, negative: phosphate buffer and TGCbub7, lacking enzyme. **C** Substrate concentration experiment of Pm-CDA2 and 293HEKT cells lysate incubated with TGCbub7 at 24 °C for 48 hours. Left, alkaline cleavage gel; right, corresponding Relative Velocity versus Substrate Concentration **D** Substrate concentration experiment of Hs-AID incubated with TGCbub7 at 34 °C for 90 minutes. All product bands (28 nt) are indicated by a black arrow. L, ladder; (Hrs), incubation time in hours; (min), incubation time in minutes; (fmol/µL), substrate concentration. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

5.3.6 Potential GST-Pm-CDA2 activity on ssDNA

To determine if the tag and expression system had any effect on Pm-CDA2 activity, Pm-CDA2 was also cloned into the GST-tagged pGEX plasmid and expressed in our prokaryotic expression system in the same manner as described in Chapter 3.[229] Two protein preparations each originating from the transfection of two clones (for a total of four preparations) were purified on GST columns and their purification confirmed on SDS-PAGE gels (**Figure 5.9A**). An initial deamination activity test on TGCbub7 (**Figure 2.3A**; 16 hours at 37 °C) was negative (**Figure 5.9B**). A second activity test was performed on all four GST-Pm-CDA2 preparations and GST-Hs-AID as a positive control, incubated with TGC-, TAC-, and ssTGCbub7 (**Figure 2.3A**) at 25 °C for 16.5 hours. Hs-AID demonstrated activity on all substrates, as expected (**Figure 5.9C–D**). One Pm-CDA2 preparation caused degradation while the others were negative on TGC- and TACbub7 (**Figure 5.9C, D**); all four Pm-CDA2 preparations showed potential activity on TTC and GAC motifs in ssTGCbub7 (**Figure 5.9E)**. Hs-AID and Pm-CDA2 were also incubated with dsTGCbub7 and TG(mC) (**Figure 2.3B** and **C**) at 25 °C for 3 hours; none were active on dsTGCbub7, while only Hs-AID was slightly active on TG(mC), as previously demonstrated (**Figure 5.9E, F**).[184]

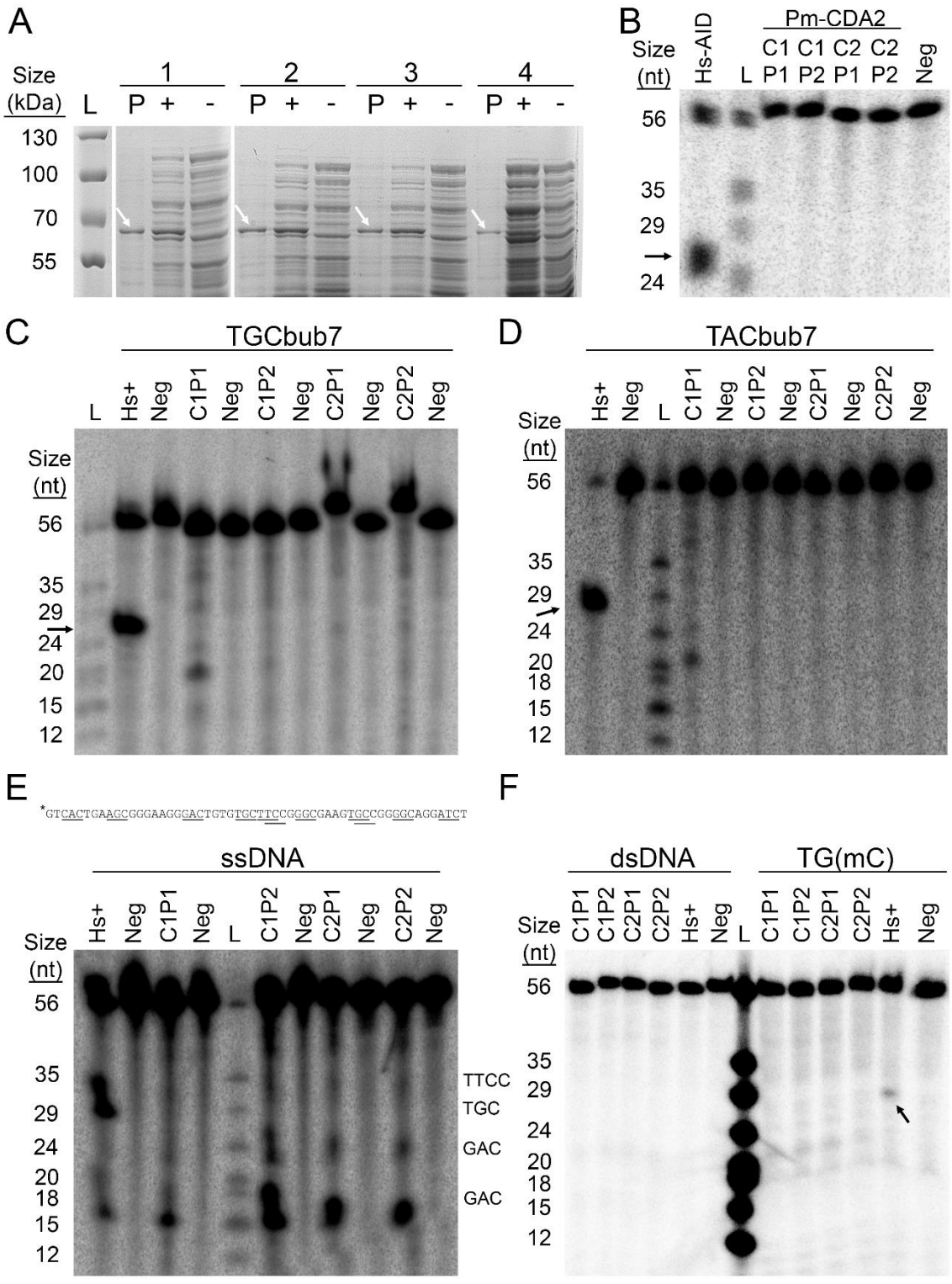


Figure 5.9. GST-Pm-CDA2 activity on various substrates

Legend: Figure 5.9

A SDS PAGE gels of four unique clones, one preparation each, of GST-Pm-CDA2, ~ 64.6 kDa (white arrows). L, ladder; P, purified protein; +, induced cell culture sample; -, uninduced cell culture sample; 1, 2, 3, 4, clone number. **B**-**E** Alkaline cleavage gels of different clones of GST-Pm-CDA2 incubated with 50 fmol of TGCbub7 (**B, C)**, TACbub7 (**D**), ssTGCbub7 (**E**), dsTGCbub7 (**F**), or TG(mC) (**F**), in phosphate buffer pH 7.2, for 3 hours (**F**), 16 hours (**B**), or 16.5 hours (**C, D, E**) at 37 ℃ (**B**) or 25 ℃ (**C, D, E, F**). H+, Hs-AID; Neg, phosphate buffer pH 7.2 and substrate, no enzyme; C#P#, clone # preparation # of Pm-CDA2; L, ladder. In (**E**), letters on right side of gel denote possible cytidine-containing motifs deaminated by Pm-CDA2 and Hs-AID. The ssDNA sequence is found above the gel; the * indicates 5' labelling; and all possible XXC target motifs are underlined. In (**F**), the product band produced by Hs-AID activity on TG(m)C is indicated by a black arrow. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

To determine if, like its His-tagged eukaryotically-expressed counterpart, prokaryotically-expressed GST-tagged Pm-CDA2 substrate specificity was thermosensitive, GST-Pm-CDA2 was tested in a similar experiment as **Figure 5.10** at three different temperatures (18 °C, 25 °C, and 37 °C), on the same three substrates (TGCbub7 [**Figure 5.10A**], TACbub7 [**Figure 5.10B**], and ssTGCbub7 [**Figure 5.10C**]), at two different durations (3 and 16 hours) with a GST-Hs-AID control. The results were similar to those seen in **Figure 5.9**: no GST-Pm-CDA2 activity was seen in the 7 nt bubble substrates, while some possible activity was seen on GAC and TGC motifs in the ssTGCbub7 substrate, while GST-Hs-AID was active on all substrates (**Figure 5.10**).

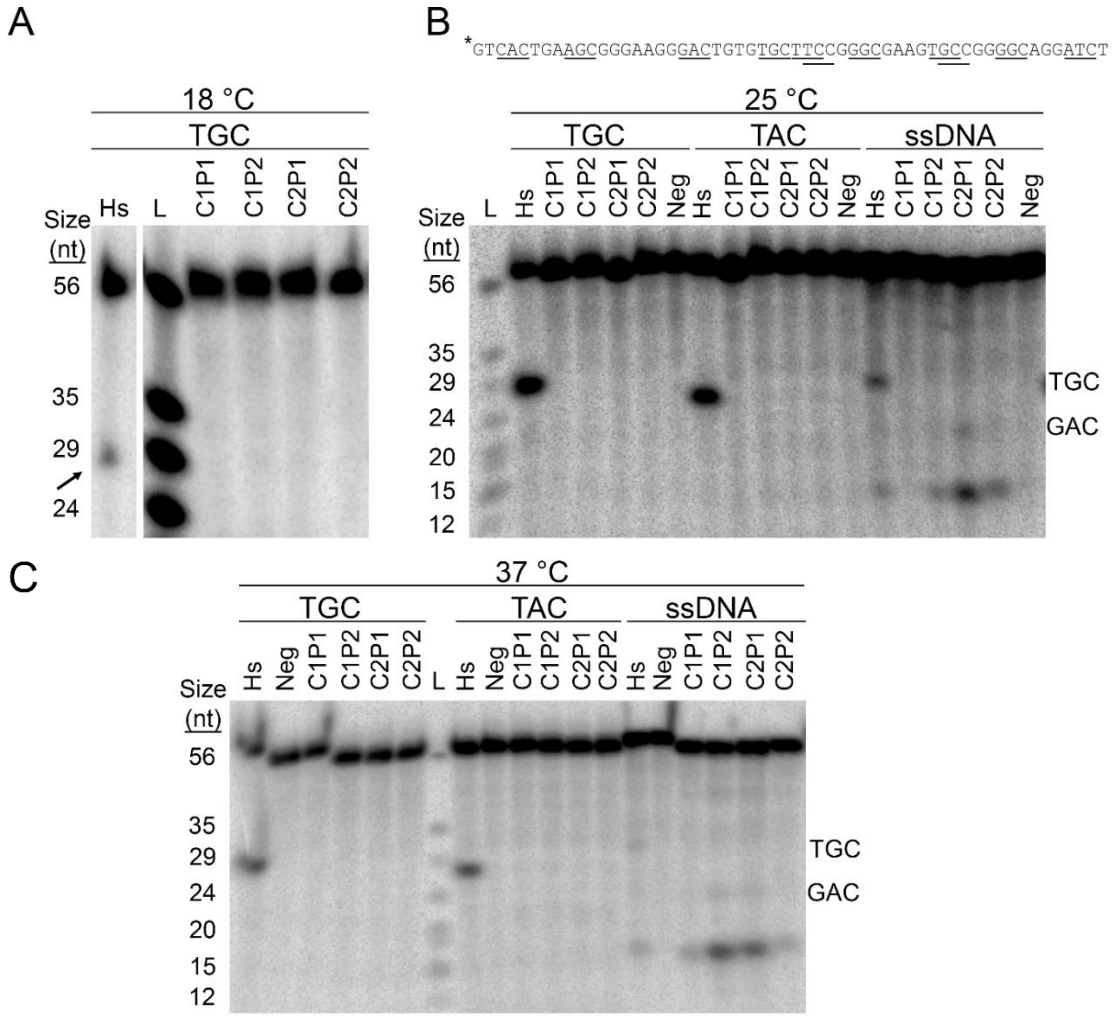


Figure 5.10. GST-Pm-CDA2 activity at different temperatures

Alkaline cleavage gels of four preparations of GST-Pm-CDA2 (C1P1, C1P2, C2P1, and C2P2) and GST-Hs-AID incubated with TGCbub7 (TGC), TACbub7 (TAC) and single-stranded TGCbub7 (ssDNA, see sequence above gel in **B**; *, 5' radioactive labelling; all possible XXC target motifs are underlined) at 18 °C (**A**), 25 °C (**B**), and 37 °C (**C**) for 3 hours (**B, C**) or 16 hours (**A**) in pb pH 7.2. The black arrow in (**A**) indicates the TGC product band at 28 nt. "TGC" and "GAC" on the right side of gels in **B** and **C** indicate possible deaminase activity on those motifs on the ssDNA substrate. The ~18nt bands in the ssDNA lanes are unlikely to be caused by cytidine deaminase activity as there are no likely XXC motifs in the ssDNA substrate. L, ladder. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

Finally, GST-Pm-CDA2 activity was tested in phosphate buffers with different additives. GST-Pm-CDA2 was incubated for 16 hours at 37 °C and 25 °C with TACbub7 and ssTGCbub7 in four buffers containing $ZnCl_2$, RNase A, both, or neither, all at pH 7.2. GST-Pm-CDA2 showed possible activity on TAC and GAC in the $ZnCl_2$ buffer, TTCC and GAC in the RNase A buffer, GGC, TGC, GAC, and AGC in the buffer containing both. GST-Pm-CDA2 may have showed activity on TRC and GAC motifs in the buffer with neither $ZnCl_2$ nor RNase A, but this could not be confirmed as the bands are close to, but not directly, at the correct size (**Figure 5.11**).

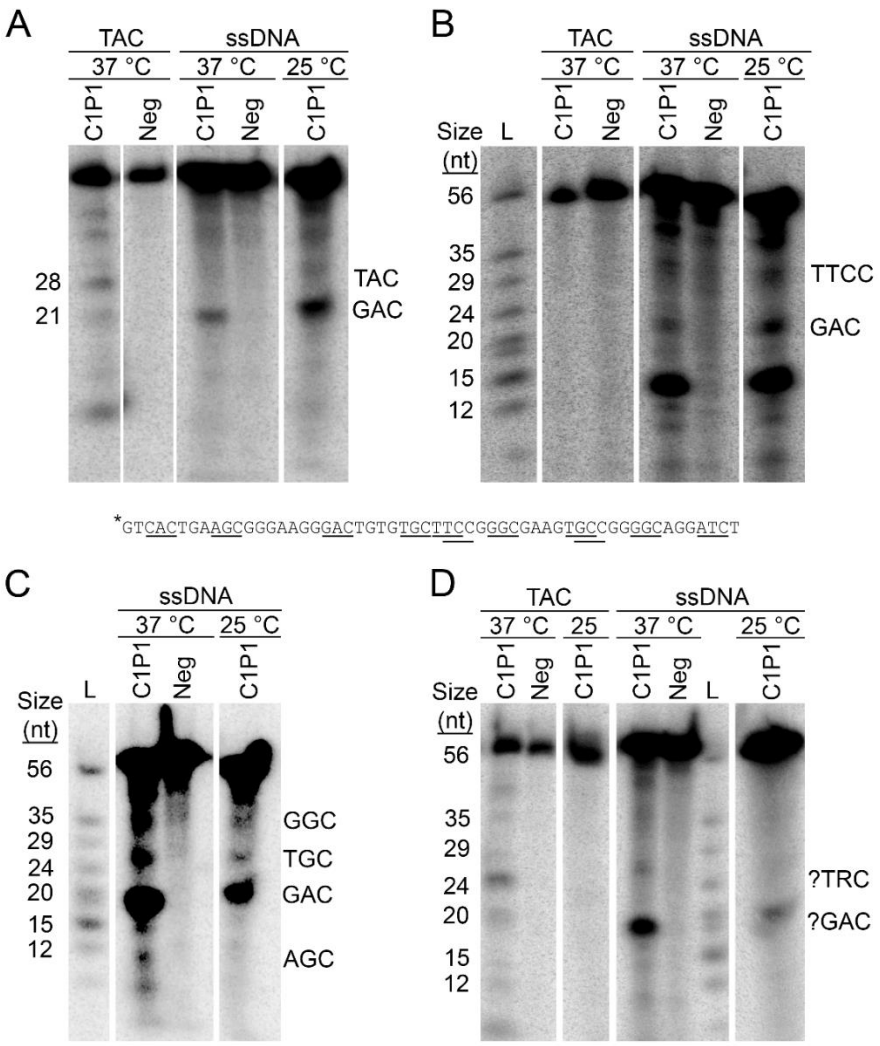


Figure 5.11. GST-Pm-CDA2 activity in different buffers

Alkaline cleavage gels of one preparation of GST-Pm-CDA2 (C1P1) incubated with TACbub7 (TAC) or single-stranded TGCbub7 (ssDNA, see sequence in the middle of all gels; *, 5' radioactive labelling; all possible XXC target motifs are underlined) at 25 °C or 37 °C for 16 hours in 100 mM phosphate buffer with RNase A (**A**), $ZnCl_2$ (**B**), both (**C**), or neither (**D**). Motifs on the right side of the gel indicate possible deamination activity at those cytidines in the TAC or ssDNA substrates; the motifs with the ? indicate possible activity with more uncertainty, as the bands may be near the correct size but not exact. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes. L, ladder; Neg, buffer + substrate, lacking enzyme.

5.3.7 Cytidine deaminase activity of His-tagged Lp-CDA2 proteins remains undetermined

During the same experiments that found CDA1L1 genes in *Lampetra planeri* described in Chapter 4, one close CDA2 ortholog was found in both *L. planeri* and *L. japonica* species, along with three CDA2 variants (**Figure 4.3** and **Figure 5.12**).[127] The Lp-CDA2 ortholog has 84.3% amino acid similarity to Pm-CDA2, whereas the second variant (Lp-CDA2_v2) has 99.4% amino acid similarity to Lp-CDA2. Lp-CDA2_v3.1 and Lp-CDA2_v3.2 are splice variants of Lp-CDA2_v2, with the 4th and last exon of Lp-CDA2_v2 edited and two extra exons added. The *Lp-CDA2_v2, _v3.1,* and *_v3.2* genes were cloned into our pcDNA3.1/V5-8xHis-TOPO plasmid and expressed in our HEK293T cell eukaryotic expression system.

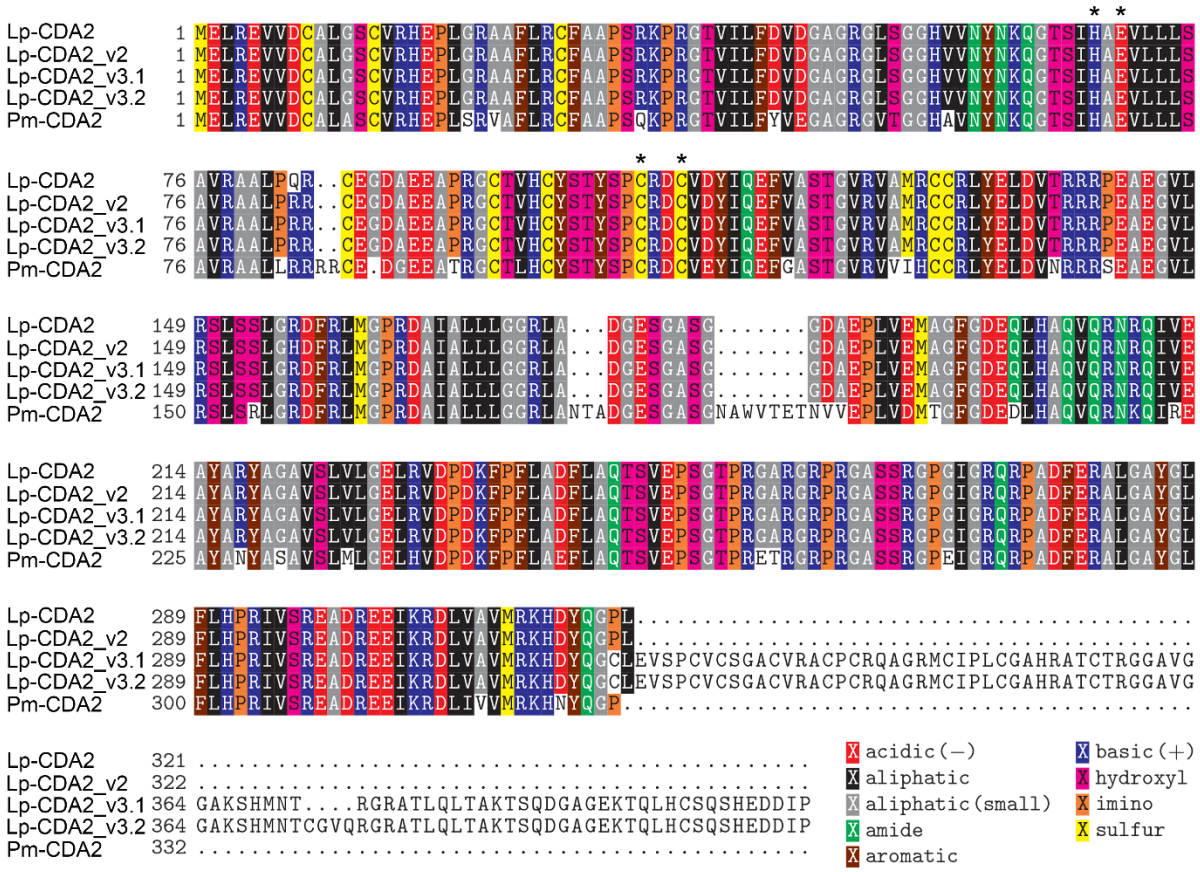


Figure 5.12. Comparison of CDA2 genes from *L. planeri* and *P. marinus*

Alignment of the amino acid sequences of Pm-CDA2, Lp-CDA2, Lp-CDA2 variant 2 (Lp-CDA2_v2), and the splice variants of Lp-CDA2_v2 (Lp-CDA2_v3.1 and Lp-CDA2_v3.2). Potential zinc coordinating amino acids are indicated with a *.

His-tagged Lp-CDA2 ortholog expression in lysates was confirmed by western blot (**Figure 5.13A**), and initial cytidine deaminase activity tests were performed on two preparations (each derived from a separate clone) of each ortholog for 16 hours at pH 6.4 and pH 7.1, 14 °C and 22 °C, on five 7 nt and 11 nt bubble substrates (AGCbub11, TTCbubB, ATCbubB, CTCbubB, TGCbub7, and TACbub7; see **Figure 2.3A**, **D**, **E**). **Figure 5.13B** demonstrates clean negative controls (untransfected HEK293T cell lysate and a "no enzyme" reaction), indicating there is no background deaminase activity in either the lysate or reagents. A Hs-AID-His lysate was transfected and expressed concurrently with the Lp-CDA2 orthologs, and in the same experiment exhibited cytidine deaminase activity on all substrates in all conditions, albeit at varying levels (**Figure 5.13B)**. Neither Lp-CDA2_v2 (**Figure 5.13C**), Lp-CDA2_v3.1 (**Figure 5.13D**), nor Lp-CDA2_v3.2 (**Figure 5.13E**) exhibited any perceivable activity in these conditions on the tested substrates.

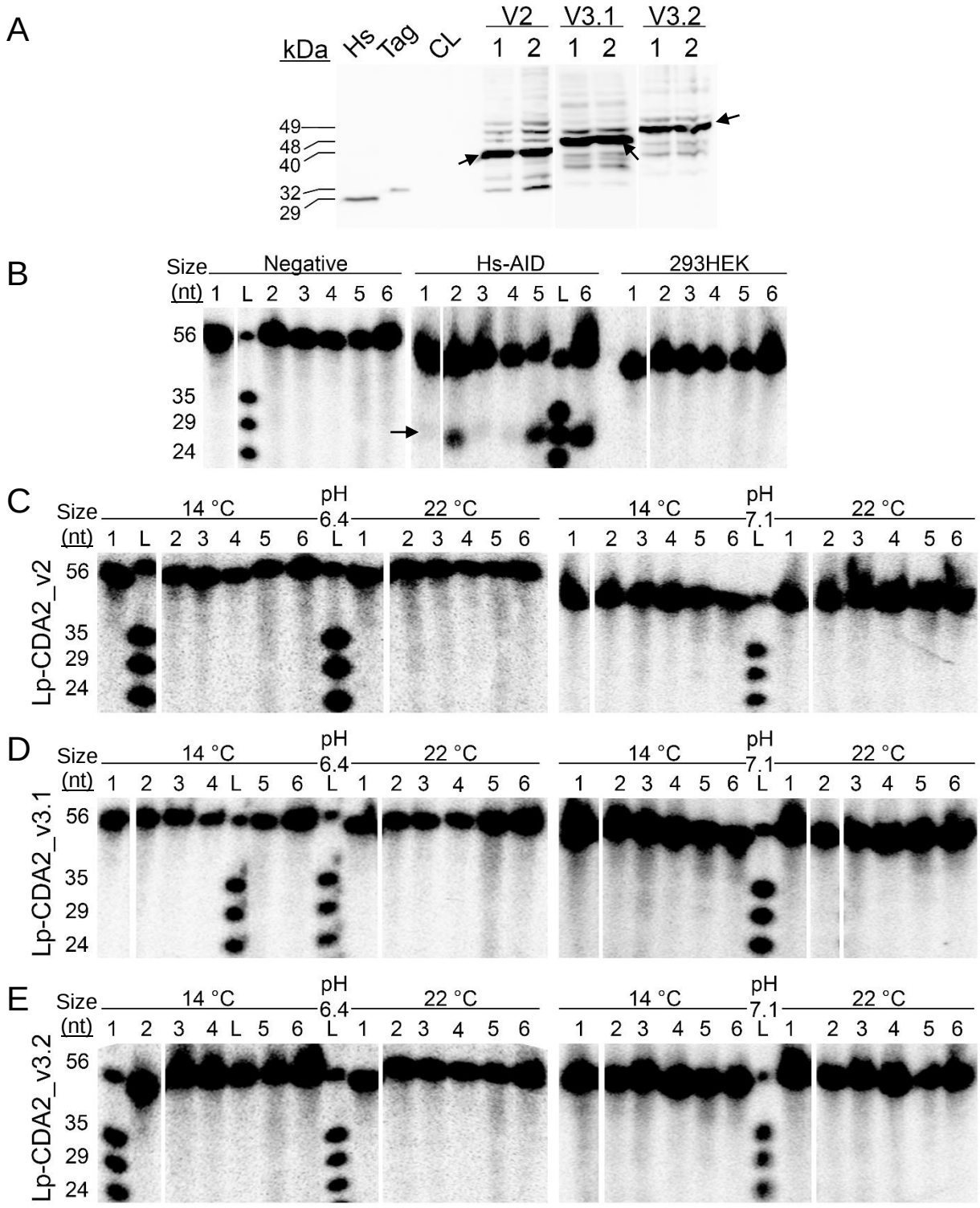


Figure 5.13. Lp-CDA2-His variants were not active in phosphate buffer pH 6.4 or 7.1

Legend: Figure 5.13

A Western blots of Hs-AID-His (Hs, 28.6 kDa), GST-V5-8xHis tag (Tag, 31.8 kDa), cell lysate lacking plasmid (CL), and His-tagged LpCDA2_v2 (V2, 39.6 kDa), LpCDA2_v3.1 (V3.1, 48.3 kDa), and LpCDA2_v3.2 (V3.2, 48.7 kDa), using anti-V5 epitope antibody. The bands of the latter three are indicated by black arrows on the SDS PAGE gel. **B-E** Alkaline cleavage gels of His-tagged enzymes incubated for 16 hours in phosphate buffer pH 6.4 or 7.1 at 14 °C or 22 °C with radioactively labelled AGCbub11 (1), TTCbubB (2), ATCbubB (3), CTCbubB (4), TGCbub7 (5), or TACbub7 (6). **B** Control experiments at pH 6.4, the product bands are indicated by a black arrow; Negative, pb and substrate, lacking enzyme; 293HEK, 293HEKT cell lysate lacking enzyme. **C**, Lp-CDA2_v2-His. **D**, Lp-CDA2_v3.1-His. **E** Lp-CDA2_v3.2-His. L, ladder. The white spaces denote removed areas of the imaged gels that contained empty or irrelevant lanes.

Additional experiments with new and repeated parameters were performed with the His-tagged Lp-CDA2 lysates to test their cytidine deaminase activity. Neither Lp-CDA2 variant was active on TGCbub7 nor GTCbub7 (**Figure 2.3A**) after 16 hours incubation at pH 7.1 at 14 °C (**Figure 5.14A**), which were similar to favourable conditions for Pm-CDA2-His (GTCbub7, 18 °C; **Figure 5.3**). A longer incubation (24 hours) with double the amount of Lp-CDA2 variant enzyme at neutral pH and a range of temperatures (25–37 °C) on different substrates (TGC-, TTC-, TAC-, and GTCbub7 (**Figure 2.3A**) also failed to show any cytidine deaminase activity of the Lp-CDA2 variants, while Hs-AID-His was positive on all substrates and the negative controls were clean (**Figure 5.14B**; Lp-CDA2_v2, **C**; Lp-CDA2_v3.1, **D**; Lp-CDA2_v3.2, **E**).

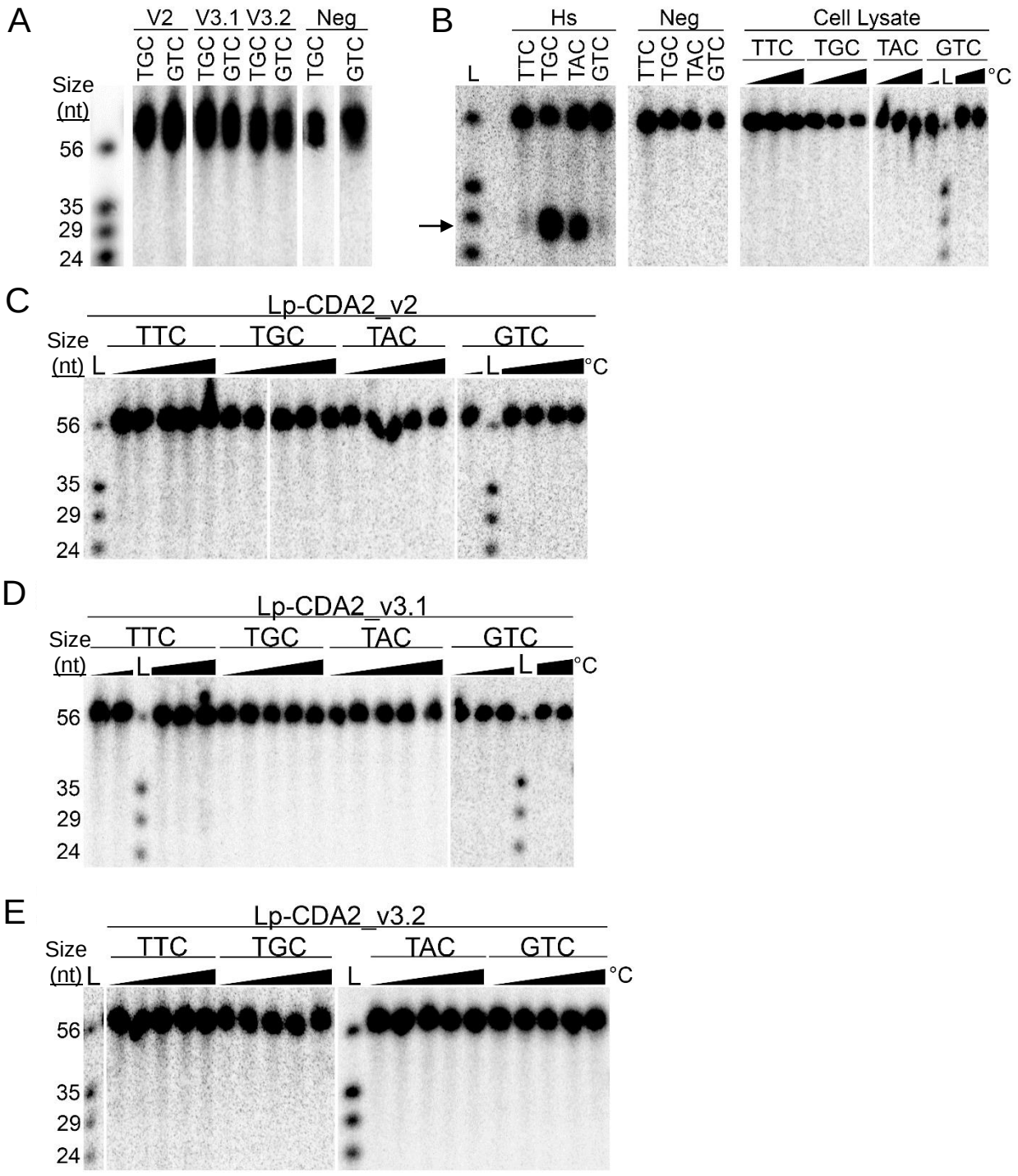


Figure 5.14. Lp-CDA2 orthologs are not active at lower pH on certain 7 bubble oligonucleotides

Legend: Figure 5.14

A Alkaline cleavage gels of 5 µL each of Lp-CDA2_v2 (V2), Lp-CDA2_v3.1 (V3.1), and Lp-CDA2_v3.2 (V3.2) lysates incubated with 50 fmol of radioactively-labelled TGCbub7 and GTCbub7 for 16 hours in pb pH 7.1 at 14 °C. Neg, negative reactions of TGC- and GTCbub7 incubated with just pb pH 7.1 at the same temperature and time. **B-E** Alkaline cleavage gels of 5 µL Hs-AID-His lysate (Hs, product bands indicated by a black arrow), and 10 µL pb pH 5.64 (Negative), untransfected 293HEKT cell lysate lacking enzyme (Cell Lysate, **B**), Lp-CDA2_v2 (LpCDA2v2, **C**), _v3.1 (LpCDA2v3.1, **D**), and _v3.2 (LpCDA2v3.2, **E**) lysate incubated with 50 fmol radioactively-labelled TTC-, TGC-, TAC-, and GTCbub7 in pb pH 5.64 with 0.8 units of Uracil DNA Glycosylase Inhibitor (UGI) for 4 hours at 25 °C (Hs-AID) or 24 hours at 25–37 °C (negative controls and Lp-CDA2 orthologs).

His-tagged, APOBEC proteins expressed in our eukaryotic expression system are often only active in HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) buffer, which contains RNase A and Triton X-100 that work to disassemble RNA-protein interactions that may hinder AID/APOBEC activity. Therefore, to determine if the Lp-CDA2-His variants also require this buffer to show enzymatic activity in our alkaline cleavage assays, the variants were incubated with TTC-, TAC-, GTC-, and AGCbub7 (**Figure 2.3A**) in HEPES buffer pH 7.4 at 16 °C, 25 °C, and 37 °C for 24 hours. Lp-CDA1_1spl was successfully used as a positive control (**Figure 5.15A**). The untransfected 293HEKT cell lysate was negative. A possible band of the correct size was seen with Lp-CDA2_v3.1 incubated with AGCbub7 at 37 °C, while all other Lp-CDA2 reactions were negative (**Figure 5.15B**).

Finally, the Lp-CDA2 variants were incubated with multiple oligonucleotides (TTC-, TGC-, GTC-, and AGCbub7; **Figure 2.3A**) at 18 °C, 25 °C, and 37 °C in phosphate buffer containing UGI to prevent any DNA repair mechanisms remaining in the HEK293T cell lysate from removing any uridines caused by CDA activity and therefore preventing CDA activity from being detected in our alkaline cleavage assay. The Lp-CDA2 variants were also incubated for 60 hours to ensure any minute activity could be detected. Furthermore, a second AGCbub7 reaction was included, which was treated with endonuclease V after AID incubation; if the enzymes acted as adenosine deaminases instead of cytidine deaminases, and deaminated the adenosine to an inosine, the endonuclease V would cleave the phosphodiester bond two positions 3' of the mutated adenosine, resulting in a 27 nt band (see **Figure 2.3A**). AGUbub7 and IGCbub7 were incubated with UDG and endonuclease V, respectively, and then NaOH, resulting in nearly 100% cleavage of the oligonucleotide, providing positive bands for comparison. The Lp-CDA2 variants were not active on any substrate in any of the aforementioned conditions (**Figure 5.15A-D**).

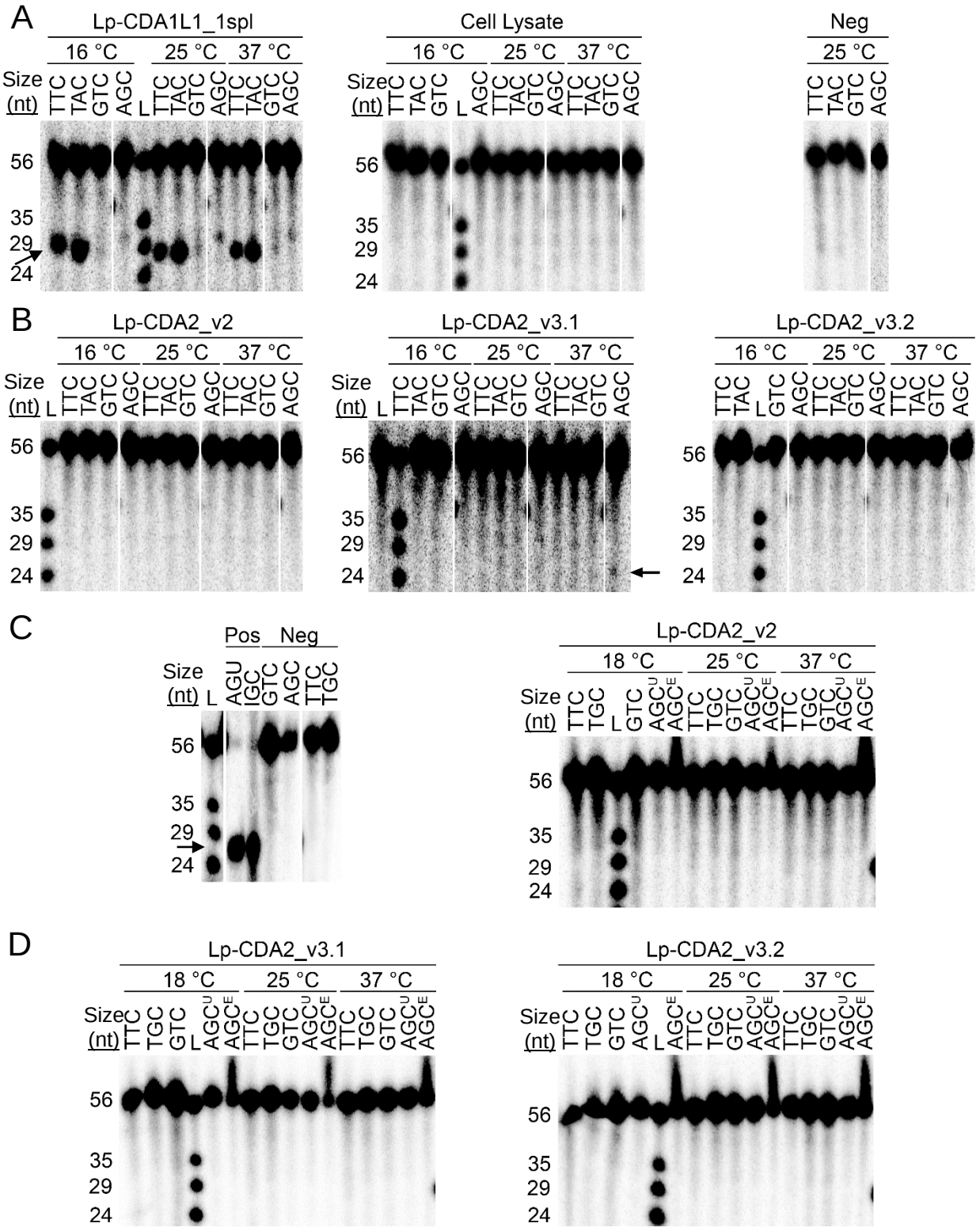


Figure 5.15. Lp-CDA2 lysate cytidine deaminase activity in HEPES buffer and adenosine deamination activity in phosphate buffer

Legend: Figure 5.15

A, B Alkaline cleavage gels of HEPES experiment: 10 µL of Lp-CDA1L1_1spl-His lysate, 293HEKT cell lysate lacking Lp enzyme, HEPES buffer pH 7.4, and Lp-CDA2_v2, _v3.1, and _v3.2 incubated with 50 µL radioactively-labelled TTC-, TAC-, GTC-, and AGCbub7 in HEPES buffer pH 7.4 at 16 °C, 25 °C, and 37 °C for 24 hours. Possible band of correct size indicated by a black arrow in Lp-CDA2_v3.1 gel; product bands of the correct size in the positive gel (**A**) were indicated by a black arrow. **C, D** Alkaline cleavage gels of longer incubation experiment with adenosine deamination test: 10 µL of Lp-CDA2_v2-His, _v3.1-His, and _v3.2-His lysates were incubated with radioactively labelled TTC-, TGC-, GTC-, and AGCbub7 at 18 °C, 25 °C, or 37 °C in pb pH 7.1 and 0.4 units of UGI for 60 hours. The negative samples contained just pb and substrate. The Lp lysates and negative reactions were treated with UDG as previously described; a second reaction with AGCbub7 was treated with 5 units of endonuclease V in NEB buffer #4 at 37 °C for 1 hour. The positive samples included AGUbub7 and IGCbub7 incubated with just UDG or endonuclease V, respectively (product bands indicated by a black arrow).

5.3.8 Cytidine deaminase activity of GST-tagged Lp-CDA2 proteins remains undetermined

As the His-tagged Lp-CDA2 variant lysate cytidine deaminase activity was inconclusive in the parameters tested, GST-tagged Lp-CDA2 variants were expressed and purified in our prokaryotic *E. coli* expression system, as previously described in Chapter 3 and in our previous works.[63,68,183–185,229] Proper expression and purity was confirmed on SDS-PAGE gels (**Figure 5.16A**). As an initial activity test, ~ 0.3 µg of each Lp-CDA2 variant was incubated with TGC- and TACbub7 (**Figure 2.3A**) for 16 hours at 14 °C, 25 °C, and 37 °C. As the temperature increased, the oligonucleotides became more degraded, making it difficult to confirm activity, despite some product bands being around the appropriate size (**Figure 5.16B–D**). Specifically, Lp-CDA2_v2 and Lp-CDA2_v3.1 appear to be active on TACbub7 at 25 °C (**Figure 5.16C**).

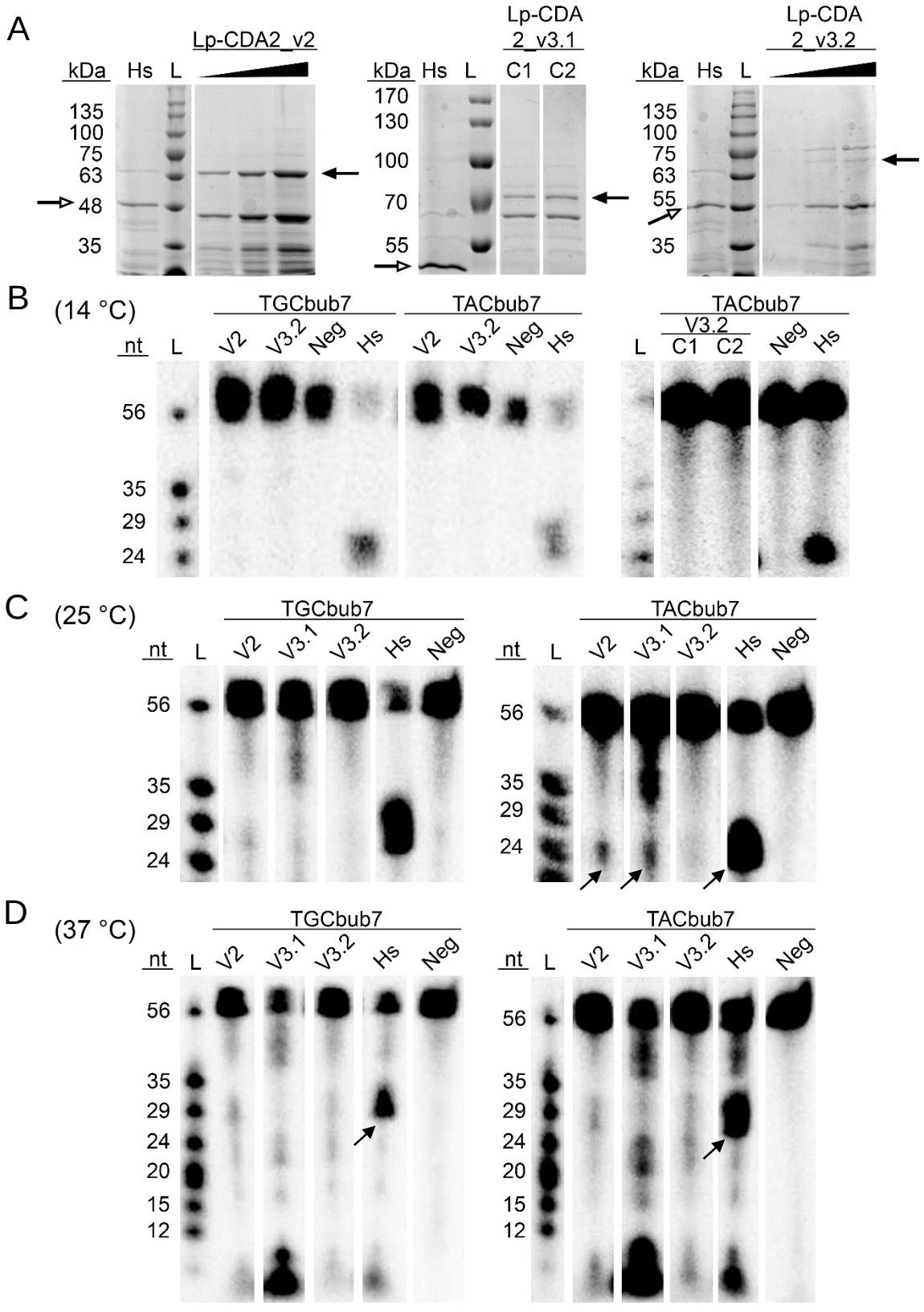


Figure 5.16. GST-tagged Lp-CDA2 variants activity on bubble substrates is inconclusive

Legend: Figure 5.16

A SDS PAGE gels of purified GST-tagged Lp-CDA2_v2 (63 kDa), Lp-CDA2_v3.1 (72 kDa), and Lp-CDA2_v3.2 (72 kDa; black arrow), with GST-Hs-AID for reference (52 kDa, white arrow). **B**-**D** Alkaline cleavage gels of GST-LpCDA2_v2 (V2), _v3.1 (V3.1), and _v3.2 (V3.2) and GST-Hs-AID (Hs) incubated with TGCbub7 and TACbub7 for 16 hrs in pb pH 7.2 at 14 °C (**B**), 25 °C (**C**), and 37 °C (**D**). Neg, substrate and buffer, lacking enzyme. Possible product bands are indicated by a black arrow (the product bands in the Hs-AID lanes are confirmed).

Three further, separate experiments were performed with slightly varying conditions to determine if the GST-tagged and purified Lp-CDA2 were active cytidine deaminases. **Figure 5.17A-C** shows the results of incubating 0.1–0.16 µg of Lp-CDA2_v2, _v3.1, and _v3.2 with five different 7 nt-bubble substrates (CCC-, CTC-, TGC-, GTC-, and TTCbub7; **Figure 2.3A**) radioactively labelled at the 5' or 3' end (as there were multiple C's in some substrates) at 14–20 °C for 8–24 hours. The positive controls included GST-tagged, prokaryotically-expressed, and purified Hs-AID, Pm-CDA1, and Pm-CDA1L1, the latter of which was identified in the same experiments in Chapter 4 that identified Lp-CDA1 variants and only differs from Pm-CDA1 by two nt mutations, one which is silent (a4g, c124a). Lp-CDA2_v2 and Lp-CDA2_v3.1 may have some activity at 20 °C for 16 hours on 3' labelled CCCbub7, but the bands were very faint (**Figure 5.17B**).

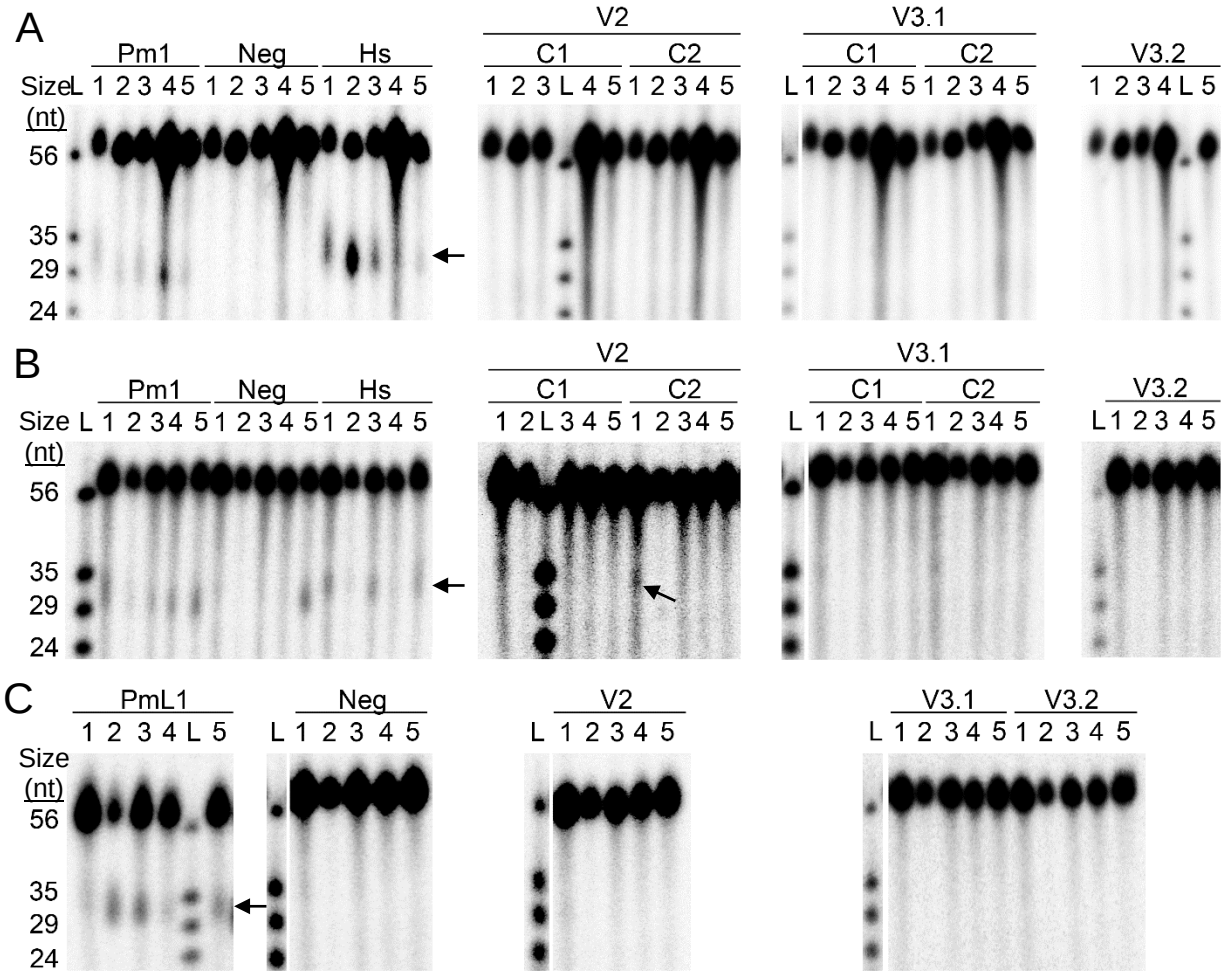


Figure 5.17. Activity of prokaryotic expressed, purified GST-Lp-CDA2 variants could not be determined

Alkaline cleavage gels of three separate experiments, where preparations of 1–2 clones (C1, C2) of GST-tagged, purified Lp-CDA2_v2 (V2), _v3.1 (V3.1), and _3.2 (V3.2), and one preparation each of GST-tagged and purified Hs-AID (Hs), Pm-CDA1 (Pm1), and Pm-CDA1L1 (PmL1) were incubated with 5' (**A**) or 3' (**B, C**) radioactively-labelled CCC- (1), CTC- (2), TGC- (3), GTC- (4), and TTCbub7 (5) in pb pH 7.1 at various temperatures and durations. **A** Lp variants: 0.1 µg, 20 °C, 8 hrs. Hs: 0.12 µg, 31 °C, 1 hr. Pm1: 0.12 µg, 14.5 °C, 4 hrs. **B** Lp variants: 0.16 µg, 20 °C, 16 hrs. Hs: 0.09 µg, 31 °C, 1 hr. Pm1: 0.09 µg, 14.5 °C, 4 hrs. **C** Lp variants: 0.16 µg, 14 °C, 24 hrs. PmL1: 0.17 µg, 14 °C, 2 hrs. **A, B, C** Neg, negative: just pb and substrate, incubated in same conditions as Lp variants. Possible product bands are indicated by a black arrow.

5.4 Discussion

In this chapter, both His- and GST-tagged CDA2 orthologs were expressed and their cytidine deaminase activity tested under various conditions. Multiple preparations of one clone of His-tagged, purified Pm-CDA2 was found to be active on partially single-stranded DNA on WRC motifs and on the non-WRC motif, GTC. Furthermore, Pm-CDA2-His' substrate specificity profile was found to be temperature dependent. This is in contrast to other AID orthologs, where their substrate specificity profile remains consistent; for example, Hs-AID displays a preference for WRC motifs no matter the temperature. This could indicate that Pm-CDA2 is more structurally flexible than other AID orthologs, having multiple temperature-sensitive conformations or "microstates" that are conducive to cytidine deaminase activity: the 25 °C conformations accept WRC motifs into the catalytic pocket in deamination-favourable positions, whereas the conformations at higher temperatures accept GTC into the catalytic pocket.

Sea lamprey can tolerate a wide range of temperatures (1–20 °C) and experience different environmental conditions as they metamorphosize from ammocoetyes to juveniles to adult lamprey.[324,374,375] The structural flexibility of Pm-CDA2-His at various temperatures could reflect unique, temperature-dependent immunological needs. Further biological investigation of lamprey immunity at its various life stages is required to explore this theory.

Maximum Pm-CDA2-His catalytic activity on TGCbub7 appears to be ~ 1000x slower than maximum Hs-AID-His on the same substrate, which is quite a statement considering that Hs-AID is already a relatively "slow" enzyme compared to other enzymes.[226,240,278] However, Pm-CDA2-His appears to exhibit higher activity on GTCbub7 at higher temperatures compared to TGCbub7 at lower temperatures, suggesting the former as its optimal temperature and substrate.

A thermosensitivity experiment with Pm-CDA2-His on GTCbub7 using temperatures beyond 37 °C would likely determine its optimal temperature, and thus kinetics experiments performed at this higher optimal temperature may reveal the true extent to Pm-CDA2-His' activity.

That being said, as a sea lamprey enzyme, Pm-CDA2 is unlikely to ever be at 37 °C or higher; as a DNA-mutating enzyme, Pm-CDA2 could have evolved to exhibit sub-optimal catalytic activity at physiological conditions, thereby protecting the sea lamprey's genomic integrity. On the other hand, the conditions of the alkaline cleavage experiment that demonstrated this higher activity on GTCbub7 at 37 °C could lack other factors that are present *in vivo*, which may affect Pm-CDA2's thermosensitivity profile. For example, a relative of Pm-CDA2, human APOBEC1, requires an accessory protein (ACF) to fold its substrate into a catalytically favourable position at physiological pH, which naturally occurs at higher temperatures.[237] Pm-CDA2 could interact with a second protein which may fold non-WRC containing ssDNA into a more catalytically-favourable form at lower temperatures, or indeed, could influence the structural formation of Pm-CDA2 itself into a conformation similar to that into which it folds alone at higher temperatures. As suggested in Chapter 4, analysis of co-transcribed genes could indicate proteins that may interact with CDA2. A pull-down assay or a BioID assay on cell lysate containing a tagged CDA2 could also identify interacting proteins. In both cases, a lamprey VLRB cell line grown at cooler temperatures (~ 15–20 °C) would be optimal.

It remains unknown why one preparation from one transfection of one clone of His-tagged Pm-CDA2 was active and the others were not. All clones were re-sequenced and were genetically identical, thus the one clone that was active did not carry mutations that could have affected its activity. The western blots of different clones also showed similar band densities, therefore the one clone that was active was not so because it was of a higher concentration than the inactive

clones. It is possible that there could be mutations in other portions of the plasmid that allowed for more efficient post-translational modifications for folding of the protein; to know for sure, the whole plasmid could be sequenced. Regardless, the activity of other CDA2 orthologs suggests that the activity displayed by this clone of Pm-CDA2-His reflects true cytidine deaminase activity.

Multiple preparations of multiple clones of GST-tagged Pm-CDA2 appear to have exhibited cytidine deaminase activity on ssDNA, specifically on GAC, AGC, TTCC, and GGC motifs. This reflects the optimal activity on larger bubble substrates, ssDNA, and both WRC and non-WRC motifs displayed by His-tagged Pm-CDA2. The addition of $ZnCl_2$ appears to have a positive effect on GST-Pm-CDA2 activity; in 2015, Marx *et al.*[376] suggested that zinc, in addition to being required for zinc-dependent cytidine deamination, may play a role in increasing cytidine deaminase activity of A3G by binding to loop 3 of the active C terminal domain, thereby placing Asp244 in the correct orientation to influence substrate binding towards more a more productive position within the catalytic pocket. Like all active enzymes with a ZDD domain, Pm-CDA2 has an Asp at a similar position in loop 3 (N59); the addition of $ZnCl_2$ to the alkaline cleavage reaction may have ameliorated GST-Pm-CDA2 activity. Further experimentation with $ZnCl_2$ buffers and N59 mutants can better address this hypothesis.

Both His-tagged and GST-tagged Lp-CDA2 variants were inactive in the parameters tested. Future experiments using parameters similar to those tested with GST-Pm-CDA2, such as the addition of $ZnCl_2$ to the buffer and with more ssDNA and larger bubble DNA substrates, may reveal cytidine deaminase activity. What has yet to be tested on all of these potential enzymes is their activity on RNA. Fellow relatives of the AID/APOBEC family human A1 and the yeast CDD1 have demonstrated activity on both RNA and DNA.[132,271] One way to test this theory would be a similar assay to the alkaline cleavage experiment, instead using fluorescently-labelled RNA,

which could be translated and labelled immediately prior to incubation with the enzyme to prevent degradation. Another way would be to incubate the enzyme with RNA, reverse transcribe said RNA after incubation, and sequence the resulting DNA; however, this method would need to take superfluous mutations caused by the reverse transcriptase into account.

In summation, the results of these experiments suggest that Pm-CDA2 may be an active cytidine deaminase on ssDNA, with relatively low catalytic activity and a unique substrate specificity profile. These data are supported by previous works which demonstrated that Pm-CDA2 is expressed by activated VLRB+ B cell-like cells and is required for VLRB antibody assembly.[370,373] To my knowledge, this is the first report of Pm-CDA2 being an active enzyme, giving further credence to the theory of lamprey adaptive immunity involving multiple AID-like proteins. There is much more work that must be done to further confirm Pm-CDA2's activity and to more thoroughly biochemically characterize this interesting enzyme.

Chapter 6 Discussion

6.1 Summary of results

The research described in this dissertation has provided novel insight into a number of AID orthologs. Chapter 3 focused on four AID enzymes from species spanning fish evolution: CDA1 in the lamprey, one of the few remaining jawless vertebrates; AID in the nurse shark, a cartilaginous fish, which evolved from an early-diverged gnathostome lineage; pufferfish AID, containing the ray-finned fish-specific region in its protein sequence; and AID in the coelacanth, one of our closest fish ancestors before the divergence of the tetrapods.[148,280,293] Studying AID orthologs from this wide range of species, each with unique adaptive immunity characteristics, enabled us to view AID evolution on a macro scale. While each species has different antibody loci organization, and while two lack canonical CSR, a vital antibody diversification mechanism initiated by AID in most jawed vertebrates, AID orthologs from all four fish species were active cytidine deaminases on partially single-stranded bubble substrates with low enzymatic activity.[229] These common characteristics indicate that since the divergence of jawed vertebrates, AID has maintained sub-optimal activity on structurally similar substrates, likely as a safeguard against promiscuous mutagenic activity. The different characteristics of these AID orthologs, such as their unique thermosensitivity and substrate specificity profiles, suggest that AID has co-evolved with the thermoregulatory needs and Ig loci of the animal in which it is expressed.

Chapters 4 and 5 focused on AID orthologs in the lamprey, an evolutionarily significant animal with a unique adaptive immune system. Both the lamprey and hagfish, the only two remaining extant jawless vertebrates, exhibit characteristics of adaptive immunity, such as delayed type hypersensitivity and an increased immune response upon second immunization, yet they lack

the canonical pillars of jawed vertebrate immunity, such as RAG and Ig proteins.[293] Upon discovery of VLRs, it became apparent that jawless vertebrates have lymphocyte-like cells which act similarly to gnathostome B and T cells. The presence of AID-like proteins expressed in these cells further suggests that these lamprey CDAs may act in a similar role to gnathostome AID: mutating antibody loci to initiate their diversification. Surprisingly, multiple CDAs were found in the lamprey, a rare characteristic in jawed vertebrates, most of whom carry only one AID gene with multiple APOBEC genes.[127,129] Furthermore, while it appeared that the presence of either *CDA1* and *CDA1L1_4* was necessary, each individual had a unique combination of the remaining CDA genes, similar to allelic differences in APOBEC genes across human populations.[127,377] Combined with the difference in surface charges between CDA1L1_4 and the remaining CDA1L1s, which is similar to the difference in surface charges between AID and APOBEC3, respectively, these findings suggest that these CDAs may play a similar role in lamprey adaptive and innate immunity as AID and the APOBEC3s, respectively.[127]

Further investigation into the biochemical characterization of the CDA1L1 orthologs revealed that all were active cytidine deaminases with optimal activity on a non-WRC substrate, TTC. These results confirm our previous work which described Pm-CDA1 as having a unique substrate preference for a non-WRC motif,[229] suggesting the target *in vivo* sequence for lamprey CDAs may be different from that of gnathostome AID, which uniformly prefer WRC motifs. Each enzyme exhibited unique thermosensitivity and optimal pH profiles, which was unexpected due to these proteins belonging to the same animal. These different characteristics suggest these enzymes occupy unique intracellular locations, similar to how AID is active in the nucleus and APOBECs are active in the cytoplasm.[108,293] An additional possibility is that these enzymes are active at

different developmental stages of the lamprey, which are dependent on environmental temperature changes.

The *in vitro* characterization of CDA2 and its variants also identified unique properties of these lamprey AID orthologs, despite being somewhat inconclusive. Preparations from one transfection of Pm-CDA2-His expressed in our eukaryotic system displayed temperature-dependent substrate specificity profiles: a novel characteristic never before seen in any AID ortholog. GST-tagged, prokaryotically-expressed Pm-CDA2 also displayed unique substrate preferences for both WRC and non-WRC motifs, however, its activity could not be confirmed with the same certainty as that of Pm-CDA2-His. Both GST-tagged and the aforementioned His-tagged preparations of Pm-CDA2 appeared to be even less active than Hs-AID, which is already thousands of times less active than the average enzyme.[278] None of the His-tagged, eukaryotically-expressed Lp-CDA2 enzymes were active in our alkaline cleavage experiments. Combined, these data are the first to suggest that CDA2 is an active cytidine deaminase.

6.2 Limitations

There are a few notable limitations of the methodologies used in this thesis. First, all enzymes were expressed in bacterial or human cells, not the same species from which the proteins originate. It can thus be assumed that these proteins were expressed in cellular conditions (such as salt, pH, etc.) foreign to their natural state, which may have affected their folding and/or post-translational modification, the latter of which is experienced by all of the studied proteins but is lacking in our bacterial expression system. Furthermore, these cells produce the target protein at either 16 or 37 °C, which is not the temperature at which most of these poikilothermic species reside, which may again affect protein expression, post-translational modification, and folding, and therefore biochemical activity. Thus, it will be important to purify these proteins in cells and conditions closer to their natural state in order to ensure that the results seen in this thesis have not been affected by the purification conditions.

Another limiting factor of this research is the fact that all proteins were tagged with either eight histidines at the C terminus or GST at the N terminus. GST's large size and histidine's positive charge at neutral pH could influence biochemical activity by obstructing or changing the confirmation of the catalytic pocket, or by affecting protein dimerization. I have already ameliorated this concern by creating and successfully expressing Hs-, Dr-, and Gc-AID with both tags, with limited differences seen between them. That being said, the results in this thesis could be confirmed by expressing the CDA1L1 proteins with an N terminal GST tag. I could also confirm that the tags did not greatly affect enzymatic activity by switching the terminal at which the tag is attached (i.e., 8xHis at the N terminus and GST at the C terminus), the idea being that perhaps one end of the protein is more important then the other in oligomerization and could be affected differently by either tag.

Another limitation of my *in vitro* investigations is that I did not explore all biochemical properties, such as optimal salt concentration, possible cofactors, or all substrate conditions (more bubble sizes, different DNA formations such as stem loops, etc.). Therefore, I could have missed optimal conditions in which to see enzymatic activity from those proteins that displayed unconfirmed or no activity. Increasing the number of parameters tested would certainly give us a more detailed view of the biochemical properties of these AID orthologs. However, to produce the most relevant results in a timely manner, I investigated the most commonly explored enzyme characteristics for AIDs (temperature, incubation time, substrate concentration, WRC-motif substrate preferences); this enabled comparison between previous work done my myself, my colleagues, and others in the field. That being said, I did include properties that are not commonly studied when investigating AID and AID ortholog biochemical activity, such as pH, which revealed interesting similarities between the different LpCDA1s and the AID/APOBECs (see Chapter 4).

A limitation of our *in silico* 3D protein modelling is that the current system requires the use of mouse and human APOBEC models as bases upon which the algorithm builds the 3D models of both human and non-human AID orthologs. I have already demonstrated in chapter 3 that the different AID orthologs are genetically dissimilar to human AID, which suggests that using models of human and mouse APOBEC may inadvertently create inaccurate, "humanized" 3D models of these fish AID orthologs. My colleagues have demonstrated that it is possible to produce an accurate human AID 3D model based on mouse and human APOBEC models, despite the genetic differences between human and mice, and AID and APOBECs[252]; however, *in silico* modeling can be improved by further XRC and NMR imaging of more non-human enzymes to more accurately create non-human, non-mouse protein structures.

Finally, while the ethos of this thesis is that biochemical characterization of enzymes using *in vitro* and *in silico* methods is a worthy endeavour, I also argue for a more comprehensive study of proteins, which also includes cellular and *in vivo* research. Therefore, one large limitation of this thesis is that it only includes the former methods while excluding the latter. That being said, there were many novel discoveries described in this thesis that could only have been found using *in vitro* and *in silico* methods, such as the difference in optimal pH for the LpCDA1s and the temperature-dependent substrate specificity of Pm-CDA2-His. This final limitation is discussed further in Section 6.4.

6.3 Future Directions

The work described herein has revealed that AID orthologs in fish species have vastly different biochemical characteristics that have likely evolved in tandem with their heterogenetic immune receptor targets, along with their intra- and extracellular environments, such as a lower basal temperature in poikilotherms. This is the first work to have shown an AID ortholog preferring a non-WRC motif or having such a cold-adaptive temperature preference. These results provide groundwork for future experiments with other AID orthologs, such as considering the basal body temperature of the animal to whom the AID ortholog belongs when designing experiments and expanding the substrate motifs tested beyond the classic WRC motifs. The biochemical characteristics that remained the same between the AID orthologs, such as lethargic activity and high binding affinity, may also direct how future work on AID is performed, such as considering longer incubation times to ensure any minute activity is captured. In other words, this work has expanded the expected conditions in which AID is believed to be active and encourages future work to consider broader possibilities for this enzyme.

The differences and similarities between the biochemical characteristics of AID orthologs can reveal a lot about the different roles these enzymes play; for example, while zebrafish AID is assuredly responsible for the mutations that initiate secondary antibody diversification mechanisms, its unique ability to mutate methylated C's suggests a second role in animal development, a revelation that was only possible due to *in vitro* experimentation.[184–186] Further investigation into comparative enzymology regarding these immunologically relevant enzymes, such as the newly discovered SNADs and NADs,[124] may continue to uncover new roles for these proteins outside of antibody diversification. One potential avenue is determining if SNADs are secreted, and if so, are they active mutators in the extracellular environment; secreted mutators

may serve a role in innate immunity against transposons or other foreign genetic material, similar to primate APOBEC3G.

The *in silico* work regarding the fish AID structures in Chapter 3 has confirmed another biochemical characteristic consistent between AID orthologs: an often-closed catalytic pocket that rarely opens to allow for substrate deamination. Combined with demonstrated high substrate binding affinities, these data explain the low catalytic rates of all AID orthologs. I have suggested in Chapter 3 and in the associated published work[229] that these properties have been maintained across hundreds of millions of years of evolution to protect the genomic integrity of the animal. To confirm this hypothesis, *in situ* and *in vivo* experiments could be developed using upmutants of AID orthologs to determine if more active versions would result in increased DNA damage compared to the naturally occurring AIDs.

The shape of the catalytic pocket, also determined by *in silico* experimentation, can also inform which substrates the AID ortholog prefers, such as with the lamprey Pm-CDA1, which appears to have differences in loops 2 and 8, and a more compact catalytic pocket compared to human AID. This computational modeling methodology may be optimized to the point in which new AID orthologs can be tested for their optimal substrates *in silico,* thus narrowing down the vast numbers of possible substrates for these potentially promiscuous enzymes before *in vitro* experiments, saving time and resources.

In Chapter 4 I not only discovered an expansion of CDA1-like enzymes in the freshwater lamprey, similar to that seen in primate A3s,[129] but that these enzymes exhibited different catalytic rates, temperature preferences, and optimal pH's. While the different optimal pH's can be perhaps explained by the different cellular localizations of these enzymes, such as in lysosomes versus the

cytosol or nucleus, the reason behind different temperature preferences is more difficult to theorize. Is it simply a matter of evolution towards suboptimal activity for genomic protection, or are there other factors at play? Furthermore, why do some individuals have certain CDA1L1s and not others? To solve these conundrums, cellular and whole-organism work will have to be completed. Cell culture of lamprey lymphocyte-like cells will need to be optimized and the cells possibly immortalized. I will then need to determine how to activate VLRA/C^{+} cells grown in culture to express CDA1L1 proteins or create an inducible cell line. Performing pull-down or BioID assays may reveal protein co-factors that influence the CDA1L1 orthologs in different ways, such as altering the optimal temperature of the enzyme-substrate interaction. CDA1L1 variant relationships with these cofactors would be assessed in *in vitro* assays similar to those described in this thesis. Why certain individuals have different CDA1L1 orthologs can only be elucidated through housing lampreys in the lab and sequencing their genomes before and after immunological and/or developmental events. Lamprey have been notoriously difficult to maintain in the lab, but recent work[373] has given hope that soon we will have another model animal organism to work with. As previously stated, *in vitro* and *in silico* work, while having more advantages than perhaps most in the biomedical field give them credit for, can only go so far; by combining what we know from *in vitro* and *in silico* work with *in vivo* and cellular experimentation, we can gain better insight into these AID orthologs.

Unlike CDA1 and CDA1L1 orthologs, CDA2 has proven to be an elusive subject of study. While one clone of His-tagged Pm-CDA2 exhibited obvious deaminase activity with a unique temperature-sensitive substrate preference, all other clones of Pm-CDA2, and both His- and GST-tagged CDA2 ortholog enzymes showed inconclusive activity. The one consistency is that a majority of the substrates on which CDA2 activity may have occurred were non-WRC substrates,

similar to the CDA1 enzymes, and were on ssDNA and 7–13 nt sized bubbles. The addition of $ZnCl_2$ also appeared to improve GST-Pm-CDA2 activity. Further *in vitro* experimentation with more ssDNA and larger-bubble substrates, with the inclusion of $ZnCl_2$, may reveal more optimal conditions for CDA2 activity. Pull-down or BioID assays of Pm-CDA2 in lamprey T cell-like cells could reveal potential co-factors that, when combined *in vitro* with purified Pm-CDA2, could improve enzyme activity enough to allow for better protein characterization.

As mentioned above regarding CDA1 orthologs, improved *in silico* research into the 3D structure of CDA2 enzymes may reveal why the substrates used thus far have not been very successful. *In silico* modeling of CDA2 has been difficult due to it not aligning well with known APOBEC models; perhaps investigating other models on which to base the CDA2 structure, such as adenosine deaminases, may improve modeling results. XRC and NMR imaging of non-mammalian AID and APOBEC proteins could improve *in silico* modelling as well by providing a base that is closer to CDA2's true structure than human APOBEC.

The last future direction with great potential is testing all aforementioned AID orthologs for deaminase activity on RNA. Most APOBECs are active on RNA, and many other AID/APOBEC orthologs have displayed promiscuous activity on both RNA and DNA substrates[132,271]; it is possible that the AID orthologs described in this dissertation may also show substrate promiscuity or a singular preference for RNA, which could suggest a secondary role similar to human APOBEC3s in viral immunity or oncogenesis.[108]

6.4 In support of *in vitro* and *in silico* enzyme characterization

The AID/APOBEC family is an excellent case-in-point for the main argument of this thesis, which is that in order to best understand an enzyme, all aspects of its character must be investigated, including its orthologs in non-human species, using multiple methodologies, such as *in vitro* and *in silico* work. For example, lamprey likely never reach an internal body temperature of 37 °C or above, yet *in vitro* work using conditions beyond what is found in nature has revealed that Pm-CDA2-His achieves higher catalytic activity on a non-WRC motif at these temperatures compared to WRC motifs, a first in the AID literature. These results, combined with the finding that CDA1 and its variants prefer non-WRC motifs, have the potential to shift the current zeitgeist in AID research, which focuses on WRC motifs when looking for AID activity with regards to antibody mutations or oncogenesis. By focusing on WRC motifs alone, researchers could be missing possible AID activity, especially when investigating non-human AIDs.

Without resorting to time-consuming and expensive crystallization processes that require mutation of the protein, *in silico* modelling efficiently solved the CDA1L1 variant structures, which suggested that CDA1L1 proteins have different optimal pH profiles. These data were confirmed using *in vitro* experimentation not possible using cell culture or *in vivo* work, due to lethal pH changes, and revealed that these various CDA1L1 variants likely play vastly different roles in lamprey immunity, such as with mammalian AID and APOBECs. This novel theory is a major contribution to the ongoing study of apparent convergent evolution of agnathan and gnathostome immune proteins.

The results described in this dissertation thus have far-reaching implications throughout AID and enzyme research, expanding what was once thought possible with these enzymes while

further elucidating the evolution of adaptive immunity. These findings highlight the advantages of *in silico* and *in vitro* methodologies, findings that will provide the foundations for further *in situ* and *in vivo* work. This dissertation demonstrates that *in silico* and *in vitro* experimentation of non-human enzymes provides immense value to protein characterization and biological science as a whole, and that this process should be just as highly valued as *in vivo* study of human biology.

6.5 Conclusion

While this work has elucidated multiple biochemical characteristics of many AID orthologs, these results have also invoked plenty of questions about the nature of these proteins and their role in adaptive immunity and AIS evolution. The answers to these questions will come through a multidimensional investigation, involving the study of various orthologs from different non-human species using multiple methodologies. As this dissertation has shown, AID research still has many surprises left to discover.

Chapter 7 References

1. Owen J, Punt J, Stranford S, Jones P. *Kuby Immunology*. 7th ed. W. H. Freeman and Company; 2013.

2. Paul WE. *Fundamental Immunology*. Lippincott Williams & Wilkins; 2003.

3. Nimmerjahn F, Ravetch JV. Antibody-mediated modulation of immune responses. *Immunol Rev*. 2010;236:265-275. doi:10.1111/j.1600-065X.2010.00910.x

4. Sun JC, Beilke JN, Lanier LL. Adaptive immune features of natural killer cells. *Nature*. 2009;457(7229):557-561. doi:10.1038/nature07665

5. Dunkelberger JR, Song WC. Complement and its role in innate and adaptive immune responses. *Cell Res*. 2010;20(1):34-50. doi:10.1038/cr.2009.139

6. Schroeder HW, Cavacini L. Structure and function of immunoglobulins. *J Allergy Clin Immunol*. 2010;125(2, Supplement 2):S41-S52. doi:10.1016/j.jaci.2009.09.046

7. Dudley DD, Chaudhuri J, Bassing CH, Alt FW. Mechanism and Control of V(D)J Recombination versus Class Switch Recombination: Similarities and Differences. In: Alt FW, ed. *Advances in Immunology*. Vol 86. Academic Press; 2005:43-112. doi:10.1016/S0065-2776(04)86002-4

8. Little AJ, Matthews A, Oettinger M, Roth DB, Schatz DG. Chapter 2 - The Mechanism of V(D)J Recombination. In: Alt FW, Honjo T, Radbruch A, Reth M, eds. *Molecular Biology of B Cells (Second Edition)*. Academic Press; 2015:13-34. doi:10.1016/B978-0-12-397933-9.00002-3

9. Allman D, Pillai S. Peripheral B cell subsets. *Curr Opin Immunol*. 2008;20(2):149-157. doi:10.1016/j.coi.2008.03.014

10. Herzenberg LA, Herzenberg LA. Toward a layered immune system. *Cell*. 1989;59(6):953-954. doi:10.1016/0092-8674(89)90748-4

11. Montecino-Rodriguez E, Dorshkind K. B-1 B Cell Development in the Fetus and Adult. *Immunity*. 2012;36(1):13-21. doi:10.1016/j.immuni.2011.11.017

12. Pieper K, Grimbacher B, Eibel H. B-cell biology and development. *J Allergy Clin Immunol*. 2013;131(4):959-971. doi:10.1016/j.jaci.2013.01.046

13. Durandy A, Revy P, Imai K, Fischer A. Hyper-immunoglobulin M syndromes caused by intrinsic B-lymphocyte defects. *Immunol Rev*. 2005;203:67-79. doi:10.1111/j.0105-2896.2005.00222.x

14. Erickson RP. Somatic gene mutation and human disease other than cancer. *Mutat Res Mutat Res*. 2003;543(2):125-136. doi:10.1016/S1383-5742(03)00010-3

15. Zan H, Casali P. Regulation of Aicda expression and AID activity. *Autoimmunity*. 2013;46(2):83. doi:10.3109/08916934.2012.749244

16. The European Bioinformatics Institute. Cytidine Deaminase. Mechanism and Catalytic Site Atlas. Published 2020. Accessed October 11, 2020. https://www.ebi.ac.uk/thornton-srv/m-csa/entry/97/

17. Feng Y, Seija N, D I Noia JM, Martin A. AID in Antibody Diversification: There and Back Again. *Trends Immunol*. 2020;41(7):586-600. doi:10.1016/j.it.2020.04.009

18. Tran TH, Nakata M, Suzuki K, et al. B cell-specific and stimulation-responsive enhancers derepress Aicda by overcoming the effects of silencers. *Nat Immunol*. 2010;11(2):148-154. doi:10.1038/ni.1829

19. Yadav A, Olaru A, Saltis M, Setren A, Cerny J, Livák F. Identification of a ubiquitously active promoter of the murine activation-induced cytidine deaminase (AICDA) gene. *Mol Immunol*. 2006;43(6):529-541. doi:10.1016/j.molimm.2005.05.007

20. Dedeoglu F, Horwitz B, Chaudhuri J, Alt FW, Geha RS. Induction of activation-induced cytidine deaminase gene expression by IL-4 and CD40 ligation is dependent on STAT6 and NFkappaB. *Int Immunol*. 2004;16(3):395-404. doi:10.1093/intimm/dxh042

21. Gonda H, Sugai M, Nambu Y, et al. The balance between Pax5 and Id2 activities is the key to AID gene expression. *J Exp Med*. 2003;198(9):1427-1437. doi:10.1084/jem.20030802

22. Park SR, Zan H, Pal Z, et al. HoxC4 binds to the promoter of the cytidine deaminase AID gene to induce AID expression, class-switch DNA recombination and somatic hypermutation. *Nat Immunol*. 2009;10(5):540-550. doi:10.1038/ni.1725

23. Dorsett Y, McBride KM, Jankovic M, et al. MicroRNA-155 suppresses activation-induced cytidine deaminase-mediated Myc-Igh translocation. *Immunity*. 2008;28(5):630-638. doi:10.1016/j.immuni.2008.04.002

24. de Yébenes VG, Belver L, Pisano DG, et al. miR-181b negatively regulates activation-induced cytidine deaminase in B cells. *J Exp Med*. 2008;205(10):2199-2206. doi:10.1084/jem.20080579

25. Borchert GM, Holton NW, Larson ED. Repression of human activation induced cytidine deaminase by miR-93 and miR-155. *BMC Cancer*. 2011;11:347. doi:10.1186/1471-2407-11-347

26. Wu X, Darce JR, Chang SK, Nowakowski GS, Jelinek DF. Alternative splicing regulates activation-induced cytidine deaminase (AID): implications for suppression of AID mutagenic activity in normal and malignant B cells. *Blood*. 2008;112(12):4675-4682. doi:10.1182/blood-2008-03-145995

27. Iacobucci I, Lonetti A, Messa F, et al. Different isoforms of the B-cell mutator activation-induced cytidine deaminase are aberrantly expressed in BCR-ABL1-positive acute lymphoblastic leukemia patients. *Leukemia*. 2010;24(1):66-73. doi:10.1038/leu.2009.197

28. Albesiano E, Messmer BT, Damle RN, Allen SL, Rai KR, Chiorazzi N. Activation-induced cytidine deaminase in chronic lymphocytic leukemia B cells: expression as multiple forms in a dynamic, variably sized fraction of the clone. *Blood*. 2003;102(9):3333-3339. doi:10.1182/blood-2003-05-1585

29. Methot SP, Di Noia JM. Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. *Adv Immunol*. 2017;133:37-87. doi:10.1016/bs.ai.2016.11.002

30. 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;108(45):18366-18371. doi:10.1073/pnas.1106729108

31. Methot SP, Litzler LC, Trajtenberg F, et al. Consecutive interactions with HSP90 and eEF1A underlie a functional maturation and storage pathway of AID in the cytoplasm. *J Exp Med*. 2015;212(4):581-596. doi:10.1084/jem.20141157

32. Wang Q, Kieffer-Kwon KR, Oliveira TY, et al. The cell cycle restricts activation-induced cytidine deaminase activity to early G1. *J Exp Med*. 2017;214(1):49-58. doi:10.1084/jem.20161649

33. Casellas R, Basu U, Yewdell WT, Chaudhuri J, Robbiani DF, Di Noia JM. Mutations, kataegis and translocations in B cells: understanding AID promiscuous activity. *Nat Rev Immunol*. 2016;16(3):164-176. doi:10.1038/nri.2016.2

34. Le Q, Maizels N. Activation-induced deaminase (AID) localizes to the nucleus in brief pulses. *PLoS Genet*. 2019;15(2):e1007968. doi:10.1371/journal.pgen.1007968

35. Patenaude AM, Orthwein A, Hu Y, et al. Active nuclear import and cytoplasmic retention of activation-induced deaminase. *Nat Struct Mol Biol*. 2009;16(5):517-527. doi:10.1038/nsmb.1598

36. Aoufouchi S, Faili A, Zober C, et al. Proteasomal degradation restricts the nuclear lifespan of AID. *J Exp Med*. 2008;205(6):1357-1368. doi:10.1084/jem.20070950

37. Brar SS, Watson M, Diaz M. Activation-induced cytosine deaminase (AID) is actively exported out of the nucleus but retained by the induction of DNA breaks. *J Biol Chem*. 2004;279(25):26395-26401. doi:10.1074/jbc.M403503200

38. Ito S, Nagaoka H, Shinkura R, et al. Activation-induced cytidine deaminase shuttles between nucleus and cytoplasm like apolipoprotein B mRNA editing catalytic polypeptide 1. *Proc Natl Acad Sci U S A*. 2004;101(7):1975-1980. doi:10.1073/pnas.0307335101

39. 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;199(9):1235-1244. doi:10.1084/jem.20040373

40. Betz AG, Milstein C, González-Fernández A, Pannell R, Larson T, Neuberger MS. Elements regulating somatic hypermutation of an immunoglobulin kappa gene: critical role for the intron enhancer/matrix attachment region. *Cell*. 1994;77(2):239-248. doi:10.1016/0092-8674(94)90316-6

41. Manis JP, Tian M, Alt FW. Mechanism and control of class-switch recombination. *Trends Immunol*. 2002;23(1):31-39. doi:10.1016/s1471-4906(01)02111-1

42. Parsa JY, Ramachandran S, Zaheen A, et al. Negative supercoiling creates single-stranded patches of DNA that are substrates for AID-mediated mutagenesis. *PLoS Genet*. 2012;8(2):e1002518. doi:10.1371/journal.pgen.1002518

43. Branton SA, Ghorbani A, Bolt BN, Fifield H, Berghuis LM, Larijani M. Activation-induced cytidine deaminase can target multiple topologies of double-stranded DNA in a transcription-independent manner. *FASEB J Off Publ Fed Am Soc Exp Biol*. 2020;34(7):9245-9268. doi:10.1096/fj.201903036RR

44. Kodgire P, Mukkawar P, North JA, Poirier MG, Storb U. Nucleosome stability dramatically impacts the targeting of somatic hypermutation. *Mol Cell Biol*. 2012;32(10):2030-2040. doi:10.1128/MCB.06722-11

45. Kuang FL, Luo Z, Scharff MD. H3 trimethyl K9 and H3 acetyl K9 chromatin modifications are associated with class switch recombination. *Proc Natl Acad Sci U S A*. 2009;106(13):5288-5293. doi:10.1073/pnas.0901368106

46. Kato L, Begum NA, Burroughs AM, et al. Nonimmunoglobulin target loci of activation-induced cytidine deaminase (AID) share unique features with immunoglobulin genes. *Proc Natl Acad Sci U S A*. 2012;109(7):2479-2484. doi:10.1073/pnas.1120791109

47. Jeevan-Raj BP, Robert I, Heyer V, et al. Epigenetic tethering of AID to the donor switch region during immunoglobulin class switch recombination. *J Exp Med*. 2011;208(8):1649-1660. doi:10.1084/jem.20110118

48. Qian J, Wang Q, Dose M, et al. B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell*. 2014;159(7):1524-1537. doi:10.1016/j.cell.2014.11.013

49. Meng FL, Du Z, Federation A, et al. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell*. 2014;159(7):1538-1548. doi:10.1016/j.cell.2014.11.014

50. Aida M, Hamad N, Stanlie A, Begum NA, Honjo T. Accumulation of the FACT complex, as well as histone H3.3, serves as a target marker for somatic hypermutation. *Proc Natl Acad Sci U S A*. 2013;110(19):7784-7789. doi:10.1073/pnas.1305859110

51. Romanello M, Schiavone D, Frey A, Sale JE. Histone H3.3 promotes IgV gene diversification by enhancing formation of AID-accessible single-stranded DNA. *EMBO J*. 2016;35(13):1452-1464. doi:10.15252/embj.201693958

52. Wang Q, Oliveira T, Jankovic M, et al. Epigenetic targeting of activation-induced cytidine deaminase. *Proc Natl Acad Sci U S A*. 2014;111(52):18667-18672. doi:10.1073/pnas.1420575111

53. Hu Y, Ericsson I, Doseth B, Liabakk NB, Krokan HE, Kavli B. Activation-induced cytidine deaminase (AID) is localized to subnuclear domains enriched in splicing factors. *Exp Cell Res*. 2014;322(1):178-192. doi:10.1016/j.yexcr.2014.01.004

54. 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;12(2):160-166. doi:10.1038/ni.1977

55. 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;31(4):474-484. doi:10.1016/j.molcel.2008.07.009

56. Kanehiro Y, Todo K, Negishi M, 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;109(4):1216-1221. doi:10.1073/pnas.1120368109

57. Kawaguchi Y, Nariki H, Kawamoto N, et al. SRSF1-3 contributes to diversification of the immunoglobulin variable region gene by promoting accumulation of AID in the nucleus. *Biochem Biophys Res Commun*. 2017;485(2):261-266. doi:10.1016/j.bbrc.2017.02.097

58. Kumar Singh A, Tamrakar A, Jaiswal A, et al. Splicing regulator SRSF1-3 that controls somatic hypermutation of IgV genes interacts with topoisomerase 1 and AID. *Mol Immunol*. 2019;116:63-72. doi:10.1016/j.molimm.2019.10.002

59. Singh AK, Tamrakar A, Jaiswal A, Kanayama N, Kodgire P. SRSF1-3, a splicing and somatic hypermutation regulator, controls transcription of IgV genes via chromatin regulators SATB2, UBN1 and histone variant H3.3. *Mol Immunol*. 2020;119:69-82. doi:10.1016/j.molimm.2020.01.005

60. Pham P, Malik S, Mak C, Calabrese PC, Roeder RG, Goodman MF. AID-RNA polymerase II transcription-dependent deamination of IgV DNA. *Nucleic Acids Res*. 2019;47(20):10815-10829. doi:10.1093/nar/gkz821

61. Methot SP, Litzler LC, Subramani PG, et al. A licensing step links AID to transcription elongation for mutagenesis in B cells. *Nat Commun*. 2018;9(1):1248. doi:10.1038/s41467-018-03387-6

62. Pavri R, Gazumyan A, Jankovic M, et al. Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell*. 2010;143(1):122-133. doi:10.1016/j.cell.2010.09.017

63. King JJ, Manuel CA, Barrett CV, et al. Catalytic pocket inaccessibility of activation-induced cytidine deaminase is a safeguard against excessive mutagenic activity. *Struct Lond Engl 1993*. 2015;23(4):615-627. doi:10.1016/j.str.2015.01.016

64. 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;161(4):762-773. doi:10.1016/j.cell.2015.03.020

65. Ribeiro de Almeida C, Dhir S, Dhir A, et al. RNA Helicase DDX1 Converts RNA G-Quadruplex Structures into R-Loops to Promote IgH Class Switch Recombination. *Mol Cell*. 2018;70(4):650-662.e8. doi:10.1016/j.molcel.2018.04.001

66. Qiao Q, Wang L, Meng FL, Hwang JK, Alt FW, Wu H. AID Recognizes Structured DNA for Class Switch Recombination. *Mol Cell*. 2017;67(3):361-373.e4. doi:10.1016/j.molcel.2017.06.034

67. Pavri R. R Loops in the Regulation of Antibody Gene Diversification. *Genes*. 2017;8(6). doi:10.3390/genes8060154

68. 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*. 2018;93:94-106. doi:10.1016/j.molimm.2017.11.012

69. Liu M, Duke JL, Richter DJ, et al. Two levels of protection for the B cell genome during somatic hypermutation. *Nature*. 2008;451(7180):841-845. doi:10.1038/nature06547

70. Álvarez-Prado ÁF, Pérez-Durán P, Pérez-García A, et al. A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. *J Exp Med*. 2018;215(3):761-771. doi:10.1084/jem.20171738

71. 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;280(5370):1750-1752. doi:10.1126/science.280.5370.1750

72. Pasqualucci L, Migliazza A, Fracchiolla N, et al. BCL-6 mutations in normal germinal center B cells: evidence of somatic hypermutation acting outside Ig loci. *Proc Natl Acad Sci U S A*. 1998;95(20):11816-11821. doi:10.1073/pnas.95.20.11816

73. Küppers R, Dalla-Favera R. Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene*. 2001;20(40):5580-5594. doi:10.1038/sj.onc.1204640

74. Senigl F, Maman Y, Dinesh RK, et al. Topologically Associated Domains Delineate Susceptibility to Somatic Hypermutation. *Cell Rep*. 2019;29(12):3902-3915.e8. doi:10.1016/j.celrep.2019.11.039

75. Klein IA, Resch W, Jankovic M, et al. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell*. 2011;147(1):95-106. doi:10.1016/j.cell.2011.07.048

76. Goldman JM, Melo JV. Chronic myeloid leukemia--advances in biology and new approaches to treatment. *N Engl J Med*. 2003;349(15):1451-1464. doi:10.1056/NEJMra020777

77. Wong S, Witte ON. The BCR-ABL story: bench to bedside and back. *Annu Rev Immunol*. 2004;22:247-306. doi:10.1146/annurev.immunol.22.012703.104753

78. Gostissa M, Yan CT, Bianco JM, Cogné M, Pinaud E, Alt FW. Long-range oncogenic activation of Igh-c-myc translocations by the Igh 3' regulatory region. *Nature*. 2009;462(7274):803-807. doi:10.1038/nature08633

79. Potter M. Neoplastic development in plasma cells. *Immunol Rev*. 2003;194:177-195. doi:10.1034/j.1600-065x.2003.00061.x

80. Robbiani DF, Bothmer A, Callen E, et al. AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. *Cell*. 2008;135(6):1028-1038. doi:10.1016/j.cell.2008.09.062

81. Takizawa M, Tolarová H, Li Z, et al. AID expression levels determine the extent of cMyc oncogenic translocations and the incidence of B cell tumor development. *J Exp Med*. 2008;205(9):1949-1957. doi:10.1084/jem.20081007

82. Ramiro AR, Jankovic M, Eisenreich T, et al. AID is required for c-myc/IgH chromosome translocations in vivo. *Cell*. 2004;118(4):431-438. doi:10.1016/j.cell.2004.08.006

83. 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;120(14):2864-2867. doi:10.1182/blood-2012-02-412791

84. Heath E, Begue-Pastor N, Chaganti S, et al. Epstein-Barr virus infection of naïve B cells in vitro frequently selects clones with mutated immunoglobulin genotypes: implications for virus biology. *PLoS Pathog*. 2012;8(5):e1002697. doi:10.1371/journal.ppat.1002697

85. Epeldegui M, Hung YP, McQuay A, Ambinder RF, Martínez-Maza O. Infection of human B cells with Epstein-Barr virus results in the expression of somatic hypermutation-inducing molecules and in the accrual of oncogene mutations. *Mol Immunol*. 2007;44(5):934-942. doi:10.1016/j.molimm.2006.03.018

86. Kalchschmidt JS, Bashford-Rogers R, Paschos K, et al. Epstein-Barr virus nuclear protein EBNA3C directly induces expression of AID and somatic mutations in B cells. *J Exp Med*. 2016;213(6):921-928. doi:10.1084/jem.20160120

87. Grande BM, Gerhard DS, Jiang A, et al. Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. *Blood*. 2019;133(12):1313-1324. doi:10.1182/blood-2018-09-871418

88. Kasar S, Kim J, Improgo R, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun*. 2015;6:8866. doi:10.1038/ncomms9866

89. Pasqualucci L, Guglielmino R, Houldsworth J, et al. Expression of the AID protein in normal and neoplastic B cells. *Blood*. 2004;104(10):3318-3325. doi:10.1182/blood-2004-04-1558

90. Duquette ML, Pham P, Goodman MF, Maizels N. AID binds to transcription-induced structures in c-MYC that map to regions associated with translocation and hypermutation. *Oncogene*. 2005;24(38):5791-5798. doi:10.1038/sj.onc.1208746

91. Richter J, Schlesner M, Hoffmann S, et al. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat Genet*. 2012;44(12):1316-1320. doi:10.1038/ng.2469

92. Lohr JG, Stojanov P, Lawrence MS, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc Natl Acad Sci U S A*. 2012;109(10):3879-3884. doi:10.1073/pnas.1121343109

93. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat Med*. 2018;24(5):679-690. doi:10.1038/s41591-018-0016-8

94. Jiao J, Jin Y, Zheng M, et al. AID and TET2 co-operation modulates FANCA expression by active demethylation in diffuse large B cell lymphoma. *Clin Exp Immunol*. 2019;195(2):190-201. doi:10.1111/cei.13227

95. Nagaoka H, Tran TH, Kobayashi M, Aida M, Honjo T. Preventing AID, a physiological mutator, from deleterious activation: regulation of the genomic instability that is associated with antibody diversity. *Int Immunol*. 2010;22(4):227-235. doi:10.1093/intimm/dxq023

96. Bödör C, Bognár A, Reiniger L, et al. Aberrant somatic hypermutation and expression of activation-induced cytidine deaminase mRNA in mediastinal large B-cell lymphoma. *Br J Haematol*. 2005;129(3):373-376. doi:10.1111/j.1365-2141.2005.05454.x

97. Nakamura M, Sugita K, Sawada Y, Yoshiki R, Hino R, Tokura Y. High levels of activation-induced cytidine deaminase expression in adult T-cell leukaemia/lymphoma. *Br J Dermatol*. 2011;165(2):437-439. doi:10.1111/j.1365-2133.2011.10342.x

98. Muñoz DP, Lee EL, Takayama S, 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;110(32):E2977-2986. doi:10.1073/pnas.1301021110

99. Che Z, Fan J, Zhou Z, et al. Activation-Induced Cytidine Deaminase Expression Facilitates the Malignant Phenotype and Epithelial-to-Mesenchymal Transition in Clear Cell Renal Cell Carcinoma. *DNA Cell Biol*. 2020;39(7):1299-1312. doi:10.1089/dna.2019.5119

100. Qiu X, Zhu X, Zhang L, et al. Human epithelial cancers secrete immunoglobulin g with unidentified specificity to promote growth and survival of tumor cells. *Cancer Res*. 2003;63(19):6488-6495.

101. Kimoto Y. Expression of heavy-chain constant region of immunoglobulin and T-cell receptor gene transcripts in human non-hematopoietic tumor cell lines. *Genes Chromosomes Cancer*. 1998;22(1):83-86. doi:10.1002/(sici)1098-2264(1998)22:1<83::aid-gcc12>3.0.co;2-o

102. Komori J, Marusawa H, Machimoto T, et al. Activation-induced cytidine deaminase links bile duct inflammation to human cholangiocarcinoma. *Hepatol Baltim Md*. 2008;47(3):888-896. doi:10.1002/hep.22125

103. Endo Y, Marusawa H, Kinoshita K, et al. Expression of activation-induced cytidine deaminase in human hepatocytes via NF-kappaB signaling. *Oncogene*. 2007;26(38):5587-5595. doi:10.1038/sj.onc.1210344

104. Kou T, Marusawa H, Kinoshita K, et al. Expression of activation-induced cytidine deaminase in human hepatocytes during hepatocarcinogenesis. *Int J Cancer*. 2007;120(3):469-476. doi:10.1002/ijc.22292

105. Hale LP. Deficiency of activation-induced cytidine deaminase in a murine model of ulcerative colitis. *PloS One*. 2020;15(9):e0239295. doi:10.1371/journal.pone.0239295

106. Vartanian JP, Henry M, Marchio A, et al. Massive APOBEC3 editing of hepatitis B viral DNA in cirrhosis. *PLoS Pathog*. 2010;6(5):e1000928. doi:10.1371/journal.ppat.1000928

107. Machida K, Cheng KTN, Sung VMH, et al. Hepatitis C virus induces a mutator phenotype: Enhanced mutations of immunoglobulin and protooncogenes. *Proc Natl Acad Sci U S A*. 2004;101(12):4262-4267. doi:10.1073/pnas.0303971101

108. Borzooee F, Joris KD, Grant MD, Larijani M. APOBEC3G Regulation of the Evolutionary Race Between Adaptive Immunity and Viral Immune Escape Is Deeply Imprinted in the HIV Genome. *Front Immunol*. 2018;9:3032. doi:10.3389/fimmu.2018.03032

109. Liang G, Kitamura K, Wang Z, et al. RNA editing of hepatitis B virus transcripts by activation-induced cytidine deaminase. *Proc Natl Acad Sci U S A*. 2013;110(6):2246-2251. doi:10.1073/pnas.1221921110

110. He X, Li J, Wu J, Zhang M, Gao P. Associations between activation-induced cytidine deaminase/apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like cytidine deaminase expression, hepatitis B virus (HBV) replication and HBV-associated liver disease (Review). *Mol Med Rep*. 2015;12(5):6405-6414. doi:10.3892/mmr.2015.4312

111. Que L, Liu G, Kitamura K, et al. Molecular characterization of AID-mediated reduction of hepatitis B virus transcripts. *Virology*. 2017;510:281-288. doi:10.1016/j.virol.2017.07.035

112. Singh GB, Byun H, Ali AF, et al. A Protein Antagonist of Activation-Induced Cytidine Deaminase Encoded by a Complex Mouse Retrovirus. *mBio*. 2019;10(4). doi:10.1128/mBio.01678-19

113. Gourzi P, Leonova T, Papavasiliou FN. A role for activation-induced cytidine deaminase in the host response against a transforming retrovirus. *Immunity*. 2006;24(6):779-786. doi:10.1016/j.immuni.2006.03.021

114. Bekerman E, Jeon D, Ardolino M, Coscoy L. A Role for Host Activation-Induced Cytidine Deaminase in Innate Immune Defense against KSHV. *PLoS Pathog*. 2013;9(11). doi:10.1371/journal.ppat.1003748

115. Martinez T, Shapiro M, Bhaduri-McIntosh S, MacCarthy T. Evolutionary effects of the AID/APOBEC family of mutagenic enzymes on human gamma-herpesviruses. *Virus Evol*. 2019;5(1). doi:10.1093/ve/vey040

116. Ishikawa C, Nakachi S, Senba M, Sugai M, Mori N. Activation of AID by human T-cell leukemia virus Tax oncoprotein and the possible role of its constitutive expression in ATL genesis. *Carcinogenesis*. 2011;32(1):110-119. doi:10.1093/carcin/bgq222

117. Matsumoto Y, Marusawa H, Kinoshita K, et al. Helicobacter pylori infection triggers aberrant expression of activation-induced cytidine deaminase in gastric epithelium. *Nat Med*. 2007;13(4):470-476. doi:10.1038/nm1566

118. Gerber AP, Keller W. An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science*. 1999;286(5442):1146-1149.

119. Schaub M, Keller W. RNA editing by adenosine deaminases generates RNA and protein diversity. *Biochimie*. 2002;84(8):791-803. doi:10.1016/S0300-9084(02)01446-3

120. Rubio MAT, Pastar I, Gaston KW, et al. An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. *Proc Natl Acad Sci U S A*. 2007;104(19):7821-7826. doi:10.1073/pnas.0702394104

121. Gerber A, Grosjean H, Melcher T, Keller W. Tad1p, a yeast tRNA-specific adenosine deaminase, is related to the mammalian pre-mRNA editing enzymes ADAR1 and ADAR2. *EMBO J*. 1998;17(16):4780-4789. doi:10.1093/emboj/17.16.4780

122. Zhou W, Karcher D, Bock R. Importance of adenosine-to-inosine editing adjacent to the anticodon in an Arabidopsis alanine tRNA under environmental stress. *Nucleic Acids Res*. 2013;41(5):3362-3372. doi:10.1093/nar/gkt013

123. 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. doi:10.1016/S0065-2776(06)94002-4

124. Krishnan A, Iyer LM, Holland SJ, Boehm T, Aravind L. Diversification of AID/APOBEC-like deaminases in metazoa: multiplicity of clades and widespread roles in immunity. *Proc Natl Acad Sci U S A*. 2018;115(14):E3201-E3210. doi:10.1073/pnas.1720897115

125. Rogozin IB, Basu MK, Jordan IK, Pavlov YI, Koonin EV. APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. *Cell Cycle Georget Tex*. 2005;4(9):1281-1285. doi:10.4161/cc.4.9.1994

126. Rogozin IB, Iyer LM, Liang L, et al. Evolution and diversification of lamprey antigen receptors: evidence for involvement of an AID-APOBEC family cytosine deaminase. *Nat Immunol*. 2007;8(6):647-656. doi:10.1038/ni1463

127. Holland SJ, Berghuis LM, King JJ, et al. Expansions, diversification, and interindividual copy number variations of AID/APOBEC family cytidine deaminase genes in lampreys. *Proc Natl Acad Sci U S A*. 2018;115(14):E3211-E3220. doi:10.1073/pnas.1720871115

128. Patel B, Banerjee R, Samanta M, Das S. Diversity of Immunoglobulin (Ig) Isotypes and the Role of Activation-Induced Cytidine Deaminase (AID) in Fish. *Mol Biotechnol*. 2018;60(6):435-453. doi:10.1007/s12033-018-0081-8

129. Conticello SG, Thomas CJF, Petersen-Mahrt SK, Neuberger MS. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Mol Biol Evol*. 2005;22(2):367-377. doi:10.1093/molbev/msi026

130. Anant S, Mukhopadhyay D, Sankaranand V, Kennedy S, Henderson JO, Davidson NO. ARCD-1, an apobec-1-related cytidine deaminase, exerts a dominant negative effect on C to U RNA editing. *Am J Physiol Cell Physiol*. 2001;281(6):C1904-1916. doi:10.1152/ajpcell.2001.281.6.C1904

131. Liao W, Hong SH, Chan BH, Rudolph FB, Clark SC, Chan L. APOBEC-2, a cardiac- and skeletal muscle-specific member of the cytidine deaminase supergene family. *Biochem Biophys Res Commun*. 1999;260(2):398-404. doi:10.1006/bbrc.1999.0925

132. Harris RS, Petersen-Mahrt SK, Neuberger MS. RNA Editing Enzyme APOBEC1 and Some of Its Homologs Can Act as DNA Mutators. *Mol Cell*. 2002;10(5):1247-1253. doi:10.1016/S1097-2765(02)00742-6

133. Mikl MC, Watt IN, Lu M, et al. Mice deficient in APOBEC2 and APOBEC3. *Mol Cell Biol*. 2005;25(16):7270-7277. doi:10.1128/MCB.25.16.7270-7277.2005

134. Powell C, Cornblath E, Goldman D. Zinc-binding domain-dependent, deaminase-independent actions of apolipoprotein B mRNA-editing enzyme, catalytic polypeptide 2 (Apobec2), mediate its effect on zebrafish retina regeneration. *J Biol Chem*. 2014;289(42):28924-28941. doi:10.1074/jbc.M114.603043

135. Sato Y, Probst HC, Tatsumi R, Ikeuchi Y, Neuberger MS, Rada C. Deficiency in APOBEC2 leads to a shift in muscle fiber type, diminished body mass, and myopathy. *J Biol Chem*. 2010;285(10):7111-7118. doi:10.1074/jbc.M109.052977

136. Sato Y, Ohtsubo H, Nihei N, et al. Apobec2 deficiency causes mitochondrial defects and mitophagy in skeletal muscle. *FASEB J Off Publ Fed Am Soc Exp Biol*. 2018;32(3):1428-1439. doi:10.1096/fj.201700493R

137. Vonica A, Rosa A, Arduini BL, Brivanlou AH. APOBEC2, a selective inhibitor of TGFβ signaling, regulates left-right axis specification during early embryogenesis. *Dev Biol*. 2011;350(1):13-23. doi:10.1016/j.ydbio.2010.09.016

138. Amemiya CT, Alföldi J, Lee AP, et al. The African coelacanth genome provides insights into tetrapod evolution. *Nature*. 2013;496(7445):311-316. doi:10.1038/nature12027

139. Nakamuta M, Chang BHJ, Zsigmond E, et al. Complete Phenotypic Characterization of apobec-1 Knockout Mice with a Wild-type Genetic Background and a Human Apolipoprotein B Transgenic Background, and Restoration of Apolipoprotein B mRNA Editing by Somatic Gene Transfer of Apobec-1. *J Biol Chem*. 1996;271(42):25981-25988. doi:10.1074/jbc.271.42.25981

140. Bishop KN, Holmes RK, Sheehy AM, Malim MH. APOBEC-mediated editing of viral RNA. *Science*. 2004;305(5684):645. doi:10.1126/science.1100658

141. 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;279(50):52353-52360. doi:10.1074/jbc.M407695200

142. Petersen-Mahrt SK, Neuberger MS. In vitro deamination of cytosine to uracil in single-stranded DNA by apolipoprotein B editing complex catalytic subunit 1 (APOBEC1). *J Biol Chem*. 2003;278(22):19583-19586. doi:10.1074/jbc.C300114200

143. Zhang J, Webb DM. Rapid evolution of primate antiviral enzyme APOBEC3G. *Hum Mol Genet*. 2004;13(16):1785-1791. doi:10.1093/hmg/ddh183

144. Sawyer SL, Emerman M, Malik HS. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol*. 2004;2(9):E275. doi:10.1371/journal.pbio.0020275

145. Tacchi L, Larragoite ET, Muñoz P, Amemiya CT, Salinas I. African Lungfish Reveal the Evolutionary Origins of Organized Mucosal Lymphoid Tissue in Vertebrates. *Curr Biol CB*. 2015;25(18):2417-2424. doi:10.1016/j.cub.2015.07.066

146. Du Pasquier L, Zucchetti I, De Santis R. Immunoglobulin superfamily receptors in protochordates: before RAG time. *Immunol Rev*. 2004;198:233-248. doi:10.1111/j.0105-2896.2004.00122.x

147. Du Pasquier L. The phylogenetic origin of antigen-specific receptors. *Curr Top Microbiol Immunol*. 2000;248:160-185.

148. Hsu E. Assembly and Expression of Shark Ig Genes. *J Immunol Baltim Md 1950*. 2016;196(9):3517-3523. doi:10.4049/jimmunol.1600164

149. Zhang SM, Adema CM, Kepler TB, Loker ES. Diversification of Ig superfamily genes in an invertebrate. *Science*. 2004;305(5681):251-254. doi:10.1126/science.1088069

150. Trancoso I, Morimoto R, Boehm T. Co-evolution of mutagenic genome editors and vertebrate adaptive immunity. *Curr Opin Immunol*. 2020;65:32-41. doi:10.1016/j.coi.2020.03.001

151. Hirano M, Guo P, McCurley N, et al. Evolutionary implications of a third lymphocyte lineage in lampreys. *Nature*. 2013;501(7467):435-438. doi:10.1038/nature12467

152. Chen H, Kshirsagar S, Jensen I, et al. Characterization of arrangement and expression of the T cell receptor gamma locus in the sandbar shark. *Proc Natl Acad Sci U S A*. 2009;106(21):8591-8596. doi:10.1073/pnas.0811283106

153. Ott JA, Castro CD, Deiss TC, Ohta Y, Flajnik MF, Criscitiello MF. Somatic hypermutation of T cell receptor α chain contributes to selection in nurse shark thymus. *eLife*. 2018;7. doi:10.7554/eLife.28477

154. Bilal S, Lie KK, Sæle Ø, Hordvik I. T Cell Receptor Alpha Chain Genes in the Teleost Ballan Wrasse (Labrus bergylta) Are Subjected to Somatic Hypermutation. *Front Immunol*. 2018;9:1101. doi:10.3389/fimmu.2018.01101

155. Ciccarese S, Vaccarelli G, Lefranc MP, et al. Characteristics of the somatic hypermutation in the Camelus dromedarius T cell receptor gamma (TRG) and delta (TRD) variable domains. *Dev Comp Immunol*. 2014;46(2):300-313. doi:10.1016/j.dci.2014.05.001

156. Hsu E, Pulham N, Rumfelt LL, Flajnik MF. The plasticity of immunoglobulin gene systems in evolution. *Immunol Rev*. 2006;210:8-26. doi:10.1111/j.0105-2896.2006.00366.x

157. Ghaffari SH, Lobb CJ. Structure and genomic organization of a second cluster of immunoglobulin heavy chain gene segments in the channel catfish. *J Immunol Baltim Md 1950*. 1999;162(3):1519-1529.

158. Dooley H, Flajnik MF. Antibody repertoire development in cartilaginous fish. *Dev Comp Immunol*. 2006;30(1-2):43-56. doi:10.1016/j.dci.2005.06.022

159. Barreto VM, Magor BG. Activation-induced cytidine deaminase structure and functions: a species comparative view. *Dev Comp Immunol*. 2011;35(9):991-1007. doi:10.1016/j.dci.2011.02.005

160. Hinds-Frey KR, Nishikata H, Litman RT, Litman GW. Somatic variation precedes extensive diversification of germline sequences and combinatorial joining in the evolution of immunoglobulin heavy chain diversity. *J Exp Med*. 1993;178(3):815-824.

161. Diaz M, Greenberg AS, Flajnik MF. Somatic hypermutation of the new antigen receptor gene (NAR) in the nurse shark does not generate the repertoire: possible role in antigen-driven reactions in the absence of germinal centers. *Proc Natl Acad Sci U S A*. 1998;95(24):14343-14348. doi:10.1073/pnas.95.24.14343

162. Dooley H, Flajnik MF. Shark immunity bites back: affinity maturation and memory response in the nurse shark, Ginglymostoma cirratum. *Eur J Immunol*. 2005;35(3):936-945. doi:10.1002/eji.200425760

163. Malecek K, Brandman J, Brodsky JE, Ohta Y, Flajnik MF, Hsu E. Somatic hypermutation and junctional diversification at Ig heavy chain loci in the nurse shark. *J Immunol Baltim Md 1950*. 2005;175(12):8105-8115.

164. Magor BG. Antibody Affinity Maturation in Fishes-Our Current Understanding. *Biology*. 2015;4(3):512-524. doi:10.3390/biology4030512

165. Eason DD, Litman RT, Luer CA, Kerr W, Litman GW. Expression of individual immunoglobulin genes occurs in an unusual system consisting of multiple independent loci. *Eur J Immunol*. 2004;34(9):2551-2558. doi:10.1002/eji.200425224

166. Greenberg AS, Avila D, Hughes M, Hughes A, McKinney EC, Flajnik MF. A new antigen receptor gene family that undergoes rearrangement and extensive somatic diversification in sharks. *Nature*. 1995;374(6518):168-173. doi:10.1038/374168a0

167. Cain KD, Jones DR, Raison RL. Antibody-antigen kinetics following immunization of rainbow trout (Oncorhynchus mykiss) with a T-cell dependent antigen. *Dev Comp Immunol*. 2002;26(2):181-190. doi:10.1016/s0145-305x(01)00063-5

168. Kaattari SL, Zhang HL, Khor IW, Kaattari IM, Shapiro DA. Affinity maturation in trout: clonal dominance of high affinity antibodies late in the immune response. *Dev Comp Immunol*. 2002;26(2):191-200.

169. Saunders HL, Oko AL, Scott AN, Fan CW, Magor BG. The cellular context of AID expressing cells in fish lymphoid tissues. *Dev Comp Immunol*. 2010;34(6):669-676. doi:10.1016/j.dci.2010.01.013

170. Saunders HL, Magor BG. Cloning and expression of the AID gene in the channel catfish. *Dev Comp Immunol*. 2004;28(7-8):657-663. doi:10.1016/j.dci.2004.01.002

171. Yang F, Waldbieser GC, Lobb CJ. The nucleotide targets of somatic mutation and the role of selection in immunoglobulin heavy chains of a teleost fish. *J Immunol Baltim Md 1950*. 2006;176(3):1655-1667.

172. Marianes AE, Zimmerman AM. Targets of somatic hypermutation within immunoglobulin light chain genes in zebrafish. *Immunology*. 2011;132(2):240-255. doi:10.1111/j.1365-2567.2010.03358.x

173. Neely HR, Flajnik MF. Emergence and Evolution of Secondary Lymphoid Organs. *Annu Rev Cell Dev Biol*. 2016;32:693-711. doi:10.1146/annurev-cellbio-111315-125306

174. Bengtén E, Clem LW, Miller NW, Warr GW, Wilson M. Channel catfish immunoglobulins: repertoire and expression. *Dev Comp Immunol*. 2006;30(1-2):77-92. doi:10.1016/j.dci.2005.06.016

175. Danilova N, Bussmann J, Jekosch K, Steiner LA. The immunoglobulin heavy-chain locus in zebrafish: identification and expression of a previously unknown isotype, immunoglobulin Z. *Nat Immunol*. 2005;6(3):295-302. doi:10.1038/ni1166

176. Hansen JD, Landis ED, Phillips RB. Discovery of a unique Ig heavy-chain isotype (IgT) in rainbow trout: Implications for a distinctive B cell developmental pathway in teleost fish. *Proc Natl Acad Sci U S A*. 2005;102(19):6919-6924. doi:10.1073/pnas.0500027102

177. Zhang T, Tacchi L, Wei Z, Zhao Y, Salinas I. Intraclass diversification of immunoglobulin heavy chain genes in the African lungfish. *Immunogenetics*. 2014;66(5):335-351. doi:10.1007/s00251-014-0769-2

178. Wakae K, Magor BG, Saunders H, et al. Evolution of class switch recombination function in fish activation-induced cytidine deaminase, AID. *Int Immunol*. 2006;18(1):41-47. doi:10.1093/intimm/dxh347

179. Ichikawa HT, Sowden MP, Torelli AT, et al. Structural Phylogenetic Analysis of Activation-Induced Deaminase Function. *J Immunol*. 2006;177(1):355-361. doi:10.4049/jimmunol.177.1.355

180. Marr S, Morales H, Bottaro A, Cooper M, Flajnik M, Robert J. Localization and differential expression of activation-induced cytidine deaminase in the amphibian Xenopus upon antigen stimulation and during early development. *J Immunol Baltim Md 1950*. 2007;179(10):6783-6789. doi:10.4049/jimmunol.179.10.6783

181. Barreto VM, Pan-Hammarstrom Q, Zhao Y, Hammarstrom L, Misulovin Z, Nussenzweig MC. AID from bony fish catalyzes class switch recombination. *J Exp Med*. 2005;202(6):733-738. doi:10.1084/jem.20051378

182. Basu U, Franklin A, Alt FW. Post-translational regulation of activation-induced cytidine deaminase. *Philos Trans R Soc Lond B Biol Sci*. 2009;364(1517):667-673. doi:10.1098/rstb.2008.0194

183. Dancyger AM, King JJ, Quinlan MJ, 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 Off Publ Fed Am Soc Exp Biol*. 2012;26(4):1517-1525. doi:10.1096/fj.11-198135

184. Abdouni H, King JJ, Suliman M, Quinlan M, Fifield H, Larijani M. Zebrafish AID is capable of deaminating methylated deoxycytidines. *Nucleic Acids Res*. 2013;41(10):5457-5468. doi:10.1093/nar/gkt212

185. Larijani M, Frieder D, Sonbuchner TM, et al. Methylation protects cytidines from AID-mediated deamination. *Mol Immunol*. 2005;42(5):599-604. doi:10.1016/j.molimm.2004.09.007

186. 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;135(7):1201-1212. doi:10.1016/j.cell.2008.11.042

187. Mussman R, Courtet M, Schwager J, Du Pasquier L. Microsites for immunoglobulin switch recombination breakpoints from Xenopus to mammals. *Eur J Immunol*. 1997;27:2610-2619.

188. Zarrin AA, Alt FW, Chaudhuri J, et al. An evolutionarily conserved target motif for immunoglobulin class-switch recombination. *Nat Immunol*. 2004;5(12):1275-1281. doi:10.1038/ni1137

189. Wilson M, Hsu E, Marcuz A, Courtet M, Du Pasquier L, Steinberg C. What limits affinity maturation of antibodies in Xenopus--the rate of somatic mutation or the ability to select mutants? *EMBO J*. 1992;11(12):4337-4347.

190. Neely HR, Guo J, Flowers EM, Criscitiello MF, Flajnik MF. "Double-duty" conventional dendritic cells in the amphibian Xenopus as the prototype for antigen presentation to B cells. *Eur J Immunol*. 2018;48(3):430-440. doi:10.1002/eji.201747260

191. Reynaud CA, Anquez V, Dahan A, Weill JC. A single rearrangement event generates most of the chicken immunoglobulin light chain diversity. *Cell*. 1985;40(2):283-291. doi:10.1016/0092-8674(85)90142-4

192. Huang T, Zhang M, Wei Z, et al. Analysis of Immunoglobulin Transcripts in the Ostrich Struthio camelus, a Primitive Avian Species. *PLOS ONE*. 2012;7(3):e34346. doi:10.1371/journal.pone.0034346

193. Lundqvist ML, Middleton DL, Radford C, Warr GW, Magor KE. Immunoglobulins of the non-galliform birds: Antibody expression and repertoire in the duck. *Dev Comp Immunol*. 2006;30(1-2):93-100. doi:10.1016/j.dci.2005.06.019

194. Moon SY, Eun HJ, Baek SK, et al. Activation-Induced Cytidine Deaminase Induces DNA Demethylation of Pluripotency Genes in Bovine Differentiated Cells. *Cell Reprogramming*. 2016;18(5):298-308. doi:10.1089/cell.2015.0076

195. Konnova EA, Swanberg M. Animal Models of Parkinson's Disease. In: Stoker TB, Greenland JC, eds. *Parkinson's Disease: Pathogenesis and Clinical Aspects*. Codon Publications; 2018. Accessed November 30, 2020. http://www.ncbi.nlm.nih.gov/books/NBK536725/

196. Vandamme TF. Use of rodents as models of human diseases. *J Pharm Bioallied Sci*. 2014;6(1):2-9. doi:10.4103/0975-7406.124301

197. Ingber DE. Is it Time for Reviewer 3 to Request Human Organ Chip Experiments Instead of Animal Validation Studies? *Adv Sci Weinh Baden-Wurtt Ger*. 2020;7(22):2002030. doi:10.1002/advs.202002030

198. Perlman RL. Mouse models of human disease. *Evol Med Public Health*. 2016;2016(1):170-176. doi:10.1093/emph/eow014

199. Minton AP. How can biochemical reactions within cells differ from those in test tubes? *J Cell Sci*. 2006;119(14):2863-2869. doi:10.1242/jcs.03063

200. Langhans SA. Three-Dimensional in Vitro Cell Culture Models in Drug Discovery and Drug Repositioning. *Front Pharmacol*. 2018;9. doi:10.3389/fphar.2018.00006

201. Markowetz F. All biology is computational biology. *PLOS Biol*. 2017;15(3):e2002050. doi:10.1371/journal.pbio.2002050

202. Gamber KM. Animal Models of Parkinson's Disease: New models provide greater translational and predictive value. *BioTechniques*. 2016;61(4):210-211. doi:10.2144/000114463

203. Elsea SH, Lucas RE. The mousetrap: what we can learn when the mouse model does not mimic the human disease. *ILAR J*. 2002;43(2):66-79. doi:10.1093/ilar.43.2.66

204. Dharmadhikari AS, Nardell EA. What Animal Models Teach Humans about Tuberculosis. *Am J Respir Cell Mol Biol*. 2008;39(5):503-508. doi:10.1165/rcmb.2008-0154TR

205. Pedroza-Gonzalez A, Xu K, Wu TC, et al. Thymic stromal lymphopoietin fosters human breast tumor growth by promoting type 2 inflammation. *J Exp Med*. 2011;208(3):479-490. doi:10.1084/jem.20102131

206. Wege AK. Humanized Mouse Models for the Preclinical Assessment of Cancer Immunotherapy. *BioDrugs Clin Immunother Biopharm Gene Ther*. 2018;32(3):245-266. doi:10.1007/s40259-018-0275-4

207. Allen TM, Brehm MA, Bridges S, et al. Humanized immune system mouse models: progress, challenges and opportunities. *Nat Immunol*. 2019;20(7):770-774. doi:10.1038/s41590-019-0416-z

208. Ito R, Katano I, Kawai K, et al. Highly sensitive model for xenogenic GVHD using severe immunodeficient NOG mice. *Transplantation*. 2009;87(11):1654-1658. doi:10.1097/TP.0b013e3181a5cb07

209. Takahashi H, Tsuboi H, Abe S, et al. Humanized NOD/SCID/IL2rγnull mice exhibit functionally augmented human regulatory T cells associated with enzymatic upregulation

of H3K27me3 in comparison with humans. *Clin Exp Immunol*. Published online February 8, 2021. doi:10.1111/cei.13583

210. Hedrich H. *The Laboratory Mouse*. Academic Press; 2004.

211. Mittelman D, Wilson JH. The fractured genome of HeLa cells. *Genome Biol*. 2013;14(4):111. doi:10.1186/gb-2013-14-4-111

212. Carter M, Shieh JC. Chapter 13 - Cell Culture Techniques. In: Carter M, Shieh JC, eds. *Guide to Research Techniques in Neuroscience*. Academic Press; 2010:281-296. doi:10.1016/B978-0-12-374849-2.00013-6

213. Moysidou CM, Barberio C, Owens RM. Advances in Engineering Human Tissue Models. *Front Bioeng Biotechnol*. 2021;8(620962). doi:10.3389/fbioe.2020.620962

214. Hochachka PW, Rupert JL, Goldenberg L, Gleave M, Kozlowski P. Going malignant: the hypoxia-cancer connection in the prostate. *BioEssays News Rev Mol Cell Dev Biol*. 2002;24(8):749-757. doi:10.1002/bies.10131

215. Hochachka PW. Defense strategies against hypoxia and hypothermia. *Science*. 1986;231(4735):234-241. doi:10.1126/science.2417316

216. Hochachka PW, Somero GN. The adaptation of enzymes to temperature. *Comp Biochem Physiol*. 1968;27(3):659-668. doi:10.1016/0010-406X(68)90605-1

217. Cooper EL. Specific tissue graft rejection in earthworms. *Science*. 1969;166(3911):1414-1415. doi:10.1126/science.166.3911.1414

218. Bruno R, Maresca M, Canaan S, et al. Worms' Antimicrobial Peptides. *Mar Drugs*. 2019;17(9). doi:10.3390/md17090512

219. Munita JM, Arias CA. Mechanisms of Antibiotic Resistance. *Microbiol Spectr*. 2016;4(2). doi:10.1128/microbiolspec.VMBF-0016-2015

220. Davies J, Davies D. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev MMBR*. 2010;74(3):417-433. doi:10.1128/MMBR.00016-10

221. Rimer J, Cohen IR, Friedman N. Do all creatures possess an acquired immune system of some sort? *BioEssays News Rev Mol Cell Dev Biol*. 2014;36(3):273-281. doi:10.1002/bies.201300124

222. Tincu JA, Menzel LP, Azimov R, et al. Plicatamide, an antimicrobial octapeptide from Styela plicata hemocytes. *J Biol Chem*. 2003;278(15):13546-13553. doi:10.1074/jbc.M211332200

223. Zhang Y, Cheng TC, Huang G, et al. Transposon molecular domestication and the evolution of the RAG recombinase. *Nature*. Published online April 10, 2019. doi:10.1038/s41586-019-1093-7

224. Huang S, Tao X, Yuan S, et al. Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. *Cell*. 2016;166(1):102-114. doi:10.1016/j.cell.2016.05.032

225. Carmona LM, Schatz DG. New insights into the evolutionary origins of the RAG proteins and V(D)J recombination. *FEBS J*. 2017;284(11):1590-1605. doi:10.1111/febs.13990

226. 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;27(1):20-30. doi:10.1128/MCB.00824-06

227. 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;27(23):8038-8048. doi:10.1128/MCB.01046-07

228. Larijani M, Martin A. The biochemistry of activation-induced deaminase and its physiological functions. *Semin Immunol*. 2012;24(4):255-263. doi:10.1016/j.smim.2012.05.003

229. 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;37(20). doi:10.1128/MCB.00077-17

230. Pancer Z, Amemiya CT, Ehrhardt GRA, Ceitlin J, Gartland GL, Cooper MD. Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature*. 2004;430(6996):174-180. doi:10.1038/nature02740

231. Buonocore F, Gerdol M. Alternative adaptive immunity strategies: coelacanth, cod and shark immunity. *Mol Immunol*. 2016;69:157-169. doi:10.1016/j.molimm.2015.09.003

232. Venkatesh B, Lee AP, Ravi V, et al. Elephant shark genome provides unique insights into gnathostome evolution. *Nature*. 2014;505(7482):174-179. doi:10.1038/nature12826

233. Peubez I, Margot S, Buffin S, et al. A novel lamprey antibody sequence to multimerize and increase the immunogenicity of recombinant viral and bacterial vaccine antigens. *Vaccine*. 2020;38(50):7905-7915. doi:10.1016/j.vaccine.2020.10.073

234. Jeffery CJ. An enzyme in the test tube, and a transcription factor in the cell: Moonlighting proteins and cellular factors that affect their behavior. *Protein Sci*. 2019;28(7):1233-1238. doi:10.1002/pro.3645

235. Petersen-Mahrt SK, Harris RS, Neuberger MS. AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature*. 2002;418(6893):99-104. doi:10.1038/nature00862

236. Borzooee F, Larijani M. Pichia pastoris as a host for production and isolation of mutagenic AID/APOBEC enzymes involved in cancer and immunity. *New Biotechnol*. 2019;51:67-79. doi:10.1016/j.nbt.2019.02.006

237. Chester A, Weinreb V, Carter CW, Navaratnam N. Optimization of apolipoprotein B mRNA editing by APOBEC1 apoenzyme and the role of its auxiliary factor, ACF. *RNA N Y N*. 2004;10(9):1399-1411. doi:10.1261/rna.7490704

238. Rogozin IB, Kolchanov NA. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim Biophys Acta*. 1992;1171(1):11-18.

239. 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;56(11):840-845. doi:10.1007/s00251-004-0748-0

240. Pham P, Bransteitter R, Petruska J, Goodman MF. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature*. 2003;424(6944):103-107. doi:10.1038/nature01760

241. Yu K, Huang FT, Lieber MR. DNA substrate length and surrounding sequence affect the activation-induced deaminase activity at cytidine. *J Biol Chem*. 2004;279(8):6496-6500. doi:10.1074/jbc.M311616200

242. 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;100(7):4102-4107. doi:10.1073/pnas.0730835100

243. Chaudhuri J, Tian M, Khuong C, Chua K, Pinaud E, Alt FW. Transcription-targeted DNA deamination by the AID antibody diversification enzyme. *Nature*. 2003;422(6933):726-730. doi:10.1038/nature01574

244. Dickerson SK, Market E, Besmer E, Papavasiliou FN. AID mediates hypermutation by deaminating single stranded DNA. *J Exp Med*. 2003;197(10):1291-1296. doi:10.1084/jem.20030481

245. Shen HM, Storb U. Activation-induced cytidine deaminase (AID) can target both DNA strands when the DNA is supercoiled. *Proc Natl Acad Sci U S A*. 2004;101(35):12997-13002. doi:10.1073/pnas.0404974101

246. Chaudhuri J, Khuong C, Alt FW. Replication protein A interacts with AID to promote deamination of somatic hypermutation targets. *Nature*. 2004;430(7003):992-998. doi:10.1038/nature02821

247. 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;4(5):452-456. doi:10.1038/ni920

248. 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;31(12):2990-2994. doi:10.1093/nar/gkg464

249. 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;43(8):1099-1108. doi:10.1016/j.molimm.2005.07.024

250. Pham P, Afif SA, Shimoda M, et al. Activation-induced deoxycytidine deaminase: structural basis favoring WRC hot motif specificities unique among APOBEC family members. *DNA Repair*. 2017;54:8-12. doi:10.1016/j.dnarep.2017.03.007

251. Kobayashi M, Takemoto M, Honjo T. The novel activation-induced deoxycytidine deaminase (AID) mutants, AIDv and AIDvΔ15 are defective in SHM and CSR. *DNA Repair*. 2017;53:1-3. doi:10.1016/j.dnarep.2017.02.015

252. King JJ, Larijani M. A Novel Regulator of Activation-Induced Cytidine Deaminase/APOBECs in Immunity and Cancer: Schrödinger's CATalytic Pocket. *Front Immunol*. 2017;8:351. doi:10.3389/fimmu.2017.00351

253. Pham P, Afif SA, Shimoda M, et al. Structural analysis of the activation-induced deoxycytidine deaminase required in immunoglobulin diversification. *DNA Repair*. 2016;43:48-56. doi:10.1016/j.dnarep.2016.05.029

254. Feig M, Harada R, Mori T, Yu I, Takahashi K, Sugita Y. Complete atomistic model of a bacterial cytoplasm for integrating physics, biochemistry, and systems biology. *J Mol Graph Model*. 2015;58:1-9. doi:10.1016/j.jmgm.2015.02.004

255. Karr JR, Sanghvi JC, Macklin DN, et al. A whole-cell computational model predicts phenotype from genotype. *Cell*. 2012;150(2):389-401. doi:10.1016/j.cell.2012.05.044

256. Isalan M. Systems biology: A cell in a computer. *Nature*. 2012;488(7409):40-41. doi:10.1038/488040a

257. Thippabhotla S, Zhong C, He M. 3D cell culture stimulates the secretion of in vivo like extracellular vesicles. *Sci Rep*. 2019;9(1):13012. doi:10.1038/s41598-019-49671-3

258. Williams A, Nowak JF, Dass R, Samuel J, Mills KL. Toward Morphologically Relevant Extracellular Matrix in Vitro Models: 3D Fiber Reinforced Hydrogels. *Front Physiol*. 2018;9. doi:10.3389/fphys.2018.00966

259. Yang Z, Xiong HR. In vitro, Tissue-Based Models as a Replacement for Animal Models in Testing of Drugs at the Preclinical Stages. *Biomed Tissue Cult*. Published online October 17, 2012. doi:10.5772/52300

260. Amemiya CT, Dorrington R, Meyer A. The coelacanth and its genome. *J Exp Zoolog B Mol Dev Evol*. 2014;322(6):317-321. doi:10.1002/jez.b.22583

261. Jaillon O, Aury JM, Brunet F, et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*. 2004;431(7011):946-957. doi:10.1038/nature03025

262. Muto T, Muramatsu M, Taniwaki M, Kinoshita K, Honjo T. Isolation, tissue distribution, and chromosomal localization of the human activation-induced cytidine deaminase (AID) gene. *Genomics*. 2000;68(1):85-88. doi:10.1006/geno.2000.6268

263. Zhao Y, Pan-Hammarström Q, Zhao Z, Hammarström L. Identification of the activation-induced cytidine deaminase gene from zebrafish: an evolutionary analysis. *Dev Comp Immunol*. 2005;29(1):61-71. doi:10.1016/j.dci.2004.05.005

264. Wedekind JE, Dance GSC, Sowden MP, Smith HC. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends Genet TIG*. 2003;19(4):207-216. doi:10.1016/S0168-9525(03)00054-4

265. Bas DC, Rogers DM, Jensen JH. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins*. 2008;73(3):765-783. doi:10.1002/prot.22102

266. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. *Proteins*. 2005;61(4):704-721. doi:10.1002/prot.20660

267. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J Chem Theory Comput*. 2011;7(2):525-537. doi:10.1021/ct100578z

268. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res*. 2011;39(Web Server issue):W270-277. doi:10.1093/nar/gkr366

269. Grosdidier A, Zoete V, Michielin O. Fast docking using the CHARMM force field with EADock DSS. *J Comput Chem*. 2011;32(10):2149-2159. doi:10.1002/jcc.21797

270. Zoete V, Cuendet MA, Grosdidier A, Michielin O. SwissParam: a fast force field generation tool for small organic molecules. *J Comput Chem*. 2011;32(11):2359-2368. doi:10.1002/jcc.21816

271. Xie K, Sowden MP, Dance GSC, 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 S A*. 2004;101(21):8114-8119. doi:10.1073/pnas.0400493101

272. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605-1612. doi:10.1002/jcc.20084

273. Knisbacher BA, Gerber D, Levanon EY. DNA Editing by APOBECs: A Genomic Preserver and Transformer. *Trends Genet TIG*. 2016;32(1):16-28. doi:10.1016/j.tig.2015.10.005

274. Pettersen HS, Galashevskaya A, Doseth B, et al. AID expression in B-cell lymphomas causes accumulation of genomic uracil and a distinct AID mutational signature. *DNA Repair*. 2015;25:60-71. doi:10.1016/j.dnarep.2014.11.006

275. Robbiani DF, Nussenzweig MC. Chromosome Translocation, B Cell Lymphoma, and Activation-Induced Cytidine Deaminase. *Annu Rev Pathol Mech Dis*. 2013;8(1):79-103. doi:10.1146/annurev-pathol-020712-164004

276. Gostissa M, Alt FW, Chiarle R. Mechanisms that promote and suppress chromosomal translocations in lymphocytes. *Annu Rev Immunol*. 2011;29:319-350. doi:10.1146/annurev-immunol-031210-101329

277. Gazumyan A, Bothmer A, Klein IA, Nussenzweig MC, McBride KM. Activation-induced cytidine deaminase in antibody diversification and chromosome translocation. *Adv Cancer Res*. 2012;113:167-190. doi:10.1016/B978-0-12-394280-7.00005-1

278. Bar-Even A, Noor E, Savir Y, et al. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*. 2011;50(21):4402-4410. doi:10.1021/bi2002289

279. 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;16(7):769-776. doi:10.1038/nsmb.1623

280. Laird DJ, De Tomaso AW, Cooper MD, Weissman IL. 50 million years of chordate evolution: seeking the origins of adaptive immunity. *Proc Natl Acad Sci U S A*. 2000;97(13):6924-6926. doi:10.1073/pnas.97.13.6924

281. Pancer Z, Saha NR, Kasamatsu J, et al. Variable lymphocyte receptors in hagfish. *Proc Natl Acad Sci U S A*. 2005;102(26):9224-9229. doi:10.1073/pnas.0503792102

282. Deng L, Velikovsky CA, Xu G, et al. A structural basis for antigen recognition by the T cell-like lymphocytes of sea lamprey. *Proc Natl Acad Sci U S A*. 2010;107(30):13408-13413. doi:10.1073/pnas.1005475107

283. Tasumi S, Velikovsky CA, Xu G, et al. High-affinity lamprey VLRA and VLRB monoclonal antibodies. *Proc Natl Acad Sci U S A*. 2009;106(31):12891-12896. doi:10.1073/pnas.0904443106

284. Kasamatsu J, Sutoh Y, Fugo K, Otsuka N, Iwabuchi K, Kasahara M. Identification of a third variable lymphocyte receptor in the lamprey. *Proc Natl Acad Sci U S A*. 2010;107(32):14304-14308. doi:10.1073/pnas.1001910107

285. Kirchdoerfer RN, Herrin BR, Han BW, Turnbough CL, Cooper MD, Wilson IA. Variable lymphocyte receptor recognition of the immunodominant glycoprotein of Bacillus anthracis spores. *Struct Lond Engl 1993*. 2012;20(3):479-486. doi:10.1016/j.str.2012.01.009

286. Das S, Hirano M, Aghaallaei N, Bajoghli B, Boehm T, Cooper MD. Organization of lamprey variable lymphocyte receptor C locus and repertoire development. *Proc Natl Acad Sci U S A*. 2013;110(15):6043-6048. doi:10.1073/pnas.1302500110

287. Li J, Das S, Herrin BR, Hirano M, Cooper MD. Definition of a third VLR gene in hagfish. *Proc Natl Acad Sci U S A*. 2013;110(37):15013-15018. doi:10.1073/pnas.1314540110

288. Holland SJ, Gao M, Hirano M, et al. Selection of the lamprey VLRC antigen receptor repertoire. *Proc Natl Acad Sci U S A*. 2014;111(41):14834-14839. doi:10.1073/pnas.1415655111

289. Luo M, Velikovsky CA, Yang X, et al. Recognition of the Thomsen-Friedenreich pancarcinoma carbohydrate antigen by a lamprey variable lymphocyte receptor. *J Biol Chem*. 2013;288(32):23597-23606. doi:10.1074/jbc.M113.480467

290. Mayer WE, Uinuk-ool T, Tichy H, Gartland LA, Klein J, Cooper MD. Isolation and characterization of lymphocyte-like cells from a lamprey. *Proc Natl Acad Sci*. 2002;99(22):14350-14355. doi:10.1073/pnas.212527499

291. Guo P, Hirano M, Herrin BR, et al. Dual nature of the adaptive immune system in lampreys. *Nature*. 2009;459(7248):796-801. doi:10.1038/nature08068

292. Bajoghli B, Guo P, Aghaallaei N, et al. A thymus candidate in lampreys. *Nature*. 2011;470(7332):90-94. doi:10.1038/nature09655

293. Hirano M. Evolution of vertebrate adaptive immunity: immune cells and tissues, and AID/APOBEC cytidine deaminases. *BioEssays News Rev Mol Cell Dev Biol*. 2015;37(8):877-887. doi:10.1002/bies.201400178

294. Hinds KR, Litman GW. Major reorganization of immunoglobulin VH segmental elements during vertebrate evolution. *Nature*. 1986;320(6062):546-549. doi:10.1038/320546a0

295. Lee SS, Tranchina D, Ohta Y, Flajnik MF, Hsu E. Hypermutation in shark immunoglobulin light chain genes results in contiguous substitutions. *Immunity*. 2002;16(4):571-582.

296. Zhu C, Lee V, Finn A, et al. Origin of immunoglobulin isotype switching. *Curr Biol CB*. 2012;22(10):872-880. doi:10.1016/j.cub.2012.03.060

297. Antonacci R, Mineccia M, Lefranc MP, et al. Expression and genomic analyses of Camelus dromedarius T cell receptor delta (TRD) genes reveal a variable domain repertoire enlargement due to CDR3 diversification and somatic mutation. *Mol Immunol*. 2011;48(12-13):1384-1396. doi:10.1016/j.molimm.2011.03.011

298. Chen H, Bernstein H, Ranganathan P, Schluter SF. Somatic hypermutation of TCR γ V genes in the sandbar shark. *Dev Comp Immunol*. 2012;37(1):176-183. doi:10.1016/j.dci.2011.08.018

299. Saha NR, Ota T, Litman GW, et al. Genome complexity in the coelacanth is reflected in its adaptive immune system. *J Exp Zoolog B Mol Dev Evol*. 2014;322(6):438-463. doi:10.1002/jez.b.22558

300. Beale RCL, Petersen-Mahrt SK, Watt IN, Harris RS, Rada C, Neuberger MS. Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. *J Mol Biol*. 2004;337(3):585-596. doi:10.1016/j.jmb.2004.01.046

301. 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*. 2008;7(1):77-87. doi:10.1016/j.dnarep.2007.08.002

302. Mu Y, Prochnow C, Pham P, Chen XS, Goodman MF. A structural basis for the biochemical behavior of activation-induced deoxycytidine deaminase class-switch recombination-defective hyper-IgM-2 mutants. *J Biol Chem*. 2012;287(33):28007-28016. doi:10.1074/jbc.M112.370189

303. Krause K, Marcu KB, Greeve J. The cytidine deaminases AID and APOBEC-1 exhibit distinct functional properties in a novel yeast selectable system. *Mol Immunol*. 2006;43(4):295-307. doi:10.1016/j.molimm.2005.05.004

304. 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;284(34):22898-22904. doi:10.1074/jbc.M109.025536

305. Carpenter MA, Rajagurubandara E, Wijesinghe P, Bhagwat AS. Determinants of sequence-specificity within human AID and APOBEC3G. *DNA Repair*. 2010;9(5):579-587. doi:10.1016/j.dnarep.2010.02.010

306. Holtz CM, Sadler HA, Mansky LM. APOBEC3G cytosine deamination hotspots are defined by both sequence context and single-stranded DNA secondary structure. *Nucleic Acids Res*. 2013;41(12):6139-6148. doi:10.1093/nar/gkt246

307. Rathore A, Carpenter MA, Demir Ö, et al. The local dinucleotide preference of APOBEC3G can be altered from 5'-CC to 5'-TC by a single amino acid substitution. *J Mol Biol*. 2013;425(22):4442-4454. doi:10.1016/j.jmb.2013.07.040

308. Papaneophytou CP, Grigoroudis AI, McInnes C, Kontopidis G. Quantification of the Effects of Ionic Strength, Viscosity, and Hydrophobicity on Protein–Ligand Binding Affinity. *ACS Med Chem Lett*. 2014;5(8):931-936. doi:10.1021/ml500204e

309. 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;285(21):16195-16205. doi:10.1074/jbc.M110.107987

310. 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;283(20):13780-13791. doi:10.1074/jbc.M801004200

311. 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;286(28):24931-24942. doi:10.1074/jbc.M111.241208

312. Holmes JA, Lin P. Thermal Niche of Larval Sea Lamprey, Petromyzon marinus. *Can J Fish Aquat Sci*. 1994;51(2):253-262. doi:10.1139/f94-026

313. Ferreira LC, Afonso AS, Castilho PC, Hazin FHV. Habitat use of the nurse shark, Ginglymostoma cirratum, off Recife, Northeast Brazil: a combined survey with longline and acoustic telemetry. *Environ Biol Fishes*. 2013;96(6):735-745. doi:10.1007/s10641-012-0067-5

314. Dekkers WJ. Review of the Asiatic freshwater puffers of the genus Tetraodon Linnaeus, 1758 (Pisces, Tetraodontiformes, Tetraodontidae). *Bijdr Tot Dierkd*. 1975;45(1):87-142.

315. Fricke H, Hissmann K. Feeding ecology and evolutionary survival of the living coelacanth Latimeria chalumnae. *Mar Biol*. 2000;136(2):379-386. doi:10.1007/s002270050697

316. Jaenicke R, Böhm G. The stability of proteins in extreme environments. *Curr Opin Struct Biol*. 1998;8(6):738-748.

317. Wintrode PL, Arnold FH. Temperature adaptation of enzymes: Lessons from laboratory evolution. In: *Advances in Protein Chemistry*. Vol 55. Evolutionary Protein Design. Academic Press; 2001:161-225. doi:10.1016/S0065-3233(01)55004-4

318. Radestock S, Gohlke H. Protein rigidity and thermophilic adaptation. *Proteins*. 2011;79(4):1089-1108. doi:10.1002/prot.22946

319. Radestock S, Gohlke H. Exploiting the Link between Protein Rigidity and Thermostability for Data-Driven Protein Engineering. *Eng Life Sci*. 2008;8(5):507-522. doi:10.1002/elsc.200800043

320. Prochnow C, Bransteitter R, Klein MG, Goodman MF, Chen XS. The APOBEC-2 crystal structure and functional implications for the deaminase AID. *Nature*. 2007;445(7126):447-451. doi:10.1038/nature05492

321. Boehm T, McCurley N, Sutoh Y, Schorpp M, Kasahara M, Cooper MD. VLR-based adaptive immunity. *Annu Rev Immunol*. 2012;30:203-220. doi:10.1146/annurev-immunol-020711-075038

322. Ratcliffe MJH. B cell development in gut associated lymphoid tissues. *Vet Immunol Immunopathol*. 2002;87(3-4):337-340. doi:10.1016/s0165-2427(02)00061-2

323. Lanning DK, Knight KL. Diversification of the Primary Antibody Repertoire by AID-Mediated Gene Conversion. *Results Probl Cell Differ*. 2015;57:279-293. doi:10.1007/978-3-319-20819-0_12

324. Docker MF, ed. *Lampreys: Biology, Conservation and Control*. Vol 1. Springer; 2015. Accessed June 30, 2020. https://www.springer.com/gp/book/9789401793056

325. Smith JJ, Timoshevskaya N, Ye C, et al. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nat Genet*. 2018;50(2):270-277. doi:10.1038/s41588-017-0036-1

326. Smith JJ, Timoshevskiy VA, Saraceno C. Programmed DNA Elimination in Vertebrates. *Annu Rev Anim Biosci*. 2021;9:173-201. doi:10.1146/annurev-animal-061220-023220

327. Timoshevskiy VA, Herdy JR, Keinath MC, Smith JJ. Cellular and Molecular Features of Developmentally Programmed Genome Rearrangement in a Vertebrate (Sea Lamprey: Petromyzon marinus). *PLoS Genet*. 2016;12(6):e1006103. doi:10.1371/journal.pgen.1006103

328. Timoshevskiy VA, Timoshevskaya NY, Smith JJ. Germline-Specific Repetitive Elements in Programmatically Eliminated Chromosomes of the Sea Lamprey (Petromyzon marinus). *Genes*. 2019;10(10). doi:10.3390/genes10100832

329. Kinsella CM, Ruiz-Ruano FJ, Dion-Côté AM, et al. Programmed DNA elimination of germline development genes in songbirds. *Nat Commun*. 2019;10(1):5468. doi:10.1038/s41467-019-13427-4

330. Chase A, Cross NCP. Aberrations of EZH2 in cancer. *Clin Cancer Res Off J Am Assoc Cancer Res*. 2011;17(9):2613-2618. doi:10.1158/1078-0432.CCR-10-2156

331. Olson MV. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet*. 1999;64(1):18-23. doi:10.1086/302219

332. Sutoh Y, Mohamed RH, Kasahara M. Origin and Evolution of Dendritic Epidermal T Cells. *Front Immunol*. 2018;9:1059. doi:10.3389/fimmu.2018.01059

333. Wu F, Chen L, Liu X, et al. Lamprey variable lymphocyte receptors mediate complement-dependent cytotoxicity. *J Immunol Baltim Md 1950*. 2013;190(3):922-930. doi:10.4049/jimmunol.1200876

334. Li J, Ma Q, Liu H, et al. Complement component C1q plays a critical role in VLRA/VLRC-mediated immune response. *Dev Comp Immunol*. 2020;111:103750. doi:10.1016/j.dci.2020.103750

335. Su P, Liu X, Pang Y, et al. The archaic roles of the lamprey NF-κB (lj-NF-κB) in innate immune responses. *Mol Immunol*. 2017;92:21-27. doi:10.1016/j.molimm.2017.10.002

336. Luo L, Lu J, Wang Q, Chen S, Xu A, Yuan S. Autophagy participates in innate immune defense in lamprey. *Fish Shellfish Immunol*. 2018;83:416-424. doi:10.1016/j.fsi.2018.09.016

337. Tokmakov AA, Kurotani A, Takagi T, et al. Multiple post-translational modifications affect heterologous protein synthesis. *J Biol Chem*. 2012;287(32):27106-27116. doi:10.1074/jbc.M112.366351

338. Ito F, Fu Y, Kao SCA, Yang H, Chen XS. Family-Wide Comparative Analysis of Cytidine and Methylcytidine Deamination by Eleven Human APOBEC Proteins. *J Mol Biol*. 2017;429(12):1787-1799. doi:10.1016/j.jmb.2017.04.021

339. Ghorbani A, King J, Larijani M. AID pH-regulation is governed by a network of DNA-binding grooves residues and isoelectric point. Published online 2021.

340. Maitland PS. *Ecology of the River, Brook and Sea Lamprey*. Conserving Natura 2000 Rivers; 2003.

341. Meeuwig MH, Bayer JM, Seelye JG. Effect of Temperature on Survival and Development of Early Life Stage Pacific and Western Brook Lampreys. *Trans Am Fish Soc*. 2005;134:19-27.

342. Garner G, Hannah DM, Sadler JP, Orr HG. River temperature regimes of England and Wales: spatial patterns, inter-annual variability and climatic sensitivity. *Hydrol Process*. 2014;28(22):5583-5598. doi:10.1002/hyp.9992

343. Marianayagam NJ, Sunde M, Matthews JM. The power of two: protein dimerization in biology. *Trends Biochem Sci*. 2004;29(11):618-625. doi:10.1016/j.tibs.2004.09.006

344. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. *Protein Sci Publ Protein Soc*. 2002;11(6):1285-1299. doi:10.1110/ps.0201402

345. Smith HC. RNA binding to APOBEC deaminases; Not simply a substrate for C to U editing. *RNA Biol*. 2016;14(9):1153-1165. doi:10.1080/15476286.2016.1259783

346. Bhagwat AS, Carpenter MA, Bujnicki JM. Is AID a monomer in solution? *DNA Repair*. 2008;7(3):349-350; author reply 351-352. doi:10.1016/j.dnarep.2007.10.009

347. 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;113(11):E1545-E1554. doi:10.1073/pnas.1601678113

348. Keim C, Kazadi D, Rothschild G, Basu U. Regulation of AID, the B-cell genome mutator. *Genes Dev*. 2013;27(1):1-17. doi:10.1101/gad.200014.112

349. van Boven M, Weissing FJ. The evolutionary economics of immunity. *Am Nat*. 2004;163(2):277-294. doi:10.1086/381407

350. Moriyama Y, Koshiba-Takeuchi K. Significance of whole-genome duplications on the emergence of evolutionary novelties. *Brief Funct Genomics*. 2018;17(5):329-338. doi:10.1093/bfgp/ely007

351. Glasauer SMK, Neuhauss SCF. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol Genet Genomics MGG*. 2014;289(6):1045-1060. doi:10.1007/s00438-014-0889-2

352. Inoue J, Sato Y, Sinclair R, Tsukamoto K, Nishida M. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proc Natl Acad Sci*. 2015;112(48):14918-14923. doi:10.1073/pnas.1507669112

353. Ghorbani A. *PhD Thesis*. Doctoral Thesis. Memorial University of Newfoundland; 2021.

354. Star B, Nederbragt AJ, Jentoft S, et al. The genome sequence of Atlantic cod reveals a unique immune system. *Nature*. 2011;477(7363):207-210. doi:10.1038/nature10342

355. Bayersdorf R, Fruscalzo A, Catania F. Linking autoimmunity to the origin of the adaptive immune system. *Evol Med Public Health*. 2018;2018(1):2-12. doi:10.1093/emph/eoy001

356. Nakatani Y, Shingate P, Ravi V, et al. Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution. *Nat Commun*. 2021;12(1):4489. doi:10.1038/s41467-021-24573-z

357. Kuraku S. Impact of asymmetric gene repertoire between cyclostomes and gnathostomes. *Semin Cell Dev Biol*. 2013;24(2):119-127. doi:10.1016/j.semcdb.2012.12.009

358. Bell EA, Cable J, Oliveira C, Richardson DS, Yant L, Taylor MI. Help or hindrance? The evolutionary impact of whole-genome duplication on immunogenetic diversity and parasite load. *Ecol Evol*. 2020;10(24):13949-13956. doi:10.1002/ece3.6987

359. Marra NJ, Richards VP, Early A, et al. Comparative transcriptomics of elasmobranchs and teleosts highlight important processes in adaptive immunity and regional endothermy. *BMC Genomics*. 2017;18:87. doi:10.1186/s12864-016-3411-x

360. Jones JH. Primates and the Evolution of Long-Slow Life Histories. *Curr Biol CB*. 2011;21(18):R708-R717. doi:10.1016/j.cub.2011.08.025

361. Uriu K, Kosugi Y, Ito J, Sato K. The Battle between Retroviruses and APOBEC3 Genes: Its Past and Present. *Viruses*. 2021;13(1):124. doi:10.3390/v13010124

362. Ito J, Gifford RJ, Sato K. Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. *Proc Natl Acad Sci U S A*. 2020;117(1):610-618. doi:10.1073/pnas.1914183116

363. Finstad J, Good RA. THE EVOLUTION OF THE IMMUNE RESPONSE. *J Exp Med*. 1964;120(6):1151-1168.

364. Perey DY, Finstad J, Pollara B, Good RA. Evolution of the immune response. VI. First and second set skin homograft rejections in primitive fishes. *Lab Investig J Tech Methods Pathol*. 1968;19(6):591-597.

365. Mayer WE, O'Huigin C, Tichy H, Terzic J, Saraga-Babic M. Identification of two Ikaros-like transcription factors in lamprey. *Scand J Immunol*. 2002;55(2):162-170. doi:10.1046/j.1365-3083.2002.01026.x

366. Haire RN, Miracle AL, Rast JP, Litman GW. Members of the Ikaros gene family are present in early representative vertebrates. *J Immunol Baltim Md 1950*. 2000;165(1):306-312. doi:10.4049/jimmunol.165.1.306

367. Shintani S, Terzic J, Sato A, et al. Do lampreys have lymphocytes? The Spi evidence. *Proc Natl Acad Sci U S A*. 2000;97(13):7417-7422. doi:10.1073/pnas.110505597

368. Herrin BR, Alder MN, Roux KH, et al. Structure and specificity of lamprey monoclonal antibodies. *Proc Natl Acad Sci U S A*. 2008;105(6):2040-2045. doi:10.1073/pnas.0711619105

369. Xiong N, Raulet DH. Development and selection of gammadelta T cells. *Immunol Rev*. 2007;215:15-31. doi:10.1111/j.1600-065X.2006.00478.x

370. Basso K, Dalla-Favera R. Roles of BCL6 in normal and transformed germinal center B cells. *Immunol Rev*. 2012;247(1):172-183. doi:10.1111/j.1600-065X.2012.01112.x

371. Raghu D, Xue HH, Mielke LA. Control of Lymphocyte Fate, Infection, and Tumor Immunity by TCF-1. *Trends Immunol*. 2019;40(12):1149-1162. doi:10.1016/j.it.2019.10.006

372. Melichar HJ, Narayan K, Der SD, et al. Regulation of gammadelta versus alphabeta T lymphocyte differentiation by the transcription factor SOX13. *Science*. 2007;315(5809):230-233. doi:10.1126/science.1135344

373. Morimoto R, O'Meara CP, Holland SJ, et al. Cytidine deaminase 2 is required for VLRB antibody gene assembly in lampreys. *Sci Immunol*. 2020;5(45). doi:10.1126/sciimmunol.aba0925

374. Beamish F. Biology of the North American Anadromous Sea Lamprey, Petromyzon marinus. *Can J Fish Aquat Sci*. 1980;37(11):1924-1943.

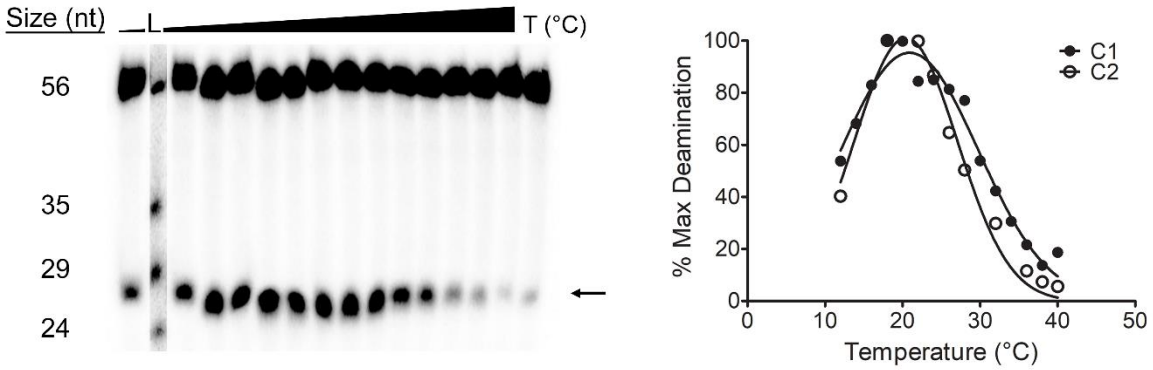
375. Lennox RJ, Bravener GA, Lin HY, et al. Potential changes to the biology and challenges to the management of invasive sea lamprey Petromyzon marinus in the Laurentian Great Lakes due to climate change. *Glob Change Biol*. 2020;26(3):1118-1137. doi:10.1111/gcb.14957

376. 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;5:18191. doi:10.1038/srep18191

377. Kidd JM, Newman TL, Tuzun E, Kaul R, Eichler EE. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet*. 2007;3(4):e63. doi:10.1371/journal.pgen.0030063

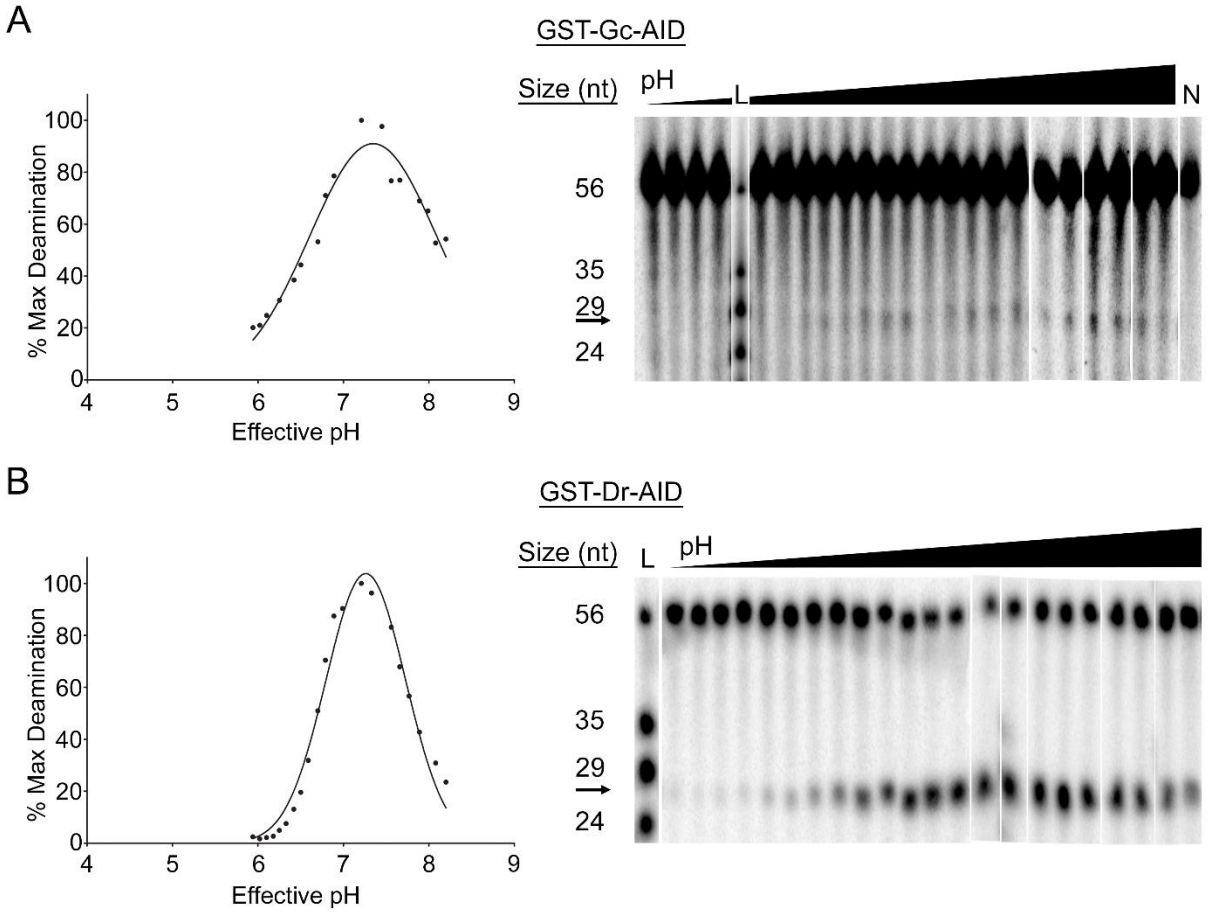
Appendices

Appendix 1 Thermosensitivity profile of prokaryotically expressed, purified, GST-Dr-AID



Left, representative alkaline cleavage gel of prokaryotically-expressed and purified GST-Dr-AID incubated with radioactively labelled TGCbub7 for 1 hour at a range of temperatures (12–40 °C). The black arrow indicates the product band at 28 nt. **Right**, graph of the results of two thermosensitivity experiments on two preparations, each of a unique clone (C1 and C2), of GST-Dr-AID. The y axis shows percentage of maximum deamination to correct for inter-preparation, -clone, and -experimental activity differences. The x axis shows temperature range in °C from 0 °C to 50 °C. The peaks of both experiments are at 20 °C and 21 °C for C1 and C2, respectively.

Appendix 2 pH profiles of prokaryotically expressed, purified, GST-tagged Gc-AID and Dr-AID



% maximum deamination graphs (**left**) of the results of pH experiments (representative alkaline cleavage gels, **right**) of prokaryotically-expressed, purified, GST-tagged Gc-AID (**A**) and Dr-AID (**B**). One preparation of one clone each of GST-Gc-AID and GST-Dr-AID were incubated individually with 50 fmol of radioactively labelled TGCbub7 for 5 (Dr) or 6 (Gc) hours at 25 °C in phosphate buffer at a range of pHs (5.94–8.20). The black arrow indicates the product band at 28 nt. The y axis in the graphs on the left show percentage of maximum deamination to maintain consistency with pH experiments in Chapter 5. The x axis show pH values from 4 to 9. The peaks of both experiments are at 7.2 for both Gc- and Dr-AID.