Intrinsic DNA Features are Determinants of Activation-Induced Cytidine Deaminase (AID) Recruitment and Activity

by Sarah Branton

A Thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science in Medicine (Immunology and Infectious Diseases)

Division of Biomedical Sciences, Faculty of Medicine

Memorial University of Newfoundland

May 2018

St. John's Newfoundland and Labrador

Abstract

While most cells strive to guard their genomic DNA from damage, B lymphocytes of the immune system actively damage their genomic DNA in order to mount a more robust antibody response. They do so by expressing the DNA-mutating enzyme activation-induced cytidine deaminase (AID). Although AID action is critical to antibody diversification, AID-mediated damage outside of antibody loci is a leading cause of leukemia/lymphomas. It is known that AID acts on single-stranded DNA and mutates genes that are highly transcribed; however, the mechanisms for AID targeting to specific genes or loci have yet to be elucidated. It has been hypothesized that one of the many factors involved in the targeting of AID is the topology of the DNA itself, as it is thought that AID can only deaminate supercoiled but not relaxed double-stranded DNA (dsDNA). We hypothesized that features of the DNA inherent to a gene (i.e. sequence, structure and topology) are important determinants of AID recruitment. Contrary to the current model that transcription significantly increases AID activity, we found that transcription is not necessary for AID activity, as AID efficiently deaminated both supercoiled and relaxed linear DNA in the absence of transcription. Moreover, DNA secondary structure may be of greater importance than primary sequence in attracting AID to its target, and that these structures may be liberated through dsDNA breathing and/or in conjunction with transcription.

Acknowledgments

I would like to express my deepest appreciation to my supervisor Dr. Mani Larijani for his continuous guidance and support throughout my Master's thesis project. Not only does he strive tirelessly to mold his students for success, his enthusiasm and passion for his field of research is infectious.

Thank you to my past and present colleagues Atefeh Ghorbani, Heather Fifield, Lesley Berghuis, Tim Caudle and Brittany Bolt for their contributions to this project.

A special thank-you to my supervisory committee members Dr. Michael Grant and Dr. Martin Mulligan for their constructive feedback.

I would also like to acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) as the funding agency of my Master's work.

Table of Contents

Abstract……....i

Acknowledgements……..…..ii

Table of Contents…….…….iii

List of Figures……vi

List of Tables……...…viii

List of Abbreviations……..…..ix

Chapter 1: Introduction……..1

1.1: The Adaptive Immune Response……..1

1.2: Primary and Secondary Antibody Diversification Processes……...4

1.3: The enzyme Activation-induced Cytidine Deaminase (AID) initiates Class Switch Recombination and Somatic Hypermutation in Mature B cells..……7

1.4: AID and Cancer……...……...10

1.5: The Role of Transcription in AID Recruitment……..……11

1.6: Targeting of AID is Influenced by both Sequence and Structure of the DNA Substrate…….18

1.7: Strand Topology and its Relation to Transcription……..……..20

1.8: Project Rationale…….23

Chapter 2: Materials and Methods…….…..……26

2.1: Preparation and Purification of the Supercoiled and Linear DNA Substrates…….....26

2.2: Topoisomerase I Assay to Verify Supercoiled DNA Topology…….....29

2.3: Southern Blot analysis to Verify Purity of Double-stranded Supercoiled and Relaxed Linear DNA……..……...30

2.4: Expression and Purification of Wildtype Human AID……...……31

2.5: Alkaline Cleavage Deamination Assay…….....32

2.6: Transcription-independent AID Activity (TIAA) Assay……..….....34

2.7: Bisulfite Deamination Assay……...…….....34

2.8: Transcription-associated AID Activity (TAAA) Assay……...……..35

2.9: Deamination-specific PCR……..……..36

2.10: Degenerate PCR, Purification and Analysis of Substrate DNA…...……37

2.11: Predicting Template DNA Secondary Structure using mfold……..40

Chapter 3: Results……..…….....41

3.1: Designing an Unbiased Assay to Examine AID Targeting and Activity…...41

3.2: Preparation of Supercoiled and Linear DNA Substrates……51

3.3: AID can Mutate Relaxed Duplex DNA in the Absence of Transcription.....57

3.4: Verifying AID Activity Using Degen-PCR……..….....60

3.5: Verification of Size and Substrate Preference of GST-AID……...……64

3.6: Confirming that AID is Responsible for the Observed Mutations…..……...66

3.7: Southern Blot Analysis of DNA templates to confirm absence of ssDNA....68

3.8: Confirming AID activity on Supercoiled DNA using AID-His…….....72

3.9: Gel Extraction After TIAA Assay……..75

3.10: Optimizing the *In Vitro* AID Activity Assay to Observe the Unaltered and Original Foot-print of AID Activity on dsDNA in the Absence of Transcription……..…….....80

3.10.1: "Protecting" Uracils using Uracil DNA Glyscosylase Inhibitor (UGI)…81

3.10.2: Determining the "Optimal" Ratio of AID:DNA to Accurately View AID Activity in the *In Vitro* TIAA Assay…….…….……90

3.10.3: AID Mutates Supercoiled DNA with a 10-100-fold Preference over Linear DNA…….....98

3.11: GST-AID Mutated both Supercoiled and Relaxed Linear dsDNA in the Degen-PCR TIAA Assay…….……108

3.12: Using Bisulfite to Generate a Model of Breathing DNA……...114

3.13: AID can act in both a Processive and Distributive Manner…….……..118

3.14: DNA Secondary Structure is more Important than Primary Sequence in AID-targeting……...122

3.15: Transcription Increases AID's Accessibility to Target DNA……..……...133

Chapter 4: Discussion……...…….....146

4.1: AID can Target and Mutate Both Supercoiled and Relaxed Linear DNA without Transcription……..…....146

4.2: AID Activity without Transcription cannot be Solely Explained by DNA Breathing……..148

4.3: AID can Mutate in both a Processive and Distributive Pattern on dsDNA without Transcription……...……150

4.4: AID can Target Both DNA Strands without Transcription……..152

4.5: Secondary Structure is a more Important Determinant of AID Targeting than Primary Sequence in the Absence of Transcription…….154

4.6: Transcription Changes the Pattern of AID Targeting…….....155

Chapter 5: Future Directions……...……...158

Chapter 6: Concluding Remarks……...…….161

References……...…….....162

Figures

Figure 1: General Antibody Structure……..….....3

Figure 2: Secondary Antibody Diversification Processes…….……..….....6

Figure 3: AID Initiates Secondary Antibody Diversification Processes……..….....9

Figure 4: Transcription Generates Secondary Structures that AID may Target….....17

Figure 5: Expression Vector Backbone…….....28

Figure 6: Assay Overview…….....47

Figure 7: Detecting C-T and G-A Mutations Using Non-specific Degenerate Primers……...49

Figure 8: The Three DNA Topologies Analyzed in the Transcription-free AID Activity Assay……...……...53

Figure 9: Preparation of Substrates……..…….54

Figure 10: Verifying Supercoiled Substrate Topology Using TopoI……...……56

Figure 11: AID Consistently Mutated the Relaxed Nicked and Linear Duplex DNA in the Absence of Transcription……58

Figure 12: Rate of AID-mediated C-T and G-A Mutations after degen-PCR…...……...63

Figure 13: GST-AID is the Correct Size and Targets its Known Substrates…….....65

Figure 14: Southern Blots of Supercoiled and Linear DNA in the presence and absence of AID……70

Figure 15: Rate and Distribution of AID-His-mediated C-T and G-A Mutations in Supercoiled DNA……...74

Figure 16: Modified Assay Schematic Including Gel Extraction……...……..76

Figure 17: Rate of AID-mediated C-T and G-A Mutations in Substrate DNA Gel Extracted after Treatment with GST-AID……...……..79

Figure 18: Alkaline Cleavage Assay to Determine UDG Activity in our GST-AID prep and Test Amount of UGI needed to Counteract UDG Activity…….……83

Figure 19: Testing UGI in the Deamination-specific PCR-based Assay…….....86

Figure 20: C-T Mutation Map of deam-PCR Amplicons with or without addition of UGI……89

Figure 21: Degen-PCR Results for Dilutions Assay……..……..94

Figure 22: Deam-PCR Results for Dilutions Assay……..……...95

Figure 23: Quantification of Degen-PCR and Deam-PCR Results for Dilutions Assay….....96

Figure 24: Deam-PCR Dilutions for Supercoiled and Relaxed Linear DNA treated with purified GST-AID……100

Figure 25: Deam-PCR Dilutions for Supercoiled and Relaxed Linear DNA treated with purified AID-His……..……104

Figure 26: Deam-PCR Dilutions for Supercoiled and Relaxed Linear DNA treated with AID-His Lysate…….……...106

Figure 27: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations in Supercoiled and Relaxed Linear Template DNA……...….....112

Figure 28: Distribution of Bisulfite-mediated C-T and G-A Mutations in Supercoiled and Relaxed Linear Template DNA…….……..117

Figure 29: Mutation Maps for all Individual Amplicons Incubated with either Bisulfite or GST-AID……...……...121

Figure 30: Analysis of Primary Sequence of GST-AID and Bisulfite-mediated Mutations…….....124

Figure 31: Secondary Structure of the Target DNA Sequence……...……126

Figure 32: Bisulfite-mediated C-T Mutations on Supercoiled Substrate Superimposed onto the Template DNA Secondary Structure…….……129

Figure 33: GST-AID-mediated C-T Mutations on Supercoiled Substrate Superimposed onto the Template DNA Secondary Structure…….....130

Figure 34: Stability of Paired or Unpaired Regions within Predicted Secondary Structures…….....132

Figure 35: RNA is Produced only in the Presence of T7 RNAP……..……..136

Figure 36: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations during *in vitro* transcription (set 1)……..……138

Figure 37: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations during *in vitro* transcription (set 2)……..……142

Figure 38: Primary Sequence Analysis of GST-AID-mediated Mutations during *in vitro* transcription…….....144

Tables

Table 1: Number and Rate of C-T and G-A Mutations after degen-PCR…….62

Table 2: C-T and G-A Mutations Observed on the Target Substrate after Incubation with the Catalytically Dead AID Mutant W80R……..…..……67

Table 3: Rate and Distribution of AID-His-mediated C-T and G-A Mutations in Supercoiled DNA……...73

Table 4: Number and Rate of Mutations in Substrate DNA Gel Extracted after Treatment with GST-AID…….....78

Table 5: Number and Rate of AID-mediated C-T Mutations from the Deam-PCR Assay…….....88

Table 6: Number of Copies of Plasmid DNA per Number of Nanograms…….....91

Table 7: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations in Supercoiled and Relaxed Linear Template DNA……111

Table 8: Rate and Distribution of Bisulfite-mediated C-T and G-A Mutations in Supercoiled and Relaxed Linear Template DNA……116

Table 9: Number and Percentage of Bisulfite- and GST-AID-mediated C-T Mutations within the Predicted Target DNA Secondary Structure…….…….....131

Table 10: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations during *in vitro* transcription (set 1)……...……...……137

Table 11: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations during *in vitro* transcription (set 2)……..……141

Abbreviations

AID: Activation-induced Cytidine Deaminase

Amp^r: Ampicillin resistance

AP: Apurinic/Apyrimidinic

APOBEC: Apolipoprotein B mRNA editing enzyme catalytic polypeptide-like

BER: Base Excision Repair

CDRs: Complementary Determining Regions

ChIP-seq: Chromatin immunoprecipitation with deep sequencing

CMV: Cytomegalovirus

C Region: Constant region

Ck: Kappa gene of constant region

CSR: Class Switch Recombination

C-terminal: Carboxy-terminal

GST-AID: Glutathione S-transferase-tagged wildtype human AID

H.D.: Heat-denatured

His-AID: Polyhistidine-tagged wildtype human AID

hnRNP: Heterogeneous nuclear ribonucleoproteins

IPTG: Isopropyl-β-D-thiogalactopyranoside

LB: Luria broth

MHC: Major histocompatibility complex

MMR: Mismatch Repair

µs: Microseconds

NHEJ: Non-homologous End Joining

Nt.: Nucleotide

N-terminal: Amino-terminal

PAF1: RNA polymerase II associated factor I

PMSF: Phenylmethylsulfonyl fluoride

PTBP2: Splicing regulator polypyrimidine tract binding protein 2

RNAP: RNA polymerase

rNTP(s): Ribonucleoside tri-phosphate(s)

RPA: Replication Protein A

SHM: Somatic Hypermutation

S region: Switch region

SSC: Saline-sodium citrate

ssDNA: Single-stranded DNA

TAAA Assay: Transcription-associated AID Activity Assay

TBE: Tris/Borate/EDTA

TE: Tris/EDTA

TIAA Assay: Transcription-independent AID Activity Assay

TopA: Topoisomerase I

TSS: Transcription start site

V region: Variable region

Vk: Kappa gene of variable region

UDG: Uracil-DNA Glycosylase

UGI: Uracil Glycosylase Inhibitor

UNG: Uracil-N-Glycosylase

I. Introduction

1.1 The Adaptive Immune Response

Over the course of evolution, vertebrates have developed an effective adaptive immune system capable of identifying, targeting, and eliminating pathogens. Unlike innate immunity, which recognizes patterns of pathogens nonspecifically, the adaptive immune system is highly specific and can form memory that triggers a faster and more robust immune response when a given pathogen is reencountered (Owen et al. 2013). The adaptive immune system consists largely of T and B lymphocytes, which recognize antigenic determinants through either T-cell receptor molecules or surface immunoglobulin, respectively. The T-cell receptor is expressed on the surface of T lymphocytes, where it recognizes processed antigenic fragments in the context of major histocompatibility complex (MHC) molecules. Antibodies are antigen-binding proteins that can either be bound to the surface of B lymphocytes, or secreted as soluble molecules that circulate in the blood. In general, antibodies are composed of four polypeptide chains – two small light chains and two larger heavy chains – linked by disulfide bonds (Figure 1). The first 110 amino acids from the N-terminus of either the heavy or the light chain form the Variable region of the antibodies. Within the Variable region are smaller regions termed Complementary Determining Regions (CDRs), which are highly variable in amino acid sequence. The variability in these regions differ between antibodies of the same class, and allow the protein structure to form a unique shape that is complementary to a given epitope. This variability governs the high specificity by which antibodies bind to antigens. The rest

of the amino acid sequence of both the heavy and light chains are kept constant within antibodies of a given class, and is therefore named the Constant region. The Constant region determines the biological properties of the immunoglobulin class. There are five different classes of antibodies – IgA, IgD, IgE, IgG and IgM. IgG is a serum immunoglobulin, representing about 80% of antibodies found in the serum. IgA is found predominantly in secretions such as breast milk, saliva, tears and mucus. IgD and IgM are the major membrane-bound antibodies on the surface of mature B cells, while IgE is responsible for the hypersensitivity reactions such as hay fever, asthma, hives and anaphylactic shock.

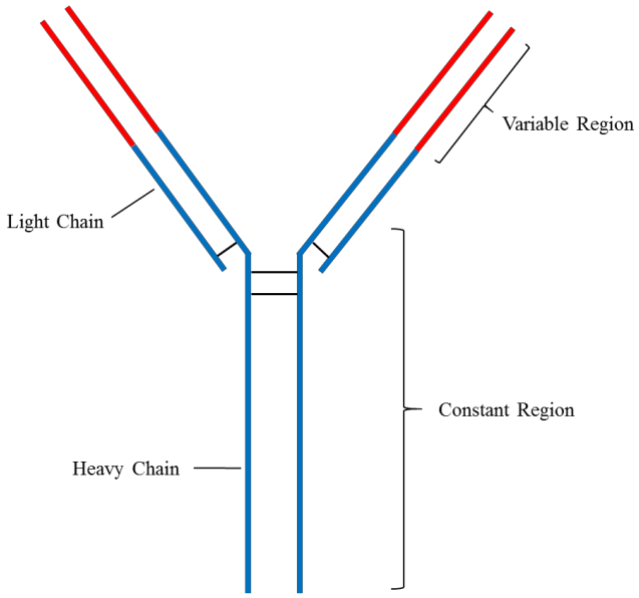


Figure 1: General Antibody Structure. Antibodies are composed of two identical large heavy chains, as well as two identical smaller light chains, linked by disulfide bonds. The first 110 amino acids from the N-terminus of both the heavy and light chains form the Variable region of the antibody. This region is highly variable, enabling it to bind antigens with high specificity. The remainder of the antibody is the constant region. There are few differences in the constant region of antibodies within a given class.

1.2 Primary and Secondary Antibody Diversification Processes

To achieve antigenic specificity, the antibody-encoding immunoglobulin genes must first undergo a primary diversification process known as V(D)J recombination (Owen et al. 2013). B cells use V(D)J recombination to assemble exons encoding immunoglobulin heavy and light chain variable regions, upstream of the corresponding constant region exons during their development (Matthews et al. 2014). V(D)J recombination involves combinatorial rearranging and joining of the V-J regions of the light chain or the V-D-J segments of the heavy chain genes, generating an abundance of low affinity antigen receptors. V(D)J recombination is regulated by allelic exclusion, ensuring that each B cell produces antibodies with a single antigenic specificity. The rearranged heavy and light chain genes will only be expressed from a single chromosome, preventing multiple copies of functional V-D-J or V-J exons in the heavy and light chain gene loci (Honjo et al. 2002).

After V(D)J recombination, naïve B cells move towards the secondary lymphoid organs, including the spleen and lymph nodes, where they are first exposed to foreign antigens (Honjo et al. 2002; DeFranco, 2016). When an antigen binds to the low-affinity receptor of a mature B cell, the cell becomes activated to undergo further genetic alterations. During secondary antibody diversification processes, the immunoglobulin loci are randomly mutated and rearranged again. One such secondary diversification process is Somatic Hypermutation (SHM), in which point mutations are induced in the V-D-J or V-J region loci (Figure 2). B cells that have randomly gained higher affinity antibodies are then preferentially selected by a limited pool of antigen, while those that have randomly gained lower affinity or stop codons undergo apoptosis (Peters and Storb, 1996). This selection

results in a pool of B cells with an overall higher affinity towards the antigen than the initial activated mature B cell. After antigenic stimulation, antibodies can also undergo Class Switch Recombination (CSR) in which double-stranded breaks occur in the Constant region of the immunoglobulin locus. When double-stranded breaks occur the C region variant (ex. Cμ) can be looped out and the Switch region can be rejoined to a different one of the other 7 variants (ex. Cγ3). CSR results in a different class of antibody (e.g. IgM to IgG) with differing biological function, without changing its antigenic specificity (i.e. V region) (Zanotti and Gearhart, 2016; Figure 2). Activated mature B cells can further differentiate into plasma or memory B cells (Berek et al. 1991). Plasma B cells secrete into the circulation large volumes of antibodies, which bind the same epitope that evoked the initial proliferation. After the primary infection, a fraction of the plasma B cells remain in the body as memory B cells. Upon repeated exposure to the pathogen, these can initiate the secondary immune response, which is both faster and more robust than the primary response when antigen was encountered for the first time.

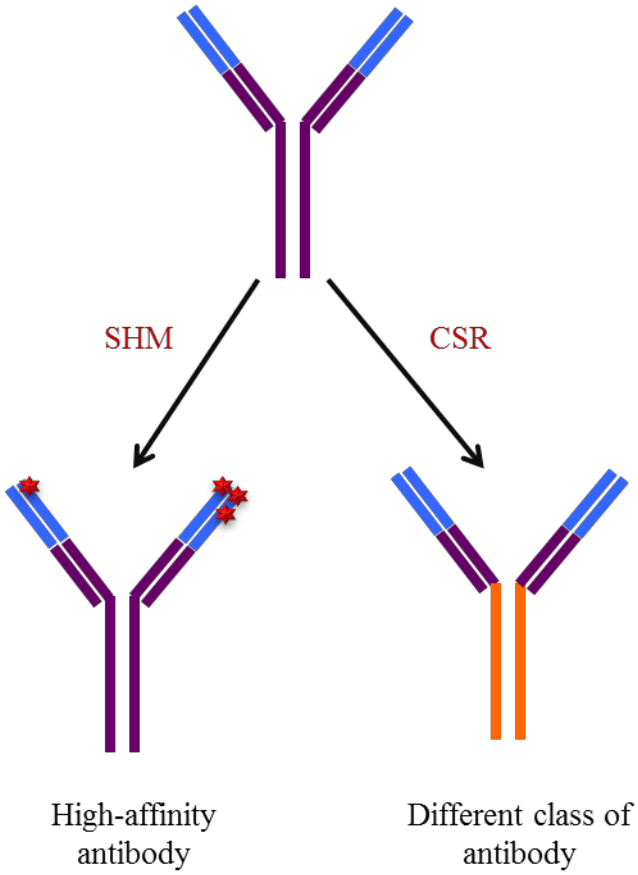


Figure 2: Secondary Antibody Diversification Processes. In the secondary lymphoid organs B cells encounter antigen through their low-affinity receptors, which stimulates them to undergo secondary antibody diversification. During somatic hypermutation, point mutations are randomly introduced into the V-J regions of the light chain and/or the V-D-J regions of the heavy chain. B cells whose antibodies have gained higher affinity toward the antigen survive, while those that gain a lower affinity undergo apoptosis. Antibodies may also undergo class switch recombination in which B cells switch from producing one isotype of antibody to another (ex. IgM to IgG), allowing them to achieve different biological activity.

1.3 The enzyme Activation-induced Cytidine Deaminase (AID) initiates Class Switch Recombination and Somatic Hypermutation in Mature B cells

The human AID gene is located on chromosome 12p13 (Honjo et al. 2002), and encodes a small 198 amino acid enzyme with a molecular mass of 24 kDa (Muramatsu et al. 1999). It is a member of the Apolipoprotein B mRNA editing enzyme catalytic polypeptide-like (APOBEC) family of cytidine deaminases that is specific to germinal center B cells (Muramatsu et al. 2000). It was first postulated (Muramatsu et al. 2007) that AID is an RNA-editing enzyme because of its primary sequence similarity to APOBEC-1, which edits RNA. Later, it was shown that AID binds and exclusively mutates deoxycytidine (dC) to deoxyuridine (dU) in single-stranded DNA (ssDNA) but not in RNA or in double-stranded DNA (dsDNA) *in vitro* (Bransteitter et al. 2003, Pham et al. 2003, Dickerson et al. 2003, Larijani et al. 2005a, Larijani and Martin 2007, Larijani et al. 2007). By mutating ssDNA regions within the Immunoglobulin loci AID initiates the secondary antibody diversification processes class switch recombination (CSR) and somatic hypermutation (SHM), allowing mature B cells to expand their antibody repertoire (Muramatsu et al. 1999, Muramatsu et al. 2000). AID is critical to achieving a robust humoral immune response as inherited defects in enzymatic function lead to Hyper IgM syndrome (Revy et al. 2000), which is characterized by recurrent and severe infections, and an increased risk for opportunistic infections and cancer.

AID's initial mutation alone is not sufficient to generate antibody diversity, instead it triggers many downstream processes that each contribute to creating diversity. When AID mutates dC to dU, a U•G mismatch is generated that can be either replicated or

recognized by the mismatch repair (MMR) or uracil-removal base excision repair (BER) pathways (Martin and Scharff 2002; Figure 3). For reasons not yet fully appreciated, when these repair pathways act downstream of AID in mature B cells, they utilize error-prone rather than high fidelity repair DNA polymerases. Consequently, more mutations are generated in the surrounding residues of the U•G mismatch, leading to an entire spectrum of mutations (i.e. from any of the four bases to any of the other three) (Larijani and Martin 2012). The MMR pathway could also engage Uracil-N-Glycosylase (UNG), which removes dU generating an abasic site (Petersen-Mahrt et al. 2002). Replication over the abasic site can lead to further transition or transversion mutations, or it can be used as a substrate for apurinic/apyrimidinic (AP) endonuclease to create a single-stranded nick (Di Noia et al. 2002). The BER pathway can then repair the nick, eliminating any mutations. Alternatively, the cleaved ends can be joined together by the Non-Homologous End Joining (NHEJ) repair system, which can also lead to CSR if the initial AID-mediated lesion occurred in the switch region of the antibody gene (Honjo et al. 2002). These error-prone repair processes create antibody diversity, so the appropriate antibody can be selected for through B cell selection. Antibodies with a high affinity for the antigen are selected, while those with low affinity for the antigen are destroyed (Rajewsky 1996).

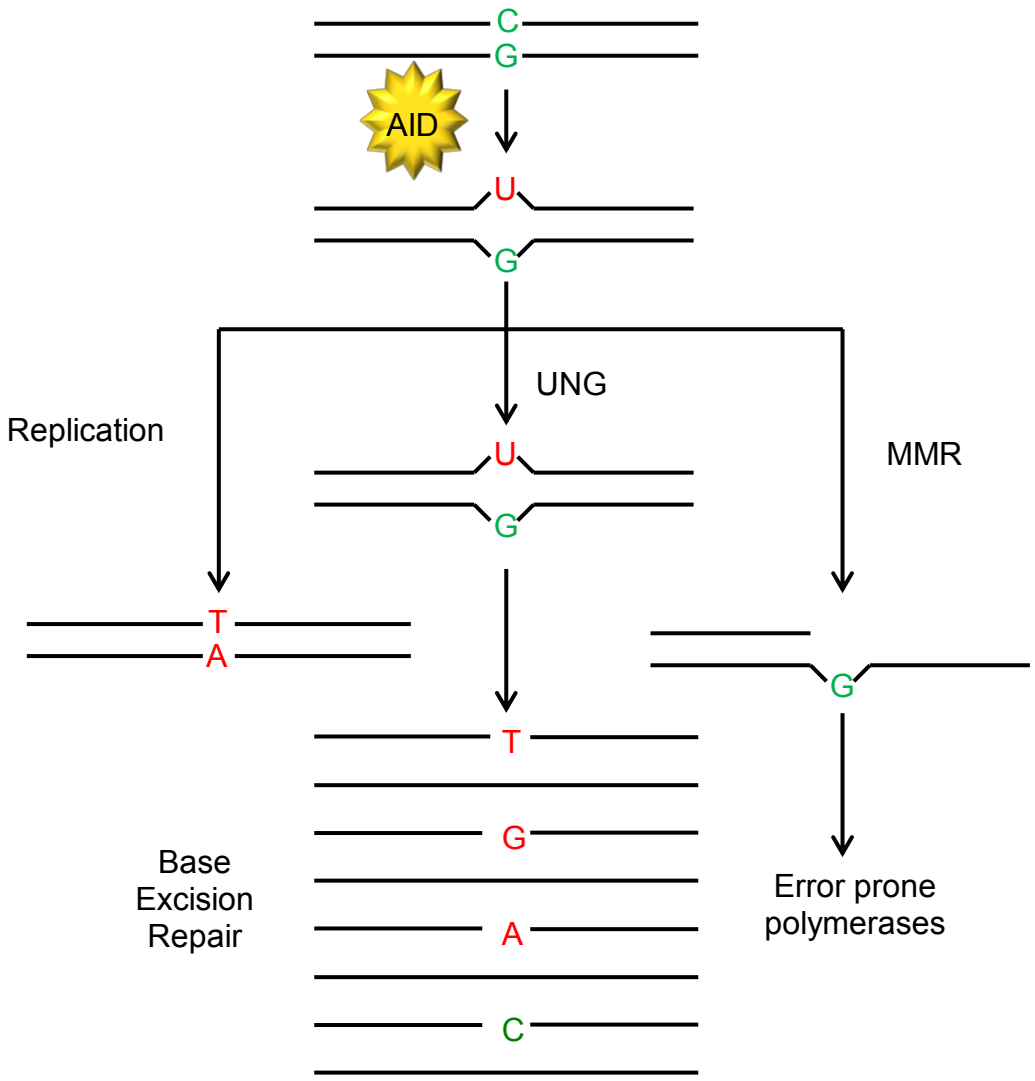


Figure 3: AID Initiates Secondary Antibody Diversification Processes. AID mutates dC to dU within Immunoglobulin genes, triggering downstream error-prone processes that lead to secondary antibody diversification. The uracil can be recognized as a T and replicated over, replacing the original C-G pair with a T-A pair. Alternatively, the BER or MMR pathways can be activated. If the BER pathway is activated, UNG removes the uracil, creating an abasic site in which error-prone polymerases fill the gap with any of the four nucleotides. If the MMR pathway is activated, endonucleases nick the DNA surrounding the U-G mismatch and remove a section of the DNA including the original mutation. Error prone polymerases can then "repair" the section of DNA, creating even more mutations and more diversity within the variable region of the antibody.

1.4 AID and Cancer

Although AID plays an important role in antibody diversification, it has genome-wide access and can mutate genes outside of the immunoglobulin loci. Chromatin immunoprecipitation (ChIP) assays in combination with deep sequencing (ChIP-seq) has shown that AID interacts with approximately 6,000 genes in stimulated murine B cells (Yamane et al. 2011). Off-target activity on non-immunoglobulin genes can lead to mutations and chromosomal translocations leading to genomic instability and cancer (Gazumyan et al. 2012; Gostissa et al. 2011; Robbiani and Nussenzweig, 2013). For example, c-MYC-IGH chromosomal translocations are frequently found in Burkitt's lymphomas (Aukema et al. 2014; Osborne et al. 2007), while BCL2-IGH translocations are found in follicular and diffuse large cell lymphomas (Xerri et al. 2016, Gomez et al. 2005). AID can also produce tumor-driving as well as secondary mutations that augment tumor progression (Kumar et al. 2014). Aberrant AID activity is associated with various lymphomas and leukemias, including Burkitt's lymphoma, diffuse large B cell lymphoma and chronic lymphocytic leukemia (Gruber et al. 2010; Müschen et al. 2000). It is also associated with numerous types of solid tissue tumors, such as lung tumors, gastric tumors and various carcinomas (Kumar et al. 2014). The mechanism behind what leads AID to either stay on-track or go off-target remains elusive. It is thought that off-target AID activity is associated with targeting of highly transcribed genes and/or increased expression levels (Klemm et al. 2009; Qian et al. 2014; Yamane et al. 2011), however the mechanism remains inconclusive. Understanding the molecular properties and mechanisms by which AID is targeted to DNA *in vitro* will allow a better understanding of its role in immunity

and oncogenesis. In the future this information may help in the development of therapies used to prevent the development and progression of these cancers.

1.5 The Role of Transcription in AID Recruitment

Currently there are three models for AID recruitment to a particular gene: 1) targeting by transcription machinery and/or specific protein cofactors serving to chaperone AID to specific loci; 2) targeting by recognition of specific primary DNA sequences; and 3) targeting due to recognition of certain ssDNA structures and topologies generated during the process of transcription. The process of transcription integrates the role(s) of co-factors, and DNA sequence, structure and topology. The precise function(s) of each feature in relationship to AID targeting remains elusive. Furthermore, the transcription of Ig genes has unique features that may enable AID targeting such as the requirement of Ig enhancers for SHM and CSR, considerable RNA polymerase (RNAP) pausing during transcription of S repeats, and displacement of newly transcribed switch region RNA by the RNA exosome complex (discussed in Qian et al. 2014). Moreover, AID has been shown to target super-enhancers in both human and mouse B cells especially those in highly transcribed genes (Meng et al. 2014, Qian et al. 2014). Altogether, understanding the process of transcription within the surrounding genomic microenvironment is crucial to determining how AID targets Ig genes and what leads it to stray.

It has been suggested that AID is associated with various elements of the transcription machinery and/or transcription-associated factors, such as RNA polymerase II (Nambu et al. 2003), the ssDNA binding protein Replication Protein A (RPA)

(Chaudhuri et al. 2004), the transcription elongation factor Spt5 (Pavri et al. 2010), RNA polymerase II associated factor I (PAF1) (Willmann et al. 2012), spliceosome-associated factor CTNNBL1 (Conticello et al. 2008), RNA binding heterogeneous nuclear ribonucleoproteins (hnRNP) (Hu et al. 2015), and splicing regulator polypyrimidine tract binding protein 2 (PTBP2) (Nowak et al. 2011). To date there are more than two dozen co-factors that have been suggested to either directly bind AID or associate indirectly through other proteins, DNA or RNA (discussed in King and Larijani 2017, Larijani and Martin 2012). Although potentially associated with AID, these co-factors are not absolute requirements for AID activity, as is has been shown numerous times that AID is fully capable of mutating ssDNA *in vitro* in the absence of any other factor (King et al. 2015; Abdouni et al. 2013; Dancyger et al. 2012; Larijani and Martin 2007; Larijani et al. 2007; Larijani et al. 2005 a,b).

It has also been suggested that AID targets certain sequences intrinsic to Ig genes. It has been shown using a DT40 chicken B cell line that CAGGTG cis-elements found in Ig enhancers are sufficient to target SHM to a transgene within 1 kb (Tanaka et al. 2010). This effect was not seen when CAGGTG was replaced with AAGGTG, and the authors suggested that the CAGGTG motif may attract AID to a target gene. An earlier study using an $Aid^{+/-}$ C57BL/6 mouse model also supported this notion with the finding that CAGGTG motifs are within approximately 2 kb of most genes that interact with AID (Liu et al. 2008). Besides directly attracting AID through motifs, sequence may play another role in changing the structure or topology of genomic DNA. For example, A-T rich regions are more susceptible to DNA melting due to the two hydrogen bonds of A-T pairs versus the three

hydrogen bonds that form G-C pairs. Multiple 46 base pair regions with greater than 72% A/T-richness exist within the immunoglobulin heavy chain locus (Webb, 2001). These regions serve as binding sites for the B cell regulator of immunoglobulin heavy chain transcription, or Bright. Bright increases Ig gene transcription three- to seven-fold in activated B cells. While not directly related to AID, increasing transcription rates at immunoglobulin loci may give AID more of an opportunity to access the secondary structures generated by transcription. Moreover, if A/T-rich regions exist within off-target genes AID may have greater access to these regions through temporary single-strandedness in breathing "naked" DNA upstream of RNAP. Another example how sequence may influence structure is through generation of stem-loops by palindromic sequences. Switch region sequences in mammals, chickens and frogs all contain short palindromic sequences that possibly form stem-loop structures during transcription (Tashiro et al. 2001). It has been suggested that the stem-loops form a secondary structure that can be recognized by CSR machinery and thus be targeted leading to efficient CSR. It cannot yet be ruled out that similar structures can also be targeted by AID.

Both SHM and CSR have two requirements – transcription and AID (Goyenechea et al. 1997, Peters and Storb 1996, Fukita et al. 1998, Pinaud et al. 2001, Betz et al. 1994, Zhang et al. 1993, Rothenfluh et al. 1993, Manis et al. 2002, Dudley et al. 2002, Muramatsu et al. 2000, Wang and Wabl 2004, Xu et al. 2012, Okazaki et al. 2002, Yoshikawa et al. 2002). In 1996, Peters and Storb first proposed a model linking SHM to transcription (Peters and Storb, 1996). They noticed that both the Vk and Ck regions of an Igk transgene could be mutated if they were preceded by a Vk promoter. Normally the C regions of Ig

genes are not mutated during SHM, showing the importance of transcription initiation at the appropriate promoter for productive SHM. Moreover, a study using the hypermutating cell line 18-81 has shown that the rate of transcription correlated with the rate of mutation, where increased transcription levels lead to increased mutation rates (Bachl et al. 2001). Early studies proposed a mutating factor to be associated with transcription, and thereby induce SHM (Peters and Storb 1996). That "mutating factor" was later identified as AID (Muramatsu et al. 2000, Revy et al. 2000).

Transcription bridges the gap between the need for AID to initiate antibody diversification and AID's requirement for ssDNA. During transcription, the structure of double-stranded genomic DNA is altered – chromatin is de-condensed, nucleosomes are remodeled, and a looser conformation is temporarily adopted (Kouzine et al. 2013). In the absence of transcription, nucleosomes are thought to physically block the targeting of AID to the DNA (Shen et al. 2009). During an *in vitro* transcription assay Shen and colleagues exposed AID to nucleosomal DNA with and without T7 polymerase. They found that AID could only mutate within nucleosomal positioning sequences during transcription. Moreover, when naked DNA was compared to nucleosomal DNA, AID-mediated mutations were only observed in the nucleosomal positioning sequences when no nucleosomes were present suggesting that AID can efficiently mutate the nucleosomal positioning sequences if it can gain access to them. The authors hypothesized that nucleosomes may prevent the "flipping-out" of cytosines produced by negative supercoiling near nucleosome positioning sequences (Shen et al. 2009). However, during elongation of the RNA transcript nucleosomes are disassembled providing AID with ample

opportunity to target the transiently naked DNA strand before reassembly after the RNA polymerase has passed. Chromatin remodeling (Clapier and Cairns 2009) and histone chaperones (Das et al. 2010) are thought to further enhance the targeting of AID through the destabilization and disassembly of nucleosomes prior to the passage of the RNA polymerase. An alternate hypothesis was later proposed that nucleosomes are not disassembled, and that AID may access its substrate through the partial unwrapping of DNA and/or nucleosome repositioning (Kodgire et al. 2012). In either case, transcription is still essential to expose "naked" genomic DNA to AID.

Even though it is known that transcription is essential for AID to gain access to chromatinized genomic DNA, the exact mechanism remains to be elucidated. Early *in vitro* transcription studies focused on the possibility of AID targeting the nontranscribed strand within the transcription bubble while its sister DNA strand interacts with RNA (Ramiro et al. 2003, Sohail et al. 2003). However, the transcription bubble created by T7 RNA polymerase (RNAP) is approximately 9 bp long, surrounded by the physically hefty elongation complex that extends ~21 bp (Huang and Sousa 2000). It is unlikely that AID would be able to physically access the nontranscribed strand within the transcription bubble. Even if AID could physically access the transcription bubble, the rate of elongation by eukaryotic RNAP II was determined to be ~23 nt/s (Shermoen and O'Farrell 1992), only allowing AID a brief instant to bind and mutate target DNA. Given that AID is a slow enzyme with a rate of deamination of 0.03 s^{-1} (Larijani and Martin 2007, Pham et al. 2011, Mak et al. 2013, Mak et al. 2015), and that only ~1 out of 100 ssDNA targets encountered by AID is actually bound over the catalytic pocket to be mutated (King et al. 2015, King

and Larijani 2017), the possibility of AID having productive access to the ssDNA within the transcription bubble itself is also unlikely.

An alternate model is that transcription generates secondary structures that can provide AID with transient ssDNA substrates. For example, transcription through G-rich CSR switch regions is known to promote formation of RNA-DNA hybrids known as R-loops (Yu et al. 2003, Fugmann and Schatz 2003). When transcription occurs using the G-rich strand as a template, the RNA product remains annealed and the C-rich top strand loops out creating a temporary ssDNA region (Figure 4). Transcription exposes naked DNA through chromatin remodeling, allowing the formation of secondary DNA structures such as stem-loops, cruciforms (Dayn et al. 1992), bubbles (Kuetche, 2016), and negative supercoiling upstream of the transversing RNAP (Krasilnikov et al. 1999; Kouzine et al. 2013; Naughton et al. 2013). It is possible that AID can target the ssDNA regions of structures such as stem-loops and bubbles, as well as the ssDNA regions generated by the loosened topology of negatively supercoiled (under-wound) DNA (Figure 4). This model would allow AID to mutate both the transcribed and nontranscribed strands, consistent with the observation that there is no strand preference for mutations in SHM (Storb et al. 1999), and both strands get mutated somewhat equally *in vivo*.

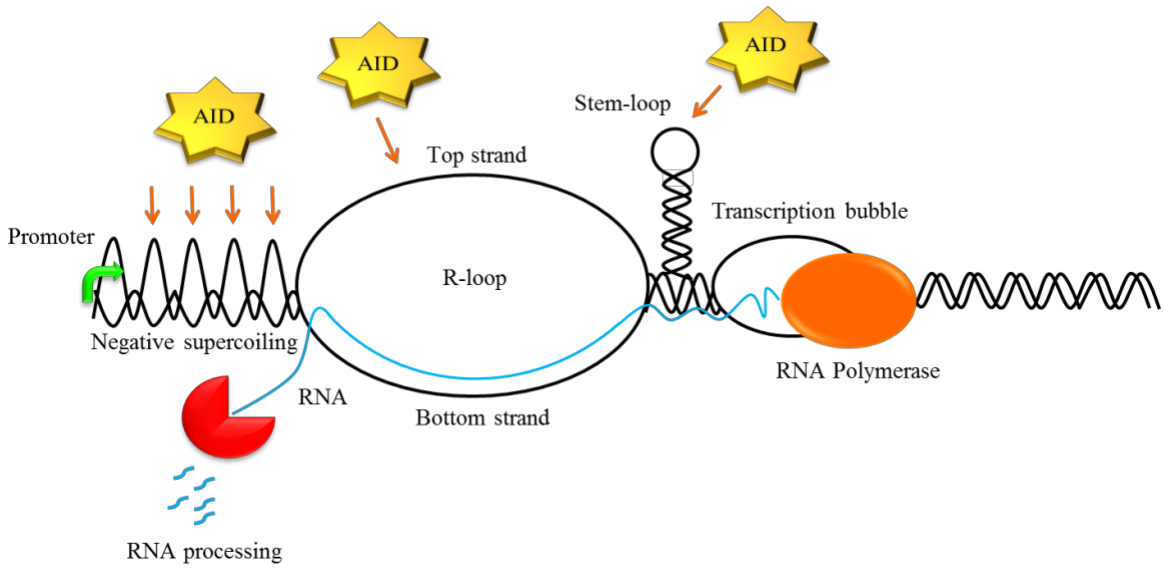


Figure 4: Transcription Generates Secondary Structures that AID may Target. Through chromatin decondensation and histone remodeling transcription liberates naked DNA including various secondary DNA structures such as R-loops, stem-loops, and negative supercoils upstream of RNAP. These structures may provide AID access to otherwise inaccessible double-stranded genomic DNA. AID may target the single-stranded region of stem-loops, or the ssDNA region of R-loops while the sister DNA strand is interacting with RNA. Negative supercoiling generated in the wake of the transversing RNAP may provide AID with access to either the transcribed or nontranscribed strand of DNA through dynamic breathing due to the torsional strain. It is possible that these secondary structures allow AID to mutate either one or both DNA strands, leading to unbiased SHM.

1.6 Targeting of AID is Influenced by both Sequence and Structure of the DNA Substrate

Numerous *in vitro* studies have shown that the biochemical substrate for AID is single-stranded DNA (ssDNA) (Bransteitter et al. 2003, Pham et al. 2003, Dickerson et al. 2003, Larijani et al. 2005, Larijani and Martin 2007, Larijani et al. 2007, Pham et al. 2007, Brar et al. 2008) on which AID acted with a small preference towards WRC (W=A/T, R=A/G) hotspot motifs (Bransteitter et al. 2003, Pham et al. 2003, Bransteitter et al. 2004, Yu et al. 2004, Larijani et al. 2005, Larijani and Martin 2007, Larijani et al. 2007, Brar et al. 2008, MacCarthy et al. 2009). Although the target substrate has been determined, there are still many questions surrounding the biochemical regulation of AID, the mechanism by which it targets the Ig loci, as well as its off-target activity on oncogenes. In terms of biochemical activity, one point of contention is whether AID acts on ssDNA in a processive or distributive manner (Coker and Petersen-Mahrt 2007; Pham et al. 2003, 2007). If AID acts in a processive manner, it will heavily mutate only a few DNA targets in a pool (Pham et al. 2003). In contrast, if AID acts in a distributive manner, a small number of mutations will occur in nearly all targets in a pool (Coker and Petersen-Mahrt 2007). Although the biochemical nature of an enzyme gives clues as to how it will act, the nature of the substrate is also crucial in determining its mechanism *in vivo*. Properties of DNA including sequence (Carpenter et al. 2010; Shen et al 2005; Larijani et al. 2007), structure (Shen et al. 2005; Larijani et al. 2005b; Larijani and Martin 2007) and topology (Shen and Storb 2004) all influence AID targeting. Each feature alone is not sufficient to determine absolute activity,

and thus the combined role(s) of all features must be considered in the context of their *in vivo* environment.

As mentioned above, AID has a 3-8-fold higher preference for mutating cytidine within degenerate 5'-WRC hotspot motifs over 5'-SYC (S=G/C, Y=C/T) cold spots, which are both prevalent in the variable (V) and switch (S) region of the immunoglobulin locus (Bransteitter et al. 2003, Pham et al. 2003, Bransteitter et al. 2004, Larijani et al. 2005a, Larijani et al. 2007). WRC motifs are frequently found on the G-rich non-template strand in the form of the palindromic sequence AGCT (Yu et al. 2004). Although AID prefers WRC motifs, Larijani and colleagues have previously shown that purified AID also mutates neutral motifs at an approximately 1.8-fold higher rate than cold spots (Larijani et al. 2005a).

AID activity has been tested on numerous targets *in vitro*, including ssDNA, dsDNA, RNA, DNA-RNA hybrids, as well as DNA-DNA bubble structures (Abdouni et al. 2017; Bransteitter et al. 2003; Larijani et al. 2007; Larijani and Martin 2007). AID has been shown to be catalytically inactive on RNA or small dsDNA oligonucleotide substrates, but it has been shown to act on ssDNA as well as small DNA-DNA structures that form ssDNA bubbles. Furthermore, AID has shown around 3-fold greater preference for WRC hotspot motifs within DNA-DNA bubbles (Bransteitter et al. 2003) and seems to prefer small bubbles 5- and 7-nucleotides in size (Larijani et al. 2007, Larijani and Martin 2007). It was also noted that although AID has an extraordinarily high affinity towards ssDNA (1-10 nM), it prefers to act on bubble structures over pure ssDNA substrate (Larijani et al. 2007). AID has also been shown to act on the DNA strand of DNA-RNA

hybrids generated by T7 polymerase during *in vitro* transcription (Canugovi et al. 2009). However, R-loops were found unnecessary for AID activity as mutations were not eliminated after the addition of RNaseH. Furthermore, we have recently found that AID mutated a TGC motif within an DNA-RNA hybrid bubble of switch region sequence with a 4-fold preference over TGC within a DNA-DNA bubble of the same sequence (Abdouni et al. 2017). This preference was not observed when the TGC substrate was surrounded by a random sequence. The fact that AID activity is so dependent on structure is evident from the finding that a WRC motif placed in a non-optimal substrate (e.g. Stemloop) is mutated less efficiently than a non-WRC motif in an optimal structure (e.g. 5-7nt. bubble) (Larijani and Martin 2007). Altogether, the above data suggests that there is not one single feature, but rather many structural components working together, that influence AID binding to its target.

1.7 Strand Topology and its Relation to Transcription

Not only does transcription modify nucleosome arrangement and DNA structure, it is also capable of modifying DNA topology (Kouzine et al. 2013). During active transcription RNAP introduces approximately 7 supercoils per second into de-chromatinized DNA (Darzacq et al. 2007). A high degree of supercoiling generates a topology highly susceptible to local denaturation and melting (Vlijm et al. 2015) unless it is relieved by DNA topoisomerases (reviewed in Timsit 2012). It has also been shown that the degree of supercoiling increases with the rate of transcription, suggesting that the introduction of supercoils may exceed the relaxation capacity of topoisomerases in highly

transcribed genes (Kouzine et al 2008). More recently, Kouzine and colleagues used psoralen photo-binding to map transcription-induced supercoiled regions in Raji human Burkitt's lymphoma cells (Kouzine et al. 2013). They found that transcription-induced supercoiling spreads ~1.5-2 kb upstream of the transcription start site (TSS) of nearly all transcribed genes. Interestingly, during SHM AID-mediated mutations begin ~100-200 bp upstream of the V region promoter and span around 2 kb (Longerich et al. 2006; Storck et al. 2011). In an *in vitro* transcription assay, AID-mediated mutations were found to begin ~80 nt. downstream of the TSS, to peak ~200-500 nt. from the TSS, and then to decrease thereafter (Besmer et al. 2006). Thus, supercoiling may provide a functional link between transcription and hypermutation, with the position of the promoter crucial to determining the distribution of supercoiling throughout a given gene.

Even without transcription, supercoiling has been shown to also play a role in AID activity. It has been demonstrated using *in vitro* assays that relaxed double-stranded DNA is a poor substrate for AID (Bransteitter et al. 2003; Chaudhuri et al. 2003; Dickerson et al. 2003; Shen and Storb 2004), while supercoiled DNA can be mutated (Shen and Storb 2004). Shen and Storb used an *in vitro* gain of function assay to show AID activity in the absence of transcription. In their assay, the start codon of the Amp^r β-lactamase gene of a circular *E. coli* plasmid (pKM2) was changed from ATG to ACG preventing antibiotic resistance. If AID reverts the start codon back to its original form, antibiotic resistance will be restored. The authors found that upon treatment with AID, antibiotic resistance was restored when the supercoiled form of pKM2 was used but not with relaxed pKM2 that had been treated with Topoisomerase I. It was concluded that AID can target dsDNA *in vitro*

but the DNA must be supercoiled (Shen and Storb 2004). Furthermore, mutations were found on both DNA strands in the clones that regained antibiotic resistance showing that both DNA strands can be mutated without transcription. WRC hotspot motifs exist on both DNA strands of immunoglobulin genes (Rogozin et al. 2001), and therefore could be potential targets for AID. A subsequent experiment connected transcription and topology. By altering the supercoiled pKM2 plasmid to include a T7 promoter upstream of the Amp^r gene (Shen et al. 2005), Shen and Storb found that there were approximately 18.4-fold more C deaminations after *in vitro* transcription than when the plasmid was not transcribed. Thus, the current model for transcription requirement is that AID can mutate supercoiled dsDNA without transcription, but transcription markedly increases the accessibility of AID to plasmid DNA.

A more recent study looked further into the function of supercoiling in allowing AID access to genomic dsDNA of both mammalian and bacterial cells (Parsa et al. 2012). First, bisulfite was used to map ssDNA regions in both *ex vivo* mouse B cells and Ramos cells, which undergo constitutive SHM in culture. Bisulfite is a chemical deaminase with activity like that of AID in that it mutates dC to dU only within ssDNA. The authors found that the V region has the highest frequency of ssDNA than any other region in the Ig locus. Furthermore, ssDNA patches of approximately 7 nucleotides in length were found on both DNA strands of the V region as well as in a control GFP sequence, but occurred at a 3.5-fold higher rate in the V region than in the control. Next the authors looked at the prevalence of ssDNA patches in the Switch regions of murine B cells that were stimulated to undergo CSR by lipopolysaccharide (LPS). Like their findings in the V region, ssDNA

patches of a median length of 7 nucleotides were found. Due to previous findings that AID has increased deaminase activity on transiently single-stranded supercoiled DNA (Shen et al. 2004), it was hypothesized that negative supercoiling may generate the observed ssDNA patches and that these patches are targeted by AID (Parsa et al. 2012). Bisulfite was next used to map ssDNA patches in an Isopropyl-β-D-thiogalactopyranoside (IPTG)-inducible gene in Wildtype [BL21(DE3)] *E. coli* and a mutant strain lacking Topoisomerase I (TopoI) [VS111(DE3)]. Hyper-negative supercoiling was found in the VS111 strain, but not in BL21 controls, corresponding to the 3-fold greater ssDNA patch density in VS111. Hyper-negative supercoiling is defined as an extremely negative supercoiled topology that is induced in plasmids containing transcriptionally-active genes in the absence of Topoisomerase I (Brill and Sternglanz 1988, Drolet et al. 1994). As in the mammalian cells, Parsa and colleagues also found ssDNA patches on both strands of DNA in *E. coli* (Parsa et al. 2012). Lastly, the authors used an IPTG-inducible AID expression vector to observe whether the increased negative supercoiling of the TopoI mutant also increases the extent of deamination by AID. A 5.8 to 12.8-fold increase in AID-mediated mutations was observed in the VS111 strain when compared to *E. coli* BL21 and parental wildtype [MG1655(DE3)] strains, respectively. Taken together these results indicate that transcription-induced negative supercoiling leads to the production of ssDNA patches, and that these are substrates for AID-mediated mutation.

1.8 Project Rationale

Although ample progress has been made in elucidating the relationship between AID targeting and transcription, thus far, the relative importance of the process of transcription itself vs DNA structures (e.g. supercoiling) in determining AID targeting is unknown. Evidence is pointing towards the genomic architecture in influencing AID targeting, where the sequence of a gene influences the structures it forms making it more or less desirable as a substrate to AID. Furthermore, past research has shown that the topology of DNA influences the targeting and deaminase activity of AID (Shen and Storb 2004, Shen et al. 2005), and that AID can mutate supercoiled but not relaxed dsDNA (Shen and Storb 2004). Thus, we hypothesized that features of DNA inherent to a gene are important determinants of AID recruitment, with DNA topology playing the most crucial role. All previous studies examining the role of transcription or DNA structure in AID targeting employed a gain of function model wherein AID would mutate an antibiotic resistance gene to regain bacterial resistance. It is possible that this assay generates biased results as AID must mutate one nucleotide on the entire plasmid in order to generate antibiotic resistance, and thus give measureable activity through colony counts. We were interested to examine whether the proposed model of transcription requirement for AID activity holds true in the absence of this selective pressure. To this end, we have developed a new assay where observation of AID-mediated deamination is not dependent on AID deaminating any particular dC residue, thus allowing us to visualize and study all AID-mediated mutations without bias. In this system purified AID, expressed in either a prokaryotic or eukaryotic purification system is incubated with a target DNA molecule of

either supercoiled, relaxed linear, or denatured topology, to observe mutation frequencies and patterns. If the results of Shen and Storb (2004) hold true, AID will mutate the supercoiled plasmid but not the linear DNA, and moreover, will do so equally on both strands.

Using our assay, we could also study AID activity on the target DNA sequence in the presence or absence of transcription. It is thought that AID targeting may be associated with progression of the elongation complex and/or transcription termination (Kodgire et al. 2013, reviewed in Chandra et al. 2015). Although RNAP pausing or stalling has been shown to contribute to AID binding, the impact of RNAP transcription dynamics on AID targeting has not been examined. We established an *in vitro* transcription assay where we can vary the rate of T7 polymerase progression by varying the concentration of rUTP or all rNTP nucleotides in the reaction. Decreasing the concentration of one or all rNTPs will temporarily stall T7 polymerase as it pauses to capture the next incoming ribonucleotide. We hypothesized that slowing down or speeding up transcription will vary the spatial and temporal window wherein AID can access transcription-induced secondary structures. Our goal was to determine whether there is an optimal rate of transcription for AID activity. In this manner, the overall goal of this project was to understand the role of transcription-induced as well as transcription-independent DNA topology in attracting AID to its target gene.

II. Materials and Methods

2.1 Preparation and Purification of the Supercoiled and Linear DNA Substrates

The target DNA sequence used to measure AID activity was obtained from Open Biosystems (Huntsville, AL). It is 1.2 kb in length, and has a G/C and A/T content of 49.6% and 50.4%, respectively. The insert was cloned into the mammalian expression vector pcDNA3.1D V5-His-TOPO (Figure 5), where eukaryotic transcription can be driven under the cytomegalovirus (CMV) promoter and prokaryotic transcription under the T7 promoter (Dang et al. 2006).

The recombinant plasmid was purified using a Maxiprep Plasmid Kit (Geneaid) according to manufacturer instructions. Although this kit is optimized for yielding supercoiled plasmid, the supercoiled fraction was further purified by cesium chloride density centrifugation. In short, 200 μg of purified plasmid DNA was re-suspended in 3.7 ml of TE buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA) followed by addition of 3.7 g of CsCl and 0.37 ml of 1 mg/ml ethidium bromide (Sigma). The solution was then transferred to an ultracentrifuge tube and the density was adjusted to 1.55 g/ml with Cesium Chloride solution. The solution was then centrifuged at 55,000 RPM in a Ti65 vertical rotor, no brake, 20°C, overnight. The supercoiled and nicked bands were visualized by UV and extracted via syringe, purified using a standard ethanol precipitation protocol, and visualized on a 0.7% agarose gel to verify purity and topology. The supercoiled DNA was further purified by gel extraction from a 0.5% agarose gel using the QIAquick Gel Extraction Kit (Qiagen). Nicked and supercoiled plasmids were aliquoted and stocks were frozen at -20°C.

Purified supercoiled DNA was linearized by treatment with SmaI (New England BioLabs) in Buffer #4 (50 mM Potassium Acetate, 20 mM Tris-acetate, 10 mM Magnesium Acetate, 1 mM DTT, pH 7.9; New England BioLabs) at 25℃ for 2.5 hours. The SmaI site is approximately 1.2 kb downstream of the target DNA sequence. Linear DNA was separated on a 0.5% agarose gel and purified using the QIAquick Gel Extraction Kit (Qiagen). Supercoiled and linear template DNA were visualized on a 0.7% agarose gel to verify the quality prior to experimental assays.

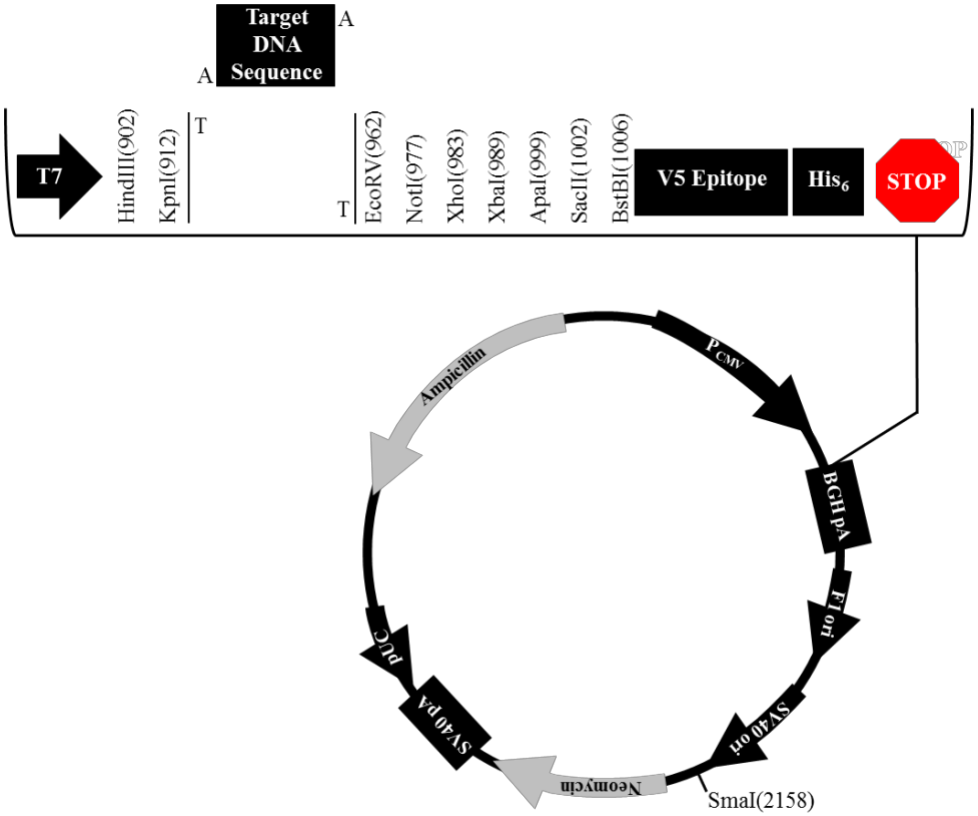


Figure 5: Expression Vector Backbone. The target DNA sequence was cloned into the pcDNA3.1 mammalian expression vector with a V5 epitope tag from SV5 paramyxovirus (bases 1020-1061) and a polyhistidine epitope at bases 1071-1088. It was inserted between KpnI(912) and EcoRV(962), as indicated above. The T7 promoter is located at position 863-882, upstream of the insert. The plasmid contains ampicillin and neomycin resistance genes. The SmaI site used for linearization of the plasmid is shown at position 2158, approximately 1.2 kb downstream of the target insert.

2.2 Topoisomerase I Assay to Verify Supercoiled DNA Topology

Although the supercoiled DNA was rigorously purified by both CsCl gradient centrifugation as well as gel extraction, we used human Topoisomerase I (Sigma) as another control to verify that the substrate is indeed supercoiled in topology. 250 ng of supercoiled substrate was incubated with ~1 unit of TopoI at 37℃ for the following times: 30 seconds, 1 minute, 5 minutes, 10 minutes, 15 minutes, 20 minutes, 25 minutes, and 30 minutes. TopoI was diluted prior to use by ½ of its stock concentration (~2 units/µl) in TopoI storage buffer (20 mM sodium dihydrogen phosphate pH 7.4, 300 mM NaCl, 50 µg/ml BSA, 50% glycerol, 50 mM imidazole). Reactions were incubated in a final concentration of 1X TopoI reaction buffer (10 mM Tris-HCl pH 7.9, 1 mM EDTA, 15 mM NaCl, 0.1% BSA, 0.1 mM Spermidine, 5% glycerol) with a final volume of 20 µl. As a control to show TopoI only acts on DNA if it is supercoiled in topology, 250 ng of relaxed linear substrate DNA was incubated with TopoI for 30 minutes alongside of the supercoiled conditions. As a negative control, TopoI was replaced with 1 µl of its storage buffer. The reactions were stopped with 2 µl of 10% SDS ($1/10^{th}$ of the reaction volume) at room temperature. DNA was purified by ethanol precipitation and re-suspended in 12 μl of TE buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA). The entire reaction was loaded and analyzed on a 1% agarose gel in 1X TAE (Tris-acetate-EDTA) buffer. The gel was electrophoresed at 150V for 3 hours, and then post-stained with 0.0025% ethidium bromide (Sigma) in 1X TAE buffer for 30 minutes followed by a wash in 1X TAE buffer for 10 minutes. Gel was imaged using a Gel logic 200 imaging system (Mandel Scientific Company Inc.) and Kodak gel imaging software.

2.3 Southern Blot analysis to Verify Purity of Double-stranded Supercoiled and Linear DNA

Two DNA oligonucleotides complementary to the target DNA sequence were used as a probe for Southern blot: Rev1: 5'-TAC AAA CCA GGT GAT CTG GAA GCG CC-3' and Rev2: 5'-AAT CTC CTT TAG CGT GCG GTG CAG GG-3'. Oligonucleotides were chosen from 5 potential oligonucleotide probes based on the least number of base pairs able to bind to each other and form hetero-dimers. 12 pmol of each oligonucleotide, and 1 μg of GeneRuler 1 kb DNA ladder (Thermo Scientific), were 5'-labeled with [γ-^{32}P] dATP with polynucleotide kinase (New England Biolabs). Following labelling, the oligonucleotides and ladder were purified through mini-Quick spin DNA columns (Roche).

100 ng of the supercoiled and linear plasmids were incubated with AID, or in AID dialysis buffer in place of AID, under native or heat-denaturing conditions (see section 2.5) to determine if there is ssDNA contaminating our prep of dsDNA. DNA was run on a 1% agarose gel in 0.5X Tris/Borate/EDTA (TBE) buffer. Both supercoiled and linear plasmid DNA were heat-denatured at 98°C for 10 minutes to form ssDNA, which was used as a size marker upon analysis via agarose gel electrophoresis. Following electrophoresis, the gel was soaked in denaturing buffer (1.5M sodium chloride, 0.5M sodium hydroxide), and the DNA was transferred to a Amersham Hybond-XL nylon membrane (GE Healthcare). The membrane was neutralized in 0.5X TBE, then cross-linked using a Spectroline Microprocessor-controlled UV Crosslinker (Fisher Scientific) for 30 s at 120 mJ/cm^2, then turned 180° and crosslinked again. The blots were probed with the labelled oligonucleotides, and washed 2 times each at 55°C with 3 different wash buffers: wash 1: 2X saline-sodium citrate (SSC), wash 2: 0.5X SSC and 0.5% sodium dodecyl sulfate

(SDS), wash 3: 0.1X SSC and 0.1% SDS. Blots were exposed to a Kodak Storage Phosphor Screen GP (Bio-Rad, Hercules, CA, USA), and visualized using a PhosphorImager (Bio-Rad, Hercules, CA, USA) and Quantity One software (Bio-Rad, Hercules, CA, USA). All Southern blot band intensities were quantified using ImageLab software (Bio-Rad).

2.4 Expression and Purification of Wildtype Human AID

Two types of purified AID were used in this study, bacterially expressed GST-AID and eukaryotically expressed AID-His. We have previously described the bacterial system for expression and purification of GST-AID (Larijani et al. 2007, Dancyger et al. 2012, Abdouni et al. 2013). In brief, *E. coli* BL21(DE3) bacteria containing the recombinant expression vector pGEX5.3 (GE Healthcare, USA) with GST-AID were induced to express AID by adding 1 mM IPTG and incubating at 16°C for 16 hours. Cells were lysed using the French pressure cell press (Thermospectronic) and then purified using Glutathione Sepharose high performance beads (Amersham). GST-AID was stored in 20 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM dithiothreitol. For eukaryotic expression, an EcoRV/KpnI fragment containing human AID-V5-His was cloned into pcDNA3.1-V5-6xHis-Topo modified to encode 2 extra His-residues in the C-terminus of the expression protein. The expression vector was transfected into HEK 293T cells. 5 x 10^5 cells were seeded on 50 x 10 cm plates, and transfected with 5µg of plasmid using Polyjet transfection reagent (Froggabio). Plates were incubated 48 hours at 37°C. Following incubation, cells were re-suspended in lysis buffer (50 mM Phosphate Buffer pH 8.2 + 500 mM NaCl, 0.2 mM PMSF, 50 µg/ml RNase A), lysed using a French pressure cell press (Thermospectronic),

and then batch bound to Nickel Sepharose beads (Amersham). Beads were serially washed with lysis buffer containing 1 mM or 30 mM Imidazole, and eluted in lysis buffer containing 500 mM Imidazole. Both preparations of AID were analyzed for purity and yield by SDS-PAGE electrophoresis followed by Coomassie Brilliant Blue staining. Western blot analysis was used to verify the relative yield and purity of eukaryotically-expressed AID. Western blots were probed with anti-V5 (Abcam) antibody, followed by secondary detection using goat anti-rabbit IgG (SantaCruz). As a final check for the quality of the purified AID, enzymatic activity was verified using an alkaline cleavage deamination assay, described below.

2.5 Alkaline Cleavage Deamination Assay

The standard enzyme assay to measure the cytidine deaminase activity of purified AID is the alkaline cleavage assay (Quinlan et al. 2017, King et al. 2015, Abdouni et al. 2013, Larijani and Martin 2007, Larijani et al. 2007). Briefly, partially single-stranded substrates containing a 7 nucleotide-long single-stranded bubble region with the WRC hotspot TGC was used as a substrate. To generate this substrate, 2.5 pmol of the target strand (the strand containing the WRC motif TGC) was 5'-labeled with [γ-^{32}P] dATP using polynucleotide kinase (New England Biolabs, USA), purified through a mini-Quick spin DNA column (Roche, Indianapolis, IN, USA), and annealed with 3-fold excess (7.5 pmol) of the complementary strand, to generate the bubble structure. Approximately 1.2 μg of GST-AID was added to 25 fmol of radio-labeled substrate in 100 mM phosphate buffer (pH 7.21), to a final volume of 10 μl per reaction. All reactions were carried out in

duplicate. The no AID control consisted of 4 μl of AID dialysis buffer (20 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM DTT), 25 fmol radiolabelled substrate in 100 mM phosphate buffer (pH 7.21) to a final volume of 10 μl. The reactions were incubated in an Eppendorf Mastercycler PCR thermal cycler (Fisher Scientific, Ontario, Canada) at 32°C for 3 hours. 32°C is the optimal temperature of Wildtype Human AID (Quinlan et al. 2017). Following incubation, the samples were heat-inactivated at 85°C for 20 minutes to deactivate AID. Next, the volume of each reaction was increased to 20 μl by the addition of 2 μl 10X Uracil-DNA Glycosylase (UDG) buffer (200 mM Tris-HCl pH 8.0, 10 mM DTT, 10 mM EDTA), 0.2 μl UDG enzyme (New England Biolabs), and 7.8 μl autoclaved milliQ dH_2O. The reactions were then incubated at 37°C for 30 minutes. Next, 2 μl of 2M NaOH was added to each reaction, which was then further incubated at 96°C for 10 minutes. In this step, the alkali-labile abasic site is cleaved, generating a nick in the target strand of the bubble substrate and denaturing the three strands. After the incubation, 10μl of formamide-loading dye solution (95% formamide, 0.25% Bromophenol Blue) was added to each reaction.

To ensure denaturation prior to electrophoresis, samples were heated at 98°C for 3 minutes and 15 μl of each reaction was loaded onto a pre-run 14% denaturing urea-formamide-acrylamide gel (1X TBE, 25% formamide, 14% acrylamide:bisacrylamide, 7M urea). The gel was electrophoresed at 300V for 3 hours in TE buffer, then exposed to a Kodak Storage Phosphor Screen GP (Carestream Health Inc., Rochester, NY, USA) overnight. The phosphor screen was imaged using a PhosphorImager (Bio-Rad), and the gel was quantified using Image Lab (Bio-Rad).

2.6 Transcription-independent AID Activity (TIAA) Assay

AID activity was tested on 3 versions of the target plasmid: heat-denatured (ssDNA), relaxed linear, or supercoiled. Approximately 1.3 μg of GST-AID or 30 ng AID-His was incubated with 100 ng of plasmid substrate in 100 mM phosphate buffer (pH 7.21) and approximately 4 x 10^{-4} units of Uracil-DNA Glycosylase inhibitor (UGI) (1/2000th of stock concentration; New England BioLabs), in a final volume of 10 μl per reaction. The no AID controls consisted of 100 ng of the supercoiled plasmid in 100 mM phosphate buffer (pH 7.21), AID dialysis buffer (20 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM DTT) and 4 x 10^{-4} units of UGI (1/2000th of stock concentration; New England BioLabs), in a final volume of 10 μl per reaction. All reactions were carried out in triplicate. For the heat-denatured (ssDNA) conditions, the supercoiled or relaxed linear DNA was heat-denatured at 98°C for 10 minutes, snap-cooled on slush ice for 1 minute, then AID and UGI were added as described above. The samples were incubated 32°C for 4 hours in an Eppendorf Mastercycler Epigradient PCR thermal cycler (Fisher Scientific, Ontario, Canada). The reactions were stored at -20°C.

2.7 Bisulfite Deamination Assay

Standard bisulfite ssDNA mapping was used to foot-print the single-stranded regions induced by breathing within the double-stranded supercoiled and relaxed linear substrates. As a positive control for bisulfite, which can only deaminate C to U in ssDNA, the assay was carried out using heat-denatured plasmid substrate. For the positive control, 2 μg of supercoiled plasmid in a volume of 54 μl with 2 mM Tris-Cl (pH 7.5) was denatured

by adding 6 μl of 3M NaOH to a final volume of 60 μl and incubated at 37°C for 15 minutes. 430 μl of a 3.6M sodium bisulfite (pH 5.0)/0.5 mM Hydroquinone solution was added and incubated in a thermocycler for 2 cycles of 95°C for 4 minutes and 55°C for 4 hours, followed by 95°C for 4 minutes and 55°C for 2 hours. Under native conditions, 430 μl of a 3.6M sodium bisulfite (pH 5.0)/0.5 mM Hydroquinone solution was added to 2 μg of relaxed linear or supercoiled plasmid substrate to a final volume of 490 μl and incubated at 37°C for 4 hours. All treated DNA was purified using a QIAquick PCR purification kit (Qiagen) as per manufacturer's instructions, and eluted in 54 μl of 10 mM Tris (pH 8.5). Next, 6 μl of 3M NaOH was added and incubated at 37°C for 15 minutes. DNA was purified by ethanol precipitation and re-suspended in 50 μl of TE buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA).

2.8 Transcription-associated AID Activity (TAAA) Assay

The target DNA sequence described above can be transcribed by the T7 promoter (Figure 5). The plasmid was transcribed *in vitro* using the MEGAscript T7 Transcription Kit (Ambion). 100 ng of supercoiled plasmid was transcribed in 1X transcription buffer, with 1 µl of T7 polymerase, varying concentrations of rNTPs (3.75 mM or 0.375 mM) or rUTP (3.75 mM, 0.375 mM, 0.075 mM, 0.0375 mM, 0.1875 mM, or 0.009375 mM), 1 µg of RNase A (Sigma), 1 µl of UGI diluted to 1/2000th of stock concentration (New England BioLabs), and 1.3 μg of GST-AID in a final volume of 20 µl. The no AID control contained 4 µl of AID dialysis buffer (20 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM DTT) instead of AID, and the no transcription control contained 1 µl of ultrapure H_2O in place of T7

polymerase. Reactions were incubated at 32°C for 4 hours. The reactions were stored at -20°C.

2.9 Deamination-specific PCR

To detect AID-mediated mutations, 1µl of either the AID only or cell-free transcription reactions was amplified by deamination-specific nested PCR (deam-PCR) using Taq DNA polymerase and mutation-specific primers: Sense fwd 4: 5'-GGG ATA TAG GGG TTT TTT GAG GTT TGG TAT TAT TTA AAT-3', Sense fwd 5: 5'-TTT ATT TTG GTT TTG TGG TAA TTG ATT GTT TGT TAA TAG G-3', Sense rev 2: 5'- ACA CAA CCA ACT TTC ATT CCA ACC ACA AAC TTT CAA TA-3', Sense rev 3: 5'-CCA ACT TTC ATT CCA ACC ACA AAC TTT CAA TAA ATT-3'. Each reaction contained: 10 µM of the forward and reverse primers (Integrated DNA Technologies Inc.), 10 mM dNTP (Invitrogen), 1 µl of either the AID only or AID with cell-free transcription reactions, Taq DNA polymerase and 10X PCR buffer (100 mM Tris pH 8.3, 500 mM KCl, 15 mM $MgCl_2$). The nested PCR reactions were 37 cycles of 94°C for 30 seconds, 58°C for 30 seconds, and 72°C for 1 minute with the A3DE sense fwd 4 and rev 3 primers, followed by a second round of PCR using 1µl of sample from the first round with primers sense fwd 5 and rev 2 and an annealing temperature of 57°C. The PCR products were then analysed on an 1% agarose gel.

To determine the relative amount of highly mutated DNA in each reaction, the AID only or AID with cell-free transcription reactions were diluted prior to deam-PCR. 1.3 µg of GST-AID or 30 ng of AID-His were mixed with 4 x 10^{-4} units of UGI (1/2000th of stock

concentration; New England BioLabs) and incubated with supercoiled or relaxed linear substrates in the following amounts: 100 ng, 50 ng, 20 ng, 2 ng, 0.2 ng, 0.02 ng and 0.002 ng. The reactions were incubated at 32℃ for 4 hours. The reactions were then serially diluted in ultrapure H_2O. For example, the dilutions for the 100 ng reaction were as follows: 1, 1/2, 1/5, 1/50, 1/500, 1/5,000, 1/50,000 and 1/500,000. As the amount of DNA in the reaction decreased, the dilutions were stopped at one dilution before the last (e.g. the last dilution for the 50 ng reaction was 1/50,000, for the 20 ng reaction 1/5,000, etc.) because the amount of DNA at that dilution is already well under the detection limit of an agarose gel. 1 µl of each dilution was added to the deamination-specific PCR reaction, and the PCR was carried out as described above. The PCR products were analyzed on a 1% agarose gel.

2.10 Degenerate PCR, Purification and Analysis of Substrate DNA

To study bisulfite and AID activity, we used degenerate primers, which bind to wild-type as well as mutated sequences with equal preference, to amplify the target region of the substrate DNA from either the AID alone or AID with cell-free transcription reactions. The primers used for the PCR are degenerate+GG reverse (5'-GGT TTT ATT TTY ATT YTA TTY ATT YA-3') and degenerate+GG forward (5'-GGA TTT YAT TTY ATT YTT ATT YTT TTA-3'), where Y= C/T. Each reaction contained: 10 µM of the forward and reverse primers (Integrated DNA Technologies Inc.), 10 mM dNTP (Invitrogen), 1 µl of either the AID alone or AID with cell-free transcription reactions, Taq DNA polymerase and 10X PCR buffer (100 mM Tris pH 8.3, 500 mM KCl, 15 mM $MgCl_2$). The PCR reactions were incubated in an Eppendorf Mastercycler Nexus PCR

thermal cycler (Fisher Scientific, Ontario, Canada) for 35 cycles of 94°C for 30 seconds, 38°C for 30 seconds, and 72°C for 1 minute. The 1.2 kb-long PCR products were analyzed on a 0.8% agarose gel, imaged using a Gel logic 200 imaging system (Mandel Scientific Company Inc.) and Kodak gel imaging software, then subsequently cloned for sequencing analysis. A Topo TA cloning kit (Invitrogen, California, USA) was used to clone the fresh PCR product into the supplied vector for transformation. 4 μl of each fresh PCR product was added to 1μl of 1.2M NaCl and 0.06M $MgCl_2$ salt before being cloned into 1μl of the pCR 2.1-Topo vector (Invitrogen). The reactions were then incubated at 22.5 °C for 35 minutes in the PCR thermal cycler. Following TA cloning, the vectors were transformed into One Shot TOP10 Chemically Competent *E. coli.* (Invitrogen). The total volume of each reaction (6 μl) was added to a tube of Top10 cells (50 μl), for a total volume of 56 μl. The transformed bacteria were then incubated on ice for 30 minutes, heat-shocked in a 42°C water bath (Fisher Scientific, OH, USA) for 1 minute, and incubated on ice for 2 minutes. 250 μl of S.O.C. growth media (Invitrogen) was added, and then the bacteria were incubated in a 37°C shaker (Orbital Shaker, ThermoScientific, IL, USA), 225 rpm, for 1 hour. The cells were plated on LB agar plates (Miller, Sigma-Aldrich Co.) containing 40 μl of 20 mg/ml X-gal and 50 μg/ml kanamycin. Following incubation, 76.5 μl of each reaction was plated on 4 plates, which were incubated at 37°C overnight.

The next day a colony check PCR was performed using the degenerate primers described above, to ensure that the white-colored colonies contained the 1.2 kb insert. The white recombinant colonies were picked using a pipette tip and added directly into a PCR reaction mix of: 10X PCR buffer, 10 mM dNTP, 10 μM of the forward and reverse

degenerate primers specific for the insert, and Taq DNA polymerase. The pipette tip from each PCR reaction was streaked onto another LB agar plate containing 50 μg/ml kanamycin and incubated at 37°C overnight. The reactions were amplified in a PCR thermal cycler for 35 cycles at 96°C for 30 seconds, 38°C for 30 seconds, and 72°C for 1 minute. The PCR products were verified on a 0.8% agarose gel. Colonies containing 1.2 kb PCR amplicon inserts were then used to make starter cultures for mini-preparation of the plasmid. Colonies were grown in 5ml of Luria broth (LB; Fisher) containing 100 μg/ml ampicillin at 37°C, 225 rpm, overnight. The cultures were pelleted in an IEC Centra-8R centrifuge (International Equipment Company, USA) for 15 min, 3500 rpm, at 4°C. The pCR2.1-Topo plasmid containing the 1.2 kb insert was purified using a Geneaid high-speed plasmid mini kit. After elution, the concentration of DNA was determined using a ThermoScientific Nanodrop 2000 Spectrophotometer. 2-3μg of purified DNA from each clone was digested with EcoRI to perform a second check that the target insert is present. 2μg of DNA from each positive clone was added to a 96-well plate, lyophilized at 65°C for 20 minutes, and sent to Macrogen (Seoul, Korea) for Sanger sequencing. 50-100 amplicons from each reaction were sequenced.

All sequenced data was analyzed using Seqman analysis software (DNASTAR). The sequencing chromatogram raw data was checked visually in the Seqman analysis software to ensure there was a clear peak for each mutated residue, leaving no doubt that each mutation was bona fide. If the chromatogram was not of high quality (i.e. no clear peaks) the sample was discarded. Low quality sequences (i.e. no clear peaks on the chromatogram) were also excluded. Furthermore, only independent amplicons were

considered and duplicate sequences (although uncommon, <1%) were considered only once. Once the sequences of all amplicons were checked, further analysis of mutation frequency and placement was conducted using Microsoft Excel and GraphPad Prism 6.

2.11 Predicting Template DNA Secondary Structure using mfold

The secondary sequence for the top strand of our target DNA substrate was modelled using the online-based DNA-folding software 'mfold' (Zuker 2003). The secondary structure was modelled at 37℃ and in 100 mM salt conditions. The window size was set at 25 nucleotides and the folding was limited such that only bases within 50 nucleotides of each other could pair. This was chosen based on the recommendations of the 'mfold' software for DNA templates greater than 1999 nt. in length (Zuker 2003). The p-num file generated by mfold provided a detailed output on the paired or unpaired nature of each base in the computed folding.

III. Results

3.1 Designing an Unbiased Assay to Examine AID Targeting and Activity

Although one study reported targeting of supercoiled DNA in the absence of transcription (Shen and Storb 2004), four other studies have demonstrated that AID mutation rates are substantially (4 to 100-fold) higher when the plasmid DNA substrate was being transcribed (Shen et al. 2005, Besmer et al. 2006, Canugovi et al. 2009, Shen et al. 2009). These latter results have been interpreted as being consistent with a role for transcription in attracting AID activity *in vivo*. Although significant insights into the role of transcription in regulating AID activity have been obtained, the assays employed to measure AID activity have two major limitations. The first limitation is that these studies relied on a well-established antibiotic resistance assay, which is highly selective in that it selects for mutations on a single specific codon that reverts the antibiotic (typically ampicillin or kanamycin) resistance. Thus, although AID is presumably acting at many positions on either DNA strand, this assay only allows for sequencing of AID-mutated targets in which one specific position on one of the two strands has been mutated. The second limitation of these studies was that some examined AID targeting in the presence or absence of transcription, whilst others examined the role of DNA topologies on AID targeting in the absence of transcription; however, the role of topologies and transcription together has not been examined to date. Given that AID can sensitively recognize specific ssDNA topologies, and that its activity is highly sensitive to transcription, we sought to establish a model system in which the combined effect of transcription and DNA topologies on AID targeting could be studied. We designed this model system with the goal of

addressing the aforementioned shortcomings of the previously-employed antibiotic resistance screen: first, that it ought to be capable of measuring AID-mediated mutations in the presence or absence of transcription on a wide range of DNA substrate topologies including denatured ssDNA, relaxed linear and supercoiled DNA. Second, unlike the antibiotic resistance screens, AID-mediated mutations ought to have no functional consequence for detection. This would ensure that there is no bias for measurement of AID-mediated mutations, in regard to either strand or the positional context of the mutation.

We chose the pcDNA3.1D V5-His-TOPO plasmid (Figure 5, pg. 28) to carry our target insert predominantly because of the close proximity of the T7 promoter, which is located <60 nt. upstream of our target sequence, and which is important for examining the effect of transcription on AID targeting in our transcription-associated AID activity (TAAA) assay. Since AID mutations start ~80 nt. downstream of the TSS during *in vitro* transcription with peak occurrence of mutations ~200-500 nt. downstream (Besmer et al. 2006), a promoter location ~60 nt. upstream of the insert means that we should catch the peak of mutations in our target DNA sequence. Furthermore, the plasmid has numerous restriction sites that we could utilize for restriction digests and/or altering the plasmid substrate. Our target sequence was inserted between KpnI(912) and EcoRV(962). The KpnI restriction site is approximately 30 nt. downstream of the promoter. We also utilized SmaI to generate the relaxed linear substrate because the SmaI(2158) restriction site is ~1200 bp downstream of the end of our target insert. It should be sufficiently far away to avoid excessive AID processing of the ends of the target insert due to increased breathing rates of the ends of the plasmid.

Our target insert was chosen based on its length of 1.2 kb and its near equal G/C (49.6%) and A/T (50.4%) content. A target sequence of greater than 1 kb was necessary for viewing accurate AID activity in our cell-free transcription assay since, as mentioned earlier, AID begins deaminating ~80 nt. into its target gene, with peak activity ~200-500 nt. from the TSS (Besmer et al. 2006). Moreover, during SHM *in vivo* AID-mediated mutations begin ~100-200 bp upstream of the V region promoter and span around 2 kb (Longerich et al. 2006; Storck et al. 2011). Since our target DNA sequence is under the control of the T7 promoter (Figure 5, pg. 28), we wanted to ensure there was sufficient length from the promoter to "catch" the AID-mediated mutations and thereby gain an accurate view of AID targeting during transcription by T7 polymerase *in vitro*. It is also important that there is a near equal G/C, A/T content to avoid bias in AID activity due to an unequal sequence distribution. Furthermore, the target sequence has 76 WRC hotspots, 68 SYC cold spots, and 159 neutral trinucleotide motifs with C at the 3' end. Having a near equal number of hot and cold spots avoids biasing AID "preference" towards one motif or another. Moreover, since AID has been described as catalytically lethargic (Larijani and Martin 2007, Pham et al. 2011, Mak et al. 2013, King et al. 2015, Mak et al. 2015, King and Larijani 2017), we also wanted to make sure there were plenty of dC's within the target region.

To allow intricate study of AID activity, we can amplify our targeted substrate by either non-specific degenerate PCR (degen-PCR) or deamination-specific PCR (deam-PCR) (Figure 6). The degen primers contain a Y (Y=C/T) in the position of dC in both the forward and reverse primers, which bind to dG in the ends of our target DNA sequence

(Figure 7a). During the incubation with AID, we expect that some DNA sequences will be highly mutated, others may contain a few mutations and others will not have been mutated. Furthermore, DNA may be mutated on either strand, both strands, or neither strand. The degen primers are A-T-rich and thus do not depend on nor are inhibited from binding due to AID-mediated C-T mutation in the primer site. However, they can amplify mutated DNA if the mutations occurred on the opposite strand (Figure 7a). For example, if AID mutates dC to dU on the top strand Taq will pair the dU with dA so the G's in the original bottom strand primer site now become A's. Since there is a Y in the position that now has A, it can still bind and thus amplify both mutated and wildtype DNA. Since not all Y's in a given primer will be all C or all T, a relaxed annealing temperature of 38℃ is used to allow binding even if there are a couple of mismatches. Using our degen primers, we can amplify a combination of wildtype and mutated DNA independent of whether the DNA was mutated, in what position of the target sequence the mutations are located, and which strand the mutations are on. Mutations on the top strand (nontranscribed strand) will appear as C-T mutations when aligned with the forward DNA sequence, while those on the bottom (transcribed) strand will appear as G-A mutations (Figure 7b). During the first round of PCR, if there are dC to dU mutations on either strand Taq polymerase will pair the dU with dA. In subsequent rounds of PCR, the dA will get paired with dT. Since we only see the result of the top strand upon sequencing, we will only observe the sequences outlined in green in Figure 7b. If we observe C-T mutations in the top strand it is indicative of AID activity on the nontranscribed strand. Alternatively, if there are G-A mutations on the top strand it is indicative of AID activity on the transcribed strand. In this manner, both C-T

and G-A mutations will not occur on the same strand. Thus, sequencing of degen-PCR products after incubation with AID in the TAAA or transcription-independent AID activity (TIAA) assay would reveal a pool of substrate sequences representative of the overall AID reaction. Following degen-PCR, the PCR products are then TA cloned and transformed into chemically competent *E. coli* (Figure 6). Next, each white recombinant colony is screened for the target insert and then 50-100 amplicons per reaction are sent for sequencing (Figure 6).

In parallel to degen-PCR, we can also selectively amplify highly mutated sequences using deam-PCR (Figure 6), which we have previously established (Larijani et al. 2005a/b, Larijani et al. 2007, Larijani and Martin 2007, Quinlan et al. 2017). The reverse primer contains dA in the position complementary to dC (Figure 7c). Therefore, when AID mutates the template DNA dC becomes dU, which binds dA in the reverse primer. Once Taq polymerase amplifies the mutated DNA, it pairs dU with dA allowing the forward primer to bind pairing dT with dA in the position that was originally dC (Figure 7c). In this way, only the nontranscribed or top strand is amplified and we therefore only observe C-T mutations when compared back to the original sequence upon sequencing analysis (Figure 6). Deam-PCR is a nested PCR with the second set of primers located slightly inward from the first set, and thereby only a few strands of heavily mutated DNA are required to obtain a PCR product (Figure 7c). We can also use deam-PCR in a semi-quantitative way to estimate the efficiency of AID on different DNA substrates in the TIAA or TAAA assays upon quantification of band intensity (Figure 6). However, we can also use deam-PCR in a nearly fully quantitative manner by serially diluting the AID-DNA reactions and altering

the PCR annealing temperature to make it more or less stringent for highly mutated DNA. Diluting the AID-DNA reactions allows us to determine targeting efficiency on various substrates by quantifying and comparing band intensities of the different conditions as the DNA is diluted. Varying the annealing temperature allows the PCR to be more or less stringent for highly mutated DNA as increasing the temperature (e.g. 65℃) will force primers to anneal only to DNA that is completely complementary (i.e. highly mutated), whereas decreasing the annealing temperature (e.g. 35℃) will make the PCR more relaxed and thus amplify a combination of wildtype, moderately mutated, and highly mutated DNA. Comparing the PCR bands at stringent temperatures along with the dilutions data will provide us with quantitative information on the efficiency of AID targeting on various substrates, as stated above. Moreover, since the deam-PCR only amplifies highly mutated DNA under stringent PCR conditions and the degen-PCR amplifies a combination of mutated and wildtype substrate, we can use both sets of primers to gain two different perspectives of the same test tube reaction.

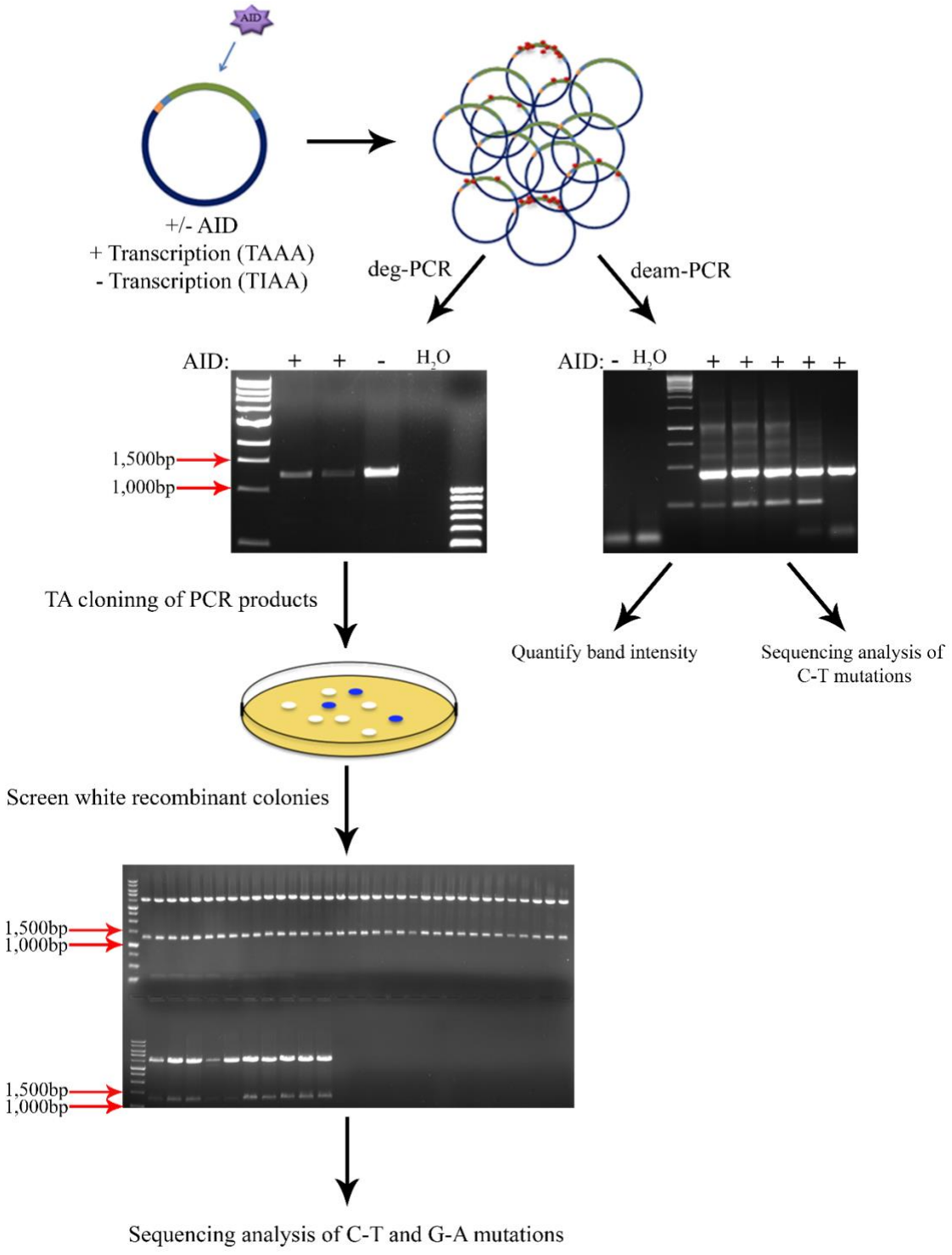


Figure 6: Assay Overview. DNA substrate is incubated with or without AID in either the TIAA or TAAA assay. After the incubation period we can use degen- or deam-PCR to

amplify the template DNA. Degen-PCR amplifies a combination of mutated and unmutated strands, allowing an unbiased view of the AID reaction. Deam-PCR is dependent on the presence of AID, and thereby only amplifies highly deaminated DNA strands. We can determine the preference of AID for targeting different substrates using deam-PCR by quantifying the band intensity or by sequencing analysis of the C-T mutations. Alternatively, we can use degen-PCR to obtain an overall view of the AID reaction, as well as gain strand preference information, by sequencing analysis of C-T and G-A mutations. Immediately following degen-PCR, the PCR products are TA cloned into a pCR2.1-TOPO vector and then transformed into TOP10 chemically competent *E. coli*. Each colony is screened for the target insert, and vectors containing the appropriate 1.2 kb insert are prepared for sequencing analysis. 50-100 amplicons are analyzed per condition.

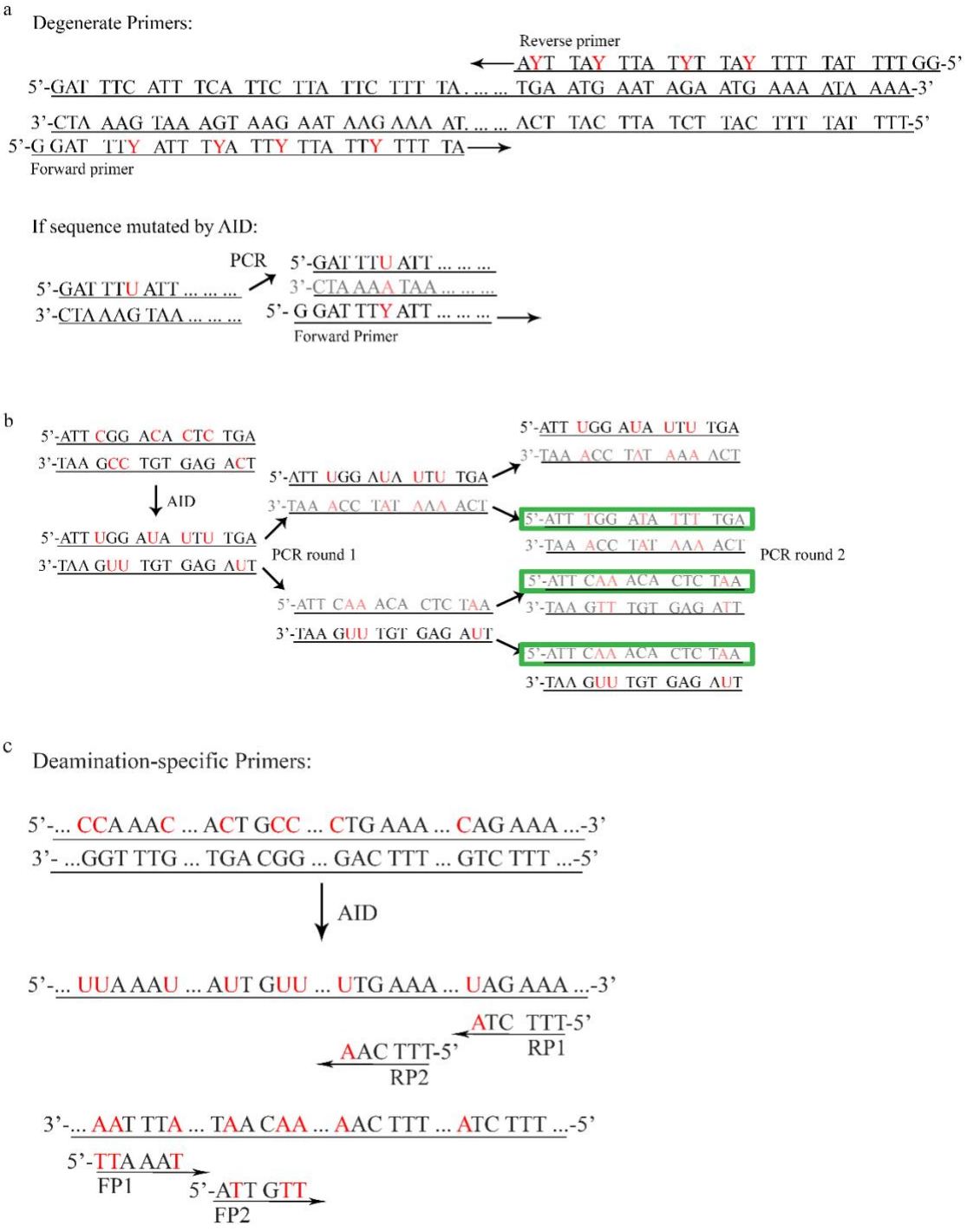


Figure 7: Detecting C-T and G-A Mutations Using Non-specific Degenerate Primers and Deam-specific Primers. a) The degen primers have a Y (Y=C/T) in place of C, allowing detection of a combination of wildtype, moderately mutated and heavily mutated target DNA at a relaxed annealing temperature of 38°C. In a wildtype sequence the Y binds

to G of the opposite strand, while if the strand opposite of the primer binding site is mutated by AID, once Taq has amplified the mutated strand and replaced dU with dA, the Y in the primer can also bind the mutated strand. The primers and their respective binding sites are A-T-rich, preventing inhibition of primer binding due to C-U mutations in the binding site. Thus, the degen primers can amplify all variations of mutated to wildtype DNA independently of where the mutations occurred in the sequence, and which strand the mutations occurred on. b) The degen-PCR also allows us to gain strand information as AID-mediated mutations originating from the top (nontranscribed) strand will appear as C-T upon sequencing and mutations originating from the bottom (transcribed) strand will appear as G-A. AID can mutate C to U in either the top or bottom DNA strand, which Taq polymerase will pair with A. On the next round of PCR, A gets paired with T. We can use this to distinguish mutations from either the top or bottom strand by only sequencing the top strand. In the above example, we would therefore only see mutations from the PCR products that are circled in green. When the PCR products with T mutations are compared back to the top strand, we see they are C-T mutations indicating AID deaminase activity on that strand. When the PCR products with A mutations are compared back to the top strand, we see they are G-A mutations, indicating that AID had acted upon the sister strand. c) Deam-PCR allows selective amplification of C-T mutations from only the top strand. In this case, AID must mutate C to U in the primer site in order to observe a PCR product. The reverse primer (RP1) must first bind and amplify the sequence as it contains dA in the position complementary to dC, which will become dU upon mutation by AID. Once Taq polymerase has amplified over all dU's and replaced them with dA's, the forward primer (FP1) can bind as it contains dT in the position complementary to dA in the primer site. Deam-PCR is a nested PCR with the second set of primers (RP2 and FP2) located slightly inward from the first set. As this is a nested PCR, only a few strands of heavily mutated DNA are required to obtain a PCR product. In this manner deam-PCR is biased, amplifying the top strand of only a few molecules exponentially.

3.2 Preparation of Supercoiled and Linear DNA Substrates

We chose to examine supercoiled and relaxed versions of our target DNA sequence because of the past report describing AID activity on supercoiled but not relaxed DNA (Shen and Storb 2004). Since we have now designed an assay that overcomes the limitations of antibiotic resistance assays, such as the one employed by Shen and Storb (2004), we were interested in further studying the influence of topology on AID targeting as well as determining the efficiency and pattern of AID-mediated mutation on either strand of the substrate. We tested AID activity on three forms of the same DNA sequence: heat-denatured, supercoiled, and linear (Figure 8). We expected that the supercoiled DNA will be targeted more efficiently than the linear DNA in our transcription-independent AID activity (TIAA) assay. We hypothesized that secondary structures such as bubbles that form during transient DNA breathing (Altan-Bonnet et al. 2003) will occur more frequently in supercoiled DNA, and thus will be more efficient at providing AID with a ssDNA substrate. Since AID is known to mutate ssDNA (Bransteitter et al. 2003, Pham et al. 2003, Dickerson et al. 2003, Larijani et al. 2005a, Larijani and Martin 2007, Larijani et al. 2007), we also heat-denatured the supercoiled and linear substrates to separate the sister DNA strands immediately prior to the incubation with AID.

Before the influence of substrate topology in our TIAA and TAAA assays was tested, we first had to generate our substrates. CsCl density centrifugation was used to separate the relaxed (nicked and linear) species from the supercoiled ones, and the fractions were then purified by ethanol precipitation (Figure 9). The supercoiled plasmid was further purified by gel extraction, then a portion was digested by SmaI to generate the relaxed

linear substrate. The relaxed linear substrate was gel purified to remove any residual undigested forms. The topologies were verified by agarose gel electrophoresis.

To verify that the topologies obtained were indeed supercoiled and relaxed linear, we incubated the substrates with TopoI (Figure 10). Eukaryotic TopoI relieves torsional strain from genomic DNA that is actively undergoing transcription by nicking, unwinding and rejoining only one strand of duplex DNA (Kim and Jinks-Robertson 2017). TopoI only relaxes supercoiled duplex DNA, and does not change the conformation of relaxed linear DNA. The supercoiled DNA was incubated with TopoI for the following timepoints to generate a gradient of relaxation: 30 seconds, 1 minute, 5 minutes, 10 minutes, 15 minutes, 20 minutes, 25 minutes, and 30 minutes. The relaxed linear substrate was incubated with TopoI for 30 minutes. To show that both the TopoI activity buffer and the TopoI storage buffer do not change the conformation of the DNA, supercoiled and relaxed linear templates were also incubated either in TopoI storage and activity buffers or in TE buffer. We found that TopoI relaxed the supercoiled substrate DNA, as indicated by the increase in higher banding with increased incubation time, and did not change the conformation of the relaxed linear substrate as we expected (Figure 10). The buffers also do not affect the conformation of the native supercoiled or relaxed linear DNA, as there was no difference between the conformation of the DNA in TopoI activity and storage buffers or in TE buffer. We thus confirmed that the topologies of the isolated supercoiled and relaxed linear DNA were in the correct conformation, and were useable to test AID activity in the TAAA and TIAA assays.

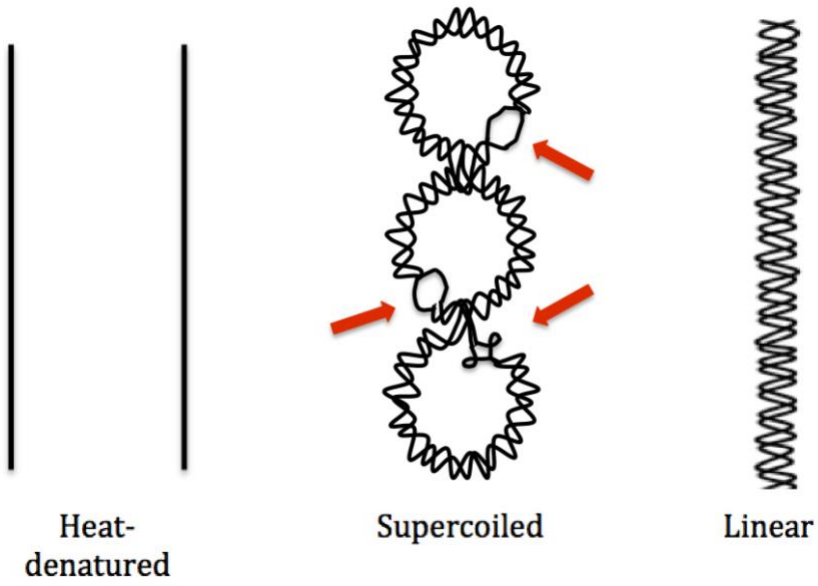


Figure 8: The Three DNA Topologies Analyzed in the Transcription-free AID Activity Assay. To determine the effect of DNA topology in AID targeting, AID activity was tested on three forms of the same DNA sequence. We expected that the supercoiled DNA will be targeted more efficiently than the linear based on literature suggesting that AID can target supercoiled, but not relaxed, plasmid DNA (Shen and Storb 2004), and because of the torsional strain induced by the topology of supercoiled DNA. We hypothesized that secondary structures such as bubbles that form during transient DNA breathing (Altan-Bonnet et al. 2003) will occur more frequently in supercoiled DNA (indicated by the arrows), and thus will be more efficient at providing AID with ssDNA substrate. Since AID is known to mutate ssDNA, the supercoiled and linear plasmids were heat-denatured at 98°C to separate the sister DNA strands immediately prior to the incubation with AID.

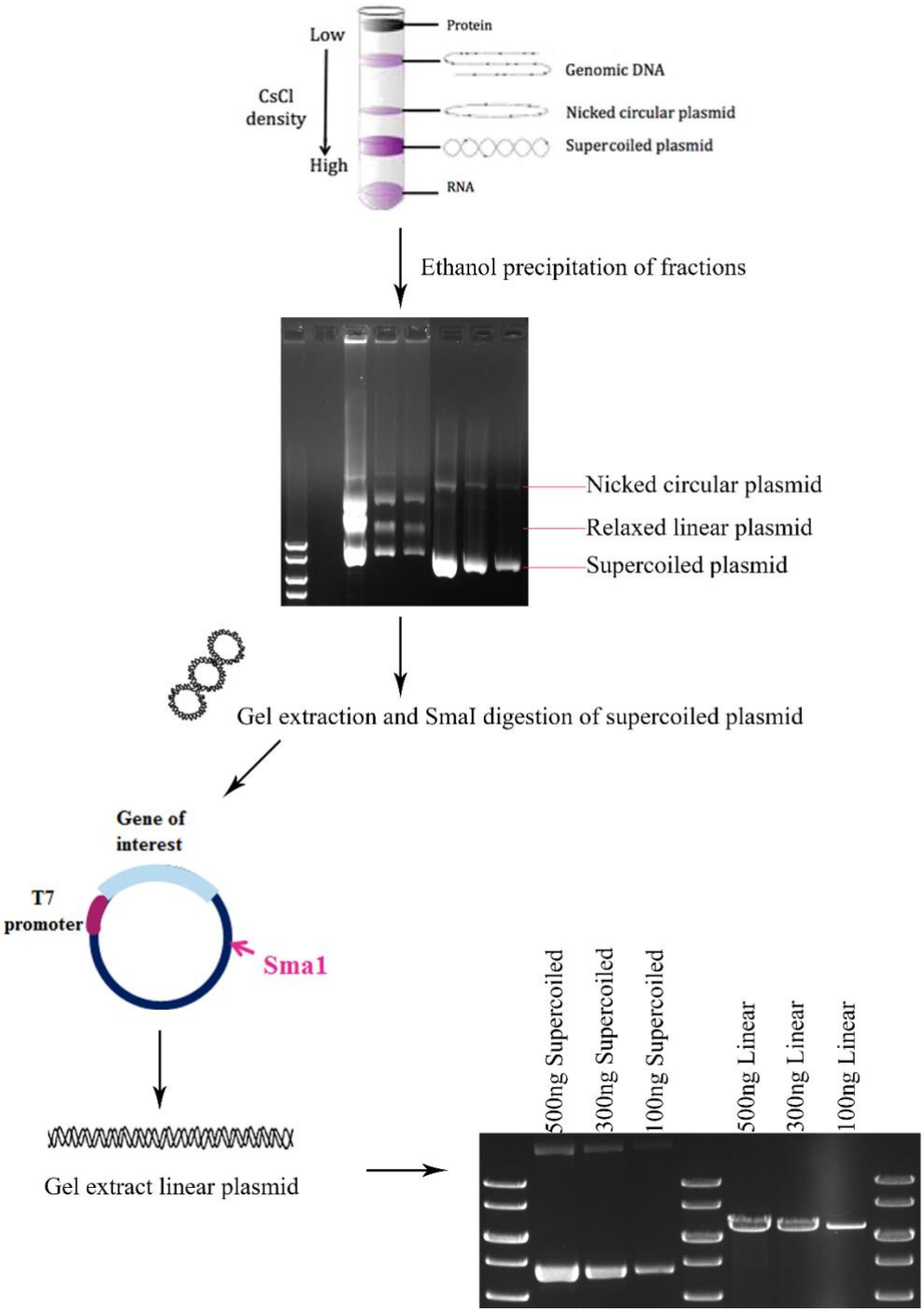


Figure 9: Preparation of Substrates. The topologies of the supercoiled and relaxed linear substrates were isolated and purified prior to the TAAA and TIAA assays. The supercoiled and relaxed fractions of the plasmid were isolated by CsCl ultracentrifugation and purified by ethanol precipitation. The supercoiled plasmid was gel extracted, then digested by SmaI

to generate the relaxed linear substrate. Following digestion, the linear band was gel purified to remove residual undigested plasmid. Topology of the purified substrates was verified by agarose gel electrophoresis.

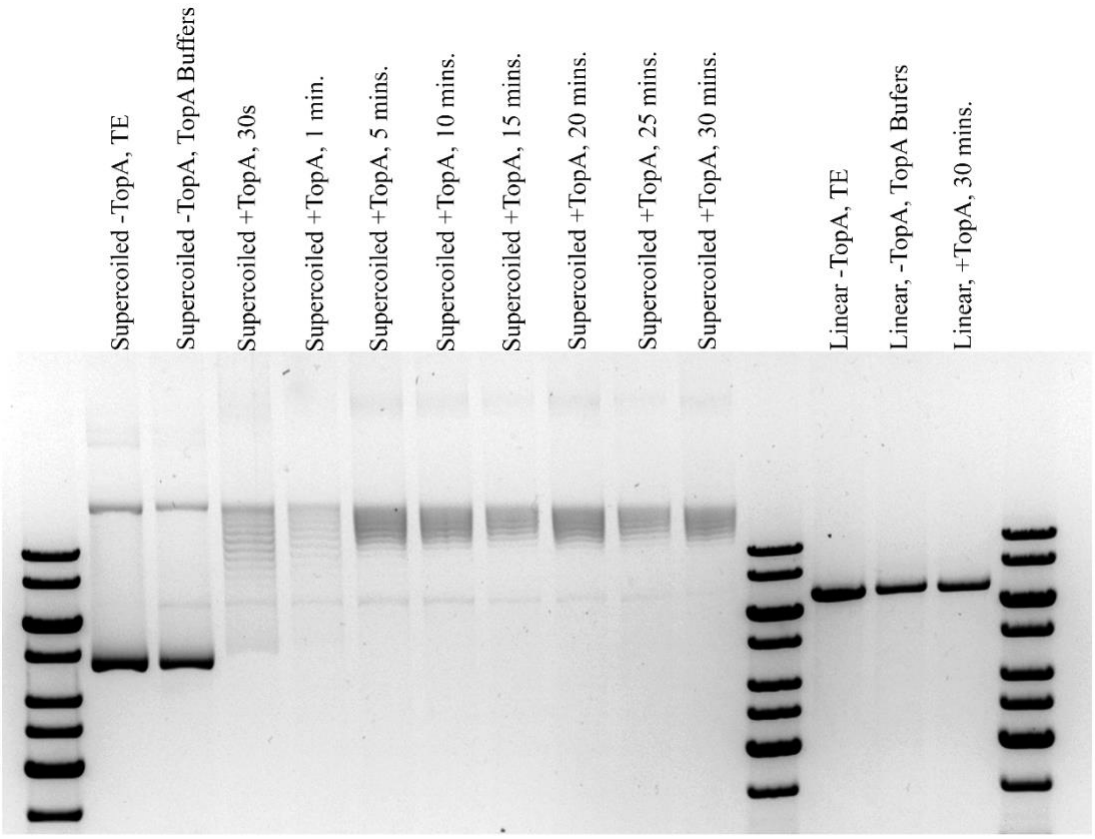


Figure 10: Verifying Supercoiled Substrate Topology Using TopoI. The topology of the purified supercoiled and relaxed linear substrates was verified by incubation with TopoI. TopoI can only relax DNA if it is supercoiled in topology. The supercoiled DNA was incubated with TopoI for the following timepoints: 30 seconds, 1 minute, 5 minutes, 10 minutes, 15 minutes, 20 minutes, 25 minutes, and 30 minutes. The relaxed linear substrate was incubated with TopoI for 30 minutes. Supercoiled and relaxed linear substrates diluted in TE buffer were used as controls to show the dsDNA in its native conformation. TopoI relaxed the supercoiled, as indicated by the increase in higher banding with increased incubation time, but did not change the conformation of the relaxed linear substrate. The agarose gel was color-inverted to clearly show the relaxed banding after treatment with TopoI.

3.3 AID can Mutate Relaxed Duplex DNA in the Absence of Transcription

Relaxed nicked, linear, supercoiled plasmid DNA (Figure 11a) and a PCR amplified fragment of our target sequence, were all tested in the TAAA assay. An RNA gel was run to verify that all of the above templates support transcription by T7 polymerase (Figure 11b). As a control for transcription, plasmids were also incubated with AID in the absence of T7 polymerase. Deam-PCR was used to check for highly mutated sequences as a consequence of AID activity on the above substrates (Figure 11c). Three independent PCRs were performed (1-3 underneath gel in Figure 11c), and it was observed that AID did indeed act consistently on supercoiled DNA with and without transcription, confirming the observation of Shen and Storb that AID can mutate supercoiled DNA in the absence of transcription (Shen and Storb 2004). AID had only minimal activity on the PCR fragment as only one faint band in the no transcription condition was observed (Figure 11c). To our surprise, AID consistently deaminated the relaxed linear and nicked circular forms of the plasmid both with and without transcription (as indicated by the red dots underneath the gel, Figure 11c), going against the grain of the past literature. Although the deam-PCR indicated that the activity was due to AID, since bands were not observed in the absence of the AID prep and the PCR water negative control was clean (Figure 11c), we sequenced the AID-supercoiled DNA +/- T7 polymerase reaction products to check for C-T mutations (Figure 11d). Indeed the sequences both with and without T7 RNA polymerase were highly mutated with all C-T mutations (Figure 11d), indicating that the deam-PCR result was indeed due to AID activity (Figure 11c).

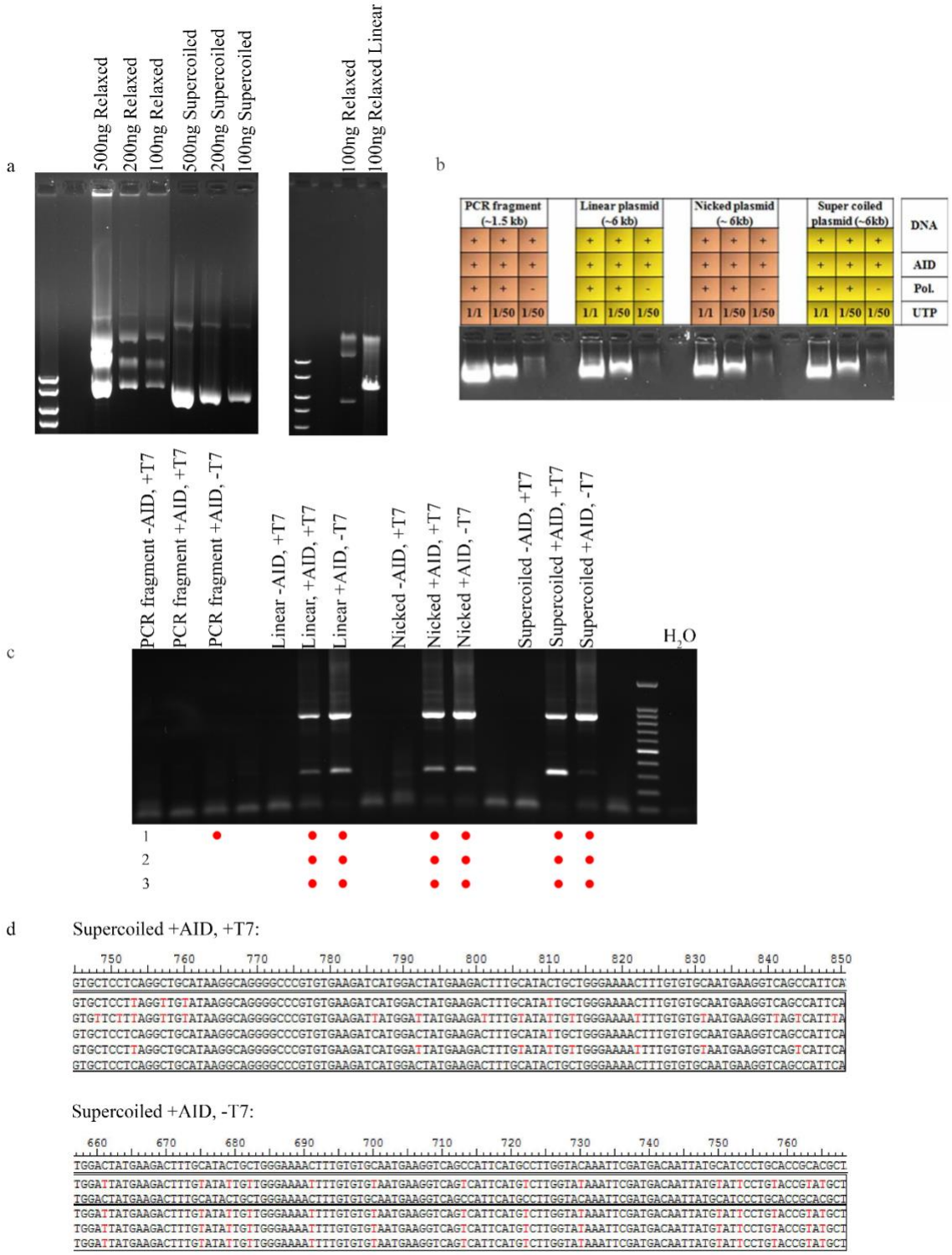


Figure 11: AID Consistently Mutated the Relaxed Nicked and Linear Duplex DNA in the Absence of Transcription. a) Agarose gel electrophoresis was used to verify the

topologies of the relaxed nicked, linear, and supercoiled DNA species prior to the TAAA assay. b) Each DNA topology (PCR fragment, relaxed linear, relaxed nicked, and supercoiled) was incubated with AID in the presence or absence of T7 polymerase. An RNA gel confirmed that T7 polymerase can transcribe each substrate in the presence of AID, and that RNA is not produced in the absence of T7 polymerase. c) Each reaction from (b) was subjected to deam-PCR to check for highly mutated sequences as a consequence of AID activity. Three independent PCRs were performed (1-3 underneath gel), and the red dots represent the presence of a band in a given PCR condition. AID acted consistently on supercoiled DNA both with and without transcription, as expected, and only highly deaminated the PCR fragment in one PCR without transcription. Surprisingly, AID consistently deaminated the relaxed linear and nicked circular forms of the plasmid both with and without transcription. There were no bands observed +T7 -AID or in the PCR H_2O negative, indicating that the PCR result is dependent on AID activity. d) To verify that the obtained PCR result was due to AID deaminase activity we sequenced the Supercoiled +AID +/-T7 conditions. All obtained sequences contained only C-T mutations confirming that the obtained result was due to AID activity.

3.4 Verifying AID Activity Using Degen-PCR

Since AID consistently acted on relaxed linear, nicked and supercoiled DNA in the absence of T7 polymerase (Figure 11), we next verified AID activity on supercoiled and relaxed linear DNA using the degen-PCR. We chose to use relaxed linear DNA as our relaxed template instead of the nicked DNA since it does not contain any supercoiled DNA and is therefore completely relaxed (Figure 9, 10). Heat-denatured supercoiled DNA was used as a control to show AID activity on ssDNA. To confirm that AID is indeed responsible for the C-T and G-A mutations that we observed after degen-PCR, we first have to rule out any mutations that could be caused by Taq polymerase during PCR amplification. Taq is quite error-prone with an established error rate between 1-20 x 10^{-5} errors/nucleotide/cycle (McInerney et al. 2014). We chose to use Taq polymerase instead of the high fidelity Pfu, which has a published error rate in the range of 1-2 x 10^{-6} errors/nucleotide/cycle (McInerney et al. 2014), because Pfu does not amplify over uracil. The ability of Taq polymerase to amplify over uracil is crucial since dU is generated by AID-mediated deamination of dC. Thus, we decided to determine the Taq error rate under the conditions of our system. As a control, AID dialysis buffer was used in place of AID. Three independent degen-PCR reactions were carried out. We combined the data from the 3 controls and determined the Taq polymerase error rate of our system to be approximately 1.42 x 10^{-5} errors/nucleotide/cycle, which is at the low end of the established error rate (McInerney et al. 2014). Since all template DNA was amplified by degen-PCR for 35 cycles, we corrected for Taq-generated C-T or G-A errors in our experimental conditions using the C-T or G-A mutation rates (mutations/nt) from the combined no AID controls.

For example, if 9 C-T mutations were detected in 139,368 nt. analyzed, then the Taq-mediated C-T error rate would be 9 C-T mutations/139,368 nt.= 6.46 x 10^{-5} mutations/nt. We then corrected for Taq errors in the experimental conditions by multiplying the C-T or G-A error rate by the number of nucleotides analyzed in a given condition. For example, in the heat-denatured supercoiled condition in Table 1 42 C-T mutations were found within 116,589 nt. sequenced. Multiplying the Taq-mediated C-T error rate (6.46 x 10^{-5} mutations/nt.) by the total number of nucleotides analyzed (116,589nt.) gives the approximate number of errors generated by Taq-polymerase ((6.46 x 10^{-5} mutations/nt.)(116,589 nt.) = 8 C-T errors). C-T and G-A mutations over and above Taq errors were considered to be generated by AID.

The degen-PCR did indeed confirm AID activity on both supercoiled and relaxed linear duplex DNA in the absence of transcription (Table 1, Figure 12). The heat-denatured supercoiled was most highly mutated as we expected, with a 3.44- and 3.62-fold higher C-T/G-A mutation rate than the supercoiled and linear DNA substrates, respectively. To our surprise, the linear DNA was mutated at nearly the same rate as the supercoiled DNA (1.05-fold difference) (Table 1, Figure 12), once again going against the grain of the literature in the field (Shen and Storb 2004).

Table 1	Number and Rate of C-T and G-A Mutations		
	GST-AID		
DNA Topology	Heat-Denatured Supercoiled	Supercoiled	Linear
C-T	42	15	13
G-A	37	14	17
Taq C-T Error Rate	6.46×10^{-5}	6.46×10^{-5}	6.46×10^{-5}
Taq G-A Error Rate	5.02×10^{-5}	5.02×10^{-5}	5.02×10^{-5}
Taq C-T Errors	8	7	7
Taq G-A Errors	6	5	6
Corrected C-T	34	8	6
Corrected G-A	31	9	11
Total C-T/G-A Mutations	65	17	17
Nucleotides Analyzed	116,589	104,804	110,087
C-T/G-A Mutation Rate	5.58×10^{-4}	1.62×10^{-4}	1.54×10^{-4}
Total Amplicons	99	91	94

Table 1: Number and Rate of C-T and G-A Mutations after degen-PCR. The Taq C-T or G-A errors were calculated by multiplying the respective Taq C-T or G-A error rate by the total number of nucleotides analyzed in a given condition. The corrected C-T or G-A mutations were obtained by subtracting the determined Taq C-T or G-A errors from the total C-T or G-A errors identified. The total C-T/G-A mutations is the sum of the corrected C-T and G-A mutations. The AID-mediated (C-T/G-A) mutation rate was found by dividing the total C-T/G-A mutations by the number of nucleotides analyzed. The supercoiled data was averaged from 5 independent AID-DNA reactions, while the linear and heat-denatured supercoiled (H-D. Supercoiled) data are from one AID-DNA reaction each.

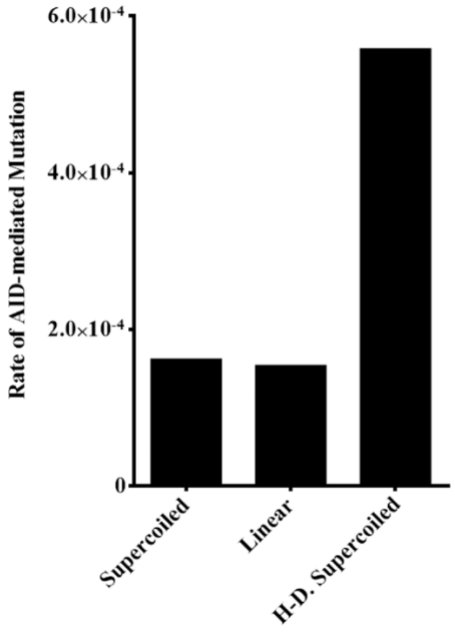


Figure 12: Rate of AID-mediated C-T and G-A Mutations after degen-PCR. The number of C-T and G-A mutations for each experimental condition was corrected for Taq error using our no AID control (Table 1). The rate of AID-mediated mutation was calculated by taking the sum of the total corrected C-T and G-A mutations and dividing it by the total number of nucleotides analyzed in each experimental condition. The supercoiled data was averaged from 5 independent AID-DNA reactions, while the linear and heat-denatured supercoiled (H-D. Supercoiled) data are from one AID-DNA reaction each.

3.5 Verification of Size and Substrate Preference of GST-AID

Since both our degen- (Table 1, Figure 12) and deam-PCR (Figure 11) results went against the literature (Shen and Storb 2004) and indicated that AID can indeed mutate relaxed linear DNA, we confirmed the size of the enzyme and its substrate preference *in vitro* to ensure that our GST-AID enzyme preparation maintains the properties of wildtype AID. We verified the size of GST-AID to be 50kDa (AID: 24 kDa, GST-tag: 26 kDa) using SDS-PAGE gel electrophoresis (Figure 13a). Next, we tested AID activity on its preferred small 7-nt.-long bubble substrate containing the 5'-WRC hotspot TGC in the standard alkaline cleavage assay (Figure 13b) (Quinlan et al. 2017, King et al. 2015, Abdouni et al. 2013, Larijani and Martin 2007, Larijani et al. 2007). We found that GST-AID was indeed active on the TGC bubble substrate (Figure 13b), so we further tested it on a 56nt. dsDNA substrate and its ssDNA counterpart (Figure 13c). We found that indeed GST-AID is not active on short dsDNA sequences, but can act on the same sequence if it is completely single-stranded, confirming the properties of wildtype AID previously described (Bransteitter et al. 2003, Pham et al. 2003, Dickerson et al. 2003, Larijani et al. 2005a, Larijani and Martin 2007, Larijani et al. 2007). Therefore, we can conclude that the activity of our GST-AID is comparable to the typical wildtype human AID, and the activity we observed (Figure 11, 12) must be true of wildtype human AID.

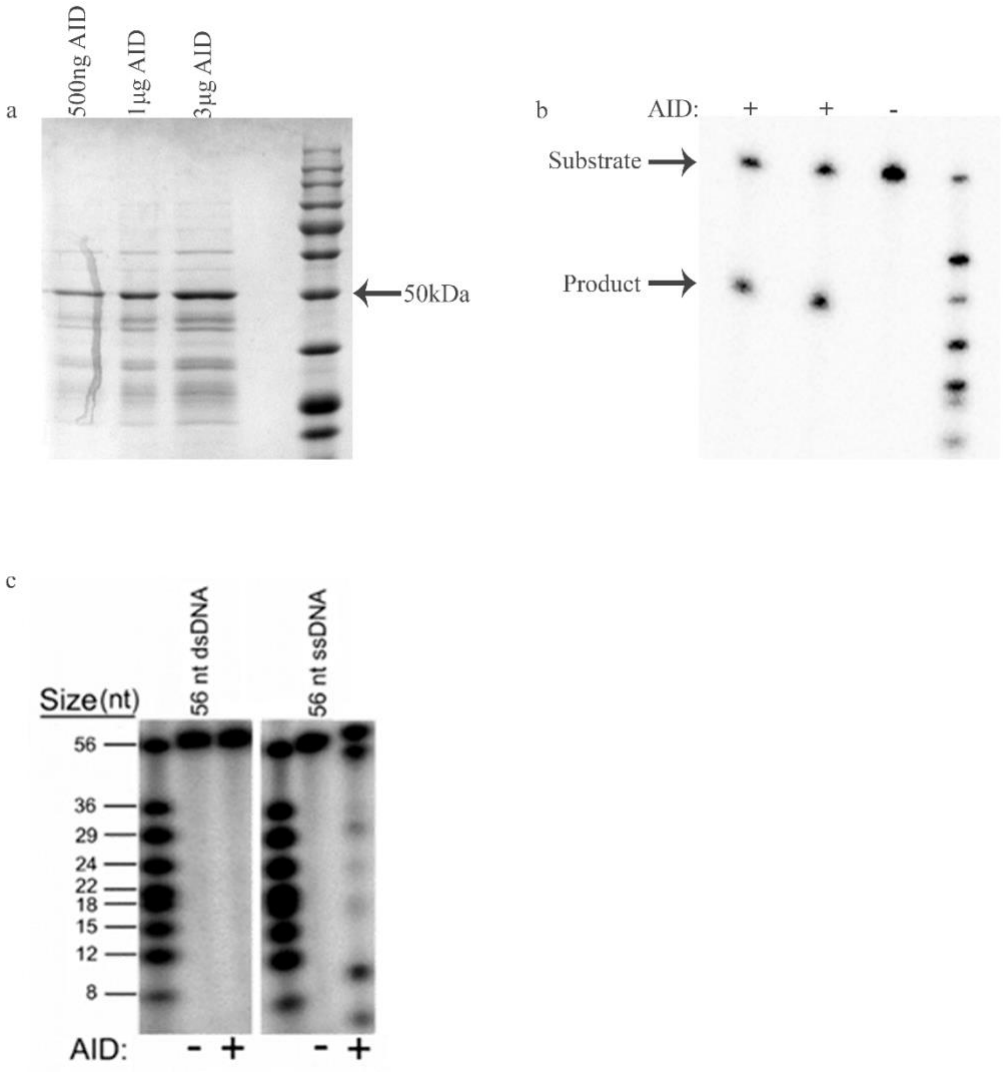


Figure 13: GST-AID is the Correct Size and Targets its Known Substrates. a) Known amounts of GST-AID (500ng, 1µg, 3µg) were loaded onto an SDS-PAGE gel and subjected to electrophoresis. The band for GST-AID migrated with the 50 kDa protein in the SDS-PAGE ruler. b) Representative alkaline cleavage gel showing that GST-AID is active on bubble substrate, and that the activity depends on the presence of AID. 50 fmol of bubble substrate was incubated in the presence (+) or absence (-) of GST-AID. c) Representative alkaline cleavage gel showing that GST-AID mutates ssDNA but not dsDNA. 50 fmol of ssDNA or dsDNA were incubated with GST-AID.

3.6 Confirming that AID is Responsible for the Observed Mutations

To ensure that there was not another DNA-processing enzyme capable of generating C-T or G-A mutations beyond Taq errors contaminating our preparation, we incubated our supercoiled substrate with a catalytically inactive AID mutant (W80R) enzyme. This version of AID contains an arginine in the place of tryptophan at position 80, obliterating enzymatic activity. W80R was incubated with supercoiled substrate DNA under the same conditions as wildtype AID, then the template DNA was subjected to degen-PCR. Once Taq errors had been accounted for, there were 4 residual mutations (1 C-T, 3 G-A) (Table 2). Thus, the frequency of mutation as a result of other potential contaminant DNA-processing enzymes in the GST-AID prep is approximately 3.74×10^{-5} (Table 2). This means that in 100,000 nt approximately 4 mutations could be due to a DNA processing enzyme other than AID.

Table 2	C-T and G-A Mutations - W80R
DNA Topology	Supercoiled
C-T	8
G-A	8
Taq C-T Error Rate	6.46×10^{-5}
Taq G-A Error Rate	5.02×10^{-5}
Taq C-T Errors	7
Taq G-A Errors	5
Corrected C-T	1
Corrected G-A	3
Total Mutations	4
Nucleotides Analyzed	106,872
Overall Mutation Rate	3.74×10^{-5}
Total Amplicons	91

Table 2: C-T and G-A Mutations Observed on the Target Substrate after Incubation with the Catalytically Inactive AID Mutant W80R. W80R was incubated with supercoiled substrate and then the template DNA was subjected to degen-PCR. The C-T or G-A Taq errors were subtracted from the total C-T or G-A mutations observed. A total of 4 C-T/G-A mutations were observed that could not be accounted for by Taq error, therefore the frequency of mutation as a result of other potential DNA-processing enzymes in the GST-AID prep is approximately 3.74×10^{-5}. The above data is obtained from one experimental reaction.

3.7 Southern Blot Analysis of DNA templates to confirm absence of ssDNA

Although the dsDNA plasmid substrates were purified using both CsCl density centrifugation and gel extraction to ensure there was no ssDNA contamination, we sought to confirm that the AID activity we observed in the absence of transcription (Figure 11, 12) was not due to any ssDNA contaminating our preparations of dsDNA templates. We chose to use Southern blot over agarose gel electrophoresis alone since it is far more sensitive and quantitative than an agarose gel alone. Furthermore, ssDNA will not be readily detected on an agarose gel using ethidium bromide stain since it does not have a double helix for ethidium to intercalate. Small amounts of ssDNA will go unnoticed on an agarose gel but can be detected using a Southern blot. Four independent Southern blots were completed using supercoiled and relaxed linear DNA incubated both with and without AID. A representative gel and its corresponding Southern blot is shown in Figure 14. To obtain a marker for ssDNA, supercoiled and linear DNA was heat-denatured immediately prior to loading the agarose gel (loaded to the right of samples, Figure 14). Relaxed linear DNA was loaded on the left of the gel, while supercoiled DNA was loaded on the right. The band indicating ssDNA in the heat-denatured markers ran at approximately 2.5 kb. A band of this size was not clearly observed in any other lane, with or without AID, nor was it present in the native marker. Upon quantifying all blots, it was determined that the percentage of ssDNA in the relaxed linear and supercoiled preparations was approximately 0.61% and 0.11%, respectively. If AID acts on 100 strands of DNA and we sequence all 100, the activity observed on <1 strand can be attributed to AID acting on ssDNA. Therefore, we

can be confident that >99% activity observed in our TIAA assay is due to AID acting on dsDNA.

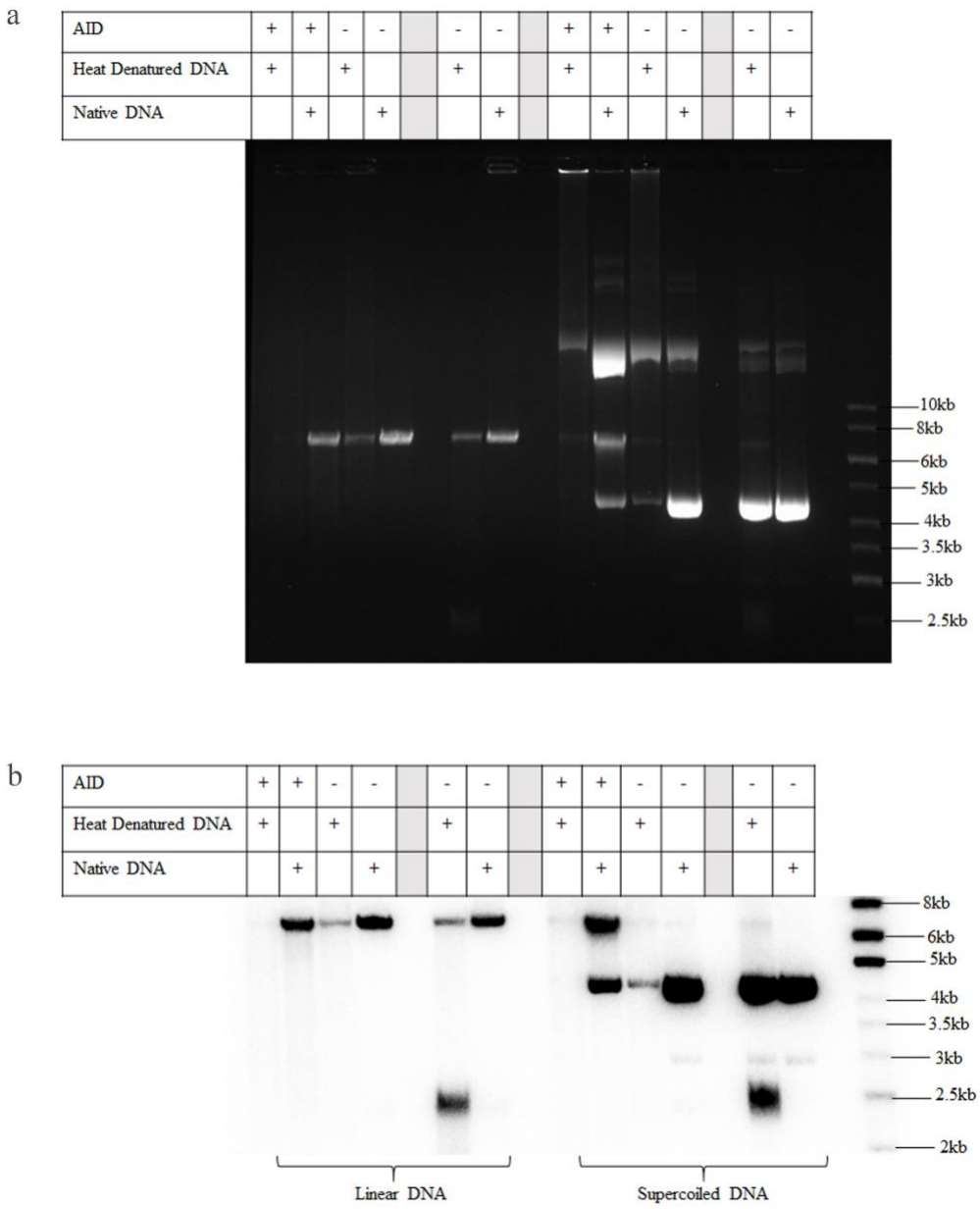


Figure 14: Southern Blots of Supercoiled and Linear DNA in the presence and absence of AID. Southern blots were used to determine whether or not ssDNA is contaminating our preparations of supercoiled and relaxed linear DNA leading to falsely observed AID activity on "dsDNA". DNA templates were blotted after incubation at 32℃ alone or in the presence of AID. As a control, supercoiled and relaxed linear templates were either heat-denatured at 98℃ or loaded in their native form as a size control for the incubated DNA. a) After incubation with or without AID, all samples and their control markers were run on a 1% agarose gel. b) The Southern blot was probed with an oligonucleotide probe complementary to an area in the middle of the target sequence. The

blot shown is representative of 4 independent experiments. Upon quantification of all 4 blots, it was determined that the percentage of ssDNA in the native linear and supercoiled preparations was 0.61% and 0.11%, respectively. Activity on less than 1% of amplicons can be attributed to AID acting on ssDNA.

3.8 Confirming AID activity on Supercoiled DNA using AID-His

To confirm AID activity on supercoiled DNA in our TIAA degen-PCR assay we used His-tagged human AID purified from HEK 293T cells. Since we used AID purified from a prokaryotic expression system (*E. coli*) in past experiments, we wanted to confirm that the activity we observed was due to the enzyme itself and not an artifact of the expression system. Small proteins of less than 100 amino acid residues can be efficiently purified from *E. coli* (Baneyx and Mujacic 2004). AID is a 24 kDa enzyme of 198-210 amino acids (reviewed in Larijani and Martin 2012), and its protein folding may not be fully supported in the absence of folding modulators such as chaperone proteins. Furthermore, the histidine tag is a small polypeptide C-terminal tag (8 histidine residues) whereas the GST tag is large 26kDa N-terminal tag. If we also observe activity on supercoiled DNA using our AID-His then we can conclude that our GST-AID is working correctly, and we can then proceed with characterizing AID activity on relaxed linear DNA.

We found that AID-His did indeed mutate the supercoiled DNA, confirming our results with GST-AID. However, AID-His was 116-fold more active than GST-AID on supercoiled duplex DNA ($1.62x10^{-4}$ vs $1.79x10^{-2}$, Table 3), and mutated the supercoiled DNA at a rate near equal to the heat-denatured supercoiled DNA (1.1-fold difference) (Table 3, Figure 15a). The distribution of AID-His-mediated mutations was also quite similar between the native and heat-denatured supercoiled DNA, 50% and 41% of the heat-denatured and native amplicons were mutated, respectively. Although in the absence of transcription we would expect the strand distribution of C-T and G-A mutations to be equal, there was a slight preference towards the top strand (C-T mutations) as 62.5% of the native

amplicons and 66.7% of the heat-denatured amplicons had C-T mutations. At this point, we can conclude that GST-AID is acting "normally", and that further work should be done to elucidate the strand preference of AID without transcription.

Table 3 Rate and Distribution of Mutations		
	AID-His	
DNA Topology	Heat-Denatured Supercoiled	Supercoiled
C-T	183	1241
G-A	83	639
Taq C-T Error Rate	6.46×10^{-5}	6.46×10^{-5}
Taq G-A Error Rate	5.02×10^{-5}	5.02×10^{-5}
Taq C-T Errors	1	7
Taq G-A Errors	1	5
Corrected C-T	182	1234
Corrected G-A	82	634
Total mutations	264	1868
Nucleotides Analyzed	16,236	104,428
Mutation Rate	1.63×10^{-2}	1.79×10^{-2}
Total Amplicons	18	117
Wildtype Amplicons	9	69
Mutated Amplicons	9	48
Amplicons with C-T Mutations	6	30
Amplicons with G-A Mutations	3	18

Table 3: Rate and Distribution of AID-His-mediated C-T and G-A Mutations in Supercoiled DNA. All error and mutation rates were determined as Table 1. The total number of amplicons were broken down into wildtype (no C-T/G-A mutations) and mutated (contained C-T/G-A mutations). The mutated amplicons were further broken down into amplicons with C-T or G-A mutations. A single amplicon cannot have both C-T and G-A mutations. The native supercoiled data was combined from 2 independent AID-DNA reactions, while the heat-denatured supercoiled (H-D. Supercoiled) data is from one AID-DNA reaction.

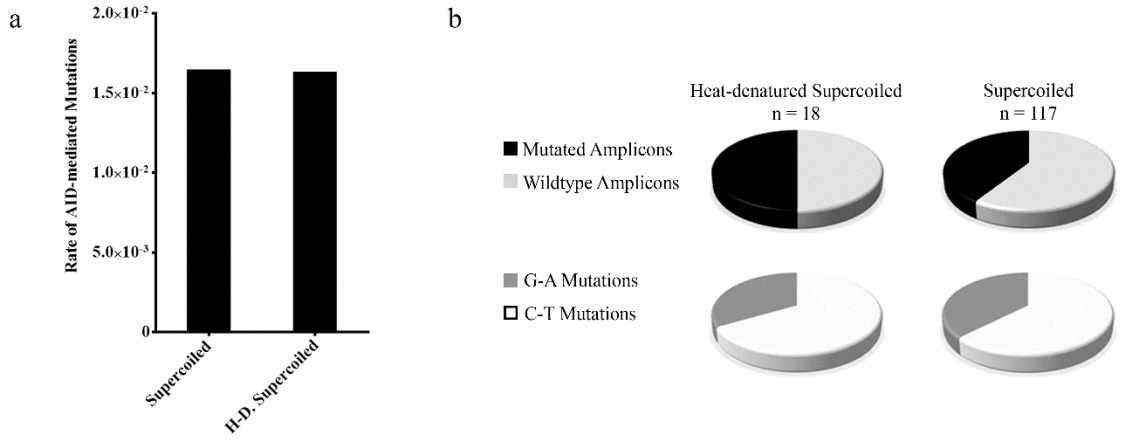


Figure 15: Rate and Distribution of AID-His-mediated C-T and G-A Mutations in Supercoiled DNA. a) The rate of AID-mediated C-T and G-A mutation on heat-denatured and native supercoiled substrate DNA. The number of C-T and G-A mutations for each experimental condition was corrected for Taq error using our no AID control (Table 3). The native supercoiled data was combined from 2 independent AID-DNA reactions, while the heat-denatured supercoiled (H-D. Supercoiled) data is from one AID-DNA reaction. b) The ratio of mutated to wildtype amplicons and amplicons with C-T or G-A mutations was plotted in pie charts, where "n" is the number of amplicons included in the analysis. The top row of pie charts show the ratio of mutated to wildtype amplicons, where mutated is shown in black and wildtype is show in grey. The bottom row of pie charts show the ratio of amplicons containing C-T mutations to those with G-A mutations, where G-A mutations are shown in dark grey and C-T mutations are shown in white.

3.9 Gel Extraction After TIAA Assay

Although the DNA templates were purified (Figure 9) and Southern blots confirmed that there was virtually no ssDNA in our reaction with AID under native conditions (Figure 14), we sought to demonstrate unequivocally that any AID activity observed is due to the enzyme acting on dsDNA under native conditions. Thus far, we have employed three independent procedures to ensure that sequencing of AID-mediated mutations on untranscribed relaxed linear and supercoiled substrates was obtained from reactions containing dsDNA substrates: first, substrates were purified by CsCl gradient centrifugation; second, linear and supercoiled substrates were gel-extracted; third, purified substrates, before and during AID incubations, were shown by Southern blotting to not contain any ssDNA contamination. These three measures provided us with confidence that AID was indeed mutating untranscribed supercoiled as well as relaxed linear dsDNA. However, to be absolutely certain, we introduced a fourth checkpoint: we incorporated a second gel extraction step after substrate incubation with AID (Figure 16), so as to ensure that if any ssDNA contaminants arose during AID incubation with linear or supercoiled DNA then they would be filtered out before degen-PCR. Yield considerations of this additional gel purification necessitated increasing the scale of the initial reactions. Thus, 10 AID-DNA reactions were pooled, followed by gel extraction of the supercoiled and relaxed linear bands.

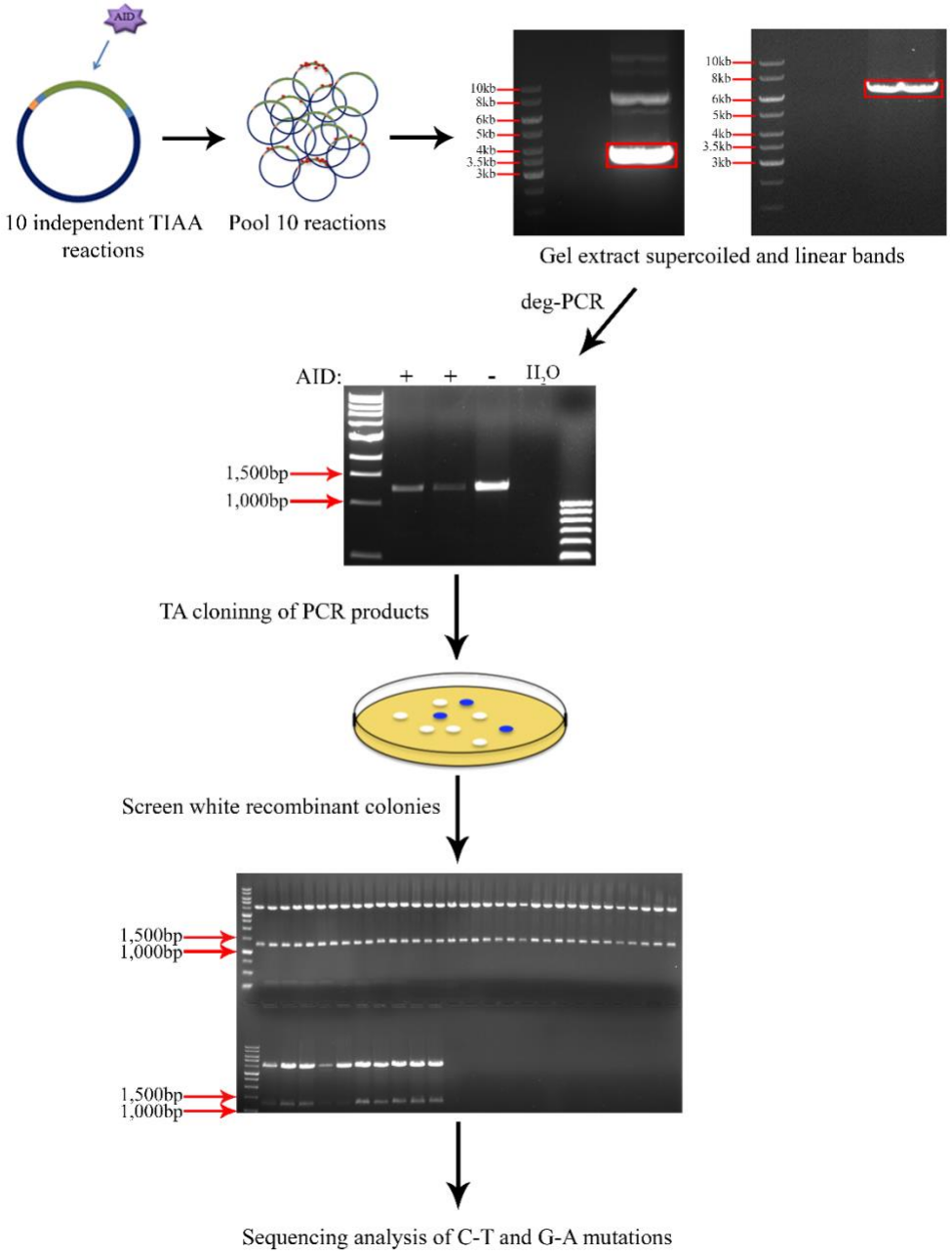


Figure 16: Modified Assay Schematic Including Gel Extraction. Ten reactions of AID and supercoiled, relaxed linear or heat-denatured DNA were incubated for 4 hours at 32°C. The reaction products were pooled and separated on an agarose gel where either the supercoiled or linear bands are gel extracted (outlined in red). Following gel extraction, the DNA templates were amplified by degen-PCR, the PCR products were TA cloned and then transformed into chemically competent *E. coli*. Each colony was screened for the target insert, the plasmid DNA was purified, and C-T and G-A mutations were identified through sequencing analysis.

We found that the rate of AID-mediated C-T/G-A mutation was 3-fold higher in the heat-denatured supercoiled than its native counterpart, but this time there were no mutations observed in relaxed linear DNA above Taq error (Table 4, Figure 17). Since the numbers of C-T and G-A mutations were low despite the large number of nucleotides analyzed (Table 4), we believed that there was still a limitation in our assay preventing us from viewing "true" AID activity.

Table 4	Number and Rate of Mutations		
	GST-AID		
DNA Topology	Heat-Denatured Supercoiled	Supercoiled	Linear
C-T	18	12	3
G-A	17	10	4
Taq C-T Error Rate	6.46×10^{-5}	6.46×10^{-5}	6.46×10^{-5}
Taq G-A Error Rate	5.02×10^{-5}	5.02×10^{-5}	5.02×10^{-5}
Taq C-T Errors	5	6	6
Taq G-A Errors	4	5	4
Corrected C-T	13	6	0
Corrected G-A	13	5	0
Total C-T/G-A Mutations	26	11	0
Bases Sequenced	71,404	91,571	86,610
C-T/G-A Mutation Rate	3.64×10^{-4}	1.20×10^{-4}	0
Total Amplicons	84	89	94

Table 4: Number and Rate of Mutations in Substrate DNA Gel Extracted after Treatment with GST-AID. The C-T/G-A mutation rate was determined as described in Table 1. The results above are combined from two independent data sets.

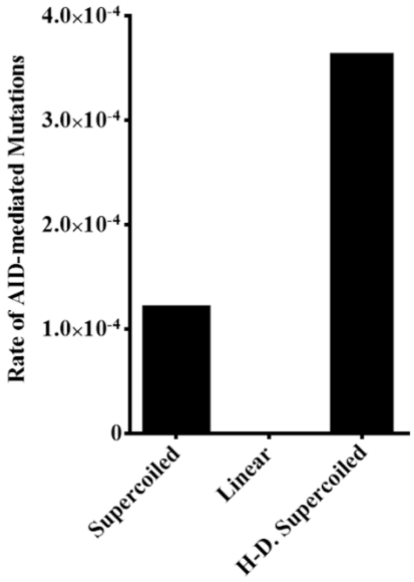


Figure 17: Rate of AID-mediated C-T and G-A Mutations in Substrate DNA Gel Extracted after Treatment with GST-AID. The number of C-T and G-A mutations for each experimental condition was corrected for Taq error using our no AID control (Table 4). The rate of AID-mediated mutation was calculated by taking the sum of "corrected" C-T and G-A mutations and dividing it by the total number of nucleotides analyzed in each experimental condition. The results above are combined from two independent data sets.

3.10 Optimizing the *In Vitro* AID Activity Assay to Observe the Unaltered and Original Foot-print of AID Activity on dsDNA in the Absence of Transcription

Thus far, we have shown that AID has modest activity on double-stranded supercoiled and relaxed linear DNA in the absence of transcription. However, the preference of AID activity between relaxed linear versus supercoiled DNA varied between individual experiments (Figure 12, Figure 17). Since the goal of designing this assay was to observe unbiased AID activity on different forms of DNA substrates, it was important to understand whether there were any biases in our assay against observing true AID activity. Our first goal was to understand if there were any enzymes contaminating our purified AID. Since GST-tagged AID cannot be purified to absolute purity and must be obtained at 90-95% purity in order to remain stable in solution (King and Larijani, 2017), we sought to examine if other possible DNA processing or repair enzymes might be residually present in the GST-AID preparations that may impact our assay. If this were the case, substrates that contain a high number of AID-generated uracils could be degraded, consistent with our lack of observation of any highly-mutated sequences, expected from AID's previously described highly processive mode of action (Pham et al. 2003). Since the first step in any repair pathway downstream of AID (BER or MMR) is uracil excision by UDG, followed by nuclease action at the abasic site, we chose to include uracil DNA glycosylase inhibitor (UGI) in our reactions. Blocking UDG activity "protects" uracils so we can observe the complete and unbiased landscape of uracils generated by AID.

Our second goal was to determine the optimal AID:DNA ratio in our assay. Too much DNA would increase the likelihood of not viewing AID-mediated mutations upon sequencing, whilst too little DNA could bias the result by either being too highly mutated or by being under the detection limit of our assay. We have previously demonstrated that more AID:ssDNA complexes are inactive due to substrate binding in a position where cytidine cannot be deaminated (King et al. 2015). Through analyzing 320 ssDNA docking clusters using 10 AID models in which the catalytic pocket conformation is open, it was found that only approximately 5% of docks positioned cytidine in an orientation that could lead to productive deamination. Based on this evidence it was estimated that approximately 1.3% of AID-DNA interactions lead to deamination.

3.10.1 "Protecting" Uracils using Uracil DNA Glyscosylase Inhibitor

The standard alkaline cleavage assay for AID activity was used to determine if there was any baseline UDG enzyme in our GST-AID prep (Figure 18). The average percentage of cleaved substrate with added UDG, as per standard alkaline cleavage protocol, was 54.7% (Figure 18, lanes 1 and 2). Lanes 3 and 4 show baseline UDG activity in the GST-AID prep, as no UDG was added to the reaction. The average percentage of cleaved substrate without further adding UDG is approximately 14.2%. These results confirmed that the GST-AID preparation does indeed contain residual UDG activity. This is not surprising, since the expression of AID in bacteria would cause genome-wide mutagenesis (Petersen-Mahrt et al. 2002, Ramiro et al. 2003) and upregulation of uracil processing

factors. We found that the baseline UDG can be effectively inhibited by addition of UGI as evidenced by the lack of cleaved substrate (Figure 18, lanes 5-14).

Lanes	1, 2	3, 4	5, 6	7, 8	9, 10	11, 12	13, 14	15, 16
AID	+	+	+	+	+	+	+	-
UDG	+	-	-	-	-	-	-	+
UGI	-	-	1	1/2	1/4	1/8	1/16	-

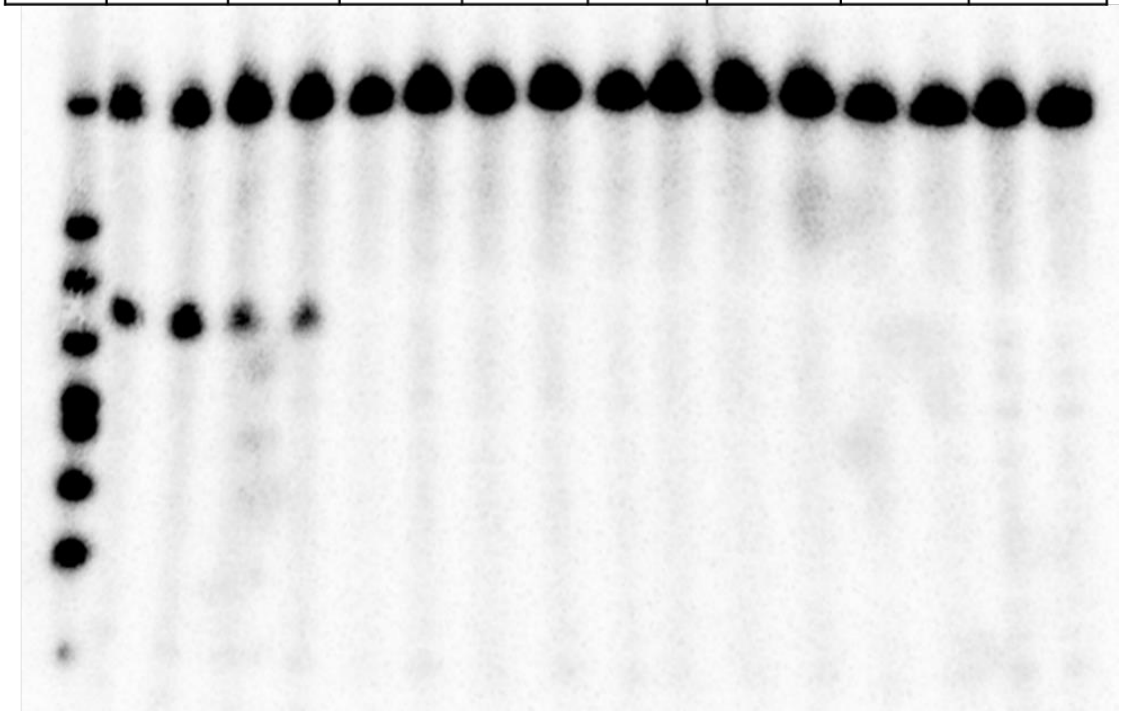


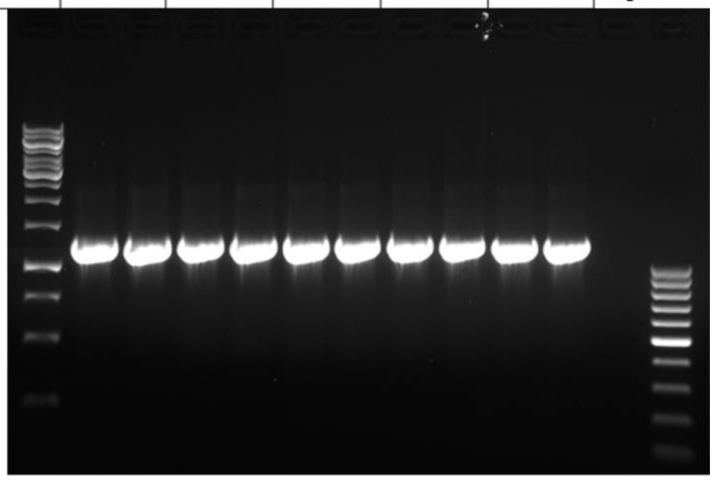
Figure 18: Alkaline Cleavage Assay to Determine UDG Activity in our GST-AID prep and Test Amount of UGI needed to Counteract UDG Activity. 50fmol of TGC bubble substrate 5'-labeled with [γ-^{32}P] dATP was incubated with 1.5µg of GST-AID and UGI for 3 hours at 32℃, then heat-inactivated at 85℃ for 20 minutes. UGI was serially diluted from 0.8 units/µl of enzyme to 0.05 units/µl (lanes 1-14). The no AID control (lanes 15-16) contained dialysis buffer instead of AID. Next, the TGC substrate was treated with either 1 unit (0.2µl) of UDG enzyme or water for 30 minutes at 37℃. After incubation, the TGC substrate was treated with 1M NaOH and ran on a denaturing polyacrylamide gel. The above gel is representative of 3 independent alkaline cleavage assays.

We next sought to test the efficacy of UGI addition in our TIAA assay using deam-PCR. GST-AID was first incubated with supercoiled DNA at 32℃ for 4 hours in the presence or absence of UGI. As a control, supercoiled DNA was heat-denatured prior to incubation with AID and/or UGI. Following incubation, the plasmid DNA was subjected to either degen-PCR (Figure 19a) or deam-PCR (Figure 19b). The degen-PCR revealed that the full 1.2kb target DNA strand was intact in all reactions and that the concentration of DNA in each reaction was consistent (Figure 19a). Therefore, any variation in band intensity of the deamination-specific PCR is due to the presence or absence of mutated DNA rather than overall template DNA. Deam-PCR revealed that AID is able to efficiently mutate heat-denatured DNA in the presence of UGI, as we did not observe a difference in band intensity with the control reaction lacking UGI (Figure 19b, lanes 5-8). On the other hand, inclusion of UGI in reactions containing AID and native supercoiled DNA resulted in a robust band indicative of highly mutated dsDNA (Figure 19b, lanes 1-4). To confirm that the inclusion of UGI had indeed protected uracils and allowed for detection of more AID-mediated mutations, the deam-PCR amplicons were sequenced and we found that there was a 1.6-fold increase in the C-T mutation frequency when UGI was added to the heat-denatured supercoiled reaction (Table 5). Since we did not obtain a PCR product upon deam-PCR of the native supercoiled +AID –UGI reaction (Figure 19), we were only able to sequence the PCR product for the native supercoiled +AID +UGI reaction. Thus, deam-PCR and sequencing confirmed that the addition of UGI facilitates the detection of AID-mediated mutations in native supercoiled DNA, but only modestly improves detection of

AID-mediated mutations when the DNA is heat-denatured. Figure 20 shows the position of the C-T mutations along the length of the deam-PCR fragment.

a

Lanes	1, 2	3, 4	5, 6	7, 8	9, 10	
AID	+	+	+	+	-	
UGI	-	+	-	+	-	
Heat-denature	-	-	+	+	+	H_2O



b

Lanes	1, 2	3, 4	5, 6	7, 8	9, 10	
AID	+	+	+	+	-	
UGI	-	+	-	+	-	
Heat-denature	-	-	+	+	+	H_2O

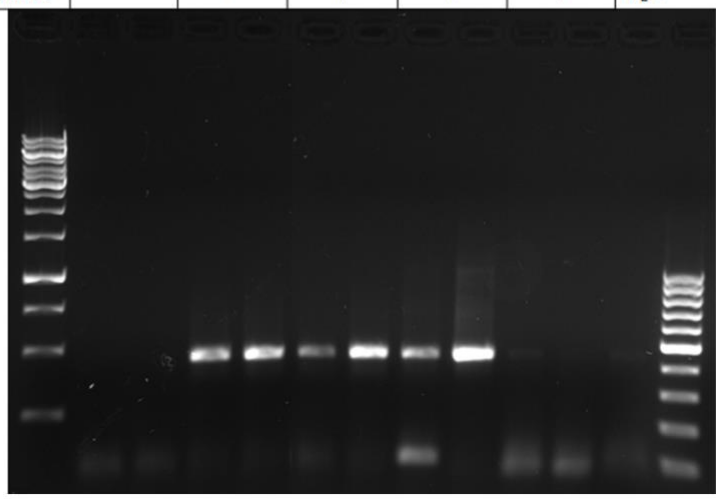


Figure 19: Testing UGI in the Deam-PCR Assay. 1.5 µg of AID was incubated with 50 ng of supercoiled DNA at 32℃ for 4 hours in the presence or absence of 0.8 units of UGI. As a control, supercoiled DNA was heat-denatured at 98℃ for 10 minutes to separate sister strands prior to incubation with AID and/or UGI. Following incubation, the plasmid DNA was subjected to either degen-PCR (a) or deam-PCR (b). a) Degen-PCR amplifies highly mutated, minimally mutated and unmutated DNA. It was used a control for deam-PCR to

show that the full 1.2 kb target DNA strand was intact in all reactions and that the concentration of DNA in each reaction is consistent. b) Deam-PCR selectively amplifies DNA containing AID-mediated C-T mutations. It allows us to test the efficiency of UGI in our PCR-based activity assay since the absence of a band means there was no detectable mutated DNA in the reaction, and the presence of a band indicates highly mutated DNA. While AID is able to efficiently deaminate heat-denatured DNA in the presence or absence of UGI (lanes 5-8), treatment of native supercoiled DNA with UGI allowed us to completely visualize a band indicating highly mutated dsDNA (lanes 1-4). This proves that AID can indeed mutate dsDNA and that the presence of UDG in our GST-AID prep was inhibiting us from visualizing true AID activity.

Table 5 Number and Rate of C-T mutations			
	GST-AID		
DNA Topology	Supercoiled +UGI	Heat-denatured Supercoiled -UGI	Heat-denatured Supercoiled +UGI
C-T Mutations	290	41	66
Taq C-T Error Rate	6.46×10^{-5}	6.46×10^{-5}	6.46×10^{-5}
Taq C-T Errors	1	0	0
Corrected C-T Mutations	289	41	66
Nucleotides Analyzed	8,078	2,646	2,226
Mutation Rate	3.58×10^{-2}	1.55×10^{-2}	2.97×10^{-2}
Total Amplicons	18	6	5

Table 5: Number and Rate of AID-mediated C-T Mutations from the Deam-PCR Assay. All error and mutation rates were determined as described in Table 1. The native supercoiled +UGI data is compiled from 2 independent reactions, the heat-denatured data is from a single reaction.

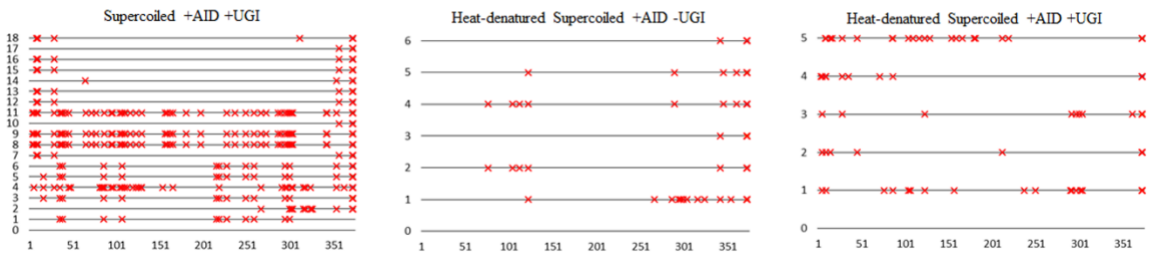


Figure 20: C-T Mutation Map of deam-PCR Amplicons with or without addition of UGI. The deam-PCR amplicons in Figure 19a were TA cloned and sent for sequencing. The red x's correspond to the location of C-T mutations along the deam-PCR fragment, where the x-axis corresponds to the nucleotide position and the y-axis corresponds to the number of independent amplicons analyzed by sequencing. The sequencing results confirm the mutations observed by deam-PCR, and shows that adding UGI to our assay allows us to observe AID activity on double-stranded supercoiled DNA. The native supercoiled +UGI data is compiled from 2 independent reactions, the heat-denatured data is from a single reaction.

3.10.2 Determining the "Optimal" Ratio of AID:DNA to Accurately View AID Activity in our *In Vitro* TIAA Assay

The ratio of enzyme:substrate in any reaction is important in determining the efficiency of catalysis. AID/APOBECs form numerous catalytically-productive as well as non-specific interactions with DNA, due to their high positive surface charge (King and Larijani, 2017). To examine the influence of the AID:DNA ratio on mutation outcomes, we altered the DNA concentration while holding the amount of AID constant. We wished to avoid having too much DNA as we thought it may hinder the detection of AID-mediated mutations. We also sought to avoid having too little DNA, which could bias the result towards high mutation rates, or possibly also towards lack of mutation detection if too little DNA did not allow for enough productive AID:DNA complexes. We calculated the number of copies of plasmid per number of nanograms DNA (Table 6). We chose 100 ng as our upper limit since this is the amount used in each reaction in all previous experiments, and the amounts after 20 ng were chosen to decrease the copy number in 10-fold increments (Table 6).

Nanograms of DNA	Copies of Plasmid	Picomoles of DNA	Ratio of AID:DNA
100	1.38×10^{10}	2.2×10^{-2}	1.1×10^{3}
50	6.88×10^{9}	1.1×10^{-2}	2.2×10^{3}
20	2.75×10^{9}	4.5×10^{-3}	5.3×10^{3}
2	2.75×10^{8}	4.5×10^{-4}	5.3×10^{4}
2×10^{-1}	2.75×10^{7}	4.5×10^{-5}	5.3×10^{5}
2×10^{-2}	2.75×10^{6}	4.5×10^{-6}	5.3×10^{6}
2×10^{-3}	2.75×10^{5}	4.5×10^{-7}	5.3×10^{7}
2×10^{-4}	2.75×10^{4}	4.5×10^{-8}	5.3×10^{8}
2×10^{-5}	2.75×10^{3}	4.5×10^{-9}	5.3×10^{9}
2×10^{-6}	2.75×10^{2}	4.5×10^{-10}	5.3×10^{10}

Table 6: Number of Copies of Plasmid DNA per Number of Nanograms. An online calculator was used to calculate the approximate number of copies of plasmid per number of nanograms DNA (URI Genomics & Sequencing Center; cels.uri.edu/gsc/cndna.html). 100 ng was chosen because this is the amount we used in all previous experiments. The amounts after 20 ng were chosen because we wanted to decrease the plasmid copy number in 10-fold increments to determine the lower limit of detectable AID activity in our assay. The size of the pcDNA3.1D/V5-His-TOPO plasmid containing our target DNA insert is 6.7 kb. The concentration of AID was held constant at approximately 120 ng/µl. 1.2 µg of AID was used in all reactions, which is equal to approximately 24 pmol.

First, we treated supercoiled and linear DNA in the amounts listed in Table 6 with 1.2 µg of GST-AID, which is approximately equal to 24 pmol of AID. Next 1µl of each reaction was amplified by degen-PCR to show the intensity of the bands as a reflection of the total amount of DNA in the reaction. We expected a positive relationship between the band intensity and amount of substrate, such that the band intensity increases with increasing amount of DNA added to the PCR. The positive relationship is confirmed in Figures 21 and 23. Thus, if there are any deviations from this pattern upon amplifying the template DNA using the deamination-specific primers, it is due to the amount of highly mutated DNA in the reaction and not reflective of the total amount of template DNA. Next, the same reactions were subjected to deam-PCR (Figure 22). We found that there was not a direct linear relationship between the amount of total DNA in the reaction and the amount of highly mutated DNA. Within the supercoiled set there was no significant difference in the average band intensity from 100 ng to 0.02 ng, indicating that there is approximately the same amount of highly mutated DNA in these reactions. Since the amount of DNA added to the deam-PCR was not equalized (i.e. there was 5,000 times less DNA added to the 0.02 ng PCR reaction than the 100 ng reaction), the deam-PCR result indicates that less DNA is actually more efficient in the TIAA assay (Figure 23). At 0.002 ng and beyond the amount of highly mutated DNA declines, as 0.002 ng has significantly less highly mutated DNA than dilutions 100 ng-0.02 ng ($p < 0.05$). However, this is not necessarily reflective of mutation efficiency since there was 10 times less DNA added to the deam-PCR reaction than the previous dilution (0.02 ng). Within the linear set, the average band intensity is significantly higher in the 20 ng ($p < 0.05$) and 2 ng ($p < 0.01$) reactions in comparison to

the 100 ng reaction indicating once again that AID can more efficiently mutate DNA when there is less (Figure 23). At 0.2 ng there was significantly less highly mutated DNA than the 100 ng ($p < 0.01$), 20 ng and 2 ng ($p < 0.05$) dilutions (Figure 23). Furthermore, although AID can highly mutate both supercoiled and relaxed linear DNA, it seems to act more efficiently on supercoiled DNA as the deam-PCR dilutions persisted two dilutions further for the supercoiled versus the linear substrate (Figure 23). The lowest consistently detected band in the degen-PCR was at 0.0002 ng for the supercoiled substrate and 0.002 ng for the relaxed linear, whereas the lowest consistently detected band in the deam-PCR was at 0.0002 ng for the supercoiled substrate and 0.02 ng for the linear substrate (Figure 23).

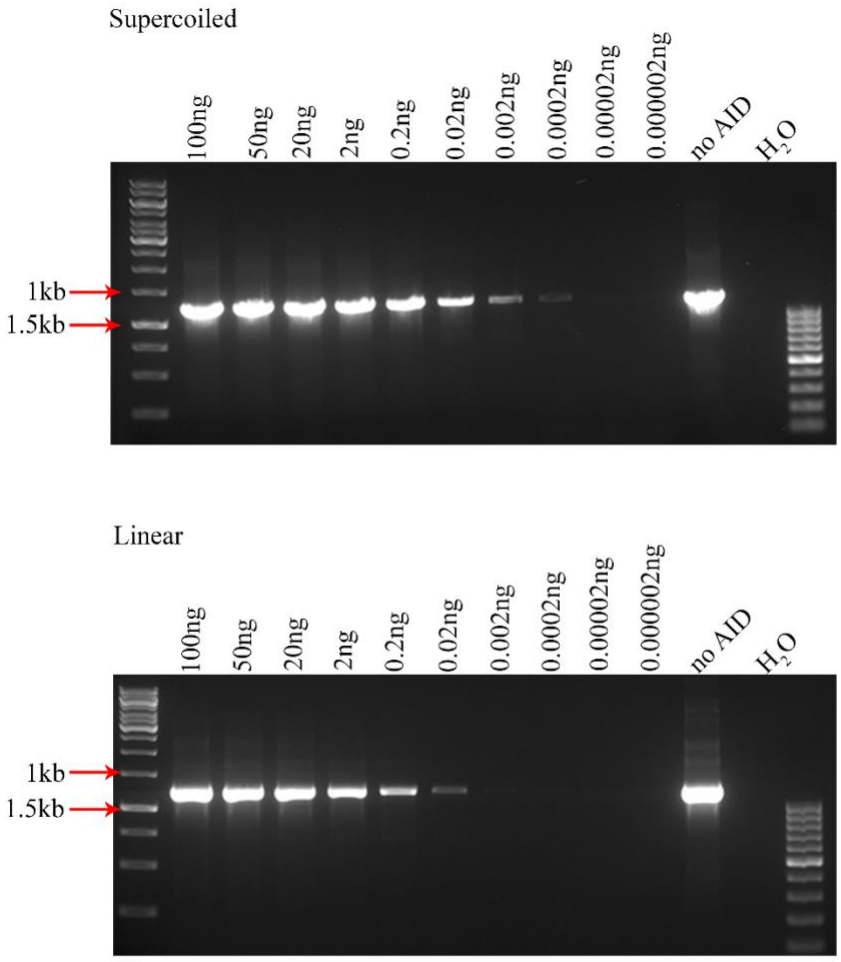


Figure 21: Degen-PCR Results for Dilutions Assay. 1.2 µg of AID was incubated with supercoiled and linear DNA in the concentrations listed in Table 6, and 1 µl of each reaction was subjected to degen-PCR. 50 ng of either the supercoiled or linear DNA that was not incubated with AID was also PCR amplified and used as a size control to show the correct band size of 1.2 kb. The initial AID reactions were done in triplicate, and the above gels are representative of the results from three independent degen-PCRs. The reactions were loaded in descending order from 100 ng to 2 x 10^{-6} ng. Since the degen-PCR has no specificity towards mutated or wildtype DNA, the intensity of the band is reflective of the concentration of the reaction such that more DNA will give you a more intense band and vice versa. If there are any deviations from this pattern upon amplifying the template DNA using the deamination-specific primers, it is due to the amount of highly mutated DNA in the reaction and not necessarily the total amount of DNA in the reaction.

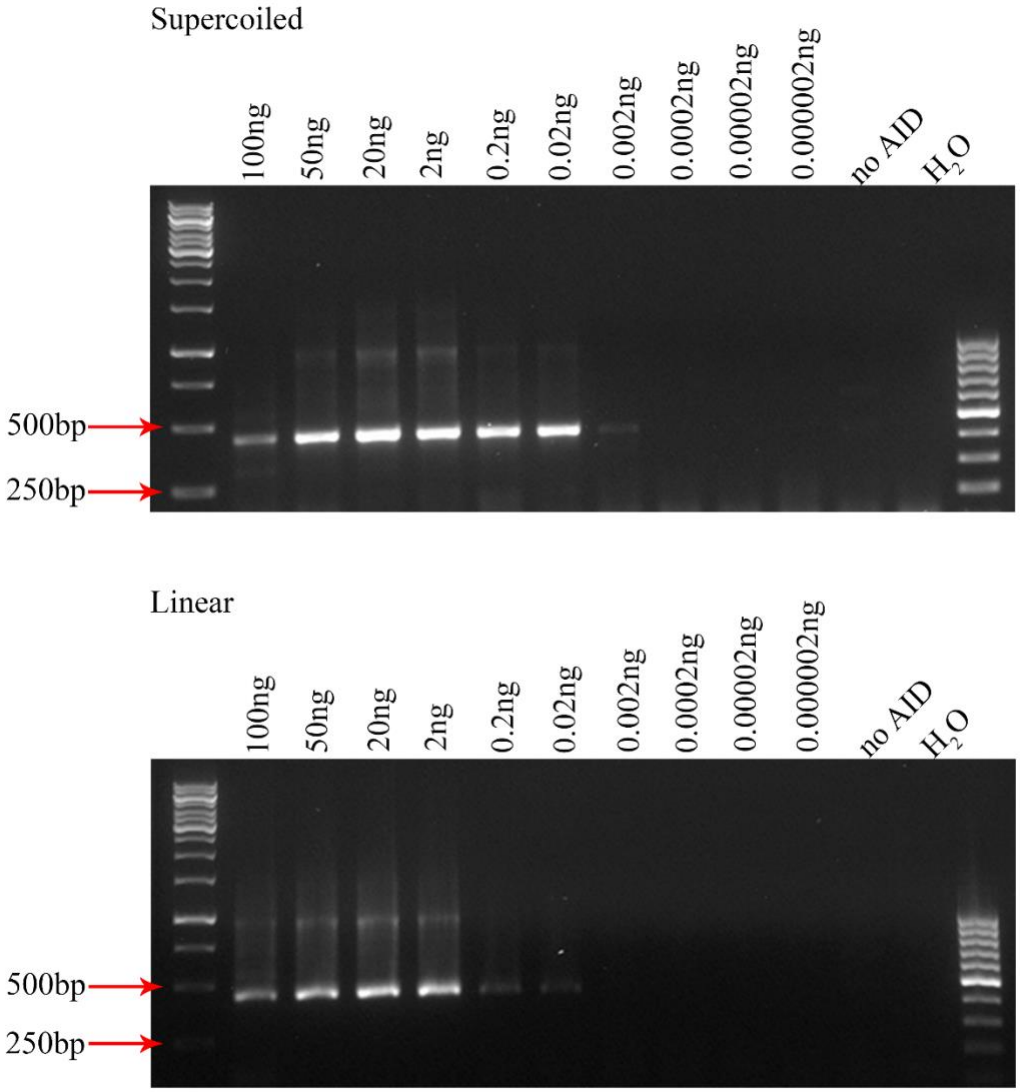


Figure 22: Deam-PCR Results for Dilutions Assay. 1.2 µg of AID mixed with 4 x 10^{-4} units of UGI was incubated with supercoiled and linear DNA in the concentrations listed in Table 6. 1 µl of each reaction was subjected to deam-PCR. The product size is 450 bp. 50 ng of either the supercoiled or linear DNA was incubated with dialysis buffer and PCR amplified to show that PCR amplification is dependent on the presence of C-T mutations. The initial AID reactions were done in triplicate, and the above gels are representative of the results from three independent deam-PCRs. The reactions were loaded in descending order from 100 ng to 2 x 10^{-6} ng. The intensity of the band reflects the amount of highly mutated DNA that was subjected to PCR.

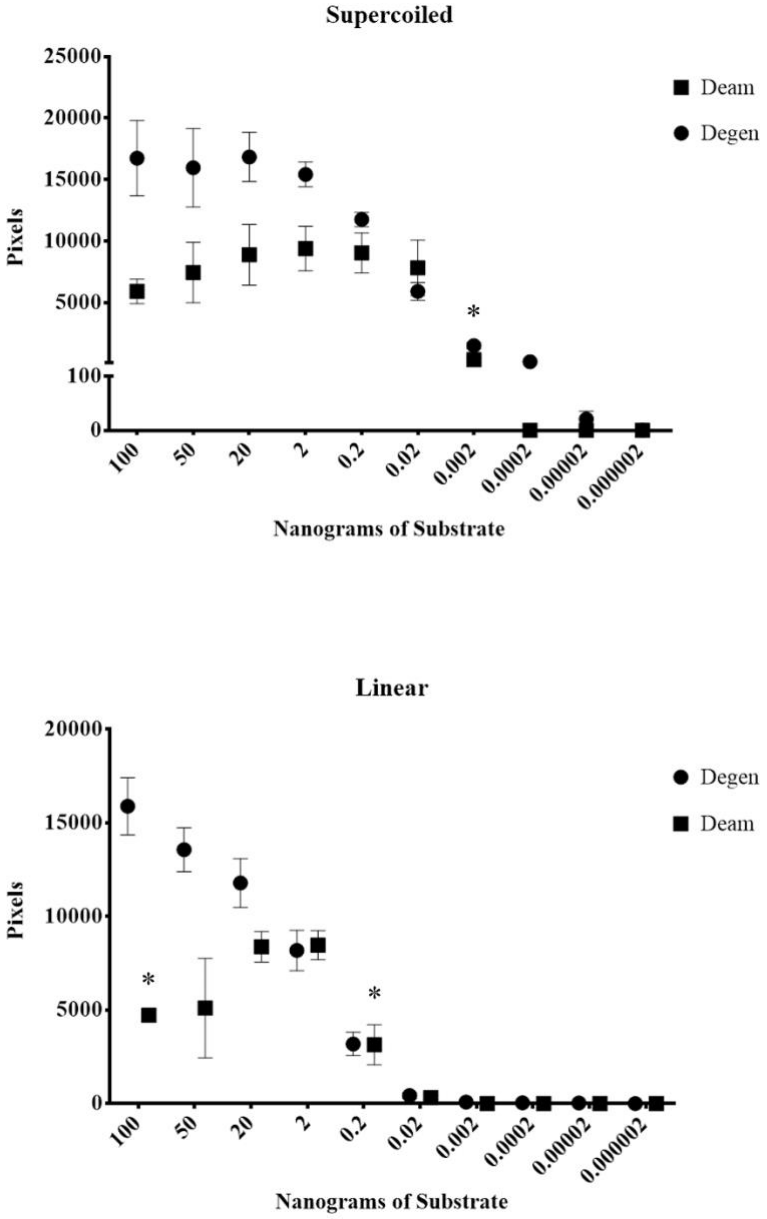


Figure 23: Quantification of Degen-PCR and Deam-PCR Results for Dilutions Assay. Three degen-PCRs and three deam-PCRs were performed using three independent reactions containing supercoiled (top) or linear (bottom) DNA. The gels were quantified using ImageJ and plotted using GraphPad Prism 6 (example gels in Figure 21, 22). The nanograms of substrate DNA are plotted on the x-axis, and the number of pixels for each band on the y-axis. Each point plotted represents the mean number of pixels from each of the three bands quantified, and the error bars show the standard error of the mean (SEM). The graph shows that the number of pixels decreases with the decreasing DNA substrate amount in the degen-PCR, but the amount of total DNA present in the reaction does not

directly reflect the amount of highly mutated DNA in the reaction as determined by deam-PCR. A two-tailed equal variance t-test was conducted to determine whether or not there was a significant difference in the amount of highly mutated DNA as the amount of substrate decreased. There was no significant difference within the dilution range 100 ng-0.02 ng for the supercoiled substrate, and the amount of highly mutated DNA significantly decreased at 0.002 ng ($p < 0.05$). For the linear substrate, there was no significant difference in highly mutated DNA from dilutions 50 ng-2 ng, but the amount of highly mutated DNA in the 100 ng and 0.2 ng reactions was significantly lower than the 20 ng and 2 ng reactions ($p < 0.05$). The lowest consistently detected band in the degen-PCR was at 0.0002 ng for supercoiled DNA and 0.002 ng for linear DNA. The lowest consistently detected band in the deam-PCR was at 0.0002 ng for the supercoiled substrate and 0.02 ng for the linear substrate.

3.10.3 AID Mutates Supercoiled DNA with a 10-100-fold Preference over Linear DNA

The "optimal" amount of substrate was not immediately clear from the first deamination-specific PCRs (Figure 22, 23), therefore we wanted to compare the amount of highly mutated DNA in each reaction. To do this we incubated AID and UGI with supercoiled and linear DNA in the amounts: 100 ng, 50 ng, 20 ng, 2 ng, 0.2 ng, 0.02 ng and 0.002 ng. Each reaction was serially diluted based on the amount (ng) of DNA added to the PCR such that the bands from each reaction could easily be compared to the others (i.e. 100ng reaction dilution: 10 ng, 5 ng, 2 ng, 0.2 ng, 0.02 ng, 0.002 ng, 0.0002 ng, 0.00002 ng; 50 ng dilution: 5 ng, 2 ng, 0.2 ng, 0.02 ng, 0.002 ng, 0.0002 ng, 0.00002 ng) (Figure 24). 1 µl of each dilution was subjected to deam-PCR. Comparing the number and intensity of the bands obtained from deam-PCR for each dilution allowed us to estimate the relative amount of mutated DNA. Furthermore, this assay provides a semi-quantitative approach to compare AID activity on the supercoiled and relaxed linear topologies.

On the supercoiled substrate, AID can heavily mutate minute amounts of DNA most efficiently as evidenced by the bands at 0.0002 ng dilutions for the 0.2 ng and 0.002 ng reactions that were not present in the reactions with larger amounts of DNA (Figure 24, top). 100 ng of supercoiled substrate appears to be as good as 50 ng, 20 ng and 2 ng, since the last band observed is for the 0.002 ng and the last consistent band observed for these reactions is at 0.02 ng (indicated by the red dots, Figure 24). For the relaxed linear substrate, 100 ng, 50 ng and 20 ng can be highly mutated at a similar efficiency since the last consistent dilution was at 0.2 ng (Figure 24, bottom). Overall AID can most efficiently highly mutate small amounts of DNA (e.g. 0.2 ng or 0.002 ng) if it is supercoiled in

topology, but needs higher amounts of relaxed linear DNA (100 ng, 50 ng, 20 ng) to be able to highly mutate it efficiently (Figure 24). Therefore, supercoiled must be a better substrate for AID than relaxed linear. If each reaction is compared between the supercoiled and linear, we see that the band representing highly mutated linear DNA disappears at 1-2 dilutions before that of the supercoiled (eg. 100 ng – 0.02 ng vs 0.002 ng; 50 ng – 0.2 ng vs 0.02 ng; 20 ng – 0.2 ng vs 0.002 ng). Furthermore, no bands were observed in the 0.2 ng, 0.02 ng or 0.002 ng reactions for the relaxed linear while there were bands observed at the 0.2 ng and 0.002 ng reactions for the supercoiled topology (Figure 24). Overall, we can conclude that supercoiled dsDNA is a 10-100 times better substrate for AID than relaxed linear dsDNA in the absence of transcription. Although we did observe that AID can efficiently mutate minute amounts of supercoiled DNA, we chose to continue using 100 ng in our future experiments. Since there was no detectable activity on the linear DNA with less than 2 ng, and there was barely a difference in activity between the 100 ng, 50 ng and 20 ng reactions in both supercoiled and linear DNA, using 100 ng will allow us to compare activity between the two topologies while providing enough DNA for downstream processes.

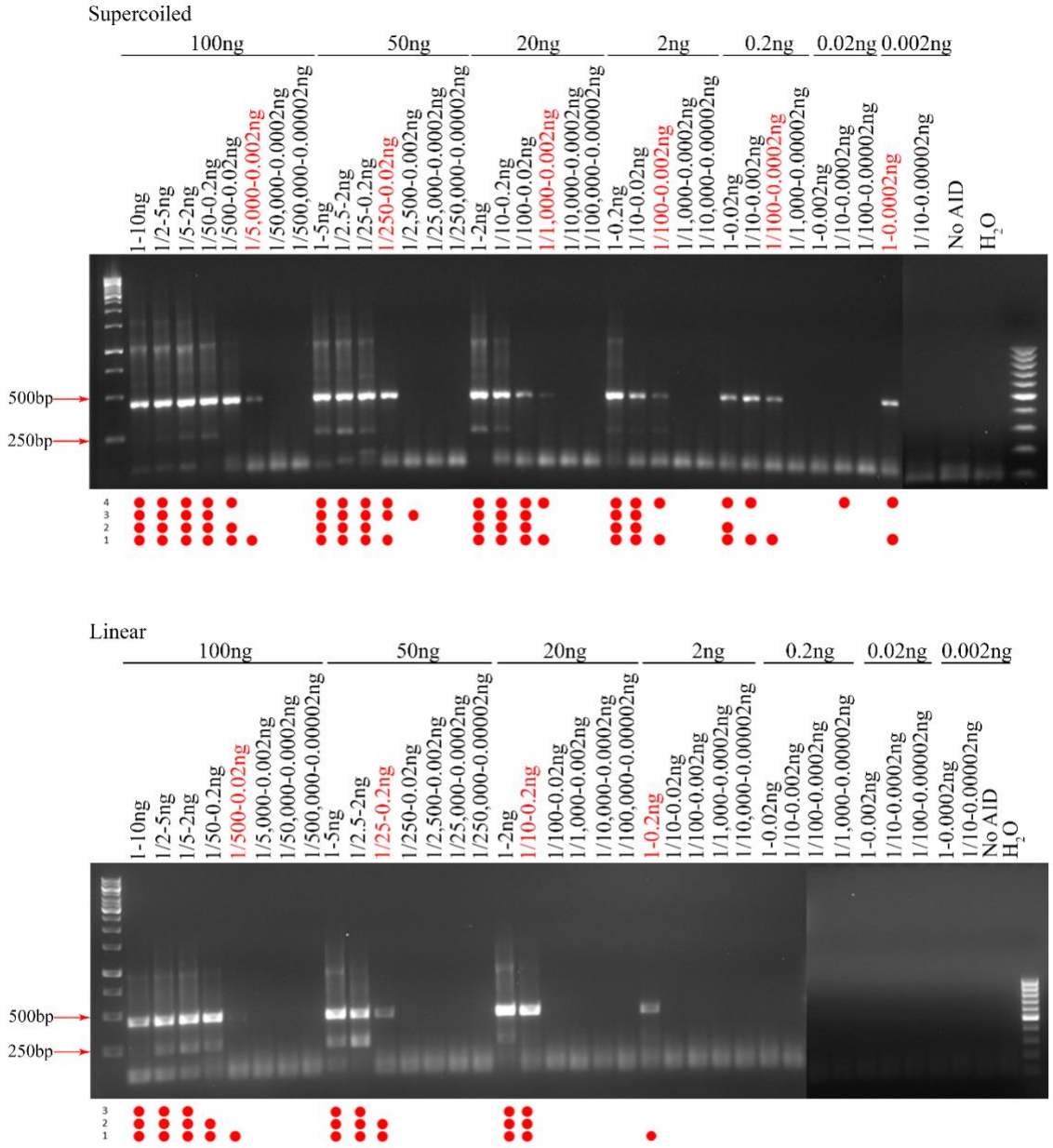


Figure 24: Deam-PCR Dilutions for Supercoiled and Relaxed Linear DNA treated with purified GST-AID. 1.2 µg of AID (4 µl) was mixed with 4 x 10^{-4} units of UGI and incubated with either supercoiled or relaxed linear DNA in the following amounts: 100 ng, 50 ng, 20 ng, 2 ng, 0.2 ng, 0.02 ng and 0.002 ng. Each reaction was serially diluted to the amounts in the figure above, and 1µl of each dilution was subjected to deam-PCR. The product size is 450 bp. 50 ng of either the supercoiled or relaxed linear DNA was incubated with AID dialysis buffer instead of AID, and PCR amplified to show that deam-PCR is dependent on the presence of AID-mediated C-T mutations. The result above is representative of 4 deam-PCRs from 4 independent reactions for the supercoiled substrate,

and 3 PCRs from 3 independent reactions for the relaxed linear substrate. The red dots underneath the gels represent the bands observed for each of the independent PCRs. The PCR bands persisted 1-2 more dilutions for each reaction in the supercoiled set than in the linear, indicating that while AID can mutate relaxed linear DNA, supercoiled is the preferred topology. The bottom halves of the two gels above were merged with the top halves to compare band intensity across dilutions.

To ensure that the 10-100 times preference for supercoiled DNA is a *bona fide* property of purified AID and not due to the particular expression or purification system of GST-AID, we tested AID purified from a different system using a different fusion tag. We tested two different preparations of AID-His: purified AID-His, and an AID-His expressing 293T whole cell lysate. If the trends in the results are the same for all three preparations of human AID (GST-AID, purified AID-His and AID-His in lysate) then we can be confident that the results are representative of true AID activity, rather than a function of expression system or purification tag. For experiments conducted with purified AID-His, and AID-His-expressing 293T whole cell lysates, two further negative controls were used to ensure the specificity of the deamination-specific PCR assay: 293T whole cell lysates from cells that were not transfected with AID expressing vectors, and the dialysis buffer in which AID is stored (Figure 25). As compared to the data obtained with purified GST-AID (Figure 24), the bands for the AID-His data set (Figure 25, 26) did not persist for as many dilutions, possibly because the 293 T cell-expressed AID preparations are considerably more dilute than the bacterially-expressed GST-AID. Nevertheless, the same trend of preference for supercoiled over relaxed linear DNA was observed. In the data set using purified AID-His, the bands persisted 1-2 more dilutions for each reaction in the set with supercoiled DNA than the set with relaxed linear DNA (e.g. 100 ng reaction: 2 ng vs 10 ng; 20 ng reaction: 0.2 ng vs 2 ng; Figure 25). Based on these results we conclude that purified AID-His also has a 10-100-fold preference for supercoiled over relaxed linear dsDNA. As a control for AID activity on ssDNA, 100 ng of both linear and supercoiled DNA were heat-denatured, diluted as the 100 ng dsDNA reactions, and subjected to deam-

PCR. An example gel showing the result from the heat-denatured linear substrate is shown at the bottom of Figure 25. In the 100 ng reaction the band disappeared at 0.02 ng for the heat-denatured linear substrate, 2 ng for the supercoiled and 10 ng for the relaxed linear. Based on these results, AID has a 100-fold preference for ssDNA over supercoiled duplex DNA and a 500-fold preference for ssDNA over relaxed linear duplex DNA (Figure 25). In the data set using the 293T cell lysate containing AID-His, the PCR bands persisted for the same number of dilutions for both the supercoiled and linear templates for the 100 ng and 20 ng reactions; however, the bands are on average approximately 3-fold brighter for supercoiled than linear DNA (Figure 26). Overall, the data from Figures 25 and 26 support the trend found in Figures 23/24 that AID does indeed target relaxed linear DNA, however it has a 10-100-fold preference for the supercoiled topology.

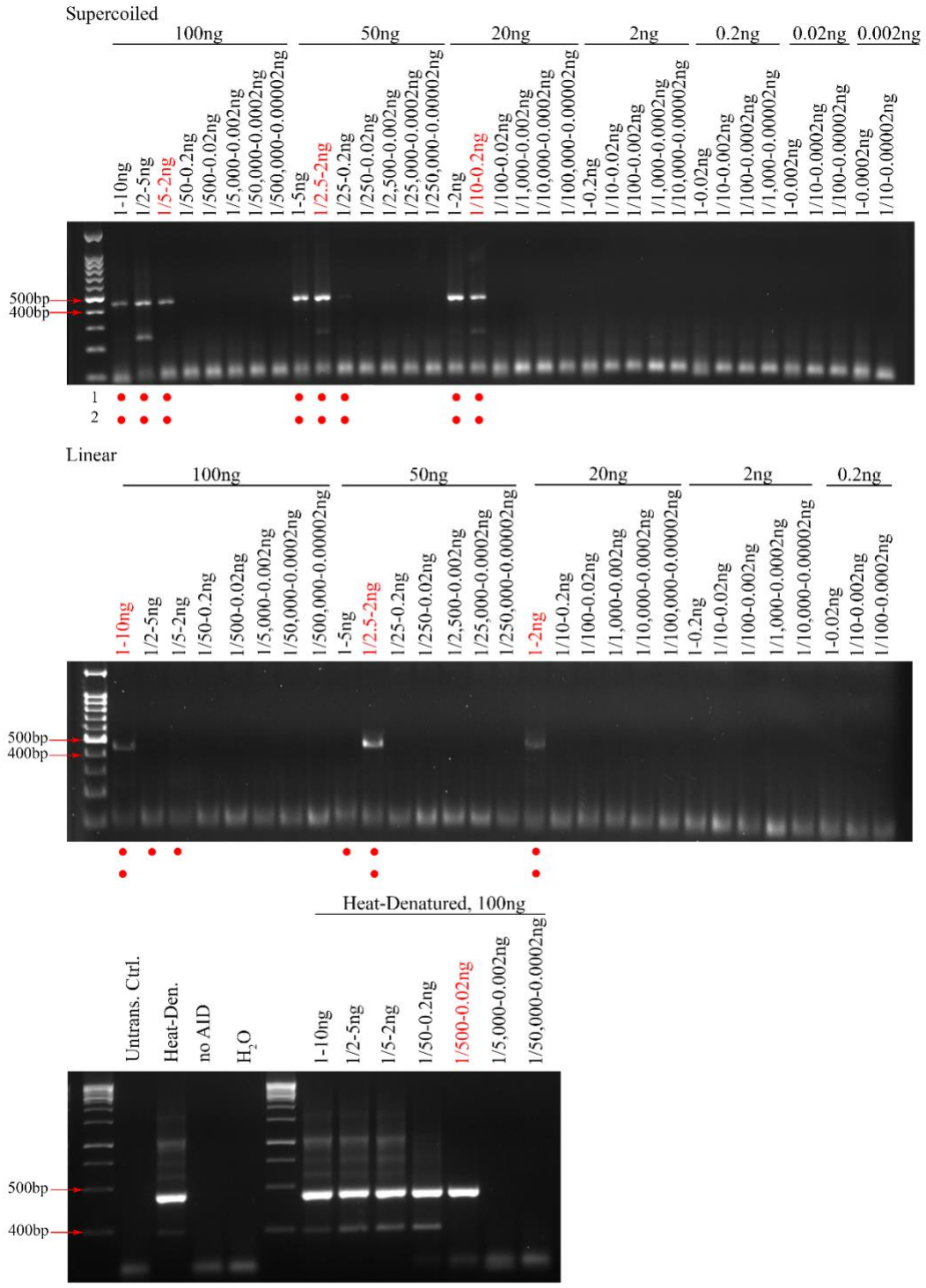


Figure 25: Deam-PCR Dilutions for Supercoiled and Relaxed Linear DNA treated with purified AID-His. His-tagged human AID was purified from HEK 293T cells. 27 ng of AID-His (4µl) mixed with 4 x 10^{-4} units of UGI was incubated with supercoiled and

relaxed linear DNA in the amounts: 100 ng, 50 ng, 20 ng, 2 ng, 0.2 ng, 0.02 ng and 0.002 ng. Each reaction was serially diluted, and 1 µl of each dilution was subjected to deam-PCR. The product size is 450 bp. 50 ng of the supercoiled and linear DNA was incubated with HEK 293T cell lysate not transfected with AID (Untrans. Ctrl.) and dialysis buffer (no AID) to show that no DNA is amplified in the absence of AID-mediated C-T mutations. Both the supercoiled and relaxed linear DNA were also heat-denatured prior to incubation with AID and then serially diluted to show the band intensity when a ssDNA substrate is present. The example above shows heat-denatured dilutions from the relaxed linear DNA. The result above is representative of 2 deam-PCRs from 2 independent reactions for both the supercoiled and relaxed linear topologies. The red dots represent the bands observed for each of the independent PCRs. The PCR bands persisted for approximately one more dilution for each reaction in the supercoiled set than the linear, indicating that AID has an approximately 10-fold preference for supercoiled dsDNA.

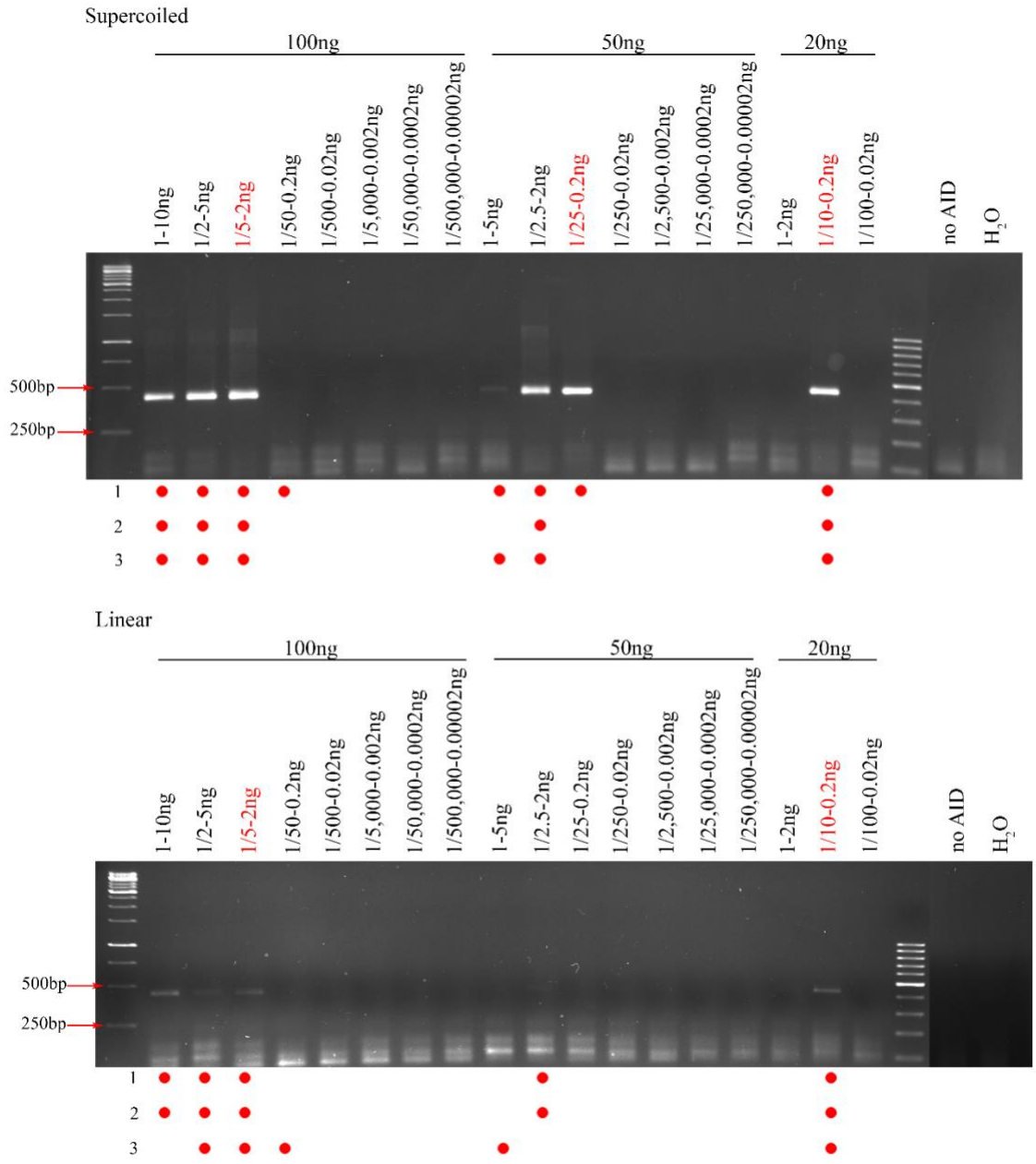


Figure 26: Deam-PCR Dilutions for Supercoiled and Relaxed Linear DNA treated with AID-His Lysate. His-tagged human AID was transfected HEK 293T cells, which were later lysed via French pressure cell press. 4 µl of the lysate was mixed with 4 x 10^{-4} units of UGI and incubated with supercoiled and relaxed linear DNA in the amounts: 100 ng, 50 ng, 20 ng, 2 ng, 0.2 ng, 0.02 ng and 0.002 ng. Each reaction was serially diluted, and 1 µl of each dilution was subjected to deam-PCR. The product size is 450 bp. 50 ng of the supercoiled and relaxed linear DNA was incubated with dialysis buffer (no AID) to show that no DNA is amplified in the absence of AID-mediated C-T mutations. The result

above is representative of 3 deam-PCRs from 3 independent reactions for both the supercoiled and relaxed linear topologies. The red dots represent the bands observed for each of the independent PCRs. Although the PCR bands persisted for nearly the same number of dilutions for both the supercoiled and relaxed linear templates, the bands representing mutated supercoiled DNA were on average approximately 2.8-fold brighter than those of the relaxed linear indicating that more mutated DNA was present in the reactions containing supercoiled substrate. The bottom halves of the two gels above were merged with the top halves to compare the controls with the experimental conditions.

3.11 GST-AID Mutated both Supercoiled and Relaxed Linear dsDNA in the Degen-PCR TIAA Assay

Having optimized our assay conditions by adding UGI (section 3.5.1) and using an appropriate amount of DNA (section 3.5.2, 3.5.3), we compared GST-AID activity on supercoiled and relaxed linear DNA topologies. We hypothesized that AID would preferentially mutate supercoiled DNA due to transiently open ssDNA bubble regions induced by the torsional constrains of the supercoils. As controls, we heat-denatured both conformations to generate fully single-stranded substrates. As expected, when heat-denatured, both the supercoiled and linear DNA were more highly mutated over their native double-stranded conformations since denaturation provides AID with a substantial amount of ssDNA to use as substrate (Table 7). The rate of AID-mediated C-T and G-A mutations was approximately 4-fold higher for the heat-denatured supercoiled than the native supercoiled, while the mutation rate for the heat-denatured linear was nearly 7-fold higher than the native linear DNA. Furthermore, the rate of AID-mediated mutations was 2.2-fold higher for the heat-denatured linear than the heat-denatured supercoiled. This is likely due to the linear DNA becoming more easily heat denatured and/or less easily re-annealed over the length of incubation time with AID, as compared to the supercoiled, which is constrained by the association of the two sister strands. Surprisingly, we found the linear and supercoiled DNA to be mutated at a nearly equal rate (Figure 27a), disagreeing with our deamination-specific PCR results that showed AID targets supercoiled DNA with a 10-100-fold preference (Figure 24, 25). AID has not previously been documented to be able

to mutate linear DNA, so we therefore further analyzed our results to look for patterns that might give more information about the biochemical activity of AID.

Our first question was: what makes AID mutate both the supercoiled and relaxed linear DNA nearly equally? To answer this question, we made mutation maps where the C-T and G-A mutations are distributed in the positions in which they occurred in the target DNA sequence (Figure 27b). The purpose of the mutation maps is to observe the overall pattern of mutation for all nucleotides analyzed, and to see if some regions are more highly targeted than others. We hypothesized that the supercoiled DNA would be mutated in patches representative of transient bubble regions, while the mutations in the relaxed linear DNA would be distributed randomly throughout the sequence. We found that the mutations were more dispersed than what we had expected, and no distinct clusters were observed in the supercoiled DNA. Furthermore, we noticed most of the mutations within the relaxed linear DNA were concentrated towards the 3' end of the template. We also noticed that unlike the heat-denatured templates which have roughly equal C-T and G-A mutations, under native conditions there were 2.3- and 9.9-fold more C-T mutations than G-A mutations in the supercoiled and linear DNA, respectively (Table 7, Figure 27b). In the absence of transcription, we did not expect to observe a strand bias since both strands should be equally exposed if transient ssDNA regions form. Next, we looked at the distribution of mutations throughout the amplicons (Figure 27c). Although the heat-denatured supercoiled and linear DNA have roughly the same pattern of mutation (Figure 27b), the ratio of mutated to unmutated amplicons differed (Figure 27c). 31.7% of the heat-denatured supercoiled amplicons were mutated, while 56.1% of the heat-denatured linear

amplicons were mutated. We expected the ratio of mutated to unmutated DNA to be the same for the heat-denatured substrates because they should be single-stranded upon exposure to AID. However, as mentioned above, the relaxed linear DNA may be more easily denatured than the supercoiled DNA providing a more readily accessible substrate for AID. The ratio of mutated to unmutated amplicons is lower in the native DNA than their heat-denatured counterparts, indicating that the total C-T and G-A mutations are located within 21.1% of the total amplicons for the supercoiled and 11.3% for the relaxed linear DNA substrate. The low mutated to unmutated amplicon ratio is likely associated with enzyme processivity, in which a small number of templates in a pool are highly mutated (Pham et al. 2003). We also noticed that the ratio of amplicons containing C-T mutations to amplicons containing G-A mutations is roughly equal across all substrates. This was expected for the heat-denatured substrates since there are roughly equal numbers of C-T and G-A mutations (Table 7, Figure 27b). Although there are roughly the same number of amplicons containing C-T and G-A mutations for the native supercoiled and relaxed linear templates, the amplicons containing C-T mutations are far more heavily mutated than those containing G-A mutations (Figure 27c). Overall, our hypothesis that AID would preferentially mutate supercoiled DNA over relaxed linear DNA was disproven because AID mutated both templates at a near equal frequency (Table 7, Figure 27a). Furthermore, the abundance of C-T mutations on the non-template strand may indicate strand preference even in the absence of transcription.

Table 7	Rate and Distribution of Mutations			
	GST-AID			
DNA Topology	Heat-Denatured Supercoiled	Supercoiled	Heat-Denatured Linear	Linear
C-T	82	74	163	89
G-A	105	32	110	9
Taq C-T Error Rate	6.46×10^{-5}	6.46×10^{-5}	6.46×10^{-5}	6.46×10^{-5}
Taq G-A Error Rate	5.02×10^{-5}	5.02×10^{-5}	5.02×10^{-5}	5.02×10^{-5}
Taq C-T Errors	4	8	3	6
Taq G-A Errors	3	6	2	5
Corrected C-T	78	66	160	83
Corrected G-A	102	26	108	4
Total Mutations	180	106	268	87
Nucleotides Analyzed	63,950	125,405	44,277	99,853
Overall Mutation Rate	2.81×10^{-3}	7.30×10^{-4}	6.05×10^{-3}	8.67×10^{-4}
Total Amplicons	82	147	82	133
Wildtype Amplicons	56	116	36	118
Mutated Amplicons	26	31	46	15
Amplicons with C-T Mutations	11	20	23	7
Amplicons with G-A Mutations	15	11	23	8

Table 7: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations in Supercoiled and Relaxed Linear Template DNA. All error and mutation rates were determined as described in Table 1. The total number of amplicons analyzed is shown at the bottom. Amplicons without C-T or G-A mutations are denoted as wildtype amplicons, while mutated amplicons had either C-T or G-A mutations. The supercoiled and linear data in the table above is combined from 2 independent reactions.

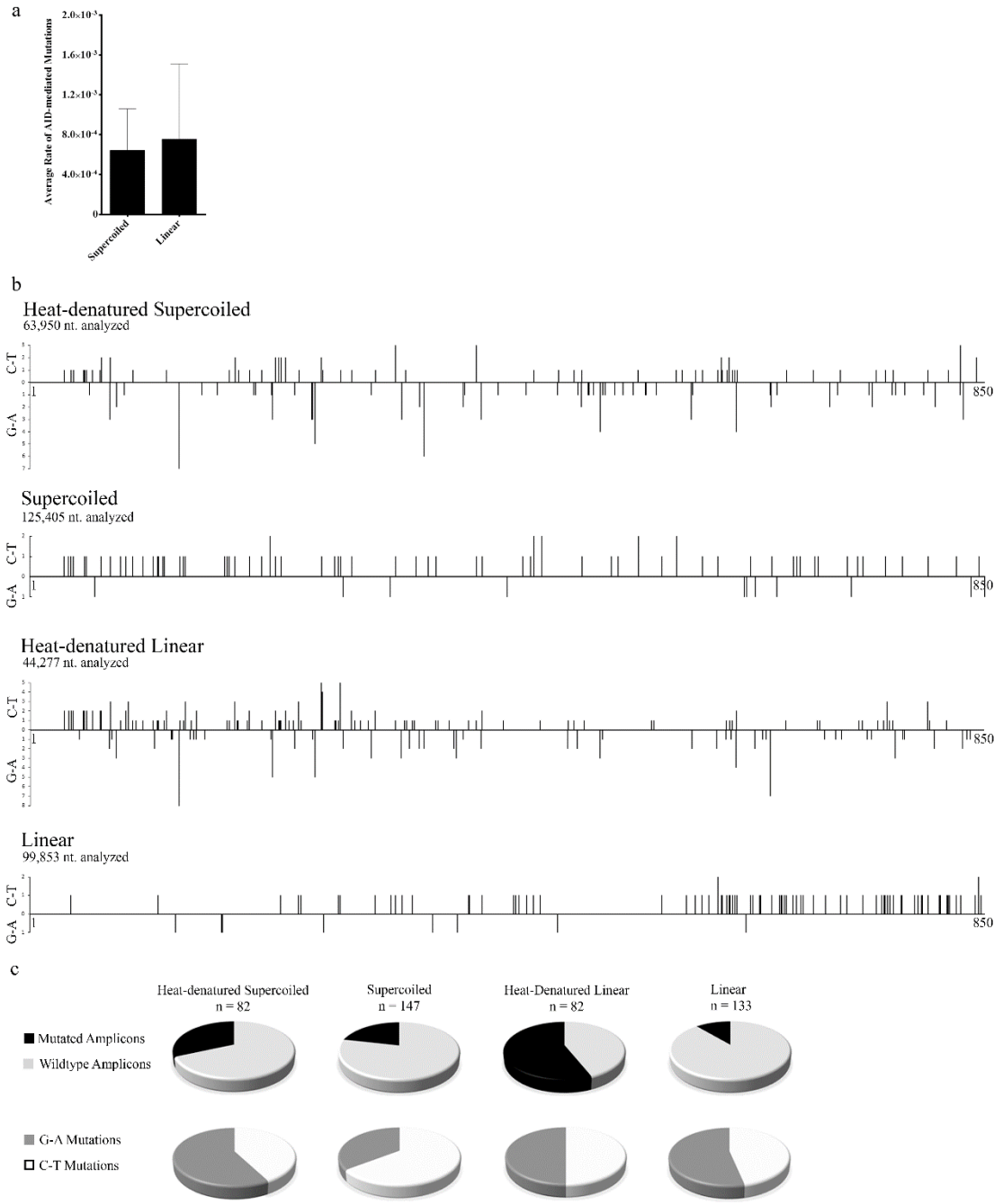


Figure 27: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations in Supercoiled and Relaxed Linear Template DNA. 1.2 µg of GST-AID (4 µl) was mixed with 4 x 10^{-4} units of UGI and incubated with 100 ng of supercoiled or relaxed linear substrate at 32℃ for 4 hours. Template DNA was subjected to degen-PCR and processed for sequencing. The data above was combined from two independent reactions for both the relaxed linear and supercoiled DNA. The heat-denatured data was from one reaction of supercoiled or relaxed linear DNA, and used as a positive control to show AID activity on

ssDNA. a) The average mutation rate was found by taking the sum of C-T and G-A mutations from each reaction condition and dividing them by the total number of nucleotides analyzed from that reaction. The error bars show the standard error of the mean (SEM). b) Each C-T and G-A mutation from the respective conditions was plotted on a line graph, where the x-axis denotes the position along the template DNA ranging from 1-850 nucleotides and the y-axis denotes the number of C-T or G-A mutations. C-T mutations are in the positive direction and G-A mutations are in the negative direction. c) The ratio of mutated to wildtype amplicons and amplicons with C-T or G-A mutations was plotted in pie charts, where "n" is the number of amplicons included in the analysis. The top row of pie charts show the ratio of mutated to wildtype amplicons, where mutated is shown in black and wildtype is shown in grey. The bottom row of pie charts show the ratio of amplicons containing C-T mutations to those with G-A mutations, where G-A mutations are shown in dark grey and C-T mutations are shown in white.

3.12 Using Bisulfite to Generate a Model of Breathing DNA

Our next question was: how does AID gain access to dsDNA without transcription? We hypothesized that AID gains access to single-stranded regions within dsDNA through DNA breathing, i.e. local denaturation and reannealing of small regions within the larger structure. Using fluorescence correlation spectroscopy in conditions similar to our *in vitro* incubation (i.e. 0.1M salt, pH 8, 37℃), bubbles 2-10 nucleotides in size were found to form in a 50 μs range (Altan-Bonnet et al. 2003). We therefore wanted to identify potential single-stranded regions within our substrate DNA during the 4-hour incubation. To map ssDNA regions, we used the chemical bisulfite, which deaminates dC to dU only within ssDNA (Shapiro et al. 1973). We are also interested in the pattern of mutation induced by bisulfite. We hypothesized that bisulfite would mutate in patches indicative of bubbles induced by breathing DNA, and that this pattern would be more pronounced in supercoiled than in relaxed linear DNA due to the higher torsional strain arising from its topology (Figure 8). We also hypothesized that bisulfite would not exhibit any strand preference since being a chemical deaminase, we should see a roughly equal number of C-T and G-A mutations.

We incubated bisulfite with our supercoiled and relaxed linear substrates and then subjected the substrate DNA to degen-PCR. Heat-denatured supercoiled was used as an ssDNA control, as bisulfite only mutates dC within ssDNA. We found that the mutation rate was 10.6-fold higher in the heat-denatured supercoiled than the native supercoiled, and the mutation rate was 1.5-fold higher in the native supercoiled than the relaxed linear DNA (Table 8). These rates were as expected based on the hypothesized amount of liberated

ssDNA due to DNA structure and topology (Figure 8). Moreover, the heat-denatured supercoiled DNA was mutated in a pattern similar to that produced by AID in which mutations span the entire length of the DNA substrate (Figure. 28a). However, the bisulfite mutation pattern differed from the AID mutation pattern in the native supercoiled and relaxed linear DNA templates. Bisulfite mutated both the supercoiled and relaxed linear substrates in a similar manner, where certain regions are more highly mutated than others representing "patches" of ssDNA as we expected. These highly mutated patches are likely in regions that have higher breathing rates either due to differing primary sequence (i.e. A-T vs G-C content) or secondary structure (i.e. ability to form bubbles, hairpins, cruciforms). In contrast, AID mutated both supercoiled and relaxed linear DNA with no apparent focus on any particular area of the DNA template (Figure 27). It is possible that AID takes advantage of DNA breathing as a way to gain access to an otherwise double-stranded substrate, but its activity is not solely restricted to highly breathing areas. Furthermore, unlike AID which mutated a small percentage of the total amplicons (~50% or less, Section 3.6), bisulfite mutated 100% of the heat-denatured amplicons and over 80% of the supercoiled and linear amplicons. Moreover, there was no strand preference found as an equal number of amplicons containing C-T and G-A mutations for the supercoiled and relaxed linear topologies were obtained as we expected. Overall, the bisulfite data shows us the regions that are single-stranded in each form of substrate (i.e. heat-denatured, supercoiled, relaxed linear). These ssDNA regions should be accessible to AID, either through complete melting of the base pairs in the case of the heat-denatured DNA or

through DNA breathing and/or formation of secondary structure in the double-stranded supercoiled and relaxed linear substrates.

Table 8 Rate and Distribution of Mutations			
	Bisulfite		
DNA Topology	Heat-Denatured Supercoiled	Supercoiled	Linear
C-T	509	101	74
G-A	332	110	47
Taq C-T Error Rate	6.46×10^{-5}	6.46×10^{-5}	6.46×10^{-5}
Taq G-A Error Rate	5.02×10^{-5}	5.02×10^{-5}	5.02×10^{-5}
Taq C-T Errors	1	3	3
Taq G-A Errors	1	3	2
Corrected C-T	508	98	71
Corrected G-A	331	107	45
Total mutations	839	205	116
Bases Sequenced	19,417	50,554	42,253
Mutation Rate	4.32×10^{-2}	4.06×10^{-3}	2.75×10^{-3}
Total Amplicons	24	62	60
Wildtype Amplicons	0	10	11
Mutated Amplicons	24	52	49
Amplicons with C-T Mutations	16	26	27
Amplicons with G-A Mutations	8	26	22

Table 8: Rate and Distribution of Bisulfite-mediated C-T and G-A Mutations in Supercoiled and Relaxed Linear Template DNA. All error and mutation rates were determined as described in Table 1. The data in the table above was collected from one reaction for each condition.

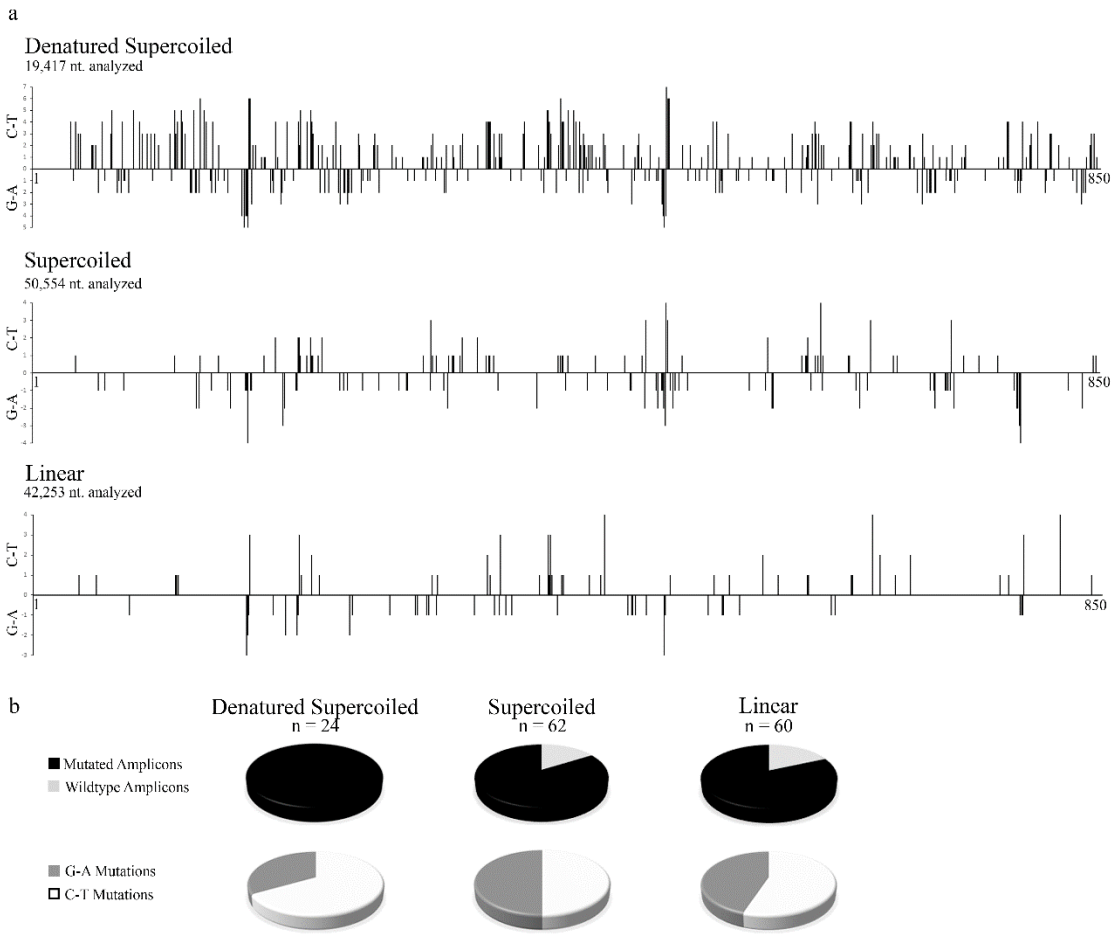


Figure 28: Distribution of Bisulfite-mediated C-T and G-A Mutations in Supercoiled and Relaxed Linear Template DNA. Bisulfite was incubated with template DNA under denaturing or native conditions (32℃, 4 hours). Template DNA was subjected to degen-PCR and processed for sequencing. The data above was collected from one reaction for each condition. a) Each C-T and G-A mutation from the respective conditions was plotted on a line graph, where the x-axis denotes the position along the template DNA ranging from 1-850 nucleotides and the y-axis denotes the number of C-T or G-A mutations. C-T mutations are in the positive direction and G-A mutations are in the negative direction. c) The ratio of mutated to wildtype amplicons and amplicons with C-T or G-A mutations was plotted in pie charts, where "n" is the number of amplicons included in the analysis. The top row of pie charts show the ratio of mutated to wildtype amplicons, where mutated is shown in black and wildtype is show in grey. The bottom row of pie charts show the ratio of amplicons containing C-T mutations to those with G-A mutations, where G-A mutations are shown in dark grey and C-T mutations are shown in white.

3.13 AID can act in both a Processive and Distributive Manner

We and others have previously described AID's activity as processive by which AID can bind and heavily mutate one strand of ssDNA as opposed to dissociating into solution to find another substrate (Pham et al. 2003, Larijani and Martin 2007, Larijani et al. 2007, reviewed in Larijani and Martin 2012). AID activity has also been described as being distributive, in which it binds and mutates each encountered substrate only once (Coker and Petersen-Mahrt 2007). Since the assays employed in the previous literature tested AID activity on either purely ssDNA or oligonucleotide bubble substrates (Pham et al. 2003, Coker and Petersen-Mahrt 2007, Larijani and Martin 2007, Larijani et al. 2007), we sought to delineate how AID acts on dsDNA of longer length. To determine if the pattern of activity is unique to AID, or simply dependent on ssDNA accessibility within dsDNA, we wanted to compare the pattern of mutation of AID to that of bisulfite. To do this, the data compiled from all amplicons in Figures 27b and 28a was disaggregated, allowing us to see the pattern of mutation for each individual amplicon (Figure 29). Although the target DNA sequence is 1.2kb in length, only up to 850bp were analyzed in the mutation maps shown because the amplicons were unidirectionally sequenced in the forward direction and data beyond 850bp was not obtained for all amplicons.

We found that the pattern of activity for bisulfite (Figure 29 top) and for GST-AID (Figure 29 bottom) was very similar on heat-denatured supercoiled substrate (Figure 29 left), in which nearly all mutated amplicons had multiple mutations with regions of closely spaced C-T or G-A mutations. However, a major difference was that bisulfite mutated

100% of the heat-denatured amplicons, while GST-AID only mutated 31.7%. The ability of AID to mutate only a fraction of the available substrate is likely due to the processive but lethargic nature of the enzyme (Goodman, 2016, Larijani et al. 2007, King et al. 2015). Bisulfite mutated the native supercoiled and relaxed linear substrates in a similar manner where most of the mutations are dispersed throughout the amplicons, while a few regions were more heavily mutated in 2-11 nucleotide patches (Figure 29 top middle and right). Heavily mutated patches were more obvious in the supercoiled amplicons than the linear (e.g. supercoiled amplicons 41, 47, 62, Figure 29 top middle), which is expected due to the topology of supercoiled DNA (Figure 8). GST-AID also mutated the native supercoiled and linear substrates in a similar way (Figure 29 bottom middle and right), but in a different pattern than that of bisulfite. On the native supercoiled substrate, 34 of 106 C-T and G-A mutations were dispersed throughout 28 of 31 mutated amplicons, while 3 amplicons contained the remaining 72 C-T or G-A mutations (Figure 29 bottom middle). Thus, 67.9% of the total mutations were contained within 2% of the total amplicons. On the relaxed linear substrate, 15 of 89 C-T and G-A mutations were dispersed throughout 14 of the 15 mutated amplicons, while 1 amplicon contained 83 closely spaced mutations (Figure 29 bottom right). Thus, 84.7% of the total mutations were contained within 0.75% of the total amplicons. Overall, our data show that AID can act in both a distributive and processive manner on dsDNA. However, the mechanism by which AID targets DNA is still unclear. Bisulfite allows us to map regions that may be available to AID during the 4-hour incubation period due to transient DNA breathing. While this may explain how AID initially gains access to its DNA substrate, it is not sufficient to explain the “processive”

activity in which AID can mutate up to 97 closely spaced nucleotides. To gain further insight into the biochemical properties of AID, our next goal was to look into the role of primary sequence and secondary structure of our DNA substrate.

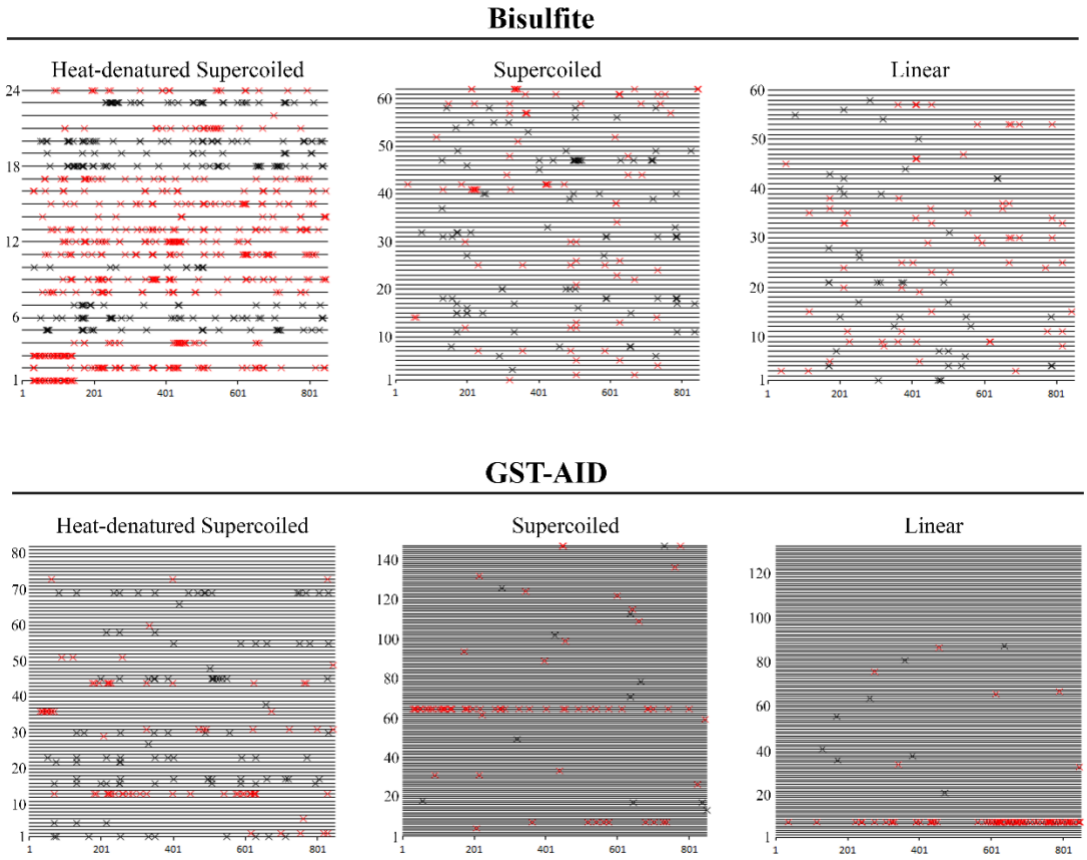


Figure 29: Mutation Maps for all Individual Amplicons Incubated with either Bisulfite or GST-AID. The data in Figures 27b and 28a was disaggregated such that the C-T or G-A mutations from each individual amplicon could be examined. The x-axis denotes the position in the target DNA sequence spanning from 1-850 nucleotides, while the y-axis denotes the number of amplicons analyzed. The red x's represent C-T mutations, while the black x's represent G-A mutations.

3.14 DNA Secondary Structure is more Important than Primary Sequence in AID-targeting

It is well documented in the literature that AID preferentially mutates at 5'-WRC (W=A/T, R=G/C) hotspot motifs (Bransteitter et al. 2003, Pham et al. 2003, Bransteitter et al. 2004, Yu et al. 2004, Larijani et al. 2005, Larijani and Martin 2007, Larijani et al. 2007, Brar et al. 2008, MacCarthy et al. 2009), and mutates at 5'-SYC (S=C/G, Y=C/T) cold spots much less frequently (Pham et al. 2003, Bransteitter et al. 2004, Larijani et al. 2005). However, it has also been shown the more important factor for determining AID targeting is the type of ssDNA presented to the enzyme. AID prefers 5-7 nt. long bubbles over stem-loop substrates or fully single-stranded DNA (Larijani et al. 2007). However, these studies were all based on the use of short oligonucleotide substrates. To determine the importance of primary sequence and secondary structure on our longer substrates, we analyzed the percentage of mutated cytidines for all possible trinucleotide motifs ending in C (Figure 30). GST-AID data was compared to that of bisulfite to compare its enzymatic sequence preference to a chemical deaminase that does not prefer any particular sequence.

We found that, on supercoiled DNA, GST-AID has a slight preference towards AGC hotspot motifs, but no strong preference for any other motif (Figure 30a). On relaxed linear DNA, GST-AID does not preferentially mutate any of the WRC hotspots, but it does have activity peaks on CAC and CGC. The GST-AID data in Figure 30a is in contrast to that of Larijani and colleagues in 2005 when they described the mutability index for GST-AID, AID-His, Ramos cells and *ung*$^{-/-}$ *msh2*$^{-/-}$ mice (Larijani et al. 2005). They found that, across all conditions, there was a strong preference for the four 5'-WRC hotspot motifs

AGC, TAC, AAC and TGC and the mutability index continuously lowered as it approached the 5'-SYC cold spots. On supercoiled DNA bisulfite activity peaks highest on AGC and GGC, while on relaxed linear DNA bisulfite activity peaks highest on TAC, CAC, GGC and CCC (Figure 30a). The sequence preference for GST-AID on native supercoiled and linear substrates is very similar to that of bisulfite (Figure 30a). However, on heat-denatured supercoiled and linear substrates GST-AID mutates most frequently at the WRC hotspots AGC and TAC (Figure 30b), and the overall pattern of mutation diverts away from that of bisulfite and becomes more similar to that found by Larijani and colleagues (Larijani et al. 2005). On denatured supercoiled substrate bisulfite activity peaks on CCC, but it does not have preference towards any other motif (Figure 30b). Since both AID and bisulfite can only mutate ssDNA regions liberated within double-stranded DNA, the difference in trinucleotide motif preference between the native and denatured substrates may be due to the accessibility of cytidines to either bisulfite or AID. One way that ssDNA may be liberated in otherwise dsDNA is through the formation of secondary structures such as hairpins, cruciforms and stem-loop structures. Interestingly, bisulfite activity peaked on native supercoiled and linear substrates when the mutated C was preceded by either GG or CC dinucleotides, which have been shown through experimental and simulation data to have unstable dinucleotide stacking interactions (Alexandrov et al. 2009). In dsDNA, the primary sequence determines the secondary structure, which then determines the areas that may be exposed as ssDNA. Therefore, in dsDNA sequences of long length secondary structure may be more important as a determinant of AID targeting than primary sequence alone.

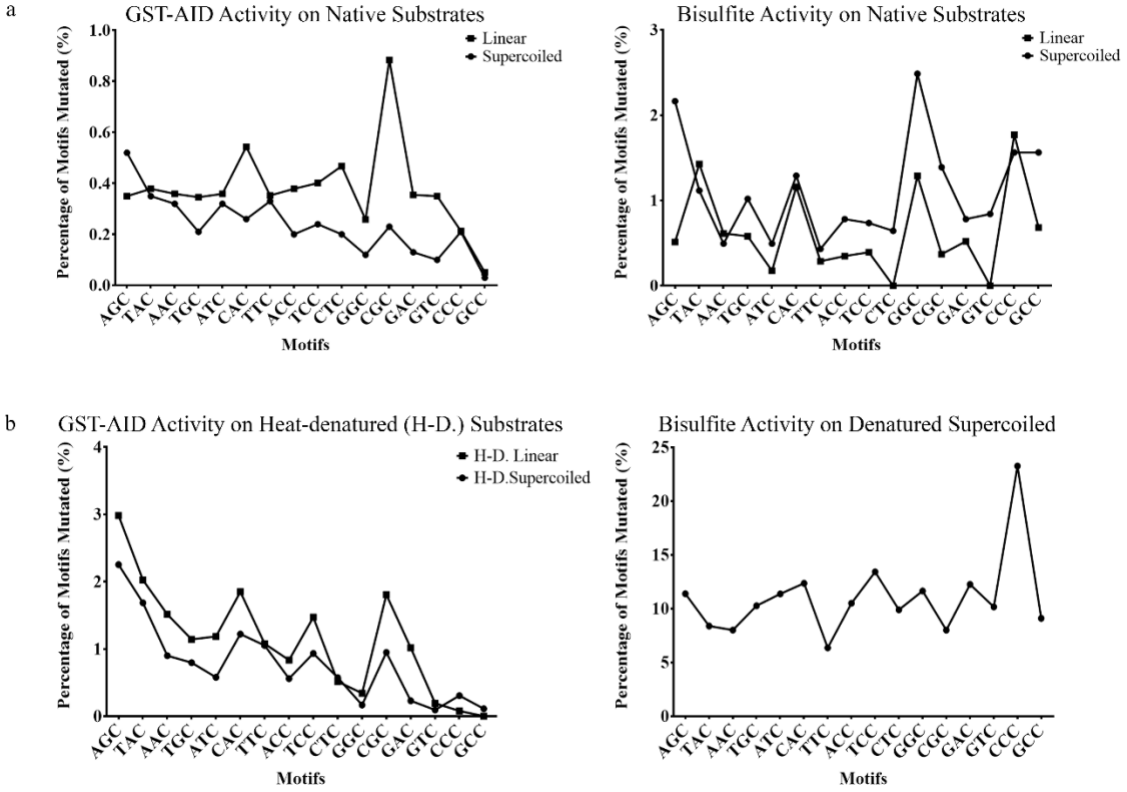


Figure 30: Analysis of Primary Sequence of GST-AID and Bisulfite-mediated Mutations. The trinucleotide motif for each GST-AID- or bisulfite-mediated C-T or G-A mutation was recorded. G-A mutations were considered as C-T mutations on the opposite (bottom) strand, so the reverse complement of the GXX (X= any nucleotide) trinucleotide motif was considered. The motifs are along the x-axis with the four WRC (W= A/T, R= A/G) hotspot motifs at the beginning. The percentage of motifs mutated (y-axis) was found by taking the number of C-T and G-A mutations at a particular motif and dividing it by the product of the number of that motif within the target DNA sequence and the number of amplicons analyzed. a) Motif data for GST-AID and bisulfite on native supercoiled and linear substrate DNA. b) Motif data for GST-AID and bisulfite on denatured substrates.

To look further into the role of secondary structure, we used the DNA-secondary structure prediction software 'mfold' to analyze the sequence of our DNA substrate (Zuker 2003). Mfold considers the input DNA to be single-stranded in nature, and therefore calculates the sequence's ability to form secondary structures with itself. The caveat is that it does not take into account DNA topology, as the input DNA is considered to be completely single-stranded. However, it still allows us to gain a picture of the secondary structures that have the potential to form and thus the areas that are more energetically likely to become single-stranded by breathing. The secondary structure was modelled at 37℃ and in 100 mM salt, conditions very similar to our assay. The window size was set at 25 nucleotides and the folding was limited such that only bases within 50 nucleotides of each other could bind. Under these constraints, mfold allows us to visualize the secondary structures that could potentially form during DNA breathing as only bases in close proximity could interact. The overall ΔG of the secondary structure produced by mfold is -71.18 kcal/mole, meaning that short regions of the template DNA sequence are able to spontaneously anneal at 37℃. The overall ΔG represents the sum of all the free energies assigned to each predicted secondary structure and base pair stacks (Zuker 2003). The predicted secondary structure is shown in Figure 31.

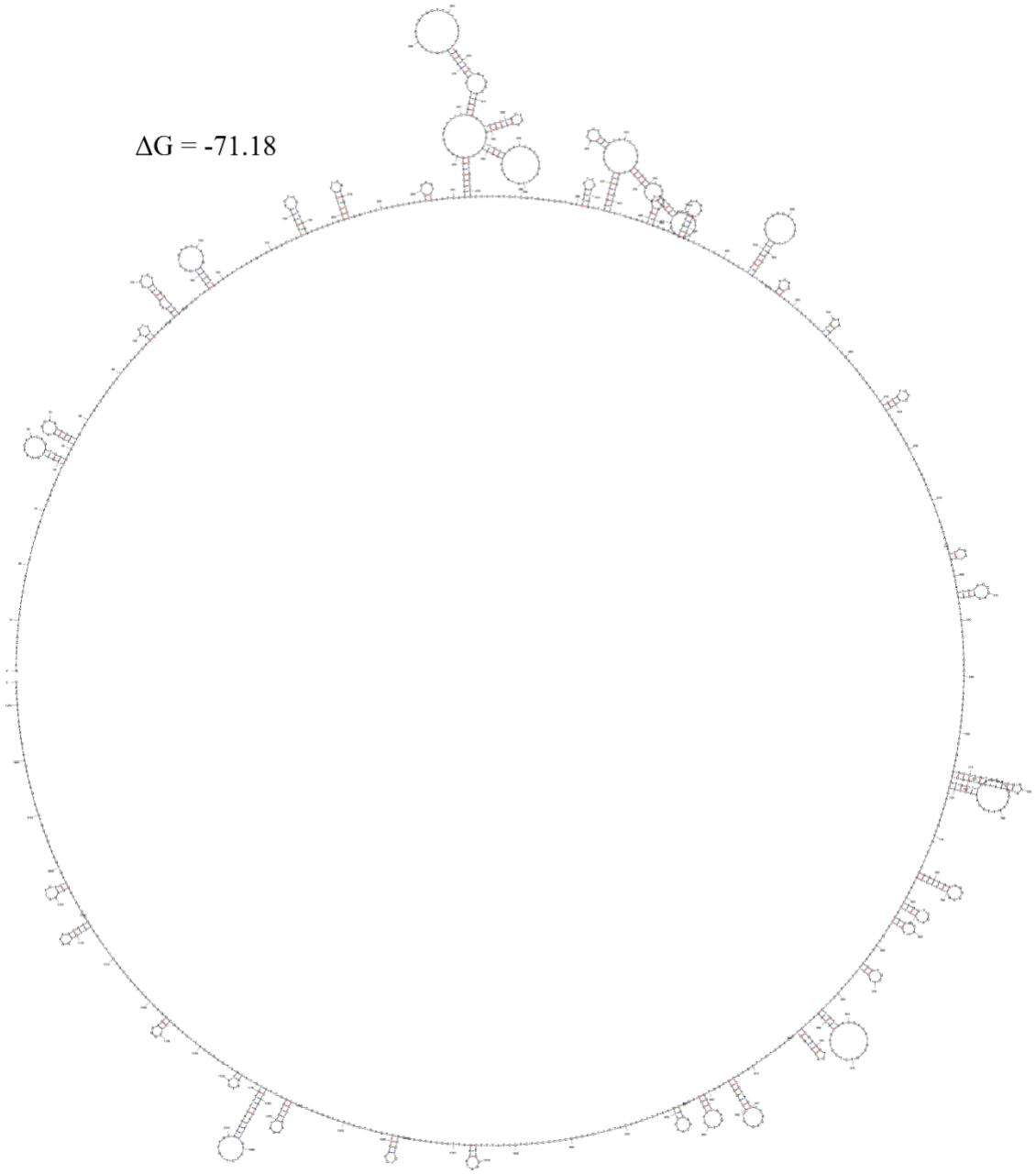


Figure 31: Secondary Structure of the Target DNA Sequence. The secondary sequence for the top strand of our target DNA substrate was modelled using the online-based DNA-folding software 'mfold'. The overall ΔG of the secondary structure shown is -71.18 kcal/mole.

The bisulfite (Figure 32) and GST-AID (Figure 33) C-T mutations from the supercoiled set were superimposed onto the structure predicted by mfold. The majority (71%) of the bisulfite-mediated C-T mutations were found within the hairpin and stem-loop structures predicted by mfold, with 43% in the stem region and 28% in the loop region (Table 9). Similarly, the majority (66%) of the GST-AID-mediated C-T mutations were found within the hairpin and stem-loop structures predicted by mfold, with 37% in the stem region and 29% in the loop region. Our initial expectation was that more mutations would fall in the loop regions since they are composed fully of ssDNA (unpaired), while the GC-rich stem regions would remain as dsDNA (paired) and inaccessible to either bisulfite or AID. Mfold provides a "p-num" value which is a measure of confidence of the paired and unpaired regions. If a loop region has a low p-num value, then there is high confidence that the region is unpaired. Similarly, if a paired region (e.g. stem region) has a low p-num value then there is high confidence that the region is paired. We plotted the bisulfite (Figure 34a) and GST-AID (Figure 34b)-mediated C-T mutations that were predicted to be within the stem or loop regions of the secondary structures (Figure 32, 33) against the p-num values of the position that the mutation occurred. The majority of both bisulfite- and AID-mediated mutations within loop regions were at positions that had a high confidence of being open. Mutations within stem regions tended to be at positions that had a lower confidence of being paired, meaning that these regions have a higher probability of fluctuating between an unpaired or paired state, consistent with the expected notion that torsional pull at either junctional end of the stem regions make them more susceptible to breathing. This high probability of breathing in the GC-rich stem regions corresponds to

the peaks in activity at CGC for GST-AID as well as GGC and CCC for bisulfite (Figure 30). Overall, our preliminary modeling data suggests that secondary structure destabilizes DNA liberating transient single-stranded regions, and likely plays a greater role than primary sequence alone in AID targeting.

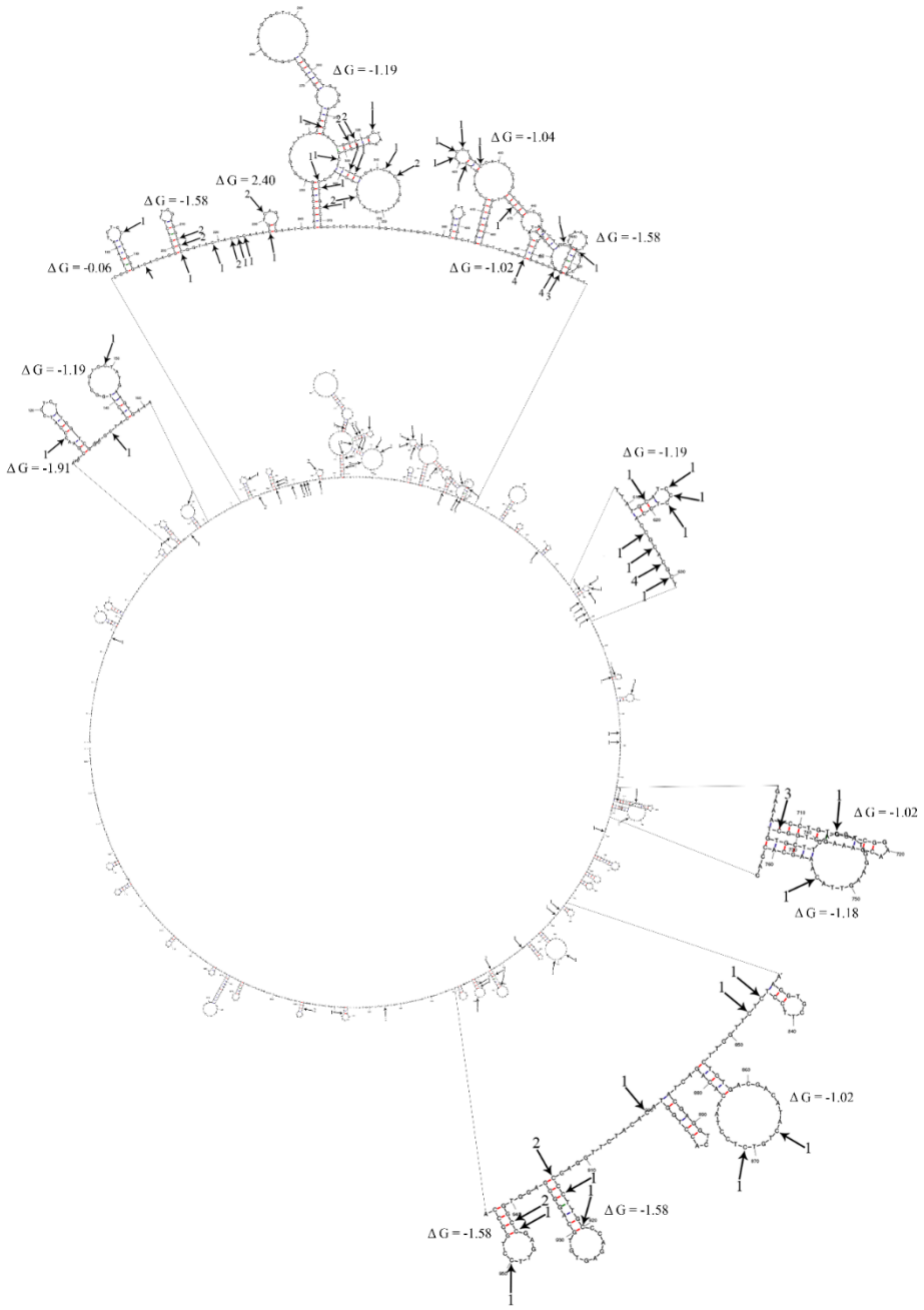


Figure 32: Bisulfite-mediated C-T Mutations on Supercoiled Substrate Superimposed onto the Template DNA Secondary Structure. Bisulfite-generated C-T mutations from all analyzed amplicons were superimposed onto the proposed template DNA secondary structure to allow visualization of the mutation location along the top strand. C-T mutations are indicated with arrows, and the numbers next to each arrow correspond to the number of mutations at that position. Highly mutated regions are enlarged and the corresponding ΔG's for the enlarged structures are shown.

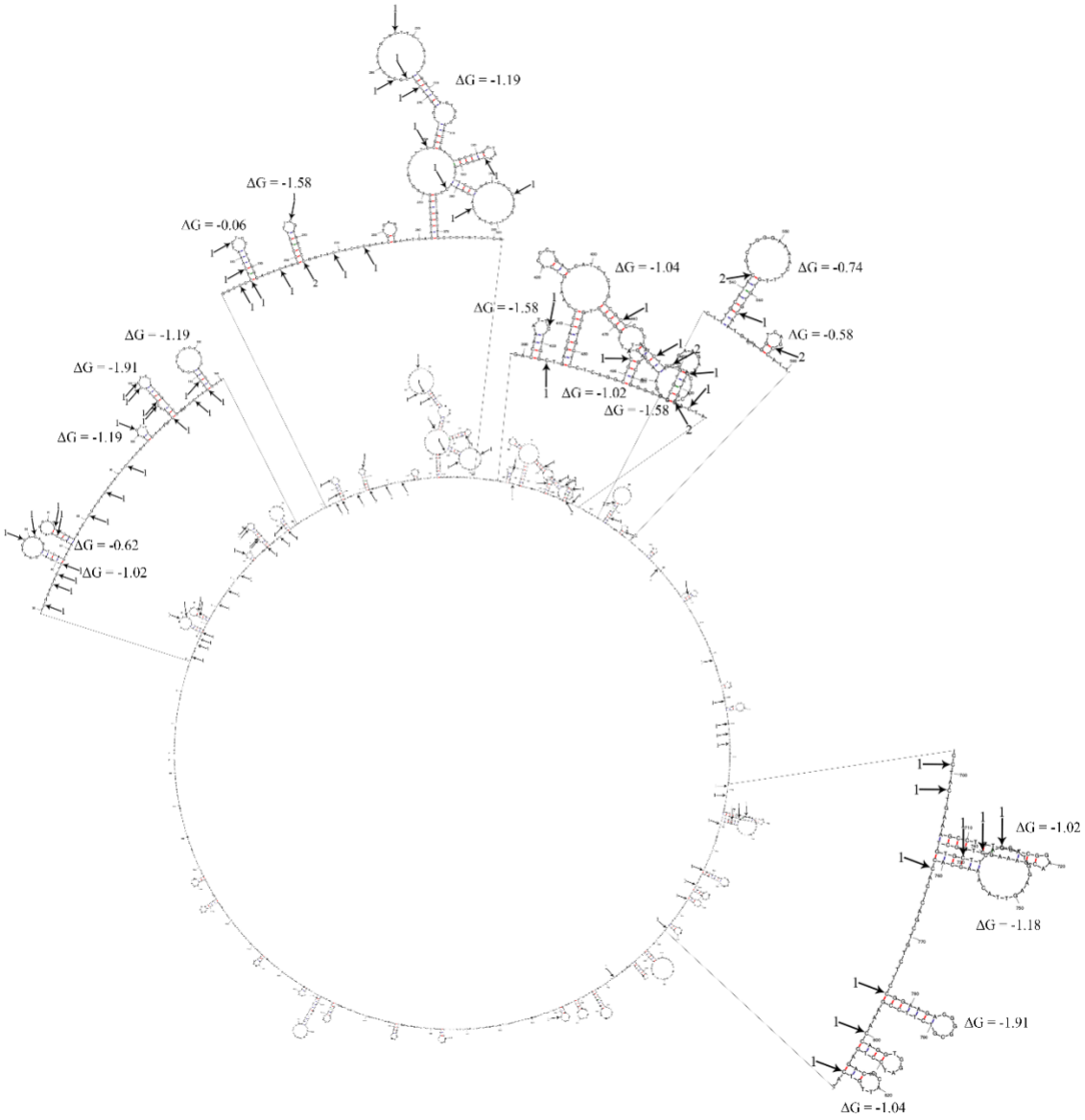


Figure 33: GST-AID-mediated C-T Mutations on Supercoiled Substrate Superimposed onto the Template DNA Secondary Structure. GST-AID-generated C-T mutations from all analyzed amplicons were superimposed onto the proposed template DNA secondary structure to allow visualization of the mutation location along the top strand. C-T mutations are indicated with arrows, and the numbers next to each arrow correspond to the number of mutations at that position. Highly mutated regions are enlarged and the corresponding ΔG's for the enlarged structures are shown.

Table 9	Secondary Structure		Outside
	Stem	Loop	
Bisulfite-mediated C-T Mutations	43	28	29
Percentage (%)	43%	28%	29%
GST-AID-mediated C-T Mutations	27	21	25
Percentage (%)	37%	29%	34%

Table 9: Number and Percentage of Bisulfite- and GST-AID-mediated C-T Mutations within the Predicted Target DNA Secondary Structure. The mutations pictured in Figures 32 and 33 are tabulated above. 71% of the bisulfite-mediated C-T mutations are predicted to be in either the stem or loop portion of the stem-loops in Figure 32. 66% of the GST-AID-mediated C-T mutations are predicted to be in either the stem or loop portion of the stem-loops in Figure 33.

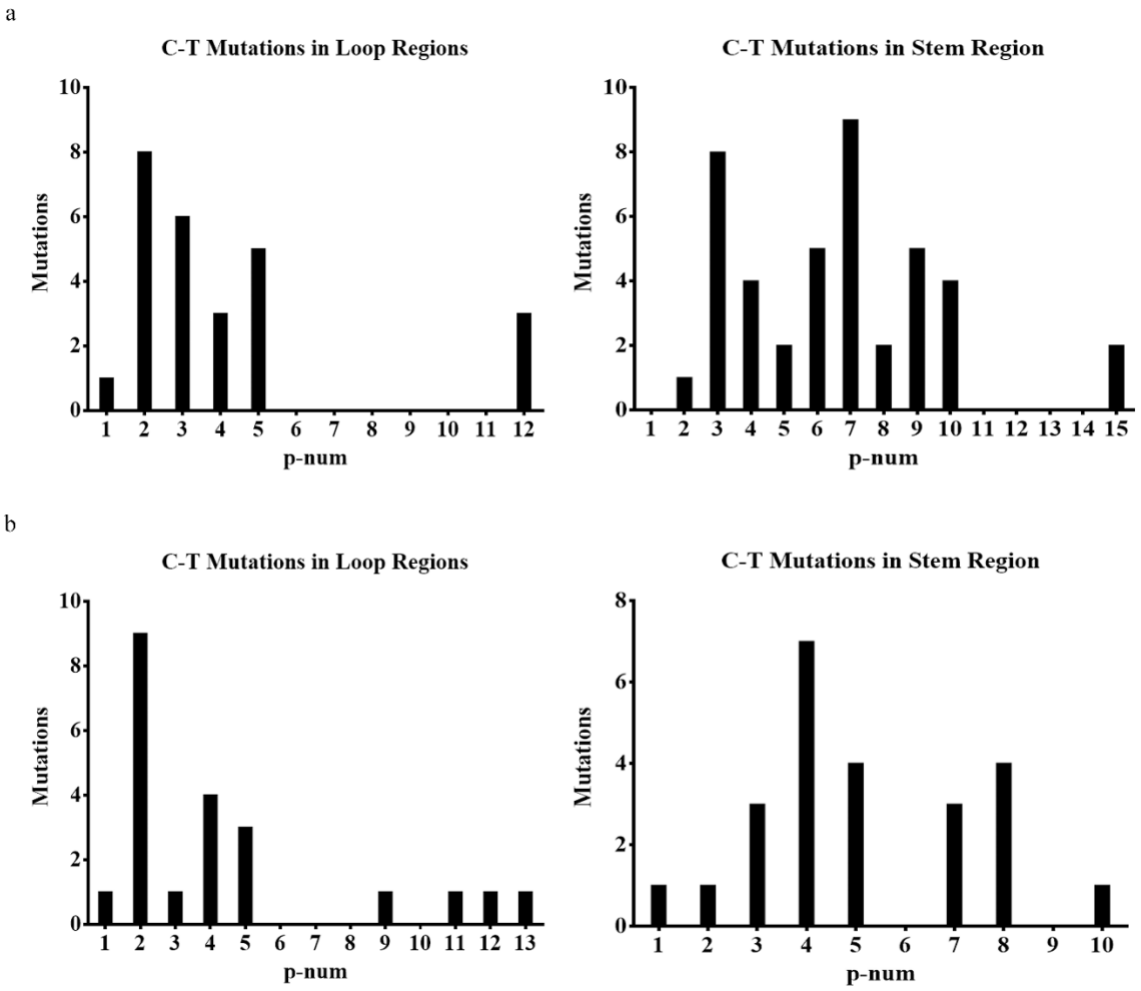


Figure 34: Stability of Paired or Unpaired Regions within Predicted Secondary Structures. The mfold software measured the confidence of the paired and unpaired regions of the predicted target DNA structure (Figure 31), termed p-num. The bisulfite (a) and GST-AID (b) mutations within the stem or loop regions of the secondary structures (Figures 32, 33) were plotted against the p-num value of the position that the mutation occurred. The majority of both bisulfite- and AID-mediated mutations within loop regions were at positions that had a high confidence of being open. Mutations within stem regions tended to be at positions that had a lower confidence of being paired.

3.15 Transcription Increases AID's Accessibility to Target DNA

Thus far we have found that AID can act on both supercoiled and relaxed linear DNA *in vitro*, and we have proposed that this activity is likely due to targeting of secondary structures made available during DNA breathing. It is thought that the purpose of transcription in AID targeting is to directly generate ssDNA through unpairing of sister-strands by the traversing RNA polymerase, and also to induce the formation of ssDNA secondary structures such as stem loops in the wake of the RNA polymerase, due to unwinding of supercoiling (Sohail et al. 2003, Shen et al. 2005, Canugovi et al. 2009, Shen et al. 2009). In these reports, AID activity was analyzed during *in vitro* transcription of a target construct by T7 RNAP. Although these assays are well-established as optimal AID targeting assays, they also rely on AID-mediated regeneration of antibiotic resistance (Sohail et al. 2003, Shen et al. 2005, Canugovi et al. 2009, Shen et al. 2009), as discussed in Sections 1.8 and 3.1. Therefore, we wanted to use *in vitro* transcription by T7 RNAP with our degen-PCR assay to compare the rate of transcription independent AID activity to the AID activity in the presence of transcription. To ensure that we would be able to measure the maximum possible rate of transcription-dependent AID activity, we used varying rates of transcription in our assay system.

We hypothesized that varying the rate of transcription would vary the availability of transcription-generated secondary structures to AID. We surmised that: 1) slowing down transcription to an "optimal speed" for AID activity would allow secondary structures to persist for a sufficient time to maximize AID activity in the 4-hour period of incubation, 2) that slowing transcription rates below the "optimal speed" would have the same effect

on AID activity as no transcription, and 3) speeding up transcription beyond the optimal speed would result in lowered AID activity since DNA secondary structures may not persist for sufficient time to support efficient mutation. We incubated supercoiled template DNA with GST-AID and with T7 polymerase *in vitro* for 4 hours. The rate of transcription by T7 polymerase can be altered by varying rNTP concentration (Ambion 2012). We chose to vary the rate of transcription by decreasing the concentration of either UTP or all rNTPs. Varying [UTP] will cause T7 RNAP to temporarily stall at dA until an UTP comes into the active site, while varying all rNTPs will cause an overall slower rate of transcription since T7 polymerase temporarily stalls at each base. RNaseA was added to the reaction to determine AID activity on the DNA target without a strand bias due to formation of RNA/DNA hybrids, as well as to ensure that AID does not bind to the RNA due to its high positive surface charge (+14) (King and Larijani 2017). Two reactions +/- T7 RNAP were used as a control for transcription to show that *in vitro* transcription is working and that RNA is being produced in the presence of T7 RNAP (Figure 35). We first did a preliminary test including three rates of transcription: 1) full speed transcription (1/1 UTP, 1/1 rNTPs), 2) $1/50^{th}$ UTP, and 3) $1/10^{th}$ of full speed transcription (1/10 UTP, 1/10 rNTPs). We compared AID activity at these rates to a no T7 RNAP condition, which was the average of three independent TIAA assay reactions using GST-AID and supercoiled substrate without transcription. We found that full-speed transcription (1/1 UTP, 1/1 NTPs) decreased AID activity 1.6-fold over no transcription (Table 10, Figure 36a). However, as transcription slowed, AID activity increased, as the mutation rate was 1.79-fold higher in the 1/50 UTP 1/1 NTPs and 5.38-fold higher in the 1/10 UTP 1/10 NTPs speed than full-

speed transcription (Table 10, Figure 36a). Furthermore, the pattern of AID targeting began to change at the slowest speed of transcription (1/10 UTP, 1/10 NTPs). In the no transcription, 1/1 UTP 1/1 NTPs and 1/50 UTP 1/1 NTPs conditions C-T and G-A mutations were found in 27-34% of the total amplicons (Table 7, Table 11, Figure 26, Figure 36b). However, in the slowest speed of transcription (1/10 UTP, 1/10 NTPs) approximately 67% of the total amplicons have C-T or G-A mutations (Table 11, Figure 36b). Slowing the speed of transcription may keep the DNA structure in a more open, accessible form for AID leading to targeting of a higher number of DNA strands. Moreover, both DNA strands can be targeted with and without transcription as both C-T and G-A mutations were observed (Table 11, Figure 36b).

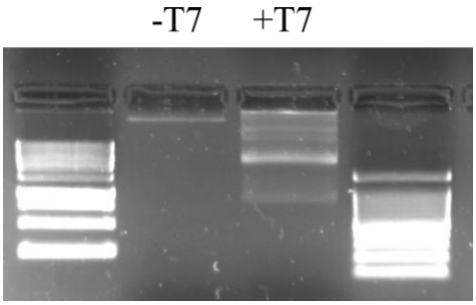


Figure 35: RNA is Produced only in the Presence of T7 RNAP. Two controls were incubated alongside of the experimental *in vitro* transcription reactions to show that RNA is indeed produced in the presence of T7 RNAP. Each reaction contained: 100ng of supercoiled substrate, 3.75mM of each rNTP, AID dialysis buffer in place of AID, and 4 x 10^{-4} units of UGI. Neither control contained RNaseA. The reaction in the first lane shows that there is no RNA production in the absence of T7 RNAP, while the send lane shows RNA production in the presence of T7 RNAP.

Table 10 Rate and Distribution of Mutations				
		GST-AID		
Transcription Condition	no T7	1/1 UTP, 1/1 NTPs	1/50 UTP, 1/1 NTPs	1/10 UTP, 1/10 NTPs
C-T	215	21	172	80
G-A	65	28	36	33
Taq C-T Error Rate	7.47×10^{-5}	7.47×10^{-5}	7.47×10^{-5}	7.47×10^{-5}
Taq G-A Error Rate	4.98×10^{-5}	4.98×10^{-5}	4.98×10^{-5}	4.98×10^{-5}
Taq C-T Errors	4	2	5	1
Taq G-A Errors	3	1	3	1
Corrected C-T	211	19	167	79
Corrected G-A	62	27	33	32
Total mutations	273	46	200	111
Nucleotides Analyzed	58,483	27,894	67,897	12,516
Overall Mutation Rate	4.67×10^{-3}	1.65×10^{-3}	2.95×10^{-3}	8.87×10^{-3}
Total Amplicons	73	34	89	15
Wildtype Amplicons	48	25	65	5
Mutated Amplicons	25	9	24	10
Amplicons with C-T Mutations	16	4	16	6
Amplicons with G-A Mutations	9	5	8	4

Table 10: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations during *in vitro* transcription (set 1). The rate of transcription was altered by altering the concentration of rUTP or all rNTPs in the reaction. All error and mutation rates were determined as described in Table 1. The data in the table above was collected from one reaction for each condition.

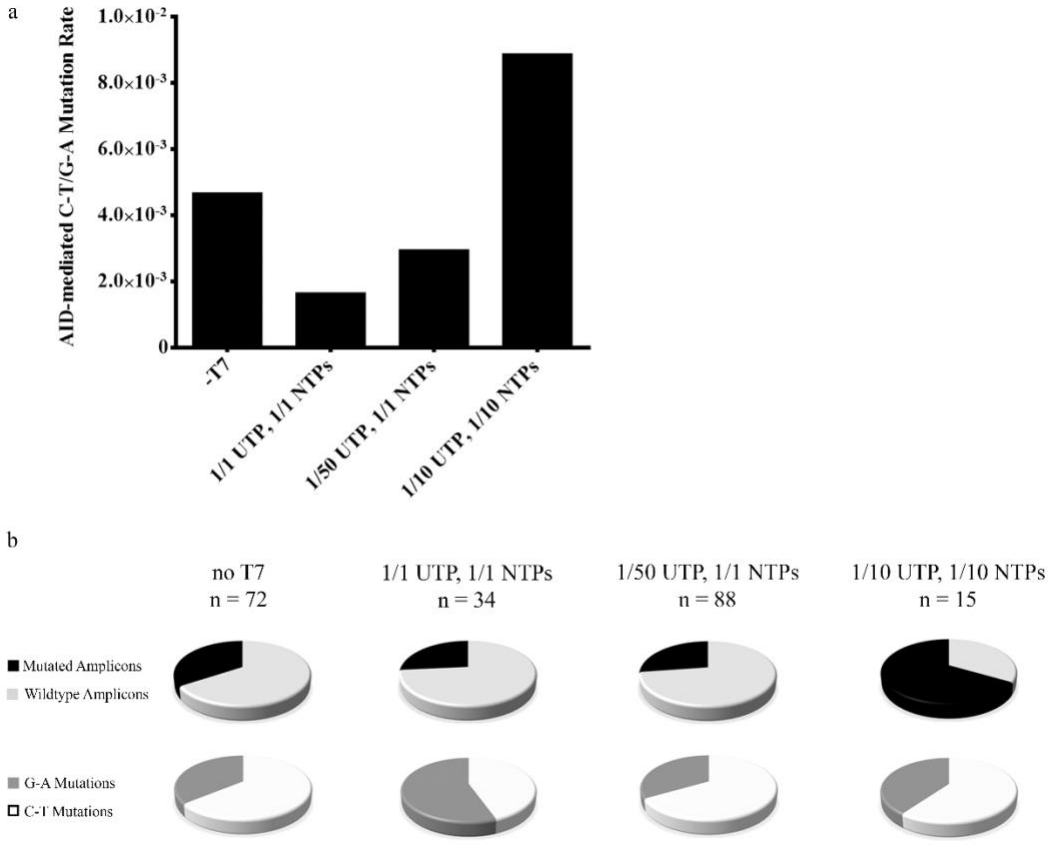


Figure 36: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations during *in vitro* transcription (set 1). 1.2µg of GST-AID (4µl) was mixed with 4 x 10^{-4} units of UGI and incubated with 100ng DNA, T7 polymerase, and rNTPs at 32℃ for 4 hours. The rate of transcription was controlled by altering the concentration of rUTP or all rNTPs in the reaction. Template DNA was subjected to degen-PCR and prepared for sequencing. Template DNA that was transcribed by T7 polymerase in the absence of AID was used to control for Taq errors. a) The average mutation rate was found by taking the sum of C-T and G-A mutations from each reaction condition and dividing them by the total number of nucleotides analyzed from that reaction. b) The ratio of mutated to wildtype amplicons and amplicons with C-T or G-A mutations was plotted in pie charts, where "n" is the number of amplicons included in the analysis. The top row of pie charts show the ratio of mutated to wildtype amplicons, where mutated is shown in black and wildtype is show in grey. The bottom row of pie charts show the ratio of amplicons containing C-T mutations to those with G-A mutations, where G-A mutations are shown in dark grey and C-T mutations are shown in white.

Next, we repeated the *in vitro* transcription assay using 9 rates of transcription where the rate of transcription was controlled by diluting only UTP (Figure 37a) or varying UTP against a lower [NTP] (Figure 37b). AID activity during transcription was compared to the average of three independent reactions of GST-AID on supercoiled DNA without T7 polymerase ("no T7"). Based on our hypothesis and our preliminary data (Figure 36), we expected AID-mediated mutations to accumulate until an "optimal speed" of transcription was reached and then decrease both above and below that speed. We found that AID activity increased 2.6-fold from no transcription to "full-speed" transcription (1/1 UTP, 1/1 NTPs; Table 11, Figure 37a,b). There was no significant difference in AID activity when only UTP was diluted by 1/10 or 1/50, but activity dropped approximately 2-fold when UTP was diluted to 1/100 (Figure 37a). At 1/200 and 1/400 UTP AID activity levels increased to near that of full speed transcription (Table 12, Figure 37a). When both UTP and rNTPs were diluted to slow the overall speed of transcription a similar pattern was observed as when we only diluted UTP to stall at dA. There was no significant difference between full speed transcription and when all rNTPs were diluted 1/10, but activity dropped approximately 4-fold when UTP was diluted by 1/100 and NTPs by 1/10 (Figure37b). AID activity increased approximately 3-fold from 1/100 UTP and 1/10 NTPs to the lowest dilution of NTPs (1/400 UTP 1/10 NTPs) (Table 11, Figure 37b).

Next, we examined the ratio of wildtype to mutated amplicons from the fastest transcription speed (1/1 UTP, 1/1 NTPs) to the slowest (1/400 UTP, 1/10 NTPs; Figure 37c). We found that transcription increased the proportion of mutated amplicons to wildtype amplicons in comparison to the no T7 condition (e.g. 34% of amplicons mutated

in the no T7 condition and 74% in the 1/1 UTP 1/1 NTPs condition), but then the ratio returns to that of no transcription as the transcription rate is slowed (1/100 UTP, 1/10 NTPs: 29% of amplicons are mutated) (Figure 37c). Furthermore, both strands of DNA can be targeted by AID as both C-T and G-A mutations were observed, but there is a clear strand bias towards the nontranscribed strand (top strand) as C-T mutations comprised more than 55% of the total mutations in every condition (Table 12; Figure 37c).

Earlier we found that secondary structure plays a greater role that primary sequence in AID targeting, as AID did not show preference for 5'-WRC hotspot motifs on native DNA but showed a slight preference for hotspots on its single-stranded counterpart (Section 3.9, Figure 30). We were therefore interested to determine substrate preference on DNA undergoing transcription in comparison to no transcription. When the transcription rates were slowed by only diluting UTP, there was a clear and consistent preference towards the four WRC hotspots: AGC, TAC, AAC and TGC (Figure 38a), similar to the mutability index described by Larijani and colleagues (Larijani et al. 2005). When the transcription rates were slowed by diluting both UTP and rNTPs, there was still a slight preference towards the WRC hotspot motifs but activity also peaked at the non-WRC motifs ATC, CAC, TCC and CGC (Figure 36b). Overall our transcription data suggests that as transcription is slowed, secondary structure plays a greater role than primary sequence in determining AID targets, and most significantly, that AID can efficiently target DNA in the absence of transcription, but only at 2-3-fold lower levels than its maximal transcription-induced mutation rates. Transcription may change the way AID targets

genomic DNA *in vivo* by unwinding DNA and creating accessibility, but further experimentation is needed to elucidate this hypothesis.

Table 11 | Rate and Distribution of Mutations												
	GST-AID											
Transcription Condition	no T7	1/1 UTP, 1/1 NTPs	1/10 UTP, 1/1 NTPs	1/50 UTP, 1/1 NTPs	1/100 UTP, 1/1 NTPs	1/200 UTP, 1/1 NTPs	1/400 UTP, 1/1 NTPs	1/10 UTP, 1/10 NTPs	1/100 UTP, 1/10 NTPs	1/400 UTP, 1/10 NTPs
C-T	215	465	641	564	284	439	478	259	106	406
G-A	65	124	105	144	90	181	16	241	23	39
Taq C-T Error Rate	7.47×10^{-5}	7.47×10^{-5}	7.47×10^{-5}	7.47×10^{-5}	7.47×10^{-5}	7.47×10^{-5}	7.47×10^{-5}	7.47×10^{-5}	7.47×10^{-5}	7.47×10^{-5}
Taq G-A Error Rate	4.98×10^{-5}	4.98×10^{-5}	4.98×10^{-5}	4.98×10^{-5}	4.98×10^{-5}	4.98×10^{-5}	4.98×10^{-5}	4.98×10^{-5}	4.98×10^{-5}	4.98×10^{-5}
Taq C-T Errors	4	4	4	4	4	4	3	3	4	4
Taq G-A Errors	3	2	3	2	3	3	2	2	2	3
Corrected C-T	211	461	637	560	280	435	475	256	102	402
Corrected G-A	62	122	102	142	87	178	14	239	21	36
Total C-T/G-A mutations	273	583	739	702	367	613	489	495	123	438
Nucleotides Analyzed	58,483	47,633	51,062	46,203	49,117	50,010	46,004	45,361	46,525	52,989
Overall Mutation Rate	4.67×10^{-3}	1.22×10^{-2}	1.45×10^{-2}	1.52×10^{-2}	7.47×10^{-3}	1.23×10^{-2}	1.06×10^{-2}	1.09×10^{-2}	2.64×10^{-3}	8.27×10^{-3}
Total Amplicons	73	65	70	49	58	57	57	52	62	70
Wildtype Amplicons	48	17	21	12	24	15	20	21	44	39
Mutated Amplicons	25	48	49	37	34	42	37	31	18	31
Amplicons with C-T Mutations	16	34	34	28	24	28	31	22	10	26
Amplicons with G-A Mutations	9	14	15	9	10	14	6	9	8	5

Table 11: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations during *in vitro* transcription (set 2). The speed of transcription was altered by altering the concentration of rUTP or all rNTPs in the reaction. All error and mutation rates were determined as described in Table 1. The data in the table above was collected from one reaction for each condition.

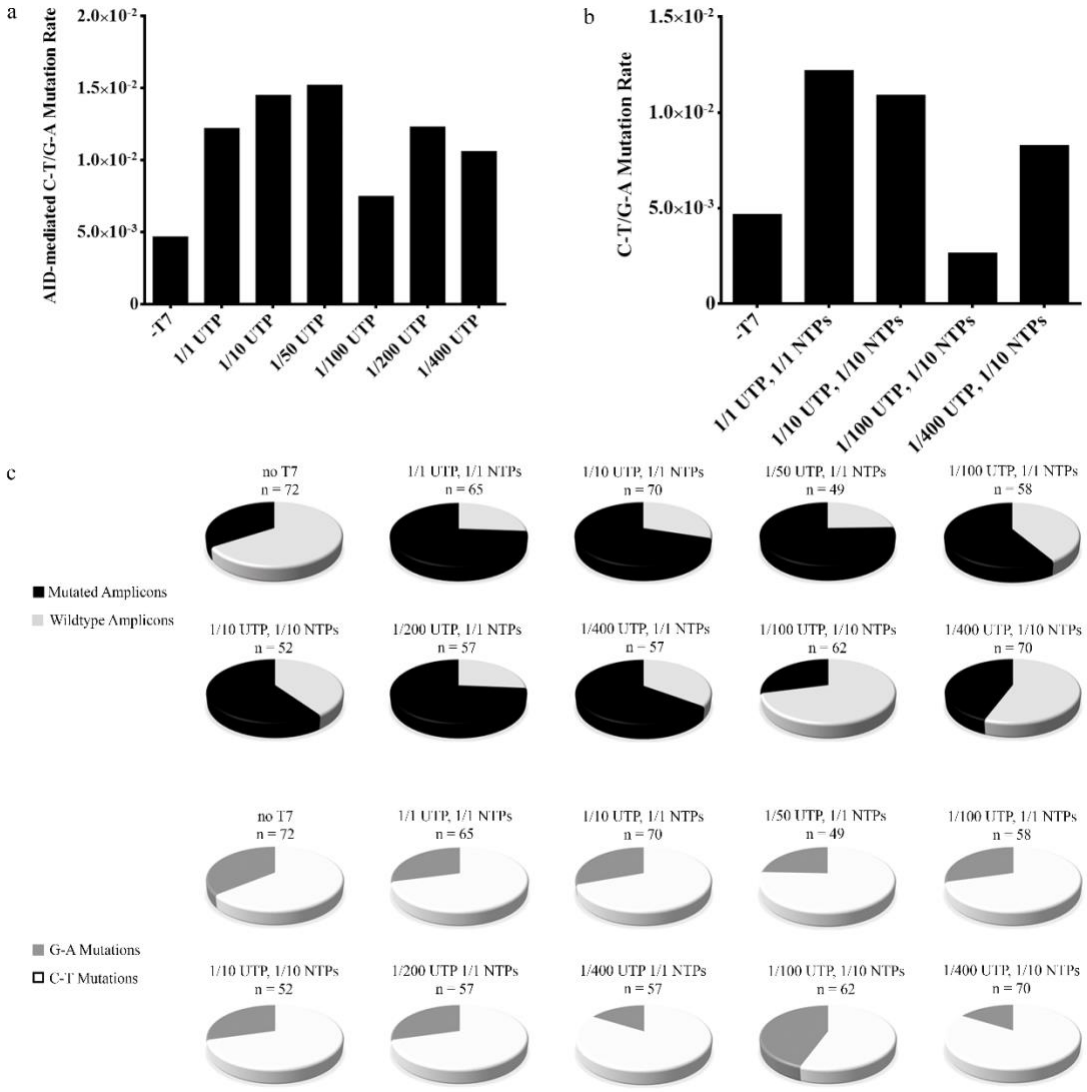


Figure 37: Rate and Distribution of GST-AID-mediated C-T and G-A Mutations during *in vitro* transcription (set 2). 1.2 µg of GST-AID (4 µl) was mixed with 4 x 10^{-4} units of UGI and incubated with 100 ng DNA, T7 polymerase, and rNTPs at 32°C for 4 hours. The rate of transcription was controlled by altering the concentration of rUTP or all rNTPs in the reaction. Template DNA was subjected to degen-PCR and prepared for sequencing. Template DNA that was transcribed by T7 polymerase in the absence of AID was used to control for Taq errors. The data from three independent +T7, no AID reactions was averaged to obtain the "Taq error control". Mutations obtained from the control were considered to be generated by Taq polymerase and subtracted from the experimental conditions. The rate and distribution of AID-mediated mutations for templates undergoing transcription was compared to the average of three independent reactions of GST-AID on supercoiled DNA without transcription. a) The average mutation rate was found by taking

the sum of C-T and G-A mutations from each reaction condition and dividing them by the total number of nucleotides analyzed. b) The ratio of mutated to wildtype amplicons and amplicons with C-T or G-A mutations was plotted in pie charts, where "n" is the number of amplicons included in the analysis. The top row of pie charts show the ratio of mutated to wildtype amplicons, where mutated is shown in black and wildtype is show in grey. The bottom row of pie charts show the ratio of amplicons containing C-T mutations to those with G-A mutations, where G-A mutations are shown in dark grey and C-T mutations are shown in white.

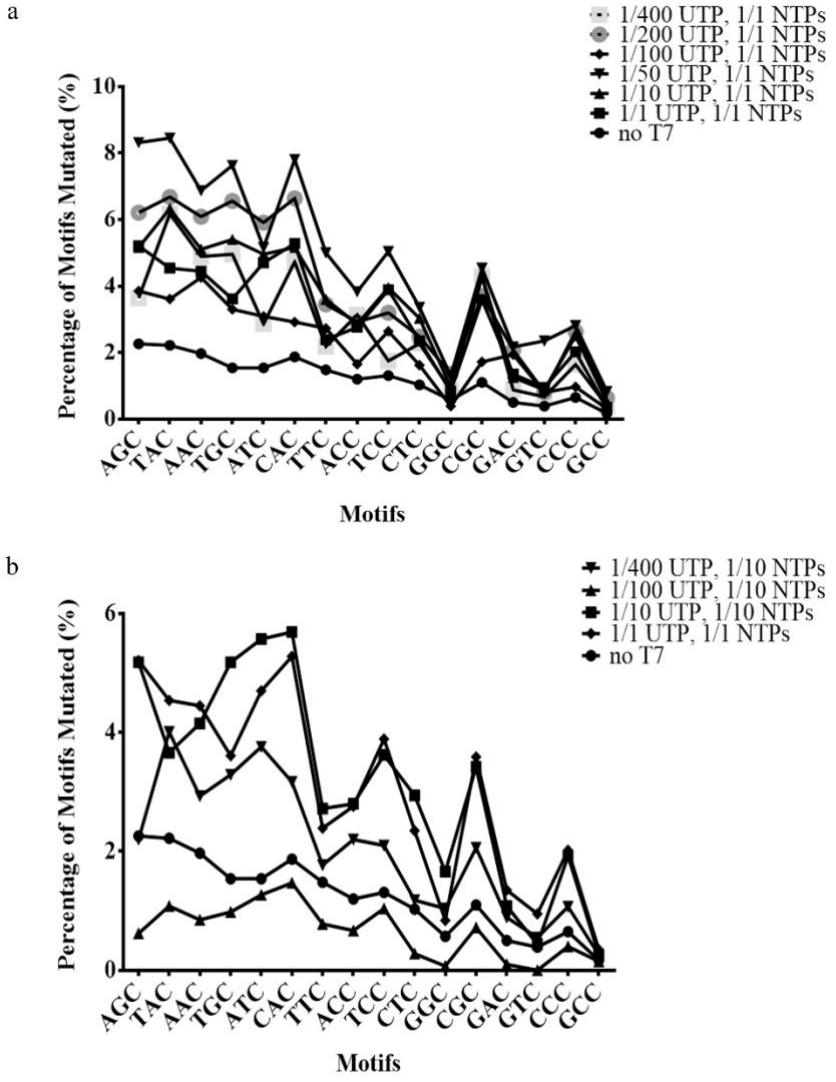


Figure 38: Primary Sequence Analysis of GST-AID-mediated Mutations during *in vitro* transcription. The trinucleotide motif for each GST-AID-mediated C-T or G-A mutation was recorded for each transcription rate. G-A mutations were considered as C-T mutations on the opposite (bottom) strand, so the reverse complement of the GXX (X= any nucleotide) trinucleotide motif was considered. The motifs are along the x-axis with the four WRC (W= A/T, R= A/G) hotspot motifs at the beginning. The percentage of motifs mutated (y-axis) was found by taking the number of C-T and G-A mutations at a particular motif and dividing it by the product of the number of that motif within the target DNA sequence and the number of amplicons analyzed. The motif data from three independent reactions of GST-AID on supercoiled DNA without transcription were averaged to obtain the "no T7" control. a) Motif data for transcription conditions where only the rUTP concentration was altered. b) Motif data for transcription conditions where the

concentration of all rNTPs was altered. Overall, AID targets WRC hotspot motifs preferentially during transcription.

IV. Discussion

4.1 AID can Target and Mutate Both Supercoiled and Relaxed Linear DNA without Transcription

Since the discovery of AID, a major goal in the field has been to elucidate its mechanism of substrate targeting. The focus of AID targeting has been centered around transcription, with numerous co-immunoprecipitation assays showing AID being able to interact with the RNA polymerase complex and transcription-associated co-factors (Nambu et al. 2003, Chaudhuri et al. 2004, Conticello et al. 2008, Pavri et al. 2010, Willmann et al. 2012, Hu et al. 2015). While the cellular environment *in vivo* is complex and it cannot be discredited that AID does indeed interact with numerous components of the transcriptional machinery, we believe that many of the keys to determining how and why AID targets some genes over others lies in the biochemical properties both of the enzyme and of its substrates. We have shown many times that AID activity does not depend on the presence of co-factors or any components of the transcriptional machinery (King et al. 2015, Abdouni et al. 2013, Dancyger et al. 2012, Larijani and Martin 2007, Larijani et al. 2007, Larijani et al. 2005 a,b). Furthermore, AID's high positive surface charge (+14) may lead it to artificially precipitate with other proteins and/or RNA *in vitro*, so we believe taking a more basic experimental approach will allow us to narrow down features of the enzyme and its substrate DNA that are involved in targeting.

Prior to our study, AID had been shown to be active on supercoiled DNA but not on relaxed DNA of the same sequence (Shen and Storb 2004). No further studies have looked deeper into the role of topology in the absence of transcription. We believed that

there was more to be uncovered, and thus wanted to study AID activity and its mutation pattern further. Interestingly, we found that AID does indeed mutate relaxed linear DNA *in vitro* (Figure 11, 12, 22, 24, 25, 26, 27), but using deam-PCR we showed that it highly mutates relaxed linear DNA 10-100-fold less than its supercoiled counterpart (Figure 24, 25, 26). Torsional strain is a property of supercoiled DNA imparted due to its natural twist and writhe, which is partially relieved by local denaturation and strand separation (Benham 1979). Local denaturation is common in supercoiled plasmids, even under conditions where base pairing is energetically favored in linear DNA. The transient partial disassociation of the helix into its two sister strands renders genomic DNA accessible to enzyme attack, and supercoilicity has long been suggested to play a role in the regulation of transcription, DNA replication, recombination and DNA repair (Wang 1974, Marians et al. 1977, Benham 1979). Linear duplex DNA also exhibits fluctuations in base pairing lasting for approximately 10^{-7} seconds, but does not form "bubbles" of 2 or more unpaired bases (reviewed in Frank-Kamenetskii and Prakash 2014). Therefore, it is not surprising that AID can highly mutate the supercoiled substrate more efficiently than the linear. Interestingly, the degen-PCR (Figure 27) result is not consistent with that of the deam-PCR (Figure 24). The degen-PCR indicated that AID mutates supercoiled and relaxed linear DNA at a near equal frequency (Table 7, Figure 27a). Perhaps AID heavily mutates (eg. >10 mutations) supercoiled DNA more efficiently than the linear, but this proportion of DNA is very small in comparison to the "lightly mutated" (eg. 1-3 mutations) and wildtype DNA that we only observed 3 heavily mutated amplicons (2% of the total) in the supercoiled set and 1 heavily mutated amplicon (<1% of the total) in the linear set.

Although the deam- and degen-PCR data seem contradictory, it may reveal how AID behaves *in vitro*, targeting few molecules of DNA and mutating them heavily while leaving few mutations elsewhere. However, with an AID:DNA ratio of 1.1×10^3, and 1.38×10^{10} copies of the same plasmid (Table 6), it remains unclear that in a pool of millions of AID and DNA molecules why only a small percentage of the substrate DNA is targeted.

4.2 AID Activity without Transcription cannot be Solely Explained by DNA Breathing

Bisulfite was used to map the single-stranded regions induced by breathing of our double-stranded plasmid. We wanted to determine the regions that are likely to have fluctuations in base pairing as we hypothesized that AID will gain access to dsDNA through breathing of sister strands in the absence of transcription. We hypothesized that the supercoiled DNA would breathe more heavily than the linear DNA due to the torsional strain imposed by its topology. While the supercoiled DNA did have approximately 1.5-fold more bisulfite-mediated mutations than the linear, it was not as great of a difference as we expected. We also expected the pattern of mutation to differ between the two topologies. Based on the loops, bubbles and cruciform structures that supercoiled DNA can form during transient fluctuations to relieve torsional strain, we expected that multiple C's in close proximity could be mutated at the same time so that the overall mutation pattern would be in small patches of nucleotides. Since linear DNA is proposed to fluctuate one base pair at a time (Frank-Kamenetskii and Prakash 2014) and there are no topological constraints on the linearized plasmid, we expected the mutations to be randomly distributed

throughout the length of the target sequence. We found that there was a slight difference in the pattern of mutation between the two topologies, with the supercoiled DNA having small regions with multiple closely spaced mutations on individual amplicons (Figure 28a, Figure 29 top). Besides these few clustered regions, the majority of the mutations were distributed individually. Although we expected the rate and pattern of bisulfite-mediated mutation to be based on topology, topology is not the only factor governing DNA breathing. Both sequence (Eslami-Mossallam et al. 2016, Furlong et al. 1989, Nishimura 1985) and secondary structure (Altan-Bonnet et al. 2003, Gough et al. 1986) influence DNA dynamics, potentially explaining the lack of major differences in DNA breathing between the supercoiled and linear topologies examined.

Although we had hypothesized that AID gains access to ssDNA regions through DNA breathing in the absence of transcription, there was a considerable difference in the pattern of AID activity in comparison to that of bisulfite. The bisulfite-mediated mutations were distributed throughout over 80% of both the supercoiled and linear amplicons (Figure 28b). On the other hand, AID mutated a small percentage of the total DNA targets. The GST-AID-mediated mutations on supercoiled DNA were contained within 21% of the amplicons, while those on the linear DNA comprised 11% of the total amplicons (Figure 27c). Although AID-His mutated supercoiled DNA at a ~25-fold higher rate than GST-AID and a ~4-fold higher rate than bisulfite, the C-T and G-A mutations were found within 41% of the total amplicons (Table 3, Figure 15b). Moreover, the pattern of mutation on individual amplicons was also markedly different. Each supercoiled amplicon mutated by bisulfite contained 1-19 C-T or G-A mutations, while those from the linear substrate

contained 1-8 mutations each. The majority of the supercoiled and linear amplicons mutated by GST-AID contained 1-4 mutations, with three supercoiled amplicons containing 9-44 mutations and one linear amplicon containing 83 mutations. Furthermore, the supercoiled amplicons mutated by AID-His contained anywhere from 1 to 97 C-T or G-A mutations. It is not possible that the supercoiled or linear duplex DNA is breathing at that high of a rate to fully explain these highly mutated amplicons. It is possible that DNA breathing provides an opening for AID to initially bind DNA, and once AID is bound that it can hold open the sister strands and thereby mutate cytidines in close proximity. Moreover, the differences in activity patterns between AID and bisulfite most likely lie in AID's intrinsic biochemical properties: high positive surface charge (reviewed in Larijani and Martin 2012), high nanomolar affinity for ssDNA (Larijani et al. 2007), and its fluctuations in catalytic pocket opening and closing (King and Larijani 2017).

4.3 AID can Mutate in both a Processive and Distributive Pattern on dsDNA without Transcription

AID's mode of targeting has been a matter of debate since AID was found to act distributively on small 40 nucleotide DNA oligonucleotide substrates (Coker and Petersen-Mahrt 2007), but processively on ssDNA 230 nucleotides in length (Pham et al. 2003). If AID does indeed act in a distributive manner, a single AID molecule will only be able to deaminate each substrate once regardless of the number of cytidines within the substrate's sequence. Alternatively, AID's activity is described as processive if it heavily mutates a small percentage of the total pool of substrate, leaving the rest untouched. Further studies

have shown that AID can mutate processively on supercoiled duplex DNA (Shen and Storb 2004), as well as DNA undergoing transcription (Bransteitter et al. 2004). During this "processive" activity, AID first targets 5'-WRC hotspots, but with increased time of reaction can later target 5'-SYC cold spots and neutral regions generating clusters of 1-10 mutations before dissociating and moving to a different position of the same substrate molecule (Bransteitter et al. 2004). Recently, AID activity during transcription has been analyzed using single-molecule fluorescence resonance energy transfer (smFRET; Senavirathne et al. 2015). AID was found to bind DNA randomly and bidirectionally scan ssDNA regions in short sliding or "hopping" movements, with a mean binding time of approximately 4.5 minutes during a single binding event.

We also found that AID can act in a seemingly "processive" manner as GST-AID heavily mutated 3 of 147 (2%) amplicons in supercoiled dsDNA and 1 of 133 (0.8%) amplicons in linear dsDNA without transcription. Furthermore, AID-His acted in a highly "processive" manner on supercoiled dsDNA, heavily mutating 34 of 117 (29%) amplicons. Although we cannot be sure that those heavily mutated DNA strands were mutated by a single AID molecule each, it is unlikely that multiple AID molecules would all target the same DNA strand in a pool of millions of potential targets. Furthermore, we also found that AID can act in a distributive manner, mutating 1 or 2 cytidines per target. The combination of both distributive and processive activity was observed most clearly with GST-AID, where 26 of 147 (18%) supercoiled amplicons and 14 of 133 (11%) linear amplicons contained 1 or 2 C-T or G-A mutations. All AID-mediated mutations occurred on a variety of trinucleotide motifs including 5'-WRC hotspots, 5'-SYC cold spots and

neutral regions (Figure 30). Without transcription, AID is likely attracted to breathing duplex DNA due to its positive surface charge (reviewed in Larijani and Martin 2012) and high affinity for ssDNA (Larijani et al. 2007). Due to the catalytic pocket inaccessibility most binding events will not lead to productive deamination (King et al. 2015), potentially explaining why AID mutated the native DNA targets in a largely distributive manner whereby multiple targets contained few mutations. Occasionally, AID may repeatedly act on the same substrate molecule, generating clusters of mutations in a processive-like manner. A combination of "processive" and distributive activity on a variety of trinucleotide motifs may explain how AID is able to initiate the diversification of a near infinite array of antibodies, each with the capacity to bind a unique epitope.

4.4 AID can Target Both DNA Strands without Transcription

During SHM both DNA strands are mutated somewhat equally (Xue et al. 2006, Foster et al. 1999, Spencer et al. 1999, Dörner et al. 1998), and therefore must be targeted by AID. However, there is contrasting evidence concerning whether or not AID preferentially targets one strand over the other. Without transcription, AID can access and mutate both strands of supercoiled duplex DNA (Shen and Storb 2004). With transcription, AID has been shown to preferentially target the nontranscribed strand (Kodgire et al. 2013, Martomo et al. 2005, Sohail et al. et al. 2003), as well as to target both strands (Besmer et al. 2006). Another body of evidence indicates that the preference of AID towards either strand is dependent on the sequence (MacCarthy et al. 2009) and/or structure of the target gene (Duvvuri et al. 2012, Shen et al. 2005).

Our results indicate that both strands should have accessible single-stranded regions in the absence of transcription, as there was an equal 50:50 distribution of bisulfite-mediated C-T and G-A mutations in the supercoiled DNA and 55% C-T and 45% G-A mutations in the linear (Table 8, Figure 28b). C-T mutations are indicative of deaminase activity on the nontranscribed strand, while G-A mutations indicate that deamination of dC on the template strand. Since both strands of our supercoiled and relaxed linear substrates were equally targeted, transient fluctuations in breathing should render both strands accessible to AID. Both GST-AID and AID-His mutated both DNA strands in the absence of transcription, as both C-T and G-A mutations were observed (Figure 15b, 27c). On the supercoiled substrate, 65% of the GST-AID-mediated mutations were C-T and 35% were G-A, indicating a preference towards the nontranscribed strand (Figure 27c). The linear substrate was targeted nearly equally with 47% C-T and 53% G-A mutations (Figure 27c). AID-His had a near identical ratio of C-T and G-A mutations as GST-AID on the supercoiled substrate, where 63% of the mutations were C-T and 37% were G-A (Figure 15b). It is likely that the slight preference towards the nontranscribed strand in the supercoiled substrate is due to either primary sequence or secondary structural features that make this strand slightly more desirable and/or accessible to AID, since bisulfite targeted the supercoiled substrate completely equally.

4.5 Secondary Structure is a more Important Determinant of AID Targeting than Primary Sequence in the Absence of Transcription

In the introduction, we suggested that there is no one single feature of DNA that is an absolute determinant of AID targeting, meaning that the sequence, structure and topology all create an environment that is favorable or unfavorable to AID. AID has been found to target and mutate structures such as bubbles (Bransteitter et al. 2003, Larijani et al. 2007, Larijani and Martin 2007), R-loops (Bransteitter et al. 2003, Canugovi et al. 2009, Abdouni et al. 2017) and supercoiling (Shen and Storb 2004) that may arise during transcription. Modelling data of AID-generated mutations on IGHV3-23 DNA has suggested that highly mutated bases on the nontranscribed strands are mostly paired, while those on the opposite strands are mostly unpaired or open (Duvvuri et al. 2012). Using mfold DNA folding software (Zuker 2003), it was predicted that these highly-mutated bases were likely stabilized in secondary structures, such as bubbles and stem-loops (Duvvuri et al. 2012). Therefore, we also modelled the top (nontranscribed) strand of our target DNA region using mfold software to determine which regions of our target DNA sequence that have the capacity to form secondary structures (Figure 31).

We found that there are several stem-loop and hairpin structures that have the potential to form in our target DNA substrate (Figure 31). Next, the bisulfite- (Figure 32) and GST-AID (Figure 33)-mediated C-T mutations were superimposed onto the proposed structure of our substrate. We found that most of the bisulfite- and GST-AID-mediated C-T mutations were located in regions predicted to form secondary structures. 71% of the bisulfite-mediated C-T mutations were found to lie within secondary structures, with 43%

in paired stem regions and 28% in unpaired loop regions (Table 9). 66% of the GST-AID-mediated C-T mutations were located within predicted secondary structures, with 37% in the stem region and 29% in the loop region (Table 9). Overall, the loop regions were predicted by mfold to remain unpaired or open, while the pairing of stem regions was predicted to fluctuate (Figure 34). Since the mfold software does not take into account pairing with the sister DNA strand or DNA topology, the predicted model may not completely represent the plasmid substrate *in vitro*. However, it still allows us to gain a picture of the areas that are energetically more likely to form secondary structures during DNA breathing. Moreover, since the majority of mutations fell within secondary structures (Figure 32, 33; Table 9), we believe that it is not a coincidence and that it is likely that at least some of these structures form transiently during DNA breathing. Furthermore, when the mutated trinucleotide motifs were plotted, GST-AID did not show preference for 5'-WRC hotspots in dsDNA but showed a small preference towards hotspots in heat-denatured DNA (ssDNA) (Figure 30). Taken together our data suggest that secondary structure is more important as a determinant of AID targeting than primary sequence alone. It is possible that secondary structures stabilize ssDNA regions in the absence of transcription, making these regions accessible to AID.

4.6 Transcription Changes the Pattern of AID Targeting

Presently no study has compared AID activity on dsDNA concurrently with and without transcription. We originally hypothesized that transcription generates secondary structures that can be targeted by AID and that slowing transcription down to an "optimal"

speed will allow these structures to persist to allow maximal AID activity during the 4-hour incubation. We thought that if transcription is too fast AID will not get sufficient opportunity to target transcription-generated secondary structures. If transcription is too slow, the rate of AID activity will be the same as without transcription. The results were a bit unexpected as GST-AID activity dropped at both 1/100 UTP and 1/100 UTP 1/10 NTPs and then increased again at later dilutions (Figure 37a,b). However, we noticed that transcription changes the way AID targets its substrate DNA. Firstly, transcription increases the proportion of amplicons that were targeted by GST-AID (Figure 37c). In the no T7 condition all AID-mediated C-T or G-A mutations occurred within 34% of the total amplicons, while 66% of the amplicons were untouched by AID. When the transcription conditions were ordered from what we considered the fastest speed of transcription (1/1 UTP, 1/1 NTPs) to the slowest (1/400 UTP, 1/10 NTPs), 59-76% of the amplicons were mutated by AID until the second last condition (1/100UTP, 1/10 NTPs) where only 29% of the amplicons were mutated by AID (Figure 37c). Thus, transcription increases the accessibility of AID to a wider range of target substrates. Once the speed slows down enough fewer overall strands are targeted, resembling AID activity without transcription. It is also interesting that transcription slightly increases AID targeting towards the nontranscribed strand, as there were 5-20% more C-T mutations than the no T7 condition (Figure 37c). Once transcription slowed to a certain point (1/100 UTP, 1/10 NTPs) both DNA strands were targeted more equally as there were 56% C-T mutations and 44% G-A mutations (Figure 37c).

In section 4.5 we suggested that secondary structure was more important than primary sequence in determining AID targeting because GST-AID did not show a preference towards WRC hotspot motifs in supercoiled and linear duplex DNA but showed a preference towards hotspots when the two substrates were heat-denatured (Figure 30). If transcription renders the DNA more accessible to AID, we would expect that GST-AID would show the same trinucleotide motif preference during transcription as it did on heat-denatured DNA. In the reactions where transcription rates were slowed by only dilution of UTP there was a clear preference towards the four WRC hotspots (Figure 38a). When transcription was slowed by diluting both UTP and all other NTPs, preference started to be placed on otherwise neutral motifs (Figure 38b). When both UTP and NTPs are diluted, the overall rate of transcription was slower than when only UTP is diluted alone. Perhaps when the overall rate of transcription is slower secondary structure becomes of greater importance for AID targeting than primary sequence. It seems that AID will have a slight preference towards 5'-WRC motifs if it can readily access them in ssDNA, but will mutate at other motifs if they are in regions that have a high rate of breathing and/or stabilized by secondary structures.

V. Future Directions

The overall goal of this project was to determine the relative importance of DNA primary sequence, secondary structure or topology in AID targeting and activity. At first, we thought DNA topology would play the most crucial role, but now it seems that secondary structure may have near equal importance. It is perhaps the interplay between all three features (i.e. sequence, structure and topology) that ultimately determines why AID targets some genes more than others. So far, we have established a protocol for our AID activity assay that is replicable. We have mapped ssDNA regions induced by breathing of our supercoiled and relaxed linear substrates using bisulfite. Furthermore, we have shown that AID prefers to heavily mutate supercoiled DNA over its relaxed linear counterpart, described AID's pattern of activity on dsDNA both with and without varying speeds of transcription. We have also generated preliminary data hinting at the importance of secondary structure in AID recruitment.

We have started to generate secondary structure modeling data using mfold to help us determine what leads AID to target some regions over others. It seems from our trinucleotide motif data (Figure 30) in combination with our mfold data (Figures 31-33) that secondary structure is more important than primary sequence in attracting AID to a particular region. In the future, we should also model the transcribed strand to see if the mutated regions align with potential secondary structures as they did with the modelled nontranscribed strand. It would also be interesting to align only mutations from single clones to get more information on AID's movement (i.e. scanning, jumping) in its search for a substrate. Data from *in vitro* transcription could also be modelled to see if mutations

cluster more around secondary structures at slower transcription speeds. It has been suggested that the sequence environment surrounding 5'-WRC hotspots and 5'-SYC cold spots greatly influences AID recruitment and activity (MacCarthy et al. 2009). As the primary sequence of DNA determines its secondary structure, it is important to also examine AID activity on different target sequences. The target DNA sequence used in this assay is a random sequence solely chosen based on its length of 1.2kb. We used it to develop our assay and gain insight into some features that may attract AID. Now that our assay is optimized to be consistent and repeatable, we can use it to examine AID activity on Ig genes and oncogenes both with and without transcription. Some genes that are of potential interest are those of the Ig variable and switch regions, as well as oncogenes such as c-MYC, BCL2 and BCL6. Comparing AID activity on its natural substrate to its activity on oncogenes and our random sequence will allow us to further delineate what sequence and/or structural features attract AID.

We have also generated some *in vitro* transcription data showing that transcription changes the pattern of strand and motif targeting. To examine any differences in AID targeting due to different sequences undergoing transcription, we could incubate AID with Ig or oncogene sequences (as suggested above) during transcription. In Section 1.7 it was noted that supercoiling spreads approximately 1.5-2kb from the TSS of most transcribed genes (Kouzine et al. 2013). During SHM, AID mutations appear 100-200bp upstream of the V region promoter and span approximately 2kb (Longerich et al. 2006, Storck et al. 2011). There could be a correlation between supercoiling of actively transcribed genes and AID activity. To examine the potential relationship between AID and supercoiling, the

distance of the promoter from our gene of interest could be varied. Since supercoiling is generated during active transcription, the further the promoter is away from the target sequence the more the supercoiling could be dissipated before reaching the target gene. We could also put the promoter on the opposite strand of the gene to observe if there are any changes in strand preference. During transcription, we found that there was a strong preference towards the nontranscribed strand as the majority of the mutations were C-T (Figure 37c). If the promoter is on the opposite strand, will AID still prefer the nontranscribed strand or will it now prefer the transcribed strand? Furthermore, we could also use a yeast SWI/SNF chromatin remodeling assay to assess AID activity during transcription-coupled DNA remodeling. Analyzing the role of sequence, supercoiling and chromatin remodeling during transcription can give us further insight into what DNA features attract AID leading it to target some genes more than others.

VI. Concluding Remarks

Here we have analyzed the influence of DNA sequence, structure and topology on AID targeting and activity both with and without transcription. We have found that transcription is not necessary for AID to effectively target and mutate supercoiled and relaxed linear DNA, and that AID may have a 10-100-fold preference for heavily mutating supercoiled over its linear counterpart. Furthermore, we have used bisulfite to map ssDNA regions induced by breathing of the supercoiled and relaxed linear substrates, suggesting that AID may initially gain access to dsDNA through these breathing regions. Although transcription is unnecessary for AID activity, it does change the pattern of AID targeting. Considering our TIAA and TAAA assay, sequence motif and preliminary modeling data altogether, it seems that DNA secondary structure and topology play a more crucial role than primary sequence in determining AID recruitment and activity.

References

Abdouni H, King JJ, Suliman M, Quinlan M, Fifield H, Larijani M (2013) Zebrafish AID is capable of deaminating methylated deoxycytidines. Nucleic Acids Res. 41:5457-5468

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

Alexandrov BS, Gelev V, Monisova Y, Alexandrov LB, Bishop AR, Rasmussen KØ, Usheva A (2009) A nonlinear dynamic model of DNA with a sequence-dependent stacking term. Nucleic Acids Res. 37:2405-2410

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

Ambion (2012) MEGAscript kit: User guide. Carlsbad, California: Life Technologies

Aukema SM, Kreuz M, Kohler CW, Rosolowski M, Hasenclever D, Hummel M, Küppers R, Lenze D, Ott G, Pott C, Richter J, Rosenwald A, Szczepanowski M, Schwaenen C, Stein H, Trautmann H, Wessendorf S, Trümper L, Loeffler M, Spang R, Luin PM, Klapper W, Siebert R (2014) Biological characterization of adult MYC-translocation-positive mature B cell lymphomas other than molecular Burkitt lymphoma. Haematologica. 99:726-735

Bachl J, Carlson C, Gray-Schopfer V, Dessing M, Olsson C (2001) Increased transcription levels induce higher mutation rates in a hypermutating cell line. J Immunol. 166:5051-5057

Baneyx F, Mujacic M (2004) Recombinant protein folding and misfolding in Escherichia coli. Nat Biotechnol. 22:1399-1408

Berek C, Berger A, Apel M (1991) Maturation of the immune response in germinal centers. Cell. 67:1121-1129

Besmer E, Market E, Papabasiliou FN (2006) The transcription elongation complex directs activation-induced cytidine deaminase-mediated DNA deamination. Mol Cell Biol. 26:4378-4385

Betz AG, Milstein C, González-Fernández A, Pannell R, Larson T, Neuberger MS (1994) Elements regulating somatic hypermutation of an immunoglobulin kappa gene: critical role for the intron enhancer/matrix attachment region. Cell. 77:239-248

Bransteitter R, Pham P, Scharff MD, Goodman MF (2003) Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proc Natl Acad Sci USA. 100:4102-4107

Bransteitter R, Pham P, Calabrese P, Goodman MF (2004) Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase. J Biol Chem. 279:51612-51621

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

Brill SJ, Sternglanz R (1988) Transcription-dependent DNA supercoiling in yeast DNA topoisomerase mutants. Cell. 54:403-411

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

Clapier CR, Cairns BR (2009) The biology of chromatin remodeling complexes. Annu Rev Biochem. 78:273-304

Carpenter MA, Rajagurubandara E, Wijesingshe P, Bhagwat AS (2010) Determinants of sequence-specificity within human AID and APOBEC3G. DNA Repair (Amst). 9:579-587

Chandra V, Bortnick A, Murre C (2015) AID targeting: old mysteries and new challenges. Trends Immunol. 36:527-535

Chaudhuri J, Khuong C, Alt FW (2004) Replication protein A interacts with AID to promote deamination of somatic hypermutation targets. Nature. 430:992-998

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

Chelico L, Sacho EJ, Erie DA, Goodman MF (2008) A model for oligomeric regulation of APOBEC3G cytosine deaminase-dependent restriction of HIV. J Biol. Chem. 283:13780-13791

Coker HA, Morgan HD, Petersen-Mahrt SK (2006) Genetic and in vitro assays of DNA deamination. Methods Enzymol. 408:156-170

Coker HA, Petersen-Mahrt SK (2007) The nuclear DNA deaminase AID functions distributively whereas APOBEC3G has processive mode of action. DNA Repair (Amst). 6:235-243

Conticello SG, Ganesh K, Xue K, Lu M, Rada C, Neuberger MS (2008) Interaction between antibody-diversification enzyme AID and splicesome-associated factor CTNNBL1. Mol Cell. 31:474-484

Dancyger AM, King JJ, Quinlan MJ, Fifield H, Tucker S, Saunders HL, Berru M, Magor BG, Martin A, Larijani M (2012) Differences in the enzymatic efficiency of human and bony fish AID are mediated by a single residue in the C terminus modulating single-stranded DNA binding. FASEB J. 26:1517-1525

Dang Y, Wang X, Esselman WJ, Zheng Y-H (2006) Identification of APOBEC3DE as another antiretroviral factor from the human APOBEC family. J Virol. 80:10522-10533

Darzacq X, Shav-Tal Y, de Turris V, Brody Y, Shenoy SM, Phair RD, Singer RH (2007) In vivo dynamics of RNA polymerase II transcription. Nat Struct Mol Biol. 14:796-806

Das C, Tyler JK, Churchill ME (2010) The histone shuffle: histone chaperones in an energetic dance. Trends Biochem Sci. 35:476-489

Dayn A, Malkhosyan S, Mirkin SM (1992) Transcriptionally driven cruciform formation in vivo. Nucleic Acid Res. 20:5991-5997

DeFranco AL (2016) The germinal center antibody response in health and disease. F1000Res. Pii:F1000 Faculty Rev-999

Dickerson SK, Market E, Besmer E, Papavasiliou FN (2003) AID Mediates Hypermutation by Deaminating Single Stranded DNA. J Exp Med. 197:1291-1296

Di Noia J, Neuberger MS (2002) Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. Nature. 419:43-48

Dörner T, Foster SJ, Farner NL, Lipsky PE (1998) Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. Eur J Immunol. 28:3384-3396

Drolet M, Bi X, Liu LF (1994) Hypernegative supercoiling of the DNA template during transcription elongation in vitro. J Biol Chem. 269:2068-2074

Dudley DD, Manis JP, Zarrin AA, Kaylor L, Tian M, Alt FW (2002) Internal IgH class switch region deletions are position-independent and enhanced by AID expression. Proc Natl Acad Sci USA. 99:9984-9989

Duvvuri B, Duvvuri VR, Wu J, Wu GE (2011) Stabilised DNA secondary structures with increasing transcription localise hypermutable bases for somatic hypermutation in IGHV3-23. Immunogenetics. 64:481-496

Eslami-Mossallam B, Schiessel H, van Noort J (2016) Nucleosome dynamics: Sequence matters. Adv Colloid Interface Sci. 232:101-113

Foster SJ, Dörner T, Lipsky PE (1999) Somatic hypermutation of VkappaJkappa rearrangements: targeting of RGYW motifs on both DNA strands and preferential selection of mutated codons within RGYW motifs. Eur J Immunol. 29:4011-4021

Frank-Kamenetskii MD, Prakash S (2014) Fluctuations in the DNA double helix: a critical review. Phys Life Rev. 11:153-170

Fugmann SD, Schatz DG (2003) RNA AIDS DNA. Nat Immunol. 4:429-430

Fukita Y, Jacobs H, Rajewsky K (1998) Somatic hypermutation in the heavy chain locus correlates with transcription. Immunity. 9:105-114

Furlong JC, Sullivan KM, Murchie AI, Gough GW, Lilley DM (1989) Localized chemical hyperreactivity in supercoiled DNA: evidence for base unpairing in sequences that induce low-salt cruiciform extrusion. Biochemistry. 28:2009-2017

Gazumyan A, Bothmer A, Klein IA, Nussenzweig MC, McBride KM (2012) Activation-induced cytidine deaminase in antibody diversification and chromosome translocation. Adv Cancer Res. 113:167-190

Gomez M, Wu X, Wang YL (2005) Detection of BCL2-IGH using single-round PCR assays. Diagn Mol Pathol. 14:17-22

Goodman MF (2016) Better living with hypermutation. Environ Mol Mutagen. 57: 421-434

Gotissa M, Alt FW, Chiarle R (2011) Mechanisms that promote and suppress chromosomal translocations in lymphocytes. Annu Rev Immunol. 29:319-350

Gough GW, Sullivan KM, Lilley DM (1986) The structure of cruciforms in supercoiled DNA: probing the single-stranded character of nucleotide bases with bisulphite. EMBO J. 5:191-196

Goyenechea B, Klix N, Yélamos J, Williams GT, Riddell A, Neuberger MS, Milstein C (1997) Cells strongly expressing Ig(kappa) transgenes show clonal recruitment of hypermutation: a role for both MAR and the enhancers. EMBO J. 16:3987-3994

Gruber TA, Chang MS, Sposto R, Müschen M (2010) Activation-induced cytidine deaminase accelerates clonal evolution in BCR-ABL1-driven B-cell lineage acute lymphoblastic leukemia. Cancer Res. 70:7411-7420

Häsler J, Rada C, Neuberger MS (2011) Cytoplasmic activation-induced cytidine deaminase (AID) exists in stoichiometric complex with translation elongation factor 1α (eEF1A). Proc Natl Acad Sci USA. 108:18366-18371

Honjo T, Kinoshita K, Muramatsu M (2002) Molecular mechanism of class switch recombination: linkage with somatic hypermutation. Annu Rev Immunol. 20:165-196

Hu W, Begum NA, Mondal S, Stanlie A, Honjo T (2015) Identification of DNA cleavage- and recombination-specific hnRNP cofactors for activation-induced cytidine deaminase. Proc Natl Acad Sci U.S.A. 112:5791-5796

Huang J, Sousa R (2000) T7 RNA polymerase elongation complex structure and movement. J Mol Biol. 303:347-358

Kim N, Jinks-Robertson S (2017) The top1 paradox: friend and foe of the eukaryotic genome. DNA Repair (Amst). S1568-7864:30205-30207

King JJ, Larijani M (2017) A novel regulator of activation-induced cytidine deaminase/APOBECs in immunity and cancer: Schrödinger's CATalytic pocket. Front Immunol. 6:351

King JJ, Manuel CA, Barrett CV, Raber S, Lucas H, Sutter P, Larijani M (2015) Catalytic pocket inaccessibility of activation-induced cytidine deaminase is a safeguard against excessive mutagenic activity. Structure. 23:615-627

Klemm L, Duy C, Iacobucci I, Kuchen S, von Levetzow G, Feldhahn N, Henke N, Li Z, Hoffmann TK, Kim YM, Hofmann WK, Jumaa H, Groffen J, Heisterkamp N, Martinelli G, Lieber MR, Casellas R, Müschen M (2009) The B cell mutator AID promotes B lymphoid blast crisis and drug resistance in chronic myeloid leukemia. Cancer Cell. 16:232-245

Kodgire P, Mukkawar P, Ratnam S, Martin TE, Storb U (2013) Changes in RNA polymerase II progression influence somatic hypermutation of Ig-related genes by AID. J Exp Med. 210:1481-1492

Kodgire P, Mukkawar P, North JA, Poirier MG, Storb U (2012) Nucleosome stability dramatically impacts the targeting of somatic hypermutation. Mol Cell Biol. 32:2030-2040

Kouzine F, Gupta A, Baranello L, Wojtowicz D, Ben-Aissa K, Liu J, Przytycka TM, Levens D (2013) Transcription-dependent dynamic supercoiling is a short-range genomic force. Nat Struct Mol Biol. 20:396-403

Kouzine F, Sanford S, Elisha-Feil Z, Levens D (2008) The functional response of upstream DNA to dynamic supercoiling in vivo. Nat Struct Mol Biol. 15:146-154

Krasilnikov AS, Podtelezhnikov A, Vologodskii A, Mirkin SM (1999) Large-scale effects of transcriptional DNA supercoiling in vivo. J Mol Biol. 292:1149-1160

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

Kumar R, DiMenna LJ, Chaudhuri J, Evans T (2014) Biological function of activation-induced cytidine deaminase (AID). Biomed J. 37:269-283

Larijani M, Frieder D, Basit W, Martin A (2005a) The mutation spectrum of purified AID is similar to the mutability index in Ramos cells and in *ung*$^{-/-}$ *msh2*$^{-/-}$ mice. Immunogenetics, 56:840-845

Larijani M, Frieder D, Sonbuchner TM, Bransteitter R, Goodman MF, Bouhassira EE, Scharff MD, Martin A (2005b) Methylation protects cytidines from AID-mediated deamination. Mol Immunol. 42:599-604

Larijani M, Martin A (2007) Single-stranded DNA structure and positional context of the target cytidine determine the enzymatic efficiency of AID. Mol Cell Biol. 27:8038-8048

Larijani M, Martin A (2012) The biochemistry of activation-induced deaminase and its physiological functions. Semin Immunol. 24:255-263

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

Liu M, Duke JL, Richter DJ, Vinuesa CG, Goodnow CC, Kleinstein SH, Schatz DG (2008) Two levels of protection for the B cell genome during somatic hypermutation. Nature. 451:841-845

Longerich S, Basu U, Alt F, Storb U (2006) AID in somatic hypermutation and class switch recombination. Curr Opin Immunol. 18:164-174

MacCarthy T, Kalis SL, Roa S, Pham P, Goodman MF, Scharff MD, Bergman A (2009) V-region mutation in vitro, in vivo, and in silico reveal the importance of the enzymatic properties of AID and the sequence environment. Proc Natl Acad Sci USA. 106:8629-8634

Mak CH, Pham P, Afif SA, Goodman MF (2013) A mathematical model for scanning and catalysis on single-stranded DNA, illustrated with activation-induced deoxycytidine deaminase. J Biol Chem. 288:29786-29795

Mak CH, Pham P, Afif SA, Goodman MF (2015) Random-walk enzymes. Phys Rev E Stat Nonlin Soft Matter Phys. 92:032717

Manis JP, Tian M, Alt FW (2002) Mechanism and control of class-switch recombination. Trends Immunol. 23:31-39

Marians KJ, Ikeda JE, Schlagman S, Hurwitz J (1977) Role of DNA gyrase in phiX replicative-form replication in vitro. Proc Natl Acad Sci U.S.A. 74:1965-1968

Martin A, Scharff MD (2002) AID and mismatch repair in antibody diversification. Nat Rev Immunol.2:605-614

Martomo SA, Fu D, Yang WW, Joshi, NS, Gearhart PJ (2005) Deoxyuridine is generated preferentially in the nontranscribed strand of DNA from cells expressing activation-induced cytidine deaminase. J Immunol. 174:7787-7791

Matthews AJ, Zheng S, DiMenna LJ, Chaudhuri J (2014) Regulation of immunoglobulin class-switch recombination: Choreography of noncoding transcription, targeted DNA deamination, and long-range DNA repair. Adv Immunol. 122:1-57

Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T (2000) Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell. 102:553-563

Muramatsu M, Nagaoka H, Shinkura R, Begum NA, Honjo T (2007) Discovery of activation-induced cytidine deaminase, the engraver of antibody memory. Adv Immunol. 94:1-36

Muramatsu M, Sankaranand VS, Anant S, Sugai M, Kinoshita K, Davidson NO, Honjo T (1999) Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. J Biol Chem. 274:18470-18476

Müschen M, Re D, Jungnickel B, Diehl V, Rajewsky K, Küppers R (2000) Somatic mutation of the CD95 gene in human B cells as a side-effect of the germinal center reaction. J Exp Med. 192:1833-1840

Nambu Y, Sugai M, Gonda H, Lee CG, Katakai T, Agata Y, Yokota Y, Shimizu A (2003) Transcription-coupled events associating with immunoglobulin switch region chromatin. Science. 302:2137-2140

Naughton C, Avlonitis N, Corless S, Prendergast JG, Mati IK, Eijk PP, Cockroft SL, Bradley M, Ylstra B, Gilbert N (2013) Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nat Struct Mol Biol. 20:387-395

Nishimura Y (1985) Sequence dependent DNA conformations: Raman spectroscopic studies and a model of action of restriction enzymes. Adv Biophys. 20:59-74

Nowak U, Matthews AJ, Zheng S, Chaudhuri J (2011) The splicing regulator PTBP2 interacts with the cytidine deaminase AID and promotes binding of AID to switch-region DNA. Nat Immunol. 12:160-166

Okazaki IM, Kinoshita K, Muramatsu M, Yoshikawa K, Honjo T (2002) The AID enzyme induces class switch recombination in fibroblasts. Nature. 416:340-345

Osborne CS, Chakalova L, Mitchell JA, Horton A, Wood AL, Bolland DJ, Corcoran AE, Fraser P (2007) Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. PLoS Biol. 5:e192

Owen JA, Punt J, Stranford SA (2013) Kuby Immunology Seventh Edition. New York, NY: W.H. Freeman and Company

Pan Y, Sun Z, Maiti A, Kanai T, Matsuo H, Li M, Harris RS, Shlyakhtenko LS, Lyubchenko YL (2017) Nanoscale characterization of interaction of APOBEC3G with RNA. Biochemistry. 56:1673-1481

Parsa JY, Ramachandran S, Zaheen A, Nepal RM, Kapelnikov A, Belcheva A, Berru M, Ronai D, Martin A (2012) Negative supercoiling creates single-stranded patches of DNA that are substrates for AID-mediated mutagenesis. PLoS Genet. 8:e1002518

Pavri R, Gazumyan A, Jankovic M, Di Vigilio M, Klein I, Ansarah-Sobrinho C, Resch W, Yamane A, Reina San-Matin B, Barreto V, Nieland TJ, Root DE, Casellas R, Nussenzweig MC (2010) Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. Cell. 143:122-133

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

Petersen-Mahrt SK, Harris RS, Neuberger MS (2002) AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. Nature. 418:99-103

Pham P, Bransteitter R, Petruska J, Goodman MF (2003) Processive AID-catalyzed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature. 424:103-107

Pham P, Calabrese P, Park SJ, Goodman MF (2011) 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. 286:24931-24942

Pham P, Chelico L, Goodman, MF (2007) DNA deaminases AID and APOBEC3G act processively on single-stranded DNA. DNA Repair (Amst.). 6:689-692

Pinaud E, Khamlichi AA, Le Morvan C, Drouet M, Nalesso V, Le Bert M, Cogné M (2001) Localization of the 3' IgH locus elements that effect long-distance regulation of class switch recombination. Immunity. 15:187-199

Qian J, Wang Q, Dose M, Pruett N, Kieffer-Kwon KR, Resch W, Liang G, Tang Z, Mathé E, Benner C, Dubois W, Nelson S, Vian L, Oliveira TY, Jankovic M, Hakim O, Gazumyan A, Pavri R, Awasthi P, Song B, Liu G, Chen L, Zhu S, Feigenbaum L, Staudt L, Murre C, Ruan Y, Robbiani DF, Pan-Hammarström Q, Nussenzweig MC, Casellas R (2014) B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. Cell. 159:1524-1537

Quinlan EM, King JJ, Amemiya CT, Hsu E, Larijani M (2017) Biochemical regulatory features of activation-induced cytidine deaminase remain conserved from lampreys to humans. Mol Cell Biol. 37:pii:e00077-17

Rajewsky K (1996) Clonal selection and learning in the antibody system. Nature. 381:751-758

Ramiro AR, Stavropoulos P, Jankovic M, Nussenzweig MC (2003) Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. Nat Immunol. 4:452-456

Revy P, Muto T, Levy Y, Geissmann F, Plebani A, Sanal O, Catalan N, Forveille M, Dufourcq-Labelouse R, Gennery A, Tezcan I, Ersoy F, Kayserili H, Ugazio AG, Muramatsu M, Notarangelo LD, Kinoshita K, Honjo T, Fischer A, Durandy A (2000) Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the hyper-IgM syndrome (HIGM2). Cell. 102:565-575

Robbiani DF, Nussenzweig MC (2013) Chromosome translocation, B cell lymphoma, and activation-induced cytidine deaminase. Annu Rev Pathol. 8:79-103

Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kenkel TA (2001) Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. Nat Immunol. 2:530-536

Rothenfluh HS, Taylor L, Bothwell AL, Both GW, Steele EJ (1993) Somatic hypermutation in 5' flanking regions of heavy chain antibody variable regions. Eur J Immunol. 23:2152-2159

Senavirathne G, Bertram JG, Jaszczur M, Chaurasiya KR, Pham P, Mak CH, Goodman MF, Rueda D (2015) Activation-indiced deoxycytidine deaminase (AID) co-transcriptional scanning at single-molecule resolution. Nat Commun. 6:10209

Shapiro R, Braverman B, Louis JB, Servis RE (1973) Nucleic acid reactivity and conformation. II. Reaction of cytosine and uracil with sodium bisulfite. J Biol Chem. 248:4060-4064

Shen HM (2007) Activation-induced cytidine deaminase acts on double-strand breaks in vitro. Mol Immunol. 44:974-983

Shen HM, Poirier MG, Allen MJ, North J, Lal R, Widom J, Storb U (2009) The activation-induced cytidine deaminase (AID) efficiently targets DNA in nucleosomes but only during transcription. J Exp Med. 206:1057-1071

Shen HM, Ratnam S, Storb U (2005) Targeting of the activation-induced cytosine deaminase is strongly influenced by the sequence and structure of the targeted DNA. Mol Cell Biol. 25:10815-10821

Shen HM, Storb U (2004) Activation-induced cytidine deaminase (AID) can target both DNA strands when the DNA is supercoiled. Proc Natl Acad Sci USA. 101:12997-3002

Shermoen AW, O'Farrell PH (1992) Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. Cell. 67:303-310

Shlyakhtenko LS, Lushnikov AY, Li M, Lackey L, Harris RS, Lyubchenko YL (2011) Atomic force microscopy studies provide direct evidence for dimerization of the HIV restriction factor APOBEC3G. J Biol Chem. 286:3387-3395

Shlyakhtenko LS, Lushnikov AY, Miyagi A, Li M, Harris RS, Lyubchenko YL (2013) Atomic force microscopy studies of APOBEC3G oligomerization and dynamics. J Struct Biol. 184:217-225

Shlyakhtenko LS, Lushnikov AY, Miyagi A, Li M, Harris RS, Lyubchenko YL (2012) Nanoscale structure and dynamics of APOBEC3G complexes with single-stranded DNA. Biochemistry. 51:6432-6440

Sohail A, Klapacz J, Samaranayake M, Ullah A, Bhagwat AS (2003) Human activation-induced cytidine deaminase causes transcription-dependent, strand-biased C to U deaminations. Nucleic Acids Res. 31:2990-2994

Spencer J, Dunn M, Dunn-Walters DK (1999) Characteristics of sequences around individual nucleotide substitutions in IgVH genes suggest different GC and AT mutators. J Immunol. 162:6596-6601

Storb U, Peters A, Kim N, Shen HM, Bozek G, Michael N, Hackett J Jr, Klotz E, Reynolds JD, Loeb LA, Martin TE (1999) Molecular aspects of somatic hypermutation of Ig genes. Cold Spring Harb Symp Quant Biol. 64:227-234

Storck S, Aoufouchi S, Weill JC, Reynaud CA (2011) AID and partners: for better and (not) for worse. Curr Opin Immunol. 23:337-344

Tanaka A, Shen HM, Ratnam S, Kodgire P, Storb U (2010) Attracting AID to targets of somatic hypermutation. J Exp Med. 207:405-415

Tashiro J, Kinoshita K, Honjo T (2001) Palindromic but not G-rich sequences are targets of class switch recombination. Int Immunol. 13:495-505

Timsit Y (2012) DNA-directed base pair opening. Molecules. 17:11947-11964

Vlijm R, VD Torre J, Dekker C (2015) Counterintuitive DNA sequence dependence in supercoiling-induced DNA melting. PLoS One. 10:e0141576

Wang JC (1979) Interactions between twisted DNAs and enzymes: the effects of superhelical turns. J Mol Biol. 87:797-816

Wang CL, Wabl M (2004) DNA acrobats of the Ig class switch. J Immunol. 172:5815-5821

Webb CF (2001) The transcription factor, Bright, and immunoglobulin heavy chain expression. Immunol Res. 24:149-161

Willmann KL, Milosevic S, Pauklin S, Schmitz KM, Rangam G, Simon MT, Maslen S, Skehel M, Robert I, Heyer V, Sciavo E, Reina-San-Martin B, Petersen-Mahrt SK (2012) A role of the RNA pol II-associated PAF complex in AID-induced immune diversification. J Exp Med. 209:2099-2111

Xerri L, Dirnhofer S, Quintanilla-Martinez L, Sander B, Chan JK, Campo E, Swerdlow SH, Ott G (2016) The heterogeneity of follicular lymphomas: from early development to transformation. Virchows Arch. 468:127-139

Xu Z, Zan H, Pone EJ, Mai T, Casali P (2012) Immunoglobulin class-switch DNA recombination: induction, targeting, and beyond. Nat Rev Immunol. 12:517-531

Xue K, Rada C, Neuberger MS (2006) The in vivo pattern of AID targeting to immunoglobulin switch regions deduced from mutation spectra in msh2-/- ung-/- mice. J Exp Med. 203:2085-2094

Yamane A, Resch W, Kuo N, Kuchen S, Li Z, Sun HW, Robbiani DF, McBride K, Nussenzweig MC, Casellas R (2011) Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. Nat Immunol. 12:62-69

Yoshikawa K, Okazaki IM, Eto T, Kinoshita K, Muramatsu M, Nagaoka H, Honjo T (2002) AID enzyme-induced hypermutation in an actively transcribed gene in fibroblasts. Science. 296:2033-2036

Yu K, Chedin F, Hsieh CL, Wilson TE, Lieber MR (2003) R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. Nat Immunol. 4:442-451

Yu K, Huang FT, Lieber MR (2004) DNA substrate length and surrounding sequence affect the activation-induced deaminase activity at cytidine. J Biol Chem. 279:6496-6500

Zanotti KJ, Gearhart PJ (2016) Antibody diversification caused by disrupted mismatch repair and promiscuous DNA polymerases. DNA Repair (Amst). 38:110-116

Zhang J, Bottaro A, Li S, Stewart V, Alt FW (1993) A selective defect in IgG2b switching as a result of targeted mutation of the I gamma 2b promoter and exon. EMBO J. 12:3529-3537

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31:3406-3415