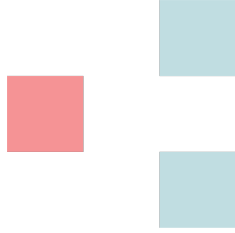
Mutagenic mechanisms in normal and neoplastic B cells: from AID-induced diversification to genome-wide patterns

Sepúlveda Yáñez, J.H.

Reserva Nacional Magallanes, Chile

CHAPTER 1

General Introduction



1.1 | CREATING ANTIBODY DIVERSITY

The adaptive immune system is responsible for the specific recognition and elimination of pathogens and other foreign structures. This recognition is mediated by immunoglobulin (Ig) molecules, which are expressed on the surface of B cells and secreted as antibodies. Immunoglobulins have the capacity to recognize a virtually limitless number of antigens with high affinity and remarkable specificity. However, the mechanism by which B cells manage to create the vast number of receptors remained a major mystery to immunologists for decades. The solution to this problem was revealed in the 1970s when Susumu Tonegawa, who won the Nobel Prize for this discovery, found that germline cells contained multiple gene segments scattered along three chromosomes. These segments, called variable (V), diversity (D), and joining (J), combine to form the heavy (H) and light (L) chain V regions in the Ig genes in a process called V(D)J recombination. [1] During V(D)J recombination, combinatorial and junctional sequence diversity generates variable regions, which are highly variable in their sequences and contain the antigen-binding portions of the antibody. V regions are assembled through V(D)J recombination of VH, DH, and JH genes for the heavy chain, and VL and JL genes for the light chain. This process involves three steps: the recognition of recombination signal sequences (RSS), the generation of double-strand breaks (DSB) by the Recombination-activating genes ***(RAG1–RAG2)***, and finally, the repair of the broken DNA by ligating the strands into a recombined configuration. [2, 3]

V(D)J recombination occurs prior to antigen exposure in the bone marrow but produces antibodies that are low in affinity and usually not effective in inactivating pathogenic organisms. The encounter with the antigen leads to the proliferation, differentiation, and migration of B cells from the bone marrow to secondary lymphoid organs, where the affinity maturation of antibodies occurs in germinal center (GC) reactions through somatic hypermutation (SHM). This process is induced by an enzyme called activation-induced cytidine deaminase (AID). [4]

1.2 | ACTIVATION-INDUCED CYTIDINE DEAMINASE (AID)

Activation-induced cytidine deaminase (AID) was identified by Tasuku Honjo in 1999 in murine B lymphoma CH12F3-2 cells before and after stimulation aimed to induce class switch recombination (CSR). These experiments revealed increased expression of a novel gene transcript in the switch-induced CH12F3-2 cells, the reason why this protein takes its name. [5]

The AID gene ***(AICDA)*** encodes a small 26 kDa protein. The primary structure of AID consists of N-terminal and C-terminal halves connected by a linker sequence, and a cytidine deaminase (CD) catalytic domain. [6] Due to the homology of this catalytic domain with the RNA (mRNA) editing enzyme APOBEC-1, it was initially suggested that AID may function as an RNA editing deaminase rather than as a DNA cytidine deaminase. However, it is now widely accepted in the field that AID functions primarily as a DNA cytidine deaminase.

The function of this protein is essential for two processes in activated B cells: class switch recombination (CSR) and somatic hypermutation (SHM). [7, 8] SHM is a critical biological

process that introduces point and sometimes tandem mutations [9] in the variable region of immunoglobulin (Ig) genes and is responsible for the generation of high-affinity antibodies. This process starts with the DNA deamination of cytosine (C) to uracil (U) by AID during the G1 phase of the cell cycle, although the presence of Us and abasic sites can continue into the S phase. [10, 11]

AID is unique in its processive activity, allowing it to catalyze multiple deaminations of deoxycytidine to deoxyuridine on single-stranded DNA (ssDNA) in a single binding event. This is in contrast to distributive enzymes, which dissociate after each catalytic event. [12, 13] The molecular basis for AID's processivity is not fully understood but is thought to involve specific interactions with the DNA substrate and potential cofactors. AID may undergo conformational changes that enable it to slide along the DNA while remaining bound, thereby facilitating multiple deamination events. [13, 14] This processive activity has profound implications for antibody diversity. By introducing multiple uracils within a localized DNA region, AID sets the stage for a cascade of downstream DNA repair and mutational events, increasing the likelihood of generating multiple mutations within a short stretch of DNA [15, 16], which is crucial for the rapid evolution of antibodies. [17]

There are specific DNA sequences preferentially deaminated by AID. The main motifs are denominated canonical motif WRCY (W = A or T, R = A or G, Y = C or T), non-canonical motif WA (W = A or T), and RCG (R= A or G). [18]

The WRCY motif (W = A or T, R = A or G, Y = C or T) is one of the most well-studied motifs directly targeted by AID during somatic hypermutation (SHM). [13, 19] Initially, the WRCY and its reverse complement RGYW were identified as hotspots for mutations from G and C bases in immunoglobulin genes even before the discovery of AID. A strong bias toward nucleotide substitutions within the quadruplet motif was described in human peripheral B cells. This study indicated that mutations of WRCY accounted for 37% of all nucleotide substitutions, emphasizing its significance in SHM processes. [20, 21] In particular, these motifs are not uniformly distributed, but occur with high frequency in variable regions of immunoglobulins, making them prime targets for AID-mediated mutations. Further studies have shown that mutations at G-C pairs in DGYW/WRCH motifs in yeast share significant similarities with those in immunoglobulin genes, reinforcing the importance of the WRCY motif in SHM. [18, 22] Furthermore, the frequency of WRCY motifs has been found to be concentrated in the complementarity-determining regions (CDRs) of immunoglobulins, which are crucial for antigen binding. [23]

The WA motif, also referred as the non-canonical AID motif, is considered a hallmark footprint left by DNA Polymerase eta (Polη), where the promiscuous action of this polymerase is responsible for mutations in the regions in proximity with the deamination site, which is defined as a signature highly tied with AID activity. [24] Polη is unique among polymerases for its ability to bypass certain types of DNA lesions, such as thymine dimers, which stall other polymerases. This lesion-bypass ability is a double-edged sword; while it allows DNA replication to continue in the presence of damage, it also introduces mutations at a higher

rate than other polymerases. [25] It has been shown that the absence of Polη in patients with gene alterations significantly reduces the number of mutations from A and T in the Ig genes; however, mutations are not completely absent, which could mean that other error-prone polymerases such as Polζ [26], Polι [27] and Polμ [28] may also be involved in somatic hypermutation.

Lastly, the RCG motif was first identified in follicular lymphoma and described as the ability of AID to act at genomic sites containing methylated cytosine within a novel hybrid nucleotide motif that overlaps the CpG methylation site [29] . This suggests that AID-mediated, CpG-methylation-dependent mutagenesis is a common feature not only in lymphomas but potentially in other types of human cancers as well. This motif adds another layer of complexity to our understanding of AID targeting and its role in both normal and pathological conditions.

After the deamination process, several potential scenarios could unfold. On one hand, the U:G mismatches cannot be detected by DNA repair mechanisms and get replicated during S-phase. Alteratively, B cell can activate error-prone mechanisms like BER and MMR or both together in their non-canonical flavors, that can introduce different types of mutations. Half of the mutations observed in the human V region will be C>T substitutions. The other half of the substitution can be from C to bases other than T produced by BER. Furthermore, mutations in A and T that are not the direct action of AID, where MMR repair is involved. [23, 30]

The base excision repair (BER) and mismatch repair (MMR) pathways are important for maintaining cell genomic stability. However, these pathways become error-prone in germinal center (GC) B cells where immunoglobulin genes undergo SHM. As a result, there is a significant increase in the number of mutations that accumulate in the V(D)J regions. In fact, it is estimated that over 50% of the mutations in these regions are not the direct result of the biochemical action of AID, which initiates the mutagenic process, but rather these mutations depend on the error-prone BER and MMR pathways. [31]

1.3 | BASE EXCISION REPAIR (BER)

BER pathway starts with the recognition of U:G mismatchs by uracil DNA glycosylase (UNG2). UNG2 recognizes uracils through specific interactions with single- or double-stranded DNA. [14] BER can also be initiated by SMUG which has been shown to play a UNG2 backup role in SHM. [32, 33] These uracil N-glycosylase proteins flip the uracil out of the DNA helix and into its active site, cleaving the N-glycosidic bond. This action removes the uracil base and leaves behind an abasic site, also known as an AP site. [34]

At this point, the BER pathway can diverge. Typically a high-fidelity repair mechanism, BER becomes error-prone during SHM. This is promoted by apurinic/apyrimidinic endonuclease (APE1) [30] which recognizes the AP site and then cleaves the sugar backbone to produce a single-strand break (SSB). [35] The proteins recruited will depend on the extent of the gap. If it corresponds to one nucleotide, DNA ligase III or XRCC1, part of the ”short-patch” BER

(also called ”single-nucleotide BER”), will be recruited. If the gap is larger (2-10 nt), the long-patch BER mechanism will come into action, with flap endonuclease 1 (FEN1) removing the displaced strand, followed by DNA ligase I or DNA ligase III. [36–39] Specialized error-prone DNA polymerases like Pol ζ and Pol ι are often recruited to these sites. These polymerases are known for their ability to insert bases opposite abasic sites, and their error-prone activity contributes to the diversification of the antibody repertoire. [26, 27]

SHM can also incorporate tandem mutations that are thought to occur when the error-prone polymerases slip or stall at the abasic site, resulting in substitutions at adjacent nucleotides. This phenomenon adds another layer of complexity to the already intricate process of SHM. [40, 41]

1.4 | MISMATCH REPAIR (MMR)

In MMR, the MutSα complex, composed of the MSH2/MSH6 heterodimer, recognizes a single U:G mismatch. Extended portion of mismatches, which could be generated by the deamination of cytosine clusters or by insertions/deletions are recognized by MSH2/MSH3 heterodimer (MutSβ). However, this latter complex does not seems essential for SHM, since the loss of MSH3 has a low impact on SHM levels. [42] Upon encountering a mismatch, the MutSα complex undergoes a conformational change, effectively ”locking” onto the DNA at the site of the mismatch. [43] MutSα then recruits the MutLα complex, consisting of MLH1 and PMS2, to the site of the mismatch. In the context of SHM, the MutLα complex has an unconventional role. Instead of facilitating the repair of the mismatch, it introduces nicks near the mismatch, which serve as entry points for the exonuclease EXO1. [44] EXO1 removes a stretch of nucleotides around the mismatch, creating a gap. [45] The size of the gap created by the excised DNA has been suggested to be as long as 200 base pairs. [46] A monoubiquitinated PCNA is recruited by the MutSα complex to the replication foci. Monoubiquitinated PCNA functions as a molecular switch, determining the recruitment of error-prone translesion synthesis (TLS) polymerases as polymerase Pol η, which amplifies C:G mutations initiated by AID at WRCY hotspots to include A:T mutations primarily at WA hotspots. [47–49] While Polη is often associated with MMR in SHM, other polymerases like Polz, and REV1 can also be involved. These polymerases are error-prone and introduce mutations while filling the gap, contributing to the diversification of the antibody repertoire. [36, 50–52]

Although BER and MMR are related to specific substitution patterns, these two pathways also work synergistically to extend AID-mediated mutagenesis. They represent different ways to deal with the encounter of an uracil in the DNA sequence; while MMR detects it as a simple mismatch, BER utilizes specific proteins that can detect very specific DNA base modifications. UNG can sense and cleave the uracil in single-stranded DNA, even when it is not causing a mismatch. From this perspective, BER is more complementary to the specific action of AID and the level of mutagenic propagation is concise. On the other hand, MMR recognizes the U:G mismatch in double-stranded DNA and is a long patch repair process that can act at a larger distance from the uracil introduced by AID. [11]

1.5 | FOLLICULAR LYMPHOMA

Lymphoid neoplasms are classified in precursor cell and mature cell neoplasm according to the most recent WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. [53] Although the proportion of B and T cells is fairly similar in the human body, the majority of lymphomas (95%) are of B lineage. B-cell lymphomas are frequently named according to the cytomorphological maturation stage, the histological growth pattern, and/or the typical location of the neoplastic cell type within the lymphoid follicle. Important examples of these principles include follicular lymphoma (FL), mantle cell lymphoma (MCL), and diffuse large B-cell lymphoma (DLBCL). [54]

FL is a paradigmatic mature B-cell lymphoma with an age-standardized incidence in Western countries of 2–4 per 100,000 person-years. [55, 56] The average survival of FL patients is approximately 15 years, with most patients showing a slowly progressing form of the disease. [57, 58] Despite this long survival FL it is currently incurable. A subset of patients exhibits a more aggressive clinical course, including those with multiple relapses, high-risk clinical behavior (evidenced by disease progression within 24 months after therapy), or transformation into a high-grade lymphoma, such as DLBCL. Unraveling the biological mechanisms contributing to FL's clinical diversity highly needed.

The origin of FL can be traced back to an aberration during the pre-B cell stage in the bone marrow. This aberration involves a failure in the V(D)J recombination process, leading to the t(14;18)(q32;q21) translocation. [59] The t(14;18) translocation results in the BCL2 oncogene being positioned under the control of IGH regulatory regions, consequently leading to BCL2 overexpression. [60] This translocation is considered the genetic hallmark of the disease and is observed in approximately 85% of FL cases. However, it is important to note that B cells carrying a BCL2 translocation are also present in the majority of healthy individuals. Subsequently, the affected B cell undergoes multiple rounds of somatic hypermutation (SHM) within the germinal center (GC), contributing to the genetic diversification of immunoglobulin genes. [61] As described above, Activation-induced cytidine deaminase (AID) plays a crucial role in this process. [62, 63] The physiological objective of SHM is to generate a B cells that express a BCR with enhanced affinity to cognate antigen. These cells then receive survival signals facilitated by this increased ability to bind and to process antigen on the surface of follicular dendritic cells (FDCs.) [64] The presentation of processed peptides to T follicular helper (Tfh) cells drives antibody affinity maturation. In the context of FL, however, cells expressing BCL2 fail to undergo apoptosis. This impairment allows the persistance of cells with low to moderate BCR affinity and contributes to early lymphomagenesis.

As previously mentioned, the t(14;18)(q32;q21) translocation alone is not sufficient to induce the development of FL. Thus, it is clear that additional factors are necessary for lymphomagenesis to occur. In recent years, two key factors have been identified: genetic factors and the tumor microenvironment (TME).

Within the spectrum of genetic alterations, extensive studies have focused on specific somatic mutations. Somatic mutations have been identified in several pathways, including

signaling pathways with mutations in genes such as ***mTOR***, ***Jak/STAT*** and ***NFkB*** (15-30% of cases), immune recognition (***TNFRSF14***, 28-40% of cases; ***CTSS***, 6% of cases), and cell cycle regulators (10-15% of cases). [65–67] In the last decade, somatic mutations in genes associated with diverse components of epigenetic mechanisms have also shown a high prevalence (>90%). These mutations include inactivating mutations in ***KMT2D***, ***CREBBP***, and ***EP300***, as well as gain-of-function mutations in ***EZH2***. Overall, these epigenetic modifiers appear to contribute to FL pathogenesis by enhancing GC formation, reducing further differentiation, or modulating interactions with the FL microenvironment. [67, 68] Despite advances in understanding the mutational landscape of FL, associations between these mutations and clinical outcomes are less clear compared to DLBCL genetic clusters. On the other hand, the TME appears to be another significant factor in the pathogenesis of FL. Unlike other lymphoma types that grow in diffuse sheets, FL uniquely maintains a three-dimensional structure, whether in the bone marrow or lymph nodes. Thus the TME, including lymphoid cells, stromal cells, and components of the extracellular matrix, seems to play a crucial role. A notable finding is the alteration in T-cell composition, particularly the CD4+ T cell subset. [55] This subset includes T follicular helper (Tfh) cells, which secrete cytokines and chemokines (IL-2, IL-4, CXCL12) promoting the growth and survival of FL cells. [69] There is also a high proportion of regulatory T cells (Tregs) enhancing immune suppression, and a specialized subset of follicular regulatory T cells (Tfr) is prevalent in FL. [70] Advances in high-resolution techniques like CyTOF have identified 'prematurely aged' T cell subsets, characterized by the loss of CD27 and CD28 markers, associated with inferior clinical outcomes. [71] Additionally, mesenchymal stromal cells in the bone marrow have been shown to promote B cell homing through chemokines CXCL12 and CXCL13, further underlining the complex interplay within the TME that promotes FL progression. [72] These cellular and molecular complexities in FL await further elucidation.

1.6 | CHRONIC LYMPHOCYTIC LEUKEMIA (CLL)

Chronic lymphocytic leukemia (CLL) is a lymphoid malignancy characterized by the monoclonal expansion of more than 5000 mature CD5 B lymphocytes per μL of peripheral blood. [73] CLL is the most common leukemia in adults in Western countries, with an age-adjusted incidence rate of 4.9 cases per 100,000 people per year. With a pronounced difference between the incidence in men (6.8 cases per 100,000/year) and women (3.5 cases per 100,000/year). [74] CLL can be classified according to the mutational status of the B-cell receptor in two major subgroups: a subgroup with the IGHV genes unmutated (U-CLL) which the identity with the germline reference is $\geq$ 98% and a mutated (M-CLL) subgroup. The cells of origins differ between these two groups, U-CLL originates from unexperienced germinal center B cells and M-CLL post germinal center B cells. [75]

CLL always passes through a precursor condition stage, called CLL-phenotype monoclonal B lymphocytosis (MBL), before progression to overt CLL. Conversely, only a small minority of MBL eventually develops CLL. [76] Indeed, the risk of progression to CLL is only 1-2% per

year, and the risk of developing other B-cell malignancies is 0.1-0.5% per year. [77] MBL is characterized by the subclinical expansion of mature B cells with CD5+CD20lowCD79low CLL phenotype. MBL is further classified into low-count MBL (LC-MBL) and high-count MBL (HC-MBL) according to the number of cells in the peripheral blood. [78] MBL as a precursor condition of CLL represents an interesting model to study the early stages of the disease, in which the factors that drive the progression to CLL are still not clear.

In the last decade, genomic alterations have been associated with the pathogenesis of CLL. Germline variants have show to increasing 2-8 fold the risk of developing the disease. Forty five susceptibility loci have been identified through genome-wide association studies. [79] The majority of these loci are located in non-coding regions, with 93% mapping to active promoters or enhancers. These susceptibility variants lead to changes in the binding site of transcriptional factors in genes associated to immune response, cell survival, or Wnt signaling. [80]

The somatic mutational landscape of CLL is highly heterogeneous with no more than five genes (***SF3B1, ATM, TP53, POT1,*** and ***NOTCH1***) being recurrently mutated in at least 5% of patients; the vast majority are mutated at lower frequencies. [81] A more homogenous distribution of alterations is observed in copy number variants, with over 80% of the patients carrying at least one. The most relevant are trisomy 12, and deletions of 13q (del(13q)), 11q (del(11q)), and 17p (del(17p)). [82] Since these structural variants are acquired early in disease progression, they are clonal markers, as are a subset of mutations in driver genes. [83] Consequently, the detection of these structural variants has become standard of care in CLL diagnostics. CLL is driven by a particular mechanism not present in other types of indolent B-cell lymphomas. This mechanism is autonomous BCR signaling of the clonotypic BCR, resulting in activation of the BCR signaling cascade in the absence of any external ligand. [84] Autonomous signaling occurs through inter- or intramolecular interactions between the leukemic CDR3 regions of one BCR complex and intrinsic motifs located in the FR2 and FR3 regions of another BCR complex. It has been demonstrated that the capacity for autonomous BCR signaling is a prerequisite for CLL development. [85] The crucial role for specific properties of the BCR for CLL development and survival is also indicated by striking structural similarities in the BCR between CLL cases, also known as stereotypes. The largest stereotype subset is subset #2L, in which somatic hypermutation creates a mutation at position 110 of IGLV3-21-containing BCR light chains, which in itself is sufficient to induce autonomous signaling. [86] This particular mutation is associated with an adverse outcome and this point mutation it is present in 7-18% of the patients. [87]

1.7 | AIM OF THE STUDY

AID plays a crucial role in B-lymphocyte maturation and increasing evidence suggests that it might contribute to lymphomagenesis. However, the extent of AID's role in different lymphomas, as well as its relation with DNA repair error-prone mechanisms in the context of SHM, has not been fully established. This thesis aims to elucidate AID's role across normal B cells and in lymphomas like Follicular Lymphoma (FL) and Chronic Lymphocytic Leukemia (CLL), while also exploring the critical oncogenic process of autonomous signaling present in CLL. Our approach includes a detailed analysis of AID activity across genomic, chromatin structure, transcriptional dimensions, and BCR functional testing.

The study of SHM has traditionally focused on single nucleotide substitutions, insertions, and deletions. However, elucidating the occurrence of AID-instigated contiguous or tandem dinucleotide substitutions (TDNS), especially in human models, presents significant challenges. In **chapter 2**, we characterize TDNS in immunoglobulin genes by applying an in silico model designed to distinguish tandem mutations from independently occurring adjacent single nucleotide substitutions (SNS). This study was performed using immunoglobulin samples from both healthy donors and patients with DNA repair deficiencies. The primary objective of this work is to decipher the mechanisms involved in tandem mutation generation and to propose a model that explains this phenomenon in human immunoglobulin genes.

With the advent of new technologies such as single-cell sequencing, we have opened up new opportunities to study AID-associated mutations in the context of somatic hypermutation at single cell level. In **chapter 3**, our aim is to address a theoretical question about the possibility of detecting two different transcripts of Ig within a single cell in a B cell, considering the recent activity of AID and its association with specific transcriptional profiles that may be active in these cells. We chose follicular lymphoma as an optimal model due to its high frequency of SHM. Additionally, we used CLL as a control due to its lower SHM levels.

The most frequent indolent B-cell neoplasia are follicular lymphoma and chronic lymphocytic leukemia. FL constitutively expressed AID and present high levels of SHM. In contrast, only a subset of CLL cases exhibit AID activity, as indicated by the discrete SHM in their immunoglobulin genes. In **chapter 4**, we aim to identify the differential contributions of AID activity to the mutational landscapes of FL and CLL. This is achieved through a mutational signature analysis at both a global (genomic) and targeted (immunoglobulin) level using a combination of de novo mutational signature extraction and COSMIC signature fitting. We integrate the mutational signatures identified with three-dimensional chromatin data from B cells to define their specific chromatin compartmentalization. Furthermore, we analyze mutations in the DNA repair pathways of these lymphomas.

Chronic Lymphocytic Leukemia exhibits unique mechanisms not found in other lymphoma types. This includes the signaling of CLL cells through clonotypic B-cell receptors (BCR) expressed autonomously, without the engagement of external antigens. This factor, along with recurrent genetic aberrations, plays a significant role in the pathogenesis of CLL. In **chapter 5**, we aim to study these two factors in a precursor condition of CLL known as

Monoclonal B-lymphocytosis (MBL). Our approach involves first screening siblings of CLL patients for the presence of MBL with a CLL phenotype. This includes identifying and quantifying autonomous signaling. We also aim to compare the prevalence of inherited risk loci and acquired genetic aberrations between CLL patients and their first-degree relatives with MBL.