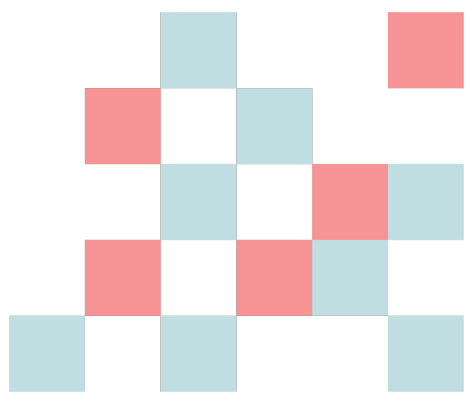
Mutagenic mechanisms in normal and neoplastic B cells: from AID-induced diversification to genome-wide patterns

Sepúlveda Yáñez, J.H.

Monte Tarn, Chile

CHAPTER 6

Summary and General Discussion



6.1 | SUMMARY

B cells have the unique capacity to recognize potentially any molecular structure in our world that constitutes an antigen through their receptor. The formation of the limitless variety of B-cell receptors with high affinity is possible due to a unique capability of B cells that seek to create mutations instead of avoiding them. This process is called somatic hypermutation (SHM) and needs two main components, the activation-induced cytidine deaminase (AID) and the DNA repair machinery. AID is a protein that produces the deamination of cytosine to uracil in DNA, the first step of SHM. In the second step, two DNA repair pathways, base excision repair and mismatch repair are involved but act in a non-canonical manner. In SHM, these DNA repair pathways are not utilized to repair the deamination faithfully, instead they permit the occurrence of mutations. The usual outcome of this process is potentially any kind of single base substitution, but this process can also create tandem mutations, i.e. two or more adjacent mutations in the DNA sequence and are produced in a single event. The presence of tandem mutations in immunoglobulin genes has been observed in several species, but has until now not been described in human ***ex vivo*** models. In **chapter 2**, we sought to detect and characterize tandem dinucleotide substitutions (TDNS) in immunoglobulin genes from peripheral blood B-cells of healthy donors and patients with DNA repair deficiencies.

Our first step to achieve this aim was to develop a computational model to distinguish tandem mutations from independently occurring adjacent single nucleotide substitutions (SNS). We generated a synthetic immune repertoire with the same V allele usage and of matching size with our BCR sequencing data. This repertoire was mutated using the same mutation rate and substitution type for each position of the original repertoire. These calculations were performed 100.000 times each for all rearrangements. Since this method introduced one mutation at a time, all observed tandem mutations were considered ”false” positive tandems. By applying this observation as correction, 46,2% of the originally detected tandem were removed, and 5775/10747 tandem mutations were maintained for the subsequent analysis. We concluded that 6% of the total mutations detected in the immunoglobulin genes represented tandems (dinucleotide until pentanucleotide), and the incidence of TDNS specifically was 4,1%.

Next, we characterize the TDNS in the V region of immunoglobulin genes topographically. We observed that the global distribution of TDNS was similar to the distribution of SNS but with some differences in FR regions. Using mutational resistance scores, we showed a negative selection of non-synonymous mutations in FR, especially in structurally essential FR residues, indicating that TDNS in FR would be counterselected since their higher potential for non-synonymous/replacement mutations could easily lead to deleterious effects on BCR integrity. As expected, TDNS have a higher potential for non-synonymous/replacement mutations than SNS, and we observed that 98,6% of TDNS encoded for at least one amino acid replacement, while only 71,6% of SNS encoded a non-synonymous mutation. Interestingly, we observed that half of all TDNS occurred in dipurine or dipyrimidine motifs and most of them (89,6%) were using one or both nucleotides from the reference. This implies that the tandem substitution

could be derived from small juxtaposition events and would explain the high incidence of inverting substitutions.

To understand which DNA repair mechanism was involved in this juxtaposition process, we evaluated patients with deficiency in MMR (4 patients) and UNG (1 patient), as well as an independent cohort of healthy donors. First, we observed that the incidence of TDNS in healthy donors was still higher than previously reported (1,9%) with a similar distribution to the first healthy donor cohort. In the MMR deficiency patients, we observed a similar amount of tandem (2,0%) than the healthy donors, suggesting that the mismatch repair machinery does not play a major role in the formation of tandem substitutions in humans. In contrast, the UNG-deficient patient showed no tandem substitutions, suggesting that the majority of tandem substitutions in humans depend on UNG activity in the BER pathway. Unfortunately, we were unable to validate this finding in more UNG-deficient patients who usually die in the first years of life. Nevertheless, we have been able to characterize and propose a cause for how tandem mutations are generated in humans.

Each B cell displays a unique B-cell receptor (BCR) on its surface, whose coding immunoglobulin gene becomes subject to modification of one to few nucleotides every time that a somatic hypermutation event takes place. This process must theoretically result in the transient coexistence of two immunoglobulin transcript populations within a single cell, a phenomenon unobserved in human single B cells until now. In **chapter 3**, we aimed to detect and quantify occurrent SHM events in individual B cells, including evidence of the active SHM mechanisms, including activation-induced cytidine deaminase and DNA repair pathways. We performed single-cell RNA and B-cell receptor sequencing on follicular lymphoma, characterized by their arrest at the germinal center stage with ongoing AID expression, and on chronic lymphocytic leukemia cells, representing a non-GC malignancy derived from either naïve or memory B cells as a control.

Initially, we established a custom cell-level variant calling pipeline using scBCR-seq data from 12 FL and 5 CLL patients. The variant calling process was followed by several filtering steps, including quality metrics assessment, doublet identification and removal, and setting thresholds for the number of transcripts and reads supporting each position (at least 10 UMIs and 5 reads), variant frequency (minimum of 20%), and exclusion of events where both transcripts are found in other cells. Using this strict filtering strategy, we detected 364 cells with two immunoglobulin gene transcripts differing by one or a few nucleotides in 7 out of the 12 FL cases, totaling 1239 SHM events. In contrast, no occurrent SHM events were detected in CLL cells. Interestingly, 268 of the 364 cells had at least one of both VDJ/VJ BCR transcripts with more than one alternative nucleotide at different positions. In 93,9% of these cells exist two main populations of transcripts, one with the most frequent variant combinations and the other with the respective alternative nucleotide at all positions.

Once the presence of occurrent SHM had been established, our next goal was to delineate the chronological development of these events and to establish the hierarchy (between original and neovariant transcripts) among the two subpopulations of immunoglobulin transcripts.

We were able to determine this hierarchy for 1,055 of the 1,239 identified SHM events. When we analyzed the type of mutations, we observed that C-to-T transitions were the most common (40.66%) events in occurrent SHM, consistent with an independent whole exome sequencing (WES) cohort and overall established V(D)J mutations detected in single FL cells. Furthermore, we noted that 27.7% of the SHM events, using the original nucleotide as a reference, occurred in the canonical AID motif (WRCY), compared to 15.6% when considering the neovariants. Similarly, the destruction of the WRCY motif was more significant when starting from the original nucleotide (79.3%) than from the neovariant (63.2%). The greater presence and destruction of the WRCY motif with the original nucleotide, statistically significant according to the Chi-square test ($p<0.001$), supports the accuracy of our original versus neovariant nucleotide annotations.

Finally, we performed a gene expression analysis using the scRNA-seq data comparing cells exhibiting occurring SHM events to those without. This analysis identified 5,766 genes differentially expressed between the two groups. Notably, activation-induced cytidine deaminase was upregulated, whereas ADAR, another deaminase but acting in RNA level, was downregulated in cells with SHM events. This supports role of AID in deaminase activity. To further explore the biological role of these differentially expressed genes, we performed an enrichment analysis. This revealed three SHM-related pathways (DNA replication, MMR, BER) that were mainly upregulated in cells with occurrent SHM and also clustered together according to the network functional representation. Since the KEGG database lacks a comprehensive gene set representing the entire SHM machinery, we created a custom annotation for the SHM pathway. Our findings indicate that the MMR pathway is more prominent than the BER pathway in cells with neovariants supported by a higher normalized enrichment score (NES) in MMR (MMR:1.93, BER:1.6). In this way, we were able to identify and quantify ongoing somatic hypermutation in single B cells for the first time.

AID plays a crucial role in generating mutations within immunoglobulin genes as part of their physiological function, as well as in other, off-target genes. As previously evaluated, the deamination reaction catalyzed by AID does not occur at random DNA sequences but is targeted to specific sequence contexts such as WRCY, WA, and RGC motifs, which are associated with downstream pathways involved in the mutation process. The contribution of AID to different indolent lymphomas has not been established in the immunoglobulin genes, along with the global genomic effect of this enzyme. Therefore, in **chapter 4**, we delineated the differential contributions of AID activity to the mutational landscapes of FL and CLL.

In our initial observations, we found that the frequency of somatic mutations in FL exceeds those in CLL, yet the six substitution patterns remain similar between both lymphoma types, with a predominant C-to-T substitution comprising 37.6% of all mutations. Despite similar proportion of AID motif contributions in these lymphomas, with the greatest contribution from WA, followed by WRCY, and then RGC, the overall contribution is higher in FL (42.7%) than in CLL (33.6%). Through mutational signature analysis at both the global (genomic) and targeted (immunoglobulin) levels, employing a combination of de novo mutational signature extraction and COSMIC signature fitting, we discovered that FL exhibits contributions from

AID-related signatures as SBS9, SBS84 and the novo signature GC. In contrast, CLL shows a higher prevalence of a cancer-related signature such SBS5, which is also present in FL, but with a smaller contribution.

Subsequently, we integrated the mutational signatures identified with three-dimensional chromatin data from B cells (lymphoblastoid B cell lines, GM12878) to elucidate their specific chromatin compartmentalization. This revealed an association between active compartments in the chromatin with signatures 3 and 6, which were observed in the global genomic landscape, and signature 84, identified within the immunoglobulin genes. Conversely, CLL does not show a distinct compartmentalization for these identified mutational signatures.

Finally, our analysis of mutations within DNA repair pathways revealed that FL has mutations linked to Fanconi anemia pathway and BER, whereas CLL is characterized by mutations in the DDR pathway. This allowed us to analyze the role of AID in the mutational landscape of FL and CLL.

So far, we have described the contribution of AID-induced mutations in a physiological context through the generation of tandem mutations in the immunoglobulin genes. We also captured the occurrence of single cell SHM events in follicular lymphoma, as well as the overall impact of AID in the mutational landscape of follicular lymphoma and chronic lymphocytic leukemia. In **chapter 5**, we aimed to understand the dynamics and hierarchy between genetic aberrations and a specific immunological driver mechanism acting in CLL, known as autonomous BCR signaling, in the progression from monoclonal B-cell lymphocytosis to chronic lymphocytic leukemia.

Through screening 191 siblings of CLL patients, we detected 34 individuals with clonal CLL-phenotype cells (CD19+CD5+CD20lowCD19low), equivalent to 17,8% of all siblings of CLL patients. Out of these, 32 were classified as low-count MBL with fewer than 500 clonal CLL-phenotype cells per µL. The BCR from the immunoglobulins of 17 low-count MBL siblings were sequenced, and we observed that 5 MBL clones could be assigned to CLL stereotypes. Furthermore, virtually all sequenced BCR contained one or both of the FR2 VRQ and FR3 YYC motifs proposed as structural requirements for autonomous BCR signaling in the majority of CLL cases.

Eleven of these 17 sequenced MBL BCRs were expressed in TKO cells and tested for calcium mobilization prior and after tamoxifen-induced reconstitution of the BCR signaling cascade and after BCR crosslinking. We observed that all tested MBL BCR from siblings of CLL patients exhibited autonomous BCR signaling activity, including the 5 BCRs assigned to CLL stereotypes. There is a discrete but significantly higher intensity of BCR signaling in CLL compared to MBL. This difference in BCR signaling strength directly correlated with the expansion of the MBL clone at the time of sampling.

In the genetic analysis of the MBL and CLL samples, 26 germline variants associated with increased incidence of CLL revealed a significantly higher prevalence of twelve risk alleles within our cohort. The initial analysis of 24 polygenic risk scores was higher in both CLL and MBL compared to the general population. However, this was not the case in the

extended analysis, where 5-8 CLL risk loci assessable in whole exome sequencing and SNP array data were incorporated. Next, we performed a CLL-associated Copy Number Variation (CNV) analysis in 15 MBL samples and 11 CLL samples. We found that 10 MBL and 11 CLL carried recurrent CLL-associated CNV. Although, in the case of CLL, the CNVs were clonal in the majority of cases (9 out of 11 cases), while in MBL, the CNVs were frequently subclonal (10 out of 11 cases). Finally, we performed a WES analysis in 10 CLL-MBL sample pairs to detect SNVs and indels. Initially, we removed variants that could be shared between siblings, which potentially could be germline variants, we referred to these as non-shared variants. In the first instance, we focused on 120 genes recurrently mutated in CLL according to the COSMIC database. It was determined that CLL and MBL present similar amount and VAF of the genes analyzed. Nevertheless, in a global analysis of the non-shared variants present in CLL and MBL samples, we observed that the prevalence of subclonal variants was significantly lower in MBL samples than in CLL cases. This allowed us to consider the contribution of two oncogenic mechanisms, i.e. genetic and immunological drivers, during the transition from MBL to CLL.

In summary, we have characterized the presence of tandem mutations in the immunoglobulin genes of healthy donors and patients with DNA repair deficiencies. We have also detected and quantified the occurrent SHM events in individual B-cells, providing evidence of the active SHM mechanism, including activation-induced cytidine deaminase and DNA repair pathways. We have delineated the differential contributions of AID activity to the mutational landscapes of FL and CLL, and we have described the dynamics and hierarchy between genetic aberrations and a specific oncogenic mechanism present in CLL, known as autonomous BCR signaling, in the progression from monoclonal B-cell lymphocytosis to chronic lymphocytic leukemia.

6.2 | GENERAL DISCUSSION

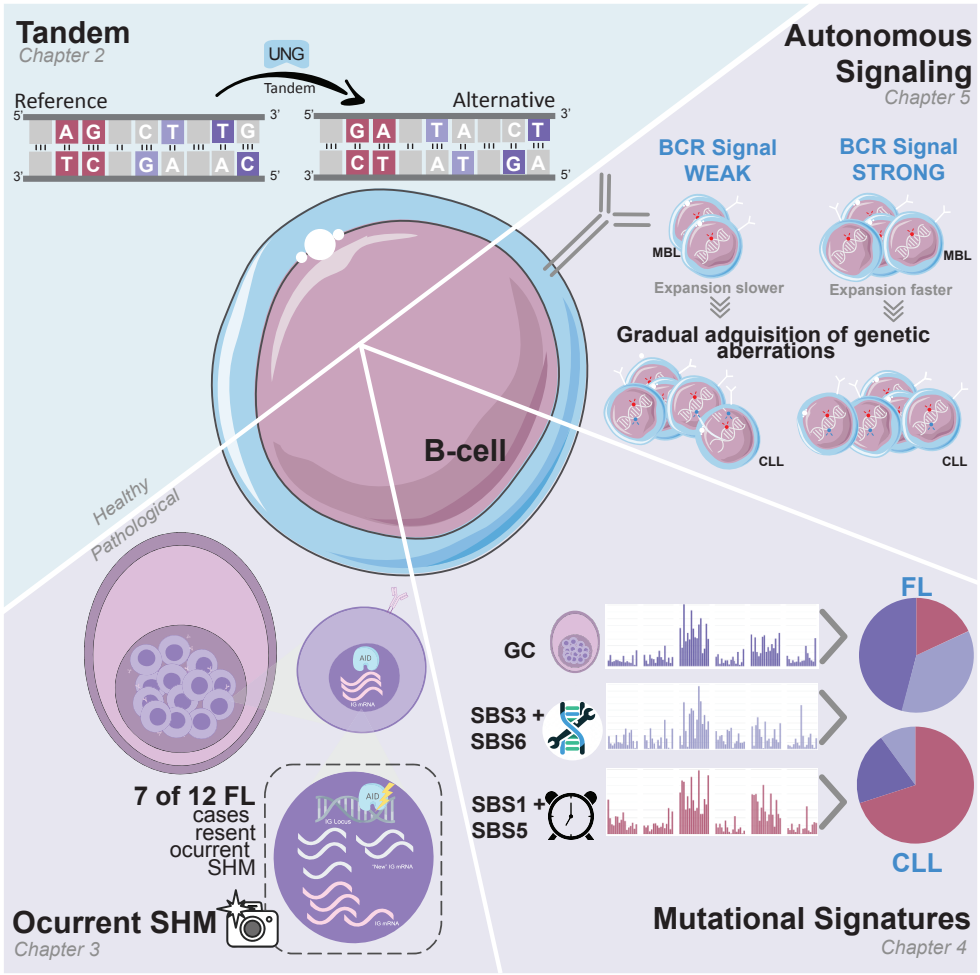


Figure 6.1. Integrated Overview of the Thesis. From Characterization of AID-Induced Mutations to the Oncogenic Processes in B-Cell Neoplasias. Chapter 2 discusses the characterization of tandem DNA substitutions in immunoglobulin genes. Chapter 3 investigates AID-associated mutations at the single-cell level. Chapter 4 compares AID's mutational impact in FL and CLL via mutational signature analysis. Chapter 5 examines autonomous BCR signaling and genetic risk in CLL and MBL.

THE CONTRIBUTION OF AID TO IMMUNE DIVERSIFICATION THROUGH TANDEM AND CONCURRENT MUTAGENESIS

Activation-induced cytidine deaminase plays a key role in the generation of antibody diversity, primarily by introducing somatic hypermutation and class switch recombination in immunoglobulin genes. Traditionally, SHM has been quantified at the cell population level

as the cumulative outcome of sequence alterations in Ig genes among selected B cells. This approach estimates the frequency of SHM in V(D)J recombination to be around 10^{-3} per base pair per generation in murine germinal center (GC) B cells, corresponding to roughly one mutation per cell division. Our findings substantially refine this view by describing two mechanisms that AID can use to generate multiple mutations simultaneously: The creation of adjacent tandem mutations and the extent of non-adjacent yet concurrent mutations.

Tandem substitutions refer in this discussion to the simultaneous alteration of two or more adjacent nucleotides in a DNA sequence. AID-induced tandems, particularly tandem dinucleotide substitutions, have been described in immunoglobulin genes of several species. The incidence of AID-induced tandem substitutions in human B cells differs enormously between species, from 1,6% in mice [122] to nearly 60% in sharks [123, 124]. In humans, the incidence of tandem dinucleotide substitutions in genome-wide was reported to range from 0,1% to 1% [125], but at the level of immunoglobulin genes, this information is lacking. In **chapter 2**, we observed that the incidence of tandem substitutions in the immunoglobulin genes of healthy donors was 6%, with 4,1% of these being tandem dinucleotide substitutions. This is higher than previously reported in humans at the genome-wide level. And this frequency may be even higher because there may be tandem with one of the bases that matches the reference that is not detected as a tandem event.

In addition to the high frequency of tandem substitutions, intrinsically this phenomenon increases the chance of amino acid replacement mutations. This was the case in almost all tandem events that we observed (98.6%). Therefore, tandem mutations are a mechanism to increase the efficiency of SHM in creating diversity of the antibody repertoire.

Recent advancements in single-cell sequencing technologies have established novel opportunities to study SHM at the single-cell level. In **chapter 3**, we have taken advantage of these technologies for the detection and quantification of recent SHM events in individual B cells. We were not only able to capture recent SHM events as hypothesized, but we also unveiled the capability of AID to perform not just singular, but multiple concurrent mutations in 73,62% of the B-cell with SHM events. This observation challenges the longstanding presumption that the action of AID is limited to introducing one mutation at a time [166]. The creation of concurrent mutations by AID most likely contributes to the high mutation burden of immunoglobulins genes acquired during the development of follicular lymphomas.

These insights into the broader mutagenic capabilities of AID indicate a more dynamic and versatile role in generating antibody diversity than previously appreciated. In particular, the ability to induce tandem and concurrent mutations may accelerate antibody affinity maturation and broaden the potential antibody diversity substantially.

THE IMPORTANCE OF THE NUCLEOTIDE CONTEXT FOR AID ACTIVITY

AID activity is preferentially constrained to specific nucleotide contexts. A canonical motif, WRCY [19, 40], exists, but also a non-canonical motif, WA [24] and the RCG motif identified originally in FL [29]. Analysis of the sequence context of somatic mutations can reveal the

footprint of AID activity. In **chapter 3**, we detected a high proportion of somatic mutations within AID motifs in FL and CLL (42% and 33% respectively). This is also the case for the tandem subtraction in **chapter 2**, where a strand bias analysis revealed a high contribution in the template strand of WRCY and WA motifs. This contrast with the contribution of mutations in these motifs in other cancer types as breast cancer (10%) and lung cancer (5%). This observation supports the essential role of AID in lymphomagenesis of FL and CLL [218, 220]. Unexpectedly, the neovariants detected in **chapter 2** exhibit a higher frequency within AID motifs (50.2%) compared to what was observed in FL samples in **chapter 3**. This discrepancy might be attributed to the fact that the neovariants represent recent mutations that have not yet been subjected to a selection process. A similar pattern is observed when we analyzed the substitution type: We observed that C-to-T transitions were the most common of occurrent mutations at genomic level for CLL and MBL in chapter 3 and similarly for neovariants in the immunoglobulin genes in **chapter 2**. This specific substitution type, C-to-T, can originate from two distinct processes, either by direct replication or by repair through the DNA repair machinery. The existence of these two different pathways to generate the same C-to-T mutation type may increase the likelihood of the occurrence of such transitions.

THE DUAL INFLUENCE OF AID AND BCR AUTONOMOUS SIGNALING IN CLL AND FL ONCOGENESIS

We have been deepening our understanding of the role of AID in creation of antibody diversity through the generation of multiple patterns of mutations. AID-induced mutagenesis in immunoglobulin genes is a physiological process, but it can also result in pathogenic consequences in the function of the BCR. In **chapter 5**, we explored a specific oncogenic mechanism present in CLL that originates within the BCR itself, known as autonomous signaling, and its role in the progression from MBL to CLL. Autonomous signaling can be the direct consequence of the action of AID by somatic hypermutation as well as by primary V(D)J recombination in cases with unmutated BCR. In our study, we were able to compare CLL-MBL sibling pairs and we observed the unequivocal presence of autonomous signaling in MBL. Furthermore, the intensity of BCR signaling correlated directly with the expansion of the MBL clone at the time of sampling. This observation suggests that the strength of BCR signaling is a key factor in the progression from MBL to CLL. However, it remains elusive whether structural features of the BCR, such as HCDR3 in conjunction with motifs in FR2 and the interaction of specific amino acids [84, 262], influence the strength of BCR signaling. Additionally, other genetic, epigenetic, and microenvironmental factors could be involved in modulating the signal. [296]

Other important oncogenic mechanisms present in B-cell malignancies are genetic/genomic alterations. In **chapter 4**, we zoom out to the global genomic landscape of FL and CLL, and we identify the main mutational processes involved in these lymphomas. The mutational signatures can be identified by non-negative matrix factorization (NMF) deconvolution techniques using the global contribution of mutations as classified by their trinucleotide context. By this approach, a specific mutational signature can be linked to a certain biological mechanism, which may be exogenous, such as tobacco exposure and UV radiation, or endogenous,

involving factors like AID, APOBEC, and DNA repair pathways. [137] Using unbiased deconvolution techniques, we were able to detect three mutational signatures, two corresponding to a mix of different mechanisms due to the complex signature extraction process [246] and a third unique signature. The first identifed signatures corresponds to a mix of the COSMIC database signatures 3 (SBS3) and 6 (SBS6), which are associated with defects in homologous recombination and DNA mismatch repair, respectively [137, 297]. The second extracted signature corresponds to a mixture of signature 1 (SBS1) associated with aging and signature 5 (SBS5) associated with carcinogenic process, both defined as clock-like signature [137, 251, 297]. The last signature does not correspond to the signatures reported by COSMIC, but shows a remarkable dominance of AID-related motifs. In particular, we were able to detect signatures 84 (SBS84) and 85 (SBS85), which are directly related to AID activity, only by targeted analysis of the immunoglobulin genes. This highlights the importance of choosing the right approach to extract mutational signatures. When we analyzed the contribution of all three extracted signatures by lymphoma type, we observed a high contribution of SBS5 in CLL. In contrast, FL presented a higher contribution of AID-related signatures as SBS9, SBS84, and the novo GC signature. This observation suggests that even when these two lymphomas share some mutational processes, their relative contributions are different and may be influenced by other oncogenic mechanisms such as BCR signaling in the case of CLL.

Further exploring genetic changes in B-cell malignancies, we found a higher number of mutations in FL compared to CLL (**chapter 4**). [298]. In contrast, closer examination of variant allele frequencies as performed in **chapter 3 and 5** reveals a higher variant allele frequency for many mutations in CLL than in FL, which is not immediately apparent at first glance, and also in MBL. The latter observation is consistent with the idea that the progression from MBL to CLL is associated with the acquisition of additional genetic alterations that drive clonal expansion [267].

THE COMPLEX INTERPLAY OF DNA REPAIR MECHANISMS IN AID-INDUCED MUTAGENESIS

Following AID-catalyzed deamination, the DNA repair machinery becomes responsible for dealing with the lesion in the DNA, primarily in an error-prone fashion. In **chapter 2**, we evaluated the role of these pathways in the formation of tandem substitutions in the immunoglobulin genes. We observed that the incidence of tandem substitutions in patients with mismatch repair (MMR) deficiency was similar to that in healthy donors. However, the only available patient with uracil DNA glycosylase (UNG) deficiency showed no tandem substitutions. This suggests that the majority of tandem substitutions in humans depend on UNG activity within the BER pathway, in contrast to reports in mice where the MMR pathway is primarily responsible for the formation of tandem substitutions [122, 129]. Analysis of the reference nucleotide context surrounding the tandems (4 nt in total) revealed that commonly the observed tandem is already present in the reference sequence, suggesting that tandem substitutions could arise from small juxtaposition events, leading to a high incidence of inversion substitutions. Considering all these factors, we proposed a molecular model for tandem

generation in the immunoglobulin genes in humans called EXPEDITE (EXtruded Pinching Effecting DIrectional Tandem Exchange) model. This model is based on a known functional property of UNG, which correspond to its ability to extrude uracil from the DNA helix. This extrusion may be followed by a pinching effect, possibly leading to a small juxtaposition event [146]. Subsequently, a translesion polymerase such a POLH will carry out DNA synthesis in the presence of the extruded site, after this the DNA strand will return to its original position. Cleavage of the AP site by an AP endonuclease creates a gap that is then faithfully repaired, resulting in tandem substitution. However, this model does not fully explain a specific common tandem type (AG to GA), suggesting that MMR may also contribute to forming this type of tandem substitution, as observed in mice [129].

In **chapter 3**, we observed through enrichment pathway analysis that MMR pathway is more prominent than the BER pathway in cells with neovariants supported by a higher normalized enrichment score (NES) in MMR (MMR:1.93, BER:1.6). This observation contrasts with the results obtained in the **chapter 2**, where we observed that the majority of tandem substitutions in humans depend on UNG activity in the BER pathway. Maybe this observed difference with respect to the DNA pathways involved in the generation of neovariants can be explained by the relatively frequent occurrence of somatic mutations in BER genes in FL biopsies (30% of FL cases, **chapter 3**) which could affect the activity of the BER pathway. In any event, our findings emphasize the complexity of DNA repair mechanisms in shaping the genomic landscape of B cells and their contribution to the generation of an unexpectedly broad spectrum of AID-induced mutations.

ADAPTING BIOINFORMATICS TOOLS FOR COMPREHENSIVE B-CELL STUDY

In the course of this thesis, we have designed and implemented a wide range of bioinformatics tools for the analysis of B cells. In **chapter 3**, we implemented a mutational signature approach that combines de novo extraction and a posteriori fitting against recognized signatures from the consensus of signatures in the COSMIC database, evaluated by non-negative least squares (NNLS). This analysis was performed both globally (genomic) and localized to the immunoglobulin genes. This approach also emerged as a way to solve the lack of consensus on the analysis and interpretation of these types of analyses in the field [122]. One of the main challenges in the extraction of mutational signatures is the task of selecting the appropriate number of signatures that can be extracted from a set of samples. First, the number of signatures is intrinsically related with the number of samples in the set. For example, if a set consists of 30 samples is not possible to extract more than 7 signatures accurately, according to cancer genome simulations [210]. Second, the process of signature extraction also is affected by the number of mutations by sample. Hypothetically, if the goal is to extract 7 or more signatures from a set of 50 samples, each sample (genome) should on average contain at least 1000 mutations. Third, the type of tumor type included in the analysis also has to be taken into account: While there are certain kinds of mutational processes expected to be involved in different types of tumor, there are others that are specific for particular tumor type such as activation-induced cytidine deaminase (AID) in B-cell lymphomas. To solve this problem,

we used the R package SigProfiler to extract the mutational signatures; this package uses a Bayesian approach to estimate the number of signatures and method considers for example the available number of samples. This tool has been used previously in the extraction of mutational signatures in cancer genomes and has been shown to be more accurate than other methods [234].

In **chapter 3**, we designed a custom cell-level variant calling pipeline for scBCR-seq data to be able to detect variants in immunoglobulin transcripts within a cell. A particular challenge related to samples of FL cells is the high SHM rate in the immunoglobulin genes. A large number of individual variants at transcript level cause the standard pipelines used for the analysis as Cell Ranger to fail in the required alignment steps. As a result, the alignment will yield low numbers of reads with presumably highly biased sequences, which tremendously limits the process of calling cell-level variants. To overcome this problem, we performed the alignment using Minimap2 and defined individual references for each patient. This approach successfully increased the number of reads and the number of detected cells with neovariants. Another challenge was to build consensus sequences per cell/UMI; an output not generated by pipelines specifically designed for immune repertoire analysis such as scRepertoire [299].

In the bioinformatics field, is not unusual to encounter suboptimal input data due to problems ranging from budget limitations over experimental design issues to limited availability of suitable samples. In **chapter 5**, we encountered the problem of lack of sample availability specifically for control samples for the genomic analysis of CLL and MBL cells. The lack of control samples led us to pool control samples with similar genetic backgrounds and sequencing techniques to our main samples, a method inspired by existing strategies in the field [300]. This approach, refined by excluding variants common among siblings, enabled a more precise genetic differentiation between CLL and MBL. This shows how bioinformatics approaches always have to be flexible and adaptable to the specific needs of the study.

POTENTIAL IMPLICATIONS, LIMITATIONS AND FUTURE RESEARCH

The understanding the molecular mechanisms involved in somatic hypermutations in B cells could contribute to the development of novel immunotherapy strategies. Aberrant somatic hypermutations and tandem mutations may lead to the formation of neoantigens. By uncovering these mutations and the pathways responsible for their generation, this research could help identify tumor-specific antigens, facilitating the development of more precise and effective treatments in cancer immunotherapy.

On the other hand, one of the limitations of this research is the lack of functional validation of the pathways involved in tandem mutation generation. Future studies could focus on validating the EXPEDITE model and identifying the enzymes involved in tandem mutation generation in humans. In our study of B cells performing SHM at the single-cell level, we were unable to explain how this mechanism evolves during the B-cell cycle. Further research could explore the cell cycle stages of B cells experiencing SHM events.

Our mutational signature analysis was limited to FL and CLL samples. Expanding this

analysis to other B-cell malignancies could be a possible direction for future research. Due to constraints related to informed consent, it was not possible to follow up with MBL patients to assess progression to CLL. A longitudinal study on MBL patients could help validate the role of autonomous BCR signaling in progression to CLL and clarify how signal strength contributes to this progression.An additional aim would be to understand the structural features of the BCR that trigger autonomous signaling. By identifying the specific structural configurations responsible for autonomous signaling, using, for example, machine learning algorithms, it could be possible to predict whether a BCR will trigger autonomous signaling and, in the best-case scenario, anticipate the likelihood of progression to CLL.

In summary, this thesis investigated the influence of AID on physiological B-cell diversification and its involvement in lymphoma, focusing on FL and CLL. The novel data highlight the role of AID and error-prone DNA repair pathways in the generation of adjacent and distant tandem mutations, in the creation of occurrent neovariants, and in genome-wide mutational signatures. Finally, the occurrence of genome-wide changes are explored in the context of immunological signalling, specifically autonomous BCR signaling in the oncogenesis of CLL. The use of bioinformatics tools for advanced analysis of high-throughput sequencing data yields unprecedented insights into the causes and consequences of genomic plasticity in normal and neoplastic B cells.