Universiteit
Leiden
The Netherlands

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

Sepúlveda Yáñez, J.H.

Citation
Sepúlveda Yáñez, J. H. (2024, November 12). *Mutagenic mechanisms in normal and neoplastic B cells: from AID-induced diversification to genome-wide patterns*. Retrieved from https://hdl.handle.net/1887/4108983

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4108983

Note: To cite this publication please use the final published version (if applicable).

MUTAGENIC MECHANISMS IN NORMAL AND NEOPLASTIC B CELLS:

From AID-induced diversification to genome-wide patterns

Julieta Haydeé Sepúlveda Yáñez

- ISBN: 978-94-6496-263-5
- Cover: Julieta Haydeé Sepúlveda Yáñez
- Lay-out: Julieta Haydeé Sepúlveda Yáñez & Diego Alvarez-Saravia
- Printed by: Gildeprint Enschede | The Netherlands

Images on the cover are in the public domain or licensed under Creative Commons Zero (CC0).

© 2024 J.H. Sepúlveda Yáñez All rights reserved. No part of this thesis may be reproduced, stored in a retrieval system of transmitted in any form or by any means, without prior permission of the author.

The research described in this thesis was financially supported by Doctorado Becas Chile (2016-72170683), ANID.

MUTAGENIC MECHANISMS IN NORMAL AND NEOPLASTIC B CELLS:

From AID-induced diversification to genome-wide patterns

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 12 november 2024
klokke 13:00 uur

door

Julieta Haydeé Sepúlveda Yáñez
geboren te Punta Arenas, Chile
in 1987

Promotor:

Prof.dr. J.H. Veelken
Prof.dr. M.A. Navarrete Signorile (Universidad de Magallanes)

Promotiecommissie:

Prof. dr. P.C.W. Hogendoorn (Secretaris)
Prof. dr. C.J.M. van Noesel (Amsterdam UMC)
Prof. dr. A.W. Langerak (Erasmus MC)
Dr. M. Griffioen

”Do not believe anything merely because presumption is in its favor, or because the custom of many years inclines you to take it as true. Do not believe anything merely on the authority of your teachers and priests. But, whatever, after thorough investigation and reflection, you find to agree with reason and experience, as conducive to the good and benefit of one and all and of the world at large, accept only that as true, and shape your life in accordance with it.”

Buddha

Contents

CHAPTER 1 GENERAL INTRODUCTION 3

1.1 Creating antibody diversity . . . 5
1.2 Activation-induced cytidine deaminase (AID) . . . 5
1.3 Base excision repair (BER) . . . 7
1.4 Mismatch repair (MMR) . . . 8
1.5 Follicular lymphoma . . . 9
1.6 Chronic lymphocytic leukemia (CLL) . . . 10
1.7 Aim of the study . . . 12

CHAPTER 2 TANDEM SUBSTITUTIONS IN SOMATIC HYPERMUTATION 15

2.1 Introduction . . . 17
2.2 Materials and methods . . . 18
2.3 Results . . . 21
2.4 Discussion . . . 30
2.5 Supplementary material . . . 34

CHAPTER 3 OCCURRENT SHM AND ASSOCIATED GENE EXPRESSION PROFILE IN SINGLE B CELLS 57

3.1 Introduction . . . 59
3.2 Materials and methods . . . 59
3.3 Results . . . 62
3.4 Discussion . . . 69
3.5 Author contributions . . . 72
3.6 Competing interests . . . 72
3.7 Supplementary material . . . 73

CHAPTER 4 MUTATIONAL SIGNATURE ANALYSIS IN INDOLENT LYMPHOMAS 85

4.1 Introduction . . . 87
4.2 Materials and Methods . . . 88
4.3 Results . . . 90
4.4 Discussion . . . 95
4.5 Conclusions . . . 99

4.6 Supplementary material 100

CHAPTER 5 AUTONOMOUS SIGNALING AND GENETIC ABERRATIONS IN MBL 117
5.1 Introduction 119
5.2 Materials and methods 119
5.3 Results 121
5.4 Discussion 125
5.5 Supplementary material 131

CHAPTER 6 SUMMARY AND GENERAL DISCUSSION 145
6.1 Summary 147
6.2 General discussion 152

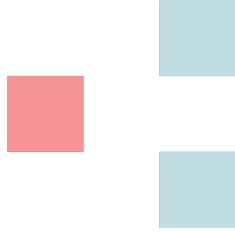
APPENDIX 161
Nederlandse samenvatting 163
List of publications and Conference presentations 169
Curriculum vitae 173
Acknowledgement 175

BIBLIOGRAPHY 177

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.

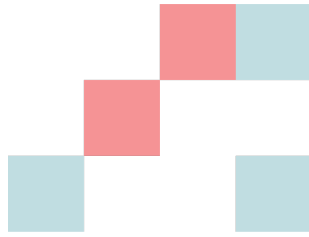
Lago Pehoe, Torres del Paine, Chile

CHAPTER 2

Tandem Substitutions in Somatic Hypermutation

Previously published as

Sepúlveda-Yáñez, J. H., Alvarez Saravia, D., Pilzecker, B., van Schouwenburg, P. A., van den Burg, M., Veelken, H., Navarrete, M. A., Jacobs, H. & Koning, M. T. Tandem Substitutions in Somatic Hypermutation. ***Frontiers in Immunology*** **12,** 807015. doi:10.3389/fimmu.2021.807015 (2022)



ABSTRACT

Upon antigen recognition, activation-induced cytosine deaminase initiates affinity maturation of the B-cell receptor by somatic hypermutation (SHM) through error-prone DNA repair pathways. SHM typically creates single nucleotide substitutions, but tandem substitutions may also occur. We investigated incidence and sequence context of tandem substitutions by massive parallel sequencing of V(D)J repertoires in healthy human donors. Mutation patterns were congruent with SHM-derived single nucleotide mutations, delineating initiation of the tandem substitution by AID. Tandem substitutions comprised 5,7% of AID-induced mutations. The majority of tandem substitutions represents single nucleotide juxtalocations of directly adjacent sequences. These observations were confirmed in an independent cohort of healthy donors. We propose a model where tandem substitutions are predominantly generated by translesion synthesis across an apyramidinic site that is typically created by UNG. During replication, apyrimidinic sites transiently adapt an extruded configuration, causing skipping of the extruded base. Consequent strand decontraction leads to the juxtalocation, after which exonucleases repair the apyramidinic site and any directly adjacent mismatched base pairs. The mismatch repair pathway appears to account for the remainder of tandem substitutions. Tandem substitutions may enhance affinity maturation and expedite the adaptive immune response by overcoming amino acid codon degeneracies or mutating two adjacent amino acid residues simultaneously.

2.1 | INTRODUCTION

To effectively counter the virtually limitless possibilities of pathogen-derived immune challenges, B lymphocytes – representing a critical arm of the adaptive immune system – can generate a virtually limitless repertoire of structural B-cell antigen receptor (BCR) variants through somatic hypermutation (SHM) [89–95]. Upon encountering an antigen, antigen-activated B cells initiate SHM, as well as class-switch recombination, through activity of activation-induced cytosine deaminase (AID) [5, 96].

AID deaminates cytosine (C) to uracil (U) preferentially in nucleotide motif WRCY (where W denotes A or T; R denotes A or G; and Y denotes C or T) on both DNA strands [97–101]. This deamination locally instigates various substitutions through various mutagenic processing pathways [100–104].

When the U remains unmodified it will instruct a template T to all polymerases, resulting in C to T and G to A transitions. However, a U in the DNA is usually efficiently detected by uracil DNA glycosylase (UNG), which cleaves the base from the sugar-phosphate backbone, thereby generating a non-instructive apyrimidinic (AP) site. The AP site normally initiates faithful base excision repair (BER) involving an AP endonuclease (APE) and POLB [105, 106]. During SHM however, the AP-site can serve as a non-instructive template for the translesion synthesis (TLS) polymerase REV1, a dCMP transferase that can only insert a C opposite the newly generated AP site, thereby enabling C to G and G to C transversions [107–109]. Alternatively, a single strand break at the AP site generated by APE allows POLH in complex with monoubiquitinated homotrimeric DNA clamp and replication processivity factor PCNA (PCNA-Ub) to access the site and generate about 8% of all A/T mutations by error-prone long-patch BER [94, 110–113].

Otherwise, the uracil is recognized as a U/G mismatch by the MSH2-MSH6 mismatch recognition complex and initiates the formation of a single-stranded gap flanking the U/G mismatch [104, 114]. Upon binding of the MSH2-MSH6 complex, exonuclease 1 excises the uracil-containing strand to initiate non-canonical mismatch repair (ncMMR), where POLH in complex with PCNA-Ub generates the majority of A/T mutations [111, 113, 115–118].

Taken together, low fidelity translesion synthesis (TLS) DNA polymerases enable the generation of a large part of the spectrum of nucleotide substitutions. In these short- and long-patch "repair" pathways, error-prone TLS DNA polymerases such as REV1, POLH, POLZ and perhaps POLI introduce mutations both at the position of the deaminated cytosine as well as in its near vicinity [115, 119–121].

In principle, all these mechanisms lead to single nucleotide substitutions (SNS). However, the occurrence of AID-instigated contiguous or "tandem" substitutions, especially tandem dinucleotide substitutions (TDNS), has been described for several species. The term tandem substitution refers to contiguous substitutions resulting from a single DNA damage event, which could in theory lead to multiple discrete single nucleotide substitutions during its repair. The contribution of tandem substitutions to all AID-induced substitutions differs per species, ranging from 1.6% in mice [122] to nearly 60% in sharks [123, 124]. Data on the frequency of

this phenomenon from human ex vivo experiments in the immunoglobulin loci are currently lacking, but genome-wide the frequency of TDNS is estimated to range from 0.1% to 1% [125–128].

The biological consequences of tandem substitutions remain unknown. One advantage for TDNS is that they overcome the redundancy in the amino acid code in positions in the BCR where non-synonymous mutations are beneficial. This would in theory allow for faster repertoire diversification and hence more effective affinity maturation and humoral immunity [122, 129].

The molecular mechanism underlying the generation of tandem substitutions is debated. Evidently, a proportion of the observed contiguous substitutions are not, in fact, single event tandem substitutions, but multiple independent SNS occurring in contiguous nucleotides [130]. This may hold especially true in and around canonical AID hot spot motifs, where consequently SHM is most efficient [124]. It appears however, that the incidence of tandem substitutions is far higher than expected from clustered SNS alone, indicating the existence of a mechanism creating contiguous substitutions in a single event [122].

Despite its potentially substantial role in the adaptive immune response, the mechanisms responsible for tandem substitutions and their relative contribution to the process of SHM has not been investigated in humans since the introduction of massive parallel sequencing. Here, we investigated the incidence of tandem substitutions in human peripheral blood B cells (PBMC) in healthy donors and patients with DNA repair deficiency and analysed their substitution motifs. We aimed to confirm association of previously implicated DNA repair mechanisms involved in the creation of tandem substitutions, as well as identify other potential mechanisms. Additionally, we investigated whether tandem substitutions indeed overcome amino acid redundancy by analysing the topographical distribution of TDNS in the V allele.

2.2 | MATERIALS AND METHODS

SAMPLE COLLECTION AND PREPARATION

Peripheral blood samples were obtained with written informed consent from twelve healthy stem cell donors in compliance with the biobanking regulations of Leiden University Medical Center. Mononuclear cells were isolated by Ficoll separation and cryopreserved in aliquots. B cells were purified from aliquots of thawed cells by removal of non-B cells with magnetic beads (B cell isolation kit II; Miltenyi Biotec, Leiden, The Netherlands), routinely yielding a purity of $>$ 99% CD19+ B cells as assessed by flow cytometry.

V(D)J LIBRARY GENERATION

Aliquots of $2x10^6$ B cells were processed according to the ARTISAN PCR protocol for unbiased amplification of BCR repertoires, which has very low amplification and sequencing error rates

of 0.126×10^{-3} [131]. Full-length IgM and IgG VDJ were amplified from all twelve donors, while IgA and IgE VDJ, VJ-kappa and VJ-lambda were amplified from six donors each [132]. Libraries were barcoded, pooled and amplified as single molecules in rolling circles on a total of fourteen SMRT cells on the RSII system (Pacific Biosciences, Menlo Park, CA, USA). Output sequence files were filtered with SMRT portal software for a minimum of eight sequencing passes. All sequences were annotated by IMGT HighV-QUEST [133].

Additional datasets of healthy donors, as well as UNG-deficient and MSH2 and MSH6-deficient patients were obtained from publicly available sources [134, 135] (Supplementary Table S1). Its amplification and sequencing methodology was previously validated and shown to be reliable for the identification VDJ of naïve and memory B cells alike [136].

SEQUENCE SELECTION

Sequences with identical V, D (if applicable), and J genes, identical CDR3 length and $\geq$ 95% pairwise identity in the nucleotide CDR3 sequence were analysed as a single sequence. Clonal expansions were reduced to the least mutated sequence to minimise the presence of amplification or sequencing errors. For the Leiden cohort healthy donors, sequences with no mutations, as well as sequences with $>$ 5% mutations in their V region were excluded from further analysis to minimize chance occurrences of consecutive mutated nucleotides whilst maintaining a reasonable number of mutations per sequence. Such selection of oligomutated sequences could not be performed for the publicly available datasets, as these datasets contained solely (often highly mutated) IgG and IgA rearrangements which, upon equally stringent selection, would have led to greatly reduced library sizes, precluding any meaningful analysis [132]. From the publicly available data, only sequences containing one of the 50 most commonly rearranged IGHV alleles were considered, offering the option to completely perform in silico correction of these sequences.

For fair comparison, sequences with >5% mutations in the Leiden cohort were grouped in bins of 5-10% mutations, 10-20% mutations and >20% mutations and analysed identically to the less mutated sequences (Supplemental Data). Because the main findings in these data sets in essence corresponded to the findings in the less mutated sequences, yet were more susceptible to methodological imperfections such as less accurate ‘“false tandem” correction (due to lower ratios of observed:simulated TDNS, see below), these sequences were not incorporated in the main analysis.

IDENTIFICATION OF MUTATIONS

We identified all SNS, TDNS, and longer contiguous substitutions in the V region of the 50 most commonly used IGHV alleles, 25 most commonly used IGKV alleles, and 25 most commonly used IGLV alleles, IGLV9-49*01 or IGLV10-54*01 (the latter two to have all IGLV families represented). We identified the germline nucleotide(s), mutated nucleotide(s), context (1 nucleotide on either side of the substitution), and their position within the V allele using

Geneious software v10.1.3 (Biomatters Ltd., Auckland, New Zealand). Insertions and deletions were disregarded in this analysis.

The similarity analysis between TDNS and the Doublet Base Substitution (DBS) Signatures deposited in the COSMIC catalog (v3.1 - June 2020) [137] was measured by cosine distance using R [138].

PREDICTED FREQUENCY OF CONTIGUOUS SUBSTITUTIONS FROM MULTIPLE MUTATION EVENTS

The contribution of contiguous substitutions resulting from two or more independently occurring adjacent SNS, was modelled in silico. To have adequate numbers of control sequences, only the 20 most abundant V alleles were used for modelling.

Considering the observed mutation frequency of SNS per position and the distribution of particular substitutions per position, germline sequences were mutated in silico to a total number of mutations matching the mutation load of sequences observed in the in vivo data. These calculations were performed 100,000 times each for all rearrangements. Output sequences were analysed for single and contiguous substitutions, SNS and TDNS motifs and nonsense mutations. Any mutation clusters thus observed were considered to represent "false" tandem substitutions and were subtracted from the actually observed frequency of tandem substitutions to obtain the true frequency of tandem substitutions in the dataset. Calculations for the complete dataset were extrapolated from this majority subset.

For the detailed modelling strategy, see: `https://github.com/CATGUMAG/tandem-substitution-simulation`.

The CMMRD patient repertoires mostly lacked IgM and both the patient and Rotterdam healthy donor repertoires were sequenced using a different approach and sequencing platform than the Leiden healthy donor cohort, which resulted in higher mutation loads and more false TDNS. Therefore, we did not pool these repertoires in our in silico mutation algorithm, but processed them separately and then independently predicted SNS mutation clusters as described for the Leiden cohort.

The paucity of unique sequences in the UNG-deficient patient's library precluded detailed in silico correction as performed for the other datasets. To obtain the best approximation of corrected tandem substitution incidence in this dataset, in silico predictions were performed on pooled sequences of the IGHV3 and IGHV4 family.

MUTATIONAL RESISTANCE SCORING

A score was assigned to each position in IGHV, IGKV or IGLV to represent the chance that any random single nucleotide substitution will cause a synonymous mutation. For each codon in the human amino acid code, all 9 possible substitutions were classified as synonymous, non-synonymous or nonsense mutations. Since nonsense mutations would normally cause the BCR to be deleted from the repertoire and therefore not be represented in this dataset,

these substitutions were disregarded and the chance of a random substitution leading to an amino acid change was calculated from the remaining options. Consecutively, for each position in the V region, proportional germline codon usage was calculated and multiplied by the chance that a codon would incur a synonymous mutation upon any random substitution. The results of all codons per position were added together to obtain the repertoire-broad chance of synonymous mutations per position (Supplementary Table S2).

This theoretically expected proportion of synonymous mutations was compared with the observed proportion of synonymous single nucleotide substitutions. When the observed frequency of synonymous mutations is higher than expected, this might indicate that BCR with non-synonymous mutations in this position have undergone negative selection and are no longer found in the repertoire. Conversely, lower observed synonymous mutation rates than expected may indicate a selective advantage for BCR altering their sequence in this position. However, it should be noted that this method assumes random substitutions across the V region whereas in reality mutations preferentially cluster around AID motifs and some substitutions occur preferentially over others [139]. This method is therefore more suited for analysis of V region-wide topographical distribution of mutations rather than scrutiny of a single position.

After calculating the permissiveness of every position to incur non-synonymous mutations, the number of TDNS per position was counted and compared to the pattern of SNS to assess whether tandem substitutions are distributed randomly or preferentially cluster in certain areas of the V region.

We applied a Quasi-Poisson regression to establish if the number of tandem substitutions for each position in the V allele is correlated with the mutational resistance score. The analysis was performed using the R Stats Package [138].

2.3 | RESULTS

6% OF SUBSTITUTIONS ARE TANDEMS

Initially, twelve healthy donor peripheral blood B-cell receptor repertoires were sequenced using full-length, unbiased, massive parallel V(D)J sequencing. Selection of unique, clonally unrelated, antigen-experienced sequences carrying up to 5% mutations yielded 13,532 VDJ, 7,952 VJ-kappa and 7.598 VJ-lambda. Comparison to the closest germline allele allowed for identification of a total of 122,878 single nucleotide substitutions (SNS), 10,735 tandem dinucleotide substitutions (TDNS) and 2,615 longer contiguous substitutions. The longest contiguous substitutions were 8 nucleotides in length (Supplementary Table S3).

Since mutation clusters of independently generated, but adjacent SNS are indistinguishable from single event tandem substitutions, we calculated the number of expected clustered SNS for each individual repertoire through in silico modeling (IGHV Figure 2.1; IGKV and IGLV and pooled IGV Supplementary Figure S1; Supplemental Data 1.1-1.23). Using the 20 most abundantly observed IGV alleles, synthetic immunoglobulin repertoires of matching size and V

allele usage distribution were mutated 100.000 times in silico. These in silico repertoires were designed to exactly match the mutation rate for each nucleotide position in the sequenced repertoire. Since all mutations within this model were introduced in single steps, observed mutation clusters in the modelled data were annotated as “false” tandem substitutions.

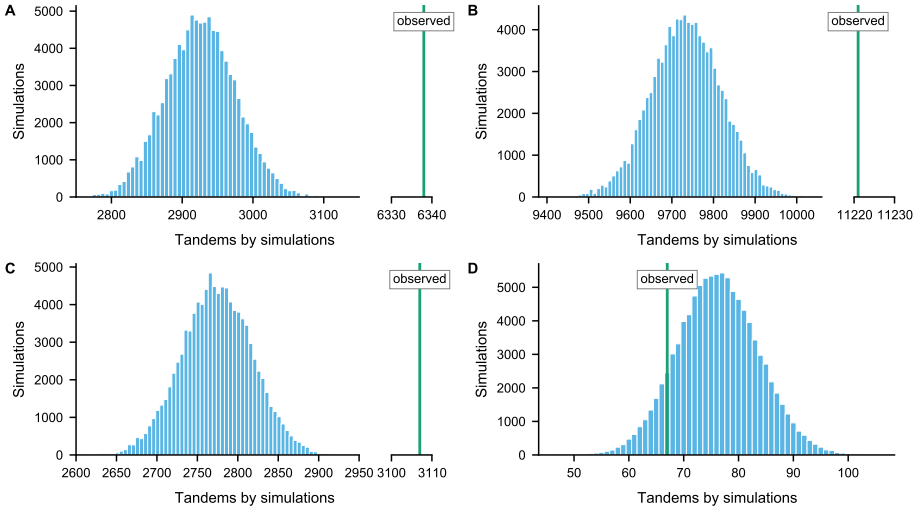


Figure 2.1. Comparison between the tandems found in the simulated datasets (distribution of 100,000 simulations) and observed values for each cohort. **(A)** Leiden healthy cohort (IGHV). **(B)** Rotterdam healthy cohort (IGHV). **(C)** Rotterdam MSH2/6 deficiency cohort (IGHV). **(D)** Rotterdam UNG deficiency cohort (IGHV). The distance between the distribution and the observed value indicates the number of tandem mutations after correction. Z-score analysis was performed to compare distributions.

Predicted “false” tandem substitutions were subtracted from the total number of observed mutation clusters to obtain the frequency of single event tandem substitutions. Of all TDNS in the in silico simulated subset, 46.2% were predicted to represent two adjacent SNS generated in independent events. Extrapolating to the whole dataset (and counting the “false” tandem substitutions as multiple SNS each), we observed 133,577 SNS and 5,775 TDNS. Therefore, the incidence of true, single event TDNS in human BCR in vivo was found to be 4.10%. Similar calculations for trinucleotide, tetranucleotide and pentanucleotide contiguous substitutions respectively showed incidence rates of 1.18%, 0.31% and 0.12%, containing 12.5%, 3.0% and 1.0% false positives, respectively. Cumulatively, 5.7% of all substitutions in human BCR were single event tandem substitutions, making their incidence rate much higher than observed in mice and in genomic human sequences outside the immunoglobulin loci (Supplementary Table S4).

THE DISTRIBUTION OF TDNS MATCHES SNS

The distribution of all SNS and TDNS occurring in the most abundantly rearranged V alleles was mapped geographically up to IMGT amino acid position C104. As expected, mutation

frequencies in complementarity determining regions (CDR) exceeded those of framework regions (FR).Specifically, structurally essential FR residues such as C23, W41 and C104 were strikingly less mutated than surrounding residues. Overall, the distribution of TDNS resembled those of SNS, indicating that mechanisms governing the generation and selection of TDNS are closely related to those of SNS (Figure 2.2, Supplementary Figure S2A and Supplementary Figure S3A. Also, the SNS follow an equal distribution pattern of mutations (Ts/Tv: 1,05) corresponding to previously described datasets (53) (Supplementary Table S5A). Despite their overall similarities, the relative abundance of TDNS differed from SNS in a number of positions. TDNS were overrepresented in a number of FR residues in the IGLV dataset (Figure 2.2, Supplementary Figure S2B) and Supplementary Figure S3B). More commonly, however, FR residues contained fewer TDNS than expected. This relative scarcity of TDNS was most profound around structurally essential FR residues, where the observed SNS generally encoded for synonymous mutations (Figure 2.2, Supplementary Figure S2C and Supplementary Figure S3C). We hypothesized that TDNS in FR would be selected against, given how their higher potential for non-synonymous/replacement mutations (see below) would easily lead to deleterious effects on BCR integrity.

To test this hypothesis, we predicted the proportion of synonymous mutations at each position (Materials and Methods) and compared this to the observed proportion of synonymous mutations at that position. Subtracting the former from the latter resulted in a mutational resistance score, where progressively higher scores represented positions with more synonymous mutations than expected. Putatively, positions with high mutational resistance scores result from negative selection of BCR with non-synonymous mutations in these positions.

Indeed, we observed relatively few TDNS in positions with higher mutational resistance scores and we confirmed a negative selection of tandem substitutions in these structurally important positions using a Quasi-Poisson regression (Figure 2.2, Supplementary Figures S2D, S3D, Supplementary Table S6, p-value<0,05).

TANDEM MUTATIONS OVERCOME CODON REDUNDANCIES

As more nucleotides are mutated at once, the chance of non-synonymous, i.e. amino acid replacement mutations, increases. To determine how often this occurred, TDNS were assigned to three groups according to their position in the codon: position 1 for substitutions of the 5' base in the coding strand and the middle base, position 2 for the middle and the coding strand 3' base, and position 3 for the 3' base of one codon and the 5' base of the subsequent codon. Substitutions in position 2 may only cause non-synonymous or nonsense mutations, whilst substitutions in position 1 and 3 may rarely cause synonymous mutations. Overall, 98.6% of TDNS encoded for at least one amino acid replacement, while only 71.6% of SNS encoded a non-synonymous mutation. Apparently, TDNS have the potential to greatly expedite amino acid changes and thereby are likely to enhance adaptive immune responses.

Additionally, TDNS in position 3 may potentially mutate two adjacent residues simultaneously, which was observed in 18.5% of such position 3-situated TDNS. It appears, that this

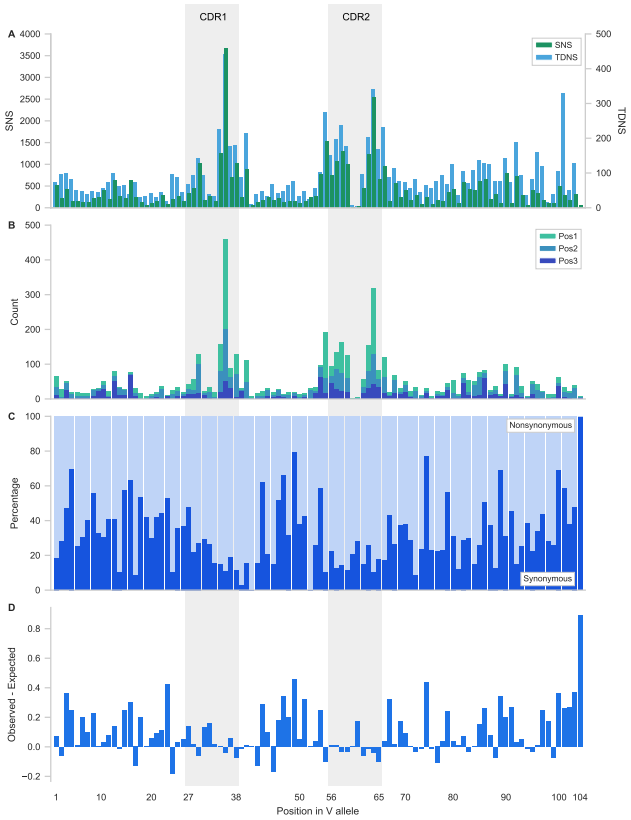


Figure 2.2. Distribution of tandem dinucleotide substitutions in the IGHV allele. **(A)** Distribution of single nucleotide substitutions (SNS) and tandem dinucleotide substitutions (TDNS) in the V allele. Amino acid positions 30-32 and 60-61 in CDR1 and CDR2 respectively, are present in only a small minority of V alleles and therefore mutations in these positions are rare events when considering the whole sequence library. **(B)** Nucleotide specific context within a codon. Relative contributions of tandem dinucleotide substitutions for VDJ. Results are split per position in the codon. Position 1 represents 5' and middle base on the coding strand, position 2 represents middle and 3' base and position 3 represents the 3' base of the codon and the 5' base of the downstream codon. **(C)** Proportion of synonymous and nonsynonymous mutations per position in the V allele. (D) Mutational resistance score. Plotted are the expected minus the observed frequencies of nonsynonymous mutations per codon position in the V allele. Increasingly higher values represent residues that are more resistant to mutation.

potential for accelerated affinity maturation comes at a price, because position 3-situated TDNS were less abundant than the other positions, supposedly due to negative selection of structurally unsound BCR or nonsense mutations. An exception was found in the IGLV library, where some hotspots that were described above, were situated in position 3 and compensated for the lower abundancy overall (Figure 2.3). Whether such relative scarcity of position 3 TDNS indeed resulted from increased negative selection pressure, could not be tested in this

reverse immunology investigation.

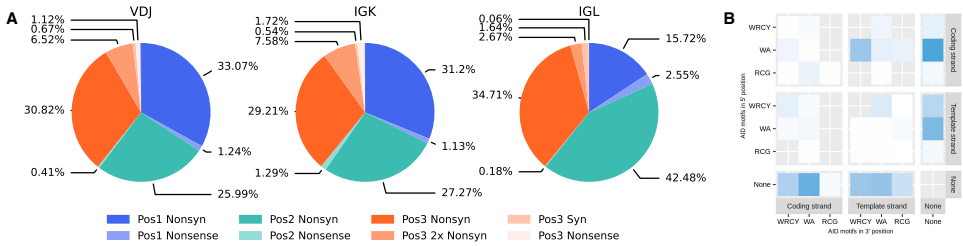


Figure 2.3. Relative contributions of synonymous, non-synonymous and nonsense tandem dinucleotide substitutions for VDJ, VJ-kappa and VJ-lambda. **(A)** Results are split per position in the codon. Position 1 represents 5' and middle base on the coding strand, position 2 represents middle and 3' base and position 3 represents the 3' base of the codon and the 5' base of the downstream codon. Tandem dinucleotide substitutions causing two separate synonymous mutations in position 3 are displayed as a separate category. **(B)** Analysis of AID motifs present in tandem substitution for IGHV sequences. The 3 motifs associated with AID activity (WRCY, WA, and RCG) were identified in position 5' or 3' of the tandem. The motifs were considered in forward (coding) and reverse (template) direction.

MOST TANDEM MUTATIONS ARE JUXTALOCATIONS

After subtraction of in silico predicted "false tandem" TDNS, substitution tables were generated for IGHV, IGKV and IGLV collections and for the complete dataset (IGV Figure 2.4; IGHV, IGKV and IGLV Supplementary Figure S4; Supplementary Tables S7-S9 and Supplemental Data 1.1-1.23). Heavy and light chain substitution tables were comparable, except for CT to TA and GA to AG TDNS, which were more prevalent in IGLV, and IGKV+IGLV, respectively. We postulate that these observations stem from differences in the reference sequences, in which these bases may be more prone to mutation, or such mutations are more permissible and lead to less negative selection.

TDNS did not occur in equal frequencies, nor did they follow substitution patterns expected from SNS tables. Instead, half (49,9%) of all TDNS occurred in dipurine or dipyrimidine motifs and virtually all mutations swapped the positions of either one of the reference bases (i.e. AG to CA; 51,5% versus 44,4% expected; Student's t-test: $p<0,0001$), or both bases at once, leading to inversions (i.e. AG to GA; 37.3% versus 11.1% expected; Student's t-test: $p<0,0001$). Only 10,4% (expected: 44,4%; Student's t-test: $p<0,0001$) of TDNS did not follow this pattern (Figure 2.4 and Supplementary Table S10A).

Additionally, we compared the TDNS signature with the recently described doublet base substitutions (DBS) signatures from COSMIC (v3.1 - June 2020). We found that the tandem substitution signature in immunoglobulins does not correspond to a previously described pattern (cosine similarity < 0.5, Supplementary Table S11).

Whenever a TDNS reference sequence was frequently found in the TDNS table, their reverse complement would also be highly prevalent. However, within such pairs, the motifs

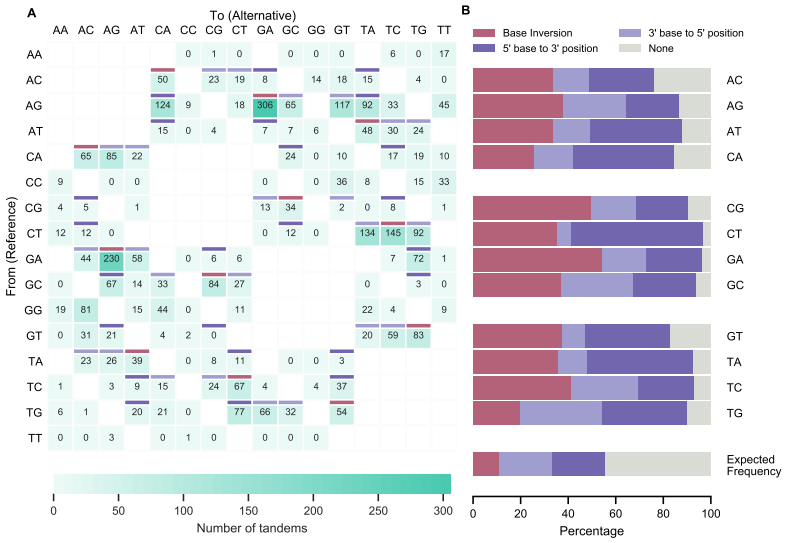


Figure 2.4. Corrected incidence of tandem dinucleotide substitutions in healthy donors. **(A)** Dinucleotide substitutions from unique IGHV, IGKV and IGLV sequences from the Leiden healthy donor cohort and corrected after in silico predictions of dinucleotide substitutions that did not occur in tandem. Burgundy cells represent sequence inversions, light and dark purple cells represent juxtalocations of the 5' and 3' base in the pair (as seen from the non-transcribed strand), respectively. For unshaded cells, juxtalocation could not be assessed due to one or more nucleotides in the reference sequence matching the mutated sequence. **(B)** Relative contribution of sequence inversions and juxtalocations.

that contained a cytosine on the transcribed/non-coding strand, which during replication serves as a template for lagging strand synthesis, were more frequently mutated than their non-transcribed partner, following the pattern previously described for AID-induced mutations [140] (Supplementary Figure S5). Indeed, a strand-bias analysis revealed that tandem substitutions preferentially occurred on the template strand, and that canonical (WRCY) and non-canonical (WA; POLH-associated) [137, 141] AID motifs were most commonly found around the 5' base of the template strand sequence (Figure 2.3). These observations implicate that tandem substitutions preferentially arise at AID-induced DNA lesions.

To identify a common mutational signature, we mapped the nucleotides directly adjacent to tandem substitutions. Scrutiny of these contexts revealed that tandem substitutions commonly already contained the mutated sequence in the overlap between the reference sequence and its context (Figure 2.5). In other words, tandem substitutions apparently derive from small juxtalocation events, where directly adjacent sequences move a single position upstream or downstream, thus explaining the abovementioned observation of one of the original bases moving to the other position in the motif. This mechanism also explains the high incidence of inverting substitutions, as they may result from both upstream and downstream juxtalocations. Although, not found in sufficient abundance to allow for statistical testing, it

appeared that tandem substitutions longer than TDNS followed similar patterns.

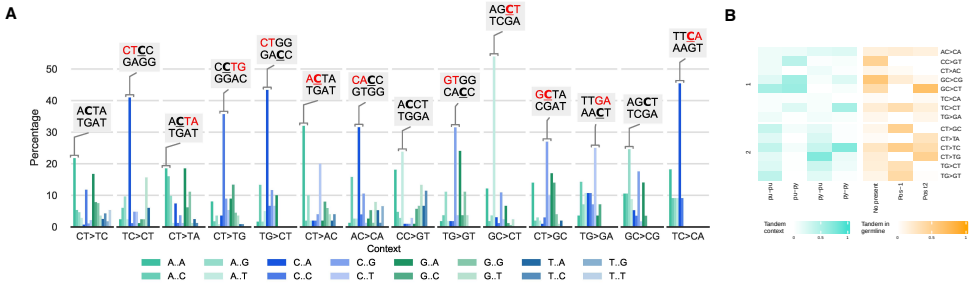


Figure 2.5. Nucleotide context of tandem dinucleotide substitutions. **(A)** The 14 most frequently found tandem dinucleotide substitutions (and its reverse complement) in IGHV from the Leiden healthy donor cohort, were grouped by all 16 possible combinations of flanking nucleotides. Labels displaying the nucleotide context (and reverse complement) with the highest incidence, in red is highlighted the presence of the tandem in the germline reference and underlined and bold the cytosine in the tandem or nearer to the tandem. **(B)** Tandem germline context analysis (left side, green heatmap). It was performed this analysis considering the base type (py: pyrimidine and pu: purine) using 1 nucleotide before and after the tandem position (the corresponding reference is represented with "-"). Tandem germline context analysis for IGHV sequences (right side, orange heatmap). The presence of the resulting tandem in the germline context was identified +/-1 nucleotide for the 14 more prevalent tandems (No present: tandem was not found in 4 nucleotide context; Pos -1: tandem was found in position -1 considering the tandem as central positions t1t2; Pos t2: tandem was found in position t2 of the tandem).

POLYDIPYRIMIDINE STRETCHES ARE FAVORED

Following the observations that TDNS occur preferentially in dipyrimidine motifs, that the cytosine-containing strand is dominantly targeted, and that tandem substitutions in majority represent single nucleotide juxtapositions, tandem substitutions should preferentially arise from polydipyrimidine stretches. Indeed, the substitution tables show that most of the dominantly observed TDNS motifs fit within this hypothesis (Figure 2.4 and Supplementary Table S10).

Notable exceptions to this rule were the commonly observed AG to GA and GC to CG inverting substitutions, which contained the mutated sequence in their germline context in less than half of instances.

VDJ TANDEM SUBSTITUTIONS IN UNG AND MMR DEFICIENCY

After AID deaminates a C into a U, immunoglobulin gene mutational patterns are governed by UNG-intiated BER and the non-canonical MMR pathway. Based on knockout mouse models, tandem substitutions were described to predominantly occur in the MMR pathway [142]. Since experimental models to mimic SHM in humans are not available, the closest approximation to

test the relative contribution of BER and MMR is to analyze BCR repertoires from DNA repair deficient patients.

We obtained a massive parallel sequencing library with VDJ from one MSH2-deficient and three MSH6-deficient patients. All patients carried biallelic defects in their respective genes, leading to Constitutional MisMatch Repair Deficiency (CMMRD) syndrome [143, 144]. An additional VDJ library was obtained from an UNG-deficient patient. As internal controls, we analyzed an independent cohort of healthy donors generated by the same methodology as these patient datasets (henceforth referred to as: Rotterdam healthy donor cohort) [135]. Sequences were filtered identically as described for the Leiden healthy donor cohort, except for the selection by mutation load (see Materials and Methods), yielding 7.654 unique VDJ sequences from the CMMRD patients and 104 unique VDJ sequences from the UNG-deficient patient.

The CMMRD patient repertoires mostly lacked IgM and both the patient and Rotterdam healthy donor repertoires were sequenced using a different approach and sequencing platform than the Leiden healthy donor cohort, which resulted in higher mutation loads and relatively more false TDNS. Therefore, we did not pool these repertoires in our in silico mutation algorithm, but processed them separately and then independently predicted SNS mutation clusters as described for the Leiden cohort.

After in silico correction, tandem dinucleotide substitutions in the Rotterdam healthy donor cohort confirmed the previously observed paradigm of juxtapositions in TCT-motifs (Supplementary Figure S6, Supplementary Table S10B, Supplemental Data 2.1-2.42). In this dataset, inversions and base swaps were less dominant in motifs other than the dipyramidine ones, putatively as a result of the more challenging in silico correction of this more artefact-prone dataset compared to the Leiden healthy donor cohort. Nevertheless, in the corrected tables, TDNS still represented 1,9% of all substitutions, almost twice the frequency of the previously highest estimate.

After removal of in silico predicted “false” TDNS, the dinucleotide substitution table of the CMMRD patients also showed great similarities to the Leiden healthy donor cohort, suggesting that the mismatch repair machinery does not play a major role in the formation of tandem substitutions in humans (Figure 2.6, Supplementary Table S10C, Supplemental Data 2.1-2.32). The corrected proportion of TDNS in all substitutions was 2,0%, comparable to the 1,9% in the equally extensively mutated sequences from the Rotterdam healthy donor cohort. Notably, the SNV show a moderately increased Ts : Tv ratio of 1,91, consistent with DNA repair deficiency during somatic hypermutation (Supplementary Table S5B).

Conversely, the previously identified substitution patterns could not be reproduced in the dinucleotide substitution table of the UNG-deficient patient. In the SNS, the Ts : Tv ratio was skewed towards transitions (3,84), consistent with the “replication without repair” pathway leading to C to T transitions in the absence of BER (Supplementary Table S5C). Rather than base swaps and inversions, contiguous substitutions dominantly also resulted from this “replication without repair” pathway, representing AID-induced C to T transitions in dinucleotide

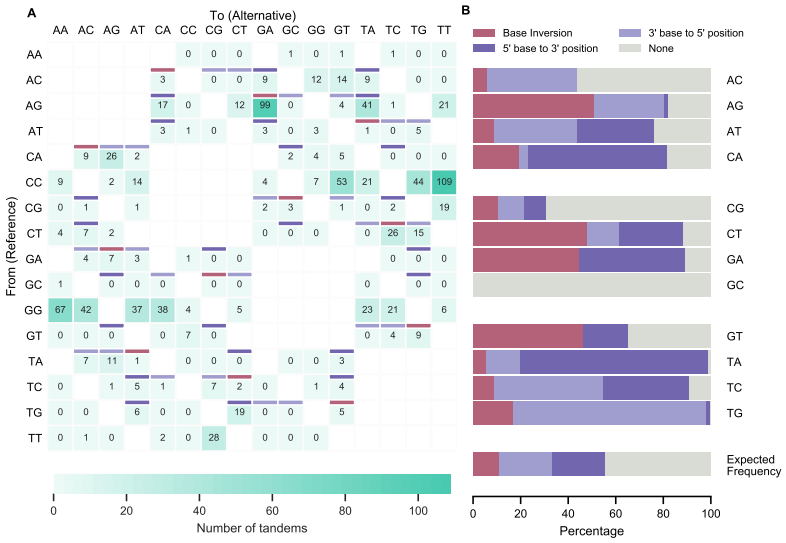


Figure 2.6. Corrected incidence of tandem dinucleotide substitutions in MSH2/6 deficiency. **(A)** Dinucleotide substitutions from unique IGHV sequences obtained from constitutional mismatch repair deficiency (CMMRD) patients and corrected after in silico predictions of dinucleotide substitutions that did not occur in tandem. Burgundy cells represent sequence inversions, light and dark purple cells represent juxtalocations of the 5' and 3' base in the pair (as seen from the non-transcribed strand), respectively. For unshaded cells, juxtalocation could not be assessed due to one or more nucleotides in the reference sequence matching the mutated sequence. **(B)** Relative contribution of sequence inversions and juxtalocations.

motifs consisting solely of cytosine and guanine bases (Figure 2.7). This observation remained consistently different from the other datasets when these were randomly sampled down to the smaller library size of this (Supplementary Figure S7). Indeed, 81% of these ubiquitous GC to AT transitions were located in AID hotspots, which are known to favor transversions only with UNG (over)expression (58). The paucity of unique sequences in this library precluded detailed in silico correction as performed for the other datasets. To obtain the best approximation of the incidence of corrected tandem substitutions in this dataset, in silico predictions were performed on pooled sequences of the IGHV3 and IGHV4 families, respectively. These analyses predicted 76 mutation clusters resulting from adjacent SNS events, of which 67 were actually observed in the dataset (Figure 2.7 and Supplemental Data 4.1-4.3). Thus, no tandem substitutions were observed in the absence of the critical, BER-initiating component UNG. Indeed, this is in line with our previous observation that tandem substitutions are more abundant in immunoglobulin loci than the genome at large, as UNG activity is mainly directed to these genetic regions. Additionally, relative absence of faithful DNA repair mechanisms during SHM makes UNG-derived AP sites more vulnerable to tandem substitutions than genome-wide events of spontaneous deamination [145].

While these findings are congruent with the hypothesis that UNG may be involved in the

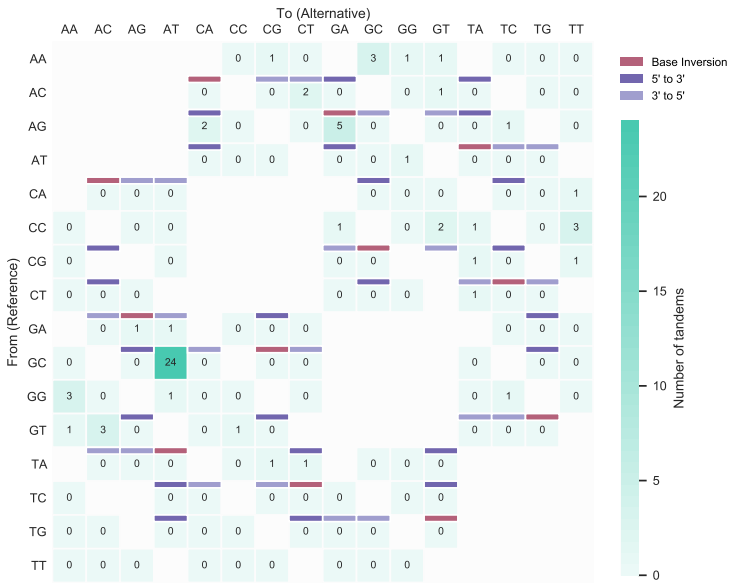


Figure 2.7. Incidence of tandem dinucleotide substitutions in UNG deficiency. All dinucleotide substitutions from unique IGHV sequences obtained from an UNG deficient patient. Burgundy cells represent sequence inversions, light and dark purple cells represent juxtalocations of the 5' and 3' base in the pair (as seen from the non-transcribed strand), respectively. For unshaded cells, juxtalocation could not be assessed due to one or more nucleotides in the reference sequence matching the mutated sequence. Results could not be reliably corrected for dinucleotide substitutions that did not occur in tandem for each mutation independently due to the size of the dataset. However, simulation of dinucleotide substitutions resulting from independent single nucleotide substitutions showed an almost identical number of dinucleotide substitutions (n=76) as the number that was observed (n=67), suggesting that this UNG deficient patients has no tandem substitution events.

formation of the majority of tandem substitutions in humans, it should be noted that the n=1 sample size precludes any definitive conclusions.

2.4 | DISCUSSION

We describe the ubiquitous presence of tandem substitutions in human V(D)J rearrangements. Tandem substitutions contribute to the acquisition of mutations during SHM in a frequency at least two to five times higher than the previously highest estimate. In fact, tandem substitution events are likely to be still underestimated by this study, since any TDNS events resulting in a mutation where one of the bases matches the reference sequence would appear as a regular SNS. Tandem substitutions have the potential to expedite the adaptive immune response by overcoming amino acid code redundancy and by incidentally mutating multiple residues at once. Clustering of such mutations around AID hotspots and their overall distribution indicates that tandem substitutions are an integral part of the SHM spectrum.

Virtually all tandem mutations adhere to a previously unrecognised substitution pattern resulting from single nucleotide juxtalocations. Confirmation of these findings in the independent Rotterdam cohort generated by a different methodical approach, albeit at a lower frequency, ensures that our observations are not a result of methodological artefacts. As an explanation for this phenomenon, we propose the EXPEDITE (EXtruded Pinching Effecting DIrectional Tandem Exchange) model. In this model, an abasic site transiently adopts an extruded configuration leading to misreading of template DNA during replication. Our data indicate that AID-mediated deamination of a cytosine in polydypirimidine stretches is the main route to the juxtalocations that underlie tandem substitutions. After recognition of the uracil, UNG cleaves the N-glycosylic bond to create an apyrimidinic (AP) site, a known risk for strand slippage [34]. Additionally, UNG creates 45° kinks at dUTP positions [146], which facilitate the extrusion and subsequent skipping of the abasic site by a DNA damage tolerant DNA polymerase. Considering the polydypirimidine motifs, such extruded positioning could be facilitated by the relatively small flanking pyrimidine bases, in a similar fashion as was recently described for flanking cytosine bases [147]. Indeed, mismatched pairs have an increased propensity for spontaneous base flipping [148], and the presence of UNG further increases the lifetime of open states following spontaneous base-pair dissociation and twists the unstacked nucleotides out of their helical conformation [149–151]. These mechanisms might be considered intrinsic to the canonical BER pathway, but lead to a previously unrecognized effect.

Subsequent repositioning of the pinched or extruded AP site or base into the strand backbone would move the sequence one base upstream. Replication by a translesion synthesis DNA polymerase capable of extending from the mismatch ensures further extension of the novel DNA strand. Although the events in this model now no longer follow the canonical BER pathway, recruitment of an AP endonuclease (APE), as a downstream component of BER, is conceivable. APE cleaves 3' of the AP site and removes the adjacent mismatched base. Finally, the newly generated gap in the template strand is filled and ligated by faithful DNA repair mechanisms, thereby completing the TDNS process (Figure 2.8). The length of the gap, minus any chance matches at its termini, determines the length of the tandem substitution. It should be stressed that the EXPEDITE model is a best estimate proposal, and the exact proteins and mechanisms involved, especially in the non-canonical mechanism following UNG involvement, are subject to speculation and await verification.

Although AP sites, mismatched bases, and uracil residues have a particular proclivity to adopt extruded configurations [147, 148, 152–154], juxtalocations could theoretically also result from spontaneous base flipping of matched base pairs, especially A-T pairs [151]. Indeed, we observe that germline motifs AT and TA, which contain no cytosine bases on either strand and would therefore remain unexplained through the UNG pathway, are the least abundant reference motifs of all tandem substitutions - yet are not absent. Furthermore, they do follow the paradigm of single nucleotide juxtalocated outcomes. Alternatively, these juxtalocations may be part of a longer tandem substitution initiated by a cytosine that is near, but not inside the motif, and other bases in this longer tandem substitution remain identical after

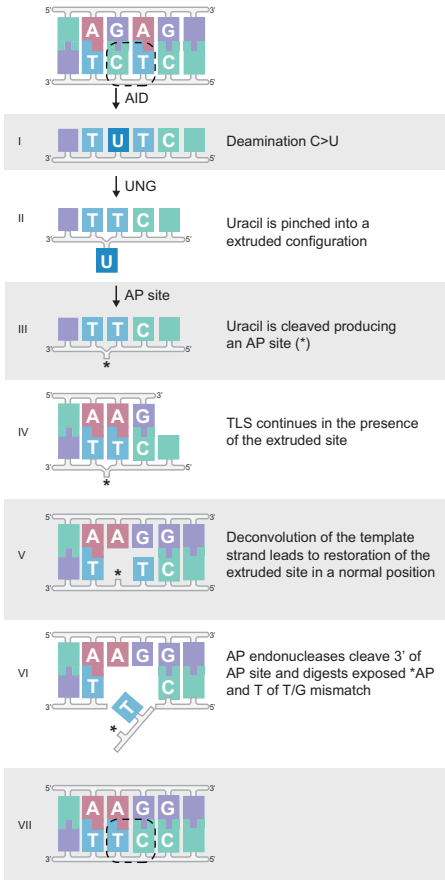


Figure 2.8. Proposed EXPEDITE model for tandem substitutions causing juxtalocation of adjacent residues during DNA replication. **(I)** Cytosine (C) is deaminated to uracil (U) in a polypyrimidine context (shown here as an example since this context typically contains the most tandem substitutions, refer to). **(II)** UNG pinches the U into an extruded configuration and **(III)** U is cleaved, leaving an AP site (*). **(IV)** If during DNA replication the resulting extruded AP site is tolerated, the juxtaposed base acts transiently as a substitute template. **(V)** After decontraction of the template strand, the AP site is reverted to a normal configuration and replication continues from the "mismatch" position containing the AP site. **(VI)** Hereafter, the AP site, as well as the mismatch, is removed by canonical base excision repair. (see Discussion). **(VII)** Final repair synthesis and ligation.

juxtalocation, therefore not registering as mutated.

Not all tandem substitutions follow the juxtalocation paradigm. Most importantly, the common (top strand) AG to GA inverting substitution contained the mutated sequence and their direct context in only less than half of cases, suggesting that these may stem from an alternative mechanism. We consider the following: deamination of the C in the bottom strand leads to U, and subsequent replication across U results in a G to A mutation on the top strand (canonical C to T on the bottom strand). Following removal of the U on the bottom strand

by either canonical UNG or canonical MMR and patch repair of the bottom strand by POLH, replication across from the new WA (POLH hotspot) introduces a C across from the A. This will generate the observed AG to GA substitution (Supplementary Figure S8). Thus it appears that similar to previous observations in mice, MMR is to some extent also involved in tandem substitution generation in humans. However, in contrast to murine studies [129], a majority of tandem substitutions in humans appear to depend on UNG activity in the BER pathway, marking a notable species difference.

A number of specific TDNS, most importantly GC to AA/TT and TC to AA, have previously been attributed to POLZ and/or POLI activity following observations in Saccharomyces cerevisiae and murine models [122, 129, 155, 156]. However, these TDNS are among the rarest in each of our datasets and therefore, at least in humans, do not seem to represent significant additional pathways beyond the mechanisms described in this manuscript. Supposedly, these differences derive from a previously remarked species difference concerning tandem substitution formation [122]. Indeed, murine translesion polymerases create tandem substitutions through the MMR pathway [129], whilst experiments in human cell lines have implicated POLI in tandem substitution formation, but rather through the involvement of UNG [157, 158]. The data in this report propose a new model where the actual tandem substitution inducing lesion, the AP site, does not serve as a non-instructive template but rather causes juxtapositioning by template flipping. Several characteristics identify POLH as the prime candidate translesion synthesis polymerase to create tandem substitution during SHM: The requirement of a large catalytic site, the open active POLH site that can accommodate non-Watson-Crick base pairs, and the high POLH error rate of $3{,}5x10^{-2}$ which is associated with its ability to bind dNTP without DNA substrate [159, 160].

All observations in this study derive from a reverse immunology approach. The lack of experimental systems to study SHM in humans currently precludes controlled experimental confirmation of our proposed mechanism. Therefore, analysis of gene-deficient patients as performed in this study serves as a preliminary attempt of experimental confirmation. When testing the EXPEDITE hypothesis against a small cohort of DNA repair deficient patients, we found an apparent absence of tandem substitutions in UNG but not in MSH2/6 deficiency. Although these findings corroborate the described model with a strict dependence on abasic sites, the small number of individuals and sequences and the more error-prone sequencing approach preclude any definitive confirmation. Therefore, the analysis of additional UNG-deficient patients with high-fidelity BCR amplification and sequencing would be highly desirable. Unfortunately, homozygous UNG deficiency is an exceptionally rare disease [161], and the national reference centers for immunodeficiency of The Netherlands (Erasmus Medical Center Rotterdam), France (Hôpital Necker-Enfants Malades, Paris), and Germany (University Medical Center Freiburg) do not have additional UNG-deficient patients in their registries and were also unable to provide us material from the three originally described patients [162]. We recommend that whenever additional research material comes available, additional B-cell receptor repertoires are sequenced and results are shared with the scientific community to elucidate the exact tole of UNG in the formation of tandem substitutions.

2.5 | SUPPLEMENTARY MATERIAL

SUPPLEMENTARY FIGURES

Figure S1. Comparison between the tandems found in the simulated datasets (distribution of 100,000 simulations) and observed values for Leiden cohort. (A) IGKV, (B) IGLV and (C) IGV (H+K+L).

Figure S2. Distribution of tandem dinucleotide substitutions in the IGKV allele.

Figure S3. Distribution of tandem dinucleotide substitutions in the IGLV allele.

Figure S4. Corrected incidence of dinucleotide tandem substitutions in Leiden healthy donors by chain.

Figure S5. Strand bias in tandem substitutions.

Figure S6. Corrected incidence of dinucleotide tandem substitutions.

Figure S7. Incidence of dinucleotide tandem substitutions in a random subset (normalized according to UNG-deficient library) of Leiden healthy donors, Rotterdam healthy donor and MSH2/6 deficiency cohort.

Figure S8. Model for AG>GA tandem substitutions.

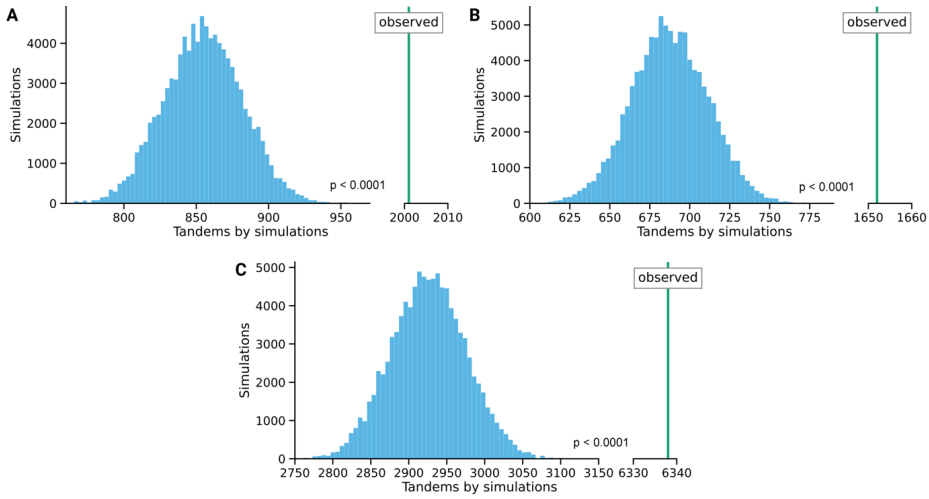


Figure S1. Comparison between the tandems found in the simulated datasets (distribution of 100,000 simulations) and observed values for Leiden cohort. (A) IGKV, (B) IGLV and (C) IGV (H+K+L).

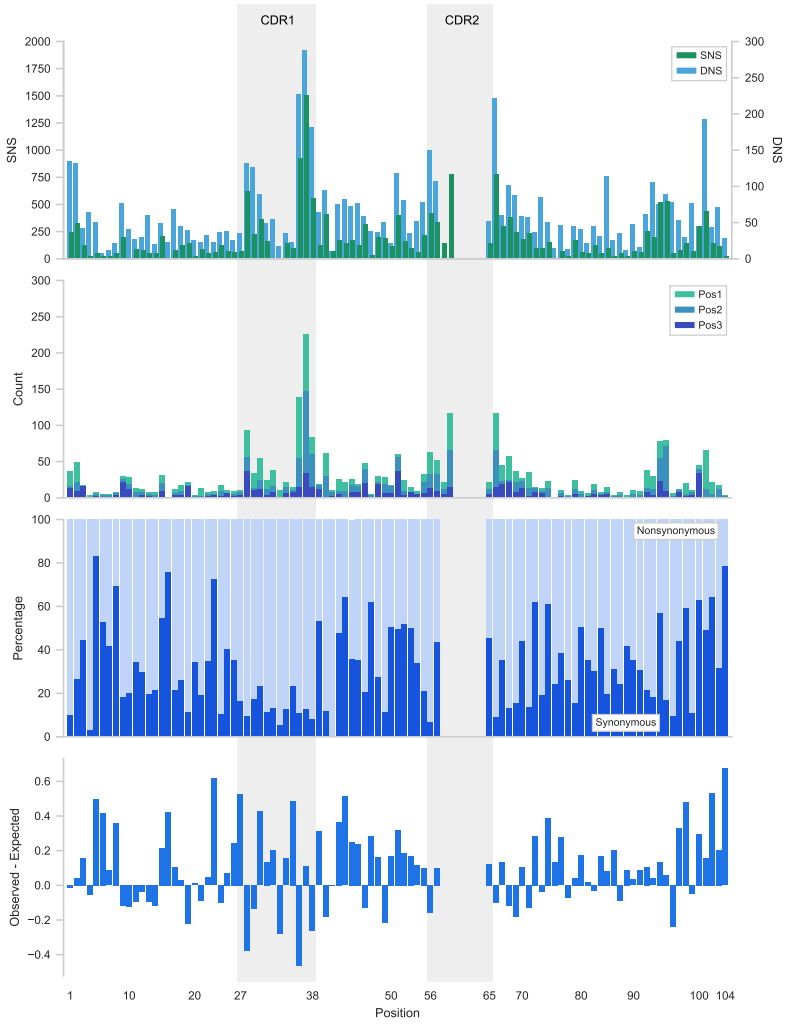


Figure S2. Distribution of tandem dinucleotide substitutions in the IGKV allele. (A) Distribution of single nucleotide substitutions (SNS) and tandem dinucleotide substitutions (TDNS) in the IGKV allele. (B) Nucleotide specific context within a codon. Relative contributions of tandem dinucleotide substitutions for VDJ. Results are split per position in the codon. Position 1 represents 5' and middle base on the coding strand, position 2 represents middle and 3' base and position 3 represents the 3' base of the codon and the 5' base of the downstream codon. (C) Proportion of synonymous and nonsynonymous mutations per position in the IGKV allele. (D) Mutational resistance score. Plotted are the expected minus the observed frequencies of nonsynonymous mutations per codon position in the IGKV allele. Increasingly higher values represent residues that are more resistant to mutation.

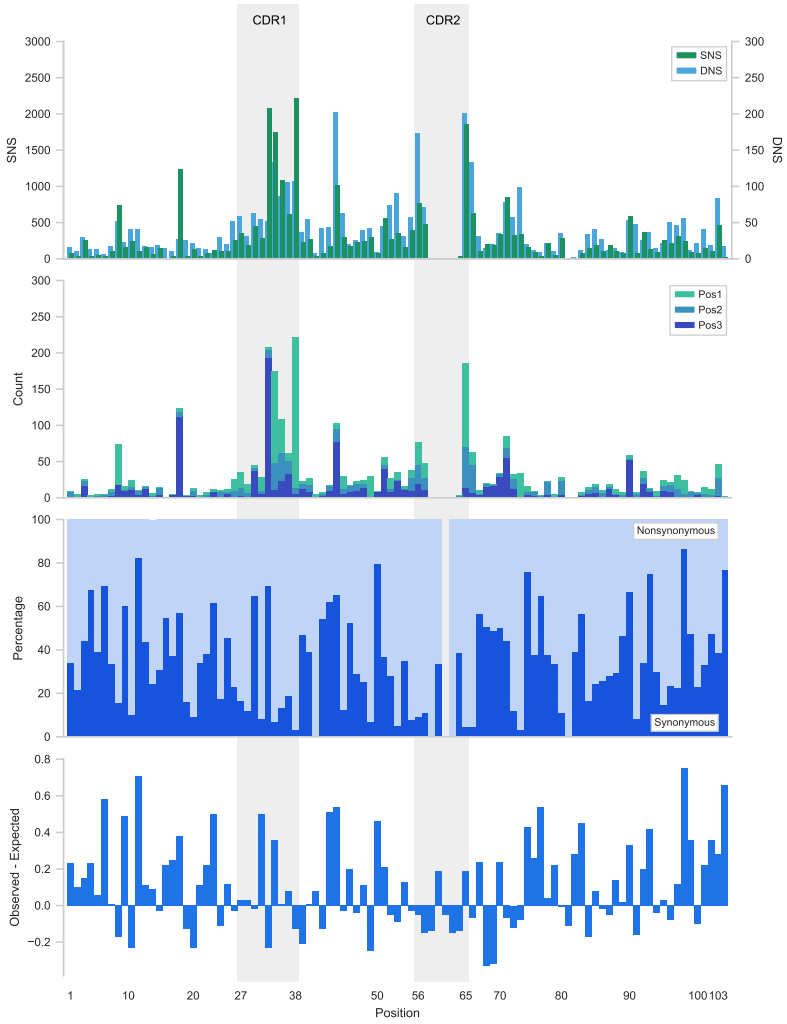


Figure S3. Distribution of tandem dinucleotide substitutions in the IGLV allele. (A) Distribution of single nucleotide substitutions (SNS) and tandem dinucleotide substitutions (TDNS) in the IGLV allele. (B) Nucleotide specific context within a codon. Relative contributions of tandem dinucleotide substitutions for VDJ. Results are split per position in the codon. Position 1 represents 5' and middle base on the coding strand, position 2 represents middle and 3' base and position 3 represents the 3' base of the codon and the 5' base of the downstream codon. (C) Proportion of synonymous and nonsynonymous mutations per position in the IGLV allele. (D) Mutational resistance score. Plotted are the expected minus the observed frequencies of nonsynonymous mutations per codon position in the IGLV allele. Increasingly higher values represent residues that are more resistant to mutation.

Figure S4. Corrected incidence of dinucleotide tandem substitutions in Leiden healthy donors by chain. (A) Data obtained from unique IGHV sequences and corrected after in silico predictions of 'false' tandems. (B) Relative contribution of inversions and position changes. (C) Data obtained from unique IGKV sequences and corrected after in silico predictions of 'false' tandems. (D) Relative contribution of inversions and position changes. (E) Data obtained from unique IGLV sequences and corrected after in silico predictions of 'false' tandems. (F) Relative contribution of inversions and position changes. Granate cells represent sequence inversions, light and dark purple cells represent juxtalocations of the 5' and 3' base in the pair (as seen from the non-transcribed strand), respectively. For grey shaded numbers, juxtalocation could not be assessed due to the reference sequence consisting of two identical nucleotides.

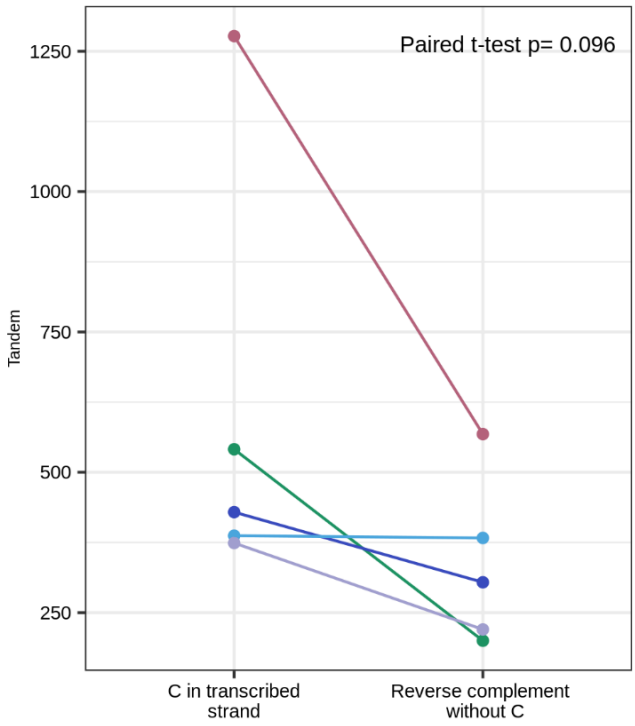


Figure S5. Strand bias in tandem substitutions. Paired dinucleotide motifs with their reverse complementary counterparts, showing that sequences with a cytosine on the transcribed strand (represented as a guanine on the coding strand motif) occur more frequently than sequences with the same motif on the non-transcribed strand.

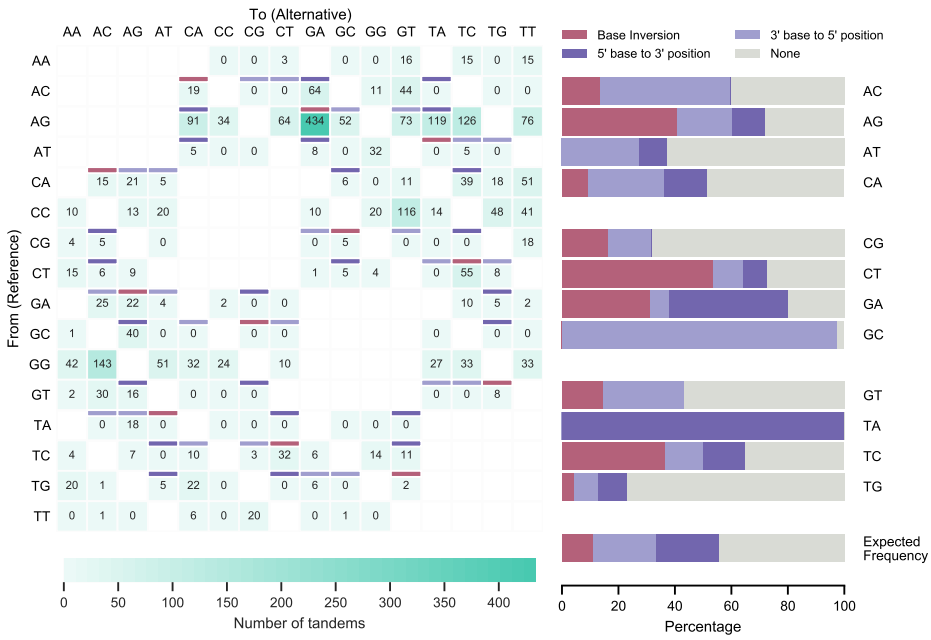


Figure S6. Corrected incidence of dinucleotide tandem substitutions. (A) Data obtained from unique IGHV sequences from the Rotterdam healthy donor cohort and and corrected after in silico predictions of 'false' tandems. Granate cells represent sequence inversions, light and dark purple cells represent juxtalocations of the 5' and 3' base in the pair (as seen from the non-transcribed strand), respectively. For grey shaded numbers, juxtalocation could not be assessed due to the reference sequence consisting of two identical nucleotides. (B) Relative contribution of inversions and position changes.

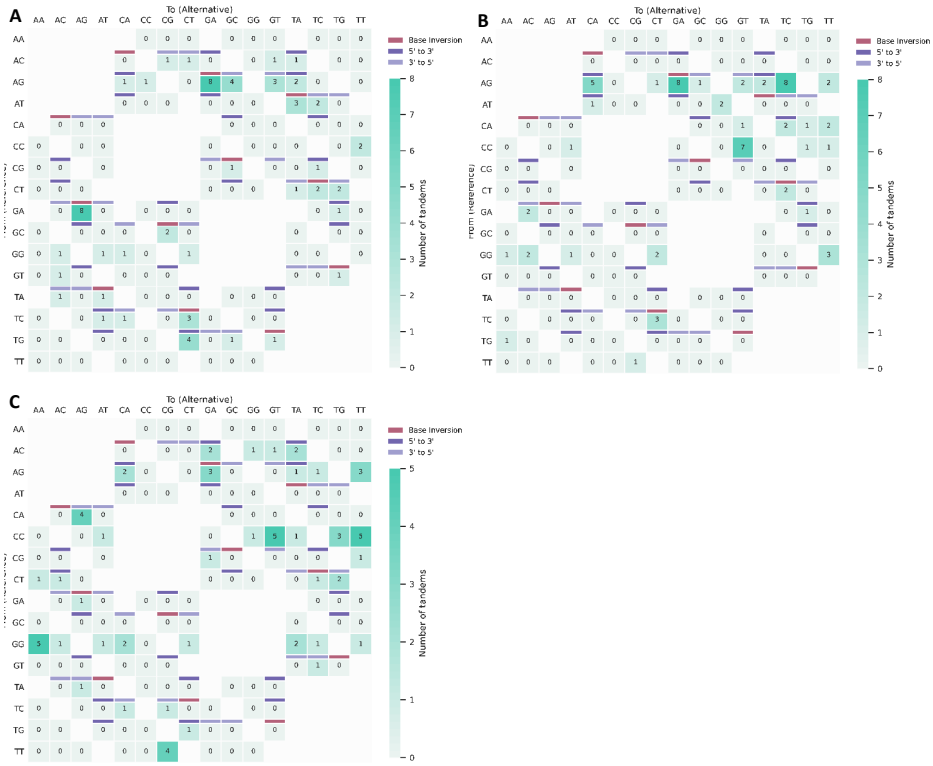


Figure S7. Incidence of dinucleotide tandem substitutions in a random subset (normalized according to UNG-deficient library) of Leiden healthy donors, Rotterdam healthy donor and MSH2/6 deficiency cohort. (A) From the Leiden healthy donor cohort, a random subset was sampled to the size of the UNG-deficient library. (B) From the Rotterdam healthy donor cohort, a random subset was sampled to the size of the UNG-deficient library. (C) From the MSH2/6-deficient cohort, a random subset was sampled to the size of the UNG-deficient library. Granate cells represent sequence inversions, light and dark purple cells represent juxtalocations of the 5' and 3' base in the pair (as seen from the non-transcribed strand), respectively. For grey shaded numbers, juxtalocation could not be assessed due to the reference sequence consisting of two identical nucleotides.

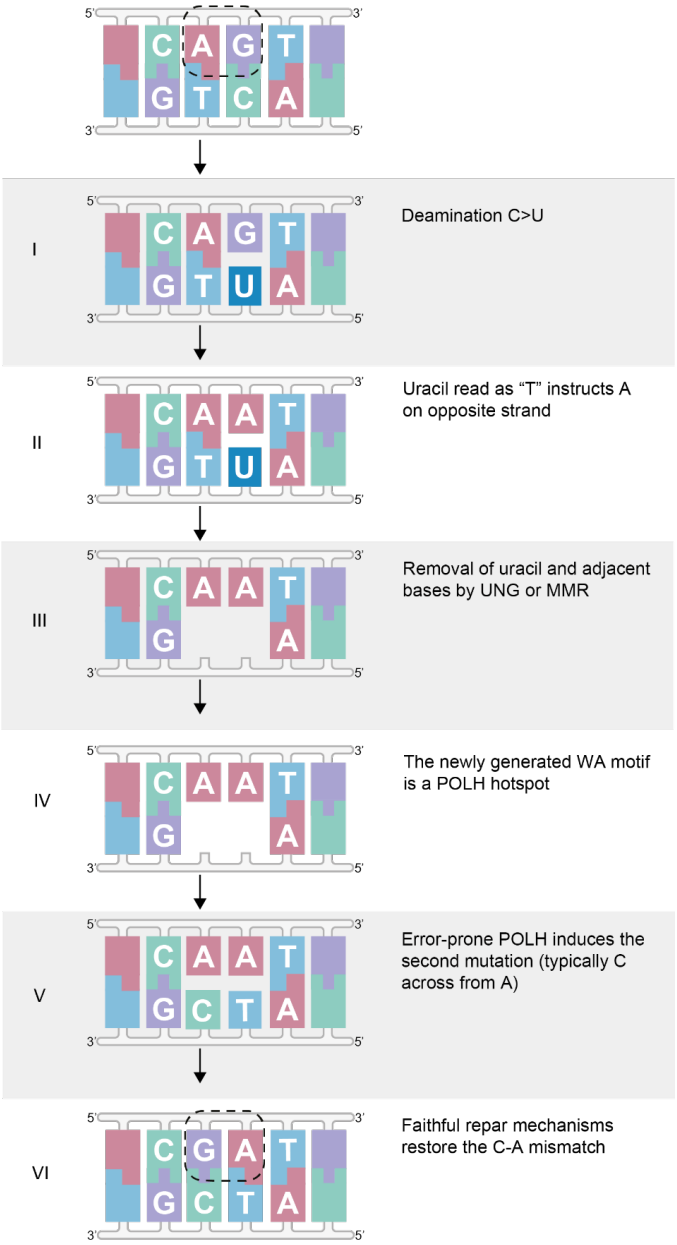


Figure S8. Model for AG>GA tandem substitutions. (I) Deamination of the C in the bottom strand leads to U. (II) Subsequent replication across U results in a G to A mutation on the top strand (canonical C to T on the bottom strand). (II) Following removal of the U on the bottom strand by either UNG or MMR. (IV) Patch repair of the bottom strand by POLH. (V) Replication across from the new WA (POLH hotspot) introduces a C across from the A. (VI) This will generate the observed AG to GA substitution.

SUPPLEMENTARY TABLES

Table S1. Cohort information

Table S2. Calculation of non-synonymous mutation chance per position.

Table S3. Number of substitution from 1-8 nucleotides length

Table S4. Incidence of single nucleotide substitutions and contiguous substitutions.

Table S5A. Relative contributions of SNS in the Leiden healthy donor dataset.

Table S5B. Relative contributions of SNS for the MSH2/6 deficient patients dataset.

Table S5C. Relative contributions of for the UNG deficient patient dataset.

Table S6. Quasi-Poisson Regression to stablish the dependent of TDNS on V allele position and mutational resistance score.

Table S7. Observed tandems in the IGHV Leiden healthy donor dataset, uncorrected for in silico predicted 'false' tandem substitutions.

Table S8. Observed tandems in the IGKV Leiden healthy donor dataset, uncorrected for in silico predicted 'false' tandem substitutions.

Table S9. Observed tandems in the IGLV Leiden healthy donor dataset, uncorrected for in silico predicted 'false' tandem substitutions.

Table S10A. Relative contributions of inversions and position changes per dinucleotide reference for the Leiden healthy donor dataset.

Table S10B. Relative contributions of inversions and position changes per dinucleotide reference for the Rotterdam healthy donor dataset.

Table S10C. Incidence of single nucleotide substitutions and contiguous substitutions.

Table S11. Incidence of single nucleotide substitutions and contiguous substitutions.

Cohort	Acquisition	Sequencing platform	Estimated error rate	Number of sequences	Number of TDNS	Tandem (2nt) by seq (mean)	Overlapping previous publications
Leiden healthy donor cohort	Sequenced in house	PacBio RSII (Pacific Biosciences)	0.126×10^{-3} per bp	5035	6338	1.3	Eur J Immunol. 2020 Dec;50(12):2099-2101. doi: 10.1002/eji.202048828. Epub 2020 Aug 27. Front Immunol . 2019 Jul 3;10:1543. doi: 10.3389/fimmu.2019.01543. eCollection 2019. Front Immunol. 2019 Sep 4;10:2092. doi: 10.3389/fimmu.2019.02092. eCollection 2019.
Rotterdam healthy donor cohort	Re-analysed from publicly available data	Roche 454	0.1 to 10×10^{-3} per bp	5112	11221	2.2	Front Immunol. 2016 Oct 17;7:410. doi: 10.3389/fimmu.2016.00410. eCollection 2016.
MSH2/6 Patient cohort	Re-analysed from publicly available data	Roche 454	0.1 to 10×10^{-3} per bp	1821	3107	1.7	Front Immunol. 2019 Aug 27;10:1913. doi: 10.3389/fimmu.2019.01913.
UNG Patient cohort	Re-analysed from publicly available data	Roche 454	0.1 to 10×10^{-3} per bp	39	67	1.7	Front Immunol. 2019 Aug 27;10:1913. doi: 10.3389/fimmu.2019.01913.

Table S1. Cohort information

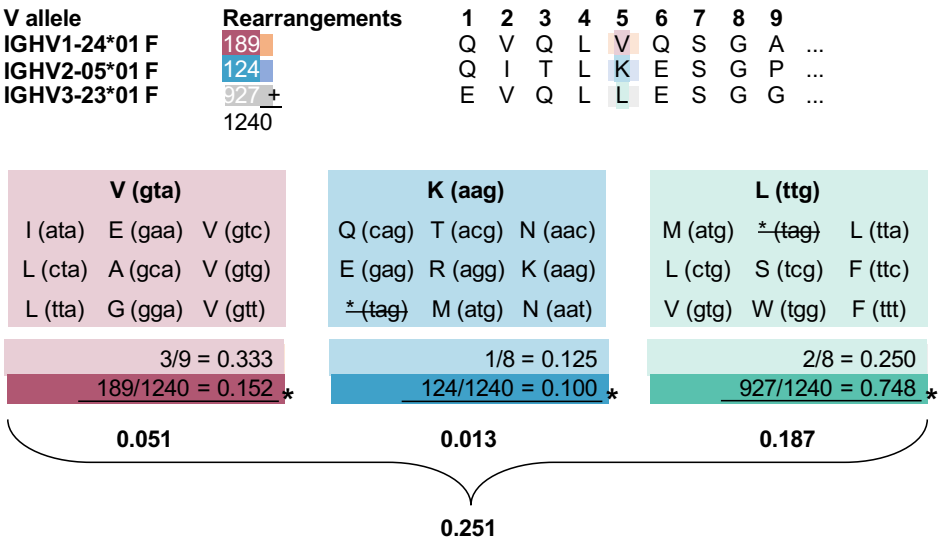


Table S2. Calculation of non-synonymous mutation chance per position. Shown is a calculation for the synonymous mutation chance of position 5 from a mix of three different IGHV alleles. IGHV1-24*01 occurs 189 times (15.2% of all sequences) and has a V in position 5. Three of the nine possible substitutions yield a synonymous mutation. Likewise, the 10% of sequences with a IGHV2-05*01 rearrangement have a one out eight chance for a synonymous mutation and the 74.8% of sequences with a IGHV3-23*01 rearrangement have a two out eight chance for a synonymous mutation. Therefore the expected proportion of synonymous mutations in position 5 across all sequences is (0.333*0.152) + (0.1*0.125) + (0.748*0.25) = 0.251.

Substitution length (nt)	IGHV	IGKV	IGLV	Total
1	60635	32071	30172	122878
2	5338	2742	2655	10735
3	942	574	398	1914
4	280	100	68	448
5	118	28	24	170
6	51	5	2	58
7	21	0	1	22
8	3	0	0	3

Table S3. Number of substitution from 1-8 nucleotides length.

Substitution length (bp)	Contiguous substitutions	In simulated subset†	Simulated *in silico*	False positives	Tandem substitutions‡	Misclassified SNS	Corrected incidence
1	122,878	n/a	n/a	n/a	n/a	n/a	94.3%
2	10,735	6338	2829	46.20%	5775	9920	4.1%
3	1914	1204	143	12.50%	1675	717	1.18%
4	448	305	8	3.00%	435	52	0.31%
5	170	100	1	1.00%	168	10	0.12%
6	58	31	0	0.00%	31	0	0.02%
7	22	16	0	0.00%	16	0	0.01%
8	3	3	0	0.00%	3	0	<0.01%

Table S4. Incidence of single nucleotide substitutions and contiguous substitutions. Substitutions from unique IGHV, IGKV and IGLV sequences from the Leiden healthy donor cohort. SNS, single nucleotide substitution. †Only the most abundantly rearranged V alleles represented datasets of adequate size to be subjected to in silico simulations. ‡Extrapolating the correction factor found in the in silico simulated subset to the complete dataset.

		Alternative			
	Healthy	A	C	G	T
Reference	A		6%	13%	6%
	C	3%		8%	14%
	G	18%	14%		5%
	T	3%	8%	4%	

Ts	0.46
Tv	0.44
Ts/Tv ratio	1.05

Table S5A. Relative contributions of SNS in the Leiden healthy donor dataset.

		Alternative			
	CMMRD	A	C	G	T
Reference	A		2%	5%	2%
	C	3%		7%	27%
	G	28%	15%		4%
	T	2%	4%	1%	

Ts	0.61
Tv	0.32
Ts/Tv ratio	1.90

Table S5B. Relative contributions of SNS for the MSH2/6 deficient patients dataset.

		Alternative			
	UNG	A	C	G	T
Reference	A		4%	10%	4%
	C	1%		2%	31%
	G	31%	4%		2%
	T	2%	7%	2%	

Ts	0.73
Tv	0.19
Ts/Tv ratio	3.84

Table S5C. Relative contributions of for the UNG deficient patient dataset.

Deviance Residuals:

Min	1Q	Median	3Q	Max
-14.47	-3.82	-1.70	1.45	23.42

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.7434	0.1313	36.131	<2e-16 ***
score	-1.3848	0.6894	-2.009	0.0474 *
regionFR	-0.9875	0.1818	-5.432	4.13e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 38.99928)

Null deviance: 4833.7 on 99 degrees of freedom
Residual deviance: 3217.8 on 97 degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

Table S6. Quasi-Poisson Regression to establish the dependent of TDNS on V allele position and mutational resistance score.

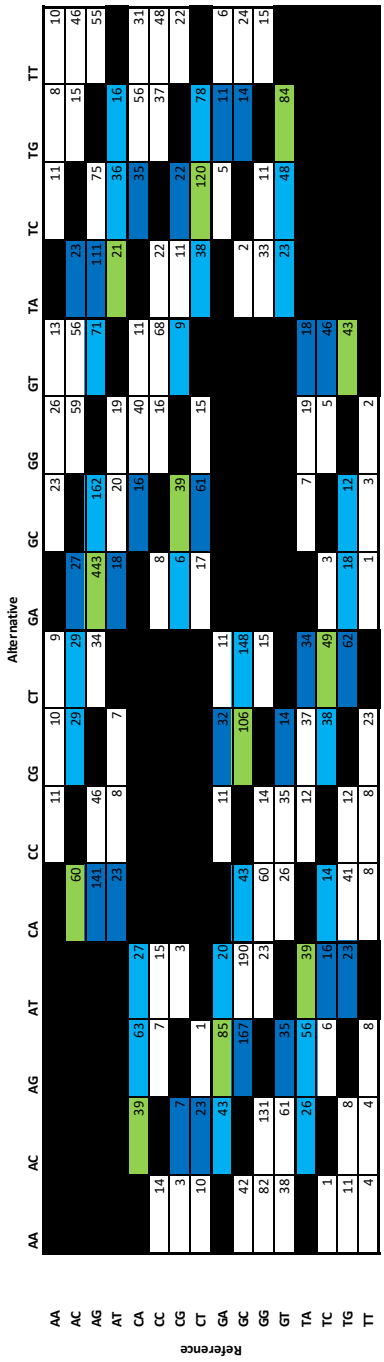


Table S7. Observed tandems in the IGHV Leiden healthy donor dataset, uncorrected for in silico predicted 'false' tandem substitutions.

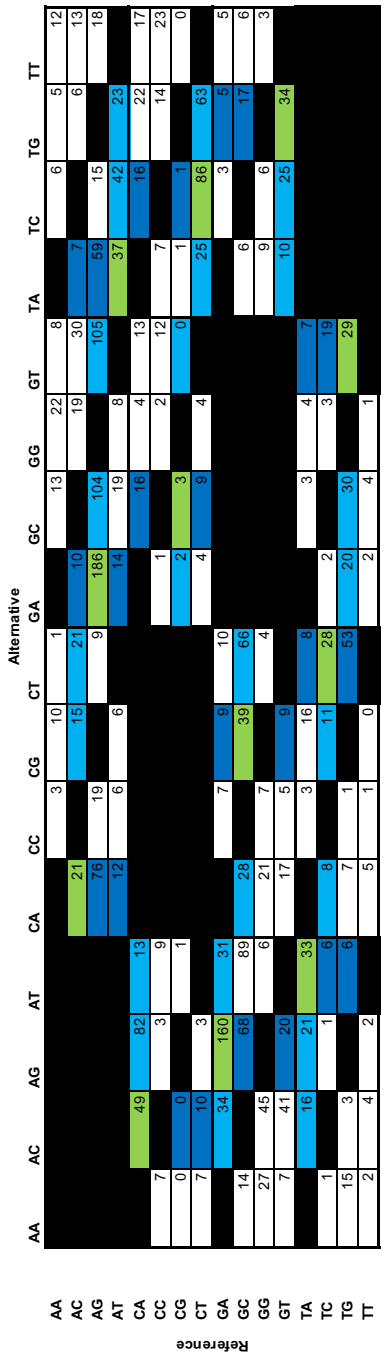


Table S8. Observed tandems in the IGKV Leiden healthy donor dataset, uncorrected for in silico predicted 'false' tandem substitutions.

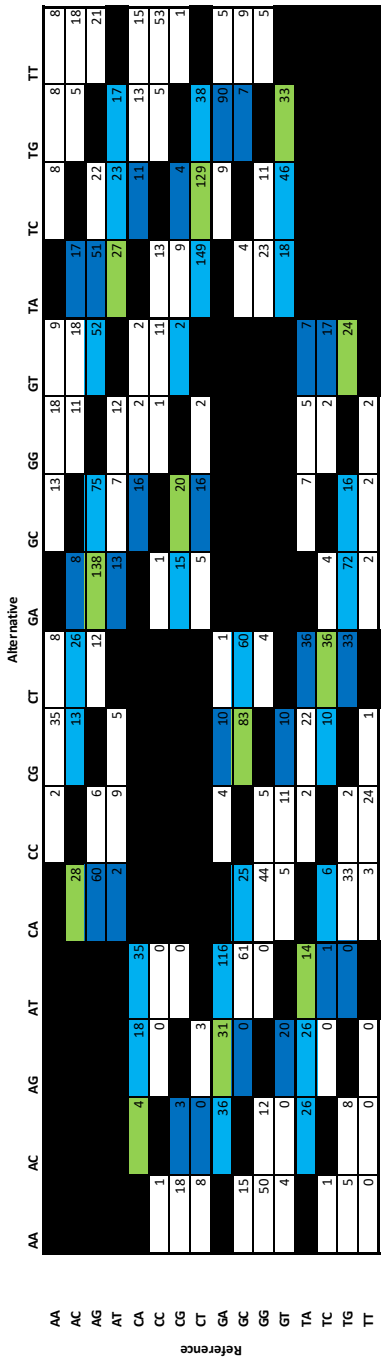


Table S9. Observed tandems in the IGLV Leiden healthy donor dataset, uncorrected for in silico predicted 'false' tandem substitutions.

Reference	Inversion	3' base to 5' position	5' base to 3' position	Rest
AA				
AC	33%	27%	16%	24%
AG	40%	21%	26%	13%
AT	34%	38%	16%	13%
CA	26%	42%	16%	16%
CC				
CG	47%	22%	19%	11%
CT	35%	55%	6%	3%
GA	54%	24%	19%	4%
GC	36%	26%	32%	6%
GG				
GT	36%	36%	10%	18%
TA	36%	45%	13%	7%
TC	40%	23%	28%	8%
TG	20%	35%	35%	10%
TT				
Total	**37.3%**	**31.8%**	**20.4%**	**10.4%**
Expected	*11.1%*	*22.2%*	*22.2%*	*44.4%*

Table S10A. Relative contributions of inversions and position changes per dinucleotide reference for the Leiden healthy donor dataset.

Reference	Inversion	3' base to 5' position	5' base to 3' position	Rest
AA				
AC	11%	0%	38%	51%
AG	40%	12%	19%	29%
AT	0%	11%	33%	56%
CA	8%	17%	24%	50%
CC				
CG	14%	0%	21%	64%
CT	47%	22%	8%	23%
GA	40%	35%	8%	18%
GC	0%	0%	48%	52%
GG				
GT	5%	0%	25%	70%
TA	0%	100%	0%	0%
TC	34%	14%	16%	35%
TG	3%	10%	8%	78%
TT				
Total	**26.1%**	**11.7%**	**22.8%**	**39.4%**
Expected	*11.1%*	*22.2%*	*22.2%*	*44.4%*

Table S10B. Relative contributions of inversions and position changes per dinucleotide reference for the Rotterdam healthy donor dataset.

Reference	Inversion	3' base to 5' position	5' base to 3' position	Rest
AA				
AC	8%	0%	26%	66%
AG	51%	3%	29%	17%
AT	0%	38%	44%	19%
CA	23%	56%	9%	11%
CC				
CG	10%	3%	5%	82%
CT	44%	37%	10%	9%
GA	58%	33%	0%	8%
GC	n/a	n/a	n/a	n/a
GG				
GT	58%	0%	0%	42%
TA	0%	100%	0%	0%
TC	11%	30%	38%	22%
TG	33%	6%	39%	22%
TT				
Total	**35.6%**	**18.8%**	**20.8%**	**24.8%**
Expected	*11.1%*	*22.2%*	*22.2%*	*44.4%*

Table S10C. Relative contributions of inversions and position changes per dinucleotide reference for the MSH2/6 deficient patient dataset.

	IGV Cosine similarity
DBS1	0.117553553
DBS2	0.104321766
DBS3	0.183792142
DBS4	0.087752493
DBS5	0.171393487
DBS6	0.230079931
DBS7	0.244419771
DBS8	0.244610783
DBS9	0.533101127
DBS10	0.126942545
DBS11	0.284857731

Table S11. Cosine similarity between IGV pattern and doublet base substitutions (DBS) from COSMIC (v3.1-June 2020)

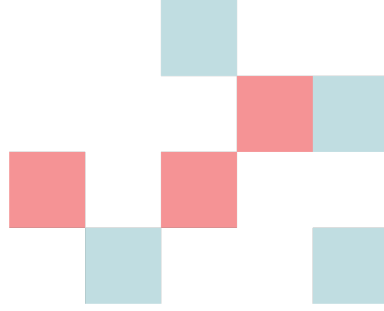
Paine Grande, Torres del Paine, Chile

CHAPTER 3

Detection of occurrent somatic hypermutation and associated gene expression profile in single B cells

Previously published as

Sepúlveda-Yáñez, J. H., Alvarez-Saravia, D., Medina, D., Quinten, E., Kloet, S., Jansen, P. M., Posthuma, E. W., Kiełbasa, S. M., van Bergen, C. A. M., Navarrete, M. A. & Veelken, H. ***Detection of occurrent somatic hypermutation and associated gene expression profile in single*** Manuscript submitted for publication



ABSTRACT

Somatic Hypermutation (SHM) in B lymphocytes is essential for antibody affinity maturation. To analyze the actual occurrence of single SHM events, we performed single-cell RNA and B-cell receptor (BCR) sequencing in follicular lymphoma (FL), a malignancy arrested at the germinal center stage. 1239 alternative V(D)J nucleotides were identified in 364 of 42,985 cells (0.84%) of 12 FL patients. 268 BCR transcripts had two or more alternative nucleotides created in tandem. 50.1% of occurrent SHM events occurred within known motifs for activation-induced deaminase (AID). No V(D)J neovariant was detected in 4439 CLL cells from 5 patients ($p<0.0001$). Neovariant-carrying FL cells expressed higher levels of AID and of the mismatch repair, base excision repair, and DNA replication pathways. Our data provide the first quantitative evidence of ongoing SHM in single cells with novel insights into the dynamics of germinal center B-cell neoplasms and the functional cooperation between AID and DNA repair pathways.

3.1 | INTRODUCTION

Somatic hypermutation (SHM) is an endogenous mutator mechanism in B lymphocytes that targets the immunoglobin V(D)J coding region and facilitates antibody affinity maturation. SHM occurs during germinal center (GC) reactions and strictly depends on the transcriptional activity of the immunoglobulin gene [164, 165]. The frequency of SHM in V(D)J was estimated to be around 10-3 per base pair per generation in murine GC B cells, which would amount to one mutation per cell division [166]. However, SHM has been quantified only as its cumulative outcome of sequence alterations of immunoglobulin (Ig) genes at the population level of selected B cells. In order to detect and quantify single SHM events in individual cells, we performed single-cell RNA sequencing of purified primary follicular lymphoma (FL) cells. FL is considered a malignancy of immortalized mature B cells that are arrested at the GC stage. FL cells constitutively express AID and undergo continuous SHM [167, 168]. Therefore, FL cells accumulate high levels of Ig sequence diversification [169, 170]. Based on assembly of cell-individual V(D)J sequences and realigning sequence reads characterized by their unique molecular identifiers (UMI), we searched with particularly rigid filtering for individual FL cells expressing two populations of Ig transcripts from the same Ig heavy or light chain locus that differed by only one nucleotide (or at most few). Detection of such alternative V(D)J variants in a single B cell would hence imply a single SHM event that had occurred very recently, i.e. within the life span of Ig mRNA transcribed prior to creation of this particular mutation in the course of ongoing SHM. This approach permitted to analyze the gene expression profile of these cells simultaneously and to test the hypothesis that genes and pathways that effectuate SHM, i.e. AID and error-prone DNA repair pathways, are active in cells with occurrent SHM.

3.2 | MATERIALS AND METHODS

SAMPLE COLLECTION AND PROCESSING

Samples with histologically confirmed infiltration of 12 follicular lymphoma (FL) and five samples derived from chronic lymphocytic leukemia (CLL) were collected with approval by the institutional review board of Leiden University Medical Center (no. B16.039) according to the Declaration of Helsinki and with written informed consent (Supp. Table S1). Single-cell suspensions were obtained by gentle mechanical disruption and mesh filtration and were cryopreserved using 10% DMSO as cryoprotectant. Thawed FL cells were purified by flow cytometry using fluorescently labeled antibodies specific for CD19 and CD10 and rested overnight, followed by removal of dead cells (MACS dead cell removal kit; Miltenyi Biotec, Bergisch Gladbach, Germany).

SINGLE-CELL RNA AND BCR ENRICHMENT LIBRARY CONSTRUCTION AND SEQUENCING

Following single-cell cDNA library generation and amplification (Chromium Next GEM Single Cell 5' Kit v1 for the initial three experiments, and v2 for the subsequent eight experiments, provided by 10X Genomics), one fraction of the library was directly sequenced for 5′ gene expression profiling. The second fraction was enriched for BCR transcripts by semi-nested amplification using 3′ constant domain primers for all BCR genes (Chromium Single Cell Human BCR Amplification Kit v1 for the initial three experiments, and v2 for the subsequent eight experiments, provided by 10X Genomics), partially digested and sequenced. Both single-cell libraries were sequenced in paired-end mode on Illumina HiSeq 4000.

SCRNA SEQUENCING DATA PREPROCESSING AND INTEGRATION OF DATASETS

FASTQ files were mapped to the GRCh38 human reference genome with Cell Ranger v3.1.0 (10x Genomics). The preprocessed Cell Ranger gene count matrix was analyzed by Seurat v.4.3.0 [171] and R v.4.1.3 software(R Core Team, 2013). Cells with fewer than 200 genes and genes expressed in fewer than three cells were filtered out as a first quality control step. Independent integration of 17 FL samples from 12 FL patients (8 experiments; Supp. Table 1, Supp. Fig. 1) and five CLL patients (2 experiments) was performed by the reciprocal Principal Component Analysis (RPCA)-based integration anchor workflow implemented in Seurat. The integrated data were scaled after mitochondrial and immunoglobulin genes were filtered out. Dimensionality reduction was performed by PCA. Doublet cells were identified through scDblFinder v1.12.0 and discarded [172]. In cases where multiple patient samples were pooled in a single chip, reads were separated according to the SNP profile and individuals' HLA type (Quinten et al., submitted).

SCBCR SEQUENCING DATA PREPROCESSING AND CALLING OF OCCURRENT SHM EVENTS.

V(D)J repertoires of each FL/CLL case were compiled as contigs per individual cell by Cell Ranger VDJ software. The total/complete/full repertoire of unique contigs served as the reference to maximize capture of single-cell B-cell receptor sequencing (scBCR-seq) reads for further analysis using a custom pipeline as described below (Supplementary Figure S1a). First, scBCR-seq FASTQ files were preprocessed by comparing cell barcodes of individual reads to the 10X Genomics cell barcode whitelist. Barcode sequences with a maximum Hamming distance of 1 from the most similar barcode of the whitelist were corrected and retained, while sequences carrying barcodes with a Hamming distance of > 1 were discarded following the reference method implemented in Cell Ranger. Likewise, unique molecular identifiers (UMIs) in reads with identical cell barcodes were corrected with a maximum permitted Hamming distance of 1 to a dominant UMI within the same cell. In the second step, these corrected scBCR reads were aligned against the case's total V(D)J Cell Ranger contig repertoire using Minimap2 v2.20 [173]. In the third step, based on the results of the alignment, a single optimal V(D)J

reference contig was identified for each cell by the highest number of mapped reads carrying that cell's bar code weighted by mapping scores. In the fourth step, the corrected scBCR-seq reads of every individual cell were realigned to that cell's single optimal V(D)J reference contig. Based on these separate, cell-individual realignments, BCR contigs were constructed for each UMI using a quality-weighted frequency strategy. In the final step, the UMI-based BCR contigs of each cell were aligned to each other, and sequence differences between these contigs were called as potential occurrent SHM events. To exclude possible artifacts, these variants in the V(D)J sequence observed in a single cell were rigorously filtered using the following criteria: (i) reads with more than 25% of the bases with a quality score below 30 were discarded; (ii) cell doublets were detected and removed by scDblFinder [172] based on scRNA gene expression data; (iii) variants must be covered by at least 10 UMIs per position with at least 5 reads per UMI, (iv) variants must be present in at least 20% of the total UMIs, and (v) one of the two alternative variant-carrying V(D)J sequence was not allowed to be exactly identical to a V(D)J sequence observed in any other cell to definitely exclude the possibility of artifacts originating from unrecognized cell doublets. Values for the filter (iii) described above were calculated using statistical approaches. First, the distribution of the values for the parameters (UMI and reads per UMI) was split by diagnosis (CLL or FL). The distribution of UMI and reads for each subject was obtained and the independence between them was checked by estimating Pearson's coefficient. Once the evaluation was made, each distribution was standardized using a z-score applying the central limit theory. Then, the median of each distribution was calculated, and a distribution of the medians was built, generating an estimator's distribution. Finally, the estimator's distribution median was obtained and used to generate the filter criterion. This filtering procedure ensured that only high-quality events were retained for downstream analysis. Finally, alternative V(D)J nucleotides detected within a single cell were annotated as the bona fide "original" variant when they were identical to the corresponding nucleotide of the respective IGV of IGJ germ-line sequence. The alternative variant was designated the bona fide "neovariant" created by occurrent SHM if this nucleotide was not observed in an identical V(D)J sequence in any other cell.

DIFFERENTIAL EXPRESSION GENE AND PATHWAY ENRICHMENT ANALYSIS.

Differential expression analysis was performed using R package DESeq2 [174] between cells with and without events. Normalization of the UMI-count matrix was performed with the function "scran:computeSumFactors" as suggested for scRNA sequencing data [175]. Next, the main DESeq2 function was run with the parameters test = "LRT", fitType = "glmGamPoi", minReplicatesForReplace = Inf, useT = TRUE, minmu = 1e-6. The genes from the differential expression analysis were ranked in decreasing order according to log2FC. This ranked gene lists were used as input for gene set enrichment analysis of KEGG pathways [176] using the gseKEGG function from ClusterProfiler v4.7.1.1 [177].

STATISTICAL INFORMATION.

All data analysis and statistic were done in R 4.2.2 using Jupyter Lab. Statistical details of experiments, including statistical significance, and statistical test are indicated in the relevant figure legends. R was used to plot figures. The graphs were then modified in Adobe Illustrator (version 28.1).

3.3 | RESULTS

IDENTIFICATION AND QUANTIFICATION OF OCCURRENT SHM EVENTS AT SINGLE CELL-LEVEL

To detect the simultaneous expression of two Ig transcripts differing by a single nucleotide originating from a recent SHM event within a single cell, we developed a custom cell-level variant calling pipeline using scBCR-seq data from 12 FL and 5 CLL patients (Figure 3.1, Supp. Fig. S1). In total, 42,985 FL cells (range: 215-8233 per biopsy depending on input; Supp. Table S1) were analyzed and subjected to rigorous filtering steps to preclude false-positive detection of SHM events arising from potential technical issues. The filtering strategy defined stringent thresholds in terms of numbers of individual Ig mRNA molecules sequenced as defined by their UMI, and reads per position obtained from the median of the estimator's distribution (Supp. Fig. S2). Essential additional aspects to ensure data quality included base quality scores, exclusion of cell doublets, Ig productivity, position within the V(D)J region, and proportion of cell-level variants (details in Methods section). After applying this filtering strategy to the single-cell cDNA libraries, 1239 alternative V(D)J variants were detected in a total of 364 single FL cells in 7 of the 12 FL cases analyzed (Supp. Table S2). The median fraction of FL cells with occurrent SHM events was 0.03% per case (range: 0% – 2.3%). Informative examples of occurrent SHM events illustrate the detected intracellular variations at defined positions along the Ig transcripts in the context of all already established acquired somatic mutations as compared to the corresponding primordial sequence Figure 3.1. No occurrent SHM events were detected in CLL cells ($p < 0.0001$; Fisher's exact test).

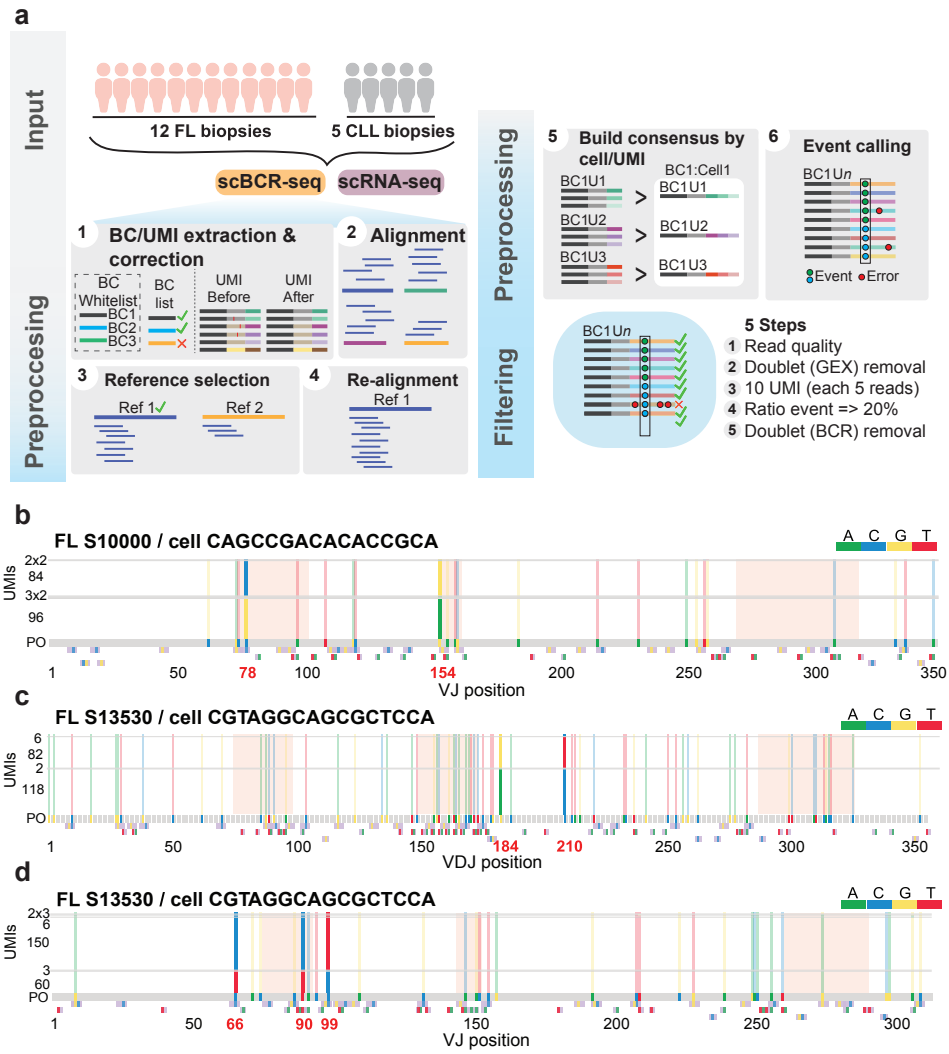
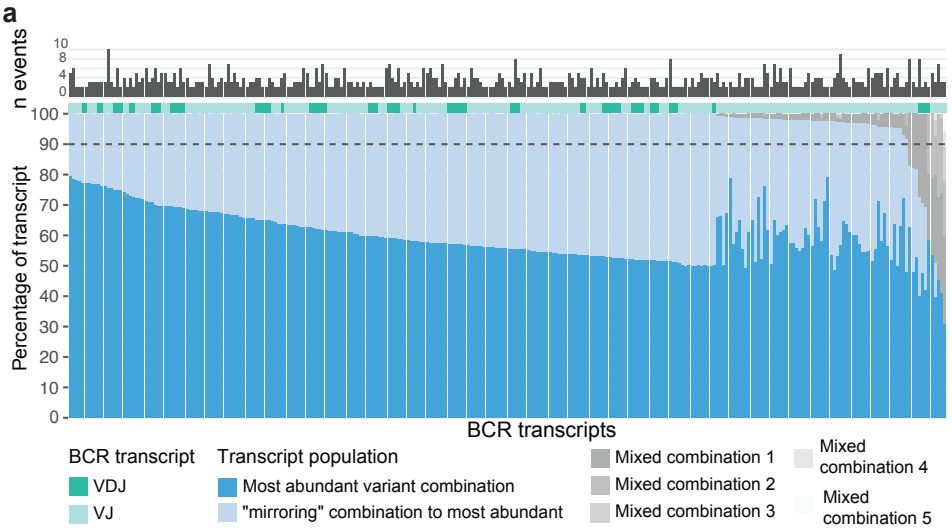
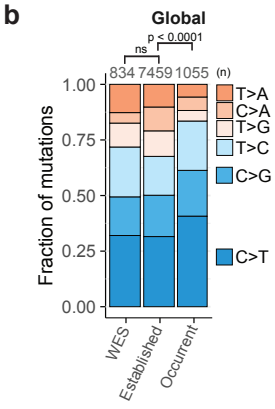
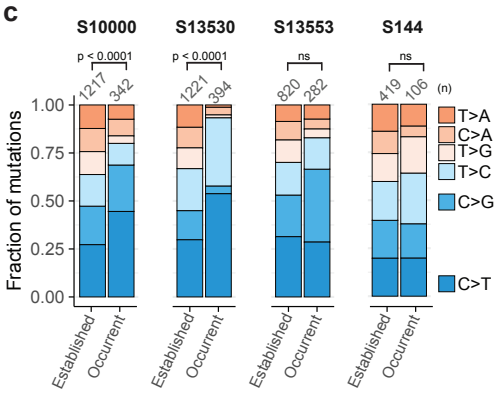

a
Input
Preproccesing
12 FL biopsies
5 CLL biopsies
scBCR-seq
scRNA-seq
1 BC/UMI extraction & correction
BC Whitelist
BC1
BC2
BC3
BC list
UMI Before
UMI After
2 Alignment
3 Reference selection
Ref 1
Ref 2
4 Re-alignment
Ref 1
Preprocessing
Filtering
5 Build consensus by cell/UMI
BC1:Cell1
BC1U1
BC1U2
BC1U3
6 Event calling
BC1Un
Event
Error
5 Steps
1 Read quality
2 Doublet (GEX) removal
3 10 UMI (each 5 reads)
4 Ratio event => 20%
5 Doublet (BCR) removal
b
FL S10000 / cell CAGCCGACACACCGCA
A C G T
UMIs
2x2
84
3x2
96
PO
1 50 78 100 154 200 250 300 350
VJ position
c
FL S13530 / cell CGTAGGCAGCGCTCCA
A C G T
UMIs
6
82
2
118
PO
1 50 100 150 184 210 250 300 350
VDJ position
d
FL S13530 / cell CGTAGGCAGCGCTCCA
A C G T
UMIs
2x3
6
150
3
60
PO
1 50 66 90 99 150 200 250 300
VJ position

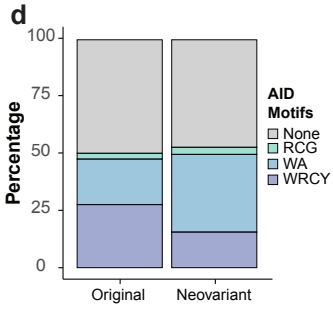
Figure 3.1. Detection of occurrent SHM events in single FL B cells. **(a)** Bioinformatics approach for calling alternative variants in V(D)J genes within single cells based on combined scRNA-seq and scBCR data. A custom pipeline was employed for preprocessing and event calling in scBCR-seq. Five filtering steps were implemented to identify cell-level variants. The number of reads and UMIs parameter was determined based on the median of the estimator's distribution (Supp. Fig. 2). The previously calculated parameters were then applied in step 3 of our filtering process. **(b-d)** Visualization of representative occurrent SHM within V(D)J sequences of individual FL cells. The x-axis corresponds to the V(D)J nucleotide sequence. The y-axis represents the number of V(D)J UMIs present in each cell with indication how many identical UMI were present. Coloured vertical bars indicate variants relative to the subject's primordial V(D)J sequence. The colours of the bars indicate the respective nucleotide changes. Bright coloured bars indicate called variants that passed all filtering steps and are considered the result of an occurrent SHM. The length of the bright bars represents the number of Ig UMIs of transcripts altered by the occurrent SHM event. CDR (complementarity-determining region) are highlighted with a reddish background. PO: Primordial sequence reconstructed from the respective germ-line IGV and IGJ alleles with the CDR3 inserted from the cell with the lowest SHM burden. AID (activation-induced cytidine deaminase) motifs are represented below the PO sequence on both corresponding DNA strands (upper row: WGRY motif, middle row: WA motif, lower row: RCG motif).

94 cells (25.8%) had a single occurrent SHM event on either heavy or light Ig transcript, and 2 cells (0.55%) single events on both heavy and light chains. In 268 single FL cells, at least one of both VDJ/VJ BCR transcripts carried more than one alternative nucleotides at different positions (Figure 3.1b-d). In 93.9% of these instances of co-occurrent SHM events on the same Ig chain of an individual cell, >90% of all V(D)J sequences carried either the most abundant combination of variants or the respective alternative nucleotide at all positions (i.e. the "mirroring" combination of variants). In only 17 V(D)J transcripts, sequences with combinations of variants that did not belong to the most abundant or its "mirroring" combination amounted to more than 10% of all UMI's Figure 3.2.


a
n events
Percentage of transcript
BCR transcripts
BCR transcript
VDJ
VJ
Transcript population
Most abundant variant combination
"mirroring" combination to most abundant
Mixed combination 1
Mixed combination 2
Mixed combination 3
Mixed combination 4
Mixed combination 5


b
Global
p < 0.0001
ns
834 7459 1055 (n)
Fraction of mutations
T>A
C>A
T>G
T>C
C>G
C>T
WES
Established
Occurrent


c
S10000
S13530
S13553
S144
p < 0.0001
p < 0.0001
ns
ns
1217 342
1221 394
820 282
419 106
(n)
Fraction of mutations
T>A
C>A
T>G
T>C
C>G
C>T
Established
Occurrent


d
Percentage
AID Motifs
None
RCG
WA
WRCY
Original
Neovariant

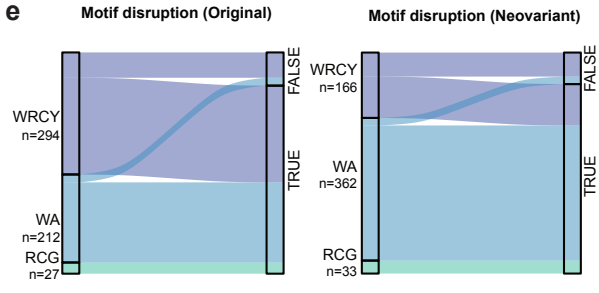

e
Motif disruption (Original)
Motif disruption (Neovariant)
WRCY n=294
WA n=212
RCG n=27
WRCY n=166
WA n=362
RCG n=33
FALSE
TRUE

Figure 3.2. Characterization of occurrent SHM events. **(a)** Combinations of multiple SHM events in Ig transcripts from individual cells. For each BCR transcript from a single cell with more than one identified SHM event, the proportion of (i) UMIs with the most abundant combination of occurrent SHM variants, (ii) UMIs with the respective alternative variant at every occurrent SHM position (the "mirroring" combination), and (iii) UMIs with any other combination of variants (up to five) is indicated. BCR transcripts are ranked according to the combined proportions of the most abundant and the mirroring combination. BCR type and number of occurrent SHM event per transcript are indicated at the top of the graph. The bar graph displays varying levels of transcript percentages across categories. Additionally, the number and location of somatic hypermutation (SHM) events in each gene are annotated above the corresponding cell. **(b)** Distribution of all 6 possible nucleotide mutations between bulk whole exome sequencing (WES) FL V(D)J sequences(Sepulveda-Yanez et al., 2021), established V(D)J variants originating from prior SHM at single cell-level, and changes from original to neovariant in occurrent SHM events across all cases of this study. The frequency of C>T mutations was compared by Chi-square test. **(c)** Distribution of all 6 possible nucleotide mutations between established V(D)J variants originating from prior SHM at single cell-level and changes from original to neovariant in occurrent SHM events in 4 individual FL samples with more than 100 occurrent SHM events. The frequency of C>T mutations was compared by Chi-square test. **(d)** Positioning of original variants and neovariants in AID motifs. **(e)** Disruption of AID motifs containing original and neovariants by occurrent SHM events.

CHARACTERIZATION OF OCCURRENT SHM EVENTS

To characterize the observed occurrent SHM events, we sought to identify which of the two alternative nucleotides existed prior to the SHM event and which was created by SHM. For isolated single variants in a V(D)J sequence, a nucleotide identical to the respective position in the corresponding germ-line IGV sequence was designated as "original" and the alternative variant as "neovariant". For 256 of all 440 (58.2%) isolated single occurrent SHM events, an original and neovariant could be distinguished. In 184 isolated occurrent SHM events (41.8%), both alternative nucleotides already differed from the germ-line sequence, and the bona fide neovariant could not be identified. For SHM events occurring in tandem, the observed combination of variants was defined as original if the majority of them corresponded to the primordial sequence, and the combination of alternative nucleotides as neovariants. In total, bona fide original and neovariant alternatives could be distinguished for 1055 of the 1239 occurrent SHM variants (85.2%). Mechanistically, AID initiates SHM by deamination of DNA cytosine residues into uracil [96, 178]. Through recognition of uracil as thymine by the DNA polymerase, C-to-T transition is the most direct AID-catalyzed mutation that does not require any additional repair pathway [111, 179, 180]. As a consequence, C-to-T transitions are the most abundant mutations created by SHM and corresponded to 32% in a separate cohort of FL samples analyzed by whole exome sequencing (WES) [181] and 31.5% of all established V(D)J mutations detected in single FL cells of this study (Figure 3.2b). For occurrent SHM events, the fraction of C-to-T transitions was even significantly higher as also seen in the two individual FL samples with the highest number of occurrent SHM (Figure 3.2c). AID preferentially catalyzes deoxycytidine deamination at canonical WRCY motifs (W = A/T; R =

A/G; Y = C/T) [40, 101], non-canonical WA motifs [182, 183], and RCG motifs [184]. Of the 1055 occurrent V(D)J variants where the original nucleotide could be identified, 27.7% were located in the dominant WRCY or any other AID hotspots, respectively. In contrast, only 15.6% of their corresponding neovariants ($p<0.0001$) were part of a WRCY motif (Figure 3.2). Furthermore, 79.3% of the WRCY motif of the original variants were destroyed by occurrent SHM events, while the hypothetical WRCY motifs of the neovariants would have been destroyed in only 63.2% Figure 3.2. This disparity in the destruction of WRCY motifs ($p<0.001$, Chi-square test) supports correctness of our annotation of occurrent SHM mutation into original or neovariant.

ESTIMATION OF SHM RATE

Based on the estimated half-life of Ig mRNA of app. 3 h [185, 186] and restriction of the observed fraction of UMI carrying an alternative variant to 0.2 – 0.8 by the filtering criteria, the actual creation of the observed neovariants can be estimated to have occurred between approximately 1 and 7 h prior to the time of sampling (Supp. Fig. S3). For FL case S13530, where 0.56% of cells had undergone a detectable SHM event during this time span of 6 h, it can be concluded that every lymphoma cell undergoes a somatic hypermutation event on average approximately every 45 days. Under the assumption that the SHM rate was constant over time and 2.98 neomutations were acquired in tandem at each SHM event, a combined heavy and light chain IGV mutation burden of 84 implies that FL S13530 was arrested at the GC stage around 3.5 years prior to sampling (Supp. Table S3).

EXPRESSION ANALYSIS AND ENRICHMENT ANALYSIS

Gene expression analysis identified 5766 genes with differential expression with an adjusted $p \leq 0.05$ in cells with occurrent SHM events (2920 genes upregulated, 2846 genes downregulated; (Figure 3.3a). ***AICDA***, the deaminase initiating SHM [187], was upregulated in cells with occurrent SHM events (Figure 3.3b). In contrast, expression of the RNA-specific deaminase ***ADAR*** was low (mean expression: -0.107), arguing against RNA editing as the cause of occurrent Ig neovariants (Figure 3.3c). KEGG pathway analysis identified mismatch repair (hsa03430, 9 enriched genes with p.adjust < 0.001), base excision repair (hsa03410, 17 enriched genes with p.adjust = 0.01), and DNA replication (hsa03030, 14 enriched genes with p.adjust < 0.001) as the three most prominently upregulated pathways in FL cells with occurrent SHM events (Figure 3.3d). Additional upregulated KEGG pathways included platinum drug resistance (hsa01524, 20 enriched genes with p.adjust = 0.001), the Fanconi anemia pathway (hsa03460, 11 enriched genes with p.adjust = 0.01), and cell cycle (hsa04110, 40 enriched genes with p.adjust = 0.004). Downregulated genes were enriched in transcriptional misregulation in cancer (hsa0520242, 14 enriched genes with p.adjust = 0.03), among others.

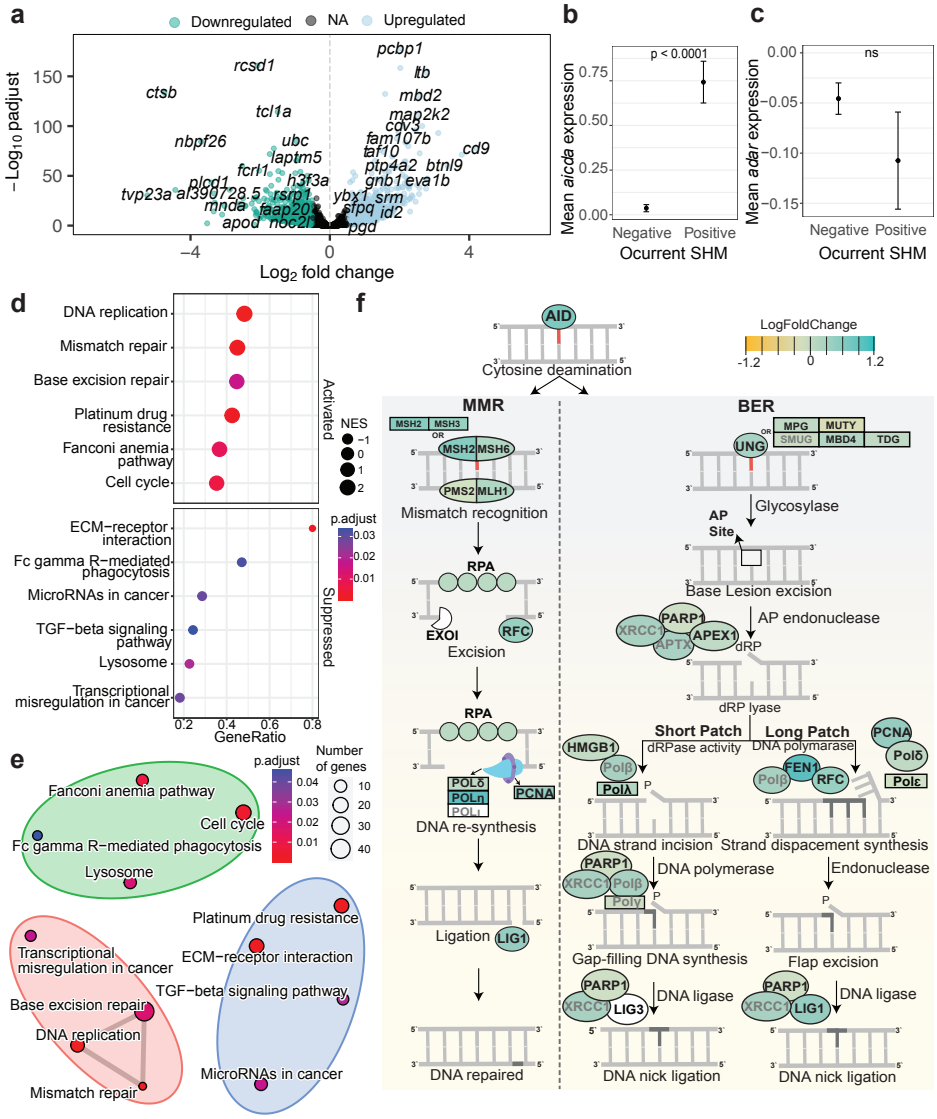

a
Downregulated
NA
Upregulated
−Log10 padjust
Log2 fold change
pcbp1
rcsd1
ltb
ctsb
mbd2
tcl1a
map2k2
cdv3
nbpf26
ubc
fam107b
cd9
laptm5
taf10
fcrl1
ptp4a2
btnl9
h3f3a
gnb1
eva1b
plcd1
tvp23a
al390728.5
rsrp1
ybx1
srm
rnnda
faap20
sfpq
id2
apod
hoe2h
pgd
b
p < 0.0001
Mean aicda expression
Negative
Positive
Ocurrent SHM
c
ns
Mean adar expression
Negative
Positive
Ocurrent SHM
d
DNA replication
Mismatch repair
Base excision repair
Platinum drug resistance
Fanconi anemia pathway
Cell cycle
Activated
NES
ECM−receptor interaction
Fc gamma R−mediated phagocytosis
MicroRNAs in cancer
TGF−beta signaling pathway
Lysosome
Transcriptional misregulation in cancer
Suppressed
p.adjust
GeneRatio
e
p.adjust
Number of genes
Fanconi anemia pathway
Cell cycle
Fc gamma R−mediated phagocytosis
Lysosome
Platinum drug resistance
ECM−receptor interaction
TGF−beta signaling pathway
Transcriptional misregulation in cancer
Base excision repair
DNA replication
Mismatch repair
MicroRNAs in cancer
f
AID
Cytosine deamination
LogFoldChange
MMR
BER
MSH2
MSH3
MSH2 MSH6
PMS2 MLH1
Mismatch recognition
RPA
EXO1
RFC
Excision
POLδ
POLη
POLι
PCNA
DNA re-synthesis
Ligation
LIG1
DNA repaired
UNG
MPG
MUTY
SMUG
MBD4
TDG
Glycosylase
AP Site
Base Lesion excision
PARP1
XRCC1
APTX
APEX1
AP endonuclease
dRP
dRP lyase
Short Patch
dRPase activity
Long Patch
DNA polymarase
HMGB1
Polβ
PolλP
FEN1
RFC
PCNA
Polδ
Polε
DNA strand incision
Strand dispacement synthesis
DNA polymerase
Endonuclease
PolγP
Gap-filling DNA synthesis
Flap excision
LIG3
LIG1
DNA ligase
DNA nick ligation

Figure 3.3. (a) Volcano plot of differential gene expression analysis (DEG) between FL cells with and without occurrent SHM. X-axis: log2 fold change (LFC). Y-axis: -log10(p.adjust). Up-regulated genes in cells with occurrent SHM are indicated by green dots, down-regulated genes by blue dots, and genes without significant differences by black dots. **(b)** Normalized expression levels of AICDA in FL cells with and without occurrent SHM. Expression levels were compared by Wilcoxon's test. **(c)** Normalized expression levels of ADAR in FL cells with and without occurrent SHM. Expression levels were compared by Wilcoxon's test. **(d)** Enrichment analysis of KEGG gene expression pathways in cells with occurrent SHM (filtered with $p < 0.5$). The size and colour of the points indicate normalized enrichment score (NES) and the p.adjust, respectively. The gene ratio demonstrates the enrichment proportion in the KEGG pathway. **(e)** Network representation of enriched pathways. The top 12 pathways from the KEGG analysis are highlighted. Each node represents a distinct pathway. The size of the node corresponds to the number of different genes enriched within the pathway. Edges represent gene overlaps or cross-talk between pathways, and the line width indicates the strength of the gene overlap. Colour coding indicates p.adjust values. The ellipses clusters indicate groups of pathways that share genes and form functional modules. **(f)** Somatic hypermutation pathway in cells with occurrent SHM. The proteins are coloured based on the log2 fold change from DEG analysis. Grey names indicate non-significant LFC, and white colouring represents protein without DE results.

The SHM-related mismatch repair, base excision repair, and DNA replication pathways belonged to the same functional gene cluster by network representation (Figure 3.3e). Since the entire SHM pathway is not defined in the KEGG gene set, we developed a custom annotation for the SHM pathway (Figure 3.3f). The uracil derived from deaminated cytosine can be modified by base excision repair (BER) [102] and/or subsequently mismatch repair (MMR) [188]. The recognition of deamination and binding to DNA mismatch is performed by either uracil DNA glycosylase (UNG) in BER [189] or by the MSH2-MSH6 mismatch recognition complex in MMR [179]. BER and MMR become error-prone in GC B cells, resulting in significantly increased number of mutations that accumulate in the V(D)J regions. It is estimated that over 50% of mutations in these regions are not the direct result of the biochemical action of AID but are created by the error-prone BER and MMR pathways [31]. In cells with occurrent SHM events, the MLH1/MSH2 genes associated with the recognition of the deamination process, as well as the POLH gene (DNA polymerase eta), which is involved in error-prone mutations, were highly expressed [190]. Overall, we observed a more prominent expression of genes related to MMR than BER in cells with neovariants supported by a higher normalized enrichment score (NES) in MMR (MMR:1.93, BER:1.6).

3.4 | DISCUSSION

By focusing on follicular lymphoma as a paradigmatic condition of expansion of clonal B cells with a GC phenotype, we here demonstrate for the first time that is possible to detect and to analyze the phenomenon of SHM at single-cell level. With rigid bioinformatic restrictions to minimize the chance of artifacts, we were able to detect the occurrence of very recent SHM events and to propose an approach to deduce the time frame for follicular lymphoma development after initiation of the malignant GC phenotype. In the informative example

of FL S13530, the deduced time of arrest at the GC stage with initiation of malignant tumor development fits the indolent nature of FL, where lymphadenopathy can exist for years before becoming symptomatic and requiring therapeutic interventions. The absence of detectable occurrent SHM events in CLL, a B-cell neoplasia characterized by a non-GC phenotype derived from either naïve or memory B cells [191] with low to absent IGHV mutation burden (Supp. Table S1), is in accordance with the GC as the predominant site of SHM and indirectly supports the validity of our findings in FL.

However, the actual rate of ongoing SHM in FL as described with our approach is quite variable and must be substantially higher in some cases. Underestimation of the actual SHM rate may be attributable to our particularly strict filtering criteria intended to avoid false-positive SHM events. In addition, current technical limitations are very likely to limit correct quantification of occurrent SHM. Across the FL samples, we noted that scBCR-seq consistently yields less analyzable VDJ sequences than light chain transcripts (p=0.0004; Supp. Table S3). Therefore, occurrent SHM events in VJ were detected in six but VDJ events in only four FL biopsies.

Furthermore, no occurrent SHM events were detected in seven FL biopsies, while these cases had accumulated similar established burdens of SHM than cases with detectable occurrent SHM. Improvement of single-cell sequencing technology can therefore be expected to permit more exact assessment of ongoing SHM than currently possible. Assignment of the individual alternative V(D)J variants into bona fide pre-SHM “original” nucleotide and the created “neovariant” based on germ-line IGV sequences supported the validity of the detected SHM events. Relevant findings include a relatively dominant proportion of C-to-T transitions, positioning of a high fraction of the original variants in the WRCY recognition motif of AID, and frequent destruction of the WRCY motif through the mutation from original to neovariant. While we noted a significant expression of POLH, known to target the WA motif, there were fewer variants in this motif compared to WRCY. This observation might be explained by considering the potential role of POLH a in a more complex, hybrid pathway. This pathway could involve elements of non-canonical mismatch repair (ncMMR) and UNG2-dependent translesion synthesis (TLS), potentially leading to a predominance of C/G transversions [192]. In the case of the RCG motif, the percentage of variants observed in this study (2.5%) closely corresponds with the findings from a separate cohort of follicular lymphoma (FL) samples analyzed by WES, where the RCG motif variants accounted for 2% [181]. Unexpectedly, the proportion of C-to-T transitions among the occurrent SHM events was even higher than in established and hence propagated V(D)J mutations. A possible explanation of this phenomenon may be the detection of the occurrent SHM event at a very recent time point when still substantial amount of transcripts carrying the original variant are present. Some of the C-to-T transitions may be not tolerated in a functional BCR and lead to eventual apoptosis of the affected cell once the abundance of the original, functional BCR falls below a critical threshold. In contrast, established SHM detected in surviving cells are selected for non-deleterious primary C-to-T transitions only. Hence, the observed fraction of C-to-T transitions in occurrent SHM events may more accurately reflect the true distribution of AID-induced mutations prior to eventual

selection of established V(D)J mutations.

The second relatively unexpected key finding was the large proportion of BCR transcripts with more than one SHM event. Distribution of BCR transcripts from the same Ig locus into two "mirroring" species as defined by their particular combination of variants in the overwhelming number of instances proves that occurrent SHM events according to our definition do not represent random observations and provides further internal validation of the observed variants. Therefore, SHM-induced neovariants are frequently created concomitantly in tandem without any intermittent mRNA transcription. This conclusion is in accordance with the reported mechanisms of AID deamination, whereby AID binds to single-stranded DNA (ssDNA) in a random and bidirectional manner, scanning it through short sliding or hopping movements. Notably, AID scans the entire length of the ssDNA multiple times during a single binding event allowing multiple deamination events, with an average bound lifetime of approximately five minutes [40, 193, 194].

Gene expression profiling of FL cells with occurrent SHM demonstrated higher expression of the known relevant pathways executing SHM, from overexpression of the initiator AID to the pathways actually creating alternative nucleotides, i.e. DNA replication, mismatch repair, and base excision repair. Therefore, the gene expression profile of FL cells provides intrinsic validation of the observation of ongoing SHM. While current evidence indicates that AID activity triggers DNA repair pathways [102, 179, 188], this phenomenon has not yet been observed in single cells ex vivo. DNA replication is crucial for effective AID deamination activity, as AID requires access to single-stranded DNA (ssDNA) located within the transcription bubble or immediately upstream of the replication fork [40, 195]. Our previous analysis of mutational signatures in FL have already identified the signature SSB6 associated with MMR in this tumor [181]. The additional overexpression of DNA replication and cell cycle genes indicate that FL cells that actively execute SHM have similarities with normal dark zone GC B cells. In contrast, FL cells without active SHM more closely resemble the gene expression profile typical of light zone GC B cells [196]. Enrichment of downregulated genes in the ECM-receptor interaction, Fc gamma R-mediated phagocytosis, and TGB-beta signaling pathways would also be related to a GC dark zone expression profile (Supp. Fig. S4) [197, 198]. Finally, the observed upregulation of the Fanconi anemia pathway in cells with neovariants is possibly related to the previously reported presence of somatic mutations in genes of this pathway in FL [181].

However, the overexpression of various KEGG pathways does not permit to compare their relative contribution to complex biological processes. To accurately represent the entire SHM pathway, we therefore generated a custom annotation and observed high expression of key genes involved in the recognition of the deamination process, such as MLH1/MSH2, as well as the error-prone mutation-associated gene, POLH, from the mismatch repair pathway. This observation suggests that cells executing SHM exhibit a more prominent expression of genes related to MMR compared to BER pathways. The error-prone repair in both MMR and BER pathways is primarily carried out by ubiquitously expressed, low-fidelity translesional DNA polymerases [199, 200]. These error-prone polymerases have a shallow and less constrained

binding pocket than high-fidelity polymerases [201]. This feature allows low-fidelity polymerases such as POLH to recognize and bypass DNA lesions during replication or repair by inserting bases opposite the lesion. Therefore, their permissive binding sites explain the lower fidelity and the higher likelihood to introduce errors during DNA synthesis [202].

In conclusion, this study describes for the first time the detection of occurrent SHM in single cells in a majority of FL samples but not in CLL used as control. In addition, we provide new insights into the molecular mechanisms of SHM. The occurrence of SHM at single-cell level was correlated with higher expression of AID and error-prone DNA repair pathways, in particular MMR. Quantification of SHM at the single-cell level facilitates improved understanding of the cellular heterogeneity and dynamics in physiological GC reactions but mat also offer novel insights into the AID-dependent shaping of the genomic landscape of germinal center B-cell neoplasms [181, 203, 204].

DATA AVAILABILITY

The data underlying Figs. 1, 2 and 3 are openly available in GEO at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE236880`.

ACKNOWLEDGMENTS

J.H.S.-Y. is funded by Doctorado Becas Chile (2016-72170683). M.A.N. acknowledges ANID for the grants Anillo ATE220016 and Fondecyt 1230298.

3.5 | AUTHOR CONTRIBUTIONS

E.Q., and C.A.M.v.B. performed the single cell sequencing experiment. J.H.S-Y., D.A-S., S.M.K. and D.M. contributed to analytical bioinformatics and statistics. H.V., M.A.N., and J.H.S-Y. wrote the manuscript. H.V. and C.A.M.v.B designed the study.

3.6 | COMPETING INTERESTS

The authors declare no competing interests.

3.7 | SUPPLEMENTARY MATERIAL

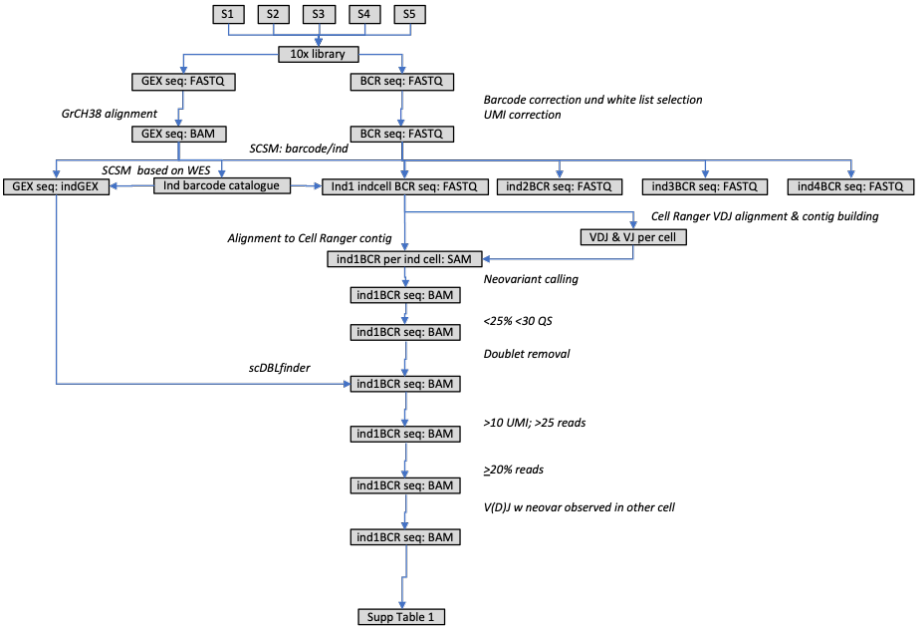
SUPPLEMENTARY FIGURES

Figure S1. Integrated Bioinformatics Pipeline for GEX and BCR Sequencing Analysis.

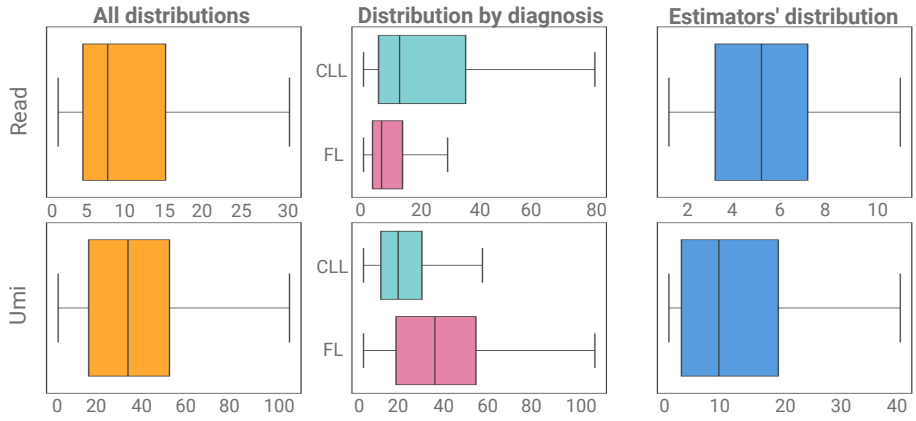
Figure S2. The distribution of reads and UMIs was analyzed across all samples and stratified by diagnosis, using the corresponding median from the estimator's distribution.

Figure S3. Estimation of SHM Event Timing through V(D)J mRNA Fraction Analysis.

Figure S4. Gene expression levels in pseudobulk of SHM positive cells and SHM negative cells.



Supplementary Figure 1. Integrated Bioinformatics Pipeline for GEX and BCR Sequencing Analysis. This schematic illustrates a pipeline used for analyzing gene expression (GEX) and B-cell receptor (BCR) sequencing data across multiple samples. The process begins with raw GEX and BCR sequencing outputs in FASTA format, which for GEX is further aligned to produce BAM files. An index barcode catalog is generated from the GEX data. BCR sequences (ind1 to ind4) undergo processing to produce SAM files after alignment and contig building, later converted to BAM files, where it is produced the neovariant calling and further filtering. The final data is compiled in Supplementary Table 1.

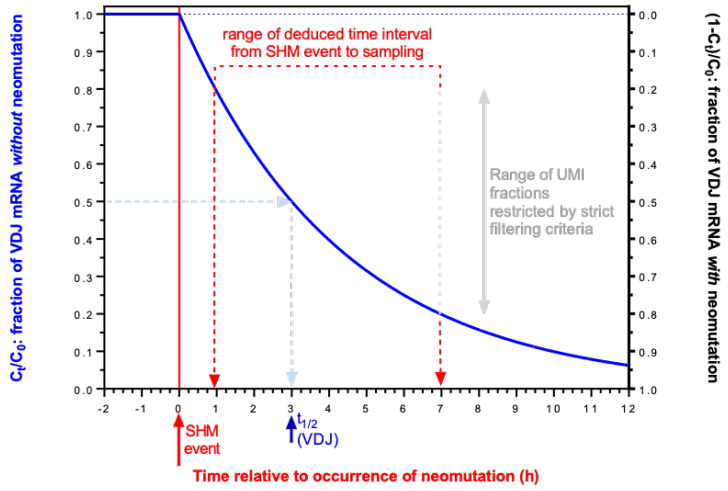


Supplementary Figure 2. The distribution of reads and UMIs was analyzed across all samples and stratified by diagnosis, using the corresponding median from the estimator's distribution.

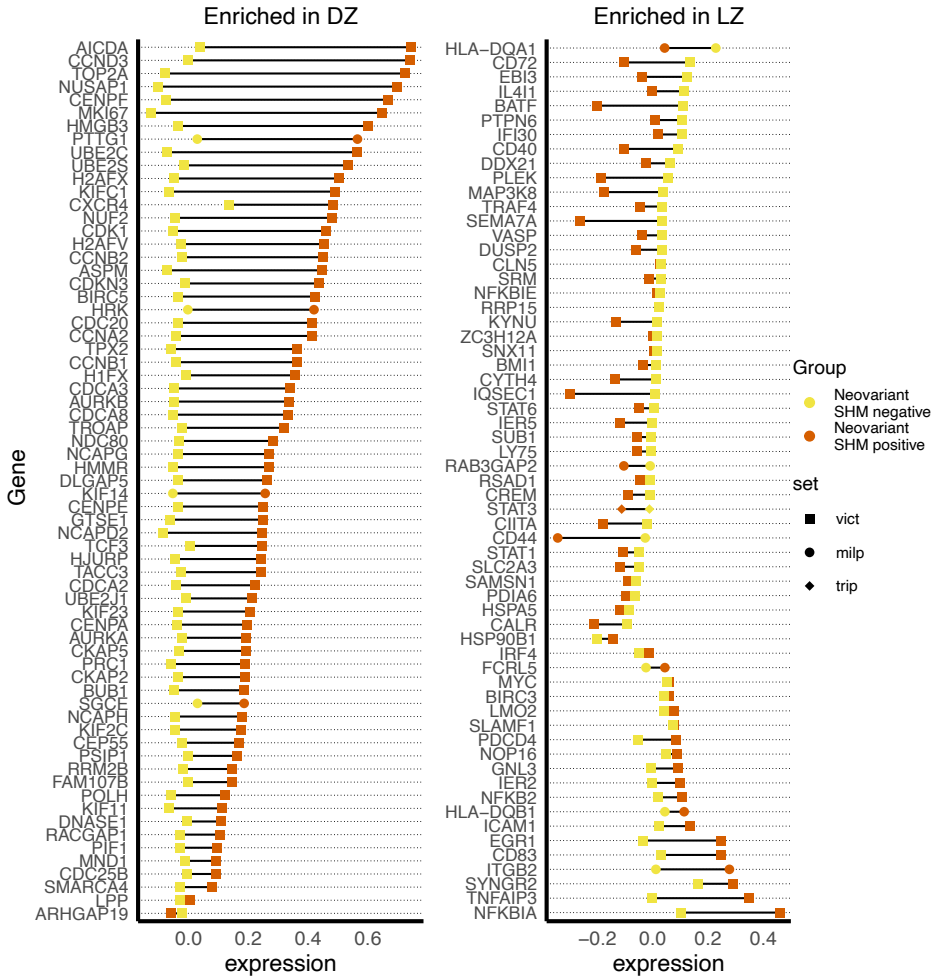
Assumptions:
- V(D)J mRNA half-life $t_{1/2}$=3h
- Rate of transcription constant over time prior, during, and after occurrence of neomutation
- First-order (=constant) decay rate λ of mRNA
Therefore mRNA concentration constant over time (C_0)

After steady-state prior to occurrence of neomutation, decrease of the concentration C_t of the VDJ mRNA without neomutation is described over time as:
$C_t/C_0 = e^{-\lambda t}$ with $\lambda = \ln(2) / t_{1/2} = \ln(2) / 3 = 0.231049$, therefore: $C_t/C_0 = e^{-0.231049t}$
- with t = h from occurrence of neomutation

To deduce time of neomutation from observed counts of UMI with and without neomutation:
$t = \ln(UMI_{w/o\ neomut}/[UMI_{w/o\ neomut}+UMI_{w\ neomut}]) / -0.231049$



Supplementary Figure 3. Estimation of SHM Event Timing through V(D)J mRNA Fraction Analysis. This graph shows a quantitative analysis of transcript dynamics post-somatic hypermutation (SHM) event. The x-axis represents the time elapsed since the occurrence of the neovariant, while both y-axis quantifies the fraction of unique molecular identifiers (UMIs) that carry or not the SHM. The curve illustrates the decay in V(D)J mRNA over time, indicative of the dilution of the V(D)J without neovariant as B cells proliferate. Two critical time points are marked: the initial SHM event and the half-life (t1/2) of the V(D)J mRNA. The red dotted lines delineate the inferred time interval between the SHM event and sample collection, with an additional range indicated for UMI fractions restricted by stringent filtering criteria.



Supplementary Figure 4. Gene expression levels in pseudobulk of SHM positive cells and SHM negative cells. Genes were analyzed that were significantly differentially expressed in single cells data and that were characteristic for dark (A) and light (B) zone germinal center B cells.

SUPPLEMENTARY TABLES

Table S1. Clinical profiles, immunoglobulin gene variability, and Single-Cell Sequencing summary

Table S2. Ocurrent somatic hypermutation in follicular lymphoma

Table S3. Immunoglobulin Gene Somatic Hypermutation dynamics and clinical correlations

	CLINICAL DATA								SINGLE-CELL RNAseq CELL NUMBERS			BCR		IGHV % identity to GL				
Patient	Diagnosis	Primary / Relapse	Tissue	Sex	Age (years)	Ann Arbor Stage	BCL-2 break	FLIPI	Input	Output gene expression	Return (%)	Output paired HC+LC BCR	Isotype	Light chain	IGHV gene	median	min	max	IGKV IGLV gene
S10000	FL	Untreated	LN	f	48	IV	yes	2	12000	18060	150.50	8233	IgM	λ	IGHV3-33	89.58	87.85	90.28	IGLV1-40
S10971	FL	Untreated	LN	m	56	III	n.d.	2	2000	1366	68.30	109	IgM	κ	IGHV3-11	83.68	82.64	84.38	IGKV3-20
S11770	FL	Untreated	LN	m	59	II	yes	0	2000	361	18.05	40	IgG	κ	IGHV3-23	97.22	96.53	97.57	IGLV3-21
S12118	FL	Relapse	LN	f	60	IV	yes	1	2000	940	47.00	609	IgG	κ	IGHV4-34	86.32	85.26	87.72	IGLV1-40
S12282	FL	Relapse	LN	m	79	IV	yes	5	2000	438	21.90	261	IgM	κ	IGHV3-11	92.01	90.28	93.06	IGKV3-20
S12390	FL	Untreated	LN	m	72	III	no	3	2000	1875	93.75	1571	IgM	κ	IGHV3-48	91.32	89.93	92.36	IGKV1-39
S12500	FL	Untreated	BM	f	66	IV	yes	2	2000	1532	76.60	1382	IgM	λ	IGHV4-39	82.47	80.76	84.54	IGLV2-14
	FL	Untreated	LN	f	66	IV	yes	2	2000	1129	56.45	1011	IgM	λ	IGHV4-39	82.13	79.38	84.19	IGLV2-14
S13530	FL	Untreated	LN	m	32	IV	yes	2	12000	14261	118.84	8220	IgG	λ	IGHV3-48	81.25	79.51	85.07	IGLV3-19
S13553	FL	Untreated	LN	m	54	III	yes	2	12000	17851	148.76	6615	IgG	κ	IGHV3-23	77.08	72.57	88.89	IGKV4-1
S144	FL	Untreated	LN	m	37	IV	yes	3	2000	1553	77.65	1214	IgM	κ	IGHV3-30	93.75	91.67	93.75	IGKV1-17
S8382	FL	Untreated	LN	m	44	IV	yes	2	2000	1486	74.30	1150	IgM	κ	IGHV3-48	87.85	86.81	89.93	IGLV2-8
S8934	FL	Relapse	LN	m	69	IV	yes	3	12000	11188	93.23	6535	IgG	κ	IGHV3-15	88.10	87.41	88.44	IGKV3-20
	FL	Relapse	LN	m	73	IV	yes	3	12000	13341	111.18	6035	IgG	κ	IGHV3-15	88.44	86.39	88.78	IGKV3-20
										Median	**77.13**	**1298**				**87.98**			
										Minimum	18.05	40				77.08			
										Maximum	150.50	8233				97.22			
												42985							
S356	CLL		PB	m	55				2000	1170	58.50	956	IgM	κ	IGHV1-69D	100.00	99.65	100.00	IGKV3-20
S4725	CLL		PB	m	58				2000	1346	67.30	1098	IgM	κ	IGHV1-46	100.00	99.31	100.00	IGKV1-5
S5714	CLL		PB	m	56				2000	1138	56.90	994	IgM	λ	IGHV4-28	97.92	97.22	97.92	IGLV3-21
S7482	CLL		PB	f	40				2000	655	32.75	375	IgM	λ	IGHV1-69D	100.00	99.31	100.00	IGLV3-9
S9856	CLL		BM	m	70				2000	622	31.10	520	IgM	λ	IGHV3-48	98.26	97.92	98.26	IGLV3-21
										Median	**56.90**	**956**				**100.00**			
										Minimum	31.10	375				97.92			
										Maximum	67.30	1098				100.00			
										FL versus CLL Mann-Whitney	*p=0.087*	*p=0.26*				*p=0.0002*			

Table S1. Clinical profiles, immunoglobulin gene variability, and Single-Cell sequencing summary.

Supplemental Table 2

order	Chip	Sample	n_cell_ID	Ig Chain	V(D)Jmutat	VDJmutatic	VJmutatior	inBoth	cell	region	subregion	position	VDJ_positic	nucl_po	original	neovariant	neovariant	original_un	neovariant	neovariant	productive	h_from_ev	codon_cc	aminoacid_	aminoacid_	aminoacid_	mutation_t	context_po	context_cc	variation_c	original_ai	neovariant_	motif_dest	strand_of_r	motif_destroyed_neovariant
1	K1B	S144	144.1	L	1	0	1	FALSE	AACCGCGA	V	FR2	262	133	A	A	C		32	52		TRUE	4.176953	XAG	K	Q		missense	CTAAG	CTAAG	A-C	WA	None	TRUE	coding	No_motif
2	K1B	S144	144.2	L	1	0	1	FALSE	AACTCAGC	V	FR1	167	61	A	A	T		10	17		TRUE	4.298879	XTC	I	F		missense	CCATC	CCATC	A-T	None	WA	No_motif	No-motif	TRUE
3	K1B	S144	144.3	L	3	0	3	FALSE	AAGGTTCG	V	CDR3	395	266	T	T	C		10	17		TRUE	4.298879	CXA	L	P		missense	TCTAC	TCTAC	T-C	WA	None	TRUE	non-coding	No_motif
4	K1B	S144	144.3	L	3	0	3	FALSE	AAGGTTCG	V	FR2	230	101	G	G	C		17	23		TRUE	3.703397	GXC	G	A		missense	AGGCT	AGGCT	G-C	WRCY	WRCY	FALSE	non-coding	FALSE
5	K1B	S144	144.3	L	3	0	3	FALSE	AAGGTTCG	V	FR2	262	133	A	A	C		17	23		TRUE	3.703397	XAG	K	Q		missense	CTAAG	CTAAG	A-C	WA	None	TRUE	coding	No_motif
6	K1B	S144	144.4	L	2	0	2	FALSE	ACACCCTC	V	FR1	152	46	G	G	A		24	19		TRUE	2.523907	XGA	G	R		missense	TAGGA	TAGGA	G-A	None	WA	No_motif	No-motif	TRUE
7	K1B	S144	144.4	L	2	0	2	FALSE	ACACCCTC	V	CDR3	383	277	A	A	T		12	18		TRUE	3.965785	XCT	T	S		missense	ATACT	ATACT	A-T	WA	None	TRUE	coding	No_motif
8	K1B	S144	144.5	L	8	0	8	FALSE	ACGCCAGA	V	CDR3	372	266	T	T	C		12	12		TRUE	3.000001	CXA	L	P		missense	TCTAC	TCTAC	T-C	WA	None	TRUE	non-coding	No_motif
9	K1B	S144	144.5	L	8	0	8	FALSE	ACGCCAGA	V	CDR3	376	270	G	G	A		12	11		TRUE	2.815799	CAX	Q	Q		silent	CAGCA	CAGCA	G-A	WRCY	WA	TRUE	non-coding	TRUE
10	K1B	S144	144.5	L	8	0	8	FALSE	ACGCCAGA	V	CDR3	383	277	A	A	C		11	10		TRUE	2.798658	XCT	T	P		missense	ATACT	ATACT	A-C	WA	None	TRUE	coding	No_motif
11	K1B	S144	144.5	L	8	0	8	FALSE	ACGCCAGA	V	FR1	131	25	T	T	C		17	14		TRUE	2.600201	XCC	S	P		missense	CATCC	CATCC	T-C	None	None	No_motif	No-motif	No_motif
12	K1B	S144	144.5	L	8	0	8	FALSE	ACGCCAGA	V	CDR1	189	83	G	G	A		17	14		TRUE	2.600201	GXC	G	D		missense	GGGCA	GGGCA	G-A	WRCY	None	TRUE	non-coding	No_motif
13	K1B	S144	144.5	L	8	0	8	FALSE	ACGCCAGA	V	CDR1	200	94	G	G	C		17	14		TRUE	2.600201	XAT	D	H		missense	ATGAT	ATGAT	G-C	None	None	No_motif	No-motif	No_motif
14	K1B	S144	144.5	L	8	0	8	FALSE	ACGCCAGA	V	CDR2	257	151	G	T	G		17	14		TRUE	2.600201	XCA	S	A		missense	CTGCA	CTTCA	T-G	None	None	No_motif	No-motif	No_motif
15	K1B	S144	144.5	L	8	0	8	FALSE	ACGCCAGA	V	FR3	284	178	T	T	G		17	14		TRUE	2.600201	XCA	S	A		missense	CATCA	CATCA	T-G	None	WRCY	No_motif	No-motif	TRUE
16	K1B	S144	144.6	L	1	0	1	FALSE	ACTTACTGT	V	FR3	299	193	A		C	T		17	13	TRUE		XAT		H	Y		GCAGT			Not_evalua	Not_evalua	No_evaluat	Not_evalua	No_evaluated
17	K1B	S144	144.7	L	4	0	4	FALSE	AGGTCATC	V	FR2	206	112	C	C	A		10	10		TRUE	3.000001	XAG	Q	K		missense	AGCAG	AGCAG	C-A	None	None	No_motif	No-motif	No_motif
18	K1B	S144	144.7	L	4	0	4	FALSE	AGGTCATC	V	FR2	240	146	A	A	C		10	10		TRUE	3.000001	TXT	Y	S		missense	CTATG	CTATG	A-C	WA	None	TRUE	coding	No_motif
19	K1B	S144	144.7	L	4	0	4	FALSE	AGGTCATC	V	FR3	252	158	G	C	G		10	10		TRUE	3.000001	AXT	T	S		missense	CAGTT	CACTT	C-G	None	WRCY	No_motif	No-motif	TRUE
20	K1B	S144	144.7	L	4	0	4	FALSE	AGGTCATC	V	FR3	256	162	G	G	C		10	10		TRUE	3.000001	TTX	L	F		missense	TTGCA	TTGCA	G-C	None	None	No_motif	No-motif	No_motif
21	K1B	S144	144.8	L	3	0	3	FALSE	AGTGAGGG	V	FR1	149	55	G	G	C		19	14		TRUE	2.3894	XTC	V	L		missense	GAGTC	GAGTC	G-C	None	None	No_motif	No-motif	No_motif
22	K1B	S144	144.8	L	3	0	3	FALSE	AGTGAGGG	V	FR3	252	158	G	G	C		17	14		TRUE	2.600201	AXT	S	T		missense	CAGTT	CAGTT	G-C	WRCY	None	TRUE	non-coding	No_motif
23	K1B	S144	144.8	L	3	0	3	FALSE	AGTGAGGG	V	FR3	314	220	A	A	G		18	12		TRUE	2.210897	XCA	T	A		missense	TCACA	TCACA	A-G	None	RCG	No_motif	No-motif	TRUE
24	K1B	S144	144.9	L	5	0	5	FALSE	AGTGTCAC	V	FR1	167	61	A	C	A		15	11		TRUE	2.380648	XTC	L	I		missense	CCATC	CCCTC	C-A	None	None	No_motif	No-motif	No_motif
25	K1B	S144	144.9	L	5	0	5	FALSE	AGTGTCAC	V	FR2	203	97	T	C	T		14	11		TRUE	2.509504	XTA	L	L		silent	ATTTA	ATCTA	C-T	None	WA	No_motif	No-motif	TRUE
26	K1B	S144	144.9	L	5	0	5	FALSE	AGTGTCAC	V	FR2	220	114	G	G	C		14	11		TRUE	2.509504	CAX	Q	H		missense	CAGAA	CAGAA	G-C	None	None	No_motif	No-motif	No_motif
27	K1B	S144	144.9	L	5	0	5	FALSE	AGTGTCAC	V	FR2	226	120	A	A	C		14	11		TRUE	2.509504	CCX	P	P		silent	CCAGG	CCAGG	A-C	None	None	No_motif	No-motif	No_motif
28	K1B	S144	144.9	L	5	0	5	FALSE	AGTGTCAC	V	FR3	304	198	A	A	C		14	11		TRUE	2.509504	GGX	G	G		silent	GGATC	GGATC	A-C	None	None	No_motif	No-motif	No_motif
29	K1B	S144	144.10	L	2	0	2	FALSE	ATCATGGG	V	FR1	110	4	A	A	T		12	23		TRUE	4.632963	XTC	I	F		missense	ACATC	ACATC	A-T	None	WA	No_motif	No-motif	TRUE
30	K1B	S144	144.10	L	2	0	2	FALSE	ATCATGGG	V	FR1	151	45	A	A	G		12	23		TRUE	4.632963	GTX	V	V		silent	GTAGG	GTAGG	A-G	WA	None	TRUE	coding	No_motif
31	K1B	S144	144.11	L	5	0	5	FALSE	CAACCAAA	V	CDR3	364	270	G	G	A		11	25		TRUE	5.131481	CAX	Q	Q		silent	CAGCA	CAGCA	G-A	WRCY	WA	TRUE	non-coding	TRUE
32	K1B	S144	144.11	L	5	0	5	FALSE	CAACCAAA	V	FR2	240	146	A	A	T		30	12		TRUE	1.456281	TXT	Y	F		missense	CTATG	CTATG	A-T	WA	WA	FALSE	coding	FALSE
33	K1B	S144	144.11	L	5	0	5	FALSE	CAACCAAA	V	FR3	274	180	A	A	T		30	12		TRUE	1.456281	TCX	S	S		silent	TCAAG	TCAAG	A-T	None	WA	No_motif	No-motif	TRUE
34	K1B	S144	144.11	L	5	0	5	FALSE	CAACCAAA	V	FR3	287	193	A	G	T		30	13		TRUE	1.558123	XAT	D	Y		missense	GCAGT	GCGAT	G-T	None	WA	No_motif	No-motif	TRUE
35	K1B	S144	144.11	L	5	0	5	FALSE	CAACCAAA	V	FR3	355	261	C	T	C		27	12		TRUE	1.591545	TTX	F	F		silent	TACTG	TTTTG	T-C	WA	None	TRUE	non-coding	No_motif
36	K1B	S144	144.12	L	1	0	1	FALSE	CAACCTCG	V	FR2	539	124	A	A	C		13	14		TRUE	3.163344	XAA	K	Q		missense	GGAAA	GGAAA	A-C	None	None	No_motif	No-motif	No_motif
37	K1B	S144	144.13	L	1	0	1	FALSE	CAAGTTGC	V	FR1	134	45	A	A	T		17	14		TRUE	2.600201	GTX	V	V		silent	GTAGG	GTAGG	A-T	WA	None	TRUE	coding	No_motif
38	K1B	S144	144.14	L	1	0	1	FALSE	CACAAACG	J	FR4	428	37	C	C	G		10	11		TRUE	3.211169	XGA	R	G		missense	AAC	AACGA	C-G	RCG	None	TRUE	coding	No_motif
39	K1B	S144	144.15	L	5	0	5	FALSE	CACATAGT	V	CDR3	372	266	T	T	C		26	50		TRUE	4.642465	CXA	L	P		missense	TCTAC	TCTAC	T-C	WA	None	TRUE	non-coding	No_motif
40	K1B	S144	144.15	L	5	0	5	FALSE	CACATAGT	V	CDR3	383	277	A	A	C		21	44		TRUE	4.890152	XCT	T	P		missense	ATACT	ATACT	A-C	WA	None	TRUE	coding	No_motif
41	K1B	S144	144.15	L	5	0	5	FALSE	CACATAGT	V	FR1	135	29	C	T	C		36	70		TRUE	4.673988	TXC	F	S		missense	CTCCC	CTTCC	T-C	None	None	No_motif	No-motif	No_motif
42	K1B	S144	144.15	L	5	0	5	FALSE	CACATAGT	V	FR1	146	40	T	T	A		36	70		TRUE	4.673988	XCT	S	T		missense	CATCT	CATCT	T-A	None	WA	No_motif	No-motif	TRUE
43	K1B	S144	144.15	L	5	0	5	FALSE	CACATAGT	V	FR3	287	181	A	A	C		36	69		TRUE	4.632963	XGG	R	R		silent	CAAGG	CAAGG	A-C	WA	RCG	TRUE	coding	TRUE
44	K1B	S144	144.16	L	1	0	1	FALSE	CAGCGACC	V	CDR2	255	149	C	C	T		31	30		TRUE	2.929624	GXT	A	V		missense	TGCTG	TGCTG	C-T	WRCY	WA	TRUE	coding	TRUE
45	K1B	S144	144.17	L	4	0	4	FALSE	CATGACAG	V	CDR3	372	266	T	T	C		11	15		TRUE	3.723025	CXA	L	P		missense	TCTAC	TCTAC	T-C	WA	None	TRUE	non-coding	No_motif
46	K1B	S144	144.17	L	4	0	4	FALSE	CATGACAG	V	CDR1	189	83	G	G	A		21	13		TRUE	2.085437	GXC	G	D		missense	GGGCA	GGGCA	G-A	WRCY	None	TRUE	non-coding	No_motif
47	K1B	S144	144.17	L	4	0	4	FALSE	CATGACAG	V	FR3	283	177	A	A	G		20	13		TRUE	2.167399	CCX	P	P		silent	CCATC	CCATC	A-G	None	RCG	No_motif	No-motif	TRUE
48	K1B	S144	144.17	L	4	0	4	FALSE	CATGACAG	V	FR3	332	226	A	T	A		18	12		TRUE	2.210897	XGC	C	S		missense	TCAGC	TCTGC	T-A	None	None	No_motif	No-motif	No_motif
49	K1B	S144	144.18	L	1	0	1	FALSE	CATTATCTC	V	FR2	249	143	T	T	C		12	22		TRUE	4.507502	AXC	I	T		missense	GATCT	GATCT	T-C	None	None	No_motif	No-motif	No_motif
50	K1B	S144	144.19	L	3	0	3	FALSE	CCAATCCTC	V	FR2	219	113	A	A	G		15	12		TRUE	2.543991	CXG	Q	R		missense	GCAGA	GCAGA	A-G	None	None	No_motif	No-motif	No_motif
51	K1B	S144	144.19	L	3	0	3	FALSE	CCAATCCTC	V	FR2	250	144	C	C	A		15	12		TRUE	2.543991	ATX	I	I		silent	ATCTA	ATCTA	C-A	None	WA	No_motif	No-motif	TRUE
52	K1B	S144	144.19	L	3	0	3	FALSE	CCAATCCTC	V	FR3	274	168	T	T	G		15	12		TRUE	2.543991	AGX	S	R		missense	AGTGG	AGTGG	T-G	None	None	No_motif	No-motif	No_motif
53	K1B	S144	144.20	L	3	0	3	FALSE	CCATTCGC	V	CDR3	372	266	T	T	C		22	12		TRUE	1.884094	CXA	L	P		missense	TCTAC	TCTAC	T-C	WA	None	TRUE	non-coding	No_motif
54	K1B	S144	144.20	L	3	0	3	FALSE	CCATTCGC	V	CDR3	378	272	A	A	G		20	11		TRUE	1.896805	CXT	H	R		missense	GCATA	GCATA	A-G	None	RCG	No_motif	No-motif	TRUE
55	K1B	S144	144.20	L	3	0	3	FALSE	CCATTCGC	J	CDR3	396	5	C	C	A		15	10		TRUE	2.210897	GXT	A	D		missense	CGCTT	CGCTT	C-A	None	None	No_motif	No-motif	No_motif
56	K1B	S144	144.21	L	4	0	4	FALSE	CCTACCATC	V	CDR1	197	92	A	A	G		39	45		TRUE	3.320746	AXT	N	S		missense	AAATG	AAATG	A-G	WA	None	TRUE	coding	No_motif
57	K1B	S144	144.21	L	4	0	4	FALSE	CCTACCATC	V	FR2	246	141	G	C	G		39	45		TRUE	3.320746	CTX	L	L		silent	CTGAT	CTCAT	C-G	None	None	No_motif	No-motif	No_motif
58	K1B	S144	144.21	L	4	0	4	FALSE	CCTACCATC	V	FR3	270	165	A	A	G		39	45		TRUE	3.320746	CAX	Q	Q		silent	CAAAG	CAACG	A-G	WA	None	TRUE	coding	No_motif
59	K1B	S144	144.21	L	4	0	4	FALSE	CCTACCATC	V	FR3	271	166	A	A	C		39	45		TRUE	3.320746	XGT	S	R		missense	AAAGT	AGAGT	A-C	None	RCG	No_motif	No-motif	TRUE
60	K1B	S144	144.22	L	4	0	4	FALSE	CCTTCCCTC	V	FR2	191	99	A	G	A		12	11		TRUE	2.815799	TTX	L	L		silent	TTAGG	TTGGG	G-A	None	WA	No_motif	No-motif	TRUE
61	K1B	S144	144.22	L	4	0	4	FALSE	CCTTCCCTC	V	FR2	214	122	G	G	C		12	11		TRUE	2.815799	GXG	G	A		missense	AGGGA	AGGGA	G-C	None	RCG	No_motif	No-motif	TRUE
62	K1B	S144	144.22	L	4	0	4	FALSE	CCTTCCCTC	V	FR2	215	123	G	G	C		12	11		TRUE	2.815799	GGX	G	G		silent	GGGAA	GGGAA	G-C	None	None	No_motif	No-motif	No_motif
63	K1B	S144	144.22	L	4	0	4	FALSE	CCTTCCCTC	V	FR3	318	226	A	A	G		11	11		TRUE	3.000001	XGC	S	G		missense	TCAGC	TCAGC	A-G	None	None	No_motif	No-motif	No_motif
64	K1B	S144	144.23	L	1	0	1	FALSE	CTAACTTGT	V	CDR1	193	83	G	G	C		16	11		TRUE	2.264663	GXC	G	A		missense	GGGCA	GGGCA	G-C	WRCY	None	TRUE	non-coding	No_motif
65	K1B	S144	144.24	L	2	0	2	FALSE	CTGGTCTG	V	FR3	313	219	C		A	C		27	18	TRUE		CTX		L	L		CTCAC			Not_evalua	Not_evalua	No_evaluat	Not_evalua	No_evaluated
66	K1B	S144	144.24	L	2	0	2	FALSE	CTGGTCTG	V	FR3	341	247	T		T	G		22	15	TRUE		XTT		F	V		ATTTT			Not_evalua	Not_evalua	No_evaluat	Not_evalua	No_evaluated
67	K1B	S144	144.25	L	4	0	4	FALSE	GAACATCTC	V	CDR3	372	266	T	T	C		14	13		TRUE	2.842598	CXA	L	P		missense	TCTAC	TCTAC	T-C	WA	None	TRUE	non-coding	No_motif
68	K1B	S144	144.25	L	4	0	4	FALSE	GAACATCTC	V	CDR3	377	271	C	C	T		12	12		TRUE	3.000001	XAT	H	Y		missense	AGCAT	AGCAT	C-T	None	WA	No_motif	No-motif	TRUE
69	K1B	S144	144.25	L	4	0	4	FALSE	GAACATCTC	V	CDR3	380	274	A	A	T		12	12		TRUE	3.000001	XAT	N	Y		missense	ATAAT	ATAAT	A-T	WA	WA	FALSE	coding	FALSE
70	K1B	S144	144.25	L	4	0	4	FALSE	GAACATCTC	J	FR4	424	33	C	C	G		11	12		TRUE	3.192392	ATX	I	M		missense	ATCAA	ATCAA	C-G	None	None	No_motif	No-motif	No_motif
71	K1B	S144	144.26	L	1	0	1	FALSE	GATGCTATC	V	CDR2	254	148	G	G	A		19	11		TRUE	1.97689	XCT	A	T		missense	ATGCT	ATGCT	G-A	None	WA	No_motif	No-motif	TRUE
72	K1B	S144	144.27	L	1	0	1	FALSE	GCACATATC	V	FR2	229	123	G	G	A		13	22		TRUE	4.286531	GGX	G	G		silent	GGGAA	GGGAA	G-A	None	None	No_motif	No-motif	No_motif
73	K1B	S144	144.28	L	2	0	2	FALSE	GGAATAAG	V	CDR1	181	92	A		A	T		37	18	TRUE		AXT		N	I		AAATG			Not_evalua	Not_evalua	No_evaluat	Not_evalua	No_evaluated
74	K1B	S144	144.28	L	2	0	2	FALSE	GGAATAAG	V	FR3	299	210	A		G	A		35	18	TRUE		GAX		E	E		GAATT			Not_evalua	Not_evalua	No_evaluat	Not_evalua	No_evaluated
75	K1B	S144	144.29	L	3	0	3	FALSE	GGCCGATA	V	FR1	158	64	A	G	A		18	12		TRUE	2.210897	XCT	A	T		missense	TCACT	TCGCT	G-A	RCG	None	TRUE	non-coding	No_motif
76	K1B	S144	144.29	L	3	0	3	FALSE	GGCCGATA	V	CDR2	247	153	A	A	T		18	12		TRUE	2.210897	GCX	A	A		silent	GCATC	GCATC	A-T	None	WA	No_motif	No-motif	TRUE
77	K1B	S144	144.29	L	3	0	3	FALSE	GGCCGATA	V	FR3	256	162	G	G	A		18	12		TRUE	2.210897	TTX	L	L		silent	TTGCA	TTGCA	G-A	None	WA	No_motif	No-motif	TRUE
78	K1B	S144	144.30	L	4	0	4	FALSE	GGCTGGTG	V	CDR3	372	266	T	T	C		31	16		TRUE	1.801178	CXA	L	P		missense	TCTAC	TCTAC	T-C	WA	None	TRUE	non-coding	No_motif
79	K1B	S144	144.30	L	4	0	4	FALSE	GGCTGGTG	V	CDR3	383	277	A	A	C		30	14		TRUE	1.657624	XCT	T	P		missense	ATACT	ATACT	A-C	WA	None	TRUE	coding	No_motif
80	K1B	S144	144.30	L	4	0	4	FALSE	GGCTGGTG	V	FR2	207	101	G	G	C		47	22		TRUE	1.661807	GXC	G	A		missense	AGGCT	AGGCT	G-C	WRCY	WRCY	FALSE	non-coding	FALSE
81	K1B	S144	144.30	L	4	0	4	FALSE	GGCTGGTG	V	FR3	300	194	G	A	C		44	21		TRUE	1.688809	TXT	Y	S		missense	CAGTG	CTATG	A-C	WA	None	TRUE	coding	No_motif
82	K1B	S144	144.31	L	3	0	3	FALSE	GGTATTGA	V	CDR2	262	156	C	G	C		45	21		TRUE	1.657624	TCX	S	S		silent	TCCAG	TCGAG	G-C	None	None	No_motif	No-motif	No_motif
83	K1B	S144	144.31	L	3	0	3	FALSE	GGTATTGA	V	FR3	271	165	A	A	G		45	21		TRUE	1.657624	CAX	Q	Q		silent	CAAAG	CAAAG	A-G	WA	None	TRUE	coding	No_motif
84	K1B	S144	144.31	L	3	0	3	FALSE	GGTATTGA	V	FR3	272	166	A	A	C		45	21		TRUE	1.657624	XGT	S	R		missense	AAAGT	AAAGT	A-C	WA	RCG	TRUE	coding	TRUE
85	K1B	S144	144.32	L	4	0	4	FALSE	GTAGGCCT	V	CDR3	376	270	G	G	A		12	10		TRUE	2.623408	CAX	Q	Q		silent	CAGCA	CAGCA	G-A	WRCY	WA	TRUE	non-coding	TRUE
86	K1B	S144	144.32	L	4	0	4	FALSE	GTAGGCCT	V	FR2	223	117	A	A	C		20	17		TRUE	2.662577	AAX	K	N		missense	AAACC	AAACC	A-C	WA	WRCY	TRUE	coding	TRUE
87	K1B	S144	144.32	L	4	0	4	FALSE	GTAGGCCT	V	CDR2	255	149	C	C	T		20	17		TRUE	2.662577	GXT	A	V		missense	TGCTG	TGCTG	C-T	WRCY	WA	TRUE	coding	TRUE
88	K1B	S144	144.32	L	4	0	4	FALSE	GTAGGCCT	V	FR3	277	171	G	G	T		20	16		TRUE	2.543991	GGX	G	G		silent	GGGGT	GGGGT	G-T	None	None	No_motif	No-motif	No_motif
89	K1B	S144	144.33	L	2	0	2	FALSE	GTCGTAAT	V	FR3	277	171	G	G	C		21	19		TRUE	2.788833	GGX	G	G		silent	GGGGT	GGGGT	G-C	None	RCG	No_motif	No-motif	TRUE
90	K1B	S144	144.33	L	2	0	2	FALSE	GTCGTAAT	V	FR3	343	237	G	G	A		18	14		TRUE	2.490226	CAX	Q	Q		silent	CAGCC	CAGCC	G-A	None	WA	No_motif	No-motif	TRUE
91	K1B	S144	144.34	L	2	0	2	FALSE	TCTTTCCGT	V	FR1	124	18	G	G	A		18	27		TRUE	3.965785	CAX	Q	Q		silent	CAGTC	CAGTC	G-A	None	WA	No_motif	No-motif	TRUE
92	K1B	S144	144.34	L	2	0	2	FALSE	TCTTTCCGT	V	FR3	342	236	A	A	T		17	22		TRUE	3.593819	CXG	Q	L		missense	GCAGC	GCAGC	A-T	None	None	No_motif	No-motif	No_motif
93	K1B	S144	144.35	L	1	0	1	FALSE	TGCCAAAA	V	FR3	276	170	G	G	C		10	10		TRUE	3.000001	GXG	G	A		missense	TGGGG	TGGGG	G-C	None	RCG	No_motif	No-motif	TRUE
94	K1B	S144	144.36	L	3	0	3	FALSE	TGCGTGGC	V	CDR3	372	266	T	T	C		17	23		TRUE	3.703397	CXA	L	P		missense	TCTAC	TCTAC	T-C	WA	None	TRUE	non-coding	No_motif
95	K1B	S144	144.36	L	3	0	3	FALSE	TGCGTGGC	V	CDR3	383	277	A	A	C		17	22		TRUE	3.593819	XCT	T	P		missense	ATACT	ATACT	A-C	WA	None	TRUE	coding	No_motif
96	K1B	S144	144.36	L	3	0	3	FALSE	TGCGTGGC	J	CDR3	395	4	G	G	C		16	17		TRUE	3.133183	XCT	A	P		missense	TCGCT	TCGCT	G-C	RCG	None	TRUE	non-coding	No_motif
97	K1B	S144	144.37	L	6	0	6	FALSE	TGCTGCTG	V	CDR3	360	266	T	T	C		16	13		TRUE	2.573944	CXA	L	P		missense	TCTAC	TCTAC	T-C	WA	None	TRUE	non-coding	No_motif
98	K1B	S144	144.37	L	6	0	6	FALSE	TGCTGCTG	V	CDR3	371	277	A	A	C		14	12		TRUE	2.679255	XCT	T	P		missense	ATACT	ATACT	A-C	WA	None	TRUE	coding	No_motif
99	K1B	S144	144.37	L	6	0	6	FALSE	TGCTGCTG	J	FR4	408	29	A	A	G		14	10		TRUE	2.332823	GXG	E	G		missense	GGAGA	GGAGA	A-G	None	None	No_motif	No-motif	No_motif
100	K1B	S144	144.37	L	6	0	6	FALSE	TGCTGCTG	V	FR1	141	47	G	A	G		23	17		TRUE	2.395099	GXA	E	G		missense	AGGAG	AGAAG	A-G	None	None	No_motif	No-motif	No_motif
101	K1B	S144	144.37	L	6	0	6	FALSE	TGCTGCTG	V	FR2	216	122	G	G	A		23	17		TRUE	2.395099	GXG	G	E		missense	AGGGA	AGGGA	G-A	None	None	No_motif	No-motif	No_motif
102	K1B	S144	144.37	L	6	0	6	FALSE	TGCTGCTG	V	FR3	267	173	T	T	C		23	17		TRUE	2.395099	GXC	V	A		missense	GGTCC	GGTCC	T-C	None	None	No_motif	No-motif	No_motif
103	K1B	S144	144.38	L	6	0	6	FALSE	TGTGGTAA	V	FR1	133	39	A	T	A		15	11		TRUE	2.380648	GCX	A	A		silent	GCATC	GCTTC	T-A	WA	None	TRUE	non-coding	No_motif
104	K1B	S144	144.38	L	6	0	6	FALSE	TGTGGTAA	V	FR1	139	45	A	A	C		15	11		TRUE	2.380648	GTX	V	V		silent	GTAGG	GTAGG	A-C	WA	None	TRUE	coding	No_motif
105	K1B	S144	144.38	L	6	0	6	FALSE	TGTGGTAA	V	CDR1	177	83	G	G	C		15	11		TRUE	2.380648	GXC	G	A		missense	GGGCA	GGGCA	G-C	WRCY	None	TRUE	non-coding	No_motif
106	K1B	S144	144.38	L	6	0	6	FALSE	TGTGGTAA	V	FR2	228	134	A	G	A		15	11		TRUE	2.380648	AXG	R	K		missense	TAAGC	TAGGC	G-A	None	WA	No_motif	No-motif	TRUE
107	K1B	S144	144.38	L	6	0	6	FALSE	TGTGGTAA	V	FR3	271	177	A	A	G		15	11		TRUE	2.380648	CCX	P	P		silent	CCATC	CCATC	A-G	None	RCG	No_motif	No-motif	TRUE
108	K1B	S144	144.38	L	6	0	6	FALSE	TGTGGTAA	V	FR3	304	210	A	A	G		15	11		TRUE	2.380648	GAX	E	E		silent	GAATT	GAATT	A-G	WA	WRCY	TRUE	coding	TRUE
109	K1B	S144	144.39	L	1	0	1	FALSE	TTAGGCAG	V	CDR2	257	151	G	G	A		10	16		TRUE	4.135536	XCA	A	T		missense	CTGCA	CTGCA	G-A	None	WA	No_motif	No-motif	TRUE
110	K1B	S144	144.40	L	7	0	7	FALSE	TTATGCTCA	V	CDR3	376	270	G	G	A		24	17		TRUE	2.317769	CAX	Q	Q		silent	CAGCA	CAGCA	G-A	WRCY	WA	TRUE	non-coding	TRUE
111	K1B	S144	144.40	L	7	0	7	FALSE	TTATGCTCA	J	FR4	426	35	A	A	G		20	18		TRUE	2.777999	AXA	K	R		missense	CAAAC	CAAAC	A-G	WA	None	TRUE	coding	No_motif
112	K1B	S144	144.40	L	7	0	7	FALSE	TTATGCTCA	J	FR4	418	27	G		A	T		11	13	TRUE		GTX		V	V		GTGGA			Not_evalua	Not_evalua	No_evaluat	Not_evalua	No_evaluated
113	K1B	S144	144.40	L	7	0	7	FALSE	TTATGCTCA	V	FR1	152	46	G		G	C		34	25	TRUE		XGA		G	R		TAGGA			Not_evalua	Not_evalua	No_evaluat	Not_evalua	No_evaluated
114	K1B	S144	144.40	L	7	0	7	FALSE	TTATGCTCA	V	FR1	179	73	G		T	G		34	25	TRUE		XCA		S	A		GGGCA			Not_evalua	Not_evalua	No_evaluat	Not_evalua	No_evaluated

Table S2. Ocurrent somatic hypermutation in follicular lymphoma.

. .

				IG HEAVY CHAIN ESTIMATES											
Patient	Diagnosis	Isotype	Cell numbers input	median established IGHV mutations per cell	min	max	Cell output HC SHM analysis	Cells with IGHV neovariants	Total IGHV neovariants	Mean IGHV neovariants per SHM event per cell	months for all cells to undergo 1 IGHV SHM event	SHM events to reach median established IGHV mutations	Months to acquire established IGHV mutations	Years to acquire established IGHV SHM
S10000	FL	IgM	12000	30	22	57	8387	45	105	2.33	1.55	12.86	19.97	1.66
S10971	FL	IgM	2000	47	45	50	99							
S11770	FL	IgG	2000	8	7	10	35	0	0					
S12118	FL	IgG	2000	39	35	42	710							
S12282	FL	IgM	2000	23	20	29	214							
S12390	FL	IgM	2000	25	22	29	1372							
S12500	FL	IgM	2000	51	45	56	1444	1	1	1.00	12.03	51.00	613.70	51.14
	FL	IgM	2000	52	46	60	919							
S13530	FL	IgG	12000	54	43	64	8337	51	159	3.12	1.36	17.32	23.60	1.97
S13553	FL	IgG	12000	66	32	79	6563	0	0					
S144	FL	IgM	2000	18	18	24	1150	0	0					
S8382	FL	IgM	2000	35	29	38	895							
S8934	FL	IgG	12000	35	33	38	6399	3	3	1.00	17.78	35.00	622.13	51.84
	FL	IgG	12000	34	33	40	6134							
					Median		**1261.00**			*IGHV versus IGKV/IGLV*		*p=0.0004*		
					Minimum		35.00			*Wilcoxon matched-pair signed rank test*				
					Maximum		8387.00							

IG LIGHT CHAIN ESTIMATES											
Patient	median established IGKV IGLV mutations per cell	min	max	Cell output LC SHM analysis	Cells with IGKV/IGLV neovariants	Total IGKV/IGLV neovariants	Mean IGKV/IGLV neovariants per SHM event per cell	months for all cells to undergo 1 IGKV/IGLV SHM event	SHM events to reach median established IGKV/IGLV mutations	Months to acquire established IGKV/IGLV mutations	Years to acquire established IGKV/IGLV SHM
S10000	12	10	74	8917	76	187	2.46	0.98	4.88	4.77	0.40
S10971	39	38	41	1071							
S11770	25	23	30	202	1	1	1.00	1.68	25.00	42.08	3.51
S12118	29	26	34	713							
S12282	14	13	17	398							
S12390	27	26	31	1469							
S12500	69	65	76	1424	1	1	1.00	11.87	69.00	818.80	68.23
	67	65	85	947							
S13530	30	20	35	8396	46	137	2.98	1.52	10.07	15.32	1.28
S13553	31	7	35	8199	91	275	3.02	0.75	10.26	7.70	0.64
S144	7	4	11	1483	32	78	2.44	0.39	2.87	1.11	0.09
S8382	11	10	13	1176							
S8934	21	16	30	8664	0	0					
	18	16	25	7869							
		Median		**1446.50**							
		Minimum		202.00							
		Maximum		8917.00							

COMBINED ESTIMATES										
Patient	Median established IGV mutations per cell	Cell output	Cells with IGV neovariants	Total IGV neovariants	Mean IGV neovariants per SHM event per cell	months for all cells to undergo 1 SHM event	SHM events to reach median established IGV mutations	Months to acquire established IGV mutations	Years to acquire established IGV SHM
S10000	42	8233	127	403	3.17	0.54	13.24	7.15	0.60
S10971									
S11770									
S12118									
S12282									
S12390									
S12500	120	1382	1	2	2.00	11.52	60.00	691.00	57.58
S13530	84	8220	46	137	2.98	1.49	28.20	42.00	3.50
S13553									
S144									
S8382									
S8934									

Table S3. Immunoglobulin gene somatic hypermutation dynamics and clinical correlations.

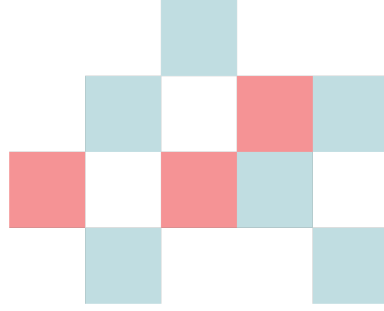
Lago Toro, Chile

CHAPTER 4

Integration of Mutational Signature Analysis with 3D Chromatin Data Unveils Differential AID-Related Mutagenesis in Indolent Lymphomas

Previously published as

Sepulveda-Yanez, J. H., Alvarez-Saravia, D., Fernandez-Goycoolea, J., Aldridge, J., van Bergen, C. A., Posthuma, W., Uribe-Paredes, R., Veelken, H. & Navarrete, M. A. Integration of Mutational Signature Analysis with 3D Chromatin Data Unveils Differential AID-Related Mutagenesis in Indolent Lymphomas. ***International Journal of Molecular Sciences* 22,** 13015. doi:10.3390/ijms222313015 (2021)



ABSTRACT

Activation-induced deaminase (AID) is required for somatic hypermutation in immunoglobulin genes, but also induces off-target mutations. Follicular lymphoma (FL) and chronic lymphocytic leukemia (CLL), the most frequent types of indolent B-cell tumors, are exposed to AID activity during lymphomagenesis. We designed a workflow integrating de novo mutational signatures extraction and fitting of COSMIC (Catalogue Of Somatic Mutations In Cancer) signatures, with tridimensional chromatin conformation data (Hi-C). We applied the workflow to exome sequencing data from lymphoma samples. In 33 FL and 30 CLL samples, 42% and 34% of the contextual mutations could be traced to a known AID motif. We demonstrate that both CLL and FL share mutational processes dominated by spontaneous deamination, failures in DNA repair, and AID activity. The processes had equiproportional distribution across active and nonactive chromatin compartments in CLL. In contrast, canonical AID activity and failures in DNA repair pathways in FL were significantly higher within the active chromatin compartment. Analysis of DNA repair genes revealed a higher prevalence of base excision repair gene mutations ($p = 0.02$) in FL than CLL. These data indicate that AID activity drives the genetic landscapes of FL and CLL. However, the final result of AID-induced mutagenesis differs between these lymphomas depending on chromatin compartmentalization and mutations in DNA repair pathways.

4.1 | INTRODUCTION

Cancer, that is, the uncontrolled proliferation of transformed cells, is generally attributable to acquired or inherited genetic variants that affect crucial cellular pathways [206]. Acquired mutations can be caused by environmental influences or stochastic errors in DNA replication [207, 208]. Particular mutagenic mechanisms generate distinguishable mutational signatures across a cancer cell's genome [209, 210].

Somatic hypermutation (SHM) represents an endogenous mutator mechanism in B lymphocytes. Physiologically, SHM targets immunoglobulin genes (IG) and is dependent on the expression of activation-induced deaminase (***AICDA***) in germinal center reactions. AID induces the deamination of deoxycytidine into deoxyuridine [211]. Subsequent activation of DNA repair mechanisms may result in faithful base repair but may alternatively lead to genetic variants, particularly through the engagement of the error-prone alternative base excision repair (BER) and mismatch repair (MMR) pathways [102, 212].

AID-induced SHM targets preferentially distinct sequence motifs: Canonical C>T/G transitions occur in WRCY motifs [101], and non-canonical A>C transversions in WA motifs [117]. A third pattern, it is characterized by C>T transitions in RCG motifs [213]. Aberrant AID activity on non-IG target genes has been implicated in the pathogenesis of various types of lymphoma [97, 214]. However, the genome-wide consequences of AID-associated mutational signatures have not yet been analyzed specifically per type of B-cell lymphoma.

In the genome-wide context, the three-dimensional (3D) chromatin structure could play an important role in the activity of the different mutational mechanisms. The introduction of High-throughput Chromosome Conformation Capture (Hi-C) [215] allows the identification of different states of the genome structure at the sub-chromosomal scale. Genomic regions can be assigned according to Hi-C to two compartments: the active compartment (compartment A) includes genomic regions characterized by transcription or epigenetics marks associated with open chromatin (H3K36me3), high density of genes, and DNase I hypersensitivity [215]. In contrast, the inactive compartment (compartment B) represents the condensed DNA regions [215, 216]. Hi-C maps from pro-B cells revealed that up to 96% of canonical AID target regions can be assigned to compartment A in the mouse genome [214].

Follicular lymphoma (FL) and chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL) are the most frequent types of indolent B-cell neoplasia [217]. Consistent with a malignancy of germinal center B-cells, FL cells constitutively express AID, and FL cells acquire very high levels of SHM in their IG genes. AID expression in FL can be correlated to both physiological and aberrant SHM [218, 219]. In contrast, CLL cells do not reside in germinal centers. Nevertheless, a subset of CLL has been exposed to AID activity as indicated by modest SHM of their immunoglobulin genes [220, 221]. Our aim was to identify AID-induced mutagenesis depending on chromatin compartmentalization in FL and CLL. We analyzed the contribution of AID to mutational signatures in whole-exome sequencing data of FL and CLL cases and the relationship between these signatures and the B-cell chromatin structure.

4.2 | MATERIALS AND METHODS

PATIENT CHARACTERISTICS AND SAMPLE ACQUISITION

Cryopreserved viable cells from 15 blood, bone marrow, and lymph node samples with infiltration by FL (***n*** = 9), CLL (***n*** = 3), or CLL-phenotype monoclonal B lymphocytosis (***n*** = 3) were obtained from the biobank of the LUMC Department of Hematology. This study was conducted in accordance with the Helsinki Declaration, all samples had written informed consent, and the study was performed with IRB approval of the local ethics committee (no. B16.039). Single cell suspensions were obtained by gentle mechanical disruption and mesh filtration and were cryopreserved using 10% DMSO as cryoprotectant. The remaining tissue was cultured in low-glucose DMEM to obtain stromal cell cultures for isolation of germinal DNA from nonmalignant cells. Thawed single cells were purified by flow cytometry using fluorescently labeled antibodies specific for CD19 and CD10 for FL, and CD19, CD5 and CD20dim for CLL and rested overnight followed by removal of dead cells using MACS dead cell removal kit. Additional whole-exome sequencing (WES) data were obtained for 24 FL and 24 CLL samples and their germ-line reference from European Genome-phenome Archive (https://ega-archive.org, last accessed on 2021-11-30) [222] as provided by Barts Cancer Institute, London (EGAD00001001301) [223], and from Institute Gustave Roussy, Villejuif (ERP003635) [224]. Patients' characteristics are shown in Supplementary Table S1.

LIBRARY PREPARATION AND SEQUENCING

Genomic DNA was isolated using QIAamp DNA Mini kit (Qiagen, Hilden, Germany). Samples were sequenced (HiSeq 4000 instrument, Illumina, San Diego, USA) in paired-end mode on Illumina (2 × 101 bp) using TrueSeq DNA exome kit (v.6) (Illumina, San Diego, USA). Mean coverage for every tumor and germline sample is depicted in Supplementary Table S2.

SEQUENCE ALIGNMENT AND VARIANT CALLING

Paired-end reads were aligned to the human reference genome sequence GRCh38 using BWA–MEM (V0.715-r1140) [225]. Alignment metrics and insert size distribution were gathered specifically through the CollectAlignmentSummaryMetrics and CollectInsertSizeMetrics tools from Picard (v2.12.1) [226]. Duplicate fragments were marked and removed using Picard (v2.12.1) tool MarkDuplicates. Local realignment was performed around indels to improve SNP calling in these conflictive areas with IndelRealigner tool. To avoid recalibration biases that might affect samples independently of each other, base quality scores were recalibrated using the BaseRecalibrator tool, with standard hard filtering parameters or Variant Quality Score Recalibration (VQSR) according to Genome Analysis Toolkit (GATK) [226] Best Practices.

Variant calling was performed on mpileup output files using Varscan (V2.3.9) [227] to WES data from tumor and patient-matched normal samples with a minimal variant frequency of 0.2, a somatic ***p***-value of 0.05, and minimum coverage of 10×. Filtered variants were then annotated applying Annovar (v.2016Feb01) [228] based on versions of the 1000 Genomes

Project (2015 Aug) [229], the Exome Aggregation Consortium (ExAC) [230], and predictions of functional importance from SIFT (Sorting Intolerant From Tolerant) [231] and by applying LRT (Likelihood Ratio Test) [232]. Variants were filtered and associated with DNA repair pathways by the genes related to BER, MMR, Fanconi anemia (FA), and DNA damage response (DDR) including their known variants. Quality control metrics were assessed using FastQC (v0.11.2) [233] before and after the alignment workflow and were reviewed to identify potential low-quality data files (Figure 2.1).

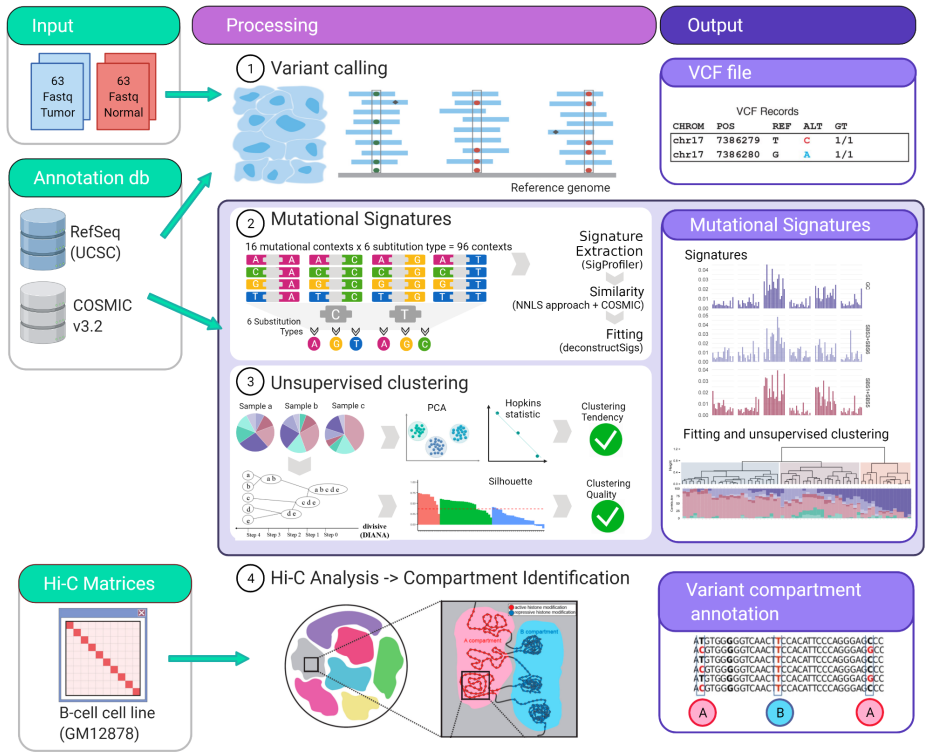


Figure 4.1. Schematic pipeline workflow. The input data are indicated in green boxes. The middle panel show the different analysis performed and the outputs are indicated in purple boxes. Figure created with BioRender.com.

MUTATIONAL SIGNATURES ANALYSIS

1. Variant classification in 96 contexts: Every single nucleotide variant (SNV) was represented for example, as C>A, C>T (by convention beginning with the pyrimidine) within a trinucleotide context (e.g., "GCT>GAT"). This approach yields 96 different possible contexts (six substitution type C>A, C>G, C>T, T>A, T>G, T>C with four possible bases immediately 5' and 3' to each substitution). This classification annotation was performed

for mutations within localized regions (IG loci) as well as for genome-wide mutations (WES) (Figure 4.1).

2. Signature extraction, similarity and fitting: Global and localized mutational signatures were defined by a workflow encompassing a three-step procedure, starting with a de novo signature extraction, followed by an similarity analysis, to allow a final fitting approach. de novo mutational signature extraction was generated by a non-negative matrix factorization (NMF) using R-package SigProfiler [234]. One of the practical drawbacks of the multiplicative NMF algorithm is that the task of selecting the appropriate number of sources is left to the user. Using this tool automatically allows the identification of the optimal number of operative signatures in our dataset (Figure 4.1, Supplementary Figure S1).

The similarity between de novo extracted signatures and the consensus Single Base Substitution (SBS) signatures deposited in the COSMIC catalog (v3.2–March 2021) [137, 235] was measured by non-negative least squares (NNLS) from SciPy python library (Supplementary Tables S3–S5) [236]. Finally, signature fitting was performed to define the contribution of the matching signatures to every sample by deconstructSigs [237] Figure 4.1.

UNSUPERVISED CLUSTERING ANALYSIS

The relative contribution of the identified de novo mutational signature and signatures from COSMIC identified in the fitting approach (SBS1, SBS3, SBS5, SBS6, SBS9, SBS84) were assembled for all FL and CLL/MBL samples in a matrix and tested for clustering tendency by two methods. The first method was a principal component analysis and the second method was a Hopkins' test to test the spatial randomness of the data (H = 0.28) [238]. Subsequently, all samples of the matrix were analyzed by divisive hierarchical clustering (DIANA) [239]. Strengths of clustering and estimation of average distances between clusters were analyzed by silhouette analysis (average silhouette width: 0.35, Supplementary Figure S2) [240] (Figure 4.1).

HI-C DATA ANALYSIS AND COMPARTMENT IDENTIFICATION

Compartment identification was performed using previously published Hi-C matrices [216] from lymphoblastoid B cell lines (GM12878) by the neural network based tool SNIPER [241], using the 10% of the data for training (Figure 4.1). The obtained compartments were used to annotate each variant using Vcfanno [242].

4.3 | RESULTS

VARIANT ALLELE DENSITIES AND FREQUENCIES IN FL AND CLL

WES data of 33 FL and 27 CLL and 3 MBL tumor samples and matched germline DNA from our own and previously reported [223, 224] samples were analyzed by Varscan to detect

somatic single nucleotide variations (SNVs) present in at least two reads in a tumor sample. For tumor samples, the average on-target rate was 84.1% (range: 82.3–86.1%), the average depth of coverage was ×133.5 (range: 119.7–147.4) with an average coverage rate over ×10 of 95.3% (range: 94.2–96.5%). For germline samples the average on-target rate was 82.7% (range: 80.6–84.8%), the average depth of coverage was ×93.8 (range: 78.0%–109.6%) with an average coverage rate over 10× of 92.9% (range: 90.7–95.1%) (Supplementary Table S2).

The application of a homogeneous streamlined pipeline for all samples provided us with the appropriate mutation metrics to perform a direct comparison. In FL cases, the median number of bona fide somatic mutations was 285 per tumor (range: 248.8–309.6), corresponding to a median mutation rate of 5.8 per Mb (range: 5.1–6.3). Since MBL was indistinguishable from CLL in all analyzed metrics, MBL and CLL cases are grouped under the same category. With a median of 179 mutations per tumor (range: 150.6–207.8) and a median mutation rate of 3.5 per Mb (range 3.1–4.2), CLL/MBL cases carried significantly fewer mutations than FL (Figure 4.2A), this is also observed in the CLL mutated and unmutated groups (Supplementary Figure S3A). In addition, the median variant allelic frequency of tumor samples was significantly higher in CLL/MBL as compared to FL (Figure 4.2B); this is preserved in the CLL mutated and unmutated groups (Supplementary Figure S3B).

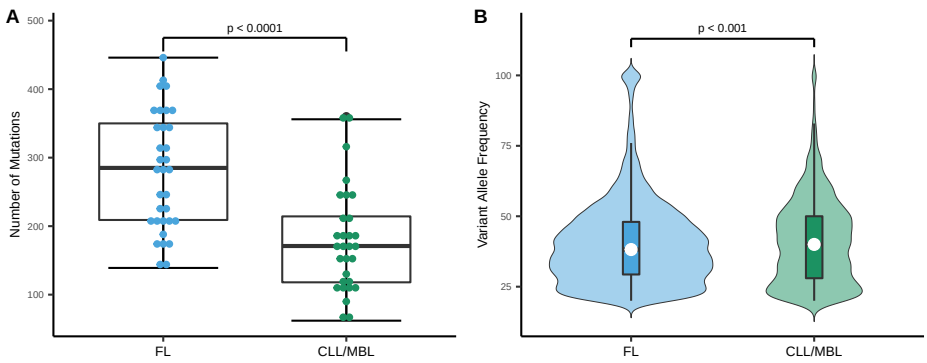


Figure 4.2. Mutational load and distribution of variant allele frequency in FL and CLL/MBL. (**A**) The mutation rate expressed as number of mutations per exome was higher in FL (***n*** = 33) than in CLL/MBL (***n*** = 30) (***t***-test, two-sided). (**B**) Violin plots depict the distinct distribution of variant allele frequency in FL and CLL/MBL. White circle: Median variant allele frequency; Colored bars: 25th and 75th percentiles; Whiskers: 5th and 95th percentiles (Wilcoxon test).

SOMATIC MUTATIONS IN CLL/MBL AND FL ARE FREQUENTLY ASSOCIATED WITH A DEAMINATION PATTERN IN AID MOTIFS

The overall pattern of nucleotide substitutions was virtually identical in FL and CLL/MBL (Figure 4.3A) and was dominated by transitions over transversions with a transition/transversion ratio of 1.6. C>T (G>A) transitions were overrepresented and comprised 37.6% of all mutations. Moreover, 35.6% of all C>T substitutions (13.4% of all somatic mutations) were observed

at CpG sites. Since methylated cytosine can easily undergo the transition to thymine [243], C>T transitions at CpG sites are generally accepted to occur through a deamination event. The second most frequent substitutions (23.7%) were T>C transitions, possibly originating from direct oxidation of thymine, or after replication of substrates containing 5-hydroxymethyl or 5-formyl-uracil [244]. Considering the observed dominance of deamination-related mutations, we then analyzed whether the mutations were associated with any of the reported AID motifs WRCY [101], WA [117] and RCG [213]. Indeed, large fractions of the mutations in both FL (42.7%) and CLL/MBL (33.6%) occurred at these AID motifs. The non-canonical WA motif alone contributed 24.7% of the mutations in FL and 18.1% in CLL/MBL. Mutations of the canonical WRCY contributed 8.8% of FL mutations and 6.7% of CLL/MBL mutations. Finally, 9.3% of mutations in FL and 8.8% in CLL/MBL involved the RCG motif. To define how frequently AID motifs would be affected by mutagenesis in human cancer without direct evidence for AID-induced mutagenesis, we analyzed a dataset of 264 samples of cutaneous melanoma and 791 samples of breast cancer The Cancer Genome atlas (TCGA) [245]. The three AID motifs combined contributed only 9.2% of the total number of mutations in cutaneous melanoma (SkCM) and 21% in breast cancer (BRCA) (FL vs. SkCM $p < 0.0001$, FL vs BRCA $p < 0.0001$, CLL/MBL vs. SkCM $p < 0.0001$, CLL/MBL vs SkCM $p < 0.0001$, Pearson's Chi-squared test) (Figure 4.3B). The contribution of mutations to these motifs in other cancer types listed in TCGA [245] is shown in Supplementary Table S6.

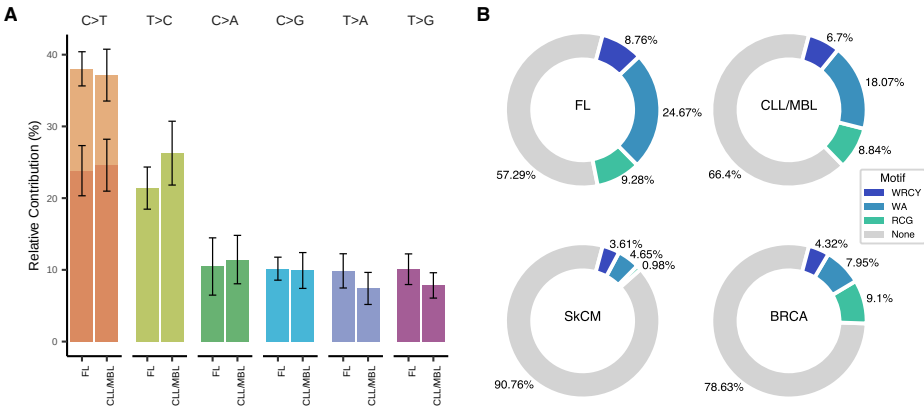


Figure 4.3. Type of substitutions and mutational patterns in FL and CLL/MBL. (**A**) The base substitution types across FL and CLL/MBL cases are dominated by transitions. (**B**) When tracing AID motifs in FL and CLL/MBL a high proportion of somatic mutations are allocated in such motifs, in strong contrast with skin cutaneous melanoma (SkCM) and Breast cancer (BRCA) that served as non-lymphoid malignancy references.

TRINUCLEOTIDE CONTEXT OF SOMATIC MUTATIONS IN FL AND CLL/MBL

All nucleotide substitutions were assigned to 96 substitution patterns according to their trinucleotide context. While FL and CLL/MBL had a similar pattern overall, FL showed a more

frequent contribution of mutation in contexts associated with the canonical (c-AID) and non-canonical (nc-AID) AID motifs than CLL/MBL (Figure 4.3B, FL: 33.43% and CLL: 24.77%). Since AID activity is expected to preferentially target immunoglobulin genes, we analyzed the mutational pattern restricted to these loci. Within these regions, the 96 substitutions pattern was closely related to the canonical AID signature (Supplementary Figure S4). This is supported by cosine similarity between the mutational pattern observed at the Ig loci and the signature SBS84 (FL: 0.79 and CLL/MBL: 0.78, Supplementary Figures S5A and S5B).

INFERRING THE ROLE OF AID AS AN UNDERLYING MUTAGENIC MECHANISM IN FL AND CLL/MBL

We performed a de novo signature extraction using the 96-trinucleotide context catalog [210]. Since the number of signatures that can be accurately extracted depends on the number of samples and the variance (number of mutations) of the dataset, we calculated the optimal number of signatures using a multiplicative NMF approach recently published [137].

In this unsupervised analysis, the mutational spectrum variance was explained by three signatures (Figure 4.4A,B, Supplementary Figure S1). The consistency and stability of the signature extraction was confirmed by repetitive extractions with bootstrapping strategies of sample subsets including categorization by neoplasm type and treatment status. Under the different tested condition, we obtained a consistent and reproducible signature extraction (Supplementary Figure S6). The global signature analysis was followed by a localized de novo signature extraction on the IG loci, where two signatures were identified (Supplementary Figure S7).

Next, we analyzed whether the extracted signatures corresponded to known mutational processes described in the COSMIC catalog (v3.2—March 2021) [137, 235]. The first signature (germinal center: GC) had a unique composition, the second signature could be attributed to the combination of mutational processes SBS1 and SBS5 (cosine similarity SBS1+SBS5 = 0.91), and the last signature was related to the combination of processes SBS3 and SBS6 (cosine similarity SBS3 + SBS6 = 0.85, Supplementary Table S4). As expected for the IG localized de novo signatures, the signatures represented the combination of the recently described signatures SBS84/SBS85 and also SBS37 [137] (Supplementary Figure S7).

To identify a per sample signature contribution, we used the signature fitting approach [246], incorporating biologically relevant signatures from the COSMIC catalogue (SBS1, SBS3, SBS5, SBS6, SBS9, SBS84) as well as the new de novo signature (GC). We also performed divisive hierarchical clustering (DIANA) for unsupervised sample classification, using seven mutational signature contributions, this analysis yielded three clusters (Figure 4.4C). FL cases were present mainly in two clusters (II and III in Figure 4.4C). Cluster two was dominated by SBS3 and SBS5 (mean contribution of SBS3 + SBS5: 61.0%). In contrast, the new GC signature contributed prominently to cluster three (mean contribution of GC signature: 63.0%). CLL cases were predominately allocated to cluster one with a dominance of SBS5 (mean contribution SBS5 in cluster 1 is 65.6%). Despite the dominance of AID-related mutations in the genomic landscapes of both FL and CLL/MBL, these data suggest that intrinsic differences in mutational

signatures exist between these types of B-cell lymphoma.

FL AND CLL SHOW DIFFERENTIAL DISTRIBUTION OF MUTATIONAL SIGNATURES ACROSS TRIDIMENSIONAL (3D) DNA COMPARTMENTS

To evaluate whether the previously identified signatures have a differential contribution at the chromatin level, we used Hi-C data from B cells [216] to allocate variants to the active nuclear compartment A and the inactive compartment B. We observed a stable distribution of the signatures across the nuclear compartments in CLL, and that distribution was independent of their mutational status (Figure 4.5A, ns). However, in FL cases, signatures associated with the DNA mismatch repair process (SBS6) were dominant in the active compartment A, and the signature related to spontaneous deamination (SBS1) in the inactive compartment (Figure 4.5A, Wilcoxon test with Bonferroni correction: 0.0003 and 0.0002 respectively). The relative contribution of SBS84 associated with mutations in the canonical AID motif (RCY, Figure 4.4B) was significantly higher in the active compartment (Figure 4.5A, Wilcoxon test with Bonferroni correction: 0.0009), whereas mutations associated with the non-canonical AID motif (WA) defined by SBS9 (Figure 4.5B) showed a significantly higher contribution within the inactive compartment (Figure 4.5A, Wilcoxon test with Bonferroni correction: 0.0024).

ANALYSIS OF MUTATIONS IN GENES INVOLVED IN DNA REPAIR

Whereas AID initiates mutagenic events by creating U:G mismatches, DNA repair mechanisms subsequently execute definite DNA alterations. Therefore, we searched for evidence of alterations in DNA repair pathways that could explain the observed difference in the substitution patterns and signatures compartment between FL and CLL/MBL. The list of genes involved in the DNA repair pathways was defined by the KEGG database and literature [247]. We found novel mutations (i.e., variants not present in germ-line or a preceding biopsy) in one or more DNA repair pathways in 23 of 33 FL samples (69.7%) and in 14 of 30 CLL/MBL samples (46.7%) (Figure 4.6A). In FL, the most frequently mutated DNA repair pathway (21 variants in 12 cases) was Fanconi anemia (FA), in particular in the gene ***FANCD2*** with even several mutations in the same patient (Supplementary Figure S8). The second frequently affected pathway in FL was BER (11 variants in 10 samples) with mutations in its main components, ***POLE*** and ***UNG***. With a single variant found in the CLL/MBL cases, the prevalence of BER mutations in FL was significantly higher than in CLL (Fisher´s exact test with Benjamini-Hochberg correction: p = 0.026; Figure 4.6B, Supplementary Figure S8). In CLL, the most frequently mutated pathway was the DNA damage response (DDR), with a total of 11 variants in 10 samples but without significant difference as compared with FL (Figure 4.6C). The main DDR genes mutated in CLL were ***TP53, ATM, and HUS1*** (Supplementary Figure S8).

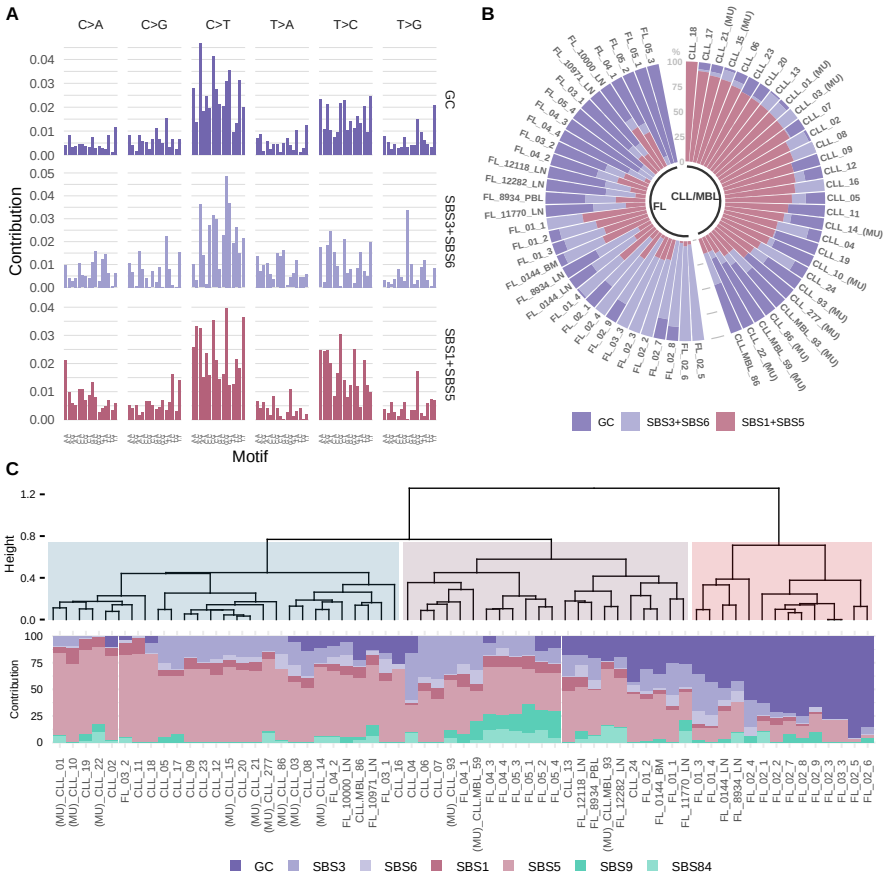


Figure 4.4. Mutational signatures analysis reveals a stable contribution of signature GC in FL. (**A**) The dataset was explained by 3 mutational signatures, a newly emerged de novo signature GC (GC) and two combinations of signatures derived from COSMIC, Single base substitution signature 3 and Single base substitution signature 6 (SBS3 + SBS6) and Single base substitution signature 1 and Single base substitution signature 5 (SBS1 + SBS5). (**B**) The prevalence of these signatures in individual tumor samples is depicted in each bar and represents an individual exome. (**C**) Fitting using de novo and COSMIC signatures. At the top an unsupervised divisive hierarchical clustering (DIANA) based on the matrix of signature contribution per sample, shows the classification of most CLL/MBL cases distant from FL. The samples with (MU) indicate CLL/MBL cases with mutated IGHV.

4.4 | DISCUSSION

Sequential acquisition of genomic alterations is considered to play a central role in oncogenesis and tumor progression. The underlying mechanisms of the mutagenic mechanisms continue to be intensively investigated and can be inferred by the analysis of mutational signatures. Here, we report the relative contribution of contextual somatic variants and mutational signatures in relationship with the genome chromatin structure for the two most prevalent

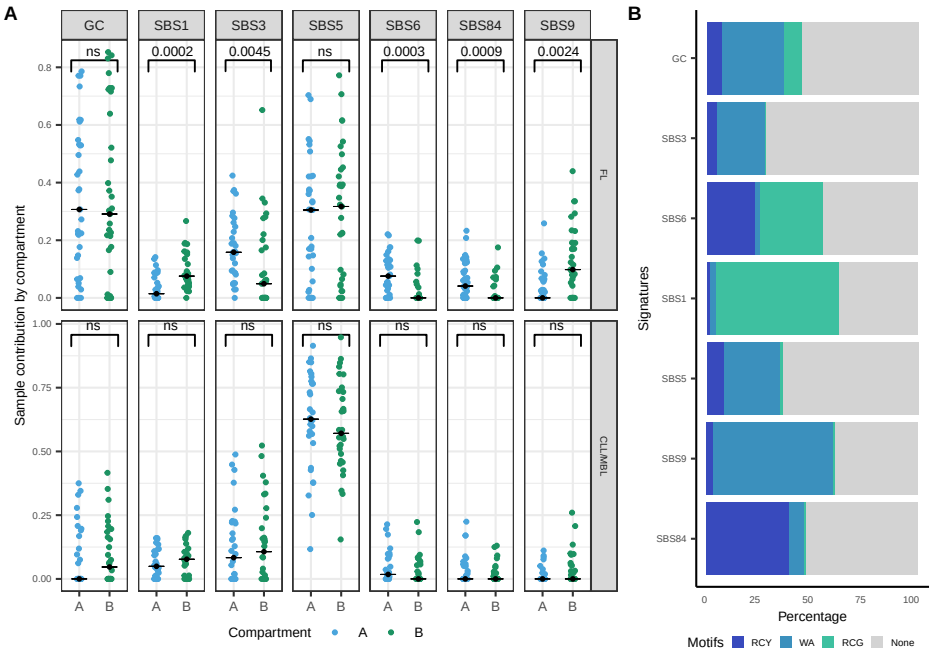


Figure 4.5. Contribution of mutational signatures on the three-dimensional (3D) chromatin structure. (**A**) Sample contribution by compartment A and B in FL and CLL/MBL cases (Wilcoxon test with Bonferroni correction), black dot: median. (**B**) Relative contribution of AID mtifs (trinucleotide) in signatures analyzed.

indolent mature B-cell neoplasms—FL and CLL.

While the genomic landscape of FL is characterized by a higher mutation density and mutation heterogeneity than CLL, the dominant genomic changes in both types of B-cell lymphoma are C>T transitions, indicating that deamination plays a central role. In both FL and CLL, an important number of somatic variants (42% and 33%, respectively) are located in DNA contexts related to AID motifs. This key observation is in remarkable contrast to the 9% and 21% observed in cutaneous melanoma and breast cancer and provides independent support for the postulated role of AID in FL and CLL development and progression [218, 220]. In FL, more mutations have been described in motifs recognized by AID and by APOBEC at relapse than that at diagnosis, and FL patients with accumulated mutations in AID-targeted genes are at high risk for transformation [248]. Besides, a recent study demonstrated that FL tumors harbor excess mutations in AID-motif overlapping the CpG methylation site [213]. Our data indicate that this finding may also extend to CLL.

Underlying mutational processes that shape cancer genomes have been identified by deconvolution of mutations classified by their context of neighboring 5' and 3' nucleotides. A set of consensus signatures has been linked to distinct mutagenic mechanisms such as aging, tobacco smoking, deficient mismatch repair, and UV light [137, 209, 249]. One of the main

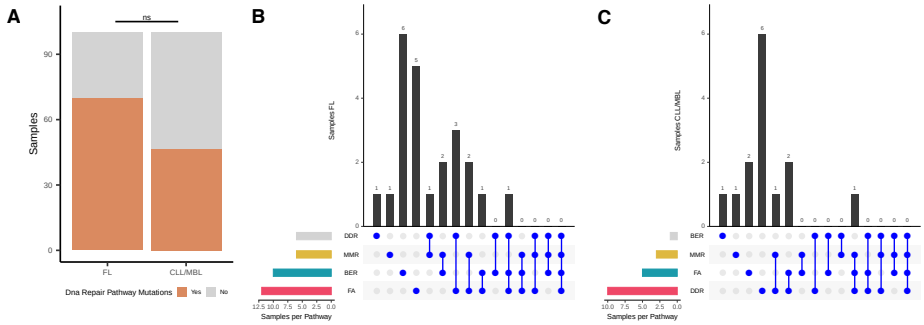


Figure 4.6. Analysis of mutations in DNA repair pathways. (**A**) Percentage of samples with at least one novel mutation (variants not present in germ-line or a preceding biopsy) in the four DNA repair pathways (Base Excision repair [BER], Mismatch Repair [MMR], Fanconi Anemia pathway [FA], and DNA Damage Response [DDR]) analyzed in FL and CLL/MBL. (**B**) Upset plot of FL samples, FA pathway was more affected in FL. (**C**) Upset plot of CLL/MBL samples, the pathway with most mutations was DDR with 6 cases.

advantages of deconvolution by NMF algorithms [250] is the lack of bias and independence on external information to extract mutational signatures. Our signatures are reliably related to FL and CLL as indicated by its consistent emergence with bootstrapping strategies. Signatures are usually complex, and they may result from a mix of different mechanisms [246]. In fact, one of the extracted signatures reflects a combination of defects in homologous recombination (SBS3 from COSMIC catalog) and DNA mismatch repair (SBS6 from COSMIC catalog) [137, 209]. The next extracted a signature correlated with the ubiquitous processes SBS1, which is associated with spontaneous deamination, and SBS5, which is present in almost any cancer type [137, 209, 251]. The third novel GC signature was characterized by a distinctive pattern with remarkable dominance of contextual somatic variants in AID motifs and was dissimilar to known COSMIC signatures (cosine similarity < 0.85).

The recently reported signatures, SBS84 and SBS85, which are related to canonical AID activity [137], were difficult to extract from global genomic data. However, these signatures can be readily detected by localized extraction in the IG loci (Supplementary Figure S7A and S5A). This apparent discrepancy indicates that the identification of certain signatures, despite their undisputed presence in the dataset, remains challenging by applying only the de novo signature extraction method and requires the addition of a fitting approach. Of note, the signature SBS9, described as related to non-canonical AID activity [141], did not emerge within global or localized extraction.

A signature analysis performed on the basis of the relative contribution of all signatures discussed above, that is, GC, SBS1, SBS3, SBS5, SBS6, SBS9, SBS84 across all analyzed FL and CLL cases, yielded three clusters that were clearly related to both lymphoma types. The cluster goodness is presented in Supplementary Figure S2. In the first cluster, whose genomic landscape was dominated by a so-called flat signature (SBS5), 81.5% of the cases (22/27) were CLL. Cases in the second cluster showed a higher proportion of AID-related signatures

(SBS9 and SBS84), and 63.6% (13/22) were FL cases. The third cluster, strongly dominated by the GC signature, had only FL cases (14 cases, 100%). These data indicate that FL and CLL share common endogenous mutagenic processes during lymphomagenesis, but additional mechanisms influence the activity and downstream consequences of AID in a lymphoma subtype-specific manner.

Since enzymatic deamination is more likely to occur in super-enhancer domains [214, 252] and since up to 96% of AID targets may be restricted to the active chromatin compartment [214], we also analyzed the contribution of particular signatures across genome compartments by integrating conformational information obtained from Hi-C data [216]. While no significantly differential contribution of signatures between compartments was observed in CLL, the canonical AID signature (SBS84) and the mismatch repair-related signature (SBS6) provided significantly higher contributions to the genomic landscape of the active compartment of FL cases. These data globally quantify the direct mutagenic effect of ongoing AID activity in FL as corroborated by the ongoing, constitutive AID overexpression in FL as opposed to CLL [218, 219, 253].

The Pol-η related signature (SBS9), that has been linked to non-canonical AID activity and CLL pathogenesis [141], as well as the spontaneous deamination-related signature SBS1 [137], were higher within the inactive compartment of FL. This finding may be indicative of the broader effects of these mutagenic mechanisms or may be reflecting mutations acquired through earlier events occurring in a mutation-prone environment such as the germinal center, and needs further investigation. Because mutagenesis and chromatin conformation are dynamic processes, future systematic analyses over time may be desirable to refine the results of this study. Nevertheless, the folding patterns of chromatin compartment domains are highly conserved within B cells, and even during malignant B-cell transformation, gene switching from the active to the inactive compartment was only 3.1% [254]. Whether specific chromatin organizations affecting CLL and FL may alter our findings, remains unanswered and will require technically challenging Hi-C experiments on human primary lymphoma cells. On the other hand, the limited number of available samples may restrict broader extrapolation of our results; however, the sample size remains similar to that of other studies addressing mutational signatures analysis [221, 246].

Although AID activity can initiate the mutagenic process, different DNA repair pathways define the outcome of the mutational cascade initiated by deamination. Mutations in these pathways can lead to an increased susceptibility to different cancers, such as diffuse large B-cell lymphoma, myeloid leukemia, breast cancer, and to cancer treatment-related toxicity [255–257]. We explored the novel hypothesis that acquired mutations in DNA repair pathways might contribute to differential mutational signatures according to lymphoma type. Indeed, a high prevalence of novel mutations was found in at least one pathway in both CLL and FL. Mutations predominantly affected FA and BER pathways in FL and DDR in CLL/MBL. A significantly higher incidence of BER pathway mutations occurred in FL and suggest the investigation of the association of DNA repair pathway mutations with AID-dependent mutational signatures in future studies.

4.5 | CONCLUSIONS

In summary, we defined the mutational processes that shape the mutational landscape of FL and CLL (global and localized) and integrated these signatures with sub-chromosomal conformation data. As indolent B-cell malignancies, CLL and FL share a common background of mutational processes. In CLL, mutational signatures are evenly distributed across chromatin compartments. In contrast, mutagenesis related to canonical AID activity and failures in DNA repair pathways in FL were more frequently found in the active chromatin compartment. The constitutive AID expression observed in FL and mutations in DNA repair pathways are candidate factors to explain these lymphoma-specific differences. Collectively, these new findings support a direct association between aberrant AID action and lymphomagenic mutations. Since certain mutators, such as ongoing endogenous deamination, are more prone to occurring in restricted areas of the tridimensional structure, the integration of genomic conformational data into signature analysis could help us to better understand the biological relevance of deconvoluted mutational processes.

4.6 | SUPPLEMENTARY MATERIAL

SUPPLEMENTARY FIGURES

Figure S1. Evaluation of the optimal signatures number.

Figure S2. Cluster silhouette analysis for divisive hierarchical clustering.

Figure S3. Mutational load and distribution of variant allele frequency in FL and CLL mutated (CLL M) and unmutated (CLL U).

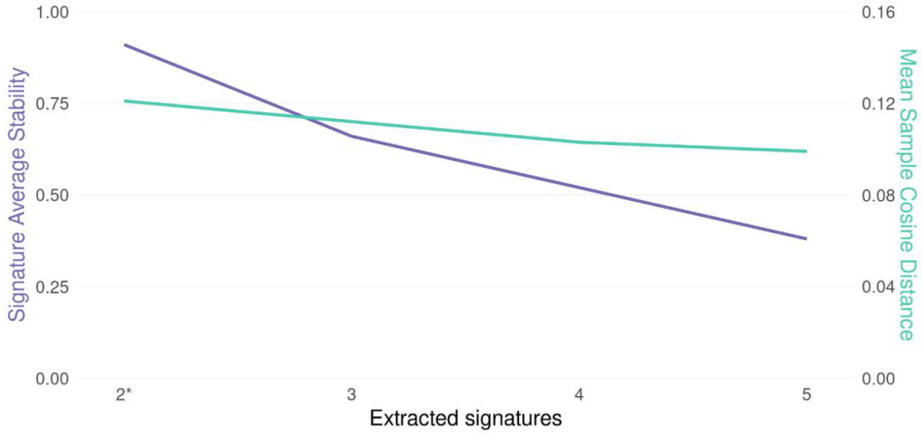
Figure S4. 96-trinucleotide mutational pattern of FL and CLL/MBL.

Figure S5. Characterization of immunoglobulin loci (IG) in FL and CLL samples.

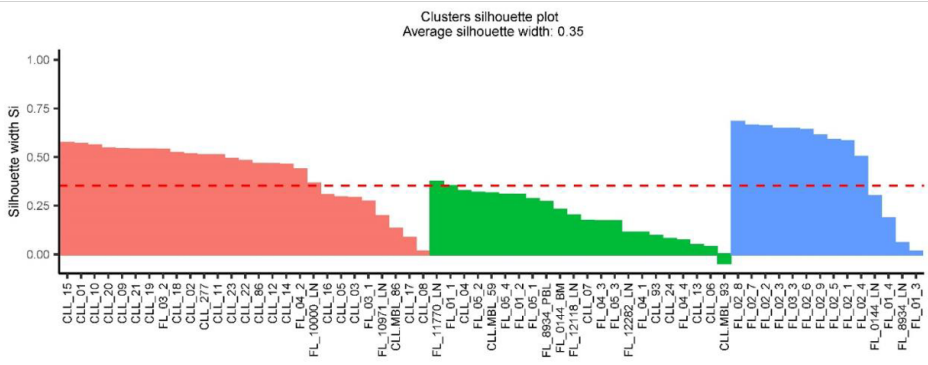
Figure S6. Consistency of signature extraction with sample bootstrapping.

Figure S7. De novo mutational signatures extraction in immunoglobulin genes loci (IG).

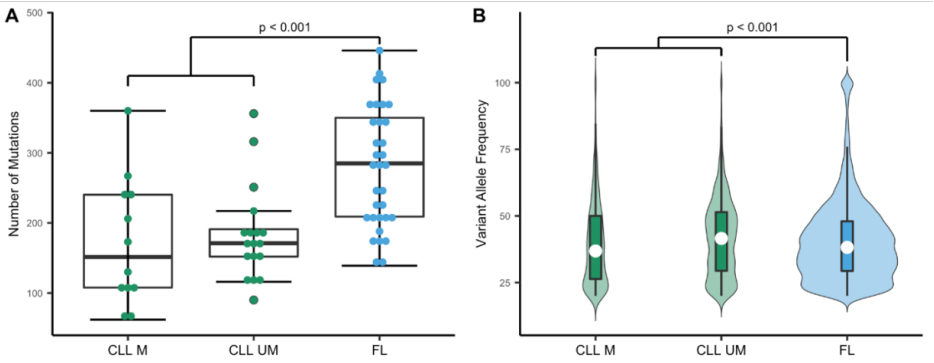
Figure S8. Heatmap of DNA repair genes mutated in FL and CLL/MBL.



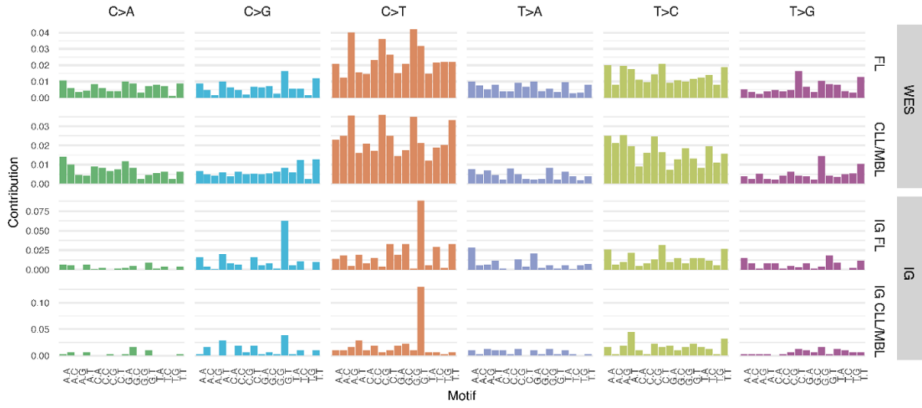
Supplementary Figure 1. Evaluation of the optimal signatures number. We calculated the residual signature average stability (average silhouette coefficient) and the mean sample cosine distance to define to optimal signatures number



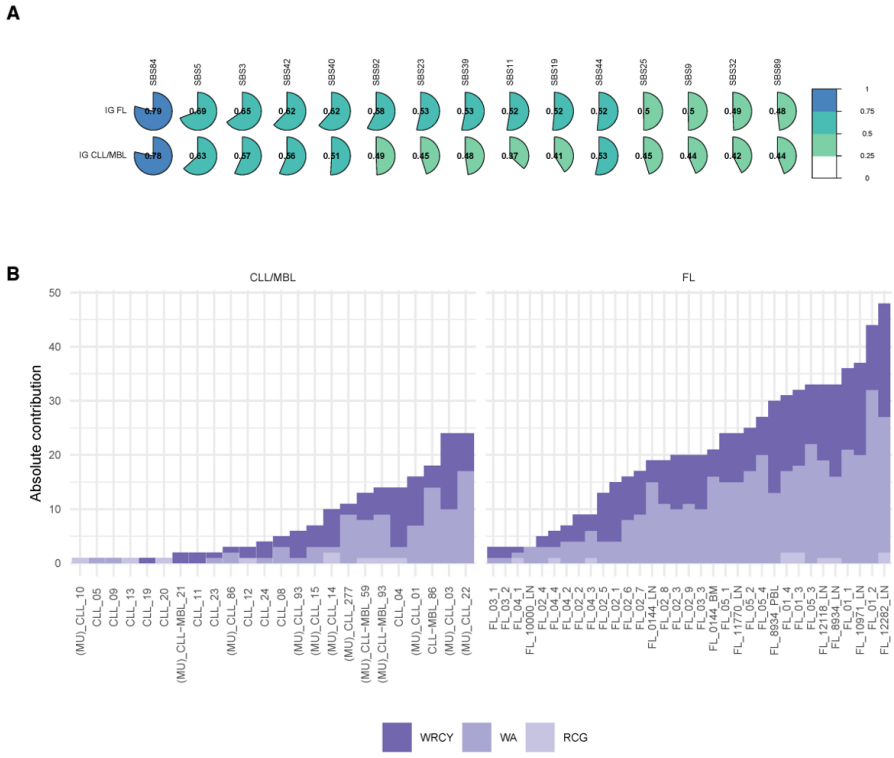
Supplementary Figure 2. Cluster silhouette analysis for divisive hierarchical clustering. Silhouette coefficient will range for each sample from -1 (if it is wrongly placed on the current cluster) to 1 (if it is correctly placed).) and the mean sample cosine distance to define to optimal signatures number



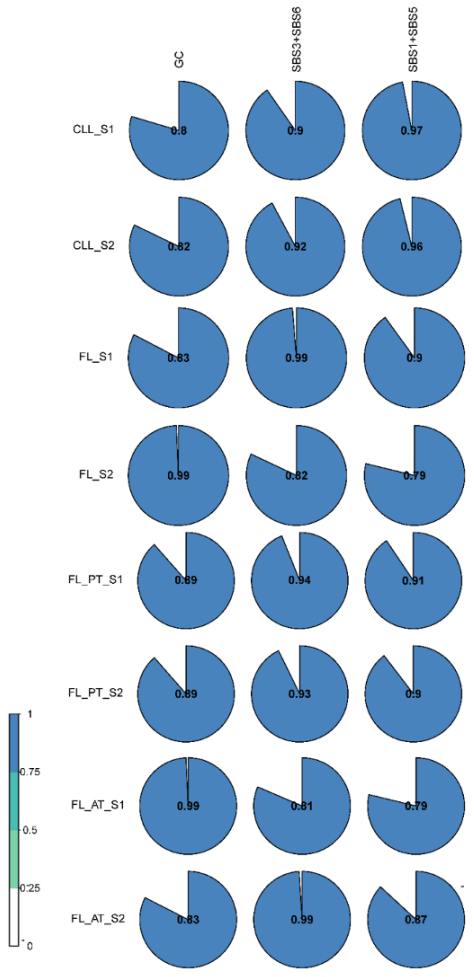
Supplementary Figure 3. Mutational load and distribution of variant allele frequency in FL and CLL mutated (CLL M) and unmutated (CLL U). A The mutation rate expressed as the number of mutations per exome was higher in FL (n=33) than in CLL M/CLL U(n=30) (t-test, two-sided). B Violin plots depict a distinct distribution of variant allele frequency in FL and CLL M/CLL UM. White circle: Median variant allele frequency; Colored bars: 25th and 75th percentiles; Whiskers: 5th and 95th percentiles (Wilcoxon test.)



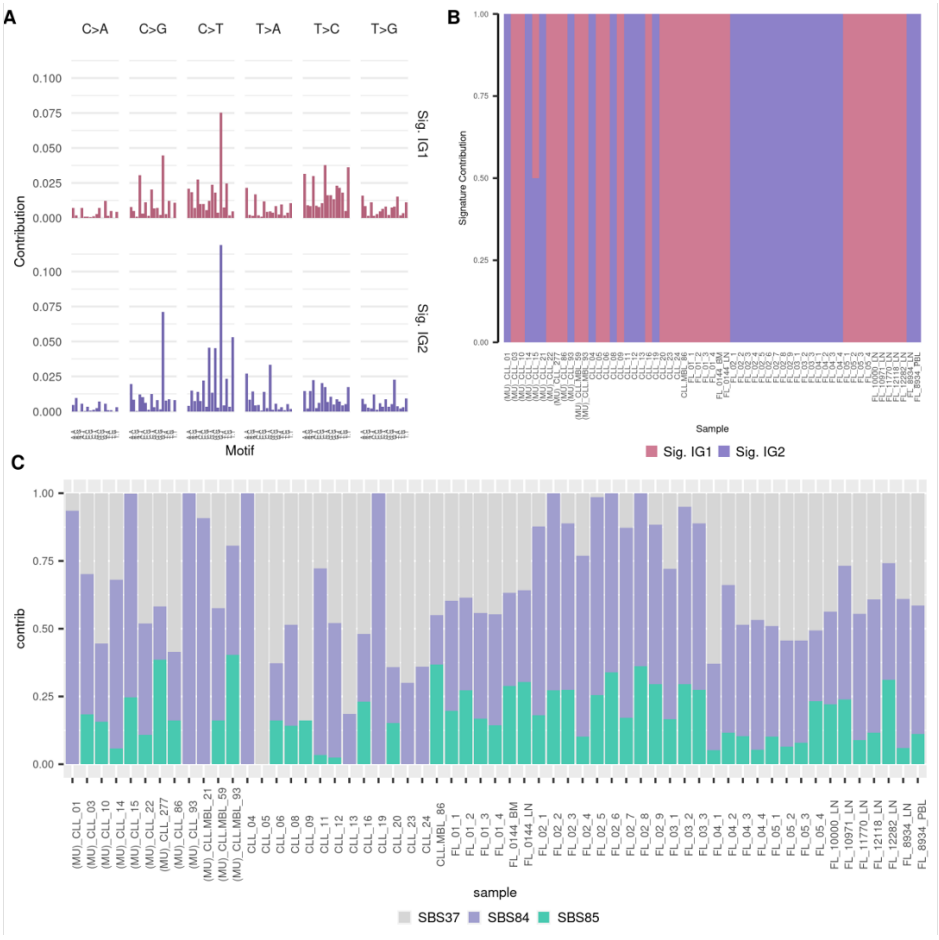
Supplementary Figure 4. 96-trinucleotide mutational pattern of FL and CLL/MBL divided on genome-wide (WES) and in immunoglobulin genes loci (IG).



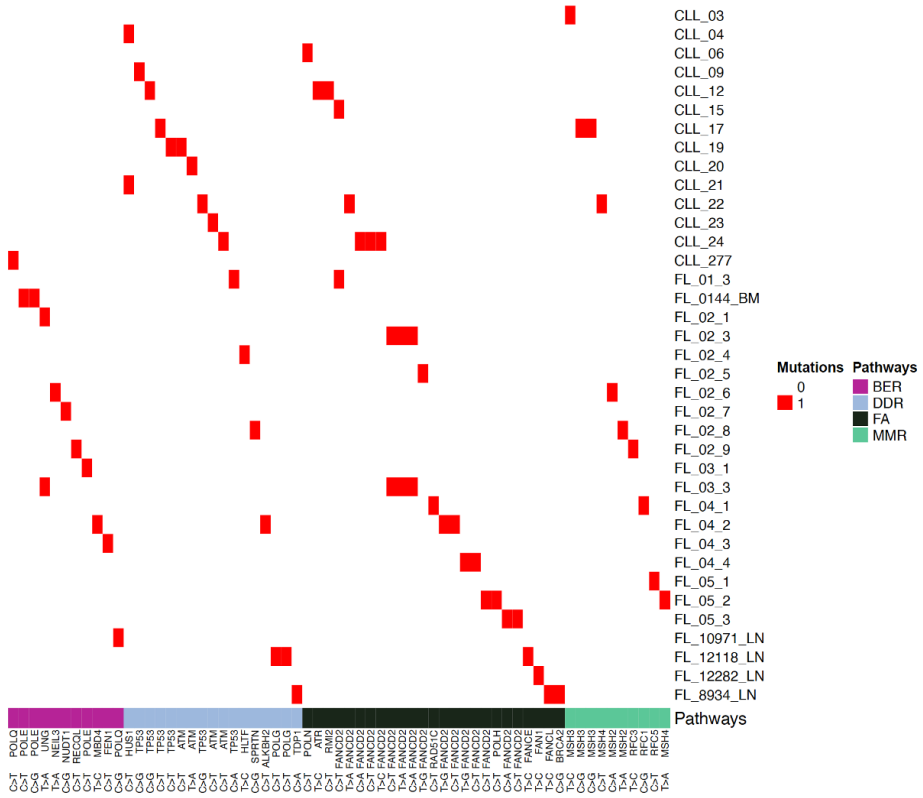
Supplementary Figure 5. Characterization of immunoglobulin loci (IG) in FL and CLL samples. A Cosine similarity measure between the mutational profile in immunoglobulin loci (IG) (showed in Fig. 4) and the Single Base Substitution (SBS) signatures from the COSMIC catalog (v3.2). It is shown the 15 most similar patterns in COSMIC. B AID motifs contribution within IG loci for each sample. (MU) indicates CLL/MBL cases with mutated IGHV.



Supplementary Figure 6. Consistency of signature extraction with sample bootstrapping. To evaluate the consistency of signature extraction, NMF was performed in different sample subsets. Extraction 1: Only CLL (chronic lymphocytic leukemia) samples. Extraction 2: Only FL (follicular lymphoma) samples. Extraction 3: FL pretreatment samples. Extraction 4: FL after treatment samples. Every signature extraction (sample subset) produced two signatures (_S1 and _S2) and it was compared independently against the complete (63 samples) de novo extraction with 3 signatures (GC, SBS3+SBS6, SBS1+SBS5). Signatures are compared using cosine similarity.



Supplementary Figure 7. De novo mutational signatures extraction in immunoglobulin gene loci (IG). A Two mutational signatures were obtained after de novo deconvolution, i.e. signature "immunoglobulin 1" (Sig. IG1), and signature "immunoglobulin 2" (Sig. IG2). B Prevalence of the de novo extracted mutational signatures to individual tumor samples. Each bar represents an individual exome. C Fitting using de novo and COSMIC signatures.



Supplementary Figure 8. Heatmap of DNA repair genes mutated in FL and CLL/MBL. The genes are divided by pathways, every column represents a defined mutation and substitution type.

SUPPLEMENTARY TABLES

Table S1. Clinical data

Table S2. Metrics WES data

Table S3. Ten top similarity between COSMIC signatures combinatory and novel signature A (WES).

Table S4. Ten top similarity between COSMIC signatures combinatory and novel signature B (WES).

Table S5. Ten top similarity between COSMIC signatures combinatory and novel signature C (WES).

Table S6. AID motifs contribution of mutations present in 6 cancer type on TCGA.

Sample ID	Age	Gender	Diagnosis	Stage	Exposed to Cytotoxic Therapy	Ig Isotype	Ig Mutation Status
CLL_01	61	female	CLL	B	N/A	N/A	mutated
CLL_02	54	male	CLL	A	N/A	N/A	unmutated
CLL_03	73	male	CLL	A	N/A	N/A	mutated
CLL_04	55	male	CLL	C	N/A	N/A	unmutated
CLL_05	56	male	CLL	B	N/A	N/A	unmutated
CLL_06	49	male	CLL	A	N/A	N/A	unmutated
CLL_07	58	male	CLL	B	N/A	N/A	unmutated
CLL_08	62	male	CLL	C	N/A	N/A	unmutated
CLL_09	88	male	CLL	A	N/A	N/A	unmutated
CLL_10	54	female	CLL	B	N/A	N/A	mutated
CLL_11	53	male	CLL	C	N/A	N/A	unmutated
CLL_12	67	male	CLL	Richter	N/A	N/A	unmutated
CLL_13	59	male	CLL	C	N/A	N/A	unmutated
CLL_14	64	male	CLL	B	N/A	N/A	mutated
CLL_15	53	female	CLL	B	N/A	N/A	mutated
CLL_16	41	female	CLL	C	N/A	N/A	unmutated
CLL_17	42	male	CLL	B	N/A	N/A	unmutated
CLL_18	59	male	CLL	B	N/A	N/A	unmutated
CLL_19	58	male	CLL	B	N/A	N/A	unmutated
CLL_20	50	male	CLL	B	N/A	N/A	unmutated
CLL_21	39	female	MBL	MBL	N/A	N/A	mutated
CLL_22	70	male	CLL	A	N/A	N/A	mutated
CLL_23	56	female	CLL	B	N/A	N/A	unmutated
CLL_24	42	male	CLL	B	N/A	N/A	unmutated
CLL_277	71	female	CLL	A	No	N/A	mutated
CLL_86	47	female	CLL	A	No	IgM	mutated
CLL_93	48	male	CLL	A	No	IgM	mutated
MBL_59	80	male	MBL	MBL	No	IgM	mutated
MBL_86	57	male	MBL	MBL	No	IgM	unmutated
MBL_93	50	female	MBL	MBL	No	IgM	mutated
FL_10000_LN	44	female	FL	IV	N/A	IgM	N/A
FL_10971_LN	56	male	FL	II	N/A	IgG	N/A
FL_11770_LN	59	male	FL	III	N/A	IgG	N/A
FL_0144_LN	37	male	FL	IV	N/A	IgM	N/A
FL_0144_BM	37	male	FL	IV	N/A	IgM	N/A
FL_8934_LN	66	male	FL	IV	N/A	IgG	N/A
FL_8934_PBL	66	male	FL	IV	N/A	IgG	N/A
FL_12282_LN	73	male	FL	IV	N/A	IgM	N/A
FL_12118_LN	51	female	FL	IV	N/A	IgG	N/A
FL_01_01	55	female	FL	IV	No	N/A	N/A
FL_01_02	55	female	FL	IV	Yes: Chlorambucil	N/A	N/A
FL_01_03	55	female	FL	IV	Yes: R-CHOP	N/A	N/A
FL_01_04	55	female	FL	IV	Yes: Bendamustine + GA101	N/A	N/A
FL_02_01	36	female	FL	IV	No	N/A	N/A
FL_02_02	36	female	FL	IV	Yes: Chl + Bexxar	N/A	N/A
FL_02_03	36	female	FL	IV	Yes: FMD then CHOP	N/A	N/A
FL_02_04	36	female	FL	IV	No	N/A	N/A
FL_02_05	36	female	FL	IV	Yes: Velcade	N/A	N/A
FL_02_06	36	female	FL	IV	No	N/A	N/A
FL_02_07	36	female	FL	IV	Yes: Radiotherapy	N/A	N/A
FL_02_08	36	female	FL	IV	Yes: Chl + Rituximab	N/A	N/A
FL_02_09	36	female	FL	IV	Yes: IFRT	N/A	N/A
FL_03_01	37	male	FL	Unk	No	N/A	N/A
FL_03_02	37	male	FL	Unk	Yes: Chlorambucil	N/A	N/A
FL_03_03	37	male	FL	Unk	No	N/A	N/A
FL_04_01	46	male	FL	IV	No	N/A	N/A
FL_04_02	46	male	FL	IV	Yes: Chlorambucil	N/A	N/A
FL_04_03	46	male	FL	IV	No	N/A	N/A
FL_04_04	46	male	FL	IV	Yes: Rituximab	N/A	N/A
FL_05_01	50	female	FL	IV	No	N/A	N/A
FL_05_02	50	female	FL	IV	Yes: BEAM-R	N/A	N/A
FL_05_03	50	female	FL	IV	Yes: Etoposide	N/A	N/A
FL_05_04	50	female	FL	IV	Yes: Velcade + Rituximab	N/A	N/A

Chl: chlorambucil; CHOP: cyclophosphamide, doxorubicin, vincristine, prednisolone; R: rituximab; GA101: obinutuzumab; FMD: fludarabine, etoposide, cytarabine, melphalan; RIC-SCT: reduced intensity conditioning stem cell transplantation; N/A: not available

Table S1. Clinical data.

Case	Type	Total Reads	Mean Coverage	On target Sequencing	On target 10x
FL_01_1	tumor	148366538	220.51	90.09	91.55
FL_01_2	tumor	116694500	154.75	91.28	97.34
FL_01_3	tumor	120151948	163.65	91.86	97.33
FL_01_4	tumor	114548762	158.43	92.17	97.39
FL_01_NO	normal	147351546	211.89	90.33	97.61
FL_02_1	tumor	109882280	144.58	91.02	97.27
FL_02_2	tumor	148524558	225.7	90.54	97.15
FL_02_3	tumor	118689014	158.23	88.99	94.47
FL_02_4	tumor	126138784	194.33	89.23	84.61
FL_02_5	tumor	118783748	173.97	90.38	96.92
FL_02_6	tumor	111401228	148.48	90.54	97.32
FL_02_7	tumor	109036698	150.66	91.8	97.14
FL_02_8	tumor	122191574	163.88	92.02	97.47
FL_02_9	tumor	130684024	176.18	90.82	97.41
FL_02_NO	normal	120578920	163.97	91.79	97.44
FL_03_1	tumor	141200834	195.96	92.41	97.68
FL_03_2	tumor	157118176	223.41	90.3	97.48
FL_03_3	tumor	107698786	163.95	96.65	84.34
FL_03_NO	normal	67184452	95.78	93	93.08
FL_04_1	tumor	118988300	164.38	91.92	97.47
FL_04_2	tumor	120262084	168.09	90.99	91.02
FL_04_3	tumor	119069140	163.2	92.53	97.48
FL_04_4	tumor	103812104	146.98	92.48	97.24
FL_04_NO	normal	113699268	182.2	96.51	97.09
FL_05_1	tumor	154134958	229.38	92.88	97.36
FL_05_2	tumor	124658200	180.59	92.65	97.19
FL_05_3	tumor	137118204	191.58	92.03	97.28
FL_05_4	tumor	99690824	119.69	80.51	95.28
FL_05_NO	normal	113518770	151.21	87.41	95.69
FL_0144_NO	normal	45467910	59.11	81.81	96.2
FL_0144_LN	tumor	130494860	136.72	70.64	98.16
FL_10000_NO	normal	49920404	65.44	81.45	96.16
FL_10000_LN	tumor	107285804	122.27	75.19	97.7
FL_10971_NO	normal	37779993	53.3	86.08	90.21
FL_10971_LN	tumor	133822353	132.93	67.03	98.13
FL_11770_NO	normal	45228495	60.24	81	94.35
FL_11770_LN	tumor	122110764	134.99	73.44	98.03
FL_0144_BM	tumor	101624412	113.22	77.52	97.5
FL_8934_NO	normal	33563972	45.84	83.14	90.2
FL_8934_LN	tumor	87562601	97.11	72.73	97.54
FL_8934_PBL	tumor	126074349	135.82	72.2	98.08
CLL_01	tumor	99975777	109.75	76.18	98.34
CLL_01_NO	normal	129779234	157.77	81.1	98.98
CLL_02	tumor	85229538	104.37	82.46	97.92
CLL_02_NO	normal	60710368	74.79	82.37	96.19
CLL_03	tumor	114674940	133.36	79.47	98.71
CLL_03_NO	normal	159444596	188.9	81.67	99.25
CLL_04	tumor	113351104	140.36	81.94	98.72
CLL_04_NO	normal	112766842	136.66	80.86	98.77
CLL_05	tumor	115682359	142.34	82.02	98.8
CLL_05_NO	normal	123990026	144.86	80.34	98.48
CLL_06	tumor	126154032	158.38	82.57	98.89
CLL_06_NO	normal	109898040	129.5	80.85	97.56

CLL_07	tumor	106298393	132.26	82.25	98.57
CLL_07_NO	normal	150116755	183.67	81.91	98.56
CLL_08	tumor	139094030	170.18	82.04	99.11
CLL_08_NO	normal	95331442	115.41	80.84	97.67
CLL_09	tumor	104454852	106.49	73.97	97.08
CLL_09_NO	normal	108021998	109.25	73.09	92.96
CLL_10	tumor	123664825	144.51	79.19	98.95
CLL_10_NO	normal	98180470	120.88	81.77	97.55
CLL_11	tumor	149233697	184.1	81.98	99.17
CLL_11_NO	normal	108345186	124.27	79.19	97.88
CLL_12	tumor	158116448	191.71	81.54	99.18
CLL_12_NO	normal	110093882	136.26	81.59	96.99
CLL_13	tumor	41849706	43.9	74.71	89.88
CLL_13_NO	normal	58636071	60.24	73.49	90.83
CLL_14	tumor	104612132	107.36	74.42	96.95
CLL_14_NO	normal	110232945	111.72	73.61	95.83
CLL_15	tumor	109906036	110.89	73.67	97.47
CLL_15_NO	normal	95832210	97.07	73.42	94.6
CLL_16	tumor	99172829	103.13	74.91	95.59
CLL_16_NO	normal	85287736	85.02	72.81	90.3
CLL_17	tumor	89581223	89.9	73.6	95.77
CLL_17_NO	normal	104387169	106.7	73.69	95.35
CLL_18	tumor	95132406	96.25	73.61	95.24
CLL_18_NO	normal	87603781	88.37	73.9	95.89
CLL_19	tumor	113777708	114.03	73.51	97.5
CLL_19_NO	normal	114298637	114.53	73.13	94.58
CLL_20	tumor	37647698	39.68	74.79	86.63
CLL_20_NO	normal	32770136	33.53	72.26	75.75
CLL_21	tumor	41585567	57.6	88.93	93.58
CLL_21_NO	normal	33680058	46.39	89.6	89.46
CLL_22	tumor	31795254	44.03	89.78	89.17
CLL_22_NO	normal	31290637	43.93	89.87	86.51
CLL_23	tumor	38790769	54.29	89.46	93.09
CLL_23_NO	normal	27253034	36.64	89.39	87.05
CLL_24	tumor	44981602	61.15	89.04	94.55
CLL_24_NO	normal	33057321	46.4	90.05	83.09
FL_12118_LN	tumor	239020845	231.43	78.16	97.45
FL_12118_NO	normal	55302794	53.05	77.83	95.12
FL_12282_LN	tumor	245363933	229.76	76.59	97.59
FL_12282_NO	normal	55601919	52.8	77.16	95.11
CLL_277	tumor	43392234	56.04	89.37	93.76
CLL_277_NO	normal	32037410	41.21	88.06	88.64
CLL_93	tumor	31590992	41.76	88.52	89.15
CLL_93_NO	normal	47606605	61.73	89.37	94.98
CLL_86	tumor	20495577	27.21	88.99	76.86
CLL_86_NO	normal	31632719	42.66	89.17	88.37
MBL_59	tumor	23177666	30.51	90.79	81.65
MBL_59_NO	normal	31016054	42.05	89.15	88.32
MBL_86	tumor	31290592	40.96	88.47	87.98
MBL_86_NO	normal	13226442	17.46	90.68	60.59
MBL_93	tumor	46616083	60.8	89.67	94.71
MBL_93_NO	normal	34519064	46.01	89.74	90.93

Table S2. Metrics WES data.

combination	similarity	proportions
SBS6+SBS40	0.84518779	SBS6=0.26;SBS40=0.74
SBS5+SBS15	0.842592001	SBS5=0.81;SBS15=0.19
SBS3+SBS6	0.838038742	SBS3=0.75;SBS6=0.25
SBS1+SBS5	0.837338189	SBS1=0.12;SBS5=0.88
SBS5+SBS6	0.834261008	SBS5=0.79;SBS6=0.21
SBS1+SBS40	0.829914249	SBS1=0.15;SBS40=0.85
SBS15+SBS40	0.823038869	SBS15=0.22;SBS40=0.78
SBS3+SBS15	0.81773249	SBS3=0.78;SBS15=0.22
SBS1+SBS3	0.817507345	SBS1=0.15;SBS3=0.85
SBS6+SBS92	0.803992034	SBS6=0.33;SBS92=0.67

Table S3. Ten top similarity between COSMIC signatures combinatory and novel signature A (WES).

combination	similarity	proportions
SBS1+SBS5	0.913996736	SBS1=0.13;SBS5=0.87
SBS5+SBS6	0.898515564	SBS5=0.8;SBS6=0.2
SBS5+SBS87	0.889060232	SBS5=0.8;SBS87=0.2
SBS5+SBS15	0.858133373	SBS5=0.89;SBS15=0.11
SBS3+SBS6	0.851716397	SBS3=0.72;SBS6=0.28
SBS6+SBS40	0.849351585	SBS6=0.29;SBS40=0.71
SBS6+SBS92	0.846014279	SBS6=0.33;SBS92=0.67
SBS5+SBS84	0.839546372	SBS5=0.9;SBS84=0.1
SBS1+SBS40	0.8388504	SBS1=0.17;SBS40=0.83
SBS5+SBS57	0.837864266	SBS5=0.89;SBS57=0.11

Table S4. Ten top similarity between COSMIC signatures combinatory and novel signature B (WES).

combination	similarity	proportions
SBS1+SBS5	0.907752268	SBS1=0.1;SBS5=0.9
SBS5+SBS6	0.894658673	SBS5=0.85;SBS6=0.15
SBS5+SBS87	0.889249105	SBS5=0.85;SBS87=0.15
SBS5+SBS15	0.88179594	SBS5=0.89;SBS15=0.11
SBS5+SBS50	0.862861657	SBS5=0.91;SBS50=0.09
SBS5+SBS51	0.862424347	SBS5=0.91;SBS51=0.09
SBS5+SBS32	0.86201106	SBS5=0.9;SBS32=0.1
SBS5+SBS94	0.860459467	SBS5=0.86;SBS94=0.14
SBS5+SBS43	0.860263858	SBS5=0.94;SBS43=0.06
SBS5+SBS58	0.859010168	SBS5=0.91;SBS58=0.09

Table S5. Ten top similarity between COSMIC signatures combinatory and novel signature C (WES).

TCGA cancer	Name	AID Motifs	Count	Percentage
hnsc_tcga	Head and Neck squamous cell carcinoma	None	51161	76.22
hnsc_tcga	Head and Neck squamous cell carcinoma	RCG	6674	9.94
hnsc_tcga	Head and Neck squamous cell carcinoma	WA	5942	8.85
hnsc_tcga	Head and Neck squamous cell carcinoma	WRCY	3348	4.99
kirc_tcga	Kidney Chromophobe	None	16015	66.29
kirc_tcga	Kidney Chromophobe	RCG	1901	7.87
kirc_tcga	Kidney Chromophobe	WA	4447	18.41
kirc_tcga	Kidney Chromophobe	WRCY	1795	7.43
luad_tcga	Lung adenocarcinoma	None	160977	77.12
luad_tcga	Lung adenocarcinoma	RCG	15728	7.54
luad_tcga	Lung adenocarcinoma	WA	17472	8.37
luad_tcga	Lung adenocarcinoma	WRCY	14547	6.97
lusc_tcga	Lung squamous cell carcinoma	None	47897	77.90
lusc_tcga	Lung squamous cell carcinoma	RCG	4528	7.36
lusc_tcga	Lung squamous cell carcinoma	WA	5312	8.64
lusc_tcga	Lung squamous cell carcinoma	WRCY	3748	6.10
ov_tcga	Ovarian serous cystadenocarcinoma	None	3977	67.73
ov_tcga	Ovarian serous cystadenocarcinoma	RCG	661	11.26
ov_tcga	Ovarian serous cystadenocarcinoma	WA	771	13.13
ov_tcga	Ovarian serous cystadenocarcinoma	WRCY	463	7.88
thca_tcga	Thyroid carcinoma	None	4511	67.17
thca_tcga	Thyroid carcinoma	RCG	809	12.05
thca_tcga	Thyroid carcinoma	WA	832	12.39
thca_tcga	Thyroid carcinoma	WRCY	564	8.40

Table S6. AID motifs contribution of mutations present in 6 cancer type on TCGA.

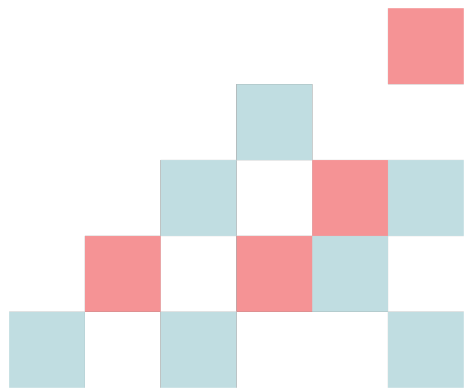
Isla Magdalena, Chile

CHAPTER 5

Autonomous B-cell receptor signaling and genetic aberrations in chronic lymphocytic leukemia-phenotype monoclonal B lymphocytosis in siblings of patients with chronic lymphocytic leukemia

Previously published as

Quinten, E., Sepúlveda-Yáñez, J. H., Koning, M. T., Eken, J. A., Pfeifer, D., Nteleah, V., de Groen, R. A., Saravia, D. A., Knijnenburg, J., Stuivenberg-Bleijswijk, H. E., Pantic, M., Agathangelidis, A., Keppler-Hafkemeyer, A., van Bergen, C. A. M., Uribe-Paredes, R., Stamatopoulos, K., Vermaat, J. S., Zirlik, K., Navarrete, M. A., Jumaa, H. & Veelken, H. Autonomous B-cell receptor signaling and genetic aberrations in chronic lymphocytic leukemia-phenotype monoclonal B lymphocytosis in siblings of patients with chronic lymphocytic leukemia. ***Haematologica*** **109.** Section: Articles, 824–834. doi:10.3324/haematol.2022.282542 (July 2023)



ABSTRACT

Clonal expansion of CD5-expressing B cells, commonly designated as monoclonal B lymphocytosis (MBL), is a precursor condition for chronic lymphocytic leukemia (CLL). The mechanisms driving subclinical MBL B-cell expansion and progression to CLL, occurring in approximately 1% of affected individuals, are unknown. An autonomously signaling B-cell receptor (BCR) is essential for the pathogenesis of CLL. The objectives of this study were functional characterization of the BCR of MBL in siblings of CLL patients and a comparison of genetic variants in MBL-CLL sibling pairs. Screening of peripheral blood by flow cytometry detected 0.2-480 clonal CLL-phenotype cells per microliter (median: 37/μL) in 34 of 191 (17.8%) siblings of CLL patients. Clonal BCR isolated from highly purified CLL-phenotype cells induced robust calcium mobilization in BCR-deficient murine pre-B cells in the absence of external antigen and without experimental crosslinking. This autonomous BCR signal was less intense than the signal originating from the CLL BCR of their CLL siblings. According to genotyping by single nucleotide polymorphism array, whole exome, and targeted panel sequencing, CLL risk alleles were found with high and similar prevalence in CLL patients and MBL siblings, respectively. Likewise, the prevalence of recurrent CLL-associated genetic variants was similar between CLL and matched MBL samples. However, copy number variations and small variants were frequently subclonal in MBL cells, suggesting their acquisition during subclinical clonal expansion. These findings support a stepwise model of CLL pathogenesis, in which autonomous BCR signaling leads to a non-malignant (oligo)clonal expansion of CD5+ B cells, followed by malignant progression to CLL after acquisition of pathogenic genetic variants.

5.1 | INTRODUCTION

Chronic lymphocytic leukemia (CLL) is a monoclonal expansion of more than 5000 mature CD5-expressing B cells per μL of peripheral blood. [259, 260] The clinical behavior of CLL is highly variable and depends strongly on structural and functional properties of the BCR as well as the presence of various recurrent genetic aberrations. [261] Autonomous signaling of the clonotypic BCR expressed by CLL cells, i.e. signaling without engagement of external antigen, is an indispensable oncogenic signal in CLL. [84] This peculiar property of CLL BCR is absent in other types of indolent B-cell lymphoma and in antigen-specific normal B cells. [84] Autonomous BCR signaling is caused by physical interaction between neighboring BCR complexes on the cell surface as demonstrated by seminal crystallographic studies in paradigmatic CLL BCR. [262] In the Eμ-TCL1 transgenic mouse model, autonomous BCR signaling is indispensable for CLL development even when all B cells primarily express an antigen-specific transgenic BCR and are stimulated by their cognate antigen. [85] CLL is characterized by remarkable skewing in the BCR immunoglobulin (BCR IG) gene repertoire, culminating in the existence of subsets of cases with restricted BCR IG features. [263] The largest immunogenetic CLL subset, stereotyped subset #2L, is defined by expression of a BCR with a light chain utilizing the IGLV3-21 gene that carries a G110R mutation at the junction between the variable and constant domains. [86] This seminal IGLV3-21 G110R mutation facilitates autonomous BCR signaling. [262] Numerous gene polymorphisms predispose to the development of CLL. [264–266] Formally, CLL develops through a precursor stadium, an subclinical expansion of mature B cells with the characteristic CD5+CD20lowCD79low CLL phenotype, commonly referred to as CLL-phenotype monoclonal B lymphocytosis (MBL). [76] The prevalence of MBL is approximately 100 times higher than CLL and is relatively increased in siblings of CLL patients. [77, 261] Similar to cases of indolent CLL, MBL cells carry relatively few genetic aberrations. [78, 267] However, the mechanisms leading to benign, longitudinally stable clonal expansion of CLL-phenotype B cells, and the hierarchy and sequence of events causing eventual progression to overt CLL are not fully elucidated yet. In order to investigate the early stages in the ontogeny of CLL, we screened siblings of CLL patients for the presence of MBL with CLL phenotype. Specifically, we sought to clarify whether these MBL cells express BCR with autonomous signaling capacity. In addition, we compared the prevalence of inherited risk loci and acquired genetic aberrations between CLL patients and their first-degree MBL relative.

5.2 | MATERIALS AND METHODS

PATIENTS, PROBANDS, AND SAMPLES

Siblings of CLL patients were asked to provide a single blood sample with informed consent. The study was conducted according to the Declaration of Helsinki and was approved by the Ethical Committees of both participating institutions (Leiden: P12.059; Freiburg: 44/08).

DETECTION, ISOLATION, AND PROCESSING OF CLL-PHENOTYPE CELLS

MBL was detected as discrete CD19+CD5+CD20lowCD79low populations (Figure 5.1) by six-color flow cytometry. DNA and RNA were isolated from MBL cells sorted to >95% purity (Allprep DNA/RNA mini kit; Qiagen, Hilden, Germany). Frequencies of MBL cells were calculated from gated cells and absolute lymphocyte counts.

IDENTIFICATION OF CLONAL BCR TRANSCRIPTS

VDJ and VJ sequences were determined by ARTISAN PCR. [268, 269] Ig sequences were analyzed by Geneious 10.2.6 (Geneious, Auckland, New Zealand) and IMGT/V-QUEST. [270, 271] Assignment to CLL stereotypes was performed by an established purpose-built algorithm. [263, 272]

ANALYSIS OF BCR SIGNALING ACTIVITY

Murine TKO pre-B cells were transduced with retroviral pMIZCC and pMIZYN vectors encoding paired Ig heavy and light chain sequences from individual cases. [84, 86] TKO cells are genetically deficient of ***Rag2, Lambda5,*** and ***Slp65***. ***Slp65*** function is reconstituted by a 4-hydroxytamoxifen-inducible ***Slp65-ERT2*** fusion gene. [74, 84, 273] Calcium mobilization was measured in Indo-1 AM-loaded (ThermoFisher, Waltham, MA), live-gated TKO cells by flow cytometry prior and after addition of 2 µM 4-hydroxytamoxifen (Sigma Aldrich, St Louis, MO). Maximum BCR signaling was measured after addition of crosslinking anti-IgM or anti-IgG antibodies (clones 2022-01 and 2042-01; Southern Biotech, Birmingham, AL). Autonomous BCR signaling was quantitated with correction for totally unresponsive cells during BCR crosslinking (supplementary material) at least twice at approximately 3 and 4 weeks after transduction.

GENETIC CHARACTERIZATION

Variants in 52 B-cell lymphoma-related genes were analysed by the LYMFv1 custom targeted sequencing panel on a Ion S5 system (ThermoFisher). [274] Sequencing reads were aligned to the GRCh37/hg19 human reference genome with TMAP 5.0.7 software under default parameters. CLL-associated copy number variations (CNV) were detected by SNP arrays. [275] Whole exome sequencing (WES) was performed with SureSelect Human All Exon V7 kit (Agilent, Santa Clara, CA) capture on the HiSeq2000 (Illumina, San Diego, CA) platform to an average coverage of 50x. If necessary, DNA was amplified by isothermal alkaline genome amplification (REPLI-g kit; Qiagen). Detected variants were annotated for predicted pathogenicity and occurrence in both CLL and MBL siblings (supplementary material). The global distribution of variant allele frequencies in CLL and MBL was estimated using the Wilcoxon rank sum test with continuity correction (Mann–Whitney U test) function 'wilcox.test()' in R. Monoallelic and subclonal variants present in CLL and MBL samples from siblings were tested using the Wilcoxon signed-rank test function 'wilcox.test(paired=TRUE)'. Both tests were only per-

formed after checking the corresponding test assumptions. Genotyping of inherited CLL risk loci was performed by targeted Sanger sequencing of 24 CLL susceptibility alleles genes and loci covered by WES and SNP array. [266, 276–280] Observed frequencies of risk alleles were compared to the most appropriate reference population from the gnomAD database by Fisher exact test. [281] CLL risk SNP were compared with WES-based data of Northwestern Europeans in gnomAD v2.1.1 and with WGS-based data of non-Finnish Europeans in gnomAD v3.1.2. Polygenic risk scores (PRS) were calculated for individual MBL probands and CLL patients for their genotyped risk loci based on the most recent reported odds ratios. [265, 266] The extent of genetic analyses was restricted in individual cases by insufficient amounts of DNA due to the paucity of clonal CLL-phenotype B cells.

5.3 | RESULTS

PREVALENCE AND FREQUENCY OF CLL-PHENOTYPE CELLS IN SIBLINGS OF CLL PATIENTS

Peripheral blood samples from 191 siblings of CLL patients were received and analyzed at two academic medical centers. Flow cytometry revealed a discrete population of CLL-phenotype CD19+CD5+CD20lowCD19low cells, in the majority with evident light chain restriction, in 34 siblings (17.8%) of 26 CLL patients (Figure 5.1, Supplementary Table 1). The median age of siblings with CLL-phenotype cells was 68 years (range: 43-80 years). The absolute number of CLL-phenotype MBL cells ranged from 0.2 to 1863 per µL blood, and 32 siblings (94%) were classified as low-count MBL with less than 500 clonal CLL-phenotype cells per µL (Figure 5.2A). [282]

SEQUENCE CHARACTERISTICS OF CLONALLY DOMINANT BCR IG OF CLL-PHENOTYPE CELLS IN CLL PATIENTS AND THEIR HEALTHY SIBLINGS

BCR IG transcripts from highly purified CLL-phenotype cells were sequenced in 17 siblings of CLL patients. In two siblings, two different productive VDJ gene rearrangements were identified within the CLL-phenotype MBL cell population (MBL01.1, MBL04.1); all other siblings carried an apparently monoclonal MBL population. Five MBL clones in healthy siblings could be assigned to CLL stereotyped subsets according to expanded immunogenetic criteria (Supplementary Table 1). [263, 272] In particular, the BCR of MBL07.1 and MBL27.2 belonged to CLL subset #2 and #2L, respectively, based on expression of a BCR light chain utilizing the IGLV3-21 gene with the G110R mutation. [86] An IgG-expressing MBL fulfilled the criteria of CLL subset #4. [283] These MBL BCR were predicted to exert autonomous signaling based on structural analyses. [262] Furthermore, all sequenced MBL 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. [84, 284] The somatic mutation status of the clonotypic rearranged IGHV gene did not correlate to the degree of expansion of CLL-phenotype cells (Figure 5.2B). In addition, IGHV gene mutational status could differ between

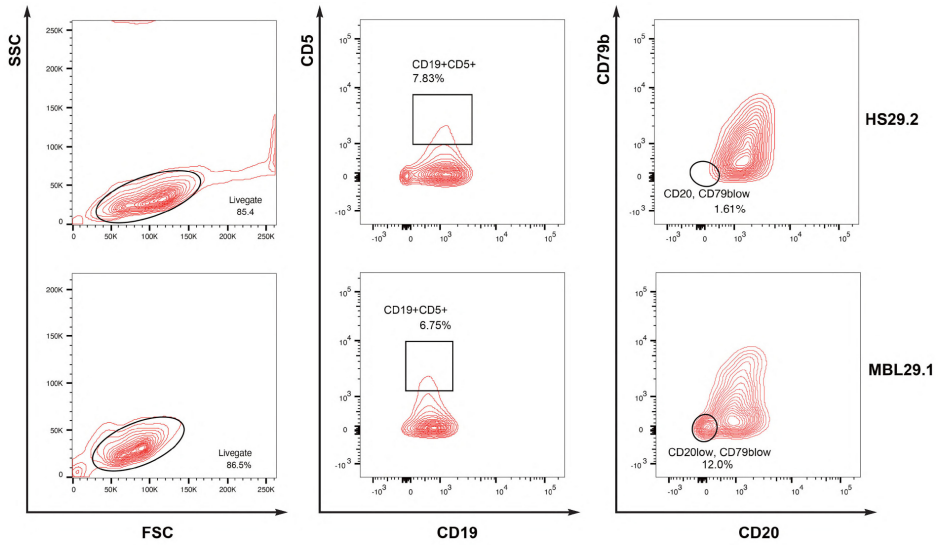


Figure 5.1. Detection of CLL-phenotype MBL cells. Flow cytometry screening of two healthy siblings of CLL patient 29.1. Upper row: Healthy sibling HS29.2 without detectable CLL-phenotype cells. Lower row:: Healthy sibling MBL29.1 with the lowest detected expansion of CLL-phenotype cells (CD19+CD5+CD20lowCD79low) in this study. SSC: Side scatter. FSC: Forward scatter.

CLL patients and their siblings; in particular, three non-subset 2L MBL (17.6%) expressed unmutated BCR IG whereas the corresponding CLL BCR was mutated (Figure 5.2C).

FUNCTIONAL CHARACTERISTICS OF MBL BCR FROM SIBLINGS OF CLL PATIENTS

BCR from clonally dominant MBL cells of 11 siblings of CLL patients, including all five MBL assigned to a CLL stereotyped subset, 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. As predicted, [86] both CLL subset #2L MBL BCR showed robust calcium flux upon addition of 4-hydroxytamoxifen without antigen exposure or BCR crosslinking (Figure 5.3). In addition, all other tested MBL BCR from siblings of CLL patients had autonomous BCR signaling activity. On average, autonomous BCR signaling was slightly stronger in CLL compared to MBL in each individual sibling pair ((Figure 5.4A). There was no quantitative correlation between BCR signaling activity of CLL and MBL sibling pairs ((Figure 5.4B). BCR signaling strength was positively correlated with MBL cell counts at the time of sampling (Figure 5.4C).

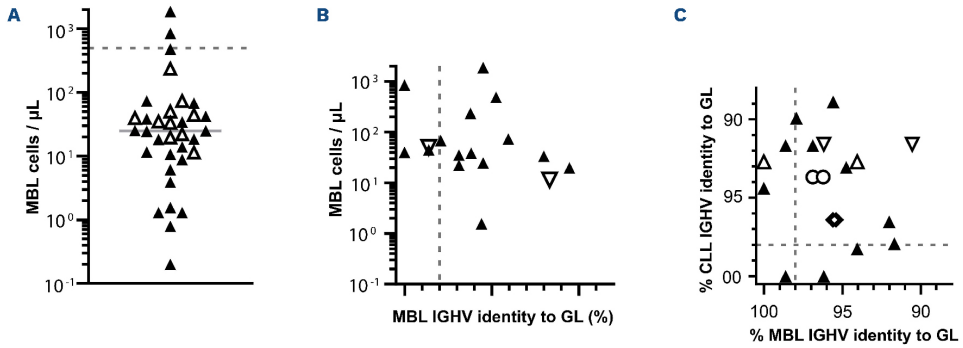


Figure 5.2. Characteristics of MBL cells in CLL siblings. A. MBL counts of all 34 siblings with a detectable CLL-phenotype B-cell population (CD19+CD5+CD20lowCD79low). Solid line: median. Dotted line: Commonly accepted threshold between low- and high-count MBL. Open triangles: MBL selected for functional BCR testing. B. IGHV gene SHM status (x axis) and counts of MBL cells (y axis). Dotted line: Commonly accepted threshold between mutated and unmutated BCR status in CLL. Red symbols: MBL cells from two CLL siblings expressing a CLL subset 2 BCR predicted to exert autonomous signaling.10 GL: Germ line. C. IGHV gene identity to germline in MBL cells (x axis) and CLL cells (y axis) in siblings. Dotted lines: Commonly accepted boundary between mutated and unmutated BCR status in CLL. Red and orange symbols: Bona fide biclonal MBL based on expression of two immunoglobulin heavy chains (MBL01.1, MBL04.1). Green and blue symbols: CLL patients (CLL24.1, CLL30.1) with two siblings with MBL, respectively (MBL24.1, MBL24.2; MBL30.1; MBL30.2). GL: Germ line.

COMPARATIVE GENETICS OF CLL PATIENTS AND MBL SIBLINGS

Analysis of known germ line variants from up to 26 different gene loci associated with increased incidence of CLL revealed a significantly higher prevalence of twelve risk alleles within our cohort (Supplementary Table 2). A polygenic CLL risk score composed of 24 risk loci was higher in both CLL and MBL sibling than the average reference score calculated from reference population data (Figure 5.5C). [281] There was no evidence for a higher PRS in CLL cases in a matched-pair comparison to their MBL siblings. Taken together, these findings suggest that CLL risk loci predispose to clonal expansion of CLL phenotype cells in both low-count MBL and CLL. Vice versa, we did not find evidence for a particular association of CLL risk alleles with malignant progression from MBL to clinical CLL. Fifteen MBL samples with sufficient DNA were genotyped by SNP arrays to detect recurrent CLL-associated CNV. Ten MBL carried recurrent CLL-associated CNV, including a del(17)(p13-q21) in one case and del(13)(q14) in nine cases. One additional MBL carried large areas of loss of heterozygosity, and another sample a del(14)(q22.2) in combination with a trisomy 18q12, resulting in a total of eleven MBL samples with informative CNV (Table 1). Eleven of 15 genotyped CLL cases within this study carried informative CNV, suggesting a similar overall prevalence of CNV between CLL cases and MBL in their siblings. Quantitative analysis of the SNP array data indicated that all malignant cells harbored the respective CNV in nine of eleven CLL samples. In contrast, individual CNV were present in only a minor fraction of the highly

purified MBL cells in ten of eleven samples (p=0.0019; Fisher exact test). We corroborated these quantitative CNV data with analysis of small variants, i.e. SNV and indels from WES in ten CLL-MBL sample pairs. Identical variants detected in a CLL case and its MBL sibling were presumed to be inherited and were considered noninformative with respect to differential pathogenesis between MBL and CLL. In contrast, variants that were not shared between siblings, i.e. variants observed exclusively in either a CLL sample or its MBL sibling, could either be inherited or acquired somatically and were considered as potentially informative on differential MBL-CLL pathogenesis. Initially, we focused on 120 genes recurrently mutated in CLL according to the COSMIC database. As determined by targeted panel sequencing and/or WES analysis in ten CLL-MBL sample pairs, CLL and MBL carried similar numbers of unique gene variants with predicted pathogenic relevance. Variants in these genes were infrequent, and their variant allele frequency (VAF) did not differ significantly between CLL and MBL (Figure 5.5B).

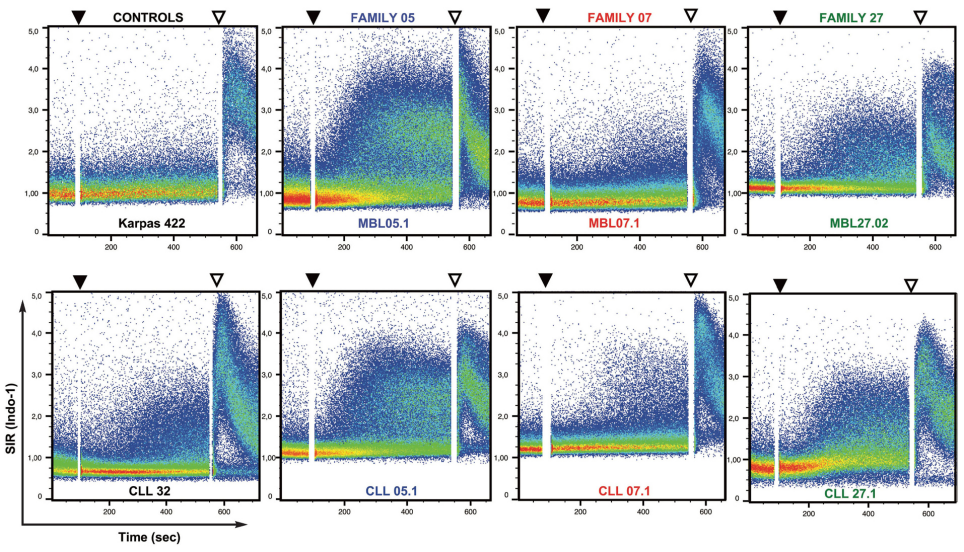


Figure 5.3. Autonomous BCR signaling of CLL and MBL cells in three representative sibling pairs. Calcium mobilization was measured by flow cytometry as Indo-1 AM signal intensity ratios of BCR-transduced TKO cells. Black triangle: Time point of addition of 4-hydroxytamoxifen to reconstitute a functional BCR signaling cascade. White triangle: Time point of addition of anti-IgM/IgG crosslinking antibody. K422: TKO cells transduced with BCR of lymphoma cell line Karpas 422 as negative control. CLL 32: TKO cells transduced with BCR of CLL 32 as positive control.7 MBL05.1: TKO cells transduced with BCR of MBL cells in a sibling expressing a CLL subset #60 BCR.5, 23 MBL07.1, MBL27.2: TKO cells transduced with BCR of MBL cells in siblings expressing a CLL subset #2 BCR. TKO cells transduced with CLL BCR are depicted below the MBL of their respective siblings.

Genome-wide, VAF of non-shared variants were significantly lower in MBL samples than in CLL cases ((Figure 5.7A). In order to test whether this difference could be attributed to a higher fraction of subclonal variants in MBL compared to CLL, we separately compared in

sibling pairs the numbers of non-shared monoallelic variants (VAF 0.43-0.55), presumably carried by all cells in a given sample, and numbers of monoallelic subclonal variants (VAF 0.1-0.33). In accordance with this hypothesis, monoallelic variants were detected at similar frequencies in CLL samples and their respective sibling MBL samples (Figure 5.7B). In contrast, the prevalence of subclonal variants was higher in MBL cells compared to their sibling CLL counterpart (Figure 5.7C).

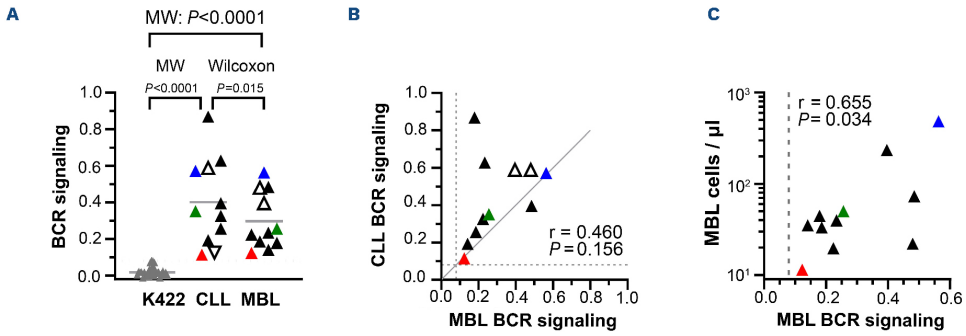


Figure 5.4. Quantitative analysis of autonomous BCR signaling of CLL and MBL cells. A. Autonomous BCR signaling strength in TKO cells transduced with clonal BCR. K422: All measurements of Karpas 422 BCR (negative control) performed during this study. Open triangle: CLL 32 (positive control). Orange triangles: A CLL with two MBL siblings. Other triangle colors correspond to Figure 3. Black triangles: All remaining tested CLL and MBL cases. M-W: Unpaired comparison by Mann-Whitney test. Wilcoxon: Paired comparison between CLL and MBL within siblings by Wilcoxon matched-pair signed rank test. B. Correlation of autonomous BCR signaling strength between CLL and MBL siblings. Colors correspond to Figure 3 and 4A. Dashed lines indicate maximum BCR signaling strength observed in Karpas 422 negative controls. C. MBL counts (y axis) according to autonomous BCR signaling strength (x axis). Colors correspond to Figure 3. Dashed line indicates maximum BCR signaling strength observed in Karpas 422 negative controls.

5.4 | DISCUSSION

This systematic study was undertaken to clarify whether the presence of an autonomous BCR signal could explain the different clinical behavior between non-malignant clonal expansions of cells with CLL phenotype (MBL) and clinically malignant CLL. In order to minimize potential confounding effects of inherited germ-line variants, we investigated this hypothesis in a comparative manner in siblings of CLL patients. Detection limit and range of clonal B cells with a CLL phenotype were very similar to a recently published study on members of CLL families. [261] However, the observed MBL prevalence of 17.8% was relatively high compared to previous reports on CLL siblings and in the range of individuals belonging to CLL families defined by more than one first-degree relative diagnosed with CLL. [77, 261]

Unbiased amplification of expressed heavy and light IG chain gene rearrangements revealed single dominant BCR in 15 and potential biclonality in two MBL subjects. Potential MBL oligoclonality has been recognized in MBL, especially in the low-count subtype. [285, 286]

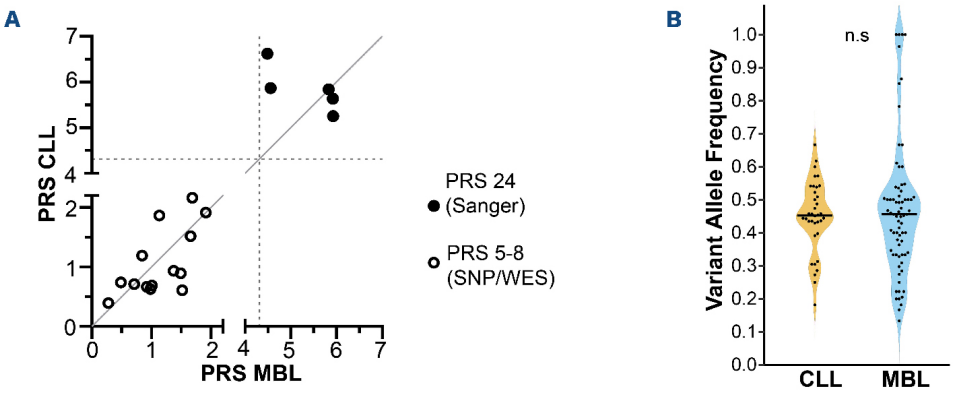


Figure 5.5. Comparison of chronic lymphocytic leukemia-associated variants in monoclonal B lymphocytosis-chronic lymphocytic leukemia siblings. (A) Polygenic risk score (PRS) for chronic lymphocytic leukemia (CLL) in probands with monoclonal B lymphocytosis (MBL) (x axis) and their CLL sibling (y axis). Depending on available data, the PRS was calculated for 24 CLL risk loci (black dots) or five to eight loci as assessed by single nucleotide polymorphism array and/or targeted Sanger sequencing (open circles). Dashed lines indicate the mean PRS of the reference European population for the 24 loci analyzed. (B) Distribution of variant allele frequencies of non-shared variants in 120 CLL driver genes as detected by whole exome sequencing in seven CLL cases and ten MBL siblings from seven families. CLL and MBL were compared by the Wilcoxon signed rank test. SNP: single nucleotide polymorphism; WES: whole exome sequencing; n.s.: not significant.

5

The identification of a fraction of MBL with unmutated BCR IG in our series is also in line with a published study. [286] Whereas low-count MBL was reported to belong only rarely to any immunogenetically defined CLL stereotype subset, [263, 286] we encountered five instances of BCR IG from low-count MBL (33.3%) that could be assigned to such CLL subsets. [263, 272] It remains hypothetical whether this apparent discrepancy is due to focusing on MBL carriers with CLL siblings in our study. In healthy persons who eventually developed CLL, app. 20% of dominant clonotypes identified during the prediagnostic period could be assigned to CLL stereotypes. [287]

With respect to the primary aim of our study, MBL BCR IG resembling certain CLL subsets were predicted to have autonomous BCR signaling activity akin to CLL. This prediction especially applied to the herein first reported low-count MBL expressing an isotype-switched BCR of CLL subset #4, and two subset #2L MBL cases expressing an IGLV3-21 light chain with the seminal G110R mutation. Both of these structural BCR IG characteristics mechanistically cause autonomous BCR signaling by facilitating steric interactions between neighboring BCR complexes on the cell surface. [86] Limited by the sample size, we found no evidence for correlations in structural features of the BCR, e.g. mutational status or IG gene usage, between CLL-MBL sibling pairs.

As already suggested by the MBL BCR IG sequences, our findings establish unequivocally by quantifiable functional testing that the BCR of MBL, i.e. a premalignant clonal expansion of

Family	MBL		CLL	
	Case	Aberration	Case	Aberration
1	MBL01.1	13q14.2q14.3x0[0.7]	CLL01.1	13q14.2q14.3x1
2	MBL02.1	13q14.2x1[0.6] 13q14.2q14.3x0[0.6] 13q14.3x1[0.6]	CLL02.1	13q14x1 8q21.3qterx3 15q26.1q26.3x1
3	MBL03.1	no CNV	CLL03.1	2p22.1p16.1x1 8q24.22x1
4	-	-	CLL04.1	no CNV
5	MBL05.1	no CNV	CLL05.1	13q14.13q14.2x1[0.5] 13q14.2q14.3x0[0.5] 13q14.3x1[0.5]
6	MBL06.1 MBL06.2	13q14.2q14.3x1 14q22.2q32.12x1[0.6] 18q12.3q23x3[0.6]	CLL06.1	14q23.2q32.13x1
7	MBL07.1 MBL07.2	13q14.2q14.3x1[0.15] 13q14.13q31.1x1[0.15]	CLL07.1	13q14.11q14.2x1
8	MBL08.1	13q14.2q14.3x1[0.85]	CLL08.1	5q33.1q35.3x3[0.3] 6q14.1q27x1 8q23.1q24.3x3 17p13.3p12x0[0,5]
9	MBL09.1	no CNV	CLL09.1	no CNV
10	MBL10.1	13q14.2q14.3x1[0.5]	CLL10.1	13q14x1
11	MBL11.1	13q14.2q14.3x1[0.7]	CLL11.1	no CNV
12	MBL12.1	13q14.2q14.3x0[0.5]	CLL12.1	13q14.2q14.3x1
13	-	-	CLL13.1	11q22.3q25x1 3q26.1q29x3 4p16.3p14x1 17q21.32q25.3x3
15			CLL15.1	no CNV
16	MBL16.1	2q36.3q37.3x1[0.7] 17p13.3q21.2x1[0.7]	CLL16.1	4p16.3p15.2x1[0.3] 4p14x1[0.3] 4p13x1[0.3] 4p13x1[0.3] 5p15.33p15.2x1[0.3] (11)cx[0.3~0.8] 13q14.11q14.2x1[0.5] 13q14.2q21.33x0[0.8] 13q21.33q22.3)x1[0.5]
Total analyzed	14		15	
With CNV	11		11	
With clonal CNV	1		9	
With subclonal CNV only	10		2	

MBL: monoclonal B lymphocytosis; CLL: chronic lymphocytic leukemia; CNV: copy number variation.

Figure 5.6. Copy number variations in monoclonal B lymphocytosis and chronic lymphocytic leukemia samples as detected by single nucleotide polymorphism array.

CD5-positive B cells with CLL phenotype, exerts antigen-independent, autonomous signaling. All known origins of autonomous BCR signaling in CLL were also encountered in our MBL cases, i.e. acquisition during primary V(D)J recombination in cases with unmutated BCR, acquisition by somatic hypermutation as exemplified best by the single G110R mutation of the IGLV3-21 gene, [86] and acquisition by class switch recombination. [262] Since our study specifically addressed CLL-MBL sibling pairs, we were able to compare the autonomous BCR signaling strength quantitatively in matched-pair fashion. This analysis revealed significantly lower intensity of BCR signaling in MBL compared to CLL. While autonomous BCR signaling is generally dependent on individual HCDR3 in conjunction with recurrent motifs in FR2 and interactions of defined amino acid restudies in some instances, [84, 86, 262, 284] quantitative

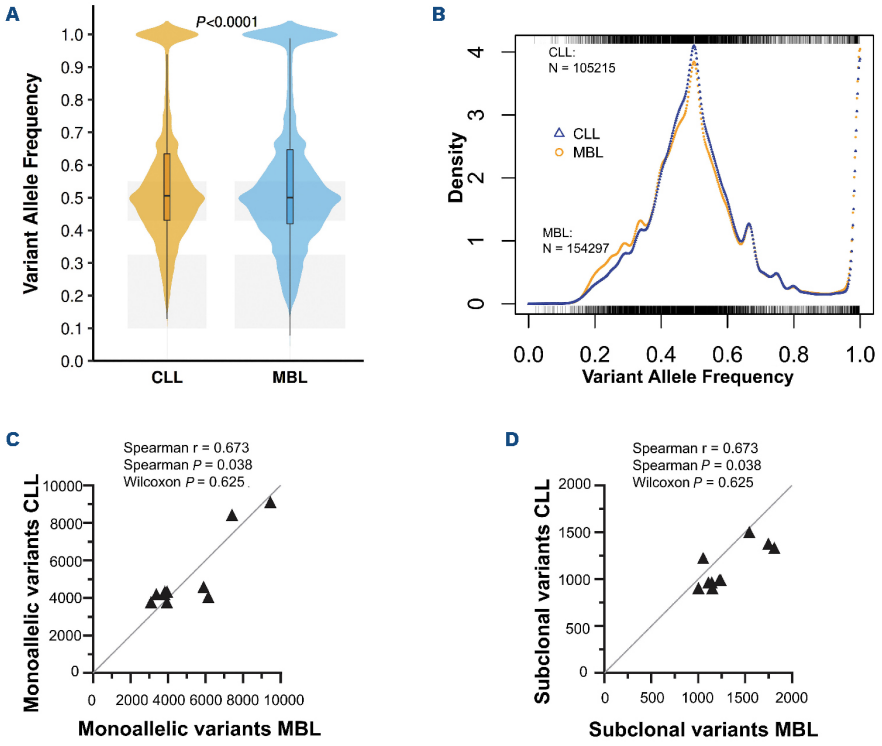


Figure 5.7. Comparison of genome-wide variants in siblings with monoclonal B lymphocytosis or chronic lymphocytic leukemia. (A) Global distribution of variant allele frequencies (VAF) in cases of chronic lymphocytic leukemia (CLL) and monoclonal B lymphocytosis (MBL). VAF were compared by the Wilcoxon rank sum test with continuity correction (Mann–Whitney U test). Shaded boxes indicate the ranges of VAF analyzed in (C) and (D). (B) Kernel density estimation distribution and rug plots of VAF in CLL and MBL. The rug plot marks along the x-axis (top and bottom) indicate positions of the VAF for CLL and MBL, respectively. A total of 105,215 CLL data points and 154,297 MBL data points were analyzed. (C) Numbers of monoallelic variants (VAF=0.43-0.55; present in all cells) of CLL and MBL samples from siblings. Correlations between CLL and MBL samples were calculated by the nonparametric Spearman correlation. CLL and MBL samples were compared by the Wilcoxon matched-pair signed rank test. (D) Numbers of subclonal variants (VAF=0.1-0.33) in CLL and MBL samples from siblings. Correlations between CLL and MBL samples were calculated by the nonparametric Spearman correlation. CLL and MBL samples were compared by the Wilcoxon matchedpair signed rank test. CLL: chronic lymphocytic leukemia; MBL: monoclonal B lymphocytosis.

differences in autonomous BCR signaling cannot be assigned yet to defined structural BCR characteristics. However, BCR signaling strength was directly correlated to the expansion of the MBL clone at the time of sampling. The novel observation requires validation in independent cohorts. Nevertheless, the quantitative differences of BCR signaling between CLL and MBL and in MBL expansion suggest a possible causal and differential role of BCR signaling in growth kinetics and subsequent clinical behavior of a neoplastic CLL-phenotype B-cell clone.

It would have been highly desirable to validate the correlation between BCR signaling strength and clonal B-cell expansion by longitudinal sampling. However, the approvals by the institutional ethics committees explicitly excluded longitudinal analyses and mandated to inform the healthy siblings with a categorical result of the screening. Since low-count MBL does not represent a definite pathological condition and rarely progresses to CLL,([261, 282] it was reported as absence of pathological findings. As a consequence, the ethical committees concluded that repeat sampling could potentially give rise to anxiety given the prior absence of pathology in the healthy siblings, and explicitly denied approval for longitudinal analysis.

In order to interpret BCR function in the context of genetic mechanisms in CLL pathogenesis, we addressed genetic susceptibility, genetic instability, and recurrent CLL-associated genetic changes in CLL cases and their MBL siblings. These studies were restricted by limited quantities of DNA extracted from the MBL samples. Nevertheless, analysis of CLL susceptibility loci indicated that both CLL and MBL probands have a similar PRS for CLL that is higher than in the general (Northern) European reference population from the gnomAD database. While these findings were predictable from the study design in CLL-MBL sibling pairs, on the other hand they do not provide any evidence that these CLL risk loci drive malignant progression rather than initial premalignant expansion of the CLL-phenotype clone. With respect to the possible inherited shared predisposition for MBL and CLL, it would have been highly desirable to strengthen this conclusion by showing lower PRS in the CLL siblings without detectable MBL cells. The study design, however, did not comprise informed consent for genetic analyses in non-MBL probands. Notwithstanding this limitation of our study, there appears no plausible causal relationship between the presence of any risk allele and the stochastic acquisition of a BCR with autonomous BCR signaling based on the known function of the genes located at the CLL risk loci.

The overall prevalence of CLL-associated CNV and gene variants was not evidently different between CLL and MBL in matched comparisons. In accordance with the literature, del(13)(q14) was the most abundant CNV in MBL, but the detected del(17)(p) has also been described occasionally. [288] Our observed low prevalence of potentially pathogenic variants in CLL driver genes is also in accordance with published data. [267] However, we noticed that CLL-associated CNV were frequently subclonal in MBL, suggesting that these aberrations might be gradually acquired at this stage. This observation is in accordance with findings from a longitudinal study in low-count MBL, wherein the categorical presence of CLL-associated CNV approximately doubled after a median interval of seven years compared to the primary sample. [289] The impression of gradual acquisition of mutations in MBL was independently supported by the observation that MBL cells carry more subclonal variants detected by next-generation sequencing than CLL, whereas the number of variants carried by all clonal cells did not differ between CLL and MBL. In CLL, the presence of subclonal drivers is associated with clinical outcome. [81]

In summary, our data demonstrate that autonomous BCR signaling operates in MBL in similarity to CLL. A hypothetical absence of this mechanism therefore does not explain the difference between benign MBL and malignant CLL. However, relatively low BCR signaling

strength in MBL may offer a plausible explanation for less aggressive expansion than in CLL. In addition, our data suggest that CLL risk loci likewise cannot account for the different clinical behavior between both conditions, and lend further support to the conclusion that prevalence and type of genetic pathogenic changes are indistinguishable between MBL and low-risk CLL. [267, 289] However, the observation of subclonal genetic CLL driver mutations in MBL supports a scenario of gradual clonal expansion driven by moderate autonomous BCR signaling, possibly leading to a certain degree of genetic instability that facilitates gradual acquisition of CLL driver mutations.

This conclusion supports the notion that MBL and CLL represent a spectrum of the same biological condition, wherein the cumulative stimuli from constitutive immune signaling through the BCR together with timing and type of eventually acquired genetic drivers govern the clonal dynamics from long-term balance between expansion and apoptosis over gradual expansion to clinically aggressive CLL.

5.5 | SUPPLEMENTARY MATERIAL

SUPPLEMENTAL METHODS

Patients, probands, and samples

CLL patients under management of the outpatient clinics at the participating centers were asked to inform their respective living siblings about this study and to provide their contact information. Approvals by the ethical committees explicitly excluded longitudinal analyses and mandated to provide siblings with a categorical screening result.

Detection, isolation, and processing of CLL-phenotype cells

Differential blood counts were obtained for each blood sample. DNA and RNA were directly isolated from sorted CLL-phenotype cells by spun-driven silica matrix columns (Allprep DNA/RNA mini kit; Qiagen, Hilden, Germany).

Identification of clonal BCR transcripts

VDJ and VJ sequences were determined by direct Sanger sequencing from gel-purified ARTISAN PCR products (Wizard SV gel and PCR clean-up kit; Promega, Madison, WI) or after molecular cloning (Topo TA cloning vector; ThermoFisher, Waltham, MA).

Analysis of BCR signaling activity

pMIZCC and pMIZYN expression vectors1 with inserted paired VDJ and VJ sequences from individual cases were synthesized (BaseClear BV, Leiden, The Netherlands) and packaged by liposomal cotransfection (FuGENE HD transfection reagent; Promega) of Phoenix-E cells with the pCL-eco helper vector. Murine TKO cells were transduced with virus-containing supernatant on fibronectin-coated plates at 48 h after transfection (RetroNectin; TaKaRa, Shiga, Japan). [86] TKO cells are arrested at the pro-B-cell stage by genetic deficiency of Rag2, Lambda5, and Slp65. Slp65 function is reconstituted by a ***Slp65-ERT2*** fusion gene that enables BCR signaling in the presence of 4- hydroxytamoxifen. [84, 273]

Calcium mobilization was measured in Indo-1 AM-loaded (ThermoFisher), live-gated TKO cells by flow cytometry as the ratio of signal intensities (SIR) at 405 and 485 nm for 90 sec. After addition of 2 μM 4-hydroxytamoxifen (Sigma Aldrich), calcium mobilization was measured for additional 7.5 min. Finally, maximum BCR signaling was induced by adding an unlabeled crosslinking anti-IgM or anti-IgG gamma antibody (clones 2022-01 and 2042-01; Southern Biotech, Birmingham, AL), and measurement was resumed for additional 3 min. Every BCR-transduced TKO population was measured at least twice at 3 approximately 3 and 4 weeks after transduction. The 2nd measurement was used for statistical analyses with two-tailed Mann-Whitney or Wilcoxon's rank sum tests for unpaired and paired comparisons, respectively, and Spearman's correlation coefficient.

Quantitative assessment of autonomous BCR signaling

The fraction of cells with a 405/485 SIR above the 95th percentile prior to addition of tamoxifen (Qaut) was determined with correction for totally unresponsive cells during BCR crosslinking. Subsequently, the calibrated median SIR observed in this Qaut fraction was determined (calSIRaut). Autonomous BCR signaling was quantitated as the arithmetic product of Qaut x

calSIRaut .

Details of recording (see Figure):

Measurements of BCR signaling activity is performed in 3 phases:

1. Baseline Phase (BP): Time interval from the start of measurements until the addition of 4-hydroxytamoxifen
2. Experimental Phase (EP): Time interval from addition of tamoxifen until the addition of cross-linking agent
3. Maximum Phase (MP): Time interval from addition of cross-linking agent to the end of measurement

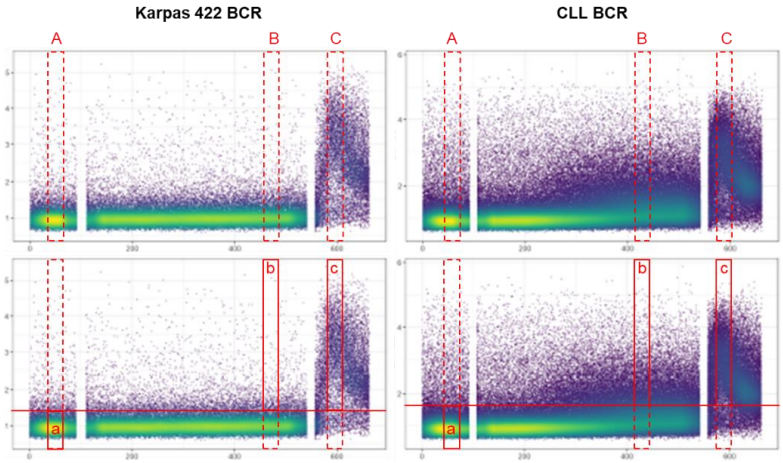
Three frames are defined and the number of events within each frame recorded:

1. A. Frame A: Stable interval in BP with number of events = NA
2. B. Frame B: Stable interval towards the end of EP with number of events = NB
3. C. Frame C: Interval starting at the time point of maximum increase signaling intensity ratio (SIR) and covering the interval of peak SIR in MP with number of events = NC.

Within each frame, a gate is defined and the number of events within each gate recorded:

a. Gate a: Part of frame A with upper limit (=SIR95) set to contain 95% of all events within frame A = an
b. Gate b: Part of frame B with lower limit set at SIR95 with number of events = nb
c. Gate c: Part of frame C with lower limit set at SIR95 with number of events = nc

Mean SIR are recorded for gates b (SIRb) and c (SIRc).



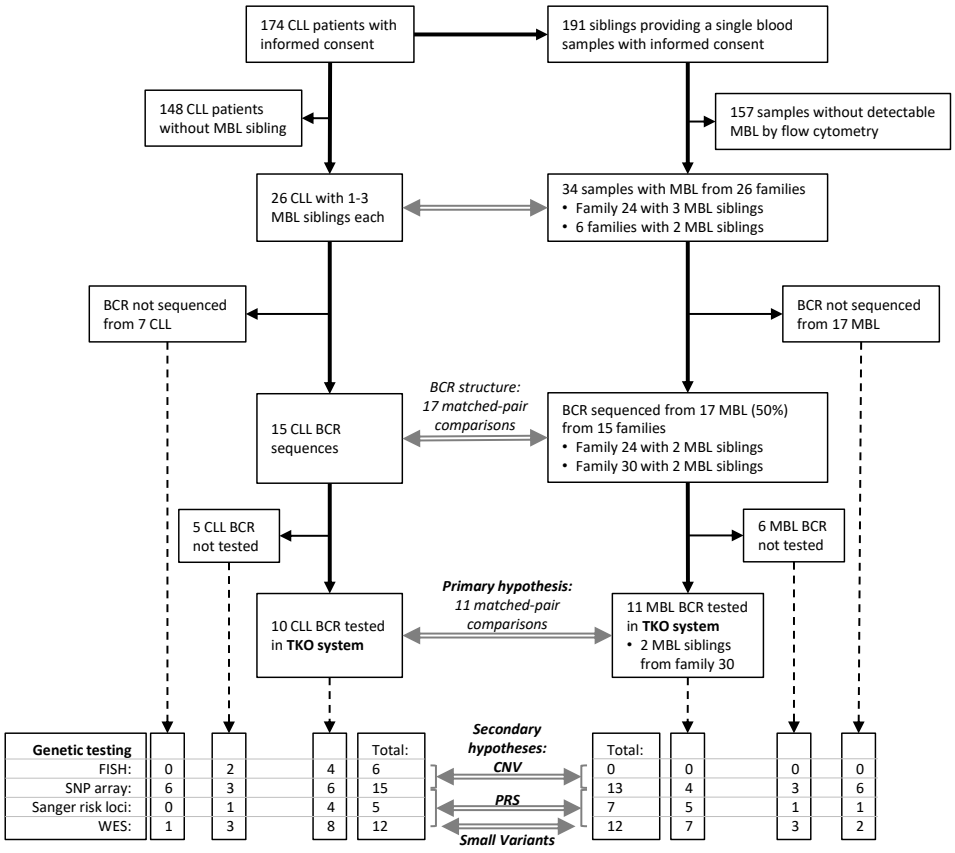
Interpretation:

- The fraction of false-positive events is set to 0.05.
- The fraction of non-responsive cells is calculated as (NC-nc) / NC.
- The fraction of cells with autonomous BCR signaling activity Qaut with correction for 5% false-positive cells and fraction of unresponsive cells is calculated as: Qaut = (nbNC – 0.05NBNC) / ncNC
- The calibrated autonomous BCR signal calSIRaut is: calSIRaut = SIRb/ SIRc
- BCR signaling strength is calculated as the product of the corrected fraction of cells with autonomous signalling and the calibrated autonomous BCR signal: QautcalSIRaut

Whole Exome Sequencing

FASTQ files from whole exome sequencing (WES) were processed using the Sarek workflow v2.7 and aligned to the human reference genome GRCh38 using Burrows Wheeler Algorithm (BWA) v0.7.17. [290, 291] Duplicated mapped reads were marked, local realignment of regions flanking indels and recalibration of base quality scores were performed to obtain more accurate bases according to the Genome Analysis ToolKit (GATK) best practices version v4.1.7.0. [292] Single nucleotide variants (SNV) and short insertions and deletions (INDELS) were called using Strelka2 v2.9.10. [293] Only high 5 confidence variants defined by quality scores (GQX) of at least 15 for SNV and 30 for INDELS were kept. The resulting variant call files were annotated by Ensembl-VEP (v103) with four filtering steps. [294] Variants were first filtered for the 120 most frequent mutated genes in CLL according the COSMIC database and an in-house cohort (Supplemental Table 3). [235, 295] Thereafter, variants were filtered based on their effect, i.e. frameshift, inframe deletion, missense, missense variant & splice region variant, splice region variant & synonymous variant, synonymous, in-frame insertion, stop gained, stop lost, frameshift variant & stop lost, missense variant & splice region variant, and coding sequence variant. Subsequently, variants were filtered for predicted deleterious effects

using a CADD phred score of >20, and benign variants annotated in Clinvar 202008 were discarded. Workflow quality control metrics were calculated and aggregated by MultiQC v1.8. 5 One sample (MBL24.3) with a high percentage (24.4%) of unmapped reads was discarded at this stage.



Supplementary Figure 1. Consort-like diagram of numbers of samples and experimental analyses. CNV: Copy number variations. PRS: Polygenic risk score.

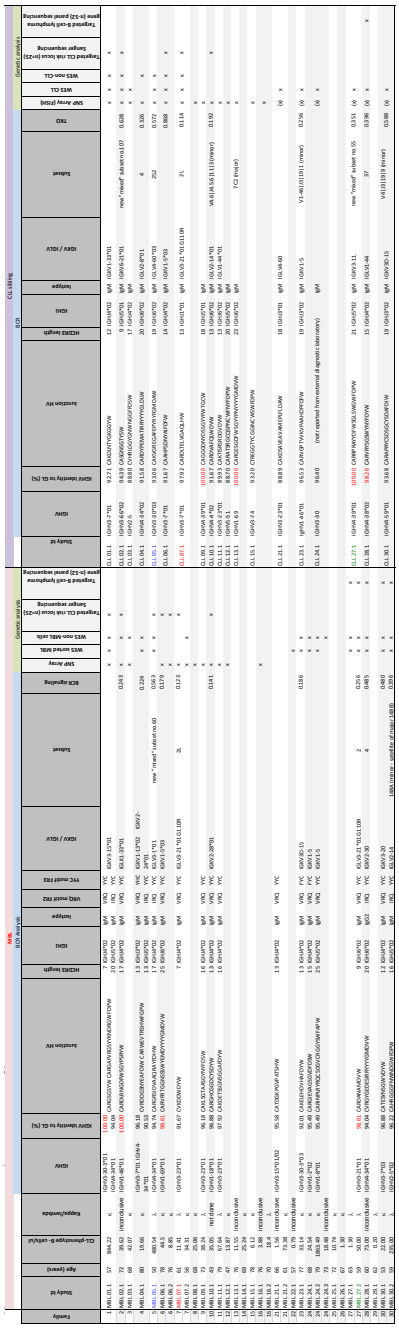
SUPPLEMENTARY TABLES

Supplementary Table 1. Characteristics and analyses of MBL and CLL in sibling pairs

Supplementary Table 2. Analysis of CLL susceptibility loci in CLL-MBL sibling pairs

Supplementary Table 3. CLL-related genes analyzed for small variants

Supplementary Table 4. Non-shared variants in 120 CLL-related genes



Supplementary Table 1. Characteristics and analyses of MBL and CLL in sibling pairs.

Supplementary Table 2. Analysis of CLL susceptibility loci in CLL-MBL sibling pairs.

ADAM19	DLG2	LINGO2	PCLO
ADGRL2	DLGAP2	LRMDA	PDE4D
ADGRL3	DMD	LRP1B	PDIA4
AGBL1	DPP10	LRRTM4	PDZD3
AGBL4	DPP6	LUZP2	PRKCH
AKAP13	DSCAM	MACROD2	PRKG1
ANKS1B	EPHA5	MAGI2	PRKN
ASTN2	EPHA6	MCL1	PTPRD
ATM	ERBB4	MDGA2	PTPRN2
ATRNL1	EYS	MYD88	PTPRT
BIRC3	FAM155A	MYH15	RBFOX1
BRINP3	FAT3	NAALADL2	RIN3
CACNA2D1	FHIT	NAGPA	ROBO1
CACNA2D3	FSTL5	NCAM2	ROBO2
CCSER1	GABRG3	NCKAP5	ROS1
CDH12	GALNTL6	NEGR1	RYR2
CDH13	GPC5	NELL1	SEMA6D
CDH18	GPC6	NKAIN2	SF3B1
CELSR1	GREB1L	NLGN1	SGCZ
CNTN4	GRID2	NOTCH1	SH3TC1
CNTN5	GRIK2	NPAS3	SNTG1
CNTNAP2	GRM7	NRG3	SPAG16
CNTNAP5	H1-2	NRXN1	SPOCK3
CSMD1	H1-3	NRXN3	THSD7B
CSMD3	H1-5	OPCML	TP53
CTNNA2	IGLL5	PARD3B	UNC5D
CTNNA3	IL1RAPL1	PARP4	USH2A
CTNND2	IL1RAPL2	PAX5	XPO1
DCC	KCNIP4	PCDH11X	ZFHX3
DGKB	KLHL1	PCDH15	ZFPM2

Supplementary Table 3. CLL-related genes analyzed for small variants.

Family	sample	CHROM	POS	SYMBOL	Gene	REF	ALT	QUAL	Allele	Consequence	IMPACT	cDNA_position	CDS_position	Protein_position	Amino_acids	Codons	Existing_variation
1	CLL01.1	chr1	49245770	AGBL4	ENSG00000186094	C	A	18	A	missense_variant&splice_region_variant	MODERATE	535	377	126	W/L	tGg/tTg	rs1370279151
1	CLL01.1	chr1	71407471	NEGR1	ENSG00000172260	T	G	46	G	missense_variant	MODERATE	1085	1040	347	Y/S	tAc/tCc	rs41289154&COSV60833851&COSV60866382
1	CLL01.1	chr2	137160352	THSD7B	ENSG00000144229	T	A	118	A	missense_variant	MODERATE	1687	1509	503	D/E	gaT/gaA	rs4954474&COSV55680598
1	CLL01.1	chr3	108470146	MYH15	ENSG00000144821	G	A	259	A	missense_variant	MODERATE	1567	1510	504	H/Y	Cac/Tac	rs9868484&COSV56308865
1	CLL01.1	chr4	61979743	ADGRL3	ENSG00000150471	C	G	73	G	missense_variant	MODERATE	1154	1156	386	L/V	Ctg/Gtg	rs125091100&COSV72232449&COSV99072504
1	CLL01.1	chr6	64307032	EYS	ENSG00000188107	T	C	47	C	missense_variant	MODERATE	6668	6129	2043	I/M	atA/atG	
1	CLL01.1	chr9	136510700	NOTCH1	ENSG00000148400	C	T	21	T	missense_variant	MODERATE	2955	2693	898	G/D	gGc/gAc	rs1166646402
1	CLL01.1	chr20	14085630	MACROD2	ENSG00000172264	C	T	122	T	missense_variant	MODERATE	568	173	58	T/I	aCt/aTt	rs29900505&COSV53960729
1	MBL01.1	chr1	216207290	USH2A	ENSG00000042781	T	A	18	A	missense_variant	MODERATE	3738	3299	1100	E/V	gAg/gTg	
1	MBL01.1	chr2	140274548	LRP1B	ENSG00000168702	C	G	15	G	missense_variant	MODERATE	13305	13018	4340	D/H	Gat/Cat	COSV101251919
1	MBL01.1	chr3	108441130	MYH15	ENSG00000144821	G	A	316	A	missense_variant	MODERATE	2903	2846	949	T/I	aCt/aTt	rs12638212&COSV56309895
1	MBL01.1	chr8	1565843	DLGAP2	ENSG00000198010	C	A	67	A	missense_variant	MODERATE	1571	1391	464	P/Q	cCa/cAa	rs2301963&COSV70102130
1	MBL01.1	chr14	92651884	RIN3	ENSG00000100599	C	T	337	T	missense_variant	MODERATE	987	835	279	R/C	Cgc/Tgc	rs117068593&COSV53649944
1	MBL01.1	chrX	31507453	DMD	ENSG00000198947	C	A	18	A	missense_variant&splice_region_variant	MODERATE	8455	8218	2740	D/Y	Gac/Tac	
2	CLL02.1	chr2	115753241	DPP10	ENSG00000175497	G	C	299	C	missense_variant	MODERATE	1156	1018	340	A/P	Gct/Cct	rs2053724&COSV59673831
2	CLL02.1	chr2	132783534	NCKAP5	ENSG00000176771	T	A	258	A	missense_variant	MODERATE	3654	3277	1093	N/Y	Aat/Tat	rs16841277&COSV58444336
2	CLL02.1	chr2	137160352	THSD7B	ENSG00000144229	T	A	99	A	missense_variant	MODERATE	1687	1509	503	D/E	gaT/gaA	rs4954474&COSV55680598
2	CLL02.1	chr2	140371210	LRP1B	ENSG00000168702	C	G	163	G	missense_variant	MODERATE	11131	10844	3615	G/A	gGa/gCa	rs76554185&COSV67206902&COSV67273739
2	CLL02.1	chr3	78667911	ROBO1	ENSG00000169855	G	T	154	T	missense_variant	MODERATE	2235	1938	646	S/R	agC/agA	
2	CLL02.1	chr3	108470146	MYH15	ENSG00000144821	G	A	1030	A	missense_variant	MODERATE	1567	1510	504	H/Y	Cac/Tac	rs9868484&COSV56308865
2	CLL02.1	chr3	108470780	MYH15	ENSG00000144821	C	T	139	T	missense_variant	MODERATE	1418	1361	454	R/Q	cGg/cAg	rs4299484&COSV56310353
2	CLL02.1	chr3	175627355	NAALADL2	ENSG00000177694	C	G	132	G	missense_variant	MODERATE	1939	1865	622	P/R	cCc/cGc	rs9866564
2	CLL02.1	chr4	161386160	FSTL5	ENSG00000168843	C	A	164	A	missense_variant	MODERATE	2533	2131	711	D/Y	Gat/Tat	rs3749598
2	CLL02.1	chr5	22078475	CDH12	ENSG00000154162	C	T	241	T	missense_variant	MODERATE	1011	202	68	V/M	Gtg/Atg	rs4371716&CM067358&COSV66394300
2	CLL02.1	chr5	157502808	ADAM19	ENSG00000135074	C	T	893	T	missense_variant	MODERATE	1382	1303	435	E/K	Gaa/Aaa	rs56384823
2	CLL02.1	chr5	157509356	ADAM19	ENSG00000135074	T	C	138	C	missense_variant	MODERATE	929	850	284	S/G	Agt/Ggt	rs14227958&CM093420&COSV57425399
2	CLL02.1	chr6	117301070	ROS1	ENSG00000047936	C	T	115	T	missense_variant	MODERATE	6905	6619	2207	D/N	Gac/Aac	rs529038&CM053396&COSV63850799
2	CLL02.1	chr8	50553159	SNTG1	ENSG00000147481	T	C	19	C	missense_variant	MODERATE	1717	790	264	S/P	Tca/Cca	
2	CLL02.1	chr8	112228859	CSMD3	ENSG00000164796	T	G	129	G	missense_variant	MODERATE	10946	10861	3621	N/H	Aat/Cat	rs1592624&COSV52174720
2	CLL02.1	chr8	112241725	CSMD3	ENSG00000164796	A	G	201	G	missense_variant	MODERATE	10548	10463	3488	L/P	cTa/cCa	rs61753736&COSV104398217
2	CLL02.1	chr10	82979088	NRG3	ENSG00000185737	G	T	81	T	missense_variant	MODERATE	1698	1551	517	R/S	agG/agT	rs761374063&COSV64583486
2	CLL02.1	chr11	92353347	FAT3	ENSG00000165323	C	T	404	T	missense_variant	MODERATE	1609	1235	412	S/F	tCt/tTt	rs10830902
2	CLL02.1	chr13	24455080	PARP4	ENSG00000102699	C	T	426	T	missense_variant	MODERATE	2769	2695	899	A/T	Gcc/Acc	rs2275660&COSV67960444
2	CLL02.1	chr15	85580133	AKAP13	ENSG00000170776	G	A	322	A	missense_variant	MODERATE	2272	2065	689	E/K	Gag/Aag	rs71771078&COSV63448301
2	CLL02.1	chr20	14085630	MACROD2	ENSG00000172264	C	T	82	T	missense_variant	MODERATE	568	173	58	T/I	aCt/aTt	rs29900505&COSV53960729
2	CLL02.1	chr21	21373867	NCAM2	ENSG00000154654	T	C	185	C	missense_variant	MODERATE	1204	1049	350	L/P	cTg/cCg	rs232518
2	CLL02.1	chr22	46535180	CELSR1	ENSG00000075275	G	C	398	C	missense_variant	MODERATE	2441	1991	664	S/W	tCg/tGg	rs4823850&COSV53084542
2	CLL02.1	chrX	32362843	DMD	ENSG00000198947	A	G	37	G	missense_variant	MODERATE	5507	5270	1757	I/T	aTc/aCc	rs201516290&COSV63773732
2	CLL02.1	chrX	105767516	IL1RAPL2	ENSG00000189108	A	G	186	G	missense_variant	MODERATE	2788	1916	639	H/R	cAt/cGt	rs368014818
4	MBL04.1	chr1	216070175	USH2A	ENSG00000042781	T	C	248	C	missense_variant	MODERATE	6414	5975	1992	Y/C	tAt/tGt	rs41303287&CM104140&COSV56354439
4	MBL04.1	chr2	115753241	DPP10	ENSG00000175497	G	C	204	C	missense_variant	MODERATE	1156	1018	340	A/P	Gct/Cct	rs2053724&COSV59673831
4	MBL04.1	chr2	132783032	NCKAP5	ENSG00000176771	G	T	281	T	missense_variant	MODERATE	4156	3779	1260	P/Q	cCa/cAa	rs13016342&COSV58445700
4	MBL04.1	chr2	132785012	NCKAP5	ENSG00000176771	C	G	291	G	missense_variant	MODERATE	2176	1799	600	S/T	aGt/aCt	rs17325719&COSV58430764&COSV58445709
4	MBL04.1	chr2	137160352	THSD7B	ENSG00000144229	T	A	351	A	missense_variant	MODERATE	1687	1509	503	D/E	gaT/gaA	rs4954474&COSV55680598
4	MBL04.1	chr3	108470146	MYH15	ENSG00000144821	G	A	155	A	missense_variant	MODERATE	1567	1510	504	H/Y	Cac/Tac	rs9868484&COSV56308865
4	MBL04.1	chr3	108470780	MYH15	ENSG00000144821	C	T	65	T	missense_variant	MODERATE	1418	1361	454	R/Q	cGg/cAg	rs4299484&COSV56310353
4	MBL04.1	chr4	61934862	ADGRL3	ENSG00000150471	A	T	15	T	missense_variant	MODERATE	303	305	102	N/I	aAc/aTc	
4	MBL04.1	chr5	22078475	CDH12	ENSG00000154162	C	T	182	T	missense_variant	MODERATE	1011	202	68	V/M	Gtg/Atg	rs4371716&CM067358&COSV66394300
4	MBL04.1	chr5	157490332	ADAM19	ENSG00000135074	A	C	523	C	missense_variant	MODERATE	2297	2218	740	S/A	Tcc/Gcc	rs10067096&COSV57429259
4	MBL04.1	chr5	157491842	ADAM19	ENSG00000135074	C	T	378	T	missense_variant	MODERATE	2058	1979	660	G/D	gGc/gAc	rs2287749
4	MBL04.1	chr7	83155814	PCLO	ENSG00000186472	C	T	142	T	missense_variant	MODERATE	1119	827	276	R/Q	cGa/cAa	rs7475932333&COSV61667530
4	MBL04.1	chr7	158138452	PTPRN2	ENSG00000155093	C	T	439	T	missense_variant	MODERATE	1128	974	325	S/N	aGt/aAt	rs11304098&COSV67024020
4	MBL04.1	chr8	1565843	DLGAP2	ENSG00000198010	C	A	69	A	missense_variant	MODERATE	1571	1391	464	P/Q	cCa/cAa	rs2301963&COSV70102130
4	MBL04.1	chr8	112228859	CSMD3	ENSG00000164796	T	G	117	G	missense_variant	MODERATE	10946	10861	3621	N/H	Aat/Cat	rs1592624&COSV52174720
4	MBL04.1	chr10	66766407	CTNNA3	ENSG00000183230	C	T	21	T	missense_variant	MODERATE	1339	1138	380	A/T	Gct/Act	rs772731501

Family	sample	CHROM	POS	SYMBOL	Gene	REF	ALT	QUAL	Allele	Consequence	IMPACT	cDNA_position	CDS_position	Protein_position	Amino_acids	Codons	Existing_variation
4	MBL04.1	chr11	20783740	NELL1	ENSG00000165973	G	A	202	A	missense_variant	MODERATE	383	245	82	R/Q	cGg/cAg	rs8176785&CM074381
4	MBL04.1	chr11	92831816	FAT3	ENSG00000165323	G	A	242	A	missense_variant	MODERATE	10050	9676	3226	V/I	Gtt/Att	rs766505847
4	MBL04.1	chr13	24455080	PARP4	ENSG00000102699	C	T	381	T	missense_variant	MODERATE	2769	2695	899	A/T	Gcc/Acc	rs2275660&COSV67960444
4	MBL04.1	chr22	46535180	CELSR1	ENSG00000075275	G	C	330	C	missense_variant	MODERATE	2441	1991	664	S/W	tCg/tGg	rs48238508&COSV53084542
5	CLL05.1	chr4	65365968	EPHA5	ENSG00000145242	C	T	234	T	missense_variant	MODERATE	2698	1951	651	A/T	Gca/Aca	rs36050417&COSV56636694
5	CLL05.1	chr4	65490601	EPHA5	ENSG00000145242	A	G	326	G	missense_variant	MODERATE	1925	1178	393	I/T	aTt/aCt	rs780551761
5	CLL05.1	chr7	14685305	DGKB	ENSG00000136267	A	C	16	C	missense_variant	MODERATE	1116	769	257	Y/D	Tat/Gat	COSV104385499
5	CLL05.1	chr11	20928414	NELL1	ENSG00000165973	C	T	120	T	missense_variant	MODERATE	1070	932	311	P/L	cCt/cTt	rs55926004
5	CLL05.1	chr20	14085630	MACROD2	ENSG00000172264	C	T	62	T	missense_variant	MODERATE	568	173	58	T/I	aCt/aTt	rs2990505&COSV53960729
5	CLL05.1	chr22	22893818	IGLL5	ENSG00000254709	G	A	44	A	missense_variant&splice_region_variant	MODERATE	563	325	109	G/S	Ggt/Agt	rs201857114&COSV101138586&COSV66537362
5	MBL05.1	chr1	71407471	NEGR1	ENSG00000172260	T	G	129	G	missense_variant	MODERATE	1085	1040	347	Y/S	tAc/tCc	rs41289154&COSV60833851&COSV60866382
5	MBL05.1	chr6	117301070	ROS1	ENSG00000047936	C	T	155	T	missense_variant	MODERATE	6905	6619	2207	D/N	Gac/Aac	rs529038&CM053396&COSV63850799
5	MBL05.1	chr7	158167251	PTPRN2	ENSG00000155093	G	C	166	C	missense_variant	MODERATE	744	590	197	A/G	gCc/gGc	rs140999965
6	CLL06.1	chr1	71407471	NEGR1	ENSG00000172260	T	G	138	G	missense_variant	MODERATE	1085	1040	347	Y/S	tAc/tCc	rs41289154&COSV60833851&COSV60866382
6	CLL06.1	chr2	115746095	DPP10	ENSG00000175497	G	A	47	A	missense_variant	MODERATE	1000	862	288	V/M	Gtg/Atg	rs36044503
6	CLL06.1	chr2	115753241	DPP10	ENSG00000175497	G	C	173	C	missense_variant	MODERATE	1156	1018	340	A/P	Gct/Cct	rs2053724&COSV59673831
6	CLL06.1	chr2	132783032	NCKAP5	ENSG00000176771	G	T	452	T	missense_variant	MODERATE	4156	3779	1260	P/Q	cCa/cAa	rs13016342&COSV58445700
6	CLL06.1	chr2	132785012	NCKAP5	ENSG00000176771	C	G	631	G	missense_variant	MODERATE	2176	1799	600	S/T	aGt/aCt	rs17325719&COSV58430764&COSV58445709
6	CLL06.1	chr2	141810341	LRP1B	ENSG00000168702	T	C	134	C	missense_variant	MODERATE	430	143	48	Q/R	cAg/cGg	rs12990449&COSV67182532
6	CLL06.1	chr3	78635874	ROBO1	ENSG00000169855	C	T	107	T	missense_variant	MODERATE	3569	3272	1091	S/N	aGc/aAc	rs35456279&COSV71397020
6	CLL06.1	chr3	108470146	MYH15	ENSG00000144821	G	A	1153	A	missense_variant	MODERATE	1567	1510	504	H/Y	Cac/Tac	rs9868484&COSV56308865
6	CLL06.1	chr3	175627355	NAALADL2	ENSG00000177694	C	G	273	G	missense_variant	MODERATE	1939	1865	622	P/R	cCc/cGc	rs9866564
6	CLL06.1	chr4	61979743	ADGRL3	ENSG00000150471	C	G	138	G	missense_variant	MODERATE	1154	1156	386	L/V	Ctg/Gtg	rs12509110&COSV72232449&COSV99072504
6	CLL06.1	chr5	22078475	CDH12	ENSG00000154162	C	T	416	T	missense_variant	MODERATE	1011	202	68	V/M	Gtg/Atg	rs4371716&CM067358&COSV66394300
6	CLL06.1	chr6	117301070	ROS1	ENSG00000047936	C	T	74	T	missense_variant	MODERATE	6905	6619	2207	D/N	Gac/Aac	rs529038&CM053396&COSV63850799
6	CLL06.1	chr7	149015000	PDIA4	ENSG00000155660	G	A	605	A	missense_variant	MODERATE	615	518	173	T/M	aCg/aTg	rs2290971&CM1511790
6	CLL06.1	chr7	158167203	PTPRN2	ENSG00000155093	C	T	30	T	missense_variant	MODERATE	792	638	213	R/H	cGc/cAc	rs1130496&COSV67046421
6	CLL06.1	chr8	1565843	DLGAP2	ENSG00000198010	C	A	16	A	missense_variant	MODERATE	1571	1391	464	P/Q	cCa/cAa	rs2301963&COSV70102130
6	CLL06.1	chr8	112228859	CSMD3	ENSG00000164796	T	G	258	G	missense_variant	MODERATE	10946	10861	3621	N/H	Aat/Cat	rs1592624&COSV52174720
6	CLL06.1	chr11	20783740	NELL1	ENSG00000165973	G	A	70	A	missense_variant	MODERATE	383	245	82	R/Q	cGg/cAg	rs8176785&CM074381
6	CLL06.1	chr13	24455080	PARP4	ENSG00000102699	C	T	639	T	missense_variant	MODERATE	2769	2695	899	A/T	Gcc/Acc	rs2275660&COSV67960444
6	CLL06.1	chr14	92651884	RIN3	ENSG00000100599	C	T	34	T	missense_variant	MODERATE	987	835	279	R/C	Cgc/Tgc	rs117068593&COSV53649944
6	CLL06.1	chr15	85580133	AKAP13	ENSG00000170776	G	A	546	A	missense_variant	MODERATE	2272	2065	689	E/K	Gag/Aag	rs71771078&COSV63448301
6	CLL06.1	chr15	85735078	AKAP13	ENSG00000170776	G	A	206	A	missense_variant	MODERATE	7576	7369	2457	G/S	Ggt/Agt	rs2241268&COSV63448267
6	CLL06.1	chr17	7670613	TP53	ENSG00000141510	A	C	178	C	missense_variant	MODERATE	1238	1096	366	S/A	Tcc/Gcc	rs17881470&CM078492&COSV52875421
6	CLL06.1	chr20	14085630	MACROD2	ENSG00000172264	C	T	80	T	missense_variant	MODERATE	568	173	58	T/I	aCt/aTt	rs2990505&COSV53960729
6	CLL06.1	chr21	21373867	NCAM2	ENSG00000154654	T	C	315	C	missense_variant	MODERATE	1204	1049	350	L/P	cTg/cCg	rs232518
6	CLL06.1	chr21	40353698	DSCAM	ENSG00000171587	C	T	41	T	missense_variant	MODERATE	1198	701	234	R/H	cGc/cAc	rs41395652
6	CLL06.1	chr22	46535180	CELSR1	ENSG00000075275	G	C	520	C	missense_variant	MODERATE	2441	1991	664	S/W	tCg/tGg	rs48238508&COSV53084542
21	CLL21.1	chr2	115753241	DPP10	ENSG00000175497	G	C	162	C	missense_variant	MODERATE	1156	1018	340	A/P	Gct/Cct	rs2053724&COSV59673831
21	CLL21.1	chr2	132783032	NCKAP5	ENSG00000176771	G	T	905	T	missense_variant	MODERATE	4156	3779	1260	P/Q	cCa/cAa	rs13016342&COSV58445700
21	CLL21.1	chr2	132785012	NCKAP5	ENSG00000176771	C	G	1173	G	missense_variant	MODERATE	2176	1799	600	S/T	aGt/aCt	rs17325719&COSV58430764&COSV58445709
21	CLL21.1	chr2	137160352	THSD7B	ENSG00000144229	T	A	843	A	missense_variant	MODERATE	1687	1509	503	D/E	gaT/gaA	rs4954474&COSV55680598
21	CLL21.1	chr2	141810341	LRP1B	ENSG00000168702	T	C	560	C	missense_variant	MODERATE	430	143	48	Q/R	cAg/cGg	rs12990449&COSV67182532
21	CLL21.1	chr2	213930019	SPAG16	ENSG00000144451	A	C	593	C	missense_variant	MODERATE	1294	1274	425	K/T	aAa/aCa	rs12623569&COSV59078990
21	CLL21.1	chr3	108470146	MYH15	ENSG00000144821	G	A	759	A	missense_variant	MODERATE	1567	1510	504	H/Y	Cac/Tac	rs9868484&COSV56308865
21	CLL21.1	chr3	108470780	MYH15	ENSG00000144821	C	T	557	T	missense_variant	MODERATE	1418	1361	454	R/Q	cGg/cAg	rs4299484&COSV56310353
21	CLL21.1	chr3	175627355	NAALADL2	ENSG00000177694	C	G	657	G	missense_variant	MODERATE	1939	1865	622	P/R	cCc/cGc	rs9866564
21	CLL21.1	chr5	157509356	ADAM19	ENSG00000135074	T	C	1981	C	missense_variant	MODERATE	929	850	284	S/G	Agt/Ggt	rs1422795&CM093420&COSV57425399
21	CLL21.1	chr6	26234599	H1-3	ENSG00000124575	G	C	49	C	missense_variant	MODERATE	389	335	112	A/G	gCg/gGg	rs140433263
21	CLL21.1	chr6	27866899	H1-5	ENSG00000184357	C	T	259	T	missense_variant	MODERATE	690	631	211	A/T	Gca/Aca	rs34144478
21	CLL21.1	chr6	117301070	ROS1	ENSG00000047936	C	T	240	T	missense_variant	MODERATE	6905	6619	2207	D/N	Gac/Aac	rs529038&CM053396&COSV63850799
21	CLL21.1	chr7	158138452	PTPRN2	ENSG00000155093	C	T	551	T	missense_variant	MODERATE	1128	974	325	S/N	aGt/aAt	rs1130498&COSV67024020
21	CLL21.1	chr8	1565843	DLGAP2	ENSG00000198010	C	A	547	A	missense_variant	MODERATE	1571	1391	464	P/Q	cCa/cAa	rs2301963&COSV70102130
21	CLL21.1	chr11	20783740	NELL1	ENSG00000165973	G	A	2263	A	missense_variant	MODERATE	383	245	82	R/Q	cGg/cAg	rs8176785&CM074381
21	CLL21.1	chr11	108310218	ATM	ENSG00000149311	G	C	283	C	missense_variant	MODERATE	5971	5821	1941	V/L	Gtt/Ctt	rs147187700&CM094667&COSV53739000&COSV53750671
21	CLL21.1	chr13	24455080	PARP4	ENSG00000102699	C	T	616	T	missense_variant	MODERATE	2769	2695	899	A/T	Gcc/Acc	rs2275660&COSV67960444

Family	sample	CHROM	POS	SYMBOL	Gene	REF	ALT	QUAL	Allele	Consequence	IMPACT	cDNA_position	CDS_position	Protein_position	Amino_acids	Codons	Existing_variation
21	CLL21.1	chr15	85735078	AKAP13	ENSG00000170776	G	A	417	A	missense_variant	MODERATE	7576	7369	2457	G/S	Ggt/Agt	rs2241268&COSV63448267
21	CLL21.1	chr15	86264612	AGBL1	ENSG00000273540	T	C	696	C	missense_variant	MODERATE	1543	1441	481	S/P	Tcc/Ccc	rs11857527&COSV66866512
21	CLL21.1	chr21	21373867	NCAM2	ENSG00000154654	T	C	1106	C	missense_variant	MODERATE	1204	1049	350	L/P	cTg/cCg	rs232518
21	CLL21.1	chr22	46535180	CELSR1	ENSG00000075275	G	C	1999	C	missense_variant	MODERATE	2441	1991	664	S/W	tCg/tGg	rs4823850&COSV53084542
22	MBL22.1	chr1	81907079	ADGRL2	ENSG00000117114	T	A	18	A	missense_variant	MODERATE	352	136	46	S/T	Tct/Act	
22	MBL22.1	chr2	213930019	SPAG16	ENSG00000144451	A	C	181	C	missense_variant	MODERATE	1294	1274	425	K/T	aAa/aCa	rs12623569&COSV59078990
22	MBL22.1	chr8	1565843	DLGAP2	ENSG00000198010	C	A	343	A	missense_variant	MODERATE	1571	1391	464	P/Q	cCa/cAa	rs2301963&COSV7010213 0
22	MBL22.1	chr8	112228859	CSMD3	ENSG00000164796	T	G	420	G	missense_variant	MODERATE	10946	10861	3621	N/H	Aat/Cat	rs1592624&COSV52174720
22	MBL22.1	chr13	94382448	GPC6	ENSG00000183098	A	G	26	G	missense_variant	MODERATE	1837	1187	396	K/R	aAa/aGa	
22	MBL22.1	chr21	40338287	DSCAM	ENSG00000171587	A	G	15	G	missense_variant	MODERATE	2094	1597	533	Y/H	Tac/Cac	
23	CLL23.1	chr5	157509356	ADAM19	ENSG00000135074	T	C	375	C	missense_variant	MODERATE	929	850	284	S/G	Agt/Ggt	rs1422795&CM093420&COSV57425399
23	CLL23.1	chr14	92651884	RIN3	ENSG00000100599	C	T	773	T	missense_variant	MODERATE	987	835	279	R/C	Cgc/Tgc	rs117068593&COSV53649944
23	CLL23.1	chr20	42104665	PTPRT	ENSG00000196090	G	T	825	T	missense_variant	MODERATE	3471	3471	1157	N/K	aaC/aaA	rs201162919&COSV61980909
23	CLL23.1	chr21	40353698	DSCAM	ENSG00000171587	C	T	210	T	missense_variant	MODERATE	1198	701	234	R/H	cGc/cAc	rs41395652
23	MBL23.1	chr1	237566732	RYR2	ENSG00000198626	A	G	624	G	missense_variant	MODERATE	3718	3380	1127	E/G	gAg/gGg	rs200525962
23	MBL23.1	chr4	8218302	SH3TC1	ENSG00000125089	G	A	238	A	missense_variant	MODERATE	1015	871	291	D/N	Gac/Aac	rs1281138
23	MBL23.1	chr4	8240900	SH3TC1	ENSG00000125089	C	A	239	A	missense_variant	MODERATE	4100	3956	1319	P/H	cCt/cAt	rs150702535
23	MBL23.1	chr6	27866899	H1-5	ENSG00000184357	C	T	278	T	missense_variant	MODERATE	690	631	211	A/T	Gca/Aca	rs34144478
23	MBL23.1	chr7	154052980	DPP6	ENSG00000130226	C	T	21	T	missense_variant	MODERATE	583	160	54	R/W	Cgg/Tgg	COSV66748904
24	MBL24.1	chr1	215728169	USH2A	ENSG00000042781	G	A	254	A	missense_variant	MODERATE	12366	11927	3976	T/M	aCg/aTg	rs142381713&CM073404&COSV56393885
24	MBL24.1	chr5	157509356	ADAM19	ENSG00000135074	T	C	353	C	missense_variant	MODERATE	929	850	284	S/G	Agt/Ggt	rs1422795&CM093420&COSV57425399
24	MBL24.1	chr6	117362658	ROS1	ENSG00000047936	G	A	555	A	missense_variant	MODERATE	3597	3311	1104	S/L	tCa/tTa	rs2229079&COSV63860865
24	MBL24.1	chr6	117404285	ROS1	ENSG00000047936	T	G	450	G	missense_variant	MODERATE	746	460	154	T/P	Act/Cct	rs1998206&COSV63854406
24	MBL24.1	chr11	92837690	FAT3	ENSG00000165323	G	T	422	T	missense_variant	MODERATE	10626	10252	3418	A/S	Gcc/Tcc	rs201449521&COSV53125665&COSV99973667
24	MBL24.1	chr13	24442624	PARP4	ENSG00000102699	G	A	74	A	missense_variant	MODERATE	3583	3509	1170	T/I	aCa/aTa	rs113301501&CM160622&COSV67960155
24	MBL24.1	chr13	24447185	PARP4	ENSG00000102699	A	G	87	G	missense_variant&splice_region_variant	MODERATE	3190	3116	1039	I/T	aTa/aCa	rs73172125&COSV67960144
24	MBL24.1	chr14	46957592	MDGA2	ENSG00000139915	C	T	632	T	missense_variant	MODERATE	1612	977	326	R/Q	cGa/cAa	rs553082374&COSV62105261
24	MBL24.1	chr17	7674917	TP53	ENSG00000141510	T	C	993	C	missense_variant	MODERATE	756	614	205	Y/C	tAt/tGt	rs1057520007&COSV52665440&COSV52677268&COSV52688506
24	MBL24.1	chr22	22888193	IGLL5	ENSG00000254709	G	A	243	A	missense_variant	MODERATE	378	140	47	S/N	aGc/aAc	rs540568540&COSV73249672
24	MBL24.2	chr2	197402110	SF3B1	ENSG00000115524	T	C	842	C	missense_variant	MODERATE	2127	2098	700	K/E	Aaa/Gaa	rs559063155&COSV59205318&COSV59216586
24	MBL24.2	chr3	108470780	MYH15	ENSG00000144821	C	T	665	T	missense_variant	MODERATE	1418	1361	454	R/Q	cGg/cAg	rs4299484&COSV63610353
24	MBL24.3	chr1	215647626	USH2A	ENSG00000042781	A	T	27	T	missense_variant	MODERATE	15126	14687	4896	L/H	cTt/cAt	
24	MBL24.3	chr1	215790251	USH2A	ENSG00000042781	A	C	15	C	missense_variant	MODERATE	10429	9990	3330	N/K	aaT/aaG	
24	MBL24.3	chr1	237569176	RYR2	ENSG00000198626	A	G	23	G	missense_variant	MODERATE	3793	3455	1152	Y/C	tAt/tGt	
24	MBL24.3	chr1	237674166	RYR2	ENSG00000198626	A	T	61	T	missense_variant	MODERATE	8999	8661	2887	E/D	gaA/gaT	
24	MBL24.3	chr2	132783534	NCKAP5	ENSG00000176771	T	A	1027	A	missense_variant	MODERATE	3654	3277	1093	N/Y	Aat/Tat	rs16841277&COSV58444336
24	MBL24.3	chr2	141062161	LRP1B	ENSG00000168702	C	A	18	A	missense_variant	MODERATE	1413	1126	376	A/S	Gca/Tca	
24	MBL24.3	chr2	197396311	SF3B1	ENSG00000115524	G	T	34	T	missense_variant	MODERATE	3313	3284	1095	A/D	gCt/gAt	
24	MBL24.3	chr3	108410799	MYH15	ENSG00000144821	C	A	16	A	missense_variant	MODERATE	4396	4339	1447	V/F	Gtc/Ttc	COSV56321473
24	MBL24.3	chr5	21783409	CDH12	ENSG00000154162	C	T	33	T	missense_variant	MODERATE	2151	1342	448	D/N	Gac/Aac	
24	MBL24.3	chr5	59893589	PDE4D	ENSG00000113448	C	G	250	G	missense_variant	MODERATE	138	34	12	A/P	Gcg/Ccg	
24	MBL24.3	chr7	81991211	CACNA2D1	ENSG00000153956	C	T	36	T	stop_gained	HIGH	2088	1770	590	W/*	tgG/tgA	rs1554341873
24	MBL24.3	chr7	147132332	CNTNAP2	ENSG00000174469	A	T	31	T	missense_variant	MODERATE	1247	1171	391	N/Y	Aac/Tac	COSV62189039
24	MBL24.3	chr7	148147676	CNTNAP2	ENSG00000174469	A	G	25	G	missense_variant	MODERATE	2816	2740	914	T/A	Acc/Gcc	
24	MBL24.3	chr8	2978629	CSMD1	ENSG00000183117	G	A	36	A	missense_variant	MODERATE	9047	8549	2850	S/F	tCc/tTc	
24	MBL24.3	chr8	14324194	SGCZ	ENSG00000185053	C	A	26	A	missense_variant	MODERATE	1053	245	82	G/V	gGa/gTa	COSV65998392&COSV66047381
24	MBL24.3	chr9	117214604	ASTN2	ENSG00000148219	A	T	3070	T	missense_variant	MODERATE	888	769	257	S/T	Tct/Act	rs139148246
24	MBL24.3	chr11	92837708	FAT3	ENSG00000165323	C	G	175	G	missense_variant	MODERATE	10644	10270	3424	P/A	Cct/Gct	
24	MBL24.3	chr13	93830302	GPC6	ENSG00000183098	T	G	63	G	missense_variant	MODERATE	1118	468	156	N/K	aaT/aaG	
24	MBL24.3	chr14	61485534	PRKCH	ENSG00000027075	T	G	30	G	missense_variant	MODERATE	1693	1311	437	N/K	aaT/aaG	
24	MBL24.3	chr15	85645868	AKAP13	ENSG00000170776	C	A	29	A	missense_variant	MODERATE	4495	4288	1430	Q/K	Caa/Aaa	
24	MBL24.3	chr18	21384368	GREB1L	ENSG00000141449	A	C	149	C	missense_variant	MODERATE	475	320	107	Q/P	cAa/cCa	rs1425668691
24	MBL24.3	chr21	40187251	DSCAM	ENSG00000171587	C	A	30	A	missense_variant	MODERATE	3156	2659	887	D/Y	Gac/Tac	
24	MBL24.3	chr22	46386404	CELSR1	ENSG00000075275	A	G	505	G	missense_variant&splice_region_variant	MODERATE	7187	6737	2246	M/T	aTg/aCg	rs749104236
24	MBL24.3	chr22	46537122	CELSR1	ENSG00000075275	C	G	27	G	missense_variant	MODERATE	499	49	17	A/P	Gcc/Ccc	
27	CLL27.1	chr1	81990901	ADGRL2	ENSG00000117114	G	A	551	A	missense_variant	MODERATE	4184	3968	1323	R/K	aGa/aAa	rs41292984&COSV54685119
27	CLL27.1	chr2	115746095	DPP10	ENSG00000175497	G	A	542	A	missense_variant	MODERATE	1000	862	288	V/M	Gtg/Atg	rs36044503

Family	sample	CHROM	POS	SYMBOL	Gene	REF	ALT	QUAL	Allele	Consequence	IMPACT	cDNA_position	CDS_position	Protein_position	Amino_acids	Codons	Existing_variation
	27 CLL27.1	chr2	197402635	SF3B1	ENSG00000115524	C	A	139	A	missense_variant	MODERATE	2027	1998	666	K/N	aaG/aaT	rs377023736&COSV59205777&COSV59205799
	27 CLL27.1	chr8	1565843	DLGAP2	ENSG00000198010	C	A	572	A	missense_variant	MODERATE	1571	1391	464	P/Q	cCa/cAa	rs230196338&COSV70102130
	27 CLL27.1	chr10	115334318	ATRNL1	ENSG00000107518	A	C	923	C	missense_variant	MODERATE	3460	3074	1025	N/T	aAt/aCt	rs1050159510
	27 CLL27.1	chr11	99819645	CNTN5	ENSG00000149972	C	T	371	T	stop_gained	HIGH	688	157	53	R/*	cGa/Tga	rs122926598&COSV54311010
	27 MBL27.1	chr4	161386160	FSTL5	ENSG00000168843	C	A	810	A	missense_variant	MODERATE	2533	2131	711	D/Y	Gat/Tat	rs3749598
	27 MBL27.1	chr9	136515524	NOTCH1	ENSG00000148400	C	T	117	T	missense_variant	MODERATE	2124	1862	621	R/H	cGc/cAc	rs138504021&COSV99487877
	27 MBL27.1	chr11	83962888	DLG2	ENSG00000150672	G	A	769	A	missense_variant	MODERATE	1723	1337	446	T/I	aCc/aTc	rs140599607
	27 MBL27.2	chr1	21579051	USH2A	ENSG00000042781	G	T	1925	T	missense_variant	MODERATE	10253	9814	3272	P/T	Ccg/Acg	COSV100239515
	27 MBL27.2	chr3	17573328	NAALADL2	ENSG00000177694	A	G	571	G	missense_variant	MODERATE	1993	1919	640	Q/R	cAa/cGa	rs62286105
	27 MBL27.2	chr9	136523093	NOTCH1	ENSG00000148400	G	A	15	A	missense_variant	MODERATE	761	499	167	P/S	Cca/Tca	rs143903832&COSV53077398
	27 MBL27.2	chr13	24447125	PARP4	ENSG00000102699	T	C	54	C	missense_variant	MODERATE	3250	3176	1059	Q/R	cAg/cGg	rs77269056&COSV67960139
	27 MBL27.2	chr13	24447185	PARP4	ENSG00000102699	A	G	194	G	missense_variant&splice_region_variant	MODERATE	3190	3116	1039	I/T	aTa/aCa	rs731721258&COSV67960144
	28 CLL28.1	chr2	61492051	XPO1	ENSG00000082898	T	C	209	C	missense_variant	MODERATE	2672	1871	624	D/G	gAt/gGt	COSV68944745
	28 CLL28.1	chr3	54569985	CACNA2D3	ENSG00000157445	G	A	672	A	missense_variant	MODERATE	931	769	257	V/M	Gtg/Atg	rs372465588&COSV55530096&COSV55536891
	28 CLL28.1	chr5	157490332	ADAM19	ENSG00000135074	A	C	836	C	missense_variant	MODERATE	2297	2218	740	S/A	Tcc/Gcc	rs10067096&COSV57429259
	28 CLL28.1	chr5	157509356	ADAM19	ENSG00000135074	T	C	462	C	missense_variant	MODERATE	929	850	284	S/G	Agt/Ggt	rs142279558&CM093420&COSV57425399
	28 CLL28.1	chr5	157575687	ADAM19	ENSG00000135074	C	T	116	T	missense_variant	MODERATE	89	10	4	G/S	Ggc/Agc	rs11465228
	28 CLL28.1	chr8	112636939	CSMD3	ENSG00000164796	A	G	695	G	missense_variant	MODERATE	3678	3593	1198	F/S	tTt/tCt	rs765163979
	28 CLL28.1	chr15	85735078	AKAP13	ENSG00000170776	G	A	413	A	missense_variant	MODERATE	7576	7369	2457	G/S	Ggt/Agt	rs224126&COSV63448267
	28 MBL28.1	chr2	211630569	ERBB4	ENSG00000178568	T	A	238	A	missense_variant	MODERATE	2244	1972	658	I/F	Att/Ttt	rs190654033
	28 MBL28.1	chr7	158138452	PTPRN2	ENSG00000155093	C	T	1271	T	missense_variant	MODERATE	1128	974	325	S/N	aGt/aAt	rs113049&COSV67024020
	28 MBL28.1	chr11	92797959	FAT3	ENSG00000165323	C	T	638	T	missense_variant	MODERATE	5320	4946	1649	P/L	cCg/cTg	rs375193261&COSV104590304&COSV53099591
	28 MBL28.1	chr13	24447185	PARP4	ENSG00000102699	A	G	21	G	missense_variant&splice_region_variant	MODERATE	3190	3116	1039	I/T	aTa/aCa	rs731721258&COSV67960144
	28 MBL28.1	chr13	91693325	GPC5	ENSG00000179399	C	T	367	T	missense_variant	MODERATE	890	464	155	A/V	gCg/gTg	rs553717&COSV65586039
	28 MBL28.1	chr15	85580133	AKAP13	ENSG00000170776	G	A	764	A	missense_variant	MODERATE	2272	2065	689	E/K	Gag/Aag	rs71771078&COSV63448301
	30 CLL30.1	chr2	124869717	CNTNAP5	ENSG00000155052	G	A	535	A	missense_variant	MODERATE	3752	3388	1130	V/I	Gta/Ata	rs1240405090
	30 CLL30.1	chr11	100061213	CNTN5	ENSG00000149972	C	G	145	G	missense_variant&splice_region_variant	MODERATE	1513	982	328	P/A	Ccc/Gcc	rs201910584&COSV54330963
	30 CLL30.1	chr13	24447185	PARP4	ENSG00000102699	A	G	63	G	missense_variant&splice_region_variant	MODERATE	3190	3116	1039	I/T	aTa/aCa	rs731721258&COSV67960144
	30 CLL30.1	chr20	14085630	MACROD2	ENSG00000172264	C	T	133	T	missense_variant	MODERATE	568	173	58	T/I	aCt/aTt	rs29905058&COSV53960729
	30 CLL30.1	chr22	22888259	IGLL5	ENSG00000254709	G	A	162	A	missense_variant&splice_region_variant	MODERATE	444	206	69	R/K	aGg/aAg	rs74548574&COSV73251123
	30 MBL30.2	chr2	211722434	ERBB4	ENSG00000178568	G	T	367	T	missense_variant	MODERATE	1114	842	281	A/E	gCa/gAa	
	30 MBL30.2	chr7	158138452	PTPRN2	ENSG00000155093	C	T	142	T	missense_variant	MODERATE	1128	974	325	S/N	aGt/aAt	rs113049&COSV67024020
	30 MBL30.2	chr15	85735078	AKAP13	ENSG00000170776	G	A	1813	A	missense_variant	MODERATE	7576	7369	2457	G/S	Ggt/Agt	rs224126&COSV63448267
	30 MBL30.1	chr1	81990901	ADGRL2	ENSG00000117114	G	A	326	A	missense_variant	MODERATE	4184	3968	1323	R/K	aGa/aAa	rs412929&COSV54685119
	30 MBL30.1	chr2	124524333	CNTNAP5	ENSG00000155052	C	T	416	T	missense_variant	MODERATE	1719	1355	452	S/L	tCg/tTg	rs177272618&COSV70497527
	30 MBL30.1	chr2	137160352	THSD7B	ENSG00000144229	T	A	1207	A	missense_variant	MODERATE	1687	1509	503	D/E	gaT/gaA	rs495447&COSV55680598
	30 MBL30.1	chr13	24442624	PARP4	ENSG00000102699	G	A	230	A	missense_variant	MODERATE	3583	3509	1170	T/I	aCa/aTa	rs113301501&CM160622&COSV67960155
	30 MBL30.1	chr18	21473100	GREBL1	ENSG00000141449	G	A	420	A	missense_variant	MODERATE	2080	1925	642	R/H	cGt/cAt	rs13435799561
	30 MBL30.1	chr22	46364248	CELSR1	ENSG00000075275	A	C	22	C	missense_variant	MODERATE	9233	8783	2928	I/S	aTc/aGc	

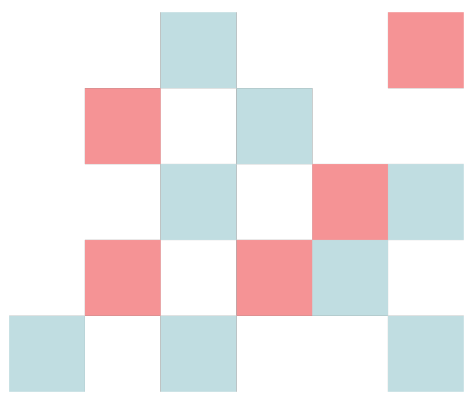
Supplementary Table 4. Non-shared variants in 120 CLL-related genes.

Monte Tarn, Chile

©JSY

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

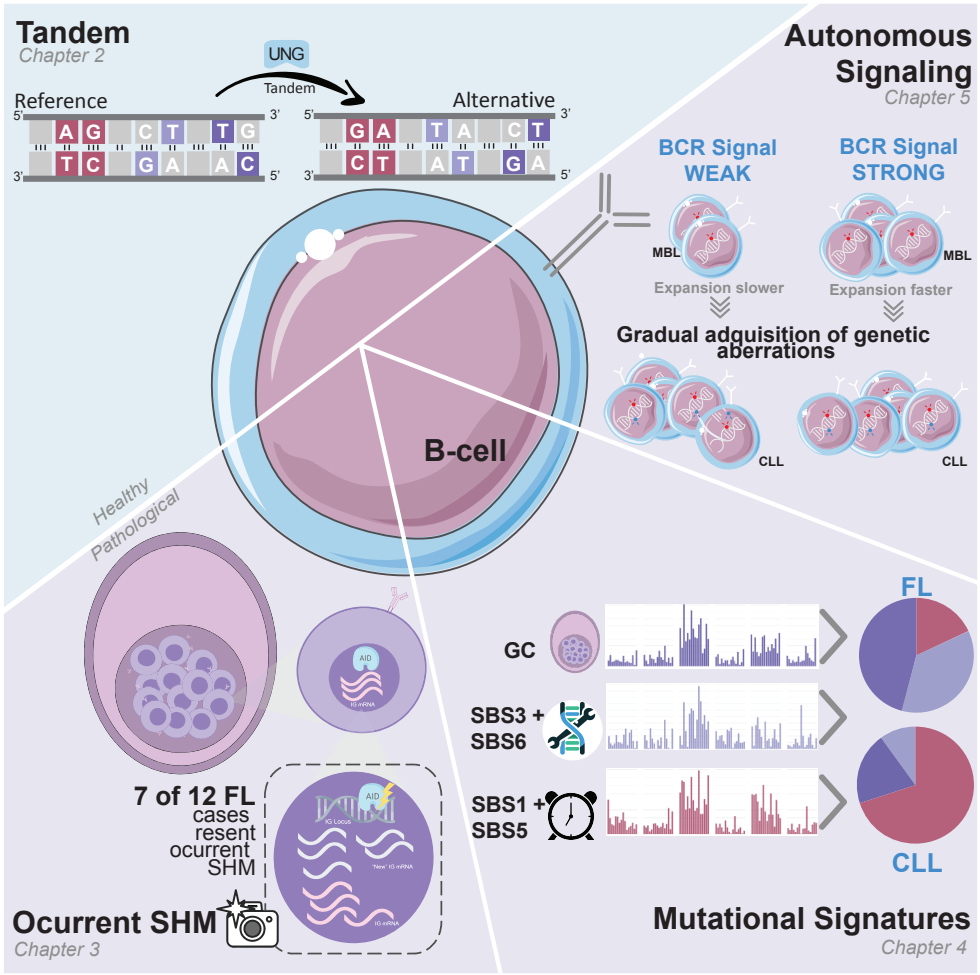


Figure 6.1. Integrated Overview of the Thesis. From Characterization of AID-Induced Mutations to the Oncogenic Processes in B-Cell Neoplasias. Chapter 2 discusses the characterization of tandem DNA substitutions in immunoglobulin genes. Chapter 3 investigates AID-associated mutations at the single-cell 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.

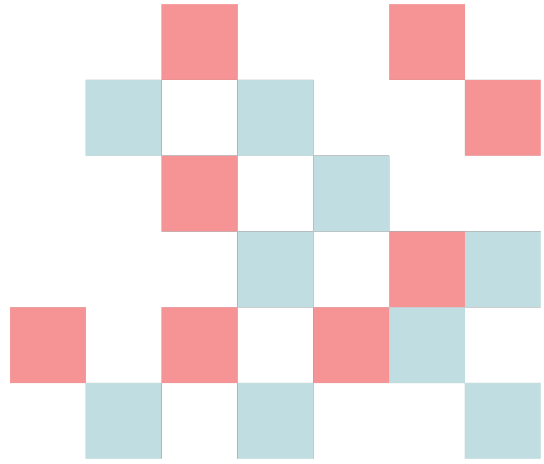
Our mutational signature analysis was limited to FL and CLL samples. Expanding this

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.

Estrecho de Magallanes, Chile

APPENDIX



NEDERLANDSE SAMENVATTING

ANTILICHAAM DIVERSITEIT

Het adaptieve immuunsysteem is een essentieel onderdeel van onze afweer en is in staat om heel gericht te beschermen tegen onder andere schadelijke indringers van buitenaf. Immunoglobulinen, ofwel antilichamen, spelen hierin een cruciale rol door te binden aan structuren (antigenen) van ziekteverwekkers en ze zo onschadelijk te maken. Immunoglobulinen zijn gespecialiseerde eiwitten die worden geproduceerd door B-cellen en tot expressie worden gebracht op het oppervlak als een B-cel receptor (BCR). De B-cel kan uitrijpen tot plasmacel die de antilichamen produceert en uitscheidt en zo zorg draagt voor het elimineren van de ziekteverwekker. Daarnaast kan de B-cel een memory cel worden die in geval van een her-infectie direct overgaat tot snelle productie van effectieve antilichamen. De specificiteit van immunoglobulinen om antigenen te herkennen wordt bepaald door twee verschillende processen. Tijdens de ontwikkeling van elke B-cel vormen combinaties van verschillende gen segmenten de zware en lichte keten van de BCR. Dit proces wordt VDJ recombinatie genoemd en plaatst een variabel (V), diversiteit (D) en joining (J) segment na elkaar zodat een aaneengesloten VDJ gen ontstaat. Tijdens deze recombinatie dragen zowel combinaties van willekeurige gen segmenten als variaties die ontstaan bij het aan elkaar koppelen van de gen segmenten bij aan de diversiteit van immunoglobulinen. Vanuit deze diversiteit kan een veelheid van antigenen tot activatie van de B-cel kan leiden. Ten tweede kan de bindingscapaciteit van immuunglobulines verbeteren, ofwel rijpen, door het verkrijgen van mutaties. Dit proces vindt plaats onder zeer strenge regulatie in kiemcentra van reactieve lymfeklieren en heet somatische hypermutatie (SHM). SHM wordt geïnitieerd door het enzym activation-induced cytidine deaminase (AID). AID deamineert de nucleobase cytidine (C) en heeft een voorkeur voor specifieke motieven in het DNA waaronder het canonieke WRCY motief en de niet-canonieke WA, en RCG motieven. De letters W, R en Y staan voor combinaties van meerdere nucleobasen.

Het fysiologische mutagene AID-proces splitst een amine-groep af van cytosine waardoor de nucleobase uracil ontstaat. Dit enzymatische proces is uiterst efficiënt omdat de hoge affiniteit van AID voor enkelstrengs DNA een bindingsduur van meer dan 5 minuten waarborgt. Daardoor kan AID meerdere deaminaties per keer uitvoeren, wat de incidentie van mutaties verhoogt en uiteindelijk een snelle rijping van de antilichamen veroorzaakt. De aanwezigheid van de nieuwe uracil nucleobase tegenover guanine is een DNA-mismatch en moet daarom hersteld worden. De B cel heeft hiervoor twee mechanismen tot haar beschikking.

DNA-HERSTELMECHANISMEN

Base Excision Repair (BER) en Mismatch Repair (MMR) zijn twee herstelmechanismen die betrokken zijn bij het repareren van DNA-schade, zoals de door AID veroorzaakte deaminatie van cytosine naar uracil. Beide mechanismen zijn betrokken op een manier waarbij fout-gevoelige DNA-polymerasen betrokken zijn. Het moleculaire mechanisme van BER

bevat UNG2 of SMUG, die delen van het DNA waarin uracil voorkomt, herkennen en losmaken uit het dubbelstrengs DNA waarna uracil weggeknipt kan worden. De deletie die daardoor ontstaat activeert andere componenten van BER, waarbij voor kleine en grote deleties verschillende herstel-routes worden bewandeld. Polymerasen zoals Pol-ζ en Pol-ι vullen de deletie op met nieuwe basen maar maken daarbij fouten. Een ander mechanisme dat uracil-guanine mismatches corrigeert is MMR. De uracil base wordt herkend door de eiwitten MSH2/MSH6 waarna exonuclease EXO1 een grote deletie veroorzaakt. Vervolgens wordt PCNA gemonoubiquitinileerd, wat vervolgens foutgevoelige translesie synthese (TLS) polymerasen zoals Pol-η, Pol-ζ en REV1 bij de reactie betrekt die de deletie opvullen. Bij normale immuun reacties zijn de fouten die BER en MMR tijdens dit DNA herstel achterlaten essentieel voor de B-cel antilichaam rijping en diversificatie van het BCR repertoire.

HEMATOLOGISCHE NEOPLASMA'S

Folliculair lymfoom (FL) is een indolente hematologische B-cel maligniteit met een incidentie van 2-4 per 100.000 mensen per jaar. Hoewel de term 'indolent' inhoudt dat progressie van de ziekte een traag verloop kent, blijft FL een ongeneselijke vorm van lymfeklierkanker. Daarnaast transformeert de ziekte op jaarbasis bij 1-2% van de patiënten naar een moeilijk behandelbaar agressief lymfoom. Een kenmerk van FL is de translocatie van BCL2 op chromosoom 18 naar het immunoglobuline zware keten locus op chromosoom 14, die op kan treden tijdens de V(D)J recombinatie. Deze translocatie leidt tot overexpressie van BCL2 waardoor de cel ontsnapt aan normale mechanismen die bedoeld zijn om ongewenste cellen te elimineren (apoptose). Op zichzelf is dit niet voldoende om de ziekte te veroorzaken maar door de aanwezigheid in kiemcentrum reacties en de doorlopende activiteit van AID, heeft FL een hoog niveau van SHM. Hierdoor blijven B-cellen met een lage tot gemiddelde immunoglobuline-affiniteit, die normaal gesproken in apoptose zouden gaan, in leven. Zo dragen ze bij aan de vroege vorming van lymfoom. De doorlopende activiteit van AID resulteert ook in mutaties in andere genen dan immunoglobulinen, wat in feite abnormale SHM is. Deze mutaties treffen vaak genen die geassocieerd zijn met signaling pathways, immuun herkenning, celcyclus regulatoren en epigenetica en met name de opeenstapeling deze afwijkingen is kenmerkend voor FL.

Een andere belangrijke factor bij de ontwikkeling van FL is het tumor micro-milieu (tumor micro-environment, TME). In tegenstelling tot andere lymfomen behoudt FL de normale 3-dimensionale morfologische kenmerken van een kiemcentrum reactie zonder zich diffuus te verspreiden door de lymfeklier. FL lijkt daarmee sterk afhankelijk van componenten van het TME zoals andere lymfoïde cellen, stromale cellen en de extracellulaire matrix. Het samenspel van genetische afwijkingen en het TME dragen bij aan bij aan het ontstaan van FL.

Chronische lymfatische leukemie (CLL) is de meest voorkomende leukemie bij volwassenen in westerse landen, met een incidentie van 6,8 mannen en 3,5 vrouwen per 100.000 mensen per jaar. CLL wordt geclassificeerd in twee groepen op basis van de mutatiestatus van de BCR, waarbij de ongemuteerde groep (98-100% identiteit met het ongemuteerde zware

keten V-gen) een agressiever klinisch verloop kent in vergelijking met de gemuteerde groep (identiteit < 98%). De ontwikkeling van CLL is vaak gekoppeld aan een fase die bekend staat als monoclonale B-cel lymfocytose (MBL) met cellen die fenotypisch niet te onderscheiden zijn van het CLL-fenotype CD19+CD5+CD20lowCD19low. Hoewel het een voorloper stadium lijkt, zal slechts 1-2% van de MBL transformeren tot CLL. MBL wordt geclassificeerd in laag-aantal MBL en hoog-aantal MBL op basis van het aantal cellen met CLL-fenotype in bloed.

De factoren die betrokken zijn bij de progressie van MBL naar CLL zijn nog onbegrepen en bieden de mogelijkheid nauwgezet te zoeken naar verschillen en overeenkomsten tussen normale en maligne cellen. Het feit dat FL en CLL beide ontstaan uit rijpe B-cellen maakt het bovendien interessant om FL en CLL in parallel te analyseren.

CLL heeft genetische veranderingen die betrokken kunnen zijn bij progressie van de ziekte. Door middel van associatiestudies zijn 45 kiembaan varianten in genomische loci geïdentificeerd die het risico op het ontwikkelen van de ziekte met 2-8 keer verhogen. Somatische mutaties zijn zeer variabel in CLL, met uitzondering van mutaties in 5 genen die frequent gemuteerd zijn (SF3B1, ATM, TP53, POT1 en NOTCH1). Copy number varianten komen daarentegen voor bij 80% van de patiënten en zijn nuttig bij de diagnostiek van de ziekte en zijn relevant voor het bepalen van de optimale behandeling. Bovendien vertoont CLL een specifiek oncogeen mechanisme dat wordt veroorzaakt door de BCR: autonome signalering. Dit mechanisme, dat een continue activerend signaal veroorzaakt zonder dat er sprake is van stimulatie door een extern antigeen, is de enige eigenschap die gezien wordt bij alle gevallen van CLL en is daarom zeer waarschijnlijk onontbeerlijk voor de ontwikkeling van deze ziekte.

DIT PROEFSCHRIFT

Ons immuunsysteem omvat een groot repertoire aan B cellen dat in staat is om een vrijwel onbeperkt scala aan antigenen te herkennen. Dit wordt bereikt door de stochastische combinatie van genen en het fysiologisch proces van somatische hypermutatie wat mutaties in de B-cel receptor veroorzaakt. De deaminatie van cytosine naar uracil door het enzym AID speelt hierbij een centrale rol. AID richt zich bij voorkeur op specifieke motieven in het DNA. Het reparatieproces ten gevolge van deze deaminaties wordt uitgevoerd door middel van BER of MMR. Het gebruikelijke resultaat van dit fout gevoelige reparatieproces is een enkele nucleotide substitutie, maar het is ook mogelijk om tandem mutaties te produceren, waarbij twee of meer aangrenzende nucleotiden gelijktijdig worden gemuteerd.

In **hoofdstuk 2** hadden we als doel om deze tandem dinucleotide substituties (TDNS) in immuunglobulinegenen te identificeren en de achterliggende mechanismen te karakteriseren bij gezonde individuen en patiënten met deficiënties in DNA herstelmechanismen. Hiervoor hebben we eerst alle mutaties in BCR repertoires vastgesteld door middel van 'massive parallel sequencing' op populaties van gezuiverde B-cellen van 12 gezonde donoren en 5 patiënten met DNA-reparatie deficiëntie waarvan 4 in MMR en 1 BER. Door eerst het resultaat van AID mutagenese te simuleren op de ongemuteerde counterparts van deze BCR repertoires konden we onderscheid maken tussen twee toevallig naastgelegen mutaties als gevolg van sequentiële

mutatie, en werkelijke TDNS als gevolg van één enkele AID-hit. Dit leidde tot het verwijderen van 46.2% van de kandidaat TDNS waardoor 5775 mutaties overbleven voor verdere analyse. Hiervan was 6% een tandem mutatie, en 4.1% een TDNS. Er was geen verschil tussen bepaalde V-genen als we de frequentie van TDNS vergeleken met de frequentie van 'single' nucleotide substituties (SNS). Bijna alle gevallen van TDNS (98,6%) hadden een aminozuur verandering tot gevolg, tegenover 71,6% van de SNS. In een groot deel (89,6%) van de TDNS werd de kiembaansequentie gedeeltelijk of helemaal gekopieerd, wat het frequent voorkomen van inverterende tandem substituties kan verklaren. Bij de analyse van BCR-sequenties van patiënten met de DNA-reparatie deficiëntie MMR vonden we een vergelijkbaar patroon van TDNS. Bij de patiënt met BER-deficiëntie daarentegen ontbraken TDNS volledig, terwijl we bij deze patiënt een vergelijkbaar niveau van SNS vonden. Dit suggereert dat tandem mutaties in BCR gerelateerd zijn aan BER-activiteit, maar omdat we dit in slechts 1 patiënt zagen dient dit verder gevalideerd te worden in een grotere patiëntengroep.

Na het bestuderen van mutagene mechanismen die ten grondslag liggen aan veranderingen in de BCR sequentie, was onze volgende vraag of het mogelijk was om het bestaan van twee immunoglobuline transcript populaties binnen één cel te detecteren die verschillen door één of enkele nucleotiden. Immers, onmiddellijk nadat AID een mutatie heeft veroorzaakt, kunnen tegelijkertijd BCR transcripten van vóór de mutatie, en ook al nieuwe transcripten mét de recent verkregen mutatie(s) in de cel aanwezig zijn. In **hoofdstuk 3** hebben we deze B-cellen gedetecteerd en gekwantificeerd. Hiervoor hebben we resultaten van single-cell RNA sequencing (scRNA-seq) en single-cell B-cell receptor sequencing gebruikt van 12 FL-patiënten en 5 CLL-patiënten als controle. In een zelf ontwikkeld bio informatica proces hebben we na zeer strenge filtering 364 cellen geïdentificeerd die inderdaad twee verschillende immunoglobuline transcripten bevatten. Deze cellen vonden we in 7 van de 12 FL-patiënten, terwijl geen enkele cel met 2 BCR transcripten werd gedetecteerd bij de CLL-patiënten. In 85% van de gevallen konden we bepalen welk transcript de oorspronkelijke en welk transcript de nieuwe varianten was. Deze mutaties bleken in 41% C-T-transities te zijn en hadden in 27% plaatsgevonden in het canonieke WRCY-motief van AID, wat consistent is met bevindingen uit exoom-brede sequencing in een onafhankelijke cohort. Daarna hebben we de volledige transcriptomen van dezelfde cellen verkregen door scRNA-seq vergeleken: cellen met 2 BCR-transcripten tegenover cellen met slechts 1 BCR transcript. Hier vonden we dat 5766 genen verschillend tot expressie kwamen. Het meest in het oog sprong de hogere expressie van AID in cellen met twee BCR-transcript varianten. Door de 5766 genen te associëren met groepen van genen waarvan in de literatuur bekend is dat deze betrokken zijn bij bepaalde cellulaire processen konden we vaststellen dat MMR prominenter is dan BER in de FL-cellen met 2 BCR-transcripten. Deze analyses, die aantonen dat FL-cellen twee verschillende BCR-transcripten kunnen bevatten waarvan er één het directe gevolg van AID is, tonen voor het eerst aan dat in FL voortdurende somatische hypermutatie plaatsvindt.

In het tweede deel van het proefschrift hebben wij ons verder verdiept in het SHM-proces, met focus op de mechanismen die normale cellen kunnen transformeren tot maligne cellen. Bij AID maken we onderscheid tussen de normale activiteit van AID om immunoglobuline

genen te muteren, en de afwijkende AID activiteit buiten de immunoglobuline regio's. In **hoofdstuk 4** hebben wij ons gericht op het identificeren van de afwijkende activiteit van AID in het totale landschap van mutaties in FL en CLL. Hoewel FL en CLL een vergelijkbaar patroon vertoonden met overwegend C-naar-T mutaties, vonden we deze vaker in een AID-motief bij FL dan bij CLL (42,7% versus 33,6%). Naast AID zijn er vele veroorzakers van mutaties, zoals chemicaliën, uv-licht en tabaksrook, en deze laten ieder hun eigen mutatie signatuur achter. Bij vergelijking van afwijkende FL en CLL mutatie signaturen met een database waarin al deze signaturen gerubriceerd zijn, bleek inderdaad dat bij FL meestal AID-gerelateerde signaturen betrokken zijn. Bij CLL daarentegen hadden kanker gerelateerde signaturen de overhand. AID-mutaties kunnen alleen plaatsvinden in actief DNA dat afgerold is en wordt afgelezen. Met de zogenaamde HiC-techniek kan onderscheid gemaakt worden tussen mutaties in actief en niet-actief DNA. In data verkregen met deze HiC-techniek troffen we het canonieke AID signatuur en ook activiteit van DNA reparatie mechanismes vaker aan in actief DNA bij FL dan bij CLL. Foutieve correctie van DNA schade met mutaties tot gevolg kan het resultaat zijn van mutaties in de reparatiegenen zelf of in genen die onderdeel uitmaken van ketens waarbinnen de reparatiegenen werken. Mutaties bij FL in deze genen vonden we vaker in de DNA reparatie ketens Fanconi anemie en BER. In CLL kwamen de meeste mutaties in de DNA damage response keten maar het niveau ervan verschilde niet met FL. Deze bevindingen kunnen verklaren waarom FL en CLL een verschillend landschap aan mutaties vertonen.

In het laatste hoofdstuk van dit proefschrift hebben wij geprobeerd de dynamiek tussen de genetische componenten en BCR-signaaloverdracht in de progressie van MBL naar CLL te bepalen. Het autonome BCR signaal bij CLL is een uniek oncogeen mechanisme dat niet ontstaat door binding aan een externe antigeen, maar wordt veroorzaakt door de herkenning van de eigen BCR. In **hoofdstuk 5** hebben we van 191 broers en zussen van CLL-patiënten een bloedmonster verkregen en gescreend op aanwezigheid van populaties van MBL cellen die herkenbaar zijn aan het CLL-fenotype (CD19+CD5+CD20lowCD19low). Wij vonden deze populaties in 34 monsters met een grote variatie in percentage CLL-fenotype cellen. In 17 gevallen konden we de volledige sequentie van de BCR achterhalen waarvan er vijf een sterke overeenkomst hadden met BCR's die we frequent in CLL vinden. Alle BCR's van MBL CLL-fenotype cellen hadden de aminozuur patronen waarvan aangenomen wordt dat deze betrokken zijn bij het ontstaan van het autonome signaal. Van 11 BCR's van MBL-individuen hebben we de sterkte van het autonome signaal gekwantificeerd. Het signaal was in alle experimenten aanwezig maar was structureel lager dan het signaal dat we in BCR's van CLL konden meten. Daarnaast vonden we een verband tussen de grootte van de MBL populatie en de sterkte van het autonome signaal. Door middel van genetische analyses vonden we een link met 12 kiembaan genotypes. Daarnaast waren 24 polygenetische risico factoren verhoogd in MBL en CLL vergeleken met de algemene bevolking. Daarnaast werd een copy number variatie (CNV) analyse uitgevoerd op 11 CLL en 15 MBL monsters. In alle monsters werden CNV's aangetroffen waarbij de CNV's in CLL voornamelijk klonaal waren en die in MBL sub-klonaal, wat suggereert dat de MBL populaties kunnen bestaan uit meerdere, verschillende

expansies van MBL cellen. Tot slot hebben we gekeken naar de somatische varianten binnen de gepaarde CLL-MBL samples. Het totaal aantal somatische varianten was gelijk in beide groepen, maar in MBL vonden we wederom een hoger aandeel aan sub-klonale somatische varianten. Deze bevindingen laten zien dat naast de fenotypische overeenkomsten tussen MBL en CLL het autonome BCR signaal ook aanwezig is in MBL en dat geëxpandeerde MBL klonen eveneens genetische afwijking bevatten die kenmerkend zijn voor CLL.

Samenvattend hebben wij mutagene mechanismen bestudeerd die geassocieerd zijn met antilichaam rijping en diversificatie. Wij analyseerden tandem mutaties bij gezonde donoren en bij patiënten met DNA-reparatie deficiëntie. Daarbij hebben wij een moleculair mechanisme voorgesteld voor het ontstaan van deze mutaties. Daarnaast hebben wij somatische hypermutatie op enkel-celniveau onderzocht, waarbij FL als model voor voortdurende SHM werd gebruikt. Dit stelde ons voor het eerst in staat om SHM op het niveau van één enkele B-cel te detecteren en te kwantificeren. Het tweede gedeelte van het proefschrift behandelde pathologische mechanismen, waaronder de afwijkende somatische hypermutatie in FL en CLL. Wij hebben het globale landschap en het effect van de pathologische activiteit van AID gekarakteriseerd en onderscheidende patronen tussen FL en CLL gedetecteerd. Ten slotte hebben we de relatie tussen genetische varianten en autonome signaaloverdracht bij individuen met MBL bestudeerd. MBL vertoonde een hoger aandeel sub-klonale somatische varianten in vergelijking met CLL. Interessant is dat alle geanalyseerde gevallen autonome signaaloverdracht vertoonden, maar met minder signaalsterkte dan bij CLL, wat de belangrijke rol van deze oncogene mechanisme in de progressie van MBL naar CLL aantoont.

LIST OF PUBLICATIONS AND CONFERENCE PRESENTATIONS

PUBLICATIONS IN THIS THESIS

Sepúlveda-Yáñez, J. H., Alvarez-Saravia, D., Fernández-Goycoolea, J., Aldridge, J., van Bergen, C. A. M., Posthuma, W., Uribe-Paredes, R., Veelken, H., & Navarrete, M. A. (2021). Integration of mutational signature analysis with 3D chromatin data unveils differential AID-related mutagenesis in indolent lymphomas. International Journal of Molecular Sciences, 22(23), 13015.

Sepúlveda-Yáñez, J. H., Alvarez Saravia, D., Pilzecker, B., van Schouwenburg, P. A., van den Burg, M., Veelken, H., Navarrete, M. A., Jacobs, H., & Koning, M. T. (2022). Tandem substitutions in somatic hypermutation. Frontiers in Immunology, 12, 807015.

Quinten, E., **Sepúlveda-Yáñez, J. H.**, Koning, M. T., Eken, J. A., Pfeifer, D., Nteleah, V., de Groen, R. A. L., Saravia, D. A., Knijnenburg, J., Stuivenberg-Bleijswijk, H. E., Pantic, M., Agathangelidis, A., Keppler-Hafkemeyer, A., van Bergen, C. A. M., Uribe-Paredes, R., Stamatopoulos, K., Vermaat, J. S. P., Zirlik, K., Navarrete, M. A., Jumaa, H., & Veelken, H. (2023). Autonomous B-cell receptor signaling and genetic aberrations in chronic lymphocytic leukemia-phenotype monoclonal B lymphocytosis in siblings of patients with chronic lymphocytic leukemia. Haematologica, 109(3), 824-834.

Sepúlveda-Yáñez, J. H., Saravia, D. A., Medina, D., Quinten, E., Kloet, S., Jansen, P. M., Posthuma, E. W. M., Kiełbasa, S. M., van Bergen, C. A. M., Navarrete, M. A., & Veelken, H. (2024). Detection of occurrent somatic hypermutation and associated gene expression profile in single. Blood Cancer Discovery. Manuscript under review.

PUBLICATIONS NOT IN THE THIS THESIS

Cabas-Mora, G., Daza, A., Soto-García, N., Garrido, V., Alvarez, D., Navarrete, M., Sarmiento-Varón, L., Yañez, J. H. S., Davari, M. D., Cadet, F., Olivera-Nappa, Á., Uribe-Paredes, R., & Medina-Ortiz, D. (2024). Peptipedia v2.0: A peptide sequence database and user-friendly web platform. A major update (p. 2024.07.11.603053). bioRxiv. https://doi.org/10.1101/2024.07.11.603053
Eken, J. A., Koning, M. T., Kupcova, K., **Sepúlveda Yáñez, J. H.**, de Groen, R. A. L., Quinten, E., Janssen, J., van Bergen, C. A. M., Vermaat, J. S. P., Cleven, A., Navarrete, M. A., Ylstra, B., de Jong, D., Havranek, O., Jumaa, H., & Veelken, H. (2024). Antigen-independent, autonomous B cell receptor signaling drives activated B cell DLBCL. Journal of Experimental Medicine, 221(5), e20230941. https://doi.org/10.1084/jem.20230941

Goles, M., Daza, A., Cabas-Mora, G., Sarmiento-Varón, L., **Sepúlveda-Yañez, J.**, Anvari-Kazemabad, H., Davari, M. D., Uribe-Paredes, R., Olivera-Nappa, Á., Navarrete, M. A., & Medina-Ortiz, D. (2024). Peptide-based drug discovery through artificial intelligence: Towards an autonomous design of therapeutic peptides. Briefings in Bioinformatics, 25(4), bbae275. https://doi.org/10.1093/bib/bbae275

Morande, P. E., Yan, X.-J., **Sepulveda, J.**, Seija, N., Marquez, M. E., Sotelo, N., Abreu, C.,

Crispo, M., Fernández-Graña, G., Rego, N., Bois, T., Methot, S. P., Palacios, F., Remedi, V., Rai, K. R., Buschiazzo, A., Di Noia, J. M., Navarrete, M. A., Chiorazzi, N., & Oppezzo, P. (2021). AID overexpression leads to aggressive murine CLL and nonimmunoglobulin mutations that mirror human neoplasms. Blood, 138(3), 246–258. https://doi.org/10.1182/blood.2020008654

Shafighi, S. D., Kiełbasa, S. M., **Sepúlveda-Yáñez, J.**, Monajemi, R., Cats, D., Mei, H., Menafra, R., Kloet, S., Veelken, H., van Bergen, C. A. M., & Szczurek, E. (2021). CACTUS: Integrating clonal architecture with genomic clustering and transcriptome profiling of single tumor cells. Genome Medicine, 13(1), 45. https://doi.org/10.1186/s13073-021-00842-w

van Bergen, C. A. M., Kloet, S. L., Quinten, E., **Sepúlveda Yáñez, J. H.**, Menafra, R., Griffioen, M., Jansen, P. M., Koning, M. T., Knijnenburg, J., Navarrete, M. A., Kiełbasa, S. M., & Veelken, H. (2023). Acquisition of a glycosylated B-cell receptor drives follicular lymphoma toward a dark zone phenotype. Blood Advances, 7(19), 5812–5816. https://doi.org/10.1182/bloodadvances.2023010725

CONFERENCE PRESENTATIONS

59th American Society of hematology (ASH) Annual Meeting and Exposition, Atlanta, United States (2017) (poster presentation) Sepulveda-Yanez,J.H., van Bergen, C.A.M., Alvarez,D., Kielbasa S.M., Mei H., Veelken, J.H., Navarrete, M.A.

The Mutation Landscape of Follicular Lymphoma and Chronic Lymphocytic Leukemia Reveals a Distinctive Frequency and Composition of Aid-Related Mutation Signatures

60th American Society of hematology (ASH) Annual Meeting and Exposition, San Diego, United States (2018) (poster presentation) Sepulveda-Yanez,J.H., Alvarez,D., Fernandez-Goycoolea, J., van Bergen, C.A.M., Veelken, J.H., Navarrete, M.A.

Differential Genome-Wide Mutational Patterns in Indolent B-Cell Lymphomas

63rd American Society of hematology (ASH) Annual Meeting and Exposition, Atlanta, United States (2021) (poster presentation) Sepulveda-Yanez,J.H., Alvarez,D., Quinten, E., Menafra, R., Kloet, S.L., Monajemi, R., Kielbasa S.M., Navarrete, M.A., van Bergen, C.A.M., Veelken, J.H.

”Snapshotting” Somatic Hypermutation in Single Follicular Lymphoma Cells

63rd American Society of hematology (ASH) Annual Meeting and Exposition, Atlanta, United States (2021) (poster presentation, co-author) Koning, M.T, Sepulveda-Yanez,J.H., Alvarez,D., Pilzecker, B., van Schouwenburg, P., van den Burg, M., Navarrete, M.A., Jacobs, H., Veelken, J.H.

Tandem Substitutions in Somatic Hypermutation

64th American Society of hematology (ASH) Annual Meeting and Exposition, New Orleans, United States (2022) (poster presentation, co-author) Gonzalez-Puelma, J., Sepulveda-Yanez,J.H., Torres-Almonacid, Alvarez,D., Marquez, M.E., Alvarez-Ojeda, H., Cardemil, D., Urive-Paredes, R., Oppezo, P., Navarrete, M.A.

Unbiased NGS Approach for IGHV Mutational Status, Clonal Space and Immune Repertoire Assessment in Chronic Lymphocytic Leukemia

64th American Society of hematology (ASH) Annual Meeting and Exposition, New Orleans, United States (2022) (poster presentation, co-author) Eken, J.A, Koning, M.T, Sepulveda-Yanez,J.H., van Bergen, C.A.M, Quinten, E., Kupcova, K., Havranek, O., De groen, R.A.L, Vermaat, J.S.P, de Jong, D., Kluin, P., Cleven, A., Jumaa., H, Veelken, J.H.

Antigen-Independent, Autonomous B-Cell Receptor Signaling in Diffuse Large B-Cell Lymphoma

64th American Society of hematology (ASH) Annual Meeting and Exposition, New Orleans, United States (2021) (poster presentation, co-author) Sepulveda-Yanez,J.H., Quinten, E., Koning, M.T, Eken, J.A, Pfeifer, D., De groen, R.A.L, Alvarez-Saravia, D., Agathangelidis, A.,van Bergen, C.A.M, Stamatopoulos, S., Vermaat, J.S.P, Zirlik, K., Navarrete, M.A., Jumaa., H, Veelken, J.H.

Autonomous BCR Signaling and Genetic Aberrations in CLL-Phenotype Monoclonal B Lymphocytosis in Siblings of CLL Patients

CURRICULUM VITAE

Julieta Haydeé Sepúlveda Yáñez was born on October 26, 1987, in Punta Arenas, Chile. In 2005, she completed her high school education at Liceo Juan Bautista Contardi. She earned her bachelor's degree in Medical Technology (2006-2010), specializing in morphophysiopathology and cytodiagnostics at the University of Concepcion, graduating cum laude and as the top of her class. In 2011, she started her master's degree in Biochemistry and Bioinformatics at the University of Concepcion, having received a national scholarship, and graduated cum laude. In 2014, she was appointed as a lecturer at the Universidad de Magallanes. In December 2016, after receiving a national doctoral scholarship, she began her PhD in the Department of Hematology at Leiden University Medical Center. Currently, she is an associate professor in the Faculty of Health Sciences at Universidad de Magallanes.

ACKNOWLEDGEMENT

I have been thinking for a long time about this section of my thesis, maybe for years. And I always come back to the same thought: my first thanks are for my family. I have been so blessed to have the family that I have. My dad, mom, and brother have always believed in me. They made me believe that anything was possible and the sky is the limit. They have always been by my side. I realize how extraordinary is my family, and I feel so fortunate.

This path has been full of people who believed in me and gave me opportunities to move forward. I would like to thank my supervisors. Hendrik gave me the freedom to investigate my own ideas and follow my own process. He provided me with this tools that I think will be priceless for my future career. I also want to thank Marcelo, in Hendrik's words, "he turns everything to gold," and I have seen him do that so many times. Additionally, I want to thank to Kees, he was there in the everyday discussion and he is an "excellent" colleague. I want to add here my gratitude to my lab mates Marvyn, Janneke and Edwin. A special thanks to Diego. I think without his support, I would still be working on my thesis. He has been an essential part of this work. Also my appreciation to all the cATG members who were part of this journey.

When I moved to the Netherlands, I was so afraid but I found an amazing community. I learned so much and felt at home living there. It was a life-changing experience. I am grateful to have chosen Leiden University Medical Center. My thanks to the entire Department of Hematology, both past and present. My deep thanks to Dyantha, Georgia, Miranda, Laura, Lois, Nadine and Tassilo, and a heartfelt thanks to Aicha for all her support. They made my stay amazing, and we shared so many memories and invaluable experiences on this path. Of course, I would also like to thank the Latin American community in the Netherlands. Special gratitude goes to Javiera for her friendship throughout this process.

This has been a long journey. None of this was fortuitous. It was completely deliberate. I dreamed of doing a PhD when I was 19 years old. I remember the exact moment: it was in a biochemistry class during my bachelor's degree in medical technology. A professor explained how the research world works, and at that moment, I knew I wanted to be a researcher. I was motivated by the naive idea that in this way, I could positively impact more lives than in the clinic. This was my driver, and I finished it. This is why I want to thank myself, and I think someone else used the perfect words for this:

"Last but not least, I wanna thank me, I wanna thank me for believing in me, I wanna thank me for doing all this hard work, I wanna thank me for having no days off, I wanna thank me for never quitting, I wanna thank me for always being a giver, and trying to give more than I receive, I wanna thank me for trying to do more right than wrong, I wanna thank me for just being me at all times" - Snoop Dogg

Bibliography

1. Hozumi, N. & Tonegawa, S. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. ***Proc Natl Acad Sci U S A* 73,** 3628–3632 (Oct. 1976).
2. Gellert, M. V(D)J Recombination: RAG Proteins, Repair Factors, and Regulation. ***Annu. Rev. Biochem.* 71.** Publisher: Annual Reviews, 101–132. doi:10.1146/annurev.biochem.71.090501.150203 (June 2002).
3. Schatz, D. G. & Spanopoulou, E. en. in ***Molecular Analysis of B Lymphocyte Development and Activation*** (eds Singh, H. & Grosschedl, R.) Series Title: Current Topics in Microbiology and Immunology, 49–85 (Springer-Verlag, Berlin/Heidelberg, 2005). doi:10.1007/3-540-26363-2_4.
4. Victora, G. D. & Nussenzweig, M. C. Germinal Centers. ***Annual Review of Immunology* 30.** _eprint: https://doi.org/10.1146/annurev-immunol-020711-075032, 429–457. doi:10.1146/annurev-immunol-020711-075032 (2012).
5. Muramatsu, M., Kinoshita, K., ***et al.*** Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. eng. ***Cell* 102.** PMID: 11007474, 553–563. doi:10.1016/s0092-8674(00)00078-7 (Sept. 1, 2000).
6. Begum, N. A., Nagaoka, H., Kobayashi, M. & Honjo, T. en. in ***Molecular Biology of B Cells*** DOI: 10.1016/B978-0-12-397933-9.00018-7, 305–344 (Elsevier, 2015).
7. Muramatsu, M., Kinoshita, K., ***et al.*** Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. eng. ***Cell* 102,** 553–563. doi:10.1016/s0092-8674(00)00078-7 (Sept. 2000).
8. Revy, P., Muto, T., ***et al.*** Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). eng. ***Cell* 102,** 565–575. doi:10.1016/s0092-8674(00)00079-9 (Sept. 2000).
9. Sepúlveda-Yáñez, J. H., Alvarez Saravia, D., ***et al.*** Tandem Substitutions in Somatic Hypermutation. ***Frontiers in Immunology* 12** (2022).
10. Sharbeen, G., Yee, C. W. Y., Smith, A. L. & Jolly, C. J. Ectopic restriction of DNA repair reveals that UNG2 excises AID-induced uracils predominantly or exclusively during G1 phase. eng. ***The Journal of Experimental Medicine* 209,** 965–974. doi:10.1084/jem.20112379 (May 2012).
11. Wang, S., Chahwan, R., Wei, L. & Scharff, M. D. en. in ***Encyclopedia of Immunobiology*** 126–133 (Elsevier, 2016). doi:10.1016/B978-0-12-374279-7.05015-3.
12. Chaudhuri, J., Tian, M., ***et al.*** Transcription-targeted DNA deamination by the AID antibody diversification enzyme. en. ***Nature* 422.** Number: 6933 Publisher: Nature Publishing Group, 726–730. doi:10.1038/nature01574 (Apr. 2003).
13. Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. en. ***Nature* 424,** 103–107. doi:10.1038/nature01760 (July 2003).
14. Methot, S. P. & Di Noia, J. M. in ***Advances in Immunology*** (ed Alt, F. W.) 37–87 (Academic Press, Jan. 2017). doi:10.1016/bs.ai.2016.11.002.
15. Di Noia, J. M. & Neuberger, M. S. Molecular Mechanisms of Antibody Somatic Hypermuta-

tion. en. *Annual Review of Biochemistry* **76,** 1–22. doi:10.1146/annurev.biochem.76.061705.090740 (June 2007).

16. Storb, U., Shen, H. M. & Nicolae, D. Somatic hypermutation: Processivity of the cytosine deaminase AID and error-free repair of the resulting uracils. *Cell Cycle* **8.** Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/cc.8.19.9658, 3097–3101. doi:10.4161/cc.8.19.9658 (Oct. 2009).
17. Keim, C., Kazadi, D., Rothschild, G. & Basu, U. Regulation of AID, the B-cell genome mutator. *Genes & Development* **27,** 1–17. doi:10.1101/gad.200014.112 (Jan. 2013).
18. Rogozin, I. B. & Diaz, M. Cutting Edge: DGYW/WRCH Is a Better Predictor of Mutability at G:C Bases in Ig Hypermutation Than the Widely Accepted RGYW/WRCY Motif and Probably Reflects a Two-Step Activation-Induced Cytidine Deaminase-Triggered Process. en. *The Journal of Immunology* **172,** 3382–3384. doi:10.4049/jimmunol.172.6.3382 (Mar. 2004).
19. Rogozin, I. B. & Kolchanov, N. A. Somatic hypermutagenesis in immunoglobulin genes: II. Influence of neighbouring base sequences on mutagenesis. en. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **1171,** 11–18. doi:10.1016/0167-4781(92)90134-L (Nov. 1992).
20. Dörner, T., Foster, S. J., Farner, N. L. & Lipsky, P. E. Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. eng. *European Journal of Immunology* **28,** 3384–3396. doi:10.1002/(SICI)1521-4141(199810)28:10<3384::AID-IMMU3384>3.0.CO;2-T (Oct. 1998).
21. Foster, S. J., Dörner, T. & Lipsky, P. E. Somatic hypermutation of VkappaJkappa rearrangements: targeting of RGYW motifs on both DNA strands and preferential selection of mutated codons within RGYW motifs. eng. *European Journal of Immunology* **29,** 4011–4021. doi:10.1002/(SICI)1521-4141(199912)29:12<4011::AID-IMMU4011>3.0.CO;2-W (Dec. 1999).
22. Mayorov, V. I., Rogozin, I. B., *et al.* Expression of human AID in yeast induces mutations in context similar to the context of somatic hypermutation at G-C pairs in immunoglobulin genes. *BMC Immunology* **6,** 10. doi:10.1186/1471-2172-6-10 (June 2005).
23. Peled, J. U., Kuang, F. L., *et al.* The Biochemistry of Somatic Hypermutation. *Annual Review of Immunology* **26,** 481–511. doi:10.1146/annurev.immunol.26.021607.090236 (Mar. 2008).
24. Matsuda, T., Bebenek, K., *et al.* Error rate and specificity of human and murine DNA polymerase eta. eng. *Journal of Molecular Biology* **312,** 335–346. doi:10.1006/jmbi.2001.4937 (Sept. 2001).
25. Pavlov, Y. I., Rogozin, I. B., *et al.* Correlation of somatic hypermutation specificity and A-T base pair substitution errors by DNA polymerase eta during copying of a mouse immunoglobulin kappa light chain transgene. eng. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 9954–9959. doi:10.1073/pnas.152126799 (July 2002).
26. Zan, H., Komori, A., *et al.* The translesion DNA polymerase zeta plays a major role in Ig and bcl-6 somatic hypermutation. eng. *Immunity* **14,** 643–653. doi:10.1016/s1074-7613(01)00142-x (May 2001).
27. Faili, A., Aoufouchi, S., *et al.* Induction of somatic hypermutation in immunoglobulin genes is dependent on DNA polymerase iota. en. *Nature* **419.** Number: 6910 Publisher: Nature Publishing Group, 944–947. doi:10.1038/nature01117 (Oct. 2002).
28. Ruiz, J. F., Lucas, D., *et al.* Overexpression of human DNA polymerase mu (Pol mu) in a Burkitt's lymphoma cell line affects the somatic hypermutation rate. eng. *Nucleic Acids Research* **32,** 5861–5873. doi:10.1093/nar/gkh929 (2004).
29. Rogozin, I. B., Lada, A. G., *et al.* Activation induced deaminase mutational signature overlaps with CpG methylation sites in follicular lymphoma and other cancers. *Scientific Reports* **6,** 38133. doi:10.1038/srep38133 (Dec. 2016).
30. Rada, C., Williams, G. T., *et al.* Immunoglobulin Isotype Switching Is Inhibited and Somatic Hypermutation Perturbed in UNG-Deficient Mice. en. *Current Biology* **12,** 1748–1755. doi:10.1016/S0960-9822(02)01215-0 (Oct. 15, 2002).
31. Di Noia, J. M. & Neuberger, M. S. Molecular Mechanisms of Antibody Somatic Hypermutation. en. *Annual Review of Biochemistry* **76,** 1–22. doi:10.1146/annurev.biochem.76.061705.090740 (June 7, 2007).
32. Di Noia, J. M., Rada, C. & Neuberger, M. S. SMUG1 is able to excise uracil from immunoglobulin

genes: insight into mutation versus repair. ***The EMBO Journal* 25.** PMID: 16407970 PMCID: PMC1383525, 585–595. doi:10.1038/sj.emboj.7600939 (Feb. 8, 2006).

33. Kemmerich, K., Dingler, F. A., Rada, C. & Neuberger, M. S. Germline ablation of SMUG1 DNA glycosylase causes loss of 5-hydroxymethyluracil- and UNG-backup uracil-excision activities and increases cancer predisposition of Ung-/-Msh2-/- mice. ***Nucleic Acids Research* 40,** 6016–6025. doi:10.1093/nar/gks259 (July 2012).
34. Au, R. Y., Ng, K. S., Chi, L. M. & Lam, S. L. Effect of an abasic site on strand slippage in DNA primer-templates. ***J Phys Chem B* 116,** 14781–7. doi:10.1021/jp308759k (2012).
35. Yousif, A. S., Stanlie, A., Mondal, S., Honjo, T. & Begum, N. A. Differential regulation of S-region hypermutation and class-switch recombination by noncanonical functions of uracil DNA glycosylase. ***Proceedings of the National Academy of Sciences* 111.** Publisher: Proceedings of the National Academy of Sciences, E1016–E1024. doi:10.1073/pnas.1402391111 (Mar. 2014).
36. Dianova, I. I., Bohr, V. A. & Dianov, G. L. Interaction of human AP endonuclease 1 with flap endonuclease 1 and proliferating cell nuclear antigen involved in long-patch base excision repair. eng. ***Biochemistry* 40,** 12639–12644. doi:10.1021/bi011117i (Oct. 2001).
37. Krokan, H. E. & Bjørås, M. Base Excision Repair. en. ***Cold Spring Harbor Perspectives in Biology* 5,** a012583. doi:10.1101/cshperspect.a012583 (Apr. 2013).
38. Ranalli, T. A., DeMott, M. S. & Bambara, R. A. Mechanism underlying replication protein a stimulation of DNA ligase I. eng. ***The Journal of Biological Chemistry* 277,** 1719–1727. doi:10.1074/jbc.M109053200 (Jan. 2002).
39. Vidal, A. E., Boiteux, S., Hickson, I. D. & Radicella, J. XRCC1 coordinates the initial and late stages of DNA abasic site repair through protein–protein interactions. ***The EMBO Journal* 20.** PMID: 11707423 PMCID: PMC125722, 6530–6539. doi:10.1093/emboj/20.22.6530 (Nov. 15, 2001).
40. Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. en. ***Nature* 424,** 103–107. doi:10.1038/nature01760 (July 2003).
41. Yeap, L.-S., Hwang, J. K., ***et al.*** Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. ***Cell* 163,** 1124–1137. doi:10.1016/j.cell.2015.10.042 (Nov. 2015).
42. Wiesendanger, M., Kneitz, B., Edelmann, W. & Scharff, M. D. Somatic Hypermutation in Muts Homologue (Msh)3-, Msh6-, and Msh3/Msh6-Deficient Mice Reveals a Role for the Msh2–Msh6 Heterodimer in Modulating the Base Substitution Pattern. ***Journal of Experimental Medicine* 191,** 579–584. doi:10.1084/jem.191.3.579 (Feb. 2000).
43. Warren, J. J., Pohlhaus, T. J., ***et al.*** Structure of the human MutSalpha DNA lesion recognition complex. eng. ***Molecular Cell* 26,** 579–592. doi:10.1016/j.molcel.2007.04.018 (May 2007).
44. Martomo, S. A., Yang, W. W. & Gearhart, P. J. A role for Msh6 but not Msh3 in somatic hypermutation and class switch recombination. eng. ***The Journal of Experimental Medicine* 200,** 61–68. doi:10.1084/jem.20040691 (July 2004).
45. Schaetzlein, S., Chahwan, R., ***et al.*** Mammalian Exo1 encodes both structural and catalytic functions that play distinct roles in essential biological processes. eng. ***Proceedings of the National Academy of Sciences of the United States of America* 110,** E2470–2479. doi:10.1073/pnas.1308512110 (July 2013).
46. Frieder, D., Larijani, M., Collins, C., Shulman, M. & Martin, A. The concerted action of Msh2 and UNG stimulates somatic hypermutation at A . T base pairs. eng. ***Molecular and Cellular Biology* 29,** 5148–5157. doi:10.1128/MCB.00647-09 (Sept. 2009).
47. Kannouche, P. L., Wing, J. & Lehmann, A. R. Interaction of Human DNA Polymerase η with Monoubiquitinated PCNA: A Possible Mechanism for the Polymerase Switch in Response to DNA Damage. ***Molecular Cell* 14,** 491–500. doi:10.1016/S1097-2765(04)00259-X (May 2004).
48. Liu, D., Keijzers, G. & Rasmussen, L. J. DNA mismatch repair and its many roles in eukaryotic cells. en. ***Mutation Research/Reviews in Mutation Research* 773,** 174–187. doi:10.1016/j.mrrev.2017.07.001 (July 2017).
49. Ouchida, R., Ukai, A., ***et al.*** Genetic analysis reveals an intrinsic property of the germinal cen-

ter B cells to generate A:T mutations. ***DNA Repair* 7,** 1392–1398. doi:10.1016/j.dnarep.2008.04.014 (Aug. 2008).

50. Delbos, F., De Smet, A., ***et al.*** Contribution of DNA polymerase η to immunoglobulin gene hypermutation in the mouse. ***Journal of Experimental Medicine* 201,** 1191–1196. doi:10.1084/jem.20050292 (Apr. 2005).
51. Jansen, J. G., Langerak, P., ***et al.*** Strand-biased defect in C/G transversions in hypermutating immunoglobulin genes in Rev1-deficient mice. eng. ***The Journal of Experimental Medicine* 203,** 319–323. doi:10.1084/jem.20052227 (Feb. 2006).
52. Zeng, X., Winter, D. B., ***et al.*** DNA polymerase η is an A-T mutator in somatic hypermutation of immunoglobulin variable genes. en. ***Nature Immunology* 2,** 537–541. doi:10.1038/88740 (June 2001).
53. Alaggio, R., Amador, C., ***et al.*** The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. en. ***Leukemia* 36.** number: 7 publisher: Nature Publishing Group, 1720–1748. doi:10.1038/s41375-022-01620-2 (July 2022).
54. Kurz, K. S., Kalmbach, S., ***et al.*** Follicular Lymphoma in the 5th Edition of the WHO-Classification of Haematolymphoid Neoplasms—Updated Classification and New Biological Data. ***Cancers* 15,** 785. doi:10.3390/cancers15030785 (Jan. 2023).
55. Carbone, A., Roulland, S., ***et al.*** Follicular lymphoma. en. ***Nature Reviews Disease Primers* 5,** 83. doi:10.1038/s41572-019-0132-x (Dec. 2019).
56. Morton, L. M., Wang, S. S., ***et al.*** Lymphoma incidence patterns by WHO subtype in the United States, 1992-2001. ***Blood* 107,** 265–276. doi:10.1182/blood-2005-06-2508 (Jan. 2006).
57. Batlevi, C. L., Sha, F., ***et al.*** Follicular lymphoma in the modern era: survival, treatment outcomes, and identification of high-risk subgroups. en. ***Blood Cancer Journal* 10.** number: 7 publisher: Nature Publishing Group, 1–12. doi:10.1038/s41408-020-00340-z (July 17, 2020).
58. Shikata, H., Yakushijin, Y., ***et al.*** Role of activation-induced cytidine deaminase in the progression of follicular lymphoma. en. ***Cancer Science* 103.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1349-7006.2011.02186.x, 415–421. doi:10.1111/j.1349-7006.2011.02186.x (2012).
59. Cleary, M. L. & Sklar, J. Nucleotide sequence of a t(14;18) chromosomal breakpoint in follicular lymphoma and demonstration of a breakpoint-cluster region near a transcriptionally active locus on chromosome 18. ***Proceedings of the National Academy of Sciences* 82.** Publisher: Proceedings of the National Academy of Sciences, 7439–7443. doi:10.1073/pnas.82.21.7439 (Nov. 1985).
60. Tsujimoto, Y., Cossman, J., Jaffe, E. & Croce, C. M. Involvement of the bcl-2 Gene in Human Follicular Lymphoma. ***Science* 228.** Publisher: American Association for the Advancement of Science, 1440–1443. doi:10.1126/science.3874430 (June 1985).
61. Klein, U. & Dalla-Favera, R. Germinal centres: role in B-cell physiology and malignancy. en. ***Nature Reviews Immunology* 8.** Number: 1 Publisher: Nature Publishing Group, 22–33. doi:10.1038/nri2217 (Jan. 2008).
62. Dickerson, S. K., Market, E., Besmer, E. & Papavasiliou, F. N. AID mediates hypermutation by deaminating single stranded DNA. eng. ***The Journal of Experimental Medicine* 197,** 1291–1296. doi:10.1084/jem.20030481 (May 2003).
63. Muramatsu, M., Sankaranand, V. S., ***et al.*** Specific Expression of Activation-induced Cytidine Deaminase (AID), a Novel Member of the RNA-editing Deaminase Family in Germinal Center B Cells *. English. ***Journal of Biological Chemistry* 274,** 18470–18476. doi:10.1074/jbc.274.26.18470 (June 1999).
64. Kranich, J. & Krautler, N. J. How Follicular Dendritic Cells Shape the B-Cell Antigenome. ***Frontiers in Immunology* 7** (2016).
65. Krysiak, K., Gomez, F., ***et al.*** Recurrent somatic mutations affecting B-cell receptor signaling pathway genes in follicular lymphoma. en. ***Blood* 129,** 473–483. doi:10.1182/blood-2016-07-729954 (Jan. 2017).
66. Okosun, J., Wolfson, R. L., ***et al.*** Recurrent mTORC1-activating RRAGC mutations in follicular lymphoma. ***Nature Genetics* 48,** 183–188. doi:10.1038/ng.3473 (Dec. 2015).
67. Perrett, M., Edmondson, C. & Okosun, J. Biology of follicular lymphoma: insights and windows of clinical opportunity. ***Hematology* 2022,** 688–

694. doi:10.1182/hematology.2022000361 (Dec. 2022).

68. Morin, R. D., Mendez-Lago, M., ***et al.*** Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. en. ***Nature* 476.** Number: 7360 Publisher: Nature Publishing Group, 298–303. doi:10.1038/nature10351 (Aug. 2011).
69. Amé-Thomas, P., Le Priol, J., ***et al.*** Characterization of intratumoral follicular helper T cells in follicular lymphoma: role in the survival of malignant B cells. ***Leukemia* 26,** 1053–1063. doi:10.1038/leu.2011.301 (May 2012).
70. Pangault, C., Amé-Thomas, P., ***et al.*** Integrative Analysis of Cell Crosstalk within Follicular Lymphoma Cell Niche: Towards a Definition of the FL Supportive Synapse. ***Cancers* 12,** 2865. doi:10.3390/cancers12102865 (Oct. 2020).
71. Yang, Z.-Z., Kim, H. J., ***et al.*** Mass Cytometry Analysis Reveals that Specific Intratumoral CD4+ T Cell Subsets Correlate with Patient Survival in Follicular Lymphoma. eng. ***Cell Reports* 26,** 2178–2193.e3. doi:10.1016/j.celrep.2019.01.085 (Feb. 2019).
72. Amé-Thomas, P., Maby-El Hajjami, H., ***et al.*** Human mesenchymal stem cells isolated from bone marrow and lymphoid organs support tumor B-cell growth: role of stromal cells in follicular lymphoma pathogenesis. ***Blood* 109,** 693–702. doi:10.1182/blood-2006-05-020800 (Sept. 2006).
73. Hallek, M., Shanafelt, T. D. & Eichhorst, B. Chronic lymphocytic leukaemia. English. ***The Lancet* 391.** Publisher: Elsevier, 1524–1537. doi:10.1016/S0140-6736(18)30422-7 (Apr. 2018).
74. Delgado, J., Nadeu, F., Colomer, D. & Campo, E. Chronic lymphocytic leukemia: from molecular pathogenesis to novel therapeutic strategies. ***Haematologica* 105,** 2205–2217. doi:10.3324/haematol.2019.236000 (July 2020).
75. Seifert, M., Sellmann, L., ***et al.*** Cellular origin and pathophysiology of chronic lymphocytic leukemia. eng. ***The Journal of Experimental Medicine* 209,** 2183–2198. doi:10.1084/jem.20120833 (Nov. 2012).
76. Landgren, O., Albitar, M., ***et al.*** B-cell clones as early markers for chronic lymphocytic leukemia. eng. ***The New England Journal of Medicine* 360,** 659–667. doi:10.1056/NEJMoa0806122 (Feb. 2009).
77. Rawstron, A. C., Yuille, M. R., ***et al.*** Inherited predisposition to CLL is detectable as subclinical monoclonal B-lymphocyte expansion. ***Blood* 100,** 2289–2290. doi:10.1182/blood-2002-03-0892 (Oct. 2002).
78. Rawstron, A. C., Bennett, F. L., ***et al.*** Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. eng. ***The New England Journal of Medicine* 359,** 575–583. doi:10.1056/NEJMoa075290 (Aug. 2008).
79. Berndt, S. I., Camp, N. J., ***et al.*** Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. en. ***Nature Communications* 7.** Number: 1 Publisher: Nature Publishing Group, 10933. doi:10.1038/ncomms10933 (Mar. 2016).
80. Speedy, H. E., Beekman, R., ***et al.*** Insight into genetic predisposition to chronic lymphocytic leukemia from integrative epigenomics. eng. ***Nature Communications* 10,** 3615. doi:10.1038/s41467-019-11582-2 (Aug. 2019).
81. Landau, D. A., Tausch, E., ***et al.*** Mutations driving CLL and their evolution in progression and relapse. en. ***Nature* 526.** Number: 7574 Publisher: Nature Publishing Group, 525–530. doi:10.1038/nature15395 (Oct. 2015).
82. Döhner, H., Stilgenbauer, S., ***et al.*** Genomic aberrations and survival in chronic lymphocytic leukemia. eng. ***The New England Journal of Medicine* 343,** 1910–1916. doi:10.1056/NEJM200012283432602 (Dec. 2000).
83. Nadeu, F., Clot, G., ***et al.*** Clinical impact of the subclonal architecture and mutational complexity in chronic lymphocytic leukemia. eng. ***Leukemia* 32,** 645–653. doi:10.1038/leu.2017.291 (Mar. 2018).
84. Dühren-von Minden, M., Übelhart, R., ***et al.*** Chronic lymphocytic leukaemia is driven by antigen-independent cell-autonomous signalling. eng. ***Nature* 489,** 309–312. doi:10.1038/nature11309 (Sept. 2012).
85. Iacovelli, S., Hug, E., ***et al.*** Two types of BCR interactions are positively selected during leukemia development in the Eμ-TCL1 transgenic mouse model of CLL. ***Blood* 125,** 1578–1588. doi:10.1182/blood-2014-07-587790 (Mar. 2015).
86. Maity, P. C., Bilal, M., ***et al.*** IGLV3-21*01 is an inherited risk factor for CLL through the acquisition of a single-point mutation enabling autonomous BCR signaling. eng. ***Proceedings of***

the National Academy of Sciences of the United States of America **117,** 4320–4327. doi:10.1073/pnas.1913810117 (Feb. 2020).

87. Stamatopoulos, B., Smith, T., ***et al.*** The Light Chain IgLV3-21 Defines a New Poor Prognostic Subgroup in Chronic Lymphocytic Leukemia: Results of a Multicenter Study. eng. ***Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*** **24,** 5048–5057. doi:10.1158/1078-0432.CCR-18-0133 (Oct. 2018).
89. Kinoshita, K. & Honjo, T. Linking class-switch recombination with somatic hypermutation. ***Nat Rev Mol Cell Biol*** **2,** 493–503. doi:10.1038/35080033 (2001).
90. Neuberger, M. S. & Milstein, C. Somatic hypermutation. ***Curr Opin Immunol*** **7,** 248–54. doi:10.1016/0952-7915(95)80010-7 (1995).
91. Papavasiliou, F. N. & Schatz, D. G. Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. ***Cell*** **109 Suppl,** S35–44. doi:10.1016/s0092-8674(02)00706-7 (2002).
92. Rajewsky, K. Clonal selection and learning in the antibody system. ***Nature*** **381,** 751–8. doi:10.1038/381751a0 (1996).
93. Seki, M., Gearhart, P. J. & Wood, R. D. DNA polymerases and somatic hypermutation of immunoglobulin genes. ***EMBO Rep*** **6,** 1143–8. doi:10.1038/sj.embor.7400582 (2005).
94. Stavnezer, J., Guikema, J. E. & Schrader, C. E. Mechanism and regulation of class switch recombination. ***Annu Rev Immunol*** **26,** 261–92. doi:10.1146/annurev.immunol.26.021607.090248 (2008).
95. Storb, U., Peters, A., ***et al.*** Molecular aspects of somatic hypermutation of immunoglobulin genes. ***Cold Spring Harb Symp Quant Biol*** **64,** 227–34. doi:10.1101/sqb.1999.64.227 (1999).
96. Muramatsu, M., Sankaranand, V. S., ***et al.*** Specific Expression of Activation-induced Cytidine Deaminase (AID), a Novel Member of the RNA-editing Deaminase Family in Germinal Center B Cells *. English. ***Journal of Biological Chemistry*** **274.** publisher: Elsevier PMID: 10373455, 18470–18476. doi:10.1074/jbc.274.26.18470 (June 25, 1999).
97. Alvarez-Prado, A. F., Perez-Duran, P., ***et al.*** A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. ***J Exp Med*** **215,** 761–771. doi:10.1084/jem.20171738 (2018).
98. Dorner, T., Foster, S. J., Farner, N. L. & Lipsky, P. E. Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. ***Eur J Immunol*** **28,** 3384–96. doi:10.1002/(SICI)1521-4141(199810)28:10<3384::AID-IMMU3384>3.0.CO;2-T (1998).
99. Pilzecker, B. & Jacobs, H. Mutating for Good: DNA Damage Responses During Somatic Hypermutation. ***Front Immunol*** **10,** 438. doi:10.3389/fimmu.2019.00438 (2019).
100. Rogozin, I. B. & Diaz, M. Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. ***J Immunol*** **172,** 3382–4. doi:10.4049/jimmunol.172.6.3382 (2004).
101. Rogozin, I. B. & Kolchanov, N. A. Somatic hypermutagenesis in immunoglobulin genes: II. Influence of neighbouring base sequences on mutagenesis. en. ***Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*** **1171,** 11–18. doi:10.1016/0167-4781(92)90134-L (Nov. 15, 1992).
102. Di Noia, J. & Neuberger, M. S. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. ***Nature*** **419,** 43–8. doi:10.1038/nature00981 (2002).
103. Neuberger, M. S., Di Noia, J. M., ***et al.*** Somatic hypermutation at A.T pairs: polymerase error versus dUTP incorporation. ***Nat Rev Immunol*** **5,** 171–8. doi:10.1038/nri1553 (2005).
104. Rada, C., Di Noia, J. M. & Neuberger, M. S. Mismatch recognition and uracil excision provide complementary paths to both Ig switching and the A/T-focused phase of somatic mutation. ***Mol Cell*** **16,** 163–71. doi:10.1016/j.molcel.2004.10.011 (2004).
105. Rada, C., Williams, G. T., ***et al.*** Immunoglobulin isotype switching is inhibited and somatic hypermutation perturbed in UNG-deficient mice. ***Curr Biol*** **12,** 1748–55. doi:10.1016/S0960-9822(02)01215-0 (2002).
106. Wiesendanger, M., Kneitz, B., Edelmann, W. & Scharff, M. D. Somatic hypermutation in MutS homologue (MSH)3-, MSH6-, and MSH3/MSH6-

deficient mice reveals a role for the MSH2-MSH6 heterodimer in modulating the base substitution pattern. *J Exp Med* **191,** 579–84. doi:10.1084/jem.191.3.579 (2000).

107. Jansen, J. G., Langerak, P., ***et al.*** Strand-biased defect in C/G transversions in hypermutating immunoglobulin genes in Rev1-deficient mice. *J Exp Med* **203,** 319–23. doi:10.1084/jem.20052227 (2006).
108. Krijger, P. H., Tsaalbi-Shtylik, A., ***et al.*** Rev1 is essential in generating G to C transversions downstream of the Ung2 pathway but not the Msh2+Ung2 hybrid pathway. *Eur J Immunol* **43,** 2765–70. doi:10.1002/eji.201243191 (2013).
109. Tissier, A., Kannouche, P., ***et al.*** Co-localization in replication foci and interaction of human Y-family members, DNA polymerase pol eta and REVl protein. *DNA Repair (Amst)* **3,** 1503–14. doi:10.1016/j.dnarep.2004.06.015 (2004).
110. Delbos, F., Aoufouchi, S., Faili, A., Weill, J. C. & Reynaud, C. A. DNA polymerase eta is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse. *J Exp Med* **204,** 17–23. doi:10.1084/jem.20062131 (2007).
111. Krijger, P. H., Langerak, P., van den Berk, P. C. & Jacobs, H. Dependence of nucleotide substitutions on Ung2, Msh2, and PCNA-Ub during somatic hypermutation. *J Exp Med* **206,** 2603–11. doi:10.1084/jem.20091707 (2009).
112. Langerak, P., Nygren, A. O., Krijger, P. H., van den Berk, P. C. & Jacobs, H. A/T mutagenesis in hypermutated immunoglobulin genes strongly depends on PCNAK164 modification. *J Exp Med* **204,** 1989–98. doi:10.1084/jem.20070902 (2007).
113. Zeng, X., Winter, D. B., ***et al.*** DNA polymerase eta is an A-T mutator in somatic hypermutation of immunoglobulin variable genes. *Nat Immunol* **2,** 537–41. doi:10.1038/88740 (2001).
114. Kunkel, T. A. & Erie, D. A. Eukaryotic Mismatch Repair in Relation to DNA Replication. *Annu Rev Genet* **49,** 291–313. doi:10.1146/annurev-genet-112414-054722 (2015).
115. Bardwell, P. D., Woo, C. J., ***et al.*** Altered somatic hypermutation and reduced class-switch recombination in exonuclease 1-mutant mice. *Nat Immunol* **5,** 224–9. doi:10.1038/ni1031 (2004).
116. Pena-Diaz, J., Bregenhorn, S., ***et al.*** Noncanonical mismatch repair as a source of genomic instability in human cells. *Mol Cell* **47,** 669–80. doi:10.1016/j.molcel.2012.07.006 (2012).
117. Rogozin, I. B., Pavlov, Y. I., Bebenek, K., Matsuda, T. & Kunkel, T. A. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. *Nat Immunol* **2,** 530–6. doi:10.1038/88732 (2001).
118. Wilson, T. M., Vaisman, A., ***et al.*** MSH2-MSH6 stimulates DNA polymerase eta, suggesting a role for A:T mutations in antibody genes. *J Exp Med* **201,** 637–45. doi:10.1084/jem.20042066 (2005).
119. Krijger, P. H., van den Berk, P. C., ***et al.*** PCNA ubiquitination-independent activation of polymerase eta during somatic hypermutation and DNA damage tolerance. *DNA Repair (Amst)* **10,** 1051–9. doi:10.1016/j.dnarep.2011.08.005 (2011).
120. Martomo, S. A., Yang, W. W., ***et al.*** Different mutation signatures in DNA polymerase eta- and MSH6-deficient mice suggest separate roles in antibody diversification. *Proc Natl Acad Sci U S A* **102,** 8656–61. doi:10.1073/pnas.0501852102 (2005).
121. Sale, J. E., Lehmann, A. R. & Woodgate, R. Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nat Rev Mol Cell Biol* **13,** 141–52. doi:10.1038/nrm3289 (2012).
122. Maul, R. W., MacCarthy, T., ***et al.*** DNA polymerase iota functions in the generation of tandem mutations during somatic hypermutation of antibody genes. *J Exp Med* **213,** 1675–83. doi:10.1084/jem.20151227 (2016).
123. Lee, S. S., Tranchina, D., Ohta, Y., Flajnik, M. F. & Hsu, E. Hypermutation in shark immunoglobulin light chain genes results in contiguous substitutions. *Immunity* **16,** 571–82. doi:10.1016/S1074-7613(02)00300-X (2002).
124. Malecek, K., Brandman, J., ***et al.*** Somatic hypermutation and junctional diversification at Ig heavy chain loci in the nurse shark. *J Immunol* **175,** 8105–15. doi:10.4049/jimmunol.175.12.8105 (2005).
125. Chen, J. M., Cooper, D. N. & Ferec, C. A new and more accurate estimate of the rate of concurrent tandem-base substitution mutations in the human germline: 0.4% of the single-nucleotide

substitution mutation rate. ***Hum Mutat* 35,** 392–4. doi:10.1002/humu.22501 (2014).

126. Francioli, L. C., Polak, P. P., ***et al.*** Genome-wide patterns and properties of de novo mutations in humans. ***Nat Genet* 47,** 822–826. doi:10.1038/ng.3292 (2015).
127. Harris, K. & Nielsen, R. Error-prone polymerase activity causes multinucleotide mutations in humans. ***Genome Res* 24,** 1445–54. doi:10.1101/gr.170696.113 (2014).
128. Schrider, D. R., Hourmozdi, J. N. & Hahn, M. W. Pervasive multinucleotide mutational events in eukaryotes. ***Curr Biol* 21,** 1051–4. doi:10.1016/j.cub.2011.05.013 (2011).
129. Saribasak, H., Maul, R. W., ***et al.*** DNA polymerase zeta generates tandem mutations in immunoglobulin variable regions. ***J Exp Med* 209,** 1075–81. doi:10.1084/jem.20112234 (2012).
130. Alexandrov, L. B., Kim, J., ***et al.*** The Repertoire of Mutational Signatures in Human Cancer. ***bioRxiv,*** 322859. doi:10.1101/322859 (2018).
131. Koning, M. T., Kielbasa, S. M., ***et al.*** ARTISAN PCR: rapid identification of full-length immunoglobulin rearrangements without primer binding bias. ***Br J Haematol* 178,** 983–986. doi:10.1111/bjh.14180 (2017).
132. Koning, M. T., Trollmann, I. J. M., ***et al.*** Peripheral IgE Repertoires of Healthy Donors Carry Moderate Mutation Loads and Do Not Overlap With Other Isotypes. ***Front Immunol* 10,** 1543. doi:10.3389/fimmu.2019.01543 (2019).
133. Alamyar, E., Duroux, P., Lefranc, M. P. & Giudicelli, V. IMGT((R)) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. ***Methods Mol Biol* 882,** 569–604. doi:10.1007/978-1-61779-842-9_32 (2012).
134. IJspeert, H., van Schouwenburg, P. A., ***et al.*** Repertoire Sequencing of B Cells Elucidates the Role of UNG and Mismatch Repair Proteins in Somatic Hypermutation in Humans. ***Front Immunol* 10,** 1913. doi:10.3389/fimmu.2019.01913 (2019).
135. IJspeert, H., van Schouwenburg, P. A., ***et al.*** Evaluation of the Antigen-Experienced B-Cell Receptor Repertoire in Healthy Children and Adults. ***Front Immunol* 7,** 410. doi:10.3389/fimmu.2016.00410 (2016).
136. Driessen, G. J., Ijspeert, H., ***et al.*** Antibody deficiency in patients with ataxia telangiectasia is caused by disturbed B- and T-cell homeostasis and reduced immune repertoire diversity. ***J Allergy Clin Immunol* 131,** 1367–75 e9. doi:10.1016/j.jaci.2013.01.053 (2013).
137. Alexandrov, L. B., Kim, J., ***et al.*** The repertoire of mutational signatures in human cancer. ***Nature* 578,** 94–101. doi:10.1038/s41586-020-1943-3 (2020).
138. R Core Team. ***R: A Language and Environment for Statistical Computing*** R Foundation for Statistical Computing (Vienna, Austria, 2021).
139. Jacobs, H., Fukita, Y., ***et al.*** Hypermutation of immunoglobulin genes in memory B cells of DNA repair-deficient mice. ***J Exp Med* 187,** 1735–43. doi:10.1084/jem.187.11.1735 (1998).
140. Sohail, A., Klapacz, J., Samaranayake, M., Ullah, A. & Bhagwat, A. S. Human activation-induced cytidine deaminase causes transcription-dependent, strand-biased C to U deaminations. ***Nucleic Acids Res* 31,** 2990–4. doi:10.1093/nar/gkg464 (2003).
141. Kasar, S., Kim, J., ***et al.*** Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. ***Nat Commun* 6,** 8866. doi:10.1038/ncomms9866 (2015).
142. Maul, R. W. & Gearhart, P. J. Refining the Neuberger model: Uracil processing by activated B cells. ***Eur J Immunol* 44,** 1913–6. doi:10.1002/eji.201444813 (2014).
143. Tesch, V. K., H, I. J., ***et al.*** No Overt Clinical Immunodeficiency Despite Immune Biological Abnormalities in Patients With Constitutional Mismatch Repair Deficiency. ***Front Immunol* 9,** 1506. doi:10.3389/fimmu.2018.01506 (2018).
144. Wimmer, K., Kratz, C. P., ***et al.*** Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium 'care for CMMRD' (C4CMMRD). ***J Med Genet* 51,** 355–65. doi:10.1136/jmedgenet-2014-102284 (2014).
145. Bahjat, M., Stratigopoulou, M., ***et al.*** DNA polymerase β prevents AID-instigated mutagenic non-canonical mismatch DNA repair. ***bioRxiv*** (2020).
146. Zharkov, D. O., Mechetin, G. V. & Nevinsky, G. A. Uracil-DNA glycosylase: Structural, thermodynamic and kinetic aspects of lesion search and

recognition. ***Mutat Res* 685,** 11–20. doi:10.1016/j.mrfmmm.2009.10.017 (2010).

147. Corbella, M., Voityuk, A. A. & Curutchet, C. How abasic sites impact hole transfer dynamics in GC-rich DNA sequences. ***Phys Chem Chem Phys* 20,** 23123–23131. doi:10.1039/c8cp03572e (2018).
148. Yin, Y., Yang, L., ***et al.*** Dynamics of spontaneous flipping of a mismatched base in DNA duplex. ***Proc Natl Acad Sci U S A* 111,** 8043–8. doi:10.1073/pnas.1400667111 (2014).
149. Cao, C., Jiang, Y. L., Stivers, J. T. & Song, F. Dynamic opening of DNA during the enzymatic search for a damaged base. ***Nat Struct Mol Biol* 11,** 1230–6. doi:10.1038/nsmb864 (2004).
150. Carr, C. E., Khutsishvili, I., Gold, B. & Marky, L. A. Thermodynamic Stability of DNA Duplexes Comprising the Simplest T –> dU Substitutions. ***Biochemistry* 57,** 5666–5671. doi:10.1021/acs.biochem.8b00676 (2018).
151. Parker, J. B., Bianchet, M. A., ***et al.*** Enzymatic capture of an extrahelical thymine in the search for uracil in DNA. ***Nature* 449,** 433–7. doi:10.1038/nature06131 (2007).
152. Fuxreiter, M., Luo, N., Jedlovszky, P., Simon, I. & Osman, R. Role of base flipping in specific recognition of damaged DNA by repair enzymes. ***J Mol Biol* 323,** 823–34. doi:10.1016/s0022-2836(02)00999-3 (2002).
153. Imhof, P. & Zahran, M. The effect of a G:T mispair on the dynamics of DNA. ***PLoS One* 8,** e53305. doi:10.1371/journal.pone.0053305 (2013).
154. Varnai, P., Canalia, M. & Leroy, J. L. Opening mechanism of G.T/U pairs in DNA and RNA duplexes: a combined study of imino proton exchange and molecular dynamics simulation. ***J Am Chem Soc* 126,** 14659–67. doi:10.1021/ja0470721 (2004).
155. Chen, J. M., Ferec, C. & Cooper, D. N. Patterns and mutational signatures of tandem base substitutions causing human inherited disease. ***Hum Mutat* 34,** 1119–30. doi:10.1002/humu.22341 (2013).
156. Stone, J. E., Lujan, S. A., Kunkel, T. A. & Kunkel, T. A. DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in Saccharomyces cerevisiae. ***Environ Mol Mutagen* 53,** 777–86. doi:10.1002/em.21728 (2012).
157. Faili, A., Aoufouchi, S., ***et al.*** Induction of somatic hypermutation in immunoglobulin genes is dependent on DNA polymerase iota. ***Nature* 419,** 944–7. doi:10.1038/nature01117 (2002).
158. Weill, J. C. & Reynaud, C. A. DNA polymerases in adaptive immunity. ***Nat Rev Immunol* 8,** 302–12. doi:10.1038/nri2281 (2008).
159. Biertumpfel, C., Zhao, Y., ***et al.*** Structure and mechanism of human DNA polymerase eta. ***Nature* 465,** 1044–8. doi:10.1038/nature09196 (2010).
160. Matsuda, T., Bebenek, K., ***et al.*** Error rate and specificity of human and murine DNA polymerase eta. ***J Mol Biol* 312,** 335–46. doi:10.1006/jmbi.2001.4937 (2001).
161. Davies, E. G. & Thrasher, A. J. Update on the hyper immunoglobulin M syndromes. ***Br J Haematol* 149,** 167–80. doi:10.1111/j.1365-2141.2010.08077.x (2010).
162. Imai, K., Slupphaug, G., ***et al.*** Human uracil-DNA glycosylase deficiency associated with profoundly impaired immunoglobulin class-switch recombination. ***Nat Immunol* 4,** 1023–8. doi:10.1038/ni974 (2003).
163. Sepúlveda-Yáñez, J. H., Alvarez-Saravia, D., ***et al.*** ***Detection of occurrent somatic hypermutation and associated gene expression profile in single*** Manuscript submitted for publication.
164. Pavri, R. & Nussenzweig, M. C. en. in ***Advances in Immunology*** (eds Alt, F. W., Austen, K. F., ***et al.***) DOI: 10.1016/B978-0-12-387663-8.00005-3, 1–26 (Academic Press, Jan. 1, 2011).
165. Rada, C. & Milstein, C. The intrinsic hypermutability of antibody heavy and light chain genes decays exponentially. eng. ***The EMBO journal* 20.** PMID: 11500383 PMCID: PMC125579, 4570–4576. doi:10.1093/emboj/20.16.4570 (Aug. 15, 2001).
166. Rajewsky, K., Förster, I. & Cumano, A. Evolutionary and somatic selection of the antibody repertoire in the mouse. eng. ***Science (New York, N.Y.)* 238.** PMID: 3317826, 1088–1094. doi:10.1126/science.3317826 (Nov. 20, 1987).
167. Hardianti, M. S., Tatsumi, E., ***et al.*** Activation-induced cytidine deaminase expression in follicular lymphoma: association between AID expression and ongoing mutation in FL. ***Leukemia* 18,** 826–831. doi:10.1038/sj.leu.2403323 (Apr. 2004).

168. Noppe, S. M., Heirman, C., ***et al.*** The genetic variability of the VH genes in follicular lymphoma: the impact of the hypermutation mechanism. en. ***British Journal of Haematology*** **107,** 625–640. doi:10.1046/j.1365-2141.1999.01732.x (1999).
169. Scherer, F., Navarrete, M. A., ***et al.*** Isotype-switched follicular lymphoma displays dissociation between activation-induced cytidine deaminase expression and somatic hypermutation. en. ***Leukemia & Lymphoma*** **57,** 151–160. doi:10.3109/10428194.2015.1037758 (Jan. 2, 2016).
170. Zuckerman, N. S., McCann, K. J., ***et al.*** Ig gene diversification and selection in follicular lymphoma, diffuse large B cell lymphoma and primary central nervous system lymphoma revealed by lineage tree and mutation analyses. ***International Immunology*** **22,** 875–887. doi:10.1093/intimm/dxq441 (Nov. 1, 2010).
171. Hao, Y., Hao, S., ***et al.*** Integrated analysis of multimodal single-cell data. English. ***Cell*** **184.** publisher: Elsevier PMID: 34062119, 3573–3587.e29. doi:10.1016/j.cell.2021.04.048 (June 24, 2021).
172. Germain, P.-L., Lun, A., Meixide, C. G., Macnair, W. & Robinson, M. D. ***Doublet identification in single-cell sequencing data using scDblFinder*** en. Tech. rep. DOI: 10.12688/f1000research.73600.2 type: article (May 16, 2022).
173. Li, H. New strategies to improve minimap2 alignment accuracy. ***Bioinformatics*** **37,** 4572–4574. doi:10.1093/bioinformatics/btab705 (Dec. 1, 2021).
174. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. ***Genome Biology*** **15,** 550. doi:10.1186/s13059-014-0550-8 (Dec. 5, 2014).
175. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. ***Genome Biology*** **17,** 75. doi:10.1186/s13059-016-0947-7 (Apr. 27, 2016).
176. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. ***Nucleic Acids Research*** **51,** D587–D592. doi:10.1093/nar/gkac963 (Jan. 6, 2023).
177. Wu, T., Hu, E., ***et al.*** clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. English. ***The Innovation*** **2.** publisher: Elsevier PMID: 34557778. doi:10.1016/j.xinn.2021.100141 (Aug. 28, 2021).
178. Petersen-Mahrt, S. K., Harris, R. S. & Neuberger, M. S. AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. en. ***Nature*** **418.** number: 6893 publisher: Nature Publishing Group, 99–104. doi:10.1038/nature00862 (July 2002).
179. Rada, C., Di Noia, J. M. & Neuberger, M. S. Mismatch Recognition and Uracil Excision Provide Complementary Paths to Both Ig Switching and the A/T-Focused Phase of Somatic Mutation. en. ***Molecular Cell*** **16,** 163–171. doi:10.1016/j.molcel.2004.10.011 (Oct. 22, 2004).
180. Shen, H. M., Tanaka, A., Bozek, G., Nicolae, D. & Storb, U. Somatic Hypermutation and Class Switch Recombination in Msh6-/-Ung-/- Double-Knockout Mice1. ***The Journal of Immunology*** **177,** 5386–5392. doi:10.4049/jimmunol.177.8.5386 (Oct. 15, 2006).
181. Sepulveda-Yanez, J. H., Alvarez-Saravia, D., ***et al.*** Integration of Mutational Signature Analysis with 3D Chromatin Data Unveils Differential AID-Related Mutagenesis in Indolent Lymphomas. en. ***International Journal of Molecular Sciences*** **22.** number: 23 publisher: Multidisciplinary Digital Publishing Institute, 13015. doi:10.3390/ijms222313015 (Jan. 2021).
182. Rogozin, I. B., Pavlov, Y. I., Bebenek, K., Matsuda, T. & Kunkel, T. A. Somatic mutation hotspots correlate with DNA polymerase η error spectrum. en. ***Nature Immunology*** **2.** number: 6 publisher: Nature Publishing Group, 530–536. doi:10.1038/88732 (June 2001).
183. Tang, C., Bagnara, D., Chiorazzi, N., Scharff, M. D. & MacCarthy, T. AID Overlapping and Polη Hotspots Are Key Features of Evolutionary Variation Within the Human Antibody Heavy Chain (IGHV) Genes. ***Frontiers in Immunology*** **11.** [Online; accessed 2024-01-05] (2020).
184. Rogozin, I. B., Lada, A. G., ***et al.*** Activation induced deaminase mutational signature overlaps with CpG methylation sites in follicular lymphoma and other cancers. ***Scientific Reports*** **6,** 38133. doi:10.1038/srep38133 (Dec. 7, 2016).
185. Genovese, C. & Milcarek, C. Increased half-life of mu immunoglobulin mRNA during mouse B cell development increases its abundancy. eng. ***Molecular Immunology*** **27.** PMID: 2119479,

733–743. doi:10.1016/0161-5890(90)90082-b (Aug. 1990).

186. Mason, J. O., Williams, G. T. & Neuberger, M. S. The half-life of immunoglobulin mRNA increases during B-cell differentiation: a possible role for targeting to membrane-bound polysomes. en. ***Genes & Development,*** 1003–1011. doi:10.1101/gad.2.8.1003 (Aug. 1988).
187. Nagaoka, H., Muramatsu, M., Yamamura, N., Kinoshita, K. & Honjo, T. Activation-induced deaminase (AID)-directed hypermutation in the immunoglobulin Smu region: implication of AID involvement in a common step of class switch recombination and somatic hypermutation. eng. ***The Journal of Experimental Medicine*** **195.** PMID: 11854365 PMCID: PMC2193625, 529–534. doi:10.1084/jem.20012144 (Feb. 18, 2002).
188. Wilson, T. M., Vaisman, A., ***et al.*** MSH2–MSH6 stimulates DNA polymerase η, suggesting a role for A:T mutations in antibody genes. ***Journal of Experimental Medicine*** **201,** 637–645. doi:10.1084/jem.20042066 (Feb. 14, 2005).
189. Rada, C., Williams, G. T., ***et al.*** Immunoglobulin Isotype Switching Is Inhibited and Somatic Hypermutation Perturbed in UNG-Deficient Mice. en. ***Current Biology*** **12,** 1748–1755. doi:10.1016/S0960-9822(02)01215-0 (Oct. 2002).
190. Chahwan, R., Edelmann, W., Scharff, M. D. & Roa, S. AIDing antibody diversity by error-prone mismatch repair. ***Seminars in immunology*** **24.** PMID: 22703640 PMCID: PMC3422444, 293–300. doi:10.1016/j.smim.2012.05.005 (Aug. 2012).
191. Seifert, M., Sellmann, L., ***et al.*** Cellular origin and pathophysiology of chronic lymphocytic leukemia. eng. ***The Journal of Experimental Medicine*** **209.** PMID: 23091163 PMCID: PMC3501361, 2183–2198. doi:10.1084/jem.20120833 (Nov. 19, 2012).
192. Pilzecker, B. & Jacobs, H. Mutating for Good: DNA Damage Responses During Somatic Hypermutation. ***Frontiers in Immunology*** **10.** [Online; accessed 2024-01-22] (2019).
193. Senavirathne, G., Bertram, J. G., ***et al.*** Activation-induced deoxycytidine deaminase (AID) co-transcriptional scanning at single-molecule resolution. ***Nature Communications*** **6.** PMID: 26681117 PMCID: PMC4703863, 10209. doi:10.1038/ncomms10209 (Dec. 18, 2015).
194. Zahn, A., Eranki, A. K., ***et al.*** Activation induced deaminase C-terminal domain links DNA breaks to end protection and repair during class switch recombination. en. ***Proceedings of the National Academy of Sciences*** **111.** [Online; accessed 2024-01-04]. doi:10.1073/pnas.1320486111 (Mar. 18, 2014).
195. Xu, Z., Zan, H., Pal, Z. & Casali, P. DNA REPLICATION TO AID SOMATIC HYPERMUTATION. ***Advances in experimental medicine and biology*** **596.** PMID: 17338180 PMCID: PMC3140876, 111–127. doi:10.1007/0-387-46530-8_10 (2007).
196. Victora, G. D., Dominguez-Sola, D., ***et al.*** Identification of human germinal center light and dark zone cells and their relationship to human B-cell lymphomas. ***Blood*** **120,** 2240–2248. doi:10.1182/blood-2012-03-415380 (Sept. 13, 2012).
197. Basso, K., Klein, U., ***et al.*** Tracking CD40 signaling during germinal center development. en. ***Blood*** **104,** 4088–4096. doi:10.1182/blood-2003-12-4291 (Dec. 15, 2004).
198. Gehringer, F., Weissinger, S. E., ***et al.*** FOXO1 Confers Maintenance of the Dark Zone Proliferation and Survival Program and Can Be Pharmacologically Targeted in Burkitt Lymphoma. ***Cancers*** **11.** PMID: 31557894 PMCID: PMC6826697, 1427. doi:10.3390/cancers11101427 (Sept. 25, 2019).
199. Wood, R. D., Gearhart, P. J., Neuberger, M. S., Tippin, B. & Goodman, M. F. A new class of errant DNA polymerases provides candidates for somatic hypermutation. ***Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*** **356.** publisher: Royal Society, 47–51. doi:10.1098/rstb.2000.0747 (Jan. 29, 2001).
200. Lehmann, A. R., Niimi, A., ***et al.*** Translesion synthesis: Y-family polymerases and the polymerase switch. eng. ***DNA repair*** **6.** PMID: 17363342, 891–899. doi:10.1016/j.dnarep.2007.02.003 (July 1, 2007).
201. Biertümpfel, C., Zhao, Y., ***et al.*** Structure and mechanism of human DNA polymerase η. en. ***Nature*** **465.** number: 7301 publisher: Nature Publishing Group, 1044–1048. doi:10.1038/nature09196 (June 2010).
202. Sale, J. E. Translesion DNA Synthesis and Mutagenesis in Eukaryotes. en. ***Cold Spring Harbor Perspectives in Biology*** **5.** Company: Cold Spring Harbor Laboratory Press Distributor:

Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press publisher: Cold Spring Harbor Lab PMID: 23457261, a012708. doi:10.1101/cshperspect.a012708 (Mar. 1, 2013).

203. Lohr, J. G., Stojanov, P., *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. ***Proceedings of the National Academy of Sciences* 109.** publisher: Proceedings of the National Academy of Sciences, 3879–3884. doi:10.1073/pnas.1121343109 (Mar. 6, 2012).
204. Morande, P. E., Yan, X.-J., *et al.* AID overexpression leads to aggressive murine CLL and non-immunoglobulin mutations that mirror human neoplasms. ***Blood* 138,** 246–258. doi:10.1182/blood.2020008654 (July 22, 2021).
206. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. ***Cell* 144,** 646–74. doi:10.1016/j.cell.2011.02.013 (2011).
207. Hainaut, P. & Pfeifer, G. P. Patterns of p53 G–>T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. ***Carcinogenesis* 22,** 367–74. doi:10.1093/carcin/22.3.367 (2001).
208. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. ***Nature* 458,** 719–24. doi:10.1038/nature07943 (2009).
209. Alexandrov, L. B., Nik-Zainal, S., *et al.* Signatures of mutational processes in human cancer. ***Nature* 500,** 415–21. doi:10.1038/nature12477 (2013).
210. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. ***Cell Rep* 3,** 246–59. doi:10.1016/j.celrep.2012.12.008 (2013).
211. Petersen-Mahrt, S. K., Harris, R. S. & Neuberger, M. S. AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. ***Nature* 418,** 99–103. doi:10.1038/nature00862 (2002).
212. Winter, D. B. & Gearhart, P. J. Altered spectra of hypermutation in DNA repair-deficient mice. ***Philos Trans R Soc Lond B Biol Sci* 356,** 5–11. doi:10.1098/rstb.2000.0742 (2001).
213. Rogozin, I. B., Lada, A. G., *et al.* Activation induced deaminase mutational signature overlaps with CpG methylation sites in follicular lymphoma and other cancers. ***Sci Rep* 6,** 38133. doi:10.1038/srep38133 (2016).
214. Qian, J., Wang, Q., *et al.* B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. ***Cell* 159,** 1524–37. doi:10.1016/j.cell.2014.11.013 (2014).
215. Lieberman-Aiden, E., van Berkum, N. L., *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. ***Science* 326,** 289–93. doi:10.1126/science.1181369 (2009).
216. Rao, S. S., Huntley, M. H., *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. ***Cell* 159,** 1665–80. doi:10.1016/j.cell.2014.11.021 (2014).
217. Teras, L. R., DeSantis, C. E., *et al.* 2016 US lymphoid malignancy statistics by World Health Organization subtypes. ***CA Cancer J Clin* 66,** 443–459. doi:10.3322/caac.21357 (2016).
218. Scherer, F., van der Burgt, M., *et al.* Selection patterns of B-cell receptors and the natural history of follicular lymphoma. ***Br J Haematol* 175,** 972–975. doi:10.1111/bjh.13901 (2016).
219. Scherer, F., Navarrete, M. A., *et al.* Isotype-switched follicular lymphoma displays dissociation between activation-induced cytidine deaminase expression and somatic hypermutation. ***Leuk Lymphoma* 57,** 151–60. doi:10.3109/10428194.2015.1037758 (2016).
220. Oppezzo, P., Vuillier, F., *et al.* Chronic lymphocytic leukemia B cells expressing AID display dissociation between class switch recombination and somatic hypermutation. ***Blood* 101,** 4029–4032. doi:10.1182/blood-2002-10-3175 (2003).
221. Burns, A., Alsolami, R., *et al.* Whole-genome sequencing of chronic lymphocytic leukaemia reveals distinct differences in the mutational landscape between IgHV(mut) and IgHV(unmut) subgroups. ***Leukemia* 32,** 332–342. doi:10.1038/leu.2017.177 (2018).
222. Lappalainen, I., Almeida-King, J., *et al.* The European Genome-phenome Archive of human data consented for biomedical research. ***Nat Genet* 47,** 692–5. doi:10.1038/ng.3312 (2015).
223. Okosun, J., Wolfson, R. L., *et al.* Recurrent mTORC1-activating RRAGC mutations in follicular lymphoma. ***Nat Genet* 48,** 183–8. doi:10.1038/ng.3473 (2016).

224. Damm, F., Mylonas, E., ***et al.*** Acquired initiating mutations in early hematopoietic cells of CLL patients. ***Cancer Discov*** **4,** 1088–101. doi:10.1158/2159-8290.CD-14-0104 (2014).
225. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. ***Bioinformatics*** **25,** 1754–60. doi:10.1093/bioinformatics/btp324 (2009).
226. McKenna, A., Hanna, M., ***et al.*** The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. ***Genome Res*** **20,** 1297–303. doi:10.1101/gr.107524.110 (2010).
227. Koboldt, D. C., Zhang, Q., ***et al.*** VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. ***Genome Res*** **22,** 568–76. doi:10.1101/gr.129684.111 (2012).
228. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. ***Nucleic Acids Res*** **38,** e164. doi:10.1093/nar/gkq603 (2010).
229. Genomes Project, C., Auton, A., ***et al.*** A global reference for human genetic variation. ***Nature*** **526,** 68–74. doi:10.1038/nature15393 (2015).
230. Lek, M., Karczewski, K. J., ***et al.*** Analysis of protein-coding genetic variation in 60,706 humans. ***Nature*** **536,** 285–91. doi:10.1038/nature19057 (2016).
231. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. ***Nat Protoc*** **11,** 1–9. doi:10.1038/nprot.2015.123 (2016).
232. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. ***Genome Res*** **19,** 1553–61. doi:10.1101/gr.092619.109 (2009).
233. Andrews, S., Krueger, F., ***et al.*** ***FastQC*** 2010.
234. Bergstrom, E. N., Huang, M. N., ***et al.*** SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. ***BMC Genomics*** **20,** 685. doi:10.1186/s12864-019-6041-2 (2019).
235. Tate, J. G., Bamford, S., ***et al.*** COSMIC: the Catalogue Of Somatic Mutations In Cancer. ***Nucleic Acids Res*** **47,** D941–D947. doi:10.1093/nar/gky1015 (2019).
236. Virtanen, P., Gommers, R., ***et al.*** SciPy 1.0: fundamental algorithms for scientific computing in Python. ***Nat Methods*** **17,** 261–272. doi:10.1038/s41592-019-0686-2 (2020).
237. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. ***Genome Biol*** **17,** 31. doi:10.1186/s13059-016-0893-4 (2016).
238. Lawson, R. G. & Jurs, P. C. New Index for Clustering Tendency and Its Application to Chemical Problems. ***Journal of Chemical Information and Computer Sciences*** **30,** 36–41. doi:DOI10.1021/ci00065a010 (1990).
239. Kaufman, L. & Rousseeuw, P. ***Finding Groups in Data: An Introduction to Cluster Analysis*** (John Wiley & Sons, 2009).
240. Lê, S., Josse, J. & Husson, F. FactoMineR: AnR-Package for Multivariate Analysis. ***Journal of Statistical Software*** **25,** Journal of Statistical Software. doi:10.18637/jss.v025.i01 (2008).
241. Xiong, K. & Ma, J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. ***Nat Commun*** **10,** 5069. doi:10.1038/s41467-019-12954-4 (2019).
242. Pedersen, B. S., Layer, R. M. & Quinlan, A. R. Vcfanno: fast, flexible annotation of genetic variants. ***Genome biology*** **17,** 1–9. doi:10.1186/s13059-016-0973-5 (2016).
243. Methot, S. P. & Di Noia, J. M. Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. ***Adv Immunol*** **133,** 37–87. doi:10.1016/bs.ai.2016.11.002 (2017).
244. Jansen, H., Verhoeven, A. J., ***et al.*** Common C-to-T substitution at position -480 of the hepatic lipase promoter associated with a lowered lipase activity in coronary artery disease patients. ***Arterioscler Thromb Vasc Biol*** **17,** 2837–42. doi:10.1161/01.atv.17.11.2837 (1997).
245. Cancer Genome Atlas Research, N., Weinstein, J. N., ***et al.*** The Cancer Genome Atlas Pan-Cancer analysis project. ***Nat Genet*** **45,** 1113–20. doi:10.1038/ng.2764 (2013).
246. Maura, F., Degasperi, A., ***et al.*** A practical guide for mutational signature analysis in hematological malignancies. ***Nature communications*** **10,** 2969. doi:10.1038/s41467-019-11037-8 (2019).
247. Das, R. & Ghosh, S. K. Genetic variants of the DNA repair genes from Exome Aggregation Consortium (EXAC) database: significance in cancer.

DNA Repair (Amst) **52,** 92–102. doi:10.1016/j.dnarep.2017.02.013 (2017).

248. Tsukamoto, T., Nakano, M., ***et al.*** High-risk follicular lymphomas harbour more somatic mutations including those in the AID-motif. ***Sci Rep*** **7,** 14039. doi:10.1038/s41598-017-14150-0 (2017).
249. Jia, P., Pao, W. & Zhao, Z. Patterns and processes of somatic mutations in nine major cancers. ***BMC Med Genomics*** **7,** 11. doi:10.1186/1755-8794-7-11 (2014).
250. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. ***Nature*** **401,** 788–91. doi:10.1038/44565 (1999).
251. Alexandrov, L. B., Jones, P. H., ***et al.*** Clock-like mutational processes in human somatic cells. ***Nat Genet*** **47,** 1402–7. doi:10.1038/ng.3441 (2015).
252. Meng, F. L., Du, Z., ***et al.*** Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. ***Cell*** **159,** 1538–48. doi:10.1016/j.cell.2014.11.014 (2014).
253. Pasqualucci, L., Guglielmino, R., ***et al.*** Expression of the AID protein in normal and neoplastic B cells. ***Blood*** **104,** 3318–25. doi:10.1182/blood-2004-04-1558 (2004).
254. Nagai, L. A. E., Park, S. J. & Nakai, K. Analyzing the 3D chromatin organization coordinating with gene expression regulation in B-cell lymphoma. ***BMC Med Genomics*** **11,** 127. doi:10.1186/s12920-018-0437-8 (2019).
255. Bogliolo, M., Bluteau, D., ***et al.*** Biallelic truncating FANCM mutations cause early-onset cancer but not Fanconi anemia. ***Genet Med*** **20,** 458–463. doi:10.1038/gim.2017.124 (2018).
256. Bret, C., Klein, B., ***et al.*** DNA repair in diffuse large B-cell lymphoma: a molecular portrait. ***Br J Haematol*** **169,** 296–9. doi:10.1111/bjh.13206 (2015).
257. Leeksma, O. C., de Miranda, N. F. & Veelken, H. Germline mutations predisposing to diffuse large B-cell lymphoma. ***Blood Cancer J*** **7,** e532. doi:10.1038/bcj.2017.15 (2017).
258. Quinten, E., Sepúlveda-Yáñez, J. H., ***et al.*** Autonomous B-cell receptor signaling and genetic aberrations in chronic lymphocytic leukemia-phenotype monoclonal B lymphocytosis in siblings of patients with chronic lymphocytic leukemia. ***Haematologica*** **109.** Section: Articles, 824–834. doi:10.3324/haematol.2022.282542 (July 2023).
259. Hallek, M., Shanafelt, T. D. & Eichhorst, B. Chronic lymphocytic leukaemia. English. ***The Lancet*** **391.** publisher: Elsevier PMID: 29477250, 1524–1537. doi:10.1016/S0140-6736(18)30422-7 (Apr. 14, 2018).
260. Delgado, J., Nadeu, F., Colomer, D. & Campo, E. Chronic lymphocytic leukemia: from molecular pathogenesis to novel therapeutic strategies. ***Haematologica*** **105.** PMID: 33054046 PMCID: PMC7556519, 2205–2217. doi:10.3324/haematol.2019.236000 (July 2, 2020).
261. Slager, S. L., Lanasa, M. C., ***et al.*** Natural history of monoclonal B-cell lymphocytosis among relatives in CLL families. ***Blood*** **137,** 2046–2056. doi:10.1182/blood.2020006322 (Apr. 2021).
262. Minici, C., Gounari, M., ***et al.*** Distinct homotypic B-cell receptor interactions shape the outcome of chronic lymphocytic leukaemia. eng. ***Nat Commun*** **8,** 15746. doi:10.1038/ncomms15746 (June 2017).
263. Agathangelidis, A., Chatzidimitriou, A., ***et al.*** Higher-order connections between stereotyped subsets: implications for improved patient classification in CLL. ***Blood*** **137,** 1365–1376. doi:10.1182/blood.2020007039 (Mar. 2021).
264. Cerhan, J. R. & Slager, S. L. Familial predisposition and genetic risk factors for lymphoma. ***Blood*** **126,** 2265–73. doi:10.1182/blood-2015-04-537498 (Nov. 2015).
265. Kleinstern, G., Camp, N. J., ***et al.*** Association of polygenic risk score with the risk of chronic lymphocytic leukemia and monoclonal B-cell lymphocytosis. ***Blood*** **131,** 2541–2551. doi:10.1182/blood-2017-11-814608 (June 2018).
266. Law, P. J., Berndt, S. I., ***et al.*** Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. ***Nat Commun*** **8,** 14175. doi:10.1038/ncomms14175 (Feb. 2017).
267. Agathangelidis, A., Ljungstrom, V., ***et al.*** Highly similar genomic landscapes in monoclonal B-cell lymphocytosis and ultra-stable chronic lymphocytic leukemia with low frequency of driver mutations. ***Haematologica*** **103,** 865–873. doi:10.3324/haematol.2017.177212 (May 2018).
268. Koning, M. T., Kielbasa, S. M., ***et al.*** ARTISAN PCR: rapid identification of full-length immunoglobulin rearrangements without primer binding

bias. *Br J Haematol* **178,** 983–986. doi:10.1111/bjh.14180 (Sept. 2017).

269. Koning, M. T., Quinten, E., *et al.* Acquired N-Linked Glycosylation Motifs in B-Cell Receptors of Primary Cutaneous B-Cell Lymphoma and the Normal B-Cell Repertoire. *J Invest Dermatol* **139,** 2195–2203. doi:10.1016/j.jid.2019.04.005 (Oct. 2019).
270. Brochet, X., Lefranc, M. P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. eng. *Nucleic Acids Res* **36,** W503–8. doi:10.1093/nar/gkn316 (July 2008).
271. Giudicelli, V., Brochet, X. & Lefranc, M. P. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* **2011,** 695–715. doi:10.1101/pdb.prot5633 (June 2011).
272. Agathangelidis, A., Darzentas, N., *et al.* Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. *Blood* **119,** 4467–75. doi:10.1182/blood-2011-11-393694 (May 2012).
273. Meixlsperger, S., Kohler, F., *et al.* Conventional light chains inhibit the autonomous signaling capacity of the B cell receptor. eng. *Immunity* **26,** 323–33. doi:10.1016/j.immuni.2007.01.012 (Mar. 2007).
274. De Groen, R. A. L., van Eijk, R., *et al.* Frequent mutated B2M, EZH2, IRF8, and TNFRSF14 in primary bone diffuse large B-cell lymphoma reflect a GCB phenotype. *Blood Adv* **5,** 3760–3775. doi:10.1182/bloodadvances.2021005215 (Oct. 2021).
275. Pfeifer, D., Pantic, M., *et al.* Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. *Blood* **109,** 1202–10 (Feb. 2007).
276. Berndt, S. I., Skibola, C. F., *et al.* Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nat Genet* **45,** 868–76. doi:10.1038/ng.2652 (Aug. 2013).
277. Crowther-Swanepoel, D., Broderick, P., *et al.* Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet* **42,** 132–6. doi:10.1038/ng.510 (Feb. 2010).
278. Di Bernardo, M. C., Crowther-Swanepoel, D., *et al.* A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nat Genet* **40,** 1204–10. doi:10.1038/ng.219 (Oct. 2008).
279. Slager, S. L., Rabe, K. G., *et al.* Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL. *Blood* **117,** 1911–6. doi:10.1182/blood-2010-09-308205 (Feb. 2011).
280. Speedy, H. E., Di Bernardo, M. C., *et al.* A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nat Genet* **46,** 56–60. doi:10.1038/ng.2843 (Jan. 2014).
281. Karczewski, K. J., Francioli, L. C., *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581,** 434–443. doi:10.1038/s41586-020-2308-7 (May 2020).
282. Strati, P. & Shanafelt, T. D. Monoclonal B-cell lymphocytosis and early-stage chronic lymphocytic leukemia: diagnosis, natural history, and risk stratification. *Blood* **126,** 454–62. doi:10.1182/blood-2015-02-585059 (July 2015).
283. Stamatopoulos, K., Belessi, C., *et al.* Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: Pathogenetic implications and clinical correlations. *Blood* **109,** 259–70. doi:10.1182/blood-2006-03-012948 (Jan. 2007).
284. Binder, M., Müller, F., *et al.* CLL B-cell receptors can recognize themselves: alternative epitopes and structural clues for autostimulatory mechanisms in CLL. eng. *Blood* **121,** 239–241. doi:10.1182/blood-2012-09-454439 (Jan. 2013).
285. Lanasa, M. C., Allgood, S. D., *et al.* Single-cell analysis reveals oligoclonality among 'low-count' monoclonal B-cell lymphocytosis. *Leukemia* **24,** 133–40. doi:10.1038/leu.2009.192 (Jan. 2010).
286. Vardi, A., Dagklis, A., *et al.* Immunogenetics shows that not all MBL are equal: the larger the clone, the more similar to CLL. *Blood* **121,** 4521–8. doi:10.1182/blood-2012-12-471698 (May 2013).
287. Kolijn, P. M., Hosnijeh, F. S., *et al.* High-risk subtypes of chronic lymphocytic leukemia are detectable as early as 16 years prior to diagnosis.

Blood **139,** 1557–1563. doi:10.1182/blood.2021012890 (Mar. 2022).

288. Fazi, C., Scarfo, L., ***et al.*** General population low-count CLL-like MBL persists over time without clinical progression, although carrying the same cytogenetic abnormalities of CLL. ***Blood*** **118,** 6618–25. doi:10.1182/blood-2011-05-357251 (Dec. 2011).
289. Criado, I., Rodriguez-Caballero, A., ***et al.*** Low-count monoclonal B-cell lymphocytosis persists after seven years of follow up and is associated with a poorer outcome. ***Haematologica*** **103,** 1198–1208. doi:10.3324/haematol.2017.183954 (July 2018).
290. Garcia, M., Juhos, S., ***et al.*** Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. ***F1000Res*** **9,** 63. doi:10.12688/f1000research.16665.2 (Sept. 2020).
291. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. ***Bioinformatics*** **25,** 1754–1760. doi:10.1093/bioinformatics/btp324 (July 2009).
292. McKenna, A., Hanna, M., ***et al.*** The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. en. ***Genome Research*** **20,** 1297–1303. doi:10.1101/gr.107524.110 (Sept. 2010).
293. Kim, S., Scheffler, K., ***et al.*** Strelka2: fast and accurate calling of germline and somatic variants. en. ***Nat Methods*** **15.** Publisher: Nature Publishing Group, 591–594. doi:10.1038/s41592-018-0051-x (Aug. 2018).
294. McLaren, W., Pritchard, B., ***et al.*** Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. ***Bioinformatics*** **26,** 2069–2070. doi:10.1093/bioinformatics/btq330 (Aug. 2010).
295. Morande, P. E., Yan, X. J., ***et al.*** AID overexpression leads to aggressive murine CLL and non-Ig mutations that mirror human neoplasms. ***Blood.*** doi:10.1182/blood.2020008654 (2021).
296. Davi, F. Prehistory of chronic lymphocytic leukemia: clues from the B-cell receptor. en. ***Haematologica*** **109.** Number: 3, 709–711. doi:10.3324/haematol.2023.283799 (2024).
297. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. ***Cell Reports*** **3,** 246–259. doi:10.1016/j.celrep.2012.12.008 (Jan. 2013).
298. Okosun, J., Bödör, C., ***et al.*** Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. ***Nature Genetics*** **46,** 176–181. doi:10.1038/ng.2856 (Dec. 2013).
299. Borcherding, N., Bormann, N. L. & Kraus, G. scRepertoire: An R-based toolkit for single-cell immune receptor analysis. ***F1000Research*** **9,** 47. doi:10.12688/f1000research.22139.2 (June 2020).
300. Hiltemann, S., Jenster, G., Trapman, J., van der Spek, P. & Stubbs, A. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. eng. ***Genome Research*** **25,** 1382–1390. doi:10.1101/gr.183053.114 (Sept. 2015).