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

Sepúlveda Yáñez, J.H.

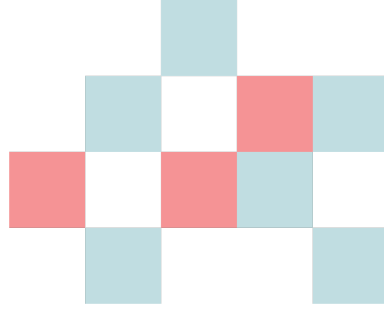
Lago Toro, Chile

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).

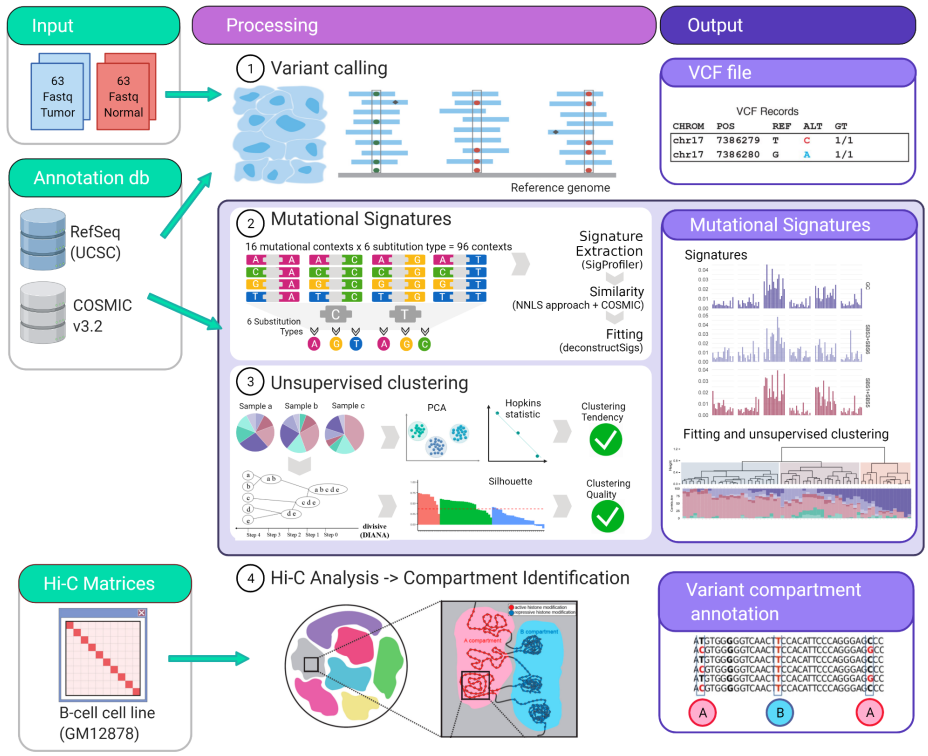


Figure 4.1. Schematic pipeline workflow. The input data are indicated in green boxes. The middle panel show the different analysis performed and the outputs are indicated in purple boxes. Figure created with BioRender.com.

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).

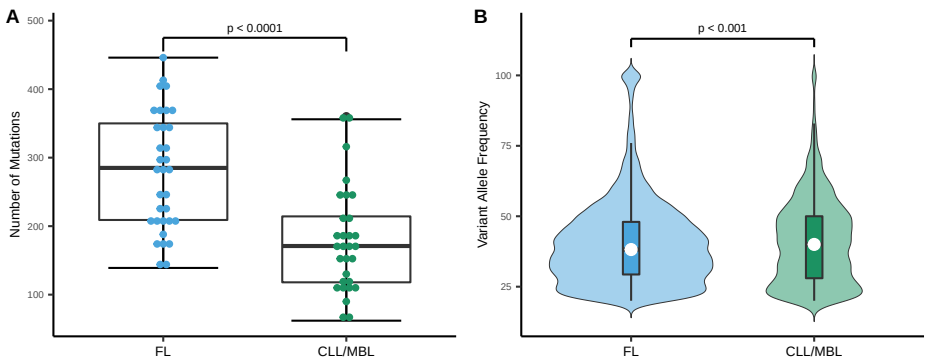


Figure 4.2. Mutational load and distribution of variant allele frequency in FL and CLL/MBL. (**A**) The mutation rate expressed as number of mutations per exome was higher in FL (***n*** = 33) than in CLL/MBL (***n*** = 30) (***t***-test, two-sided). (**B**) Violin plots depict the distinct distribution 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.

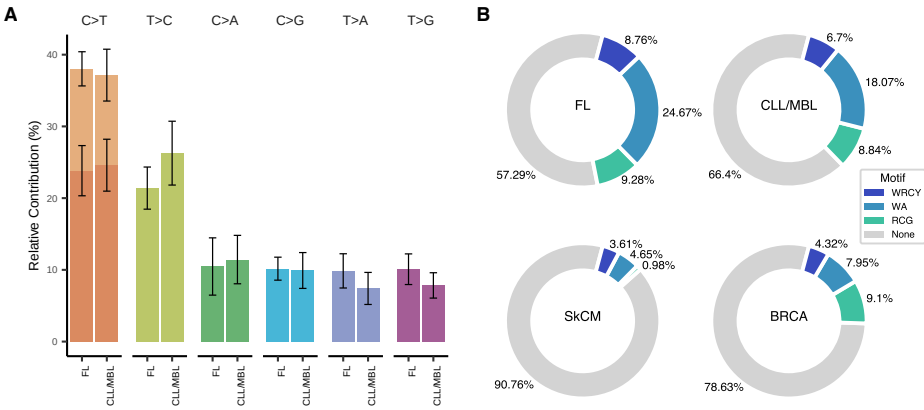


Figure 4.3. Type of substitutions and mutational patterns in FL and CLL/MBL. (**A**) The base substitution types across FL and CLL/MBL cases are dominated by transitions. (**B**) When tracing AID motifs in FL and CLL/MBL a high proportion of somatic mutations are allocated in such motifs, in strong contrast 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).

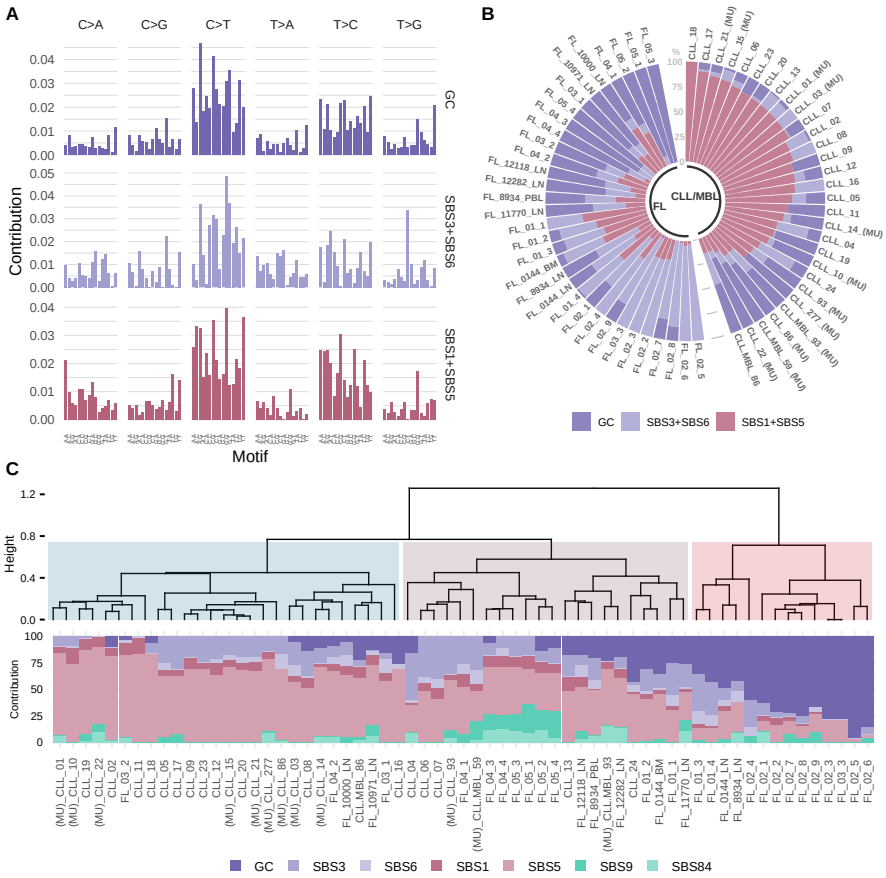


Figure 4.4. Mutational signatures analysis reveals a stable contribution of signature GC in FL. (**A**) The dataset was explained by 3 mutational signatures, a newly emerged de novo signature GC (GC) and two combinations of signatures derived from COSMIC, Single base substitution signature 3 and Single 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

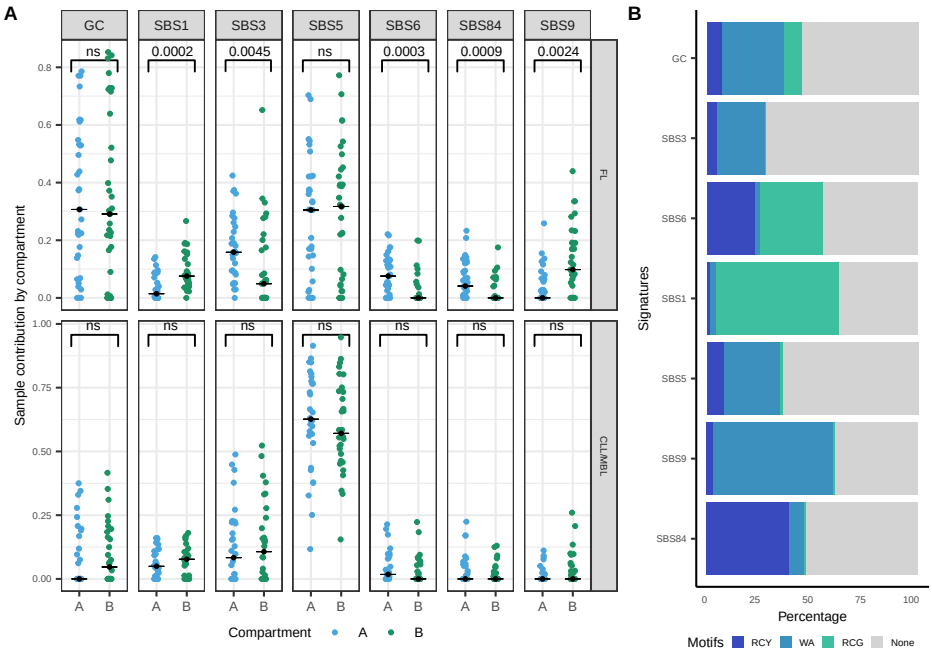


Figure 4.5. Contribution of mutational signatures on the three-dimensional (3D) chromatin structure. (**A**) Sample contribution by compartment A and B in FL and CLL/MBL cases (Wilcoxon test with Bonferroni correction), black dot: median. (**B**) Relative contribution of AID mtifs (trinucleotide) in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

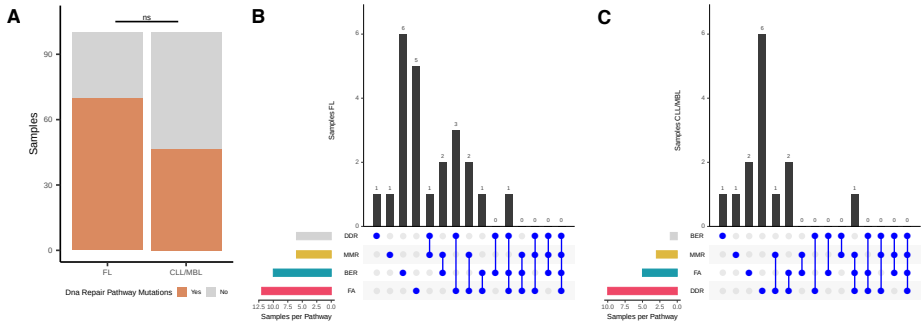


Figure 4.6. Analysis of mutations in DNA repair pathways. (**A**) Percentage of samples with at least one novel mutation (variants not present in germ-line or a preceding biopsy) in the four DNA repair pathways (Base Excision repair [BER], Mismatch Repair [MMR], Fanconi Anemia pathway [FA], and DNA Damage 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).

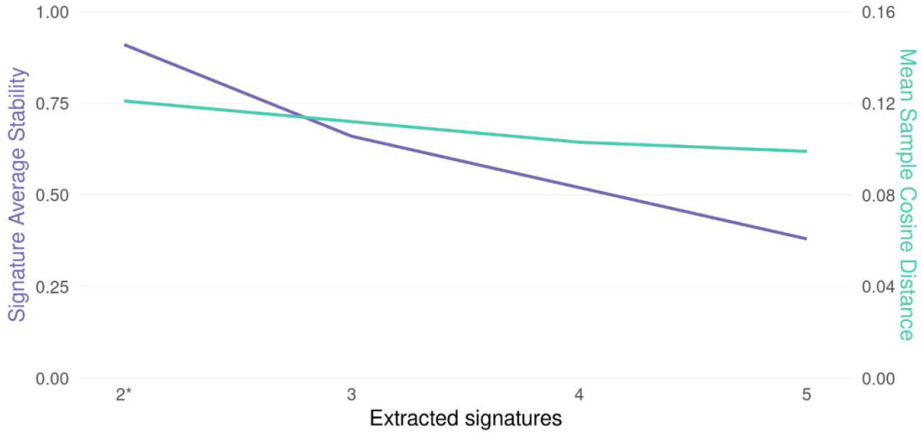
Figure S4. 96-trinucleotide mutational pattern of FL and CLL/MBL.

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

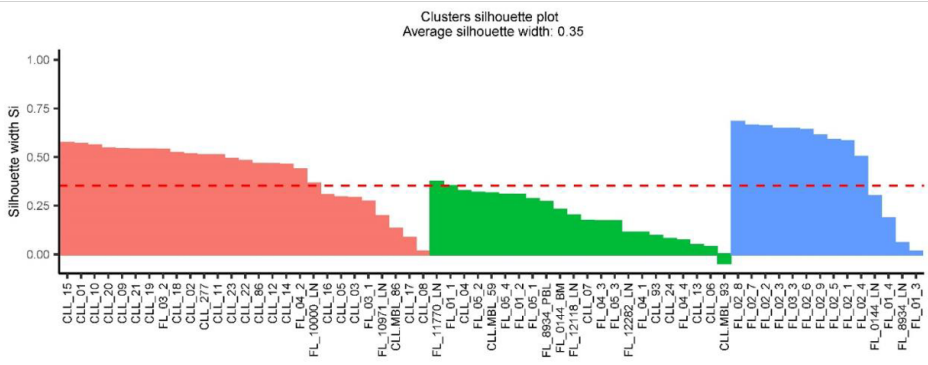
Figure S6. Consistency of signature extraction with sample bootstrapping.

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

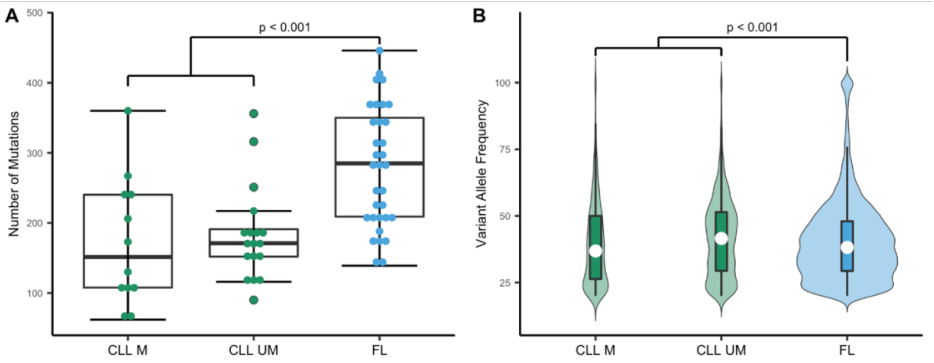
Figure S8. Heatmap of DNA repair genes mutated in FL and CLL/MBL.



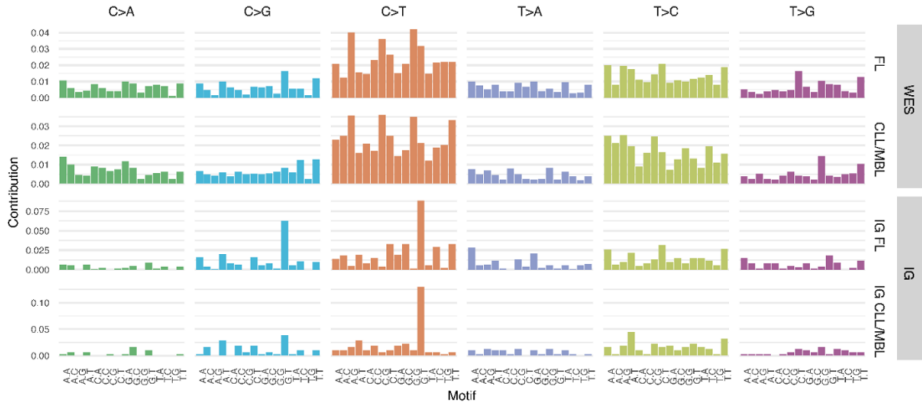
Supplementary Figure 1. Evaluation of the optimal signatures number. We calculated the residual signature average stability (average silhouette coefficient) and the mean sample cosine distance to define to optimal signatures number



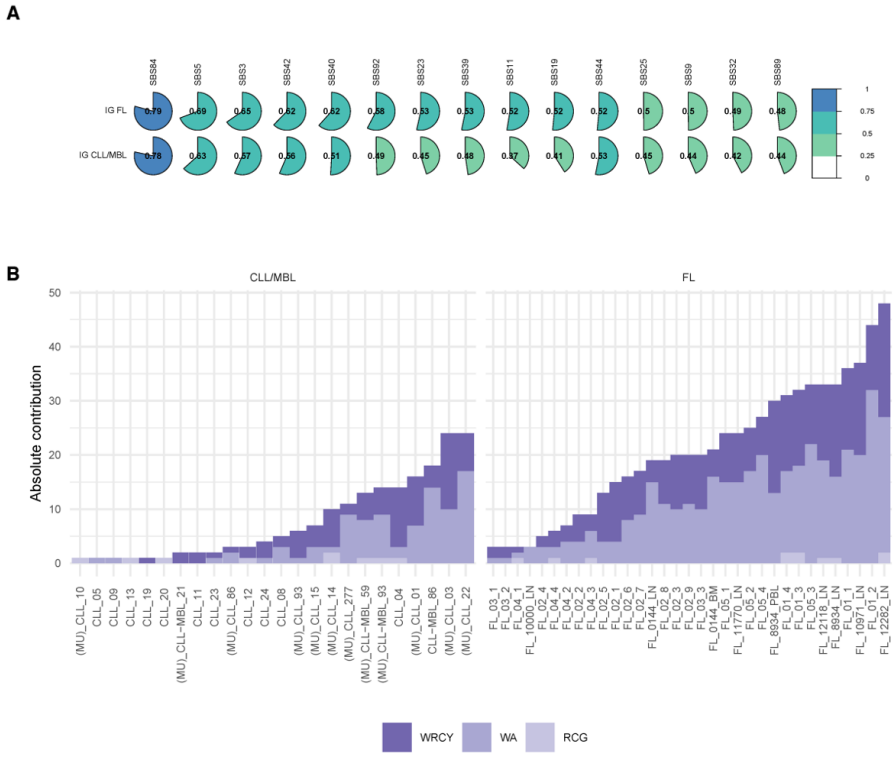
Supplementary Figure 2. Cluster silhouette analysis for divisive hierarchical clustering. Silhouette coefficient will range for each sample from -1 (if it is wrongly placed on the current cluster) to 1 (if it is correctly placed).) and the mean sample cosine distance to define to optimal signatures number



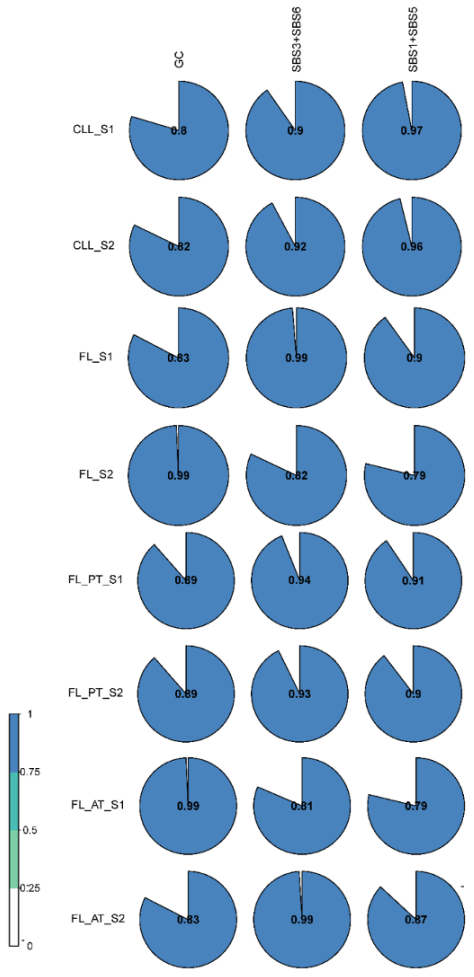
Supplementary Figure 3. Mutational load and distribution of variant allele frequency in FL and CLL mutated (CLL M) and unmutated (CLL U). A The mutation rate expressed as the number of mutations per exome was higher in FL (n=33) than in CLL M/CLL U(n=30) (t-test, two-sided). B Violin plots depict a distinct 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.)



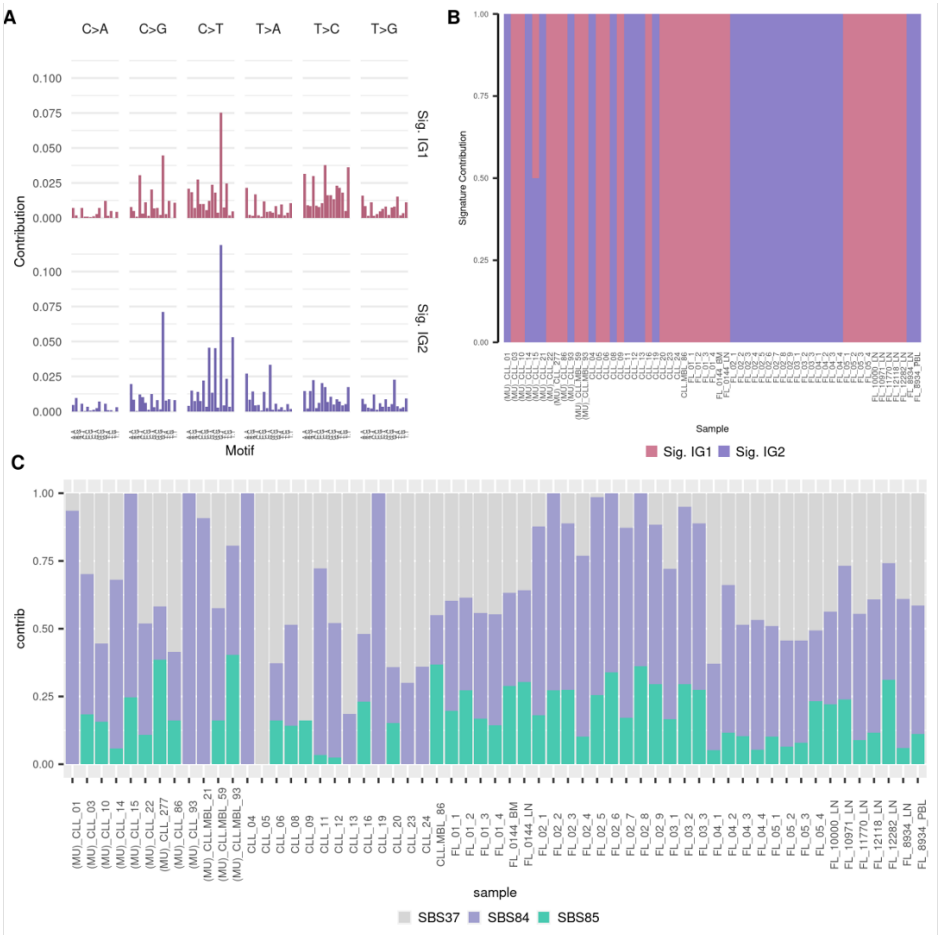
Supplementary Figure 4. 96-trinucleotide mutational pattern of FL and CLL/MBL divided on genome-wide (WES) and in immunoglobulin genes loci (IG).



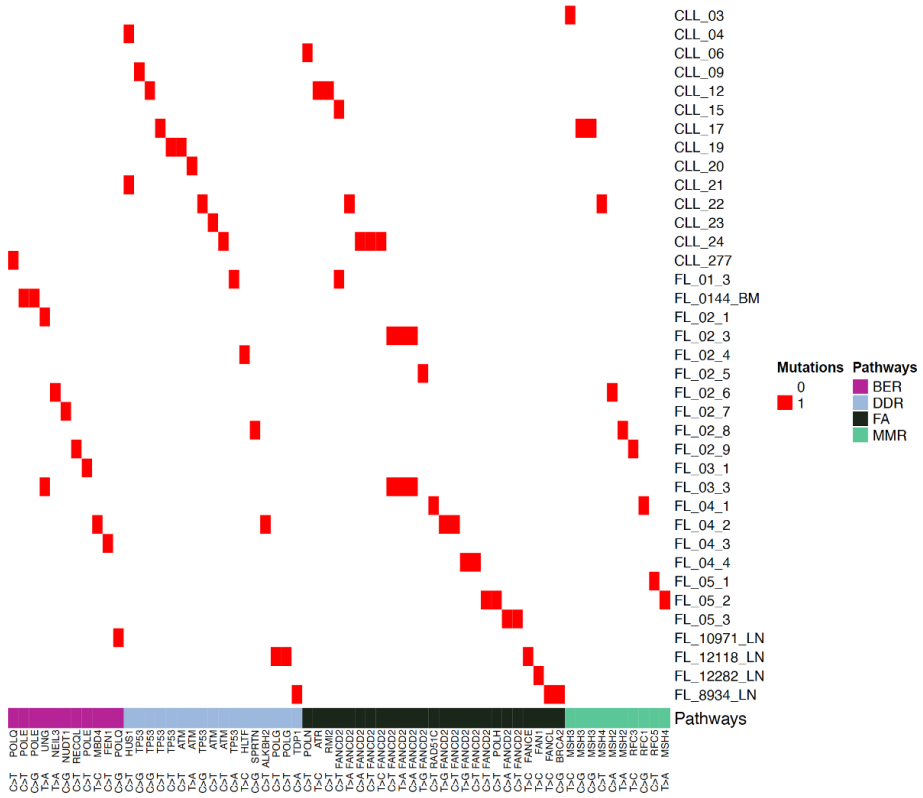
Supplementary Figure 5. Characterization of immunoglobulin loci (IG) in FL and CLL samples. A Cosine similarity measure between the mutational profile in immunoglobulin loci (IG) (showed in Fig. 4) and the Single Base Substitution (SBS) signatures from the COSMIC catalog (v3.2). It is shown the 15 most 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.