

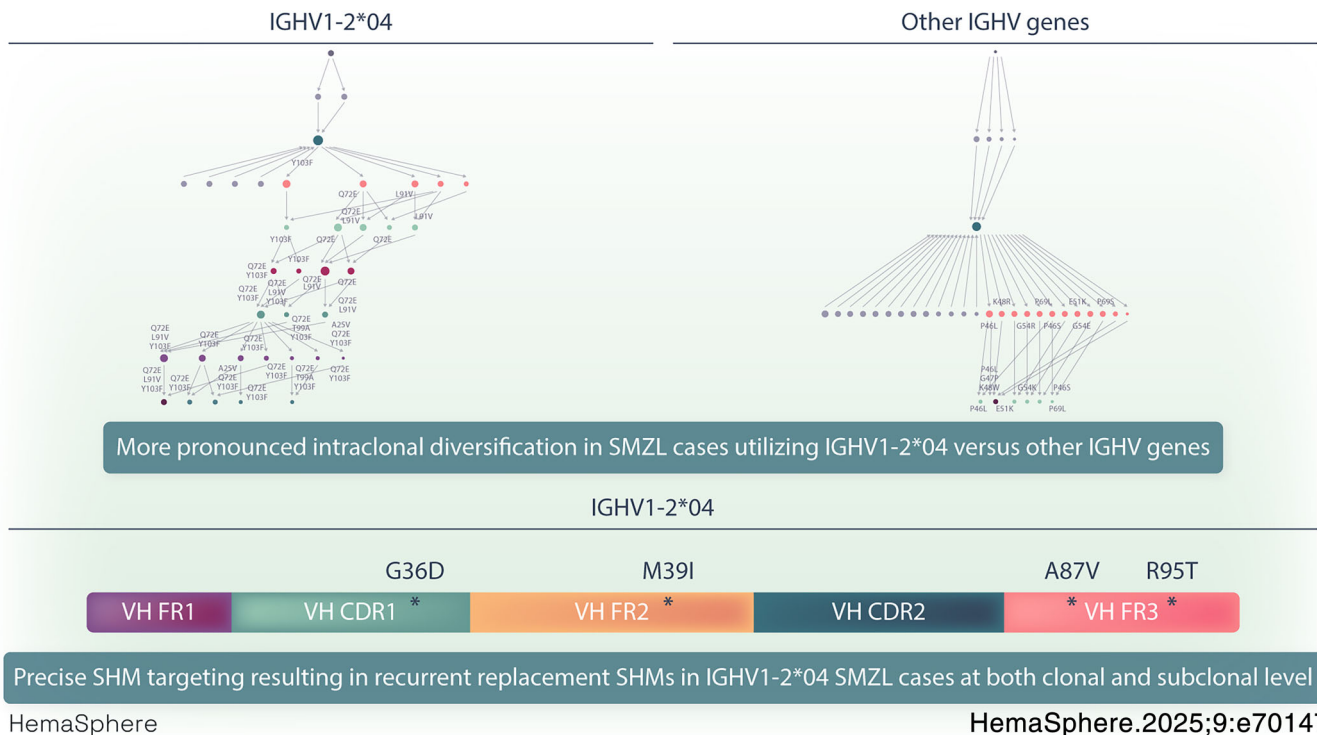




Antigen selection reflected in the subclonal architecture of the B-cell receptor immunoglobulin gene repertoire in splenic marginal zone lymphoma

Laura Zaragoza-Infante^{1,2,^} | Andreas Agathangelidis^{1,3,^}  | Anastasia Iatrou¹ |
 Valentin Junet^{4,5} | Nikos Pechlivanis^{1,6} | Maria Karypidou¹ | Triantafyllia Koletsa⁷ |
 Giorgos Karakatsoulis¹ | Alessio Bruscatto⁸ | Zandie Davis⁹ | Valeria Spina⁸ |
 Aurelie Verney^{10,11} | Eleftheria Polychronidou³ | Fotis Psomopoulos¹ |
 David Oscier⁹ | Alexandra Traverse-Glehen^{10,11} | Maria Papaioannou² |
 Paolo Ghia^{12,13}  | Davide Rossi^{8,14} | Anastasia Chatzidimitriou^{1,15} |
 Kostas Stamatopoulos^{1,15}

Graphical Abstract



Antigen selection reflected in the subclonal architecture of the B-cell receptor immunoglobulin gene repertoire in splenic marginal zone lymphoma

Laura Zaragoza-Infante^{1,2,^} | Andreas Agathangelidis^{1,3,^}  | Anastasia Iatrou¹ |
 Valentin Junet^{4,5} | Nikos Pechlivanis^{1,6} | Maria Karypidou¹ | Triantafyllia Koletsa⁷ |
 Giorgos Karakatsoulis¹ | Alessio Bruscatto⁸ | Zadie Davis⁹ | Valeria Spina⁸ |
 Aurelie Verney^{10,11} | Eleftheria Polychronidou³ | Fotis Psomopoulos¹ |
 David Oscier⁹ | Alexandra Traverse-Glehen^{10,11} | Maria Papaioannou² |
 Paolo Ghia^{12,13}  | Davide Rossi^{8,14} | Anastasia Chatzidimitriou^{1,15} |
 Kostas Stamatopoulos^{1,15}

Correspondence: Andreas Agathangelidis (agathan@biol.uoa.gr) and Kostas Stamatopoulos (kostas.stamatopoulos@certh.gr)

Abstract

Almost one-third of all splenic marginal zone lymphoma (SMZL) cases express B-cell receptor immunoglobulin (BcR IG) encoded by the IGHV1-2*04 gene, implicating antigen selection in disease ontogeny. Evidence supporting this notion mostly derives from low-throughput sequencing approaches, which have limitations in capturing the full complexity of the BcR IG gene repertoire. This hinders the comprehensive assessment of the subclonal architecture of SMZL as shaped by antigen selection. To address this, we conducted a high-throughput immunogenetic investigation of SMZL aimed at the comprehensive characterization of the somatic hypermutation (SHM) and intraclonal diversification within the IG genes. We identified significant differences in the SHM and ID profiles between cases expressing the IGHV1-2*04 gene and those expressing other IGHV genes. Specifically, IGHV1-2*04 cases displayed (i) targeted SHM resulting in recurrent replacement SHMs, and (ii) significantly more pronounced intraclonal diversification, reflecting ongoing antigen selection. Overall, our findings suggest that SMZL cases expressing the IGHV1-2*04 gene have a distinct immunogenetic signature shaped by microenvironmental pressure on the clonotypic BcR IG, corroborating the idea that this group may represent a distinct molecular variant of SMZL.

¹Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece

²First Department of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

³Department of Biology, National and Kapodistrian University of Athens, Athens, Greece

⁴Anaxomics Biotech SL, Barcelona, Spain

⁵Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Barcelona, Spain

⁶Department of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece

⁷Department of Pathology, Faculty of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

⁸Institute of Oncology Research, Laboratory of Experimental Hematology, Bellinzona, Switzerland

⁹Department of Haematology, Royal Bournemouth Hospital, Bournemouth, United Kingdom

¹⁰Centre International de Recherche en Infectiologie, INSERM U1111, Université Claude Bernard Lyon 1, CNRS UMR5308, Ecole Normale Supérieure de Lyon, Université de Lyon, Lyon, France

¹¹Department of Pathology, Hospices Civils de Lyon, Lyon, France

¹²Division of Experimental Oncology, Università Vita-Salute San Raffaele, IRCCS Ospedale San Raffaele, Milan, Italy

¹³IRCCS Ospedale San Raffaele, Milan, Italy

¹⁴Division of Hematology, Oncology Institute of Southern Switzerland, Bellinzona, Switzerland

¹⁵Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden

[^]These authors contributed equally to this work.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *HemaSphere* published by John Wiley & Sons Ltd on behalf of European Hematology Association.

INTRODUCTION

Splenic marginal zone lymphoma (SMZL) is a mature B-cell malignancy displaying considerable biological and clinical heterogeneity.¹ Despite advances in identifying key drivers of the disease, the precise ontogeny of SMZL remains to be fully elucidated. That said, micro-environmental interactions mediated through the B-cell receptor (BcR) have emerged as critical to the natural history of SMZL. This claim is supported by immunogenetic studies demonstrating biases in the BcR immunoglobulin (BcR IG) gene repertoire, alluding to selection by specific antigen(s).^{2–5} Notably, one-third of SMZL cases express clonotypic BcR IG utilizing allele *04 of the IGHV1-2 gene.^{6,7} This finding is particularly striking, given the low frequency of this IGHV gene allele in other mature B-cell lymphomas, such as chronic lymphocytic leukemia (CLL) or mantle cell lymphoma (MCL) (around 5% and 2%, respectively).^{8,9} Furthermore, SMZL cases expressing IGHV1-2*04 BcR IG are enriched for certain genomic aberrations [del (7q) as well as *NOTCH2* and *KLF2* mutations] and follow a more aggressive clinical course compared to cases utilizing other IGHV genes.^{10–15}

The germline IGHV1-2*04 gene allele is unique in encoding tryptophan (W) at position 75 of the heavy variable framework region 3 based on the IMGT unique numbering (VH FR3 75), contrasting all other alleles of IGHV1-2 and nearly all other human IGHV gene alleles that encode arginine (R) at this position (<https://www.imgt.org/IMGTrepertoire/Proteins/index.php#C>). This R residue typically forms a critical intramolecular ionic bond with an aspartic acid (D) residue at position VH FR3 98, which contributes to the stability of the BcR IG heavy chain.^{16–18} While the precise role of W-75 remains unclear, previous *in silico* analyses suggest that it could influence antigen binding.⁶ When rearranged in SMZL, IGHV1-2*04 preferentially pairs with certain IGHD genes, resulting in a particularly long heavy variable complementarity-determining region 3 (VH CDR3) with conserved amino acid (aa) motifs at the tip of the loop across unrelated cases.^{4,6} Moreover, IGHV1-2*04 rearrangements in SMZL show biased pairings with certain IG(K/L)V-J gene rearrangements.^{6,19}

The somatic hypermutation (SHM) imprint on IGHV1-2*04 gene rearrangements in SMZL is relatively modest yet distinctive, characterized by a unique pattern of recurrent replacement SHMs, often conservative, clustering within the heavy variable framework regions (VH FR).⁶ Furthermore, SMZL cases carrying the IGHV1-2*04 gene allele exhibit higher levels of intraclonal diversification compared to cases expressing other IGHV genes, indicating more pronounced and ongoing interactions with antigen(s) throughout the course of the disease.²⁰

Altogether, this evidence highlights the role of the tumor microenvironment in the pathogenesis of IGHV1-2*04 SMZL. Yet, much of the existing evidence comes from low-throughput studies, which are inherently limited in their capacity to capture the full complexity of the BcR IG gene repertoire in the context of antigen selection.

To overcome this limitation, we conducted the first high-throughput immunogenetic investigation of SMZL, focusing on the detailed characterization of SHM and intraclonal diversification. We report that IGHV1-2*04 cases exhibited distinct SHM patterns, linked to AID and Polη activity and recurrent replacement amino acid changes. Additionally, these cases displayed more pronounced intraclonal diversification compared to cases expressing other IGHV genes. These immunogenetic features suggest more intense and continuous antigen-driven selection within the tumor microenvironment, further supporting the notion that ongoing antigen interactions may shape the subclonal dynamics of the BcR IG gene repertoire in IGHV1-2*04 SMZL.

MATERIALS AND METHODS

Study group

The study group comprised 77 patients with SMZL from five collaborating Institutes in France, Greece, Switzerland, and the United Kingdom. The diagnosis was based on the 2016 WHO classification criteria and International guidelines.²¹ The study was conducted in accordance with the Declaration of Helsinki and approved by the local Ethics Committee of each participating institute.

The samples were selected based on their clonotypic BcR IGH gene rearrangement (previously characterized by Sanger sequencing). The selection was intentionally biased toward cases expressing BcR IG encoded by the IGHV1-2*04 gene ($n = 42$). Cases ($n = 35$) expressing other IGHV genes were also evaluated; of these, 8 and 6, respectively, expressed the IGHV3-23 or IGHV4-34 genes, the second and third most frequent genes in the IG gene repertoire of SMZL.⁶ Supporting Information S3: Table 1 lists the analyzed cases and provides information regarding basic demographics, the clonotypic rearranged IGHV gene and the genomic status of each case, particularly focusing on deletion of chromosome 7q as well as *KLF2* and *NOTCH2* mutations.

PCR amplification and sequencing of IGHV-IGHD-IGHJ gene rearrangements

The starting material was formalin-fixed, paraffin-embedded (FFPE) splenectomy specimens in 27/77 cases (35%) of the present cohort. In these cases, xylene-ethanol deparaffinization was followed by genomic DNA (gDNA) extraction with the QIAamp DNA FFPE Tissue Kit (QIAGEN, Germany), following the manufacturer's instructions. In the remaining cases (50/77, 65%) gDNA and total RNA were simultaneously extracted from mononuclear cells obtained from either peripheral blood (PB) or bone marrow (BM) samples using the QIAamp DNA kit (QIAGEN, Germany). More specifically, gDNA was utilized in 18 (23.4%), while RNA was utilized in the remaining 32 cases (41.6%). In the latter cases, cDNA synthesis was performed utilizing the SuperScript II RT reverse transcriptase (Invitrogen, MA, USA).

Amplification of IGHV-IGHD-IGHJ gene rearrangements was performed by polymerase chain reaction (PCR) using the Platinum Taq polymerase (ThermoFisher Scientific, MA, USA). In a total of 58 samples (32 cDNA and 26 gDNA samples, respectively), the IGHV-IGHD-IGHJ gene rearrangements were amplified using 5' IGHV Leader and 3' IGHJ primers, as described previously.^{6,19} In the remaining 19 gDNA samples, all obtained from formalin-fixed, paraffin-embedded (FFPE) tissue samples, the IGHV-IGHD-IGHJ gene rearrangements were amplified using 5' IGHV FR1 and 3' IGHJ primers following the BIOMED-2 protocol.²² PCR products were either purified using the QIAGEN DNA purification kit (QIAGEN, Germany) or gel-purified (QIAGEN, Germany).

Library preparation was performed according to the manufacturer's instructions using the NEB Next Ultra II DNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA). Paired-end NGS was performed using the MiSeq Reagent Kit v3 (2 × 300 bp) on a MiSeq Benchtop Sequencer (Illumina, CA, USA).

To ensure the reproducibility of our experimental approach, PCR amplification, and NGS, we repeated for 18 cases from all major immunogenetic groups of SMZL, namely IGHV1-2*04 ($n = 4$), IGHV3-23 ($n = 3$), IGHV4-34 ($n = 3$), and other IGHV genes ($n = 8$).

IG gene sequence analysis

Raw NGS data were quality-filtered and stitched using an in-house algorithm. Full-length, high-quality sequences were annotated with

the IMGT/HighV-QUEST tool,²³ using the IMGT numbering for the V-DOMAIN. Metadata was further analyzed with the T-cell receptor immunoglobulin profiler (tripr) tool.²⁴ Unproductive rearrangements and rearrangements lacking the VH CDR3 landmarks (C104 and W118, respectively) were discarded from further analysis.

Clonotype definitions were based on the recent guidelines by the EuroClonality NGS Working Group.²⁵ In detail, productive IG gene rearrangement sequences were assigned to clonotypes, defined as unique nucleotide (nt) sequences. The most frequent clonotype of each sample was defined as dominant. Other clonotypes were considered as close to the dominant clonotype, when they fulfilled the following criteria: (i) they utilized the same IGHV gene, (ii) they carried VH CDR3 of identical length; (iii) they encoded for identical or highly similar VH CDR3 nucleotide sequences, allowing for a maximum of two amino acid differences, (iv) they carried one or more distinct SHMs over the VH FR1-to-FR3 part of the VH domain. In the case of the 19 samples that were amplified with VH FR1 primers, codons 1–23 were not taken into consideration.

Analysis of intraclonal diversification within the IG genes

Analysis of intraclonal diversification within the IG genes was performed using the IglDivA software.²⁶ The parameters used for the analysis are listed in Supporting Information S3: Table 2. Only clonotypes with at least 10 reads were considered for this analysis.

Analysis with IglDivA provided the following types of information: (i) identification of close clonotypes of the dominant clonotype in each sample, as well as their connections, termed mutational pathways; (ii) visualization of the connections between the dominant clonotype and close clonotypes through graph networks; (iii) quantification of intraclonal diversification levels through the calculation of graph network-based metrics; and (iv) statistical comparisons between particular groups of samples. More information regarding the graph network-based metrics is provided in Supporting Information S3: Table 3.

BcR IG 3D model generation and clustering

BcR IG 3D model generation was performed using the RepertoireBuilder software.²⁷ For the generation of the IG heavy chains of each model, the sequences were either “SHM-free” (i.e., devoid of any SHM) or carried the most recurrent SHMs separately or in different combinations. Furthermore, a consensus VH CDR3 amino acid sequence was computed using the VH CDR3 of all SMZL IGHV1-2*04 cases of the present cohort and was incorporated into all the BcR IG heavy chain sequences. Given the fact that NGS of the clonotypic light chain IG sequences was beyond the scope of the present study, the two most representative IG light chain gene sequences from our cohort (as determined by Sanger sequencing) were paired with each heavy chain.¹⁹ This choice was based on (i) the high frequency of the respective genes (IGKV1-8 and IGKV1-39/1D-39) in both the present cohort as well as in SMZL in general²⁸; and (ii) the lack of SHMs, to minimize the impact of the light chain on the overall 3D BcR IG structure. Immunogenetic information regarding the utilized light chain IG sequences is provided in Supporting Information S3: Table 4.

To assess the overall structural similarity,^{29,30} the 3D BcR IG models were compared against each other and the root mean square deviation (RMSD) metric was calculated for each pairwise comparison utilizing the PDBfold software.³¹ Subsequently, RMSD values corresponding to the structural differences between each pair of 3D BcR

IG models were used for the generation of a distance matrix. Finally, hierarchical clustering was performed for the identification of groups of structurally similar 3D BcR IG. To analyze the output, a threshold was applied at 0.4 to obtain a number of fairly populated clusters that could facilitate the extraction of meaningful conclusions.

Statistical analysis

For the group comparisons, the Kruskal–Wallis and the Wilcoxon tests were applied (if appropriate). The significance level was set to 5%, all tests were two-sided, and, in case of post hoc analyses, a *p* value correction was also applied.

RESULTS

Higher levels of clonality and VH CDR3 sequence convergence in SMZL cases utilizing the IGHV1-2*04 versus other IGHV genes

A total of 14,034,541 IGHV-IGHD-IGHJ gene rearrangement sequences were obtained, with an average of 188,267 sequences per sample (Supporting Information S3: Table 5). After pre-processing and filtering, 12,144,902 (86.5%) sequences were classified as “productive” by IMGT/HighV-QUEST and subjected to downstream analysis. The average number of productive sequences was significantly higher in cases utilizing the IGHV1-2*04 (180,900) versus cases utilizing other IGHV genes (129,917) ($p = 0.04$).

All 77 analyzed samples exhibited a clear monoclonal profile. Specifically, the average cumulative frequency of the dominant clonotype and its closely related clonotypes was 93.8% (median: 97.8%; Supporting Information S3: Table 6). The remaining clonotypes were classified as follows: (i) distant clonotypes, with more than two amino acid differences within the VH CDR3 (average cumulative frequency: 1%; median: 0.4%); (ii) unrelated clonotypes, which carried different IGHV genes and VH CDR3 sequences (average cumulative frequency: 5.2%; median: 1.8%) (Supporting Information S3: Table 6).

The median cumulative frequency of the dominant clonotype and its closely related clonotypes was significantly higher ($p < 0.01$) in the IGHV1-2*04 group (98.3%, range: 83.4–99.9) compared to cases carrying other IGHV genes (93.5%, range: 61.9–99.8). A similar trend was observed when comparing the IGHV1-2*04 group to groups of cases utilizing the second and third most frequent IGHV genes in SMZL, namely IGHV3-23 (96.4%, range: 62–99.8) and IGHV4-34 (95.3%, range: 62.4–98.8), yet the differences did not reach statistical significance ($p = 0.07$ for both comparisons). Vice versa, unrelated clonotypes were significantly more frequent in the non-IGHV1-2*04 group (median: 4.2%, range: 0.1–37.5), as well as in the IGHV3-23 (3.4%, range: 0.2–37.5) and the IGHV4-34 groups (4.36%, range: 1–35) compared to the IGHV1-2*04 group (median: 1%, range: 0.1–16.2) (Supporting Information S3: Table 7) ($p < 0.05$ in all comparisons).

VH CDR3 convergence was assessed through determining the number of different clonotypes encoding the same VH CDR3 amino acid sequence in each sample. The IGHV1-2*04 group showed more pronounced VH CDR3 convergence compared to non-IGHV1-2*04 group, with a significantly higher ($p < 0.01$) median number of clonotypes encoding for the same VH CDR3 amino acid sequence (111 vs. 61; Supporting Information S3: Table 5 and Supporting Information S1: Figure 1). This was also evident in the comparisons between the IGHV1-2*04 group against either the IGHV3-23 group (111 vs. 58.5; $p < 0.01$) or the IGHV4-34 group (111 vs. 83, $p < 0.05$).

TABLE 1 Graph metrics values for comparisons between the IGHV1-2*04 SMZL group, the non-IGHV1-2*04 group at large, as well as the IGHV3-23 and IGHV4-34 groups individually. For each metric, the median value together with the range of values are shown for each group of samples.

Metric	IGHV1-2*04	Non-IGHV1-2*04	p value	IGHV3-23	p value	IGHV4-34	p value
RRC	0.10 (0.01–8.3)	0.01 (0.002–1.13)	<0.01	0.012 (0.002–0.327)	<0.01	0.028 (0.017–0.062)	NS
END	0.33 (0.1–0.7)	0.42 (0.2–0.5)	<0.05	0.48 (0.2–0.5)	NS	0.44 (0.2–0.5)	NS
MPL	3 (2–6)	2 (2–6)	<0.05	2 (2–4)	NS	2 (2–5)	NS
MML	3 (2–12)	2 (2–29)	NS	2 (2–29)	NS	5 (2–6)	NS
ADe	2.59 (1.7–3.5)	2.7 (1.2–3.2)	NS	2.8 (1.71–3.2)	NS	2.7 (2.4–2.9)	NS
ADi	2 (1–3)	2 (1–3)	NS	2 (1–2)	NS	2 (1–2)	NS

Abbreviations: ADe, average degree; ADi, average distance; END, end nodes density; MML, maximal mutational length; MPL, maximal pathway length; NS, non significant; RRC, relative reads convergence.

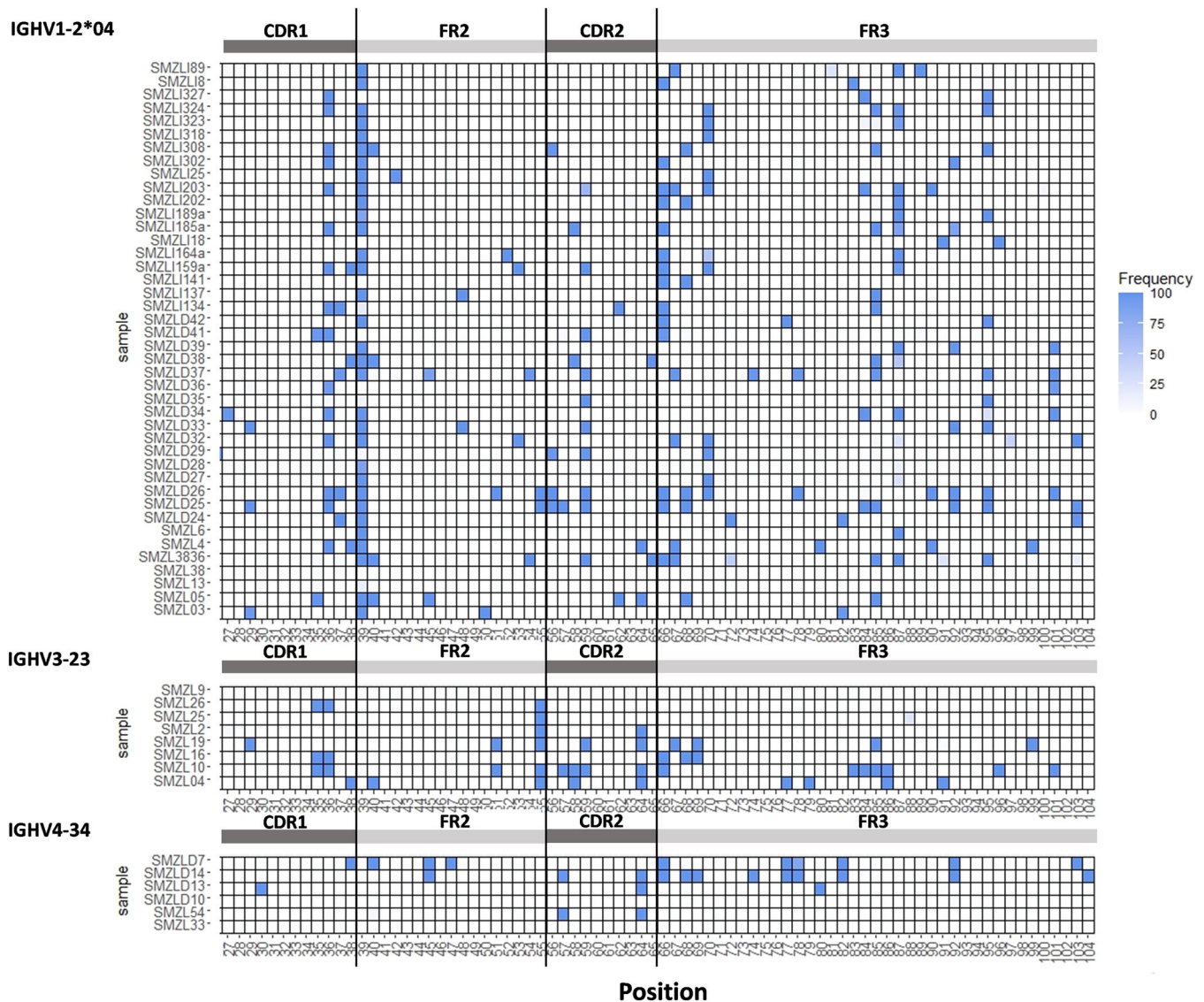


FIGURE 1 Differential SHM targeting in IGHV1-2*04 cases versus cases utilizing other IGHV genes. The topology of replacement SHMs reveals the existence of different patterns in IGHV1-2*04 cases compared to cases expressing the IGHV3-23 or IGHV4-34 genes. Rows in each table represent the codons of the clonotypic BcR IG heavy chain sequence from the start of VH FR1 until the end of VH FR3. The sequence of the first 26 codons of VH FR1 was not available in 19 IGHV1-2*04 rearrangements that were PCR-amplified utilizing VH FR1 rather than leader primers; this part of the graph is depicted with a light gray color. Columns in each table represent individual SMZL cases in each group.

To address the potential confounding effect of having significantly more BcR IG sequences in the IGHV1-2*04 group versus the group of cases utilizing other IGHV genes, we divided the former into two subgroups; this division was based on a threshold of 130,000 BcR IG sequences, which corresponds to the median of the non IGHV1-2*04 group. We did not find any significant differences in the median cumulative frequencies of either the dominant clonotype and its closely related clonotypes, nor in the frequencies of the distant and unrelated clonotypes. Differences were also insignificant for VH CDR3 convergence or any of the intraclonal diversification-related metrics (data not shown). These findings indicate that the higher number of BcR IG sequences in the IGHV1-2*04 group did not affect their subclonal immunogenetic architecture.

More pronounced intraclonal diversification in SMZL cases utilizing IGHV1-2*04 versus other IGHV genes

The IGHV1-2*04 SMZL group showed a higher level of interconnection between clonotypes, with mutational pathways converging toward few end clonotypes bearing specific combinations of SHMs. In contrast, cases expressing other IGHV genes displayed more random SHM patterns, resulting in more end clonotypes with distinct SHM profiles (Supporting Information S1: Figure 2).

To quantitatively assess these differences, we calculated six relevant graph metrics (Supporting Information S3: Table 8A), which enabled the in-depth characterization of the complexity of intraclonal diversification in each individual sample (Supporting Information S3: Table 8B). Eight samples (10.4%) exhibited very low SHM complexity, showing no connections between clonotypes, and were therefore excluded from downstream comparisons. Overall, we found significantly higher intraclonal diversification levels in the IGHV1-2*04 group versus the other SMZL cases, as shown by several graph metrics (Supporting Information S1: Figure 3 and Table 1). Intraclonal diversification levels were also higher in the IGHV1-2*04 group when compared to the IGHV3-23 or IGHV4-34 groups (Table 1). Intraclonal diversification was not evident in the VH CDR3 sequence in any of the analyzed groups, as there were no differences in its amino acid composition between the dominant clonotype and its closely related clonotypes in 67/69 samples (97.1%).

We also considered the possibility that differences in the genomic background, particularly mutations in the *KLF2* and/or *NOTCH2* genes, could affect intraclonal diversification. Virtually all cases carrying gene mutations were utilizing the IGHV1-2*04 gene; hence, the observed differences between mutated versus wildtype cases were confounded by the immunogenetic profiles. That said, cases expressing the IGHV1-2*04 and carrying gene mutations exhibited higher intraclonal diversification complexity against the wildtype group expressing other IGHV genes.

To ensure that our findings truly reflect disease biology and are not influenced by the experimental setup, we undertook the following actions: (i) we explored whether differences in the starting material could affect the reported findings by comparing ID metrics between IGHV1-2*04 cases analyzed on gDNA ($n = 6$) or RNA ($n = 32$), finding no significant differences (Supporting Information S1: Figure 4); and (ii) we repeated the PCR amplification and NGS steps for 18 cases from all major immunogenetic groups of SMZL (IGHV1-2*04 [$n = 4$], IGHV3-23 [$n = 3$], IGHV4-34 [$n = 3$], and other IGHV genes [$n = 8$]) and used the generated NGS data to replace the respective data from the original sample replicates. Remarkably, the results regarding intraclonal diversification profiles were highly concordant (Supporting Information S1: Figures 5 and 6).

AID and polymerase η hotspots were more frequently targeted by SHM in SMZL cases utilizing the IGHV1-2*04 versus other IGHV genes

We analyzed the topology of SHM by focusing on the targeting of established AID hotspots, including WRC/GYW ($W = A/T$, $R = A/G$, $Y = C/T$), WGCW, DGYW/WRCH ($D = A/G/T$, $H = T/C/A$), and the Pol η hotspot WA/TW.

In total, 322 clonal SHMs (i.e., present in the dominant clonotype) were identified in the IGHV1-2*04 group. Of these, the majority (226/322 clonal SHMs, 70%) were located either within AID hotspots (190 SHMs, 59%) or within Pol η hotspots 36 (11%). A nearly identical distribution was evident for subclonal SHMs (i.e., present in related clonotypes), as 280/398 subclonal SHMs (70%) were located either within AID hotspots ($n = 237$) or within Pol η hotspots ($n = 43$). In contrast, cases expressing other IGHV genes exhibited significantly lower targeting ($p < 0.01$) of AID and Pol η hotspots. In detail, only 48% of clonal SHMs (130/271) and 47% of subclonal SHMs (164/345) were located within AID hotspots ($n = 90$ for clonal and $n = 108$ for subclonal SHMs) or Pol η hotspots ($n = 40$ for clonal and $n = 56$ for subclonal SHMs, 15%). The proportion of SHMs (both clonal and subclonal) located in hotspots was significantly higher in the IGHV1-2*04 group compared to the IGHV3-23 and IGHV4-34 groups ($p < 0.01$ for all comparisons). Specifically, of a total of 70 clonal SHMs in the IGHV3-23 group, 39 (55.7%) were located in hotspots; in addition, 51/98 subclonal SHMs (52%) were located in hotspots. In the IGHV4-34 group, 47 clonal SHMs were identified, of which 21 (44.7%) were located in hotspots, while 28/57 subclonal SHMs (49.1%) were located in hotspots.

Distinct topology, higher frequency, and stronger selection of replacement SHMs in SMZL cases utilizing the IGHV1-2*04 versus other IGHV genes

Next, we investigated in detail replacement SHMs, both clonal and subclonal, focusing on their distribution in groups of cases expressing the IGHV1-2*04, IGHV3-23, and IGHV4-34 genes.

TABLE 2 Most frequently targeted codons in the heavy variable (VH) domain in SMZL cases expressing the IGHV1-2*04 or other IGHV genes. Only codons that were targeted by SHM (i) in $\geq 25\%$ of the cases in each particular group; and (ii) in $\geq 5\%$ of the mutated nucleotide variants in each case of every SMZL group. The numbering of VH codons follows the IMGT unique numbering.

Codon	IGHV1-2*04 ($n = 42$)		Non IGHV1-2*04 ($n = 35$)		p value
	Cases (no)	Cases (%)	Cases (no)	Cases (%)	
VH CDR1-36	16	38.1	10	28.6	NS
VH FR2-39	36	85.7	-	-	<0.01
VH CDR2-59	10	23.8	-	-	NS
VH CDR2-64	-	-	13	37.1	<0.01
VH FR3-66	14	33.3	-	-	<0.05
VH FR3-70	12	28.6	-	-	<0.05
VH FR3-85	12	28.6	-	-	<0.05
VH FR3-87	19	45.2	-	-	<0.01
VH FR3-95	13	30.9	-	-	<0.05

Abbreviation: NS, non significant.

The IGHV1-2*04 group exhibited frequent targeting by replacement SHMs across eight codons spanning the region between VH FR1 and VH FR3 (Figure 1 and Table 2). In contrast, only two codons were frequently targeted in the non-IGHV1-2*04 group. Interestingly, IGHV1-2*04 cases were characterized by frequent targeting of the VH FR3; in more detail, 5/8 (62.5%) of the frequently targeted codons were located in this region, contrasting the IGHV3-23 group (Supporting Information S3: Table 9).

Besides selectively targeting certain codons, SHM in the IGHV1-2*04 group led to the frequent introduction of recurrent replacement SHMs. Among samples subjected to intraclonal diversification analysis, the M-to-I replacement (M39I) was by far the most frequent, appearing in 31/38 cases (81.6%), in 26 in the dominant clonotype (clonal SHMs) and in the remaining 5 in the end clonotype (subclonal SHMs) (Figure 2). Additional examples are provided in Supporting

Information S3: Table 10. In addition, the VH FR2-39 M39I replacement clearly predominated over other replacements at the same position. This contrasted with the IGHV3-23 and IGHV4-34 groups, where frequently targeted codons exhibited several different amino acid replacements at similar frequencies (Supporting Information S3: Table 11).

Finally, we also addressed whether the recurrent replacement SHMs observed in the IGHV1-2*04 group were SMZL-biased, by comparing them to large datasets from (i) CLL (data extracted from the IMGT/CLL-DB, <https://www.imgt.org/CLLDBInterface/query>), and (ii) healthy donors (NCBI, BioProject PRJNA527941). According to the results, M39I and other recurrent SHMs were significantly overrepresented in IGHV1-2*04 rearrangements from SMZL compared to both CLL⁸ and the circulating normal B-cell compartment³² ($p < 0.05$ in both comparisons) (Supporting Information S3: Table 12).

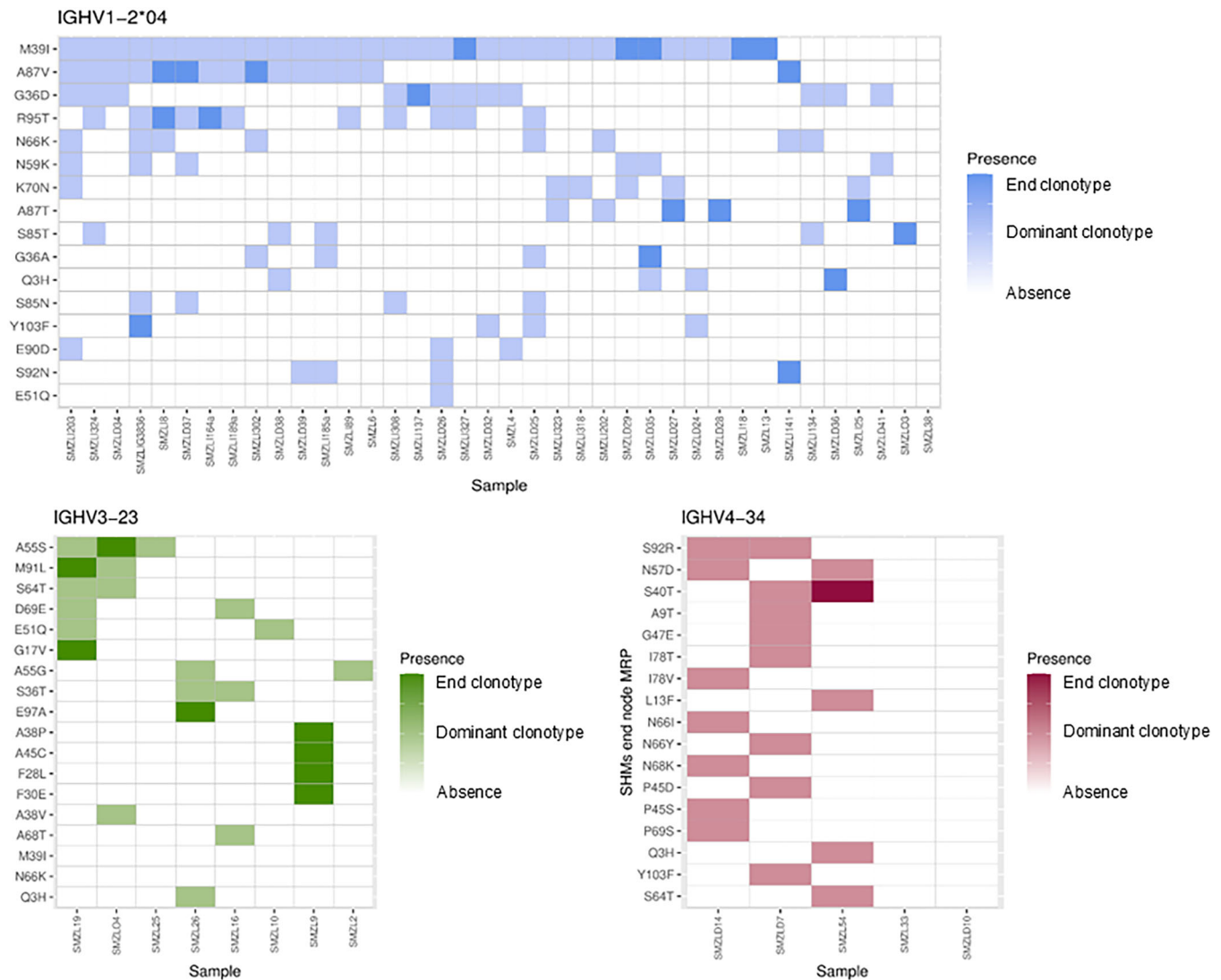


FIGURE 2 Differential distribution of recurrent replacement SHMs in IGHV1-2*04 cases versus cases utilizing other frequent IGHV genes. The IGHV1-2*04 group shows a higher frequency of recurrent replacement SHMs occurring both in the dominant clonotype as well as in the end clonotype in the context of intraclonal diversification. Some of these SHMs, namely M39I, A87V, G36D, and R95T, were found to occur simultaneously in a large number of IGHV1-2*04 cases. Co-occurrence of recurrent replacement SHMs was significantly more limited in the other sample groups, being evident only in two samples of the IGHV3-23 group. In each individual heatmap, rows represent different recurrent replacement SHMs, while columns represent individual SMZL cases of each group. A lighter shade of color indicates that the SHM was present in the dominant clonotype, while a stronger shade of color indicates that the SHM was acquired in the end clonotype (i.e., the one with the highest number of SHMs acquired in the context of intraclonal diversification of the IG genes).

Three-dimensional IG modeling in IGHV1-2*04 SMZL cases highlights replacement SHMs with a potential significant structural impact

We conducted an *in silico* analysis to investigate the potential structural impact of particular recurrent replacement SHMs in IGHV1-2*04 SMZL cases. In total, 32 distinct models of pairs of IGHV1-2*04 heavy and light chain sequences were generated and subjected to structural comparisons. Hierarchical clustering led to the identification of six individual clusters (Figure 3). The composition of these clusters was analyzed in relation to SHM patterns, with a particular focus on the most frequent replacement SHMs: M39I, G36D, A87V, and R95T. Cluster 1 included the 3D models of IGHV1-2*04 heavy chains devoid of SHMs or carrying G36D or M39I in isolation. The presence of A87V or R95T resulted in the assignment of the respective 3D BcR IG models to distinct clusters, suggesting that these SHMs may exert a more pronounced structural impact compared to G36D and M39I. Notably, no case in our cohort harbored A87V or R95T in isolation. When looking at combinations of SHMs, 3D BcR IG models carrying all possible combinations of the G36D, M39I, and R95T were grouped within the C2 and C3 clusters, indicating a high level of structural homology with the “SHM-free” 3D models. In contrast, all models carrying A87V in any possible combination were clearly separated from the “SHM-free” BcR IG 3D models, further supporting the distinct structural impact of this particular SHM.

DISCUSSION

Almost one-third of SMZL cases express BcR IG encoded by the IGHV1-2*04 gene allele.^{6,33} This remarkable restriction likely reflects functional selection and prompts investigation into the interactions between the

malignant cells and their microenvironment.³⁴ To address this issue from an immunogenetic perspective, we performed a high-throughput analysis of the subclonal immunogenetic architecture in a large SMZL cohort, intentionally enriched in cases expressing the IGHV1-2*04 gene.

IGHV1-2*04 cases displayed a significantly more clonal yet also more intraclonally diversified BcR IG gene repertoire, with pronounced sequence convergence within the VH CDR3 and recurrent replacement SHMs in particular positions within specific VH domain positions. This profile indicates a stronger and more prominent microenvironmental pressure acting on the BcR of IGHV1-2*04 SMZL cases compared to cases utilizing other IGHV genes.

Clonality assessment confirmed the monoclonal profile of SMZL in all analyzed cases, yet IGHV1-2*04 cases displayed higher levels of clonality, perhaps alluding to stronger and/or more prolonged microenvironmental triggering. In line with this, IGHV1-2*04 cases were characterized by a significantly higher degree of VH CDR3 convergence compared to non IGHV1-2*04 cases, indicative of selection pressure favoring the preservation of a specific amino acid composition.

Analysis of intraclonal diversification within the IGHV genes revealed that SMZL cases utilizing the IGHV1-2*04 gene exhibited more complex connections between clonotypes, shaped by a more prominent antigenic selection process. Using NGS combined with a purpose-built software based on graph network metrics,²⁶ we provide for the first time not only an in-depth qualitative but also quantitative assessment of intraclonal diversification dynamics. The herein reported intraclonal diversification metrics in the IGHV1-2*04 SMZL variant support a model of continuous accumulation of SHMs taking place in a progressive and specified manner, leading to the emergence of clonotypes with distinct SHM profiles. This ongoing SHM process suggests sustained microenvironmental triggering in IGHV1-2*04 cases, uniquely shaping their immunogenetic architecture.

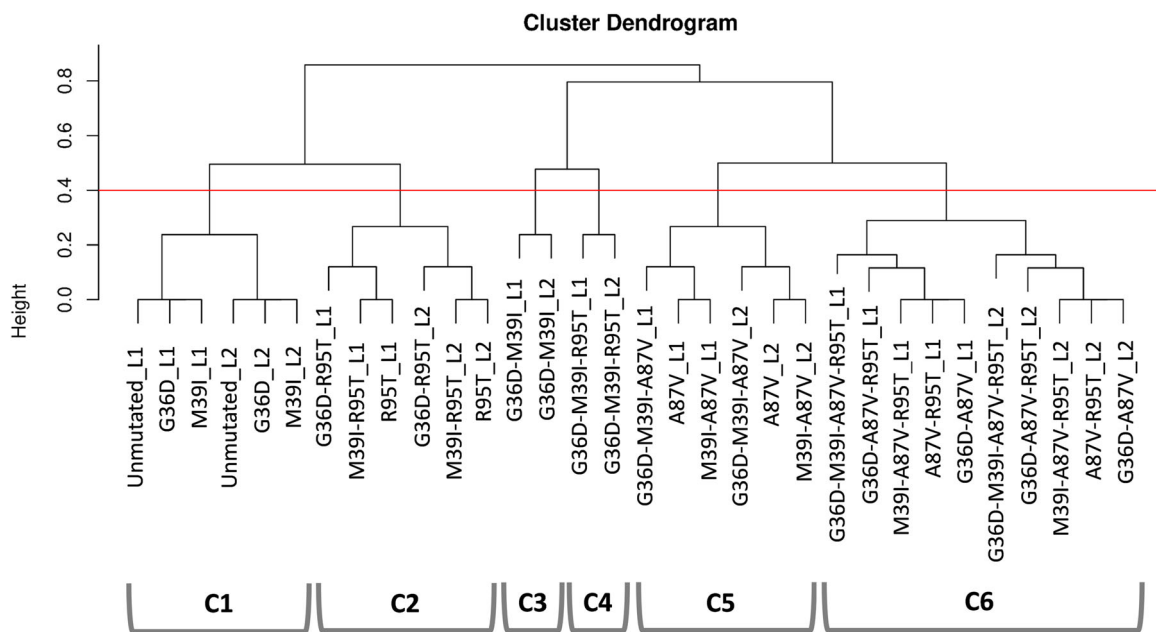


FIGURE 3 *In silico* BcR IG modeling and clustering in IGHV1-2*04 cases support a differential impact for recurrent replacement SHMs. The 3D BcR IG models of the SMZL cases carrying ID-related recurrent SHMs were grouped in six different clusters. “L1” and “L2” indicate the two different light chains that were paired with each heavy chain to eliminate the possible impact of the light chain. Each separate pair of 3D BcR IG models with a given heavy chain and a different light chain was assigned to the same cluster, indicating that the overall structure of the 3D BcR IG models is largely unaffected by the light chain partner of the IGHV1-2*04 heavy chain. At the cluster level, C1 comprised the “SHM-free” IGHV1-2*04 models (i.e., those devoid of any SHM) as well as those carrying the G36D and M39I mutations individually. Cluster C2 models were characterized by the presence of the R95T mutation, either as a single mutation or in combination with one of the G36D or the M39I mutations. Clusters C3 and C4 consisted of the models carrying the G36D-M39I and G36D-M39I-R95T SHM combinations, respectively. Clusters 5 and 6 were formed by all 3D BcR IG models carrying the A87V, again either individually or in combination with the other analyzed mutations.

IGHV1-2*04 cases were also characterized by a significantly higher frequency of recurrent replacement SHMs compared to other SMZL cases, reflected in a unique SHM topology. The most striking example concerned VH FR2 M39I, which appeared to be disease-biased, as its frequency was significantly lower in IGHV1-2*04 expressing clones in CLL⁸ and the normal B-cell repertoire in the blood circulation.³² Furthermore, AID and polymerase η hotspots were more preferentially targeted by SHM in IGHV1-2*04 cases at both the clonal and subclonal levels. The selective targeting of these hotspots may reflect evolutionary pressures favoring interactions between AID and/or Pol η induced mutations,³⁵ further arguing against a random SHM process. This suggests that the unique nature of the IGHV1-2*04 gene allele, combined with a distinct set of antigenic stimuli, contributes to the subclonal architecture of IGHV1-2*04 SMZL.

Application of an *in silico* BcR IG modeling and clustering approach, developed by our group,³⁶ in IGHV1-2*04 SMZL cases underlined the dominance of the IG heavy chain, at least in this disease variant.^{37,38} This was evident from the structural similarity of BcR IG models constructed using IGHV1-2*04 heavy chains paired with light chains utilizing different variable genes, namely IGKV1-8 and IGKV1-39/1D-39, two of the most frequently utilized IG light variable genes in SMZL.²⁸ Furthermore, structural modeling suggested a distinct impact of recurrent replacement SHMs, since BcR IG models carrying G36D and M39I clustered closely with SHM-free models, while those harboring A87V or R95T exhibited significant structural deviations. Given the high prevalence of M39I and G36D in our cohort compared to either other mature B-cell malignancies, such as CLL,⁸ or circulating normal B cells,³² our findings may indicate refinement of BcR IG structure within the context of affinity maturation; of note, this does not appear to affect the unique structural features of IGHV1-2*04 gene allele, including the distinctive W residue at position VH FR3 75.⁶

A limitation of the present work concerns the utilization of PCR-based NGS methodologies on bulk cell populations rather than single-cell approaches. While the latter would offer deeper insights into SMZL immunogenetic dynamics, lack of viable cells precluded such studies. That notwithstanding—our primary objective to comprehensively characterize intraclonal diversification at a bulk population level, still the most scalable approach for large-cohort studies—allowed reaching strong evidence for the distinctive immunogenetic architecture of the IGHV1-2*04 group in SMZL.

In conclusion, our findings support the notion that IGHV1-2*04-expressing SMZL likely represents a distinct disease variant shaped by ongoing antigenic interactions and genetic drivers. This has significant implications for the development of novel therapeutic strategies targeting the IGHV1-2*04 BcR. Conceivably, this can be achieved through, e.g., the development of chimeric antigen receptor (CAR) T cells targeting the IGHV1-2*04, similar to previous attempts by us³⁹ and others⁴⁰ with IGHV4-34 and IGLV3-21^{R110} CARs, respectively, in CLL and other B-cell malignancies. Alternatively, pharmacological compounds targeting the unique structural features of IGHV1-2*04 BcRs (i.e., the IGHV1-2*04-specific VH FR3 W75) could be identified through *in silico* docking-based screening of libraries of compounds already approved for human use.

AUTHOR CONTRIBUTIONS

Laura Zaragoza-Infante, Andreas Agathangelidis, and Anastasia Iatrou designed the research, performed the experiments, analyzed the data, and wrote the manuscript. Maria Karypidou, Triantafyllia Koletsa, Alessio Brusca, Zadie Davis, Valeria Spina, and Aurelie Verney assisted in experiments. Valentin Junet, Nikos Pechlivanis, and Eleftheria Polychronidou assisted in bioinformatics analysis. Giorgos

Karakatsoulis performed statistical analysis. Fotis Psomopoulos, David Oscier, Alexandra Traverse-Glehen, and Maria Papaioannou supervised the research. Paolo Ghia, Davide Rossi, Anastasia Chatzidimitriou, and Kostas Stamatopoulos designed and supervised the research and wrote the manuscript. All authors provided final approval of the manuscript.

CONFLICT OF INTEREST STATEMENT

P. G. reports fees for consulting and honoraria from AbbVie, AstraZeneca, BeiGene, Bristol Myers Squibb, Janssen, Loxo@Lilly, Merck Sharp & Dohme Corp., and Roche and research funding from AbbVie, AstraZeneca, and Janssen. K. S. has received honoraria from AbbVie, AstraZeneca, and Janssen and research support from AbbVie, AstraZeneca, Gilead Sciences, Janssen, and Roche. The other authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

FUNDING

The research work was supported in part by COSMIC (Combating Disorders of Adaptive Immunity with Systems Medicine), a Marie Curie European Training Network funded from the European Union's Horizon 2020 research and innovation program under grant agreement no. 765158; PureCell, funded by the Hellenic Foundation for Research and Innovation (HFRI) with project no #2810; CGI-Clinics, a European Union's Horizon 2022-2027 program under grant agreement 101057509; AIRC 5xmillie program (PI. R. Foà, grant agreement no #21198).

ORCID

Andreas Agathangelidis  <http://orcid.org/0000-0002-8467-7945>

Paolo Ghia  <http://orcid.org/0000-0003-3750-7342>

SUPPORTING INFORMATION

Additional supporting information can be found in the online version of this article.

REFERENCES

- Zucca E, Arcaini L, Buske C, et al. Marginal zone lymphomas: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2020;31(1):17-29.
- Stamatopoulos K, Belessi C, Papadaki T, et al. Immunoglobulin heavy- and light-chain repertoire in splenic marginal zone lymphoma. *Mol Med*. 2004;10(7-12):89-95.
- Traverse-Glehen A, Davi F, BenSimon E, et al. Analysis of VH genes in marginal zone lymphoma reveals marked heterogeneity between splenic and nodal tumors and suggests the existence of clonal selection. *Haematologica*. 2005;90(4):470-478.
- Zibellini S, Capello D, Forconi F, et al. Stereotyped patterns of B-cell receptor in splenic marginal zone lymphoma. *Haematologica*. 2010; 95(10):1792-1796.
- Hockley SL, Giannouli S, Morilla A, et al. Insight into the molecular pathogenesis of hairy cell leukaemia, hairy cell leukaemia variant and splenic marginal zone lymphoma, provided by the analysis of their IGH rearrangements and somatic hypermutation patterns. *Br J Haematol*. 2010;148(4):666-669.
- Bikos V, Darzentas N, Hadzidimitriou A, et al. Over 30% of patients with splenic marginal zone lymphoma express the same immunoglobulin

- heavy variable gene: ontogenetic implications. *Leukemia*. 2012;26(7):1638-1646.
7. Xochelli A, Bikos V, Polychronidou E, et al. Disease-biased and shared characteristics of the immunoglobulin gene repertoires in marginal zone B cell lymphoproliferations. *J Pathol*. 2019;247(4):416-421.
 8. Agathangelidis A, Chatzidimitriou A, Gemenetzi K, et al. Higher-order connections between stereotyped subsets: implications for improved patient classification in CLL. *Blood*. 2021;137(10):1365-1376.
 9. Hadzidimitriou A, Agathangelidis A, Darzentas N, et al. Is there a role for antigen selection in mantle cell lymphoma? Immunogenetic support from a series of 807 cases. *Blood*. 2011;118(11):3088-3095.
 10. Campos-Martín Y, Martínez N, Martínez-López A, et al. Clinical and diagnostic relevance of *NOTCH2* -and *KLF2*-mutations in splenic marginal zone lymphoma. *Haematologica*. 2017;102(8):e310-e312.
 11. Rossi D, Trifonov V, Fangazio M, et al. The coding genome of splenic marginal zone lymphoma: activation of *NOTCH2* and other pathways regulating marginal zone development. *J Exp Med*. 2012;209(9):1537-1551.
 12. Clipson A, Wang M, de Leval L, et al. *KLF2* mutation is the most frequent somatic change in splenic marginal zone lymphoma and identifies a subset with distinct genotype. *Leukemia*. 2015;29(5):1177-1185.
 13. Kiel MJ, Velusamy T, Betz BL, et al. Whole-genome sequencing identifies recurrent somatic *NOTCH2* mutations in splenic marginal zone lymphoma. *J Exp Med*. 2012;209(9):1553-1565.
 14. Arribas AJ, Rinaldi A, Mensah AA, et al. DNA methylation profiling identifies two splenic marginal zone lymphoma subgroups with different clinical and genetic features. *Blood*. 2015;125(12):1922-1931.
 15. Parry M, Rose-Zerilli MJ, Ljungström V, et al. Genetics and prognosis in splenic marginal zone lymphoma: revelations from deep sequencing. *Clin Cancer Res*. 2015;21(18):4174-4183.
 16. Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc MP. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *J Mol Recognit*. 2004;17(1):17-32.
 17. Giudicelli V, Duroux P, Rollin M, et al. IMGT® immunoinformatics tools for standardized V-DOMAIN analysis. *Methods Mol Biol*. 2022;2453:477-531.
 18. Lefranc MP. Nomenclature of the human immunoglobulin genes. *Curr Protoc Immunol*. 2001; Appendix 1:1.
 19. Bikos V, Stalika E, Baliakas P, et al. Selection of antigen receptors in splenic marginal-zone lymphoma: further support from the analysis of the immunoglobulin light-chain gene repertoire. *Leukemia*. 2012;26(12):2567-2569.
 20. Bikos V, Karypidou M, Stalika E, et al. An immunogenetic signature of ongoing antigen interactions in splenic marginal zone lymphoma expressing IGHV1-2*04 receptors. *Clin Cancer Res*. 2016;22(8):2032-2040.
 21. Swerdlow SH, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*. 2016;127(20):2375-2390.
 22. van Dongen JJM, Langerak AW, Brüggemann M, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98-3936. *Leukemia*. 2003;17(12):2257-2317.
 23. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT® 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*. 2012;882:569-604.
 24. Kotouza MT, Gemenetzi K, Galigalidou C, et al. TRIP-T cell receptor/immunoglobulin profiler. *BMC Bioinformatics*. 2020;21(1):422.
 25. Sofou E, Vlachonikola E, Zaragoza-Infante L, et al. Clonotype definitions for immunogenetic studies: proposals from the EuroClonality NGS Working Group. *Leukemia*. 2023;37(8):1750-1752.
 26. Zaragoza-Infante L, Junet V, Pechlivanis N, et al. IgIDivA: immunoglobulin intraclonal diversification analysis. *Brief. Bioinform*. 2022;23(5):bbac349.
 27. Schmitt D, Li S, Rozewicki J, et al. Repertoire builder: high-throughput structural modeling of B and T cell receptors. *Mol Syst Desi Eng*. 2019;4(4):761-768.
 28. Bikos V, Stalika E, Baliakas P, et al. Selection of antigen receptors in splenic marginal-zone lymphoma: further support from the analysis of the immunoglobulin light-chain gene repertoire. *Leukemia*. 2012;26(12):2567-2569.
 29. Webb B, Sali A. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics*. 2016;54:5.6.1-5.6.37.
 30. Kamburov A, Lawrence MS, Polak P, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci*. 2015;112(40):E5486-E5495.
 31. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D*. 2004;60(12):2256-2268.
 32. Ghraichy M, Galson JD, Kovaltsuk A, et al. Maturation of the human immunoglobulin heavy chain repertoire with age. *Front Immunol*. 2020;11:1734.
 33. Xochelli A, Bikos V, Polychronidou E, et al. Disease-biased and shared characteristics of the immunoglobulin gene repertoires in marginal zone B cell lymphoproliferations. *J Pathol*. 2019;247(4):416-421.
 34. Zaragoza-Infante L, Agathangelidis A, Papaioannou M, Chatzidimitriou A, Stamatopoulos K. The B cell receptor in marginal zone lymphoma ontogeny and evolution. *Ann Lymphoma*. 2020;4:10.
 35. Krantsevich A, Tang C, MacCarthy T. Correlations in somatic hypermutation between sites in IGHV genes can be explained by interactions between AID and/or Polη hotspots. *Front Immunol*. 2021;11:618409.
 36. Polychronidou E, Kalamaras I, Agathangelidis A, et al. Automated shape-based clustering of 3D immunoglobulin protein structures in chronic lymphocytic leukemia. *BMC Bioinformatics*. 2018;19(S14):414.
 37. Li H, Jiang Y, Prak EL, Radic M, Weigert M. Editors and editing of anti-DNA receptors. *Immunity*. 2001;15(6):947-957.
 38. Jang YJ, Stollar BD. Anti-DNA antibodies: aspects of structure and pathogenicity. *Cell Mol Life Sci*. 2003;60(2):309-320.
 39. Cohen IJ, Bochi-Layec AC, Kim KH, et al. Precision targeting of the malignant clone in diffuse large B cell lymphoma using chimeric antigen receptor T cells against the clonotypic IGHV4-34 B cell receptor. *Blood*. 2023;142(suppl 1):1020.
 40. Märkl F, Schultheiß C, Ali M, et al. Mutation-specific CAR T cells as precision therapy for IGLV3-21R110 expressing high-risk chronic lymphocytic leukemia. *Nat Commun*. 2024;15(1):993.