Distinct subtypes of diffuse large B-cell lymphoma defined by hypermutated genes

Amjad Alkodsi[1], Alejandra Cervera[1], Kaiyang Zhang[1], Riku Louhimo[1], Leo Meriranta[2,3], Annika Pasanen[2,3], Suvi-Katri Leivonen[2,3], Harald Holte[4], Sirpa Leppä[2,3], Rainer Lehtonen[1] & Sampsa Hautaniemi[1*]

[1]*Research Programs Unit, Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland*

[2]*Research Programs Unit, Applied Tumor Genomics, Faculty of Medicine, University of Helsinki, Helsinki, Finland*

[3]*Department of Oncology, Helsinki University Hospital Comprehensive Cancer Center, Helsinki,Finland*

[4]*Department of Oncology, Oslo University Hospital, Oslo, Norway*

**To whom correspondence should be addressed: sampsa.hautaniemi@helsinki.fi*

Abstract

Diffuse large B-cell lymphoma (DLBCL) is a biologically and clinically heterogeneous disease whose personalized clinical management requires robust molecular stratification. Here, we show that somatic hypermutation (SHM) patterns constitute a marker for DLBCL molecular classification. The activity of SHM mutational processes delineated the cell of origin (COO) in DLBCL. Expression of the herein identified 36 SHM target genes stratified DLBCL into four novel SHM subtypes. In a meta-analysis of patients with DLBCL treated with immunochemotherapy, the SHM subtypes were significantly associated with overall survival (1,642 patients) and progression-free survival (795 patients). Multivariate analysis of survival indicated that the prognostic impact of the SHM subtypes is independent from the COO classification and the International Prognostic Index. Furthermore, the SHM subtypes had a distinct clinical outcome within each of the COO subtypes, and strikingly, even within unclassified DLBCL. The genetic landscape of the four SHM subtypes indicated unique associations with driver alterations and oncogenic signaling in DLBCL, which suggests a possibility for therapeutic exploitation. These findings provide a biologically driven classification system in DLBCL with potential clinical applications.

Introduction

Diffuse large B-cell lymphoma (DLBCL) is the most common lymphoid neoplasm in adults. The standard immunochemotherapy regimen consisting of rituximab, cyclophosphamide, doxorubicin, vincristine and prednisone (R-CHOP) cures approximately 60% of patients[1,2]. DLBCL is classified according to the cell of origin (COO) into germinal center B-cell (GCB) or activated B-cell (ABC), while leaving out about 10% of cases unclassified. The COO subtypes have distinct gene expression profiles, and patients with the ABC subtype have significantly worse survival[3–5]. However, substantial heterogeneity in patient outcome still exists within each COO subtype, which necessitates a search for additional stratification tools that can also serve as biomarkers for tailored therapies.

The genetic landscape of primary DLBCL has been widely explored by several sequencing studies that revealed a high inter-individual heterogeneity with a wide repertoire of driver genes[6–12]. Several of DLBCL driver genes harbor mutations within the known hotspot motifs of somatic hypermutation (SHM), a process that normally targets immunoglobulin variable region during B-cell development[13]. SHM is initiated in B-cells by activation-induced cytidine deaminase (AID). AID deaminates cytosines creating uracil mismatches that can be repaired by several pathways. The choice of repair pathway determines the type of substitution, and whether error-prone polymerases such as polymerase η are recruited[14,15]. Aberrant SHM targets several cancer genes downstream of transcription start sites (TSS)[13,16], which has a key role in germinal-center derived lymphomagenesis[17].

The clinical and molecular heterogeneity of DLBCL constitute a major obstacle in patient management. Therefore, various molecular classification approaches have been suggested[10,11,18–21].

However, the previously proposed classification systems have limitations, which leaves capacity for improvement. Monti et al. used unsupervised clustering of gene expression to identify distinct groups of DLBCL, but the identified groups did not associate with clinical outcome[18]. Dybkær et al. used normal B-cell associated gene expression signatures for stratification of DLBCL, but the distinct clinical outcome was only seen within the GCB subtype[19]. Chapuy et al. and Schmitz et al. used genomic alterations to classify DLBCL into several groups with distinct clinical outcome[10,11]. However, the generalization applicability and external validation of their proposed genetic subtypes remain to be seen. Furthermore, more than half of DLBCL patients remain genetically unclassified using the classification method proposed by Schmitz et al. Most recently, Ennishi et al. and Sha et al. extracted gene expression signatures that recognize a small fraction of the GCB subtype associating with double-hit and molecular high-grade DLBCL, respectively[20,21]. Remarkably, none of the previous approaches fully exploited SHM and its transcriptional dependency.

Here, we explored the activity of SHM mutational processes in DLBCL, which showed a striking segregation with the cell of origin. Using unsupervised clustering of expression of 36 SHM target genes, we identified four distinct subtypes of DLBCL. These subtypes had a significant association with overall and progression-free survival. Finally, characterization of the genetic landscape of the identified subtypes indicated a distinct underlying biology and oncogenic signaling that could potentially be exploited therapeutically.

Materials and Methods

Patient cohorts

We used RNA and whole-genome sequencing (WGS) data of primary DLBCL obtained from the cancer genome characterization initiative (CGCI)[22] (97 patients with RNA-sequencing of which 39 had also WGS) in addition to in-house WGS data from seven primary-relapse sample pairs as a discovery cohort. For validation, we used gene expression data from the Lenz et al. cohorts[5] (233 R-CHOP treated and 181 CHOP treated patients; microarrays), Visco et al. cohort[23] (470 R-CHOP treated patients; microarrays), Schmitz et al. cohort[10] (234 patients treated with R-CHOP-like treatment; RNA sequencing), Chapuy et al. cohort[11] (101 patients treated with R-CHOP; microarrays), Arthur et al. cohort[12] (383 patients; RNA sequencing), and Reddy et al. cohort[6] (604 patients treated with R-CHOP; RNA sequencing). We also utilized genomic alterations from Schmitz et al. cohort (521 patients), Reddy et al. cohort (604 patients) and Arthur et al. cohort (153 patients).

Analysis of sequencing data from the CGCI cohort

Raw sequencing reads for 39 whole-genome sequencing DLBCL samples with matched normal controls were downloaded from dbGaP study accession *phs000532*. Sequencing reads were mapped to human reference genome hg19 using the Burrows-Wheeler aligner BWA (v0.5.9-r16). Standard bam file processing was performed including marking duplicates using PICARD tools (v1.84), base quality recalibration and realignment around indels using GATK (v2.7-2). Somatic base substitutions were detected using MuTect (v1.1.7) using matched tumor-normal pairs and a panel of normals (constructed by MuTect using 39 normal WGS samples from CGCI)

as inputs. Processing of the primary-relapse WGS data is described in Supplementary Materials and Methods.

RNA raw sequencing reads were downloaded from dbGaP study accession *phs000532*. We used Sepia pipeline[24] for quality control, alignment and quantification. Illumina adapters, when present, and low quality bases at either end were cropped using Trimmomatic (v0.329). Sequences shorter than 25 base pair long after trimming were excluded. Reads from each run were independently aligned to human-ensembl 37 and annotation version 75 with STAR (v2.4.2a modified) 2-pass mapping. Gene expression was quantified using eXpress[25] (v1.5.1). All processing steps and analyses were performed within the Anduril 2.0 computational framework[26,27].

Clustered mutation signatures analysis

Our approach in discerning clustered mutation signatures was adapted from Supek & Lehner study[28]. We categorized base substitutions into clustered when the distance to closest mutation with the same substitution and strand orientation was <1000, or unclustered otherwise (Supplementary Fig. 1A). The clustering information was incorporated with the standard 96 classes of substitutions (six possible base substitutions in 16 possible trinucleotide context) resulting in 192 classes. Next, mutation type counts per sample were normalized so that the sum of clustered mutations is 0.5 and the sum of unclustered mutations is 0.5 in each sample. This was done to balance the weight between clustered and unclustered mutations as the number of unclustered mutations is largely higher. Finally, non-negative matrix factorization (NMF) was used to discern clustered mutation signatures using the *NMF* R package[29]. The consensus of 1000 NMF runs was used. To determine the optimal number of signatures, we performed the

extraction using 2-10 signatures 200 times each, and subsequently cophenetic correlation and explained variance measurements were used to decide the optimal number of signatures (Supplementary Fig. 1B). Estimation of absolute and relative signature contribution is described in Supplementary Materials and Methods.

Identification of SHM target genes and the SHM subtypes

We restricted our search to genomic regions between transcription start sites (TSS) of protein-coding genes and 2500 basepairs downstream of TSS based on exploratory analysis visualized in Fig. 1B. In each examined target region with at least five substitutions, we tested the null hypothesis that the probability of observed number of clustered substitutions in R<u>C</u>H context (or <u>T</u>W context) is equal or less than the background rate of clustered R<u>C</u>H (or <u>T</u>W) substitutions (one-sided binomial test). The background rate was computed based on the region between 5000bp to 2500bp upstream of TSS. P-values were adjusted for multiple hypothesis testing using the false discovery rate (FDR) method, and a threshold of 0.1 was used. Binomial test p-values were computed in R using the *binom.test* function.

The identified SHM target genes were used for gene expression consensus hierarchical clustering in 97 DLBCL samples from the CGCI cohort. First, we computed the inter-quartile range (IQR) of the expression of each gene, and then removed 25% of the genes with the lowest IQR values leaving out 36 genes. Next, we performed consensus hierarchical clustering using the *ConsensusClusterPlus* R package[30] with the following parameters: *reps* = 10000, *pItem* = 0.75, *pFeature* = 0.75, *distance* = euclidean, *innerLinkage* = ward.D, and *finalLinkage* = ward.D. The number of clusters was determined by evaluating the relative change in area under the cumulative distribution function (CDF) curve (Supplementary Fig. 1C). Cluster centroids were computed as

the mean z-score of each gene in samples belonging to that cluster. All the parameters were decided independently from clinical data, and locked before performing downstream analyses.

To validate that the extracted four clusters can be reproduced in other larger DLBCL cohorts, we used a previously published method implemented in the ClusterRepro R package[31] (Supplementary Materials and Methods).

SHM subtyping in validation cohorts

Obtaining and processing (when applicable) of gene expression and genetic alterations data in validation cohorts is described in Supplementary Materials and Methods. To determine the SHM subtypes for the samples in validation cohorts, gene expression values were standardized to z-scores (zero mean and unit variance) in each dataset. Then, samples were classified to the cluster (SHM subtype) with the nearest centroid (maximum Pearson correlation coefficient).

Statistical analysis

All statistical analyses were done in R (>v3.4.3). Kaplan-Meier and multivariate Cox regression analyses were done using the *survival* (v2.43-3) R package. All statistical tests were two-sided unless otherwise specified. Correction for multiple hypotheses testing was performed when needed.

Results

Clustered mutation signatures delineate the cell of origin in DLBCL

Using a clustered mutations approach in a discovery set of 53 DLBCL samples, we extracted four highly-reproducible clustered mutation signatures (Fig. 1A, Supplementary Fig. 1B and

Supplementary Table 1). Two of these signatures were similar to the previously known Signature 5 and Signature 17 (unknown etiology)[32]. The third signature was characterized by clustered thymine mutations at TW (W is A or T) context, resembling the known motif of polymerase η and henceforth labeled Signature TW. The fourth signature was characterized by clustered cytosine mutations and showed similarity to the known RCH (R is A or G; H is C, T or A) motif of AID mediated mutagenesis[33], and henceforth labeled Signature RCH. Clustered mutations of Signature TW and Signature RCH specifically targeted the downstream region of transcription start sites (TSS) (Fig. 1B), which is a characteristic of somatic hypermutation.

The contribution of signatures TW and RCH demonstrated a striking difference between the ABC and GCB subtypes (Fig. 1C and Supplementary Table 1). ABC cases had a significantly higher relative contribution of Signature RCH ($p<0.0001$), whereas GCB cases had a significantly higher relative contribution of Signature TW ($p=0.0001$; Fig. 1C). The absolute contribution (number of contributed mutations) of the two signatures showed a similar significant contrast between ABC and GCB subtypes (Supplementary Fig. 2A). With such high contrast in signatures TW and RCH contribution between GCB and ABC subtypes, COO classification using exposure to the SHM processes was achievable. We used a classifier that assigns GCB subtype when the relative contribution of Signature TW was higher than the relative contribution of Signature RCH, and ABC subtype otherwise. That classifier assigned the correct COO in 30 out of 35 GCB and ABC cases (accuracy: 0.86; sensitivity: 0.75; specificity: 0.95; positive class: ABC), on a par with several immunohistochemical methods[34] that are considered sufficient for determination of COO in DLBCL[35].

We sought to validate the reproducibility of Signature TW and Signature RCH, and the significant difference in their contribution between ABC and GCB subtypes in an expanded set of 153 WGS analyzed with different variant calling pipelines from Arthur et al. (Supplementary Fig. 3A-B and Supplementary Table 1). A comprehensive characterization of other mutational signatures in this cohort has been done previously by Arthur et al. The difference in relative and absolute contribution of the two SHM mutational signatures between ABC and GCB cases was also significant (Fig. 1D and Supplementary Fig. 2B), and sufficient to distinguish the cell of origin in 71/91 samples with ABC/GCB subtypes (accuracy: 0.78; sensitivity: 0.74; specificity: 0.81). The two SHM mutational signatures presented here closely resemble the exome-derived signatures reported by Chapuy et al., whose contributions were also significantly different between the ABC and GCB subtypes (Supplementary Fig. 3C). However, the difference was not sufficient to distinguish the COO subtypes, which can be explained by the fact that the imbalance in the number of RCH and TW mutations between the ABC and GCB subtypes is more pronounced in the intronic and intergenic regions than in the exons (Supplementary Fig. 3D). Signature TW has a similar characteristics to the known Signature 9, whose exposure was found to be significantly higher in the GCB subtype[12]. Collectively, these results confirm previous findings and suggest that mutational spectra could be an important component in a genetic classifier that determines the cell of origin in DLBCL from non-invasive genetic testing, provided that targeted sequencing covers beyond the exonic regions. The results also emphasize that somatic hypermutation activity could be used to identify distinct disease entities in DLBCL.

Target genes of aberrant SHM define distinct transcriptional profiles in DLBCL

Somatic hypermutation requires and correlates with transcription of its target genes[15,36]. Given our finding that the two SHM-related mutational processes were differentially active between ABC and GCB subtypes, and given their transcriptional dependence, we hypothesized that the expression of SHM target genes may constitute a biologically and clinically meaningful pattern. To investigate this, we first identified the target genes of aberrant SHM. Using the discovery cohort, we identified 38 genes targeted by RCH mutations and 16 genes targeted by TW mutations (Supplementary Table 1). The majority of the targets genes we detected have been also identified by one or more of the previous studies (Supplementary Table 1) [10,12,12,16].

Next, we used unsupervised consensus hierarchical clustering of expression of 36 SHM target genes, which yielded four clusters of patients with distinct association to COO subtypes (Fig. 2A and Supplementary Table 2). We henceforth call these clusters SHM subtypes (SHM1-4) since all the genes defining their expression profiles were identified as targets of aberrant hypermutation. Fig. 2B displays the average expression of each of the 36 genes in each cluster/subtype. GCB cases predominated SHM1 and SHM3 subtypes, whereas ABC cases were the vast majority in SHM2 and small majority in SHM4 subtypes. Overall survival was different between the four groups of patients (p=0.05; Fig. 2C).

We tested whether the four extracted clusters could be reproduced in independent larger cohorts (Lenz et al., Visco et al., Arthur et al., Reddy et al. and Schmitz et al. cohorts). Our results indicated that each of the four clusters was reproducible in at least three of the tested cohorts (Supplementary Table 2).

The SHM subtypes of DLBCL are associated with patient outcome

To generalize the SHM subtypes established in the discovery cohort, we applied SHM subtyping to seven DLBCL cohorts (Supplementary Table 3 and Supplementary File 1). The four SHM subtypes retained a similar relative composition of ABC and GCB cases in all tested cohorts (Supplementary Fig. 4). SHM subtype assignment had a significant association with overall survival in Lenz et al. cohorts (R-CHOP treated: $p<0.0001$; CHOP treated: $p=0.0006$), Visco et al. cohort ($p=0.0005$), Schmitz et al. cohort ($p<0.0001$), Arthur et al. cohort ($p=0.0059$) and Reddy et al. cohort ($p=0.001$), but was not significant in Chapuy et al. cohort ($p=0.07$), as shown in Supplementary Fig. 4. Progression-free survival (PFS) was significantly different between the SHM subtypes in Visco et al. cohort ($p<0.0001$), and Schmitz et al. cohort ($p<0.0001$), but not in Chapuy et al. cohort ($p=0.46$) as shown in Supplementary Fig. 4. Time-to-progression (TTP) differed significantly between the SHM subtypes in Arthur et al. cohort ($p=0.0018$). Using the Reddy et al. cohort, we tested whether the SHM subtypes were influenced by different levels of tumor purity (Supplementary Table 3), and found no significant association ($p=0.28$; Chi-squared test).

The lack of significance in Chapuy et al. cohort could be attributed to the small cohort size and the inherent bias in the demography of the cohort (Supplementary Fig. 5). The demographies of the different cohorts also influence the relative difference in survival between the SHM subtypes. For example, SHM4 in Chapuy et al. cohort did not have any case with high International Prognostic Index (IPI) score, which explains the longer survival of SHM4 than SHM3 in Chapuy et al. cohort. Similarly, SHM1 in Reddy et al. cohort had the highest fraction of IPI 4-5 cases and the lowest fraction of IPI 0 cases among all cohorts, which could explain the worse survival of SHM1 than SHM2 in that cohort.

We conducted a meta-analysis of DLBCL patients treated with R-CHOP-like therapy. The SHM subtypes were significantly associated with overall survival (1,642 patients from five cohorts; $p<0.0001$; Fig. 3A), and progression-free survival (795 patients from three cohorts; $p<0.0001$; Supplementary Fig. 6A). The SHM subtypes, IPI scores, and COO subtypes remained significant in a multivariate analysis of overall survival using Cox proportional hazard regression model (Fig. 3B). In a multivariate analysis of progression-free survival, the SHM subtypes along with IPI scores remained significant (Supplementary Fig. 6B). These results indicate that the SHM subtypes confer an additional clinical prognostic value, not captured by the COO subtype or the IPI score.

We examined overall and progression-free survival separately within each of the COO subtypes. The SHM subtypes were significantly associated with overall survival within the GCB subtype ($p<0.0001$; Fig. 3C), the ABC subtype ($p=0.027$; Fig. 3D), and strikingly, within unclassified DLBCL ($p=0.012$; Fig. 3E). Progression-free survival was significantly different between the SHM subtypes within the GCB subtype ($p=0.002$; Supplementary Fig. 6C), the ABC subtype ($p=0.0033$; Supplementary Fig. 6D), but not within unclassified DLBCL ($p=0.21$; Supplementary Fig. 6E), which had only 100 patients with available PFS information in comparison to 226 patients with available overall survival.

Genetic characteristics of the SHM subtypes

We utilized genomic alterations data from Schmitz et al. (521 samples) and Reddy et al. (604 samples) to explore the genomic landscape characterizing the SHM subtypes. The genetic alterations attributed to each SHM subtype exhibited profound similarity between the two cohorts (Fig. 4A,

Supplementary Fig. 7 and Supplementary Table 4), highlighting the robustness and biological relevance of our classification method. The SHM1 subtype was characterized by high frequency of alterations targeting genes related to chromatin remodeling and histone modifications such as *EZH2*, *KMT2D* and *CREBBP*. *MYC* alterations and *BCL2* translocations were also at highest frequency in SHM1. Alterations in the G-protein signaling pathway such as *GNA13*, *GNAI2* and *P2RY8*, in addition to whole-chromosomal gains of chromosomes 7 and 12 were specifically common in SHM1 (Fig. 4A and Supplementary Fig. 7).

SHM2 had the highest frequency of *MYD88* (L265P) and *CD79B* in addition to other alterations that characterize the ABC subtype such as *CDKN2A* homozygous deletions, *PIM1*, *MPEG1*, *ETV6* and *IRF4* mutations. Whole-chromosomal gains of chromosomes 18 and 3 were specifically common in SHM2.

SHM3 was characterized by alterations in genes of the JAK-STAT pathway such as *SOCS1*, *STAT3* and *STAT6*. Additionally, this subtype had the highest frequency of mutations in *TNFAIP3*, *SGK1* and *IRF8*. Whole-chromosomal gains (chromosomes 7, 18 and 3) were all significantly depleted in this subtype.

SHM4 had the highest frequency of *BCL6* fusions in contrast to SHM2 that had the lowest frequency. Alterations in *CD70*, *BCL10*, *SPEN* and *MYD88* (other than L265P) were most frequent in this subtype. Additionally, mutations in the *HLA-A*, *HLA-B* and *HLA-C* genes were all significantly enriched in SHM4.

Most of the recurrent genetic alterations in DLBCL were significantly associated with the SHM subtypes as well as with the COO subtypes (Fig. 4A and Supplementary Fig. 7). However, some alterations exhibited significant association exclusively with the SHM subtypes such as the

alterations in *KMT2D* (enriched in SHM1 and depleted in SHM3), *MYC* (enriched in SHM1 and depleted in SHM4), *TNFAIP3* (enriched in SHM3 and depleted in SHM1), *TP53* (depleted in SHM3) and *CD70* (enriched in SHM4 and depleted in SHM2).

The genes most frequently altered in each of the SHM subtypes showed a clear similarity to the core genes defining the genetic subtypes reported by Schmitz et al.[10]. *BCL2* translocations and *EZH2* mutations defining the EZB subtype were most frequent in SHM1, *MYD88* (L265P) and *CD79B* mutations defining the MCD subtype were most common in SHM2, and *BCL6* translocations and *NOTCH2* mutations defining the BN2 subtype were at highest frequency in SHM4. These similarities explain the distinct relationship between the genetic and SHM subtypes displayed in Fig. 4B. Yet, substantial difference between the two classification systems exists. First, more than half of DLBCL cases remain genetically unclassified and these constituted roughly half of the cases in each SHM subtype (Fig. 4B). Second, the core alteration defining a certain genetic subtype can be found in all SHM subtypes, but at different frequencies. For example, *BCL6* fusions that characterize 73% of the BN2 subtype were found in 28% of SHM4 and 17% of non-SHM4 cases. Finally, SHM3 was not similar to any of the genetic subtypes.

We examined the prognostic impact of the SHM subtypes in relation to the genetic subtypes. The SHM subtypes differed significantly in overall survival and progression-free survival within the BN2 genetic subtype (OS: p=0.0013; PFS: p=0.0028; Fig. 4C and Supplementary Fig. 8A), and among the genetically unclassified cases (OS: p=0.0066; PFS: p=0.0022; Fig. 4D and Supplementary Fig. 8B). Taken together, the SHM subtypes provided additional prognostic value to the genetic subtypes.

Finally, we compared the SHM classification to the genetic groups defined by Chapuy et al. and the double-hit transcriptional signature (DHITsig) extracted by Ennishi et al. The C5 group characterized by chromosome 18q gain and frequent *MYD88* mutations by Chapuy et al. Had almost exclusively SHM2 subtype, while the other five groups (C0-4) did not show any marked association with the SHM subtypes (Supplementary Fig. 9A). The DHITsig-positive cases in Arthur et al. cohort were strongly enriched in SHM1 subtype (Supplementary Fig. 9B).

The patterns of somatic hypermutation across the SHM subtypes

We utilized 103 samples from Arthur et al. with both WGS and RNA-sequencing data to shed light on the patterns of somatic hypermutation of the SHM subtyping genes. Several of the SHM subtyping genes were hypermutated at different frequencies in the SHM subtypes such as *BCL2*, *MYC*, *IRF4*, *PIM1* and *CD74*, or were almost exclusively hypermutated in a specific SHM subtype such as *C1orf186*, *MPEG1* and *SOCS1* (Fig. 5A-B, Supplementary Fig. 10 and Supplementary Table 5). The specific locus of hypermutation within the target 2.5Kbp downstream of TSS was consistently the same between the four SHM subtypes in the majority of the SHM subtyping genes. Exceptionally, *BCL11A*, *KLHL6* and *DTX1* showed a pattern suggestive of positional hypermutation specificity between the four SHM subtypes (Fig. 5A-B). We examined the well-known correlation between hypermutation and elevated gene expression. Indeed, the majority of the SHM subtyping genes had significantly higher expression in samples with mutations at the target region spanning 2.5Kbp downstream of TSS (Supplementary Fig. 11).

Since some of the SHM subtyping genes did not show an association with elevated expression and/or did not show any relative difference in hypermutation between the four SHM subtypes, we

sought to assess the importance of each gene in SHM subtyping. We computed the fraction of samples that changed the assigned SHM subtype when a gene was omitted (Fig. 5C). The gene with the highest importance score was *IRF4*, followed by *DTX1*, which has been associated with survival in DLBCL[37], *BCL2*, *FCRL3* and *AICDA*. The genes whose hypermutation did not relatively differ between the SHM subtypes nor correlate with elevated expression were at the other end of the least important SHM subtyping genes. Remarkably, the maximum importance score (*IRF4*) was only 0.14, which indicates that the prognostic impact of SHM subtyping is an additive result of multiple genes.

Discussion

The SHM transcriptional program uncovered in this study by fully unsupervised analysis provides a new molecular classification system in DLBCL. The identified SHM subtypes have significantly different clinical outcome, as demonstrated in several independent cohorts with different sample collection methods, analysis platforms and data processing protocols. With independence from previously proposed genetic, transcriptomic and clinical markers, SHM subtyping captures previously unexplained clinical heterogeneity, and significantly improves prognostic stratification in DLBCL.

Our proposed classification divides the GCB class of DLBCL into two major subtypes (SHM1 and SHM3) with significantly different clinical outcome. The SHM1 group consists of mostly GCB cases with dismal outcome after standard immunochemotherapy who may benefit from alternative treatment strategies. In contrast, the SHM3 group has a superior outcome after standard R-CHOP treatment, which may be continued as the optimal treatment for the patients of this group. The ABC class of DLBCL can be divided according to SHM subtyping into two major groups: SHM4, and

SHM2 that has the worst survival among all SHM subtypes and the most acute need for modified treatment. A striking finding of this study is that unclassified DLBCL can be classified according to the SHM subtypes into four groups with distinct overall survival, which can enhance the rational treatment options for the patients with unclassified DLBCL.

The distinct genetic landscapes of the SHM subtypes reflected the unique underlying biology for each group. Remarkably, recently recognized genetic subtypes defined by Schmitz et al. were strongly associated with different SHM subtypes. However, our classifier was able to also classify patients with genetically indistinct features. The SHM1 subtype had a high frequency of *BCL2* and *MYC* alterations that were not necessarily concurrent, in addition to mutations affecting chromatin modifying genes. The SHM2 subtype was characterized by abundance of alterations in the B-cell receptor signaling pathway, which can be therapeutically manipulated with kinase inhibitors including Bruton tyrosine kinase inhibitor Ibrutinib. SHM3 had a high prevalence of mutations in the JAK-STAT pathway, and was found highly responsive to standard R-CHOP treatment in DLBCL. SHM4 had the highest rate of *BCL6* fusions in addition to mutations affecting *CD70* and *BCL10*. Retrospective evaluation of clinical trial data in the context of the SHM subtypes can shed light on the possible treatment options for each of the four SHM subtypes.

The transcriptionally defined SHM subtypes correlated with unique patterns of aberrant somatic hypermutation. However, the mutational patterns of SHM could not accurately define the SHM subtypes due to several factors such as the sparse nature of mutation data and the fact that not all highly transcribed genes are hypermutated. Somatic hypermutation per se does not fully explain the phenotypic and genetic differences between the four SHM subtypes, but is rather a proxy to identify distinct entities of the disease. The mechanism by which target genes

attract SHM is still not well understood, but is far from random. Indeed, the vast majority of the target genes have relevant functions in lymphoma and B-cell biology given their expression during germinal center reaction. Our approach of restricting the expression profiling to SHM targets exploits a non-random program in lymphomagenesis to stratify the disease into biologically relevant subtypes. The unsupervised discovery of the four subtypes ensures a natural segregation of the disease, which is advantageous to other supervised approaches that led to partially similar grouping[10,20,21]. Aberrant SHM is also prevalent to various extents in other non-Hodgkin lymphomas. Therefore, extending the SHM subtypes to other lymphomas is plausible and may shed light on novel disease groups with distinct prognosis.

Acknowledgements

This work was supported financially by the Academy of Finland (S.H., S.L.), the Sigrid Jusélius Foundation (S.H., S.L.), Finnish Cancer Foundations (S.H., S.L.), Helsinki University Hospital (S.L.), the Biomedicum Helsinki Foundation (A.A.) and the University of Helsinki graduate program (A.A.). The results published here are in part based upon data generated by the Cancer Genome Characterization Initiative (Non-Hodgkin Lymphoma project), and data generated by the Genomic Variation in Diffuse Large B Cell Lymphoma study, which was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, Department of Health and Human Services. Both datasets have been accessed through the NIH database for Genotypes and Phenotyeps (dbGAP) with accession numbers phs000532 and phs001444. Information about CGCI projects can be found at https://ocg.cancer.gov/programs/cgci. The authors thank Dr. Ryan D. Morin and Dr. Sandeep Dave for providing gene expression data. The international DLBCL consortia whose data we have used in this study are gratefully acknowledged.

The authors thank professors Ville Mustonen and Jussi Taipale for critical review of the article. Computing resources from CSC – IT Center for Science Ltd. and technical assistance from Anne Aarnio and Marika Tuukkanen are gratefully acknowledged.

Author contributions

A.A. conceptualized the study together with R. Lehtonen, S.L. and S.H.; A.A. designed the methodology, performed the analyses, made the figures and wrote the paper; A.A., A.C., K.Z. and R. Louhimo processed sequencing data; A.P., SK.L., L.M., H.H. and S.L. provided resources and materials for the in-house samples and participated in scientific discussions; S.H., S.L. and R. Lehtonen supervised the study; S.H. and S.L. acquired funding; All authors read and edited the paper.

Competing interests

The authors declare no competing interests.

REFERENCES

1. Coiffier B, Lepage E, Brière J, Herbrecht R, Tilly H, Bouabdallah R, et al. CHOP Chemotherapy plus Rituximab Compared with CHOP Alone in Elderly Patients with Diffuse Large-B-Cell Lymphoma. New England Journal of Medicine. 2002;346(4):235–242.

2. Pfreundschuh M, Trümper L, Österborg A, Pettengell R, Trneny M, Imrie K, et al. CHOP-like chemotherapy plus rituximab versus CHOP-like chemotherapy alone in youngpatients with good-prognosis diffuse large-B-cell lymphoma: a randomised controlled

trialby the MabThera International Trial (MInT) Group. The Lancet Oncology. 2006;7(5):379– 391.

3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000 feb;403(6769):503–511.

4. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. New England Journal of Medicine. 2002;346(25):1937–1947. PMID: 12075054.

5. Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, et al. Stromal Gene Signatures in Large-B-Cell Lymphomas. New England Journal of Medicine. 2008;359(22):2313–2323.

6. Reddy A, Zhang J, Davis NS, Moffitt AB, Love CL, Waldrop A, et al. Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. Cell. 2017;171(2):481 – 494.e15.

7. Morin RD, Mungall K, Pleasance E, Mungall AJ, Goya R, Huff RD, et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. Blood. 2013;122(7):1256–1265.

8. Lohr JG, Stojanov P, Lawrence MS, Auclair D, Chapuy B, Sougnez C, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proceedings of the National Academy of Sciences. 2012;109(10):3879–3884.

9. Dubois S, Viailly PJ, Mareschal S, Bohers E, Bertrand P, Ruminy P, et al. Next-Generation Sequencing in Diffuse Large B-Cell Lymphoma Highlights Molecular Divergence and Therapeutic Opportunities: a LYSA Study. Clinical Cancer Research.2016;22(12):2919–2928.

10. Schmitz R, Wright GW, Huang DW, Johnson CA, Phelan JD, Wang JQ, et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. New England Journal of Medicine. 2018;378(15):1396–1407.

11. Chapuy B, Stewart C, Dunford AJ, Kim J, Kamburov A, Redd RA, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nature Medicine. 2018 may;24(5):679–690.

12. Arthur SE, Jiang A, Grande BM, Alcaide M, Cojocaru R, Rushton CK, et al. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. Nature Communications. 2018 dec;9(1):4001.

13. Pasqualucci L, Neumeister P, Goossens T, Nanjangud G, Chaganti R, Küppers R, et al. Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. Nature.2001;412(6844):341–346.

14. Liu M, Duke JL, Richter DJ, Vinuesa CG, Goodnow CC, Kleinstein SH, et al. Twolevels of protection for the B cell genome during somatic hypermutation. Nature. 2008;451(7180):841–845.

15. Ramiro AR, Stavropoulos P, Jankovic M, Nussenzweig MC. Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. Nature immunology. 2003;4(5):452–456.

16. Khodabakhshi AH, Morin RD, Fejes AP, Mungall AJ, Mungall KL, Bolger-Munro M, et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. Oncotarget. 2012;3(11):1308.

17. Pasqualucci L, Bhagat G, Jankovic M, Compagno M, Smith P, Muramatsu M, et al. AID is required for germinal center–derived lymphomagenesis. Nature genetics. 2008;40(1):108.

18. Monti S, Savage KJ, Kutok JL, Feuerhake F, Kurtin P, Mihm M, et al. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood. 2005;105(5):1851–1861.

19. Dybkær K, Bøgsted M, Falgreen S, Bødker JS, Kjeldsen MK, Schmitz A, et al. Diffuse Large B-Cell Lymphoma Classification System That Associates Normal B-Cell Subset Phenotypes With Prognosis. Journal of Clinical Oncology. 2015;33(12):1379–1388.

20. Ennishi D, Jiang A, Boyle M, Collinge B, Grande BM, Ben-Neriah S, et al. Double-Hit Gene Expression Signature Defines a Distinct Subgroup of Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma. Journal of Clinical Oncology. 0;0(0):JCO.18.01583.

21. Sha C, Barrans S, Cucco F, Bentley MA, Care MA, Cummin T, et al. Molecular High-Grade B-Cell Lymphoma: Defining a Poor-Risk Group That Requires Different Approaches to Therapy. Journal of Clinical Oncology. 0;0(0):JCO.18.01314.

22. Morin RD, Johnson NA, Severson TM, Mungall AJ, An J, Goya R, et al. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. Nature genetics. 2010;42(2):181.

23. Visco C, Li Y, Xu-Monette ZY, Miranda RN, Green TM, Li Y, et al. Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: a report from the International DLBCL Rituximab-CHOP Consortium Program Study. Leukemia. 2012 sep;26(9):2103–2113.

24. Icay K, Chen P, Cervera A, Rantanen V, Lehtonen R, Hautaniemi S. SePIA: RNA and small RNA sequence processing, integration, and analysis. BioData mining. 2016;9(1):20.

25. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nature methods. 2013;10(1):71–73.

26. Cervera A, Rantanen V, Ovaska K, Laakso M, Nuñez-Fontarnau J, Alkodsi A, et al. Anduril 2: Upgraded large-scale data integration framework. Bioinformatics. 2019;btz133.

27. Ovaska K, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, Aittomäki V, et al. Large-scaledata integration framework provides a comprehensive view on glioblastoma multiforme.Genome medicine. 2010;2(9):65.

28. Supek F, Lehner B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. Cell. 2017;170(3):534–547.

29. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC bioinformatics. 2010;11(1):367.

30. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26(12):1572–1573.

31. Kapp AV, Tibshirani R.Are clusters found in one dataset present in another dataset? Biostatistics. 2006;8(1):9–31.

32. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013 aug;500(7463):415–421.

33. Rogozin IB, Diaz M. Cutting edge: DGYW/WRCH is a better predictor of mutability at G: C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. The Journal of Immunology. 2004;172(6):3382–3384.

34. Meyer PN, Fu K, Greiner TC, Smith LM, Delabie J, Gascoyne RD, et al. Immunohistochemical Methods for Predicting Cell of Origin and Survival in Patients With Diffuse Large B-Cell Lymphoma Treated With Rituximab. Journal of Clinical Oncology. 2011;29(2):200–207.

35. Swerdlow SH, Campo E, Pileri SA, Harris NL, Stein H, Siebert R, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood. 2016;127(20):2375–2390.

36. Bachl J, Carlson C, Gray-Schopfer V, Dessing M, Olsson C. Increased transcription levels induce higher mutation rates in a hypermutating cell line. The Journal of Immunology. 2001;166(8):5051–5057.

37. Meriranta L, Pasanen A, Louhimo R, Cervera A, Alkodsi A, Autio M, et al. Deltex-1 mutations predict poor survival in diffuse large B-cell lymphoma. Haematologica. 2017;102(5):e195–e198.

Figures

Figure 1: Clustered mutation signatures in DLBCL

(A) Four signatures of mutational processes extracted by non-negative matrix factorization (NMF). Each signature bar plot has six panels representing six possible base substitutions each broken down into 16 possible trinucleotide contexts on the horizontal axis. The stacked bars show NMF weight for clustered (red) and unclustered (blue) mutations. Counts of clustered and unclustered mutations were normalized before NMF extraction to balance the weight between the two classes (Methods). (B) Fraction of clustered mutations at RCH and TW context were computed at various distances before and after transcription start sites (TSS) of protein coding genes and smoothed using the loess method (95% confidence intervals are displayed). (C-D) Box plots comparing Signature RCH and Signature TW relative contributions between activated B-cell (ABC) and germinal center B-cell (GCB) cases in (C) the CGCI cohort and (D) the extended cohort. Box plots represent median and inter-quartile range. P-values were generated using two-sided Mann-Whitney test.

Figure 2: Discovery of four SHM subtypes using expression of 36 genes targeted by aberrant SHM

(A) Heatmap of gene expression (normalized z-scores) of 36 genes in 97 primary DLBCL cases in the CGCI cohort. Consensus hierarchical clustering of the 97 cases yielded four clusters indicated by color on top of the heat map. The columns dendrogram is based on the consensus of 10,000 iterations. (B) Panel of bar plots defining the 36 dimensional centroids of the four clusters (SHM subtypes). Each bar plot shows one-gene centroid coordinates (corresponding to average expression) of the four clusters (SHM subtypes). The border of the plot is colored with the cluster color of the maximum value. (C) Kaplan-Meier plot comparing overall survival between the four SHM subtypes in the CGCI cohort. P-value is computed by the log-rank test. A risk table showing the number of patients at risk for each group at specific time intervals is displayed below the survival plot. A stacked bar plot showing patient count in each group stratified by cell-of-origin (COO) subtype is displayed below the risk table.

Figure 3: Meta-analysis studying the relationship between the SHM subtypes and overall survival in DLBCL.

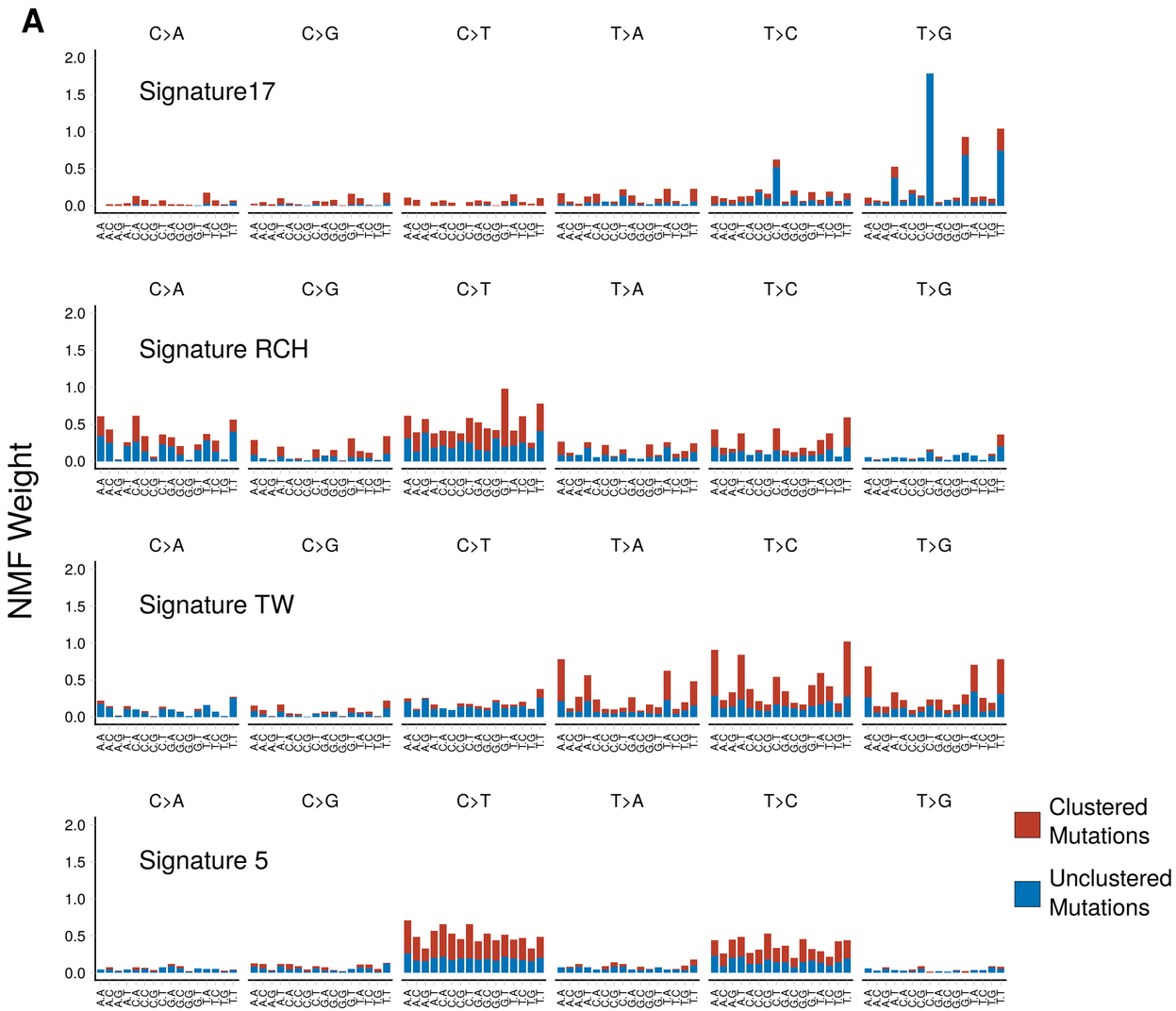
(A) Kaplan-Meier plots and risk tables comparing overall survival between the SHM subtypes in 1,642 patients treated with R-CHOP. (B) Forest plot showing hazard ratios, their confidence intervals and p-values according to multivariate Cox proportional hazards ratios models studying overall survival in relation to the SHM subtypes, the IPI scores and the COO subtypes. (C-E) Kaplan-Meier plots and risk tables comparing overall survival between the SHM subtypes within (C) the GCB subtype, (D) the ABC subtype, and (E) unclassified DLBCL.

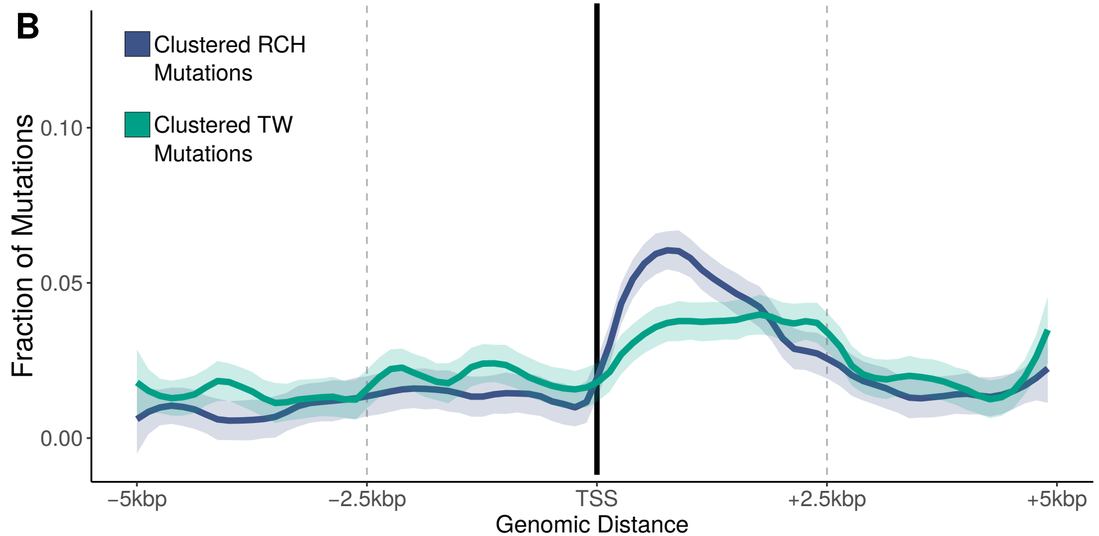
Figure 4: The genetic landscape of the SHM subtypes in DLBCL

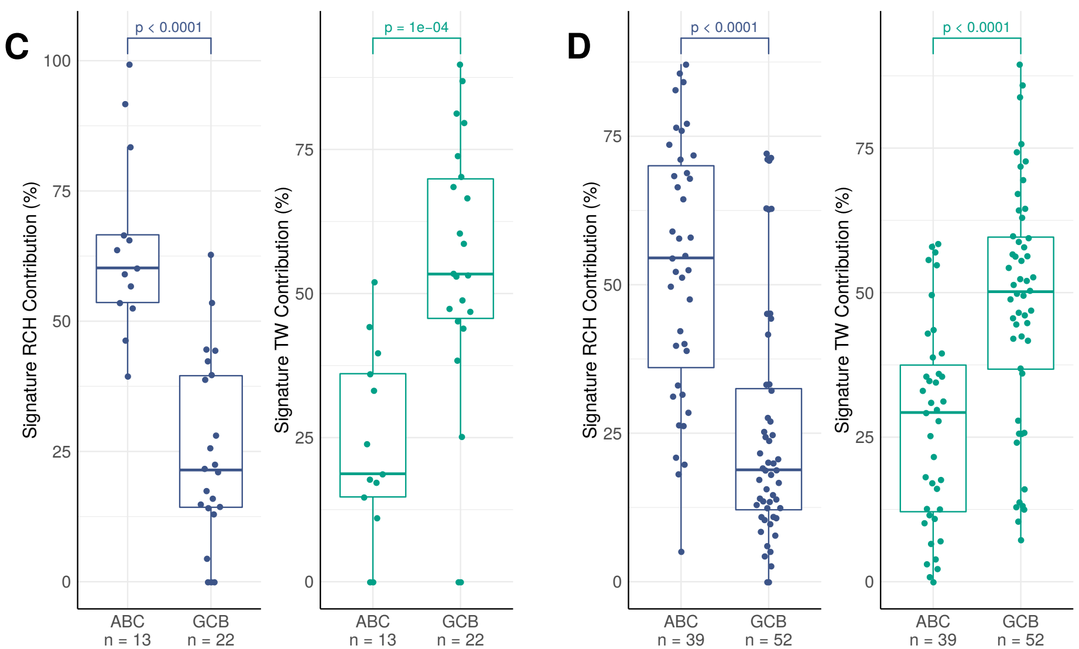
(A) Frequent genetic alterations (protein-changing mutations, amplifications, homozygous deletions and translocations) and their prevalence in each of the SHM subtypes in the Schmitz et al. Cohort. Genes are categorized and ordered according to the SHM subtype of the highest prevalence. Prevalence of genetic alterations and their significant associations are indicated at the right side of the plots. P-values were computed using the Fisher's exact test. (B) Stacked barplot showing sample counts of the SHM subtypes in Schmitz et al. cohort categorized by the genetic subtypes (color). (C-D) Kaplan-Meier plots showing overall survival between the SHM subtypes within (C) the genetic BN2 subtype and (D) genetically unclassified DLBCL.

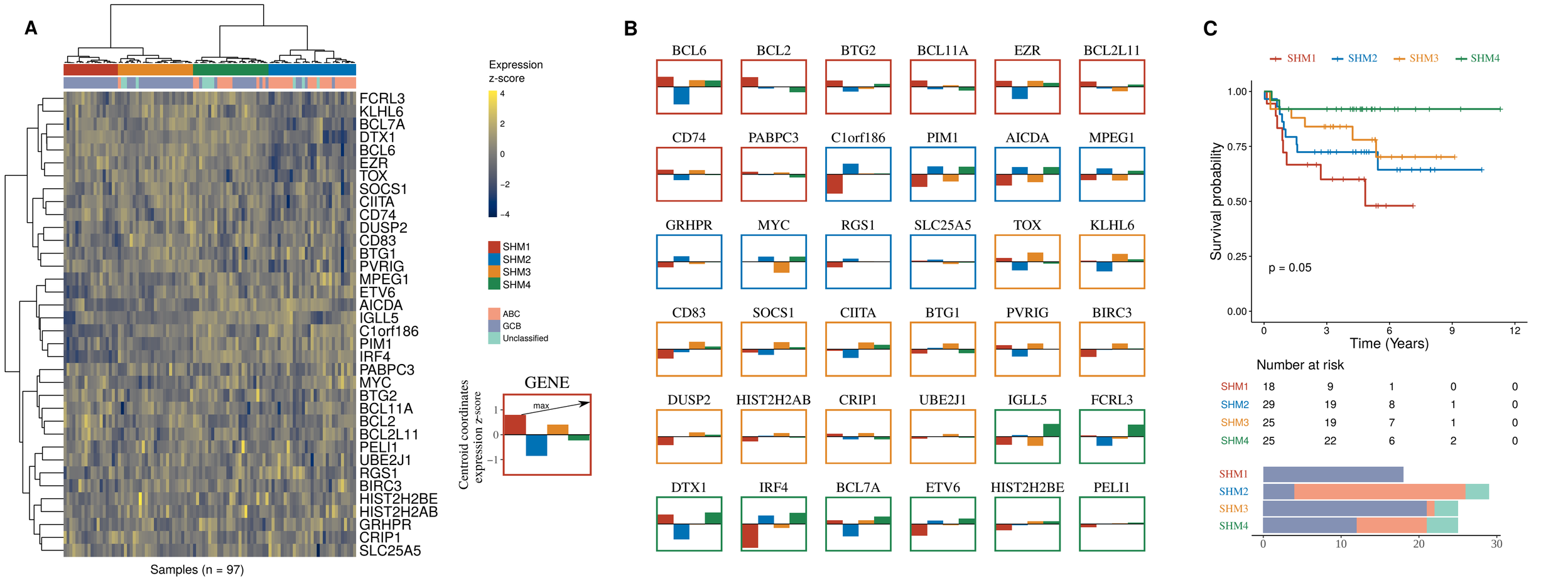
Figure 5: Patterns of somatic hypermutation in the SHM subtypes

(A) Average mutations (mutation/sample) were computed and smoothed for each SHM subtype at 25 bins in the genomic regions spanning 2.5Kbp downstream of transcription start sites (TSS) of the genes that define the SHM subtypes. (B) Prevalence of mutations in the SHM target regions (TSS to 2.5Kbp downstream TSS) in each of the SHM subtypes. Panels (A-B) were produced using the Arthur et al. cohort. (C) Bar plot showing gene importance in SHM subtyping. The importance of a certain gene is defined here as the fraction of samples that change SHM subtype classification when that gene is omitted. The importance scores were quantified using 703 R-CHOP treated DLBCLs in Lenz et al. and Visco et al. cohorts. Three genes without uniquely mapping microarray probes were omitted.

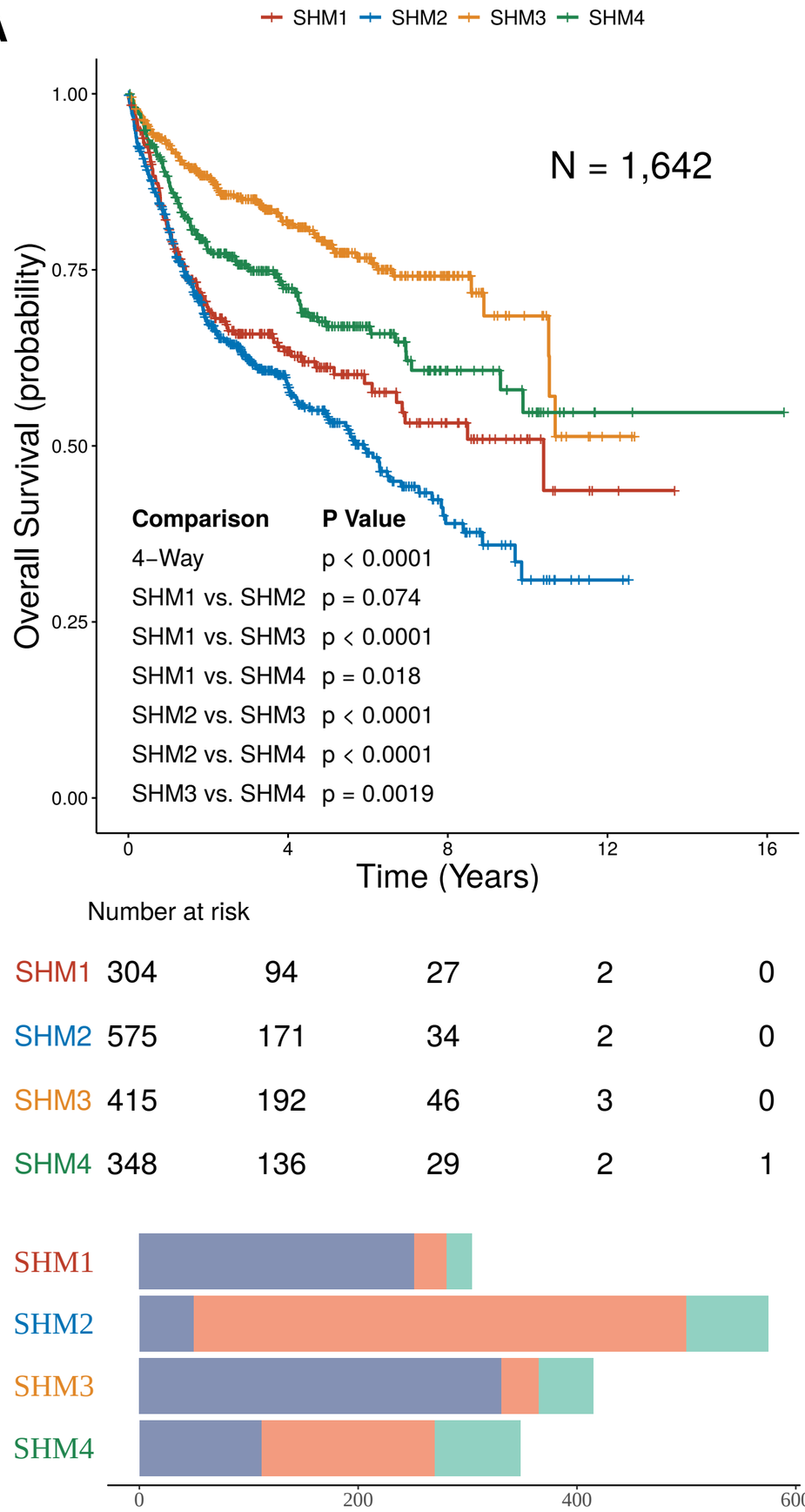

A
NMF Weight
C>A
C>G
C>T
T>A
T>C
T>G
Signature17
Signature RCH
Signature TW
Signature 5
Clustered Mutations
Unclustered Mutations


B
Clustered RCH Mutations
Clustered TW Mutations
Fraction of Mutations
0.10
0.05
0.00
−5kbp
−2.5kbp
TSS
+2.5kbp
+5kbp
Genomic Distance

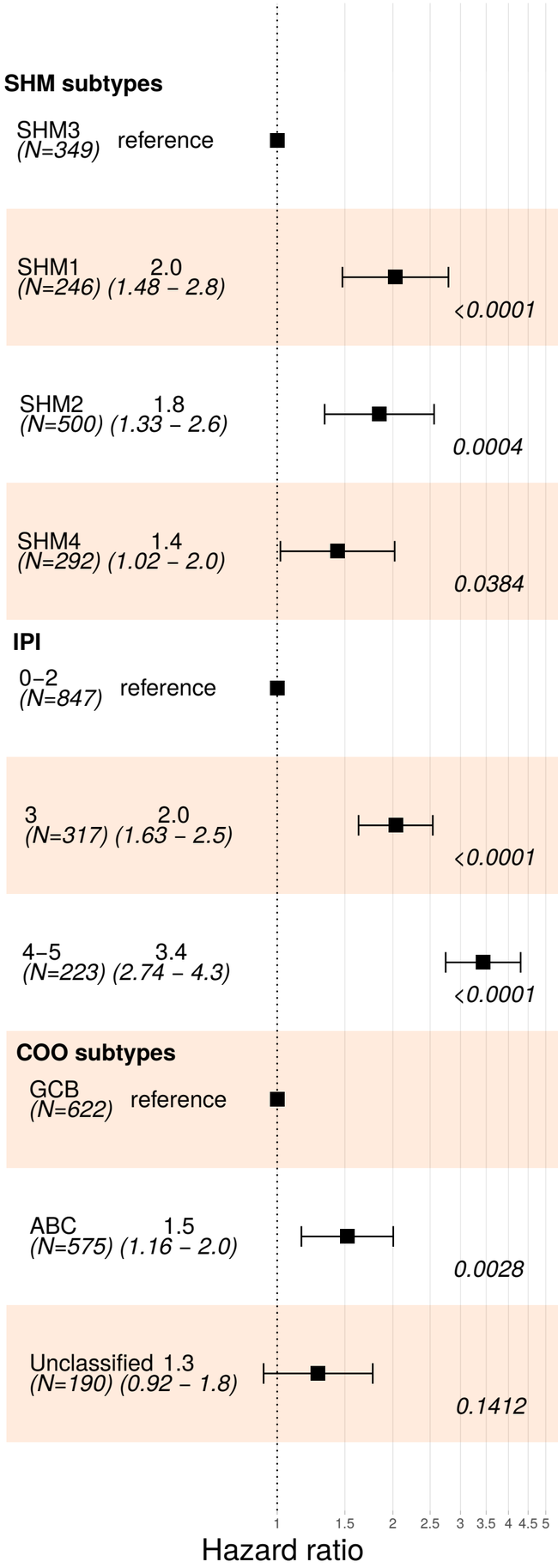

C
Signature RCH Contribution (%)
p < 0.0001
ABC n = 13
GCB n = 22
Signature TW Contribution (%)
p = 1e−04
ABC n = 13
GCB n = 22
D
Signature RCH Contribution (%)
p < 0.0001
ABC n = 39
GCB n = 52
Signature TW Contribution (%)
p < 0.0001
ABC n = 39
GCB n = 52


A
FCRL3
KLHL6
BCL7A
DTX1
BCL6
EZR
TOX
SOCS1
CIITA
CD74
DUSP2
CD83
BTG1
PVRIG
MPEG1
ETV6
AICDA
IGLL5
C1orf186
PIM1
IRF4
PABPC3
MYC
BTG2
BCL11A
BCL2
BCL2L11
PELI1
UBE2J1
RGS1
BIRC3
HIST2H2BE
HIST2H2AB
GRHPR
CRIP1
SLC25A5
Samples (n = 97)
Expression
z-score
4
2
0
−2
−4
SHM1
SHM2
SHM3
SHM4
ABC
GCB
Unclassified
GENE
max
Centroid coordinates
expression z-score
1
0
−1
B
BCL6
BCL2
BTG2
BCL11A
EZR
BCL2L11
CD74
PABPC3
C1orf186
PIM1
AICDA
MPEG1
GRHPR
MYC
RGS1
SLC25A5
TOX
KLHL6
CD83
SOCS1
CIITA
BTG1
PVRIG
BIRC3
DUSP2
HIST2H2AB
CRIP1
UBE2J1
IGLL5
FCRL3
DTX1
IRF4
BCL7A
ETV6
HIST2H2BE
PELI1
C
SHM1
SHM2
SHM3
SHM4
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.05
0
3
6
9
12
Time (Years)
Number at risk
SHM1 18 9 1 0 0
SHM2 29 19 8 1 0
SHM3 25 19 7 1 0
SHM4 25 22 6 2 0
SHM1
SHM2
SHM3
SHM4
0
10
20
30

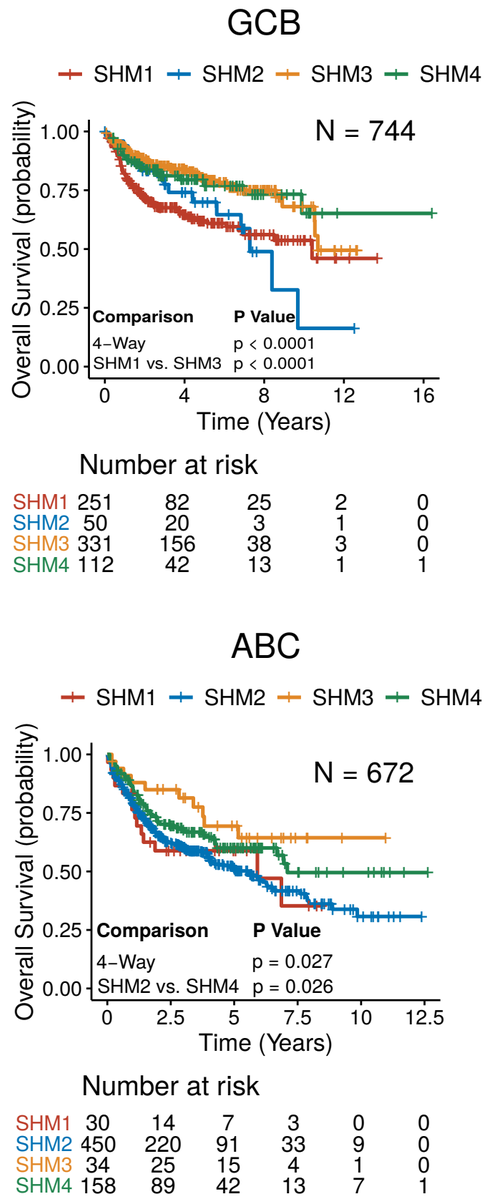
A


SHM1 SHM2 SHM3 SHM4
N = 1,642
Overall Survival (probability)
1.00
0.75
0.50
0.25
0.00
Comparison P Value
4−Way p < 0.0001
SHM1 vs. SHM2 p = 0.074
SHM1 vs. SHM3 p < 0.0001
SHM1 vs. SHM4 p = 0.018
SHM2 vs. SHM3 p < 0.0001
SHM2 vs. SHM4 p < 0.0001
SHM3 vs. SHM4 p = 0.0019
0 4 8 12 16
Time (Years)
Number at risk
SHM1 304 94 27 2 0
SHM2 575 171 34 2 0
SHM3 415 192 46 3 0
SHM4 348 136 29 2 1
SHM1
SHM2
SHM3
SHM4
0 200 400 600

B


SHM subtypes
SHM3 (N=349) reference
SHM1 2.0 (N=246) (1.48 – 2.8) <0.0001
SHM2 1.8 (N=500) (1.33 – 2.6) 0.0004
SHM4 1.4 (N=292) (1.02 – 2.0) 0.0384
IPI
0–2 (N=847) reference
3 2.0 (N=317) (1.63 – 2.5) <0.0001
4–5 3.4 (N=223) (2.74 – 4.3) <0.0001
COO subtypes
GCB (N=622) reference
ABC 1.5 (N=575) (1.16 – 2.0) 0.0028
Unclassified 1.3 (N=190) (0.92 – 1.8) 0.1412
1 1.5 2 2.5 3 3.5 4 4.5 5
Hazard ratio

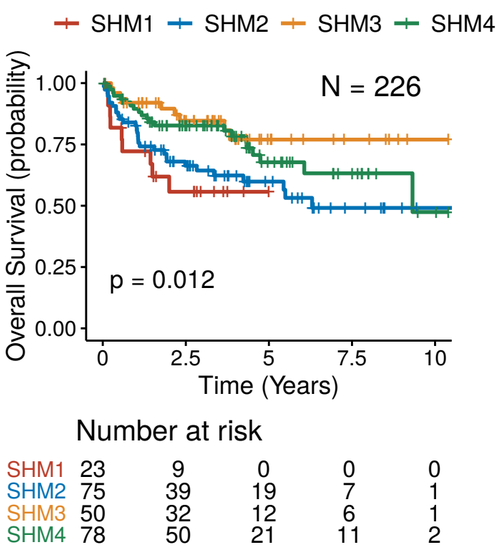
C


GCB
SHM1 SHM2 SHM3 SHM4
N = 744
Overall Survival (probability)
1.00
0.75
0.50
0.25
0.00
Comparison P Value
4−Way p < 0.0001
SHM1 vs. SHM3 p < 0.0001
0 4 8 12 16
Time (Years)
Number at risk
SHM1 251 82 25 2 0
SHM2 50 20 3 1 0
SHM3 331 156 38 3 0
SHM4 112 42 13 1 1

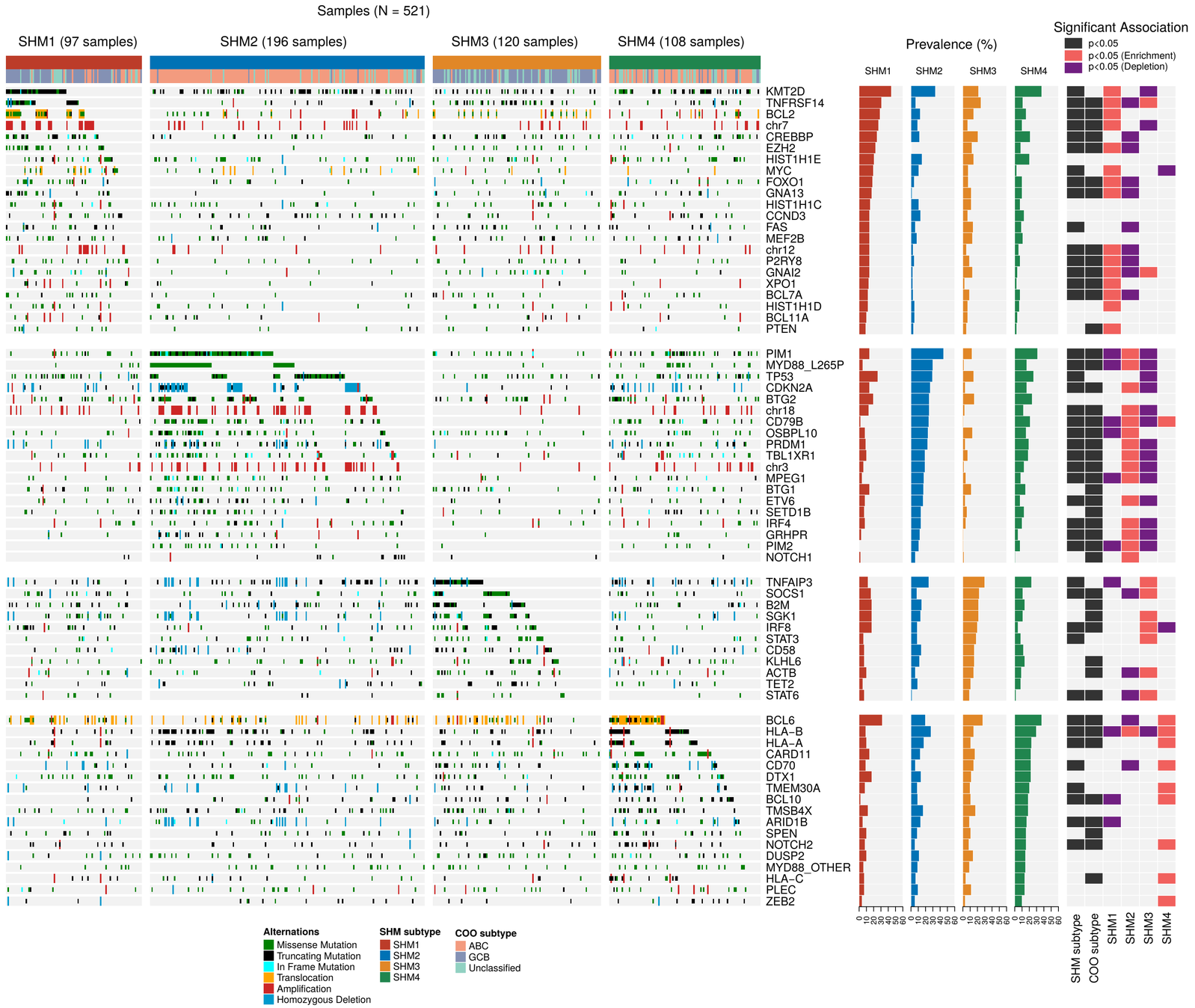
D

ABC
SHM1 SHM2 SHM3 SHM4
N = 672
Overall Survival (probability)
1.00
0.75
0.50
0.25
0.00
Comparison P Value
4−Way p = 0.027
SHM2 vs. SHM4 p = 0.026
0 2.5 5 7.5 10 12.5
Time (Years)
Number at risk
SHM1 30 14 7 3 0 0
SHM2 450 220 91 33 9 0
SHM3 34 25 15 4 1 0
SHM4 158 89 42 13 7 1

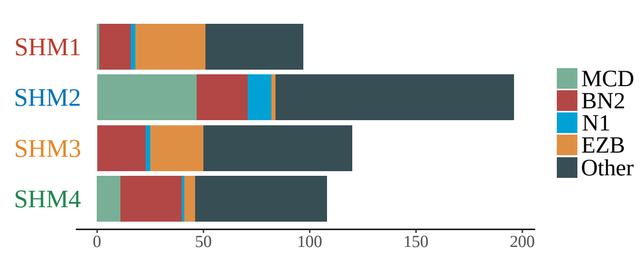
E


Unclassified
SHM1 SHM2 SHM3 SHM4
N = 226
Overall Survival (probability)
1.00
0.75
0.50
0.25
0.00
p = 0.012
0 2.5 5 7.5 10
Time (Years)
Number at risk
SHM1 23 9 0 0 0
SHM2 75 39 19 7 1
SHM3 50 32 12 6 1
SHM4 78 50 21 11 2

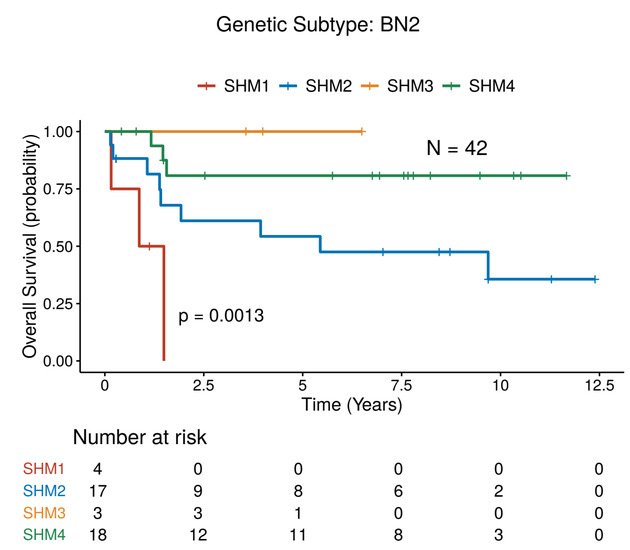
A



B



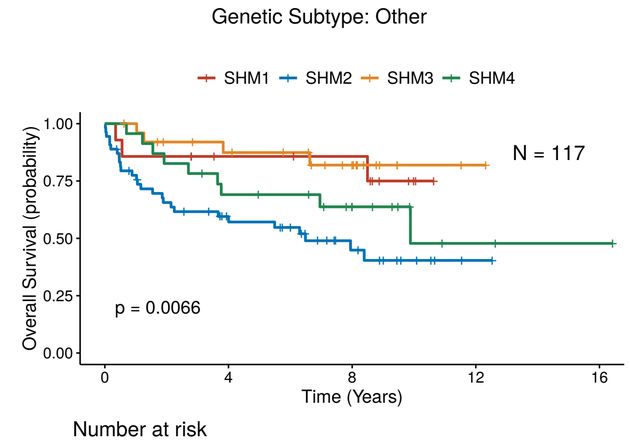
C



Number at risk

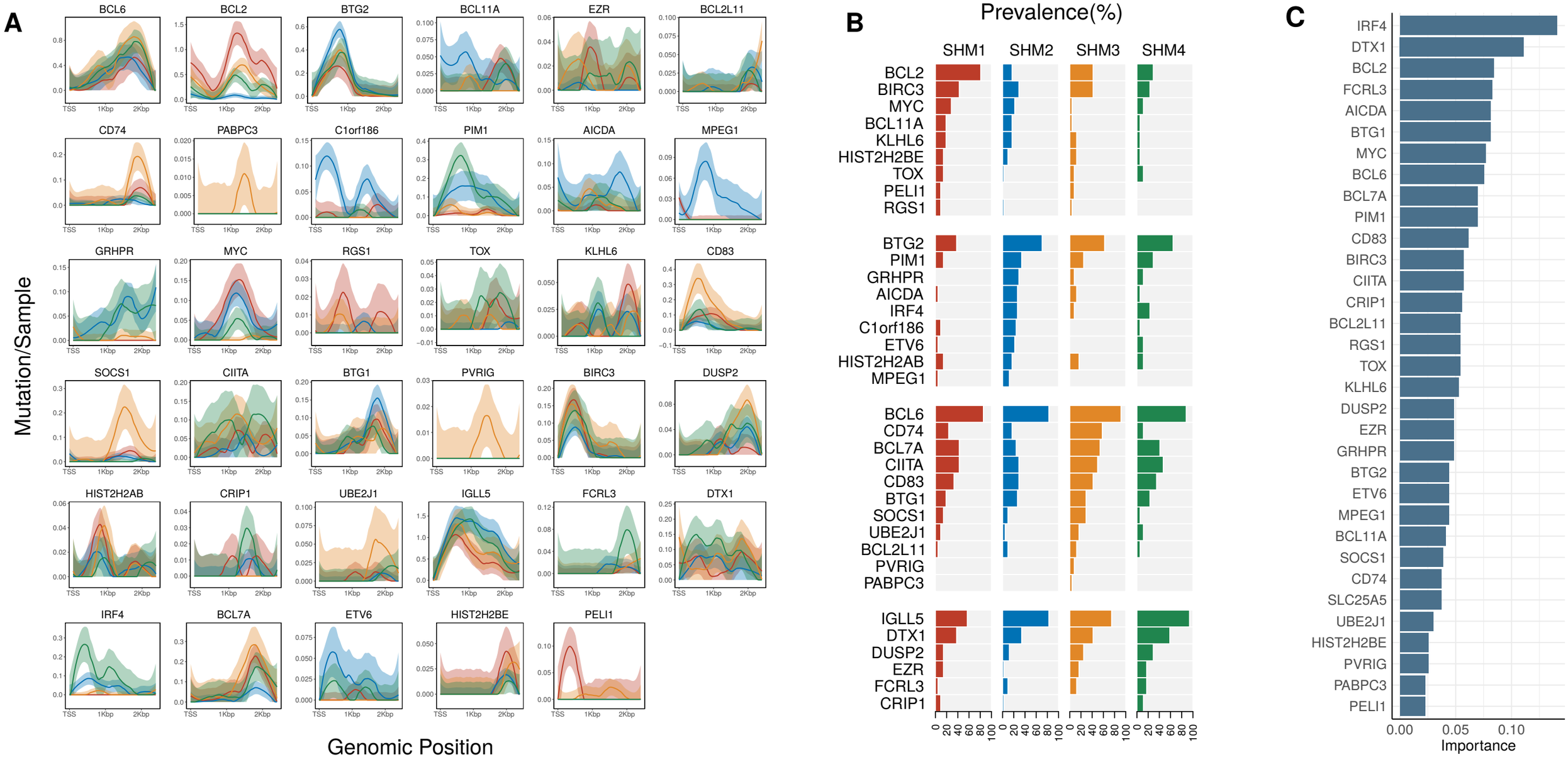
SHM1	4	0	0	0	0	0
SHM2	17	9	8	6	2	0
SHM3	3	3	1	0	0	0
SHM4	18	12	11	8	3	0

D



Number at risk

SHM1	14	9	8	0	0
SHM2	54	25	11	1	0
SHM3	26	19	11	1	0
SHM4	23	15	8	2	1


A
Mutation/Sample
Genomic Position
BCL6
BCL2
BTG2
BCL11A
EZR
BCL2L11
CD74
PABPC3
C1orf186
PIM1
AICDA
MPEG1
GRHPR
MYC
RGS1
TOX
KLHL6
CD83
SOCS1
CIITA
BTG1
PVRIG
BIRC3
DUSP2
HIST2H2AB
CRIP1
UBE2J1
IGLL5
FCRL3
DTX1
IRF4
BCL7A
ETV6
HIST2H2BE
PELI1
TSS
1Kbp
2Kbp
B
Prevalence(%)
SHM1
SHM2
SHM3
SHM4
BCL2
BIRC3
MYC
BCL11A
KLHL6
HIST2H2BE
TOX
PELI1
RGS1
BTG2
PIM1
GRHPR
AICDA
IRF4
C1orf186
ETV6
HIST2H2AB
MPEG1
BCL6
CD74
BCL7A
CIITA
CD83
BTG1
SOCS1
UBE2J1
BCL2L11
PVRIG
PABPC3
IGLL5
DTX1
DUSP2
EZR
FCRL3
CRIP1
0
20
40
60
80
100
C
IRF4
DTX1
BCL2
FCRL3
AICDA
BTG1
MYC
BCL6
BCL7A
PIM1
CD83
BIRC3
CIITA
CRIP1
BCL2L11
RGS1
TOX
KLHL6
DUSP2
EZR
GRHPR
BTG2
ETV6
MPEG1
BCL11A
SOCS1
CD74
SLC25A5
UBE2J1
HIST2H2BE
PVRIG
PABPC3
PELI1
0.00
0.05
0.10
Importance