

Additional file 1 for

Immunoglobulin somatic hypermutation has clinical impact in DLBCL and potential implications for immune checkpoint blockade and neoantigen-based immunotherapies

Supplementary figure Legends

Figure S1. Construction and clinical outcome of the diffuse large B-cell

lymphoma (DLBCL) cohort. (a) CONSORT flow diagram illustrating the construction of the training set and validation set comprising the overall study cohort. (b) Overall survival (OS) and progression-free survival (PFS) rates of the overall patients included in this study, and OS/ PFS comparison between germinal center B-cell-like (GCB) and activated B-cell-like (ABC) subtypes of DLBCL.

Figure S2. Diagram showing the numbers of cases in this mutation study that have been characterized by various biomarker studies, and survival rates of patients whose sequencing results were correlated with prognosis.

Figure S3. CONSORT flow diagram illustrating the number of cases performed for high-throughput IG sequencing and clonal sequence analysis. In the Venn diagram, green number is for cases with only heavy chain analyzed, red number is for cases with only light chain analyzed, and black number is for cases with both heavy and light chain analyzed. In the flowchart, green boxes are for heavy chain analysis, and red boxes are for light chain analysis.

Figure S4. Molecular characterization for immunoglobulin heavy chain

(IGH) gene usage in the study cohort. (a) Stacked histograms for IGHD gene usage in cases with D-J-only (left) and cases with V-D-J (right) diagnostic sequences. (b-c) Stacked histograms showing the specific IGHV gene usage pattern of diagnostic clones in overall DLBCL and

germinal-center B-cell-like (GCB) and activated B-cell-like (ABC) DLBCL subtypes. IGHV4-34 is over-represented in the ABC-DLBCL cases compared with the GCB-DLBCL cases.

Histogram showing the distribution of somatic mutations by V-gene family as measured by total reads in GCB-DLBCL and ABC-DLBCL.

Figure S5. Immunoglobulin heavy chain V gene (IGHV) somatic

hypermutation (SHM) analysis. (a) Histogram showing the distribution of IGHV somatic mutations within productive diagnostic clonal sequences. The number above the bar chart is the absolute count of the total clones. GCB-DLBCL compared with ABC-DLBCL had a significantly higher mean level of IGHV SHM. (b) High degree of IGHV SHM (SHM^{high}) was associated with a nonsignificant trend of better progression-free survival (PFS) in overall DLBCL. (c) IGHV SHM^{high} was associated with significantly better overall survival (OS) in ABC-DLBCL. (d) In GCB-DLBCL, IGHV SHM^{high} was associated with better OS with borderline significance only in cases lacking *BCL2* rearrangement (*BCL2*-R) or *MYC* rearrangement (*MYC*-R) but not in cases with *BCL2*-R or *MYC*-R.

Figure S6. Analysis for length of heavy chain complementarity

determining region 3 (HCDR3). (a) Long length of HCDR3 with a cutoff of 18 amino acids (aa) was associated with poorer progression-free survival (PFS) in DLBCL and ABC-DLBCL, and with poorer PFS and overall survival (OS) in GCB-DLBCL. (b) Short HCDR3 length was associated with significantly higher IGHV SHM degree in GCB-DLBCL, and higher IGK/L SHM degree in ABC-DLBCL. The cutoff of short length was < 16 aa in GCB-DLBCL and <18 aa in ABC-DLBCL. (c) Long HCDR3 length was associated with poorer OS in GCB-DLBCL cases of the validation set and ABC-DLBCL cases of the training set. (d) Long HCDR3 length was associated with significantly poorer PFS and OS in DLBCL in both the training and validation sets.

Figure S7. Prediction of MHC-binding peptides and frequency of T-cell exposed motifs (TCEM) for immunoglobulin diagnostic sequences in the training set and validation set. (a) Regional distribution of relatively rare neoantigens (TCEM FC>16) derived from heavy chain and light chain immunoglobulin genes in DLBCL patients in the training set (top) and validation set (bottom). Each dot represents one peptide predicted having high MHC-II-binding affinity (exceeding the -1 standard deviation threshold for MHC derived from 24 HLA-DR alleles) and relatively rare TCEM (FC>16). The color intensities of the dots are scaled on the FC scale from FC16 to the very rare FC24. (b) Cases with high degree of heavy chain or light chain IGV SHM compared with cases without had higher frequency of relatively rare TCEM (FC>16) in the training (left) and validation sets (right). (c) In ABC-DLBCL, high IGV SHM was associated with lower tissue cellularity of CD4⁺ T cells.

Figure S8. Molecular analysis for immunoglobulin heavy chain ongoing

SHM and light chain SHM. (a) Histogram showing distribution of IGHV ongoing SHM over IGHV gene families. (b) High IGHV ongoing SHM was associated with *AICDA* upregulation in the validation set. High PD-L2 expression in macrophages (in overall cohort) and *PD-L1/PD-L2* gene amplification (in the ABC subtype) were associated with a higher mean percentage of subclones with IGHV ongoing SHM in the sequence repertoire. (c) Histogram showing distribution of IGK/LV SHM within light chain diagnostic clonal sequences. (d) Compared with IGKV SHM, IGLV SHM showed more correlation with IGHV SHM.

Figure S9. Immunoglobulin light chain SHM and CDR3 analysis. (a)

IGK/LV SHM^{high} was associated with significantly worse progression-free survival (PFS) in GCB-DLBCL in both the training and validation sets. (b-c) High degree of IGK/LV SHM (SHM^{high}) was associated with significantly worse overall survival (OS) in GCB-DLBCL independent of *BCL2* and *MYC* rearrangement (R) status. The *P* value was most significant in cases with *BCL2* rearrangement. (d) A long immunoglobulin light chain kappa/lambda CDR3

(K/LCDR3) length was associated with poorer OS in overall DLBCL and ABC-DLBCL. (e) In GCB-DLBCL, IGK/LV SHM^{high} was significantly associated with higher *CTSL1* mRNA levels whereas IGHV SHM^{high} was significantly associated with lower *CTSF* mRNA levels. (f) In ABC-DLBCL, IGK/LV SHM^{high} was negatively associated with PD-1 expression in B cells, and positively associated with *AICDA* mRNA expression in the training set.

Figure S10. Comparison between different subsets of DLBCL. (a) The

training set compared with the validation set had significantly higher mean levels of some *MHC* class II and cathepsin genes' mRNA expression. (b) In the validation set overall (and the GCB subtype, figures not shown), *MYC* rearrangement (*MYC-R*) was associated with downregulation of some *MHC* class II genes. In both the training and validation sets, *MYC-R* was significantly associated with *HLA-F* and *CTSH* downregulation (only figures for the entire cohort were shown). (c) In both the training and validation sets, ABC-DLBCL compared with GCB-DLBCL had significantly higher mean levels of macrophage- and CD8⁺ T cell-infiltration and PD-L1 expression on B cells whereas lower mean *HLA-DQB2* mRNA level (only figures for overall cohort were shown). In the overall cohort (and in the training set but not the validation set), ABC- compared with GCB-DLBCL had significantly higher mean level of *CTSL1* mRNA.

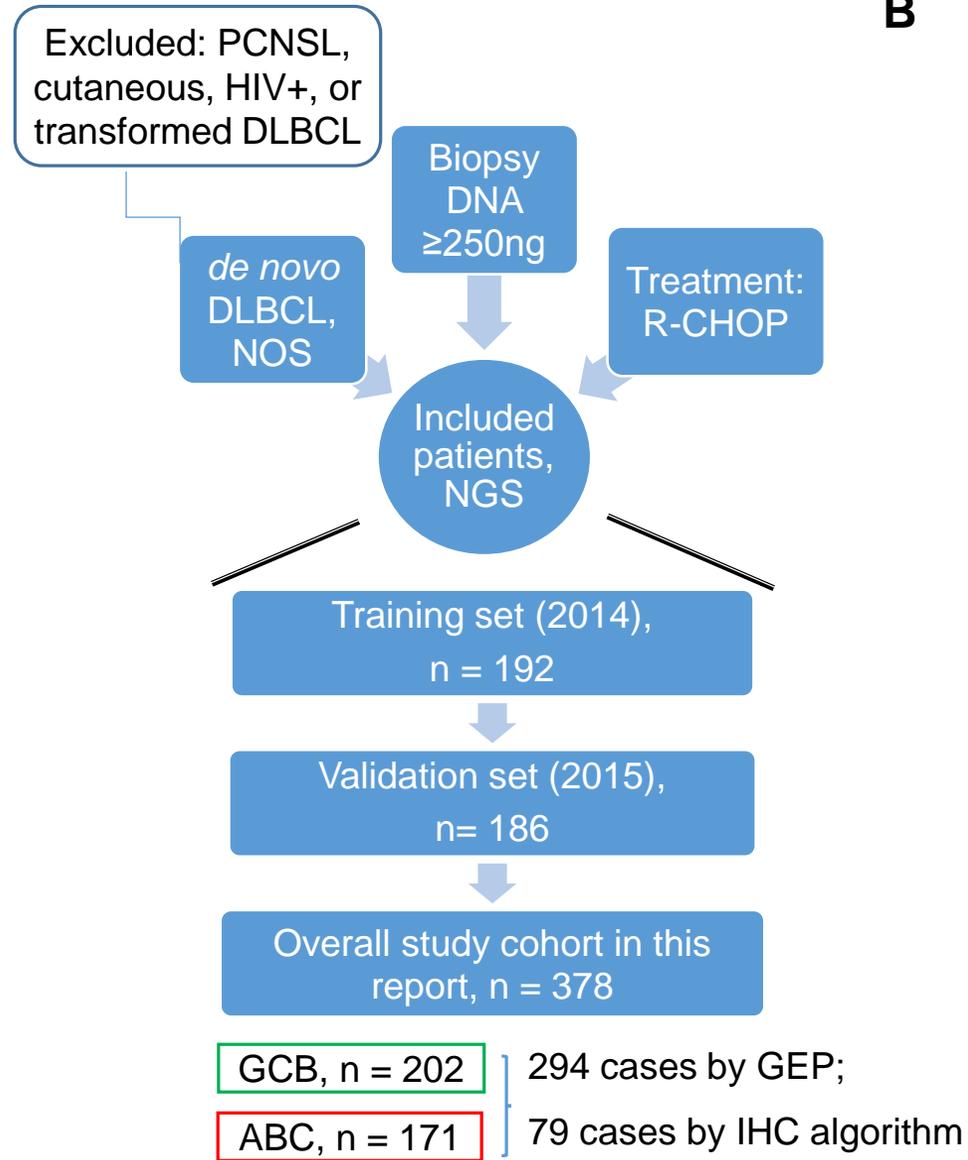
Figure S11. Light chain IGK/LV ongoing SHM analysis. (a) There were

negative associations between IGV SHM and light chain ongoing SHM. Like light chain SHM^{high}, (shown in Figure 5d) IGHV SHM^{high} was associated with lower numbers of subclonal sequences with IGK/LV ongoing SHM. Conversely, high numbers of IGK/LV subclonal sequences with ongoing SHM were associated with lower mean levels of SHM in IGV. (b) Heatmap for gene signatures associated with high ongoing IGK/LV SHM in DLBCL and GCB-DLBCL. (c) In GCB-DLBCL, high IGK/LV ongoing SHM (≥ 17 subclones) was associated with significantly poorer overall survival (OS). The adverse prognostic effect of light chain ongoing SHM was not

significant in the validation set. **(d)** In the training set, high light chain IGV ongoing SHM was associated with significantly poorer survival.

Figure S1

A



B

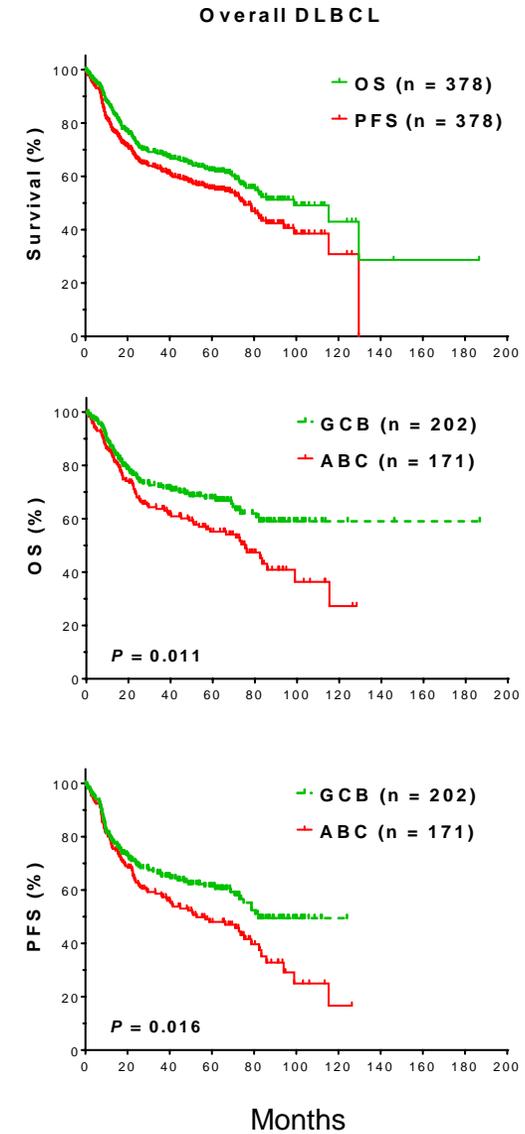


Figure S2

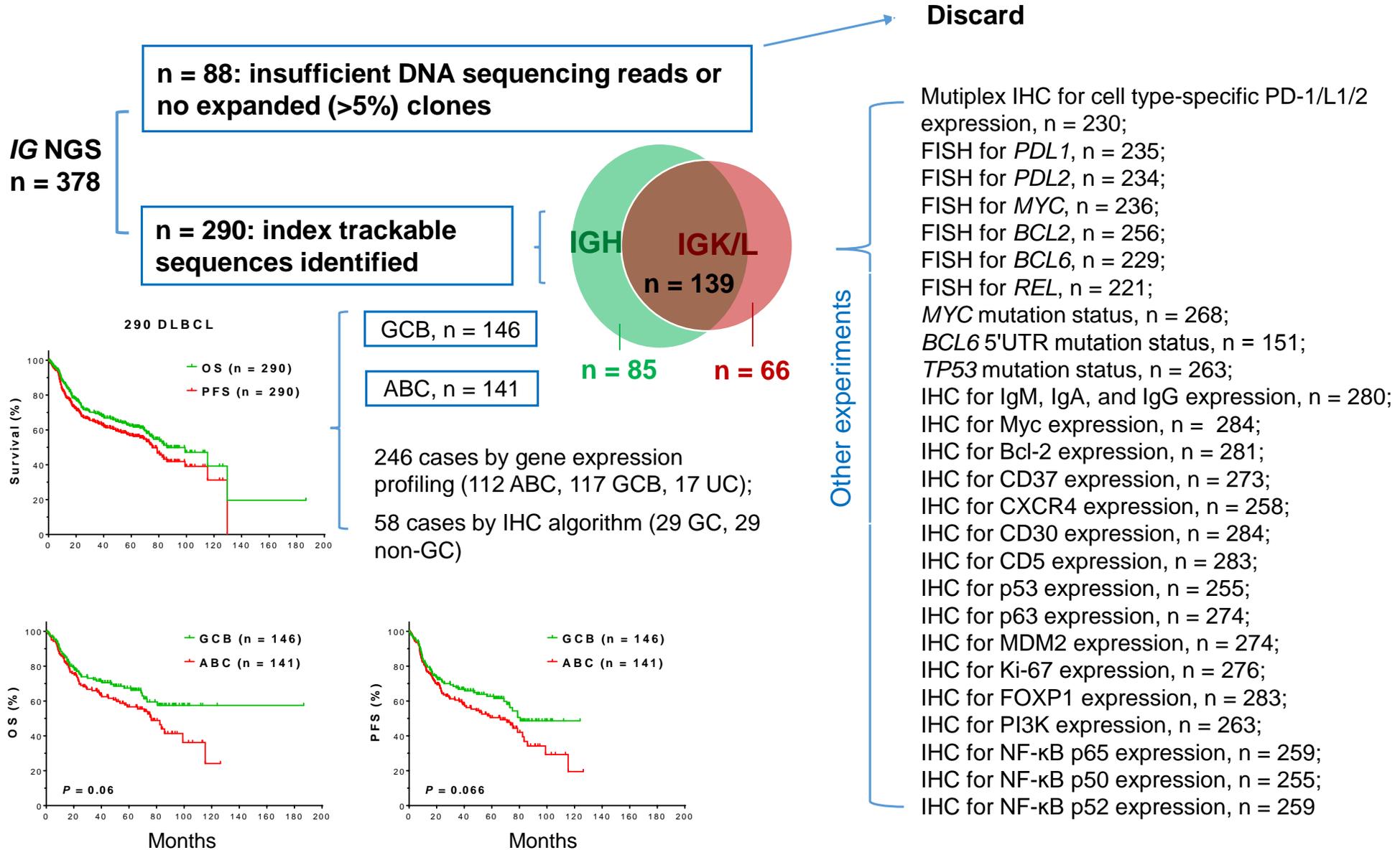


Figure S3

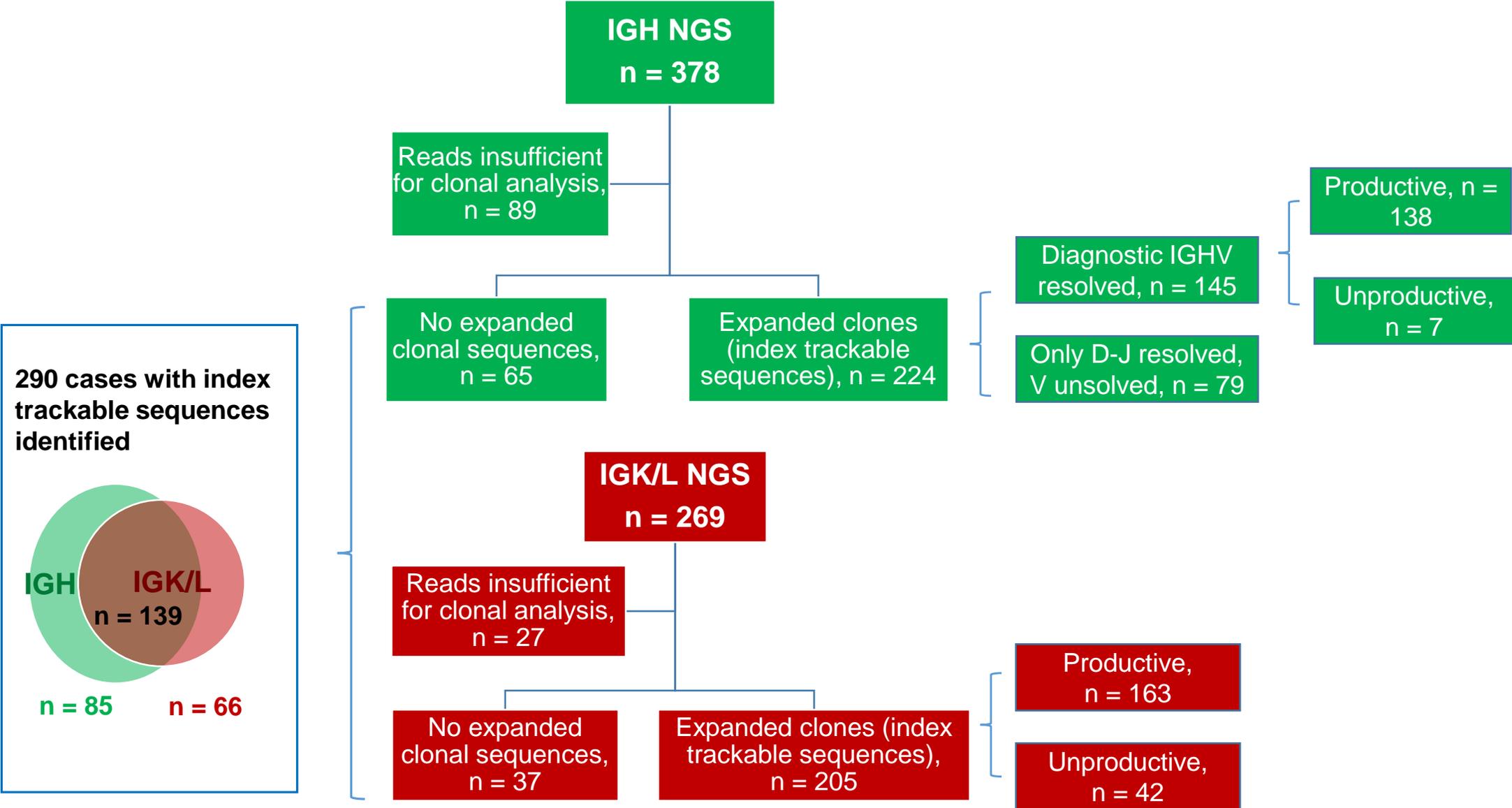
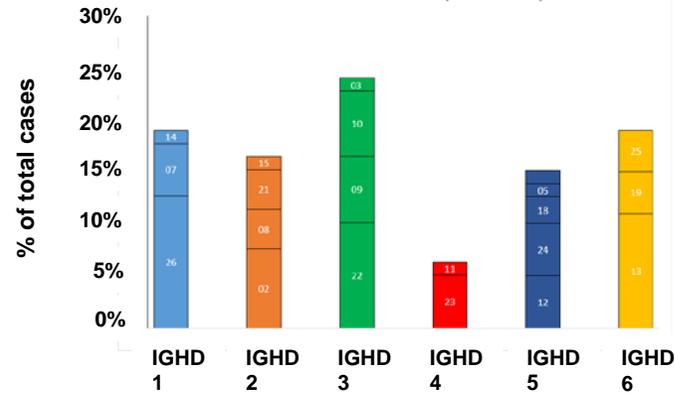
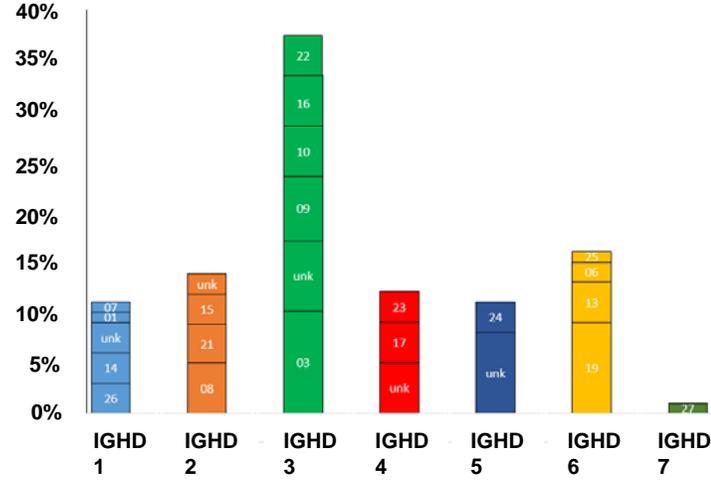


Figure S4

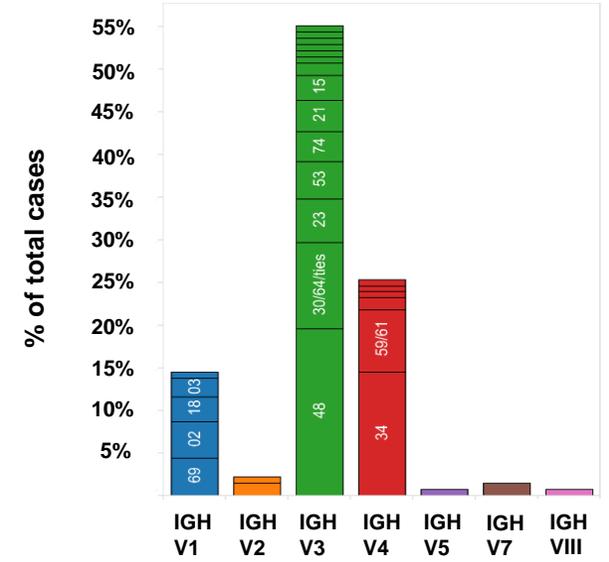
A Cases with only-D-J (V-unresolved) diagnostic sequences



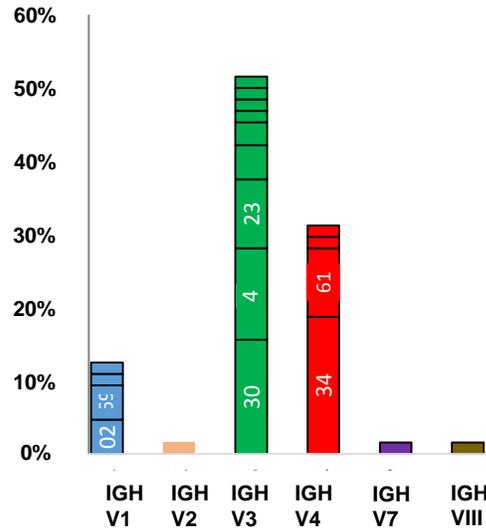
Cases with V-D-J diagnostic sequences



B DLBCL



C ABC-DLBCL



GCB-DLBCL

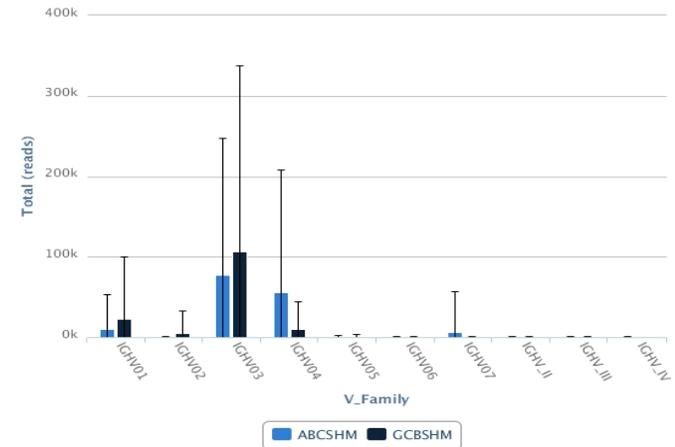
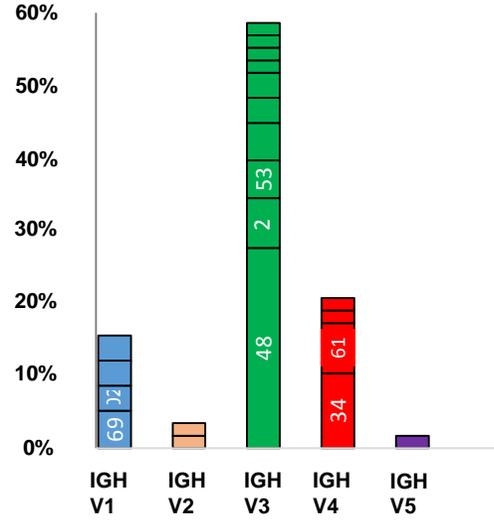


Figure S5

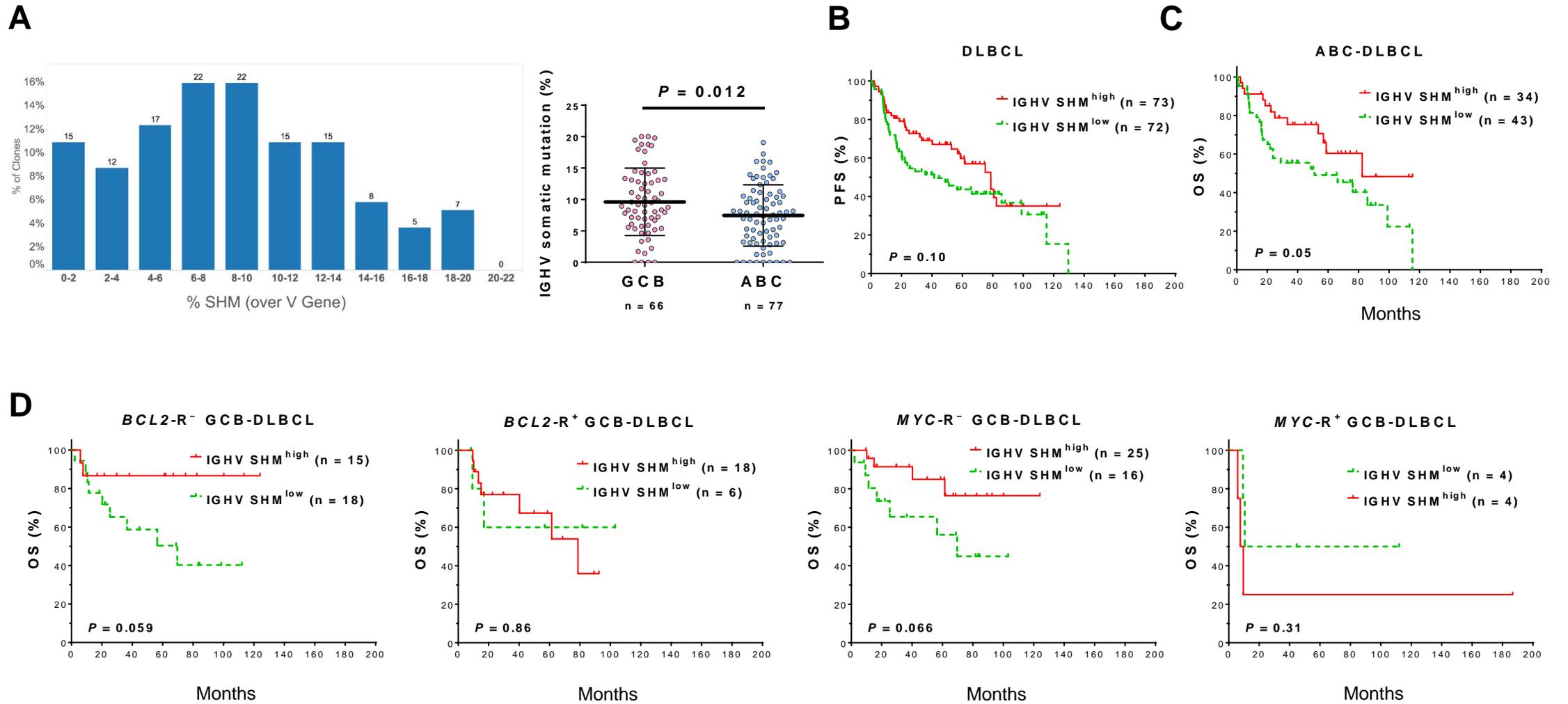


Figure S6

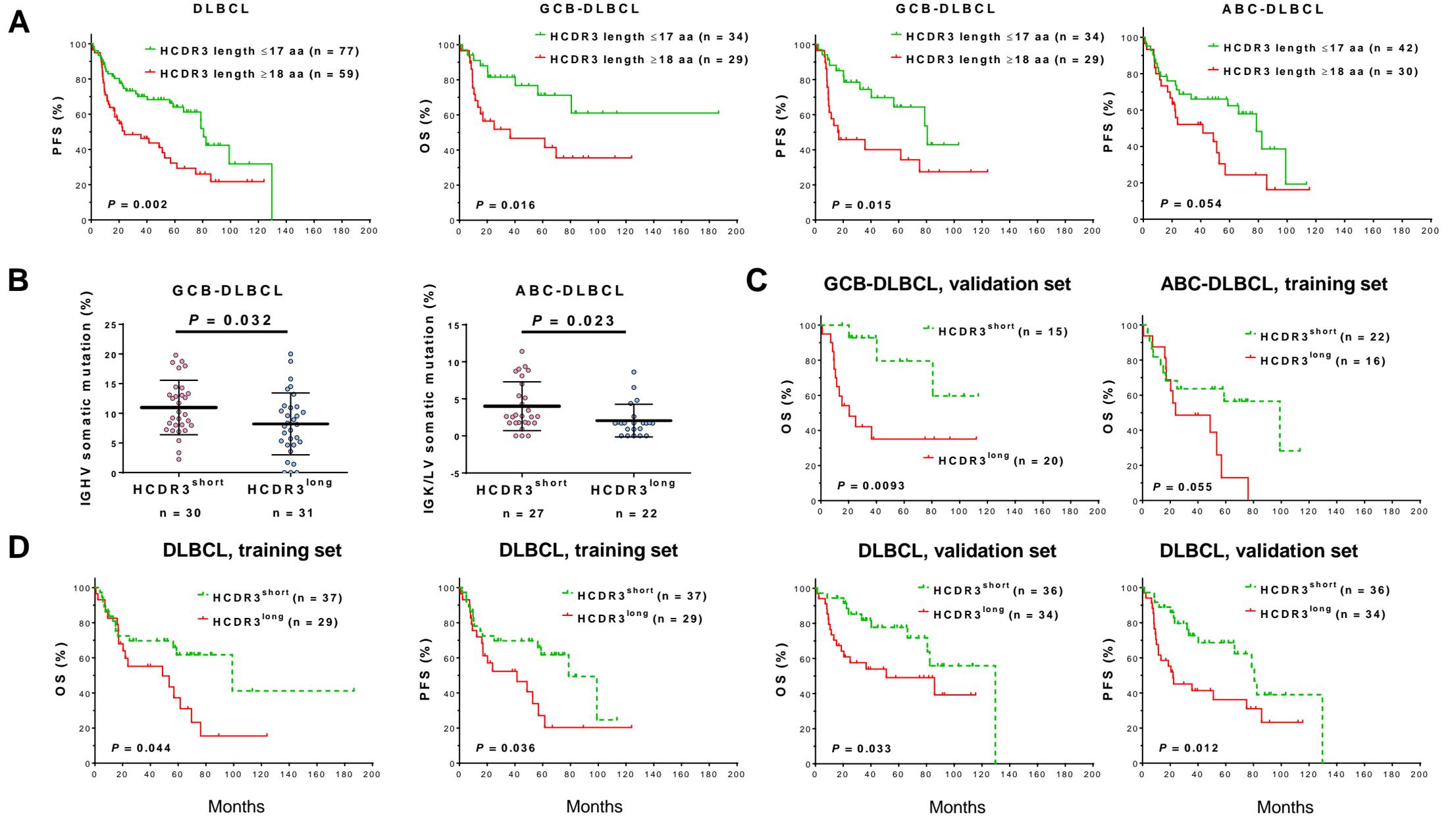


Figure S7

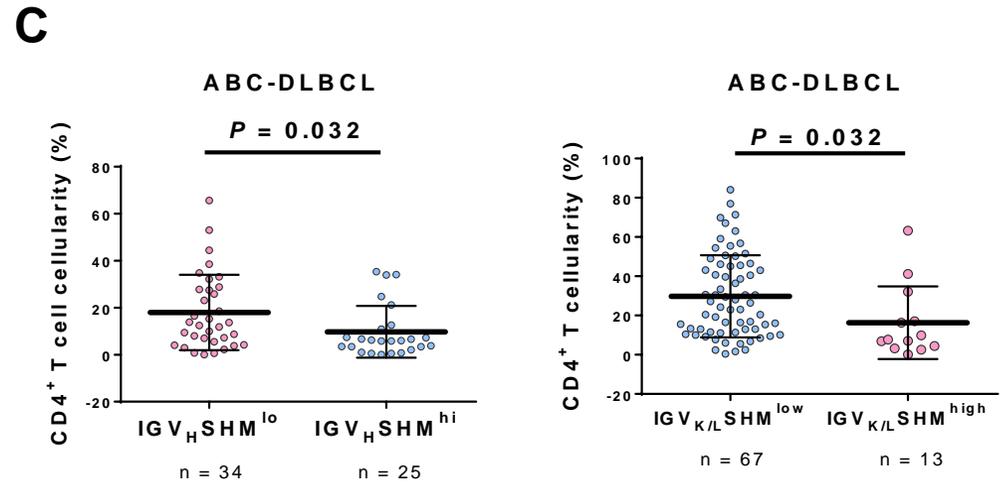
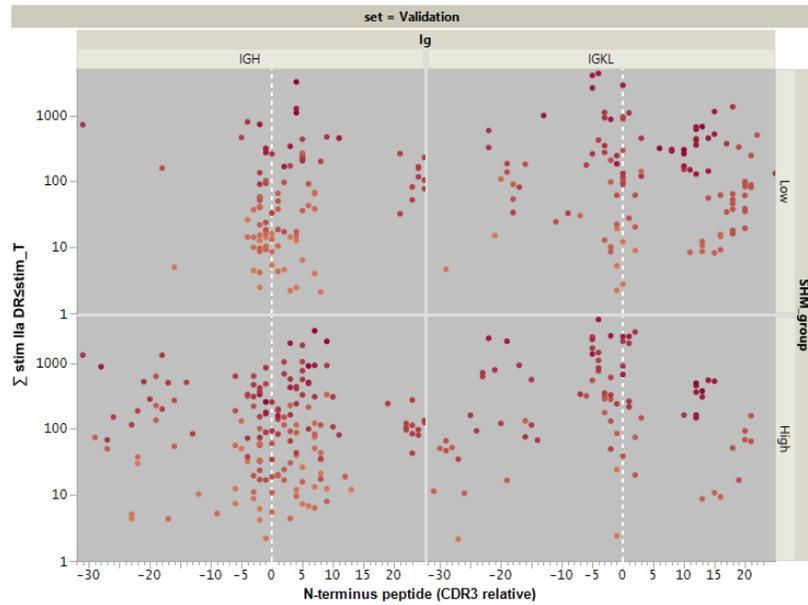
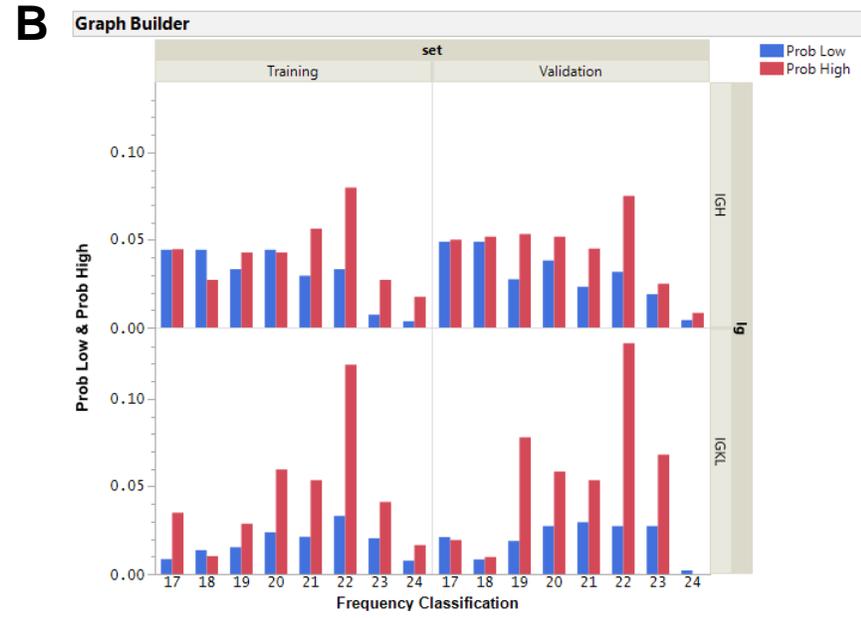
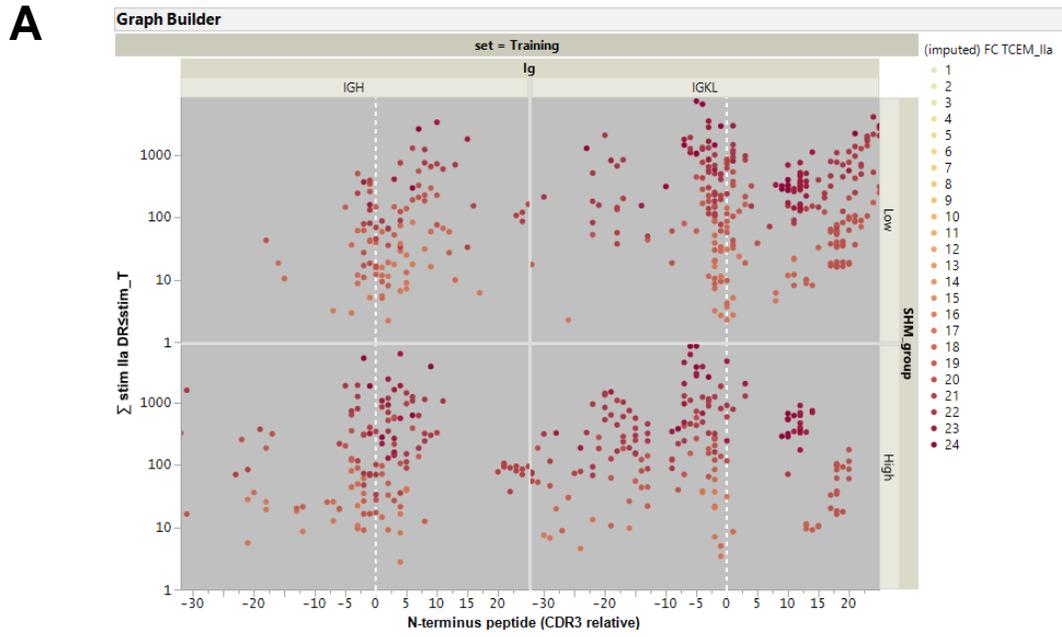
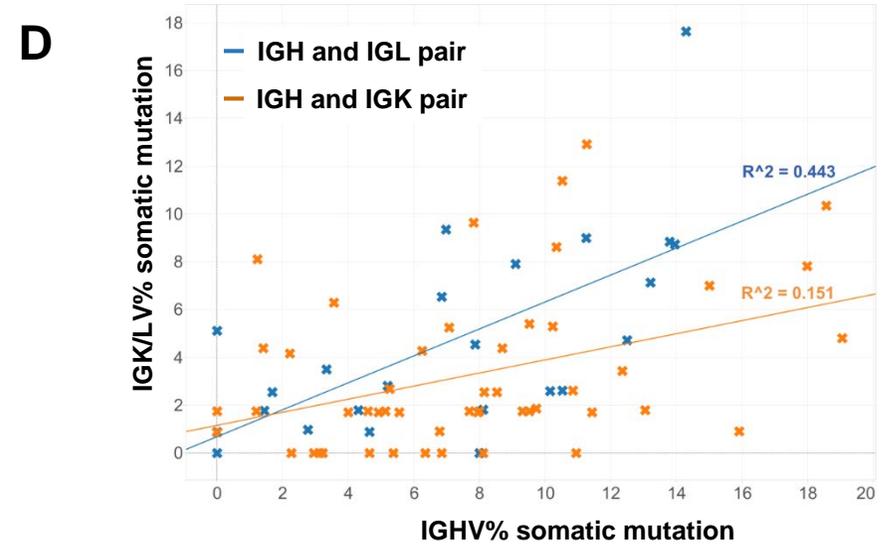
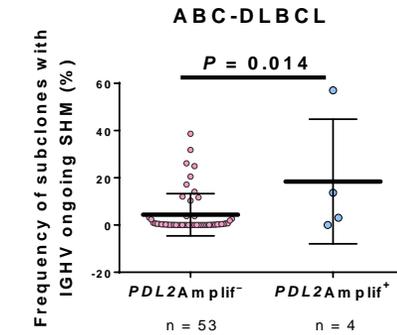
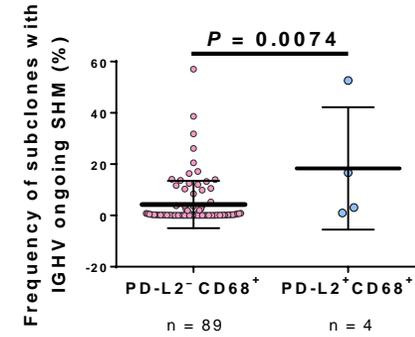
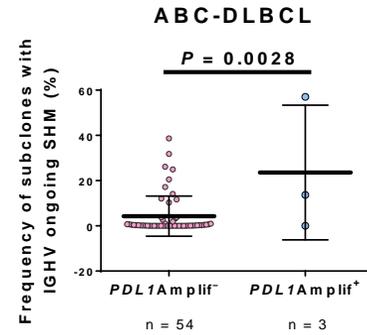
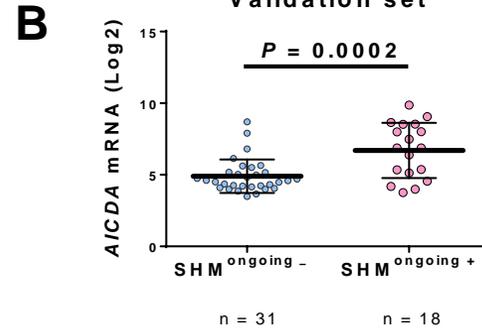
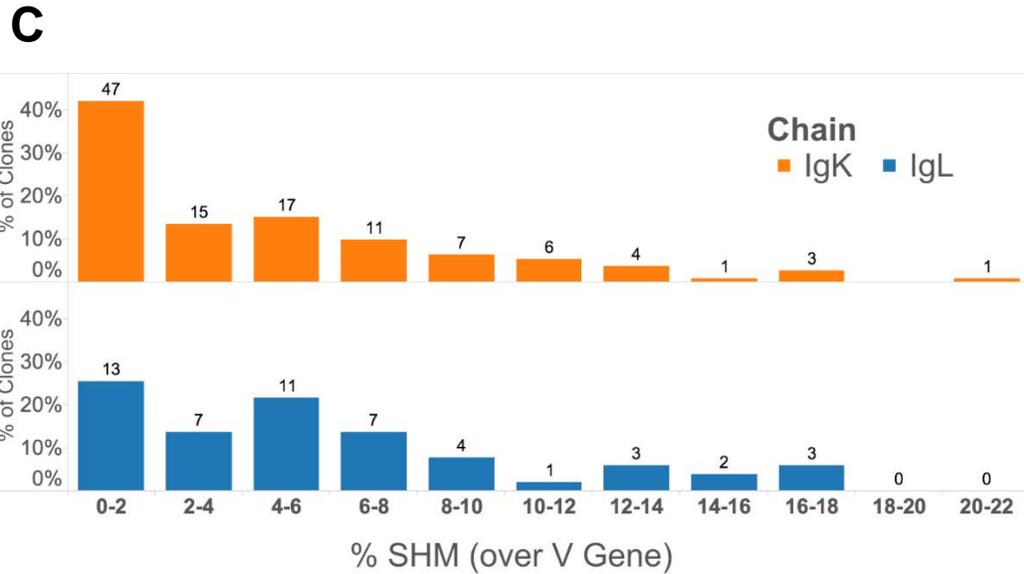
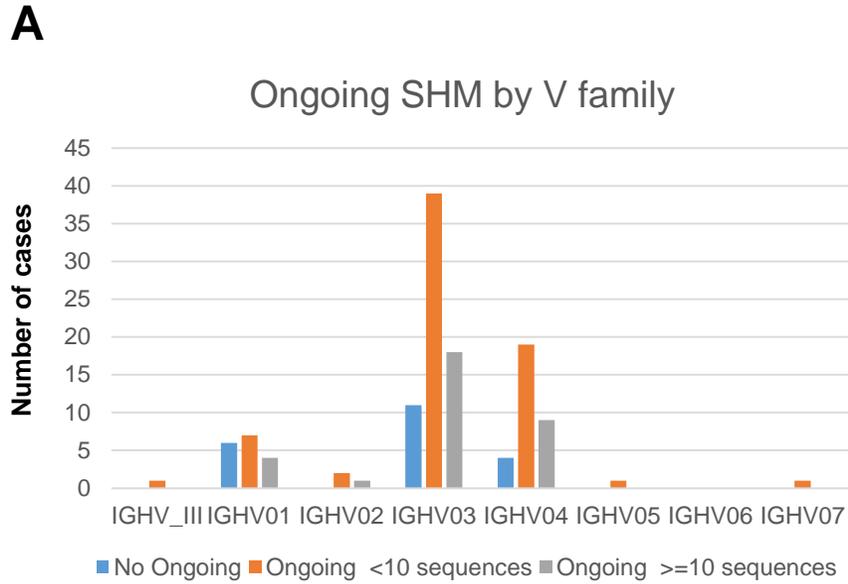
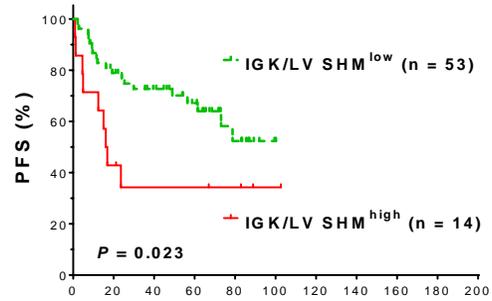


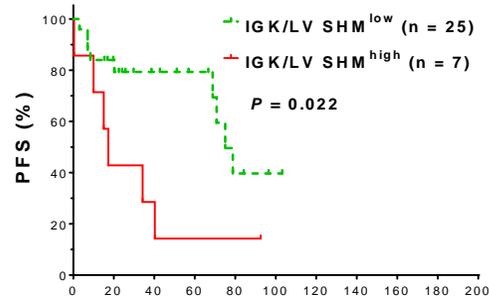
Figure S8



A GCB-DLBCL, training set

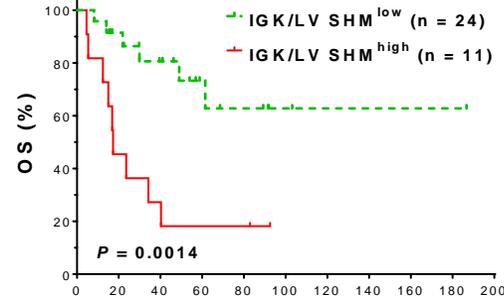


GCB-DLBCL, validation set

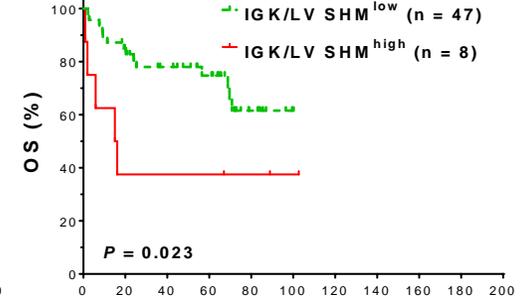


B

BCL2-R⁺ GCB-DLBCL

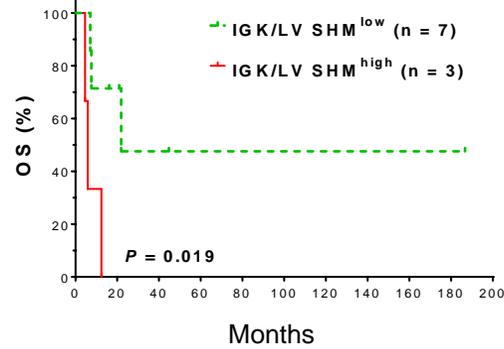


BCL2-R⁻ GCB-DLBCL

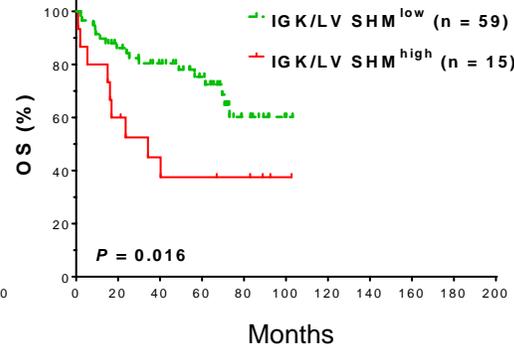


C

MYC-R⁺ GCB-DLBCL

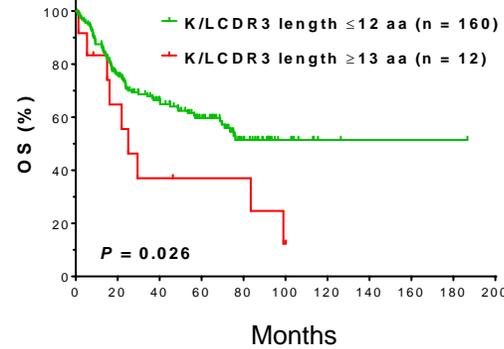


MYC-R⁻ GCB-DLBCL

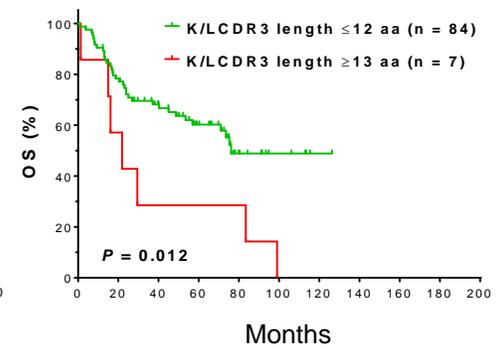


D

DLBCL

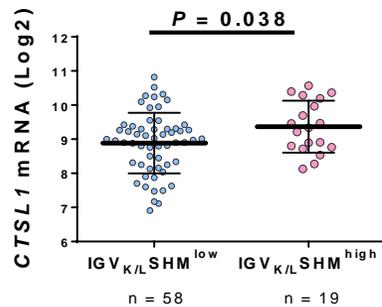


ABC-DLBCL

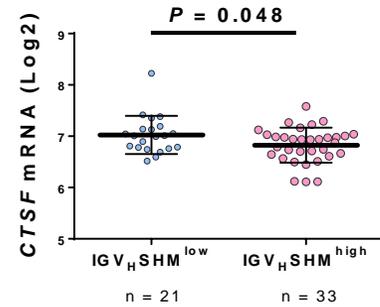


E

GCB-DLBCL

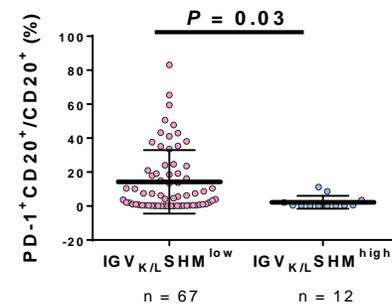


GCB-DLBCL

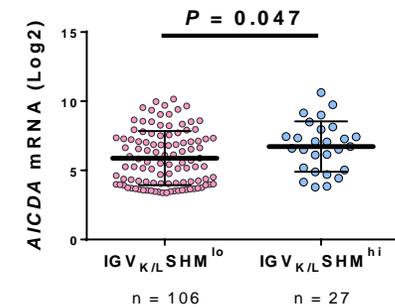


F

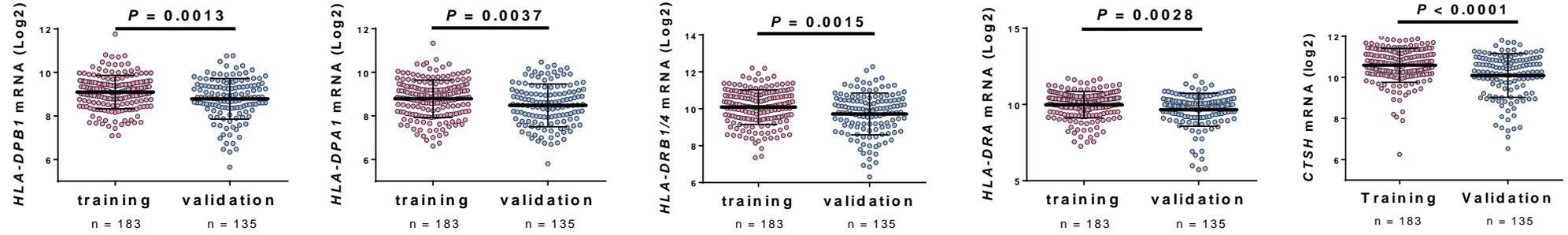
ABC-DLBCL



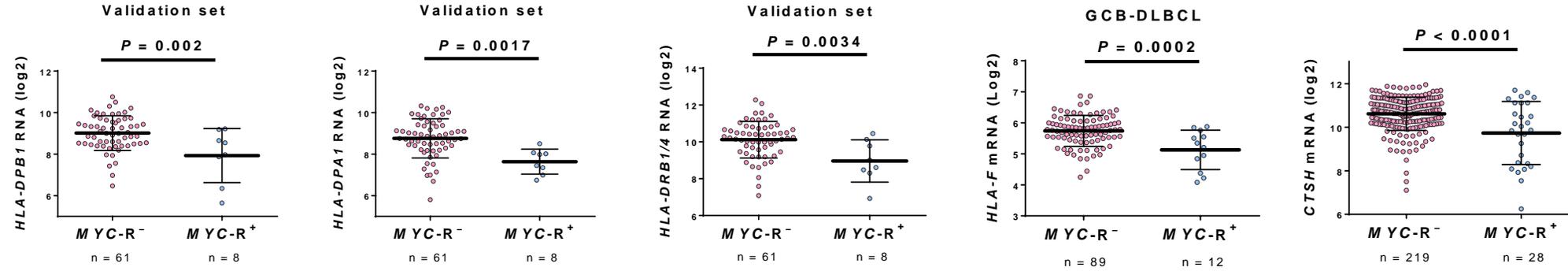
Training set ABC-DLBCL



A



B



C

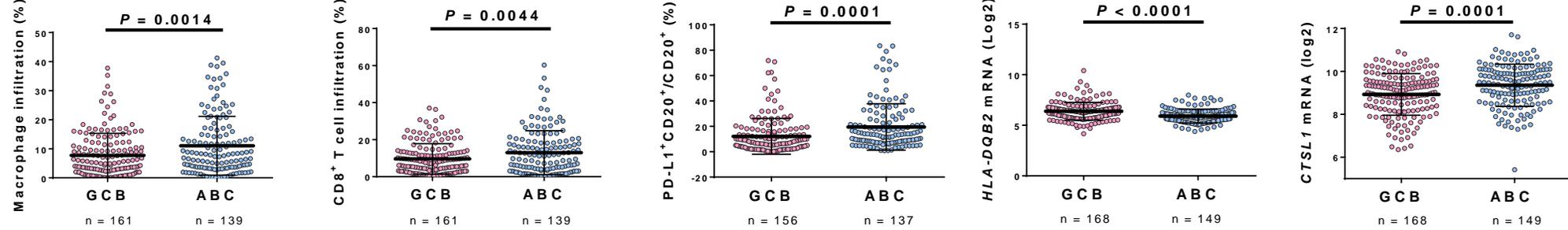


Figure S11

