Supplement

Supplementary Materials and Methods

Clinicopathological characteristics of the study group

Both conventional and immunohistochemical slides were reviewed by a panel of experienced hematopathologists in a multi-step process, where diagnosis was confirmed by two experienced hematopathologists (ACF, HM) in accordance with the current edition of the WHO classification of tumors of the hematopoietic and lymphoid tissues (1). From 80 cases meeting diagnostic and clinical criteria for HGBL-DH/TH with available biopsy specimen, 47 were selected for subsequent genomic analysis, based on tumor DNA quality and library preparation success.

Antibodies and positivity cutoffs employed in the current study are summarized in **Supplementary Table 1**. Fluorescence *in situ* hybridization (FisH) for *MYC* (including *MYC*-Ig fusion), *BCL2*, *BCL6* and chromogenic *in situ* hybridization for EBER were performed, as described (2, 3). *MYC* translocation partner was shown to be an Ig-gene in 45% of cases with sufficient tissue available following DNA extraction (9/20) which is in keeping with previous reports (4).

Clinical information was collected from the original files, and patients' performance status (Eastern Cooperative Oncology Group [ECOG]), stage, treatment modalities, therapeutic response, pattern of relapse, baseline serum levels of lactate dehydrogenase (LDH), revised international prognostic index (R-IPI)(5) and information on survival were anonymously coded alongside hematopathological assessments. Extent of disease was routinely evaluated according to the Cotswold modifications of the Ann Arbor classification (6).

Extraction of nucleic acids

Tumor as well as germline DNA (when available from biopsies taken from non-involved sites; n = 7) was extracted from three FFPE tissue sections of 5µm thickness using Maxwell® RSC DNA FFPE kit (Promega), according to the manufacturers' instructions. Quality assessment and quantification was performed on an Agilent 2100 Bioanalyzer system (Agilent Technologies).

Targeted next generation sequencing

Library preparation was carried out according to manufacturers' instructions and sequencing was performed on the Illumina MiSeq platform (Illumina, San Diego, California, USA) to a median depth of 3569x (s.d. ± 1635).

Panel resequencing data analysis

Resequencing data was processed as described above for whole-exome data, but the remove duplicates step was omitted. Variant calling was done using FREEBAYES (v1.3.2-46-g2c1e395), variants were annotated using ANNOVAR and coverage for each variant was extracted using VCF-QUERY (7). Afterwards, variants were filtered and only variants with a minimum coverage of 100, minimum variant allele frequency of 10% and population allele frequency < 0.001 in GNOMAD or POPFREQMAX database were kept for further analysis.

Variant Calling in exome sequencing data

Raw paired-end data (*fastq* format) was trimmed and quality filtered using FASTP(8) (v0.20.0; minimum length 50bp, max. unqualified bases 30%, trim tail set to 1) and trimmed reads were mapped to GRCh37/hg19 using BWA MEM (v0.7.15)(9). Resulting alignment files in *SAM* format were cleaned and sorted and converted into *BAM* format using PICARD TOOLS (v2.18.4). Next, mate-pair information was fixed, duplicates were removed and base quality recalibration was performed using PICARD TOOLS(10) and dbSNP v138. Single nucleotide variants (SNVs) and short insertions and deletions (indels) were identified following the best practices for somatic mutations calling provided by GATK(11) for version 4.1.7 or higher for matched normal and unmatched tumor samples. Briefly, GATKS MUTECT2(12) (v4.1.7.0) algorithm was applied to all *BAM* files with GNOMAD variants as germline resource and the b37 exome panel data as panel of normal. In cases, where matched normal tissue was available (Cases 1-7), Mutect2 was run in matched tumor-normal mode. Afterwards, FFPE read orientation artefacts were identified and removed according to GATKs guidelines and left-aligned filtered variants were annotated using ANNOVAR(13) (v2019Oct24). Coverage for reference and alternative alleles for each variant were extracted using VCF-QUERY (VCFTOOLS v0.1.13(14)). The top 20 frequently mutated genes (FLAGS(15)) were removed from further analysis. Somatic variants were filtered as follows: Minimum coverage of 40, minimum variant allele frequency of 10%, variant allele frequency < 0.001 in GNOMAD or POPFREQMAX database. To identify genes that are more often mutated than expected, MUTSIGCV (v1.41)(16) was applied and potential driver genes were identified using $p < 0.001$ and $q < 0.1$.

Copy number aberrations

The genomic landscape of HGBL-DH/TH was assessed for somatic copy number aberrations (SCNAs) by CONTROL-FREEC (v11.4)(17) using the tumor-only mode for samples without matched and normal tissue and in matched normal-tumor mode for samples with normal tissue available (cases 1-7). The output (ratio and reads per called window) was converted to run GISTIC2.0 (v2.0.23)(18) excluding CNA calls from chromosomes X and Y and excluding known common CNAs using Broad Institute's panel of normal (ftp://ftp.broadinstitute.org/pub/GISTIC2.0/hg19_support). GISTIC2.0 analysis was performed using default parameters.

Mutational significance and deleteriousness, network propagation, gene set enrichment and mutational cluster analysis

The effect of strong deleterious effects (CADD1.3 phred score > 20) was assessed per sample using a network propagation approach(19) applying a regularized Laplacian kernel based on STRINGDB v11(20) protein-protein interaction network (DIFFUSTATS v1.8.0(21)). Genes affected by strongly deleterious mutations were set to 1, whereas non-mutated genes were set to 0 to model the behavior of the mutation and network diffusion was performed using a parametric method with statistical normalization (*z*-scores).

The effect of potential driver genes identified by MUTSIGCV on neighboring genes was assessed using a network propagation approach (as described above). Resulting *z*-scores were used as pre-ranked input for a rank-MANOVA based statistical approach to detect enriched gene sets (MITCH R packages)(22, 23). Gene set enrichment was performed against HALLMARK gene sets and the NF-kB signaling pathway (genes were retrieved from KEGG; entry ID hsa04064). In addition, the acquired genomic data were processed through the LymphGen algorithm and the mutational patterns sequentially underwent cluster analysis and were subsequently screened manually for an enrichment in overlapping aberrations with the molecular clusters proposed by Chapuy *et al.* (24, 25). Further, a logistic regression framework was employed in order to test for significantly different numbers of mutated genes in a given cluster between HGBL harboring different cytogenetic constellations.

Statistical Analyses

If not stated differently, all statistical analyses were performed using R (v4.1.0) and TIDYVERSE (v1.3.1)(26) for data handling. Filtering of genomic regions was performed using GENOMICRANGES (v1.44.0)(27) and MAFTOOLS (v2.8.0)(28) were used to visualize the data. The number of somatic signatures and the contribution of each signature to each case was estimated based on the YAPSA (v1.18.0) package with calculation of correction factors for WES to avoid biases in the k-mer distribution introduced with the exome capture kit. Progression-free survival and overall survival (PFS, OS) were calculated from the date of diagnosis and censored at last clinical contact. Survival (PFS and OS) according to potential prognostic factors was estimated by means of the Kaplan–Meier method and univariate log-rank test. Survival analysis was carried out employing the R packages survival (v3.2-11) and survminer (v0.4.9).

Supplementary Table 1. Antibodies and positivity cutoffs employed in the current study

Antibody	**Supplier**	**Clone**	**Positivity cutoff**
Bcl2	Lab Vision	100/D5	30%
Bcl6	Dako	BG-B6p	30%
CD10	Menarini	56C6	30%
CD20	Dako	L26	-
CD30	Dako	BerH2	10%
CD38	Leica Biosystems	SPC32	-
CD138	Leica Biosystems	MI15	-
CD56	Leica Biosystems	CD564	10%
Kappa	Leica Biosystems	CH15	-
Lambda	Leica Biosystems	SHL53	-
MUM-1 (Irf4)	Dako	Mum 1P	30%
Ki-67	Dako	Mib-1	-

Supplementary Table 2. Mean sequencing coverage and proportion of targets covered by 40x or 100x coverage of each sample applied in WES.

Supplementary Table 3. Gene list for the custom AmpliSeq panel (Thermo Fisher Scientific, Waltham, Massachusetts, USA) for targeted amplicon sequencing – see separate .xlsx file

Supplementary Table 4. Genes found to be significantly mutated in HGBL-DH/TH according to the MutSig2CV algorithm – see separate .xlsx file.

Supplementary Table 5. All variants described by WES – see separate .xlsx file**.**

Supplementary Table 6. All variants described by panel based NGS – see separate .xlsx file.

Supplementary Table 7. Association between HGBL-DH/TH cases and mutational signatures derived from WES data – see separate .xlsx file.

Supplementary Table 8. LymphGen predictions for each case – see separate .xlsx file.

Supplementary Table 9. A logistic regression framework assessing the number of C3 subgroup mutations present in *MYC* rearranged with additional rearrangements of *BCL2*, *BCL6* or both.

	Estimate	Std.Error	t-value	p-value	CI_Lower	CI_Upper	DF
(Intercept)	2.5714	0.2893	8.89	$2.21*10^{-11}$	1.988	3.1544	44
BCL2/6	-0.6714	0.4004	-1.677	$1.01*10^{-01}$	-1.478	0.1356	44
BCL6	-1.7589	0.3681	-4.778	$2.00*10^{-05}$	-2.501	-1.0171	

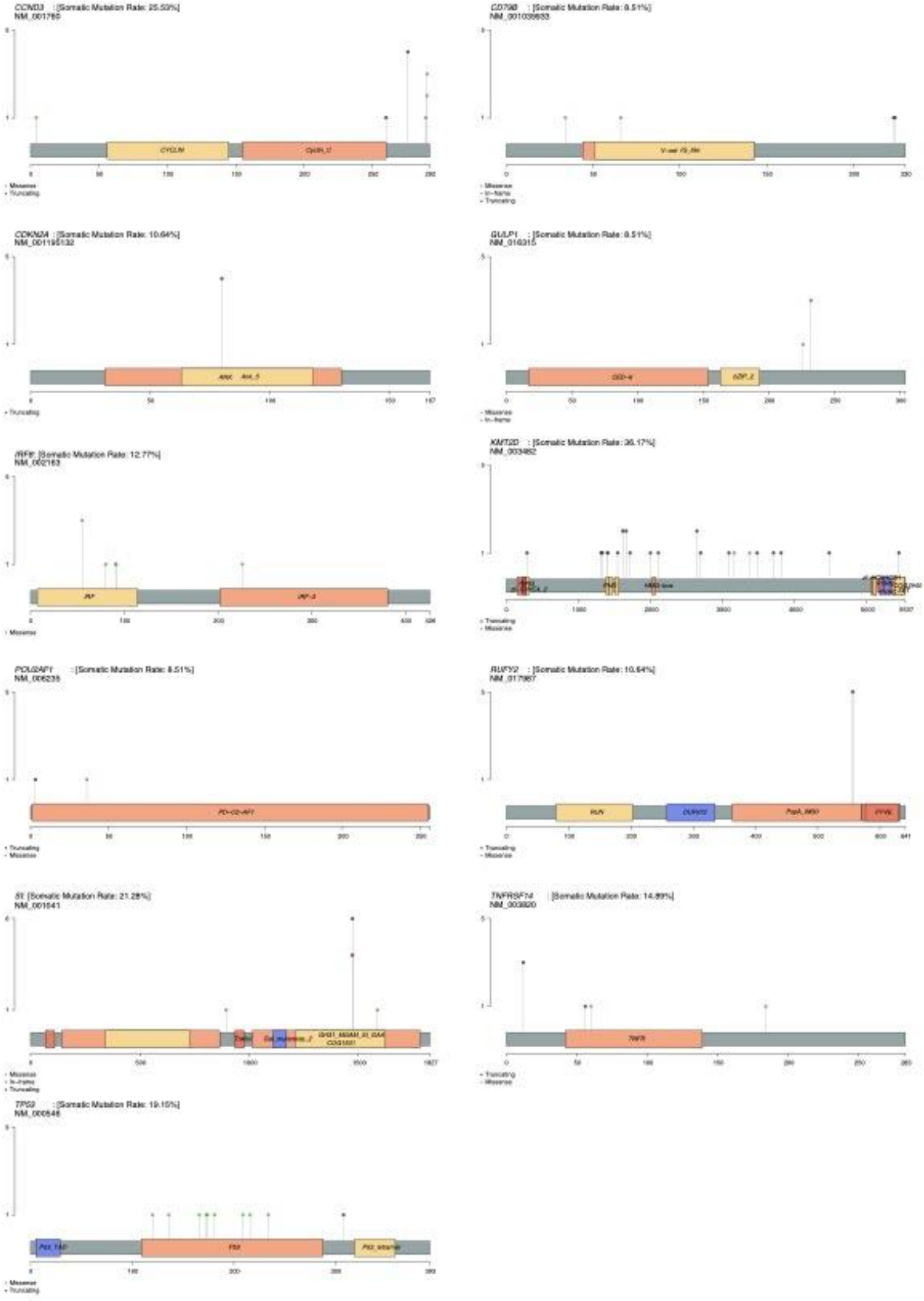
Multiple R-squared: 0.3407, Adjusted R-squared: 0.3108

F-statistic: 12.21 on 2 and 44 DF, p-value: 6.059e-05

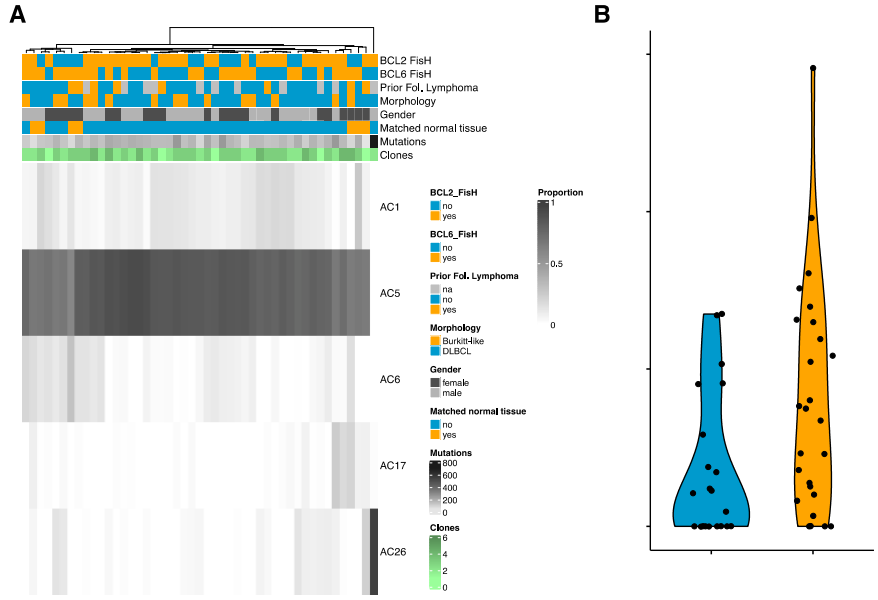
CI = 95% confidence interval, Std. Error = standard error. DF = degrees of freedom.

Supplementary Figure legends

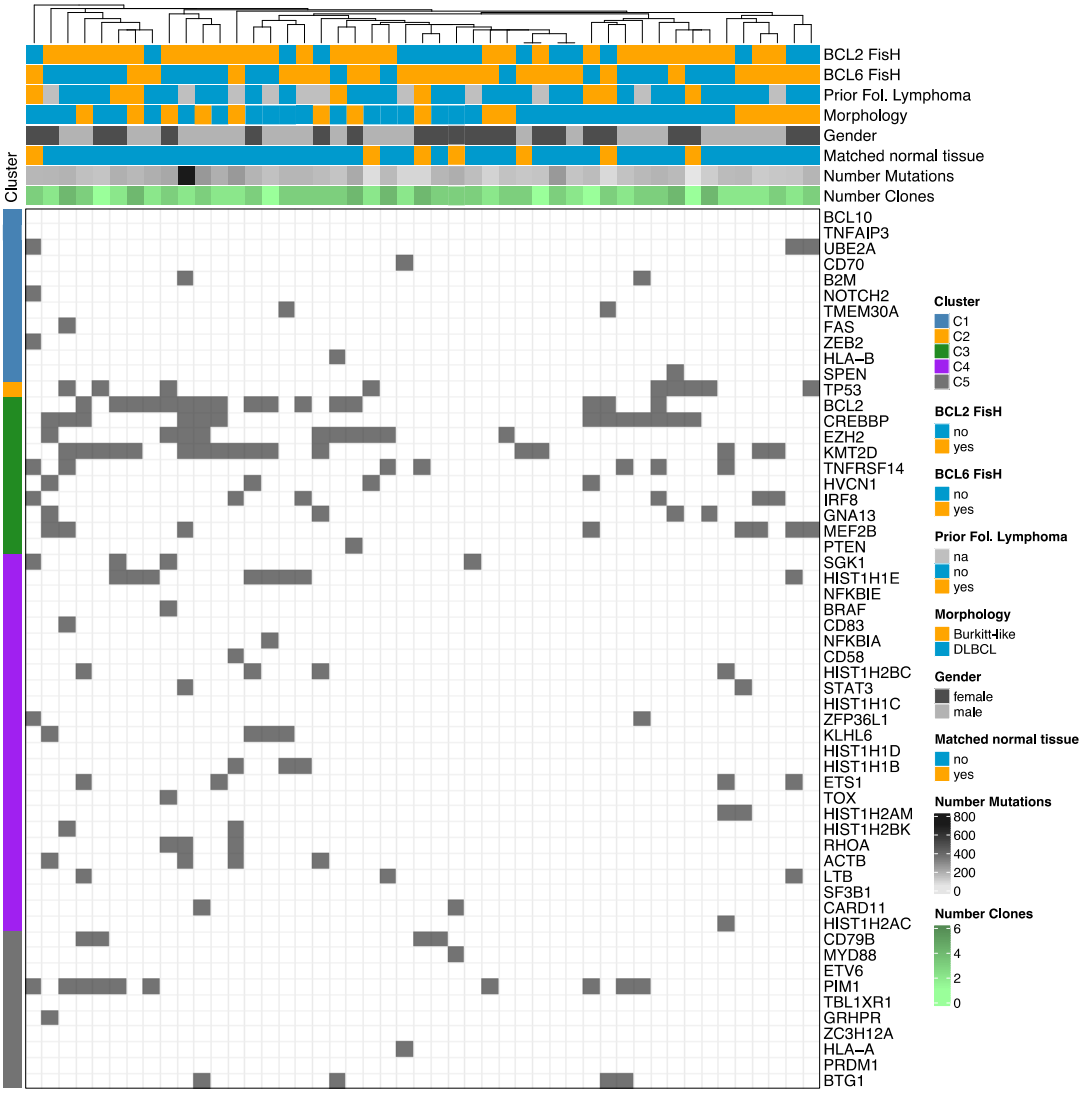
Supplementary Figure 1. Distribution of mutations within selected, significantly mutated genes in the format of lollipop plots.



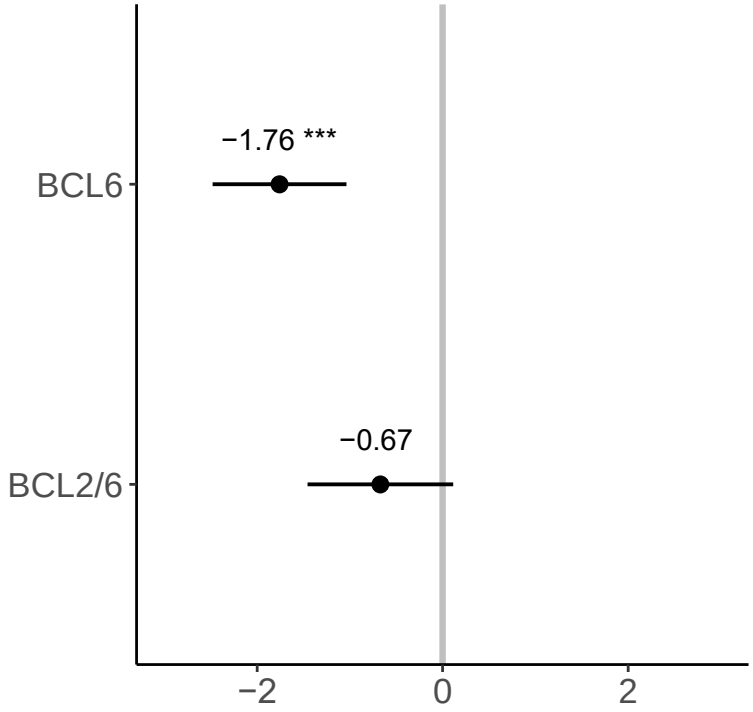
Supplementary Figure 2. Mutational signature analysis. Panel (A) depicts the profiling of mutational signatures driving HGBL-DH/TH revealed a homogenous predominance of the SBS5 signature alongside the significantly emphasized occurrence of the SBS6 signature (implicated in defective DNA mismatch repair) in patients with *BCL6* rearrangements. Additionally, covariates are shown above the plot for each sample. Panel (B) displays the proportion of the SBS6 signature according to *BCL6* FISH status, showing an elevated frequency in *BCL6* rearranged cases.



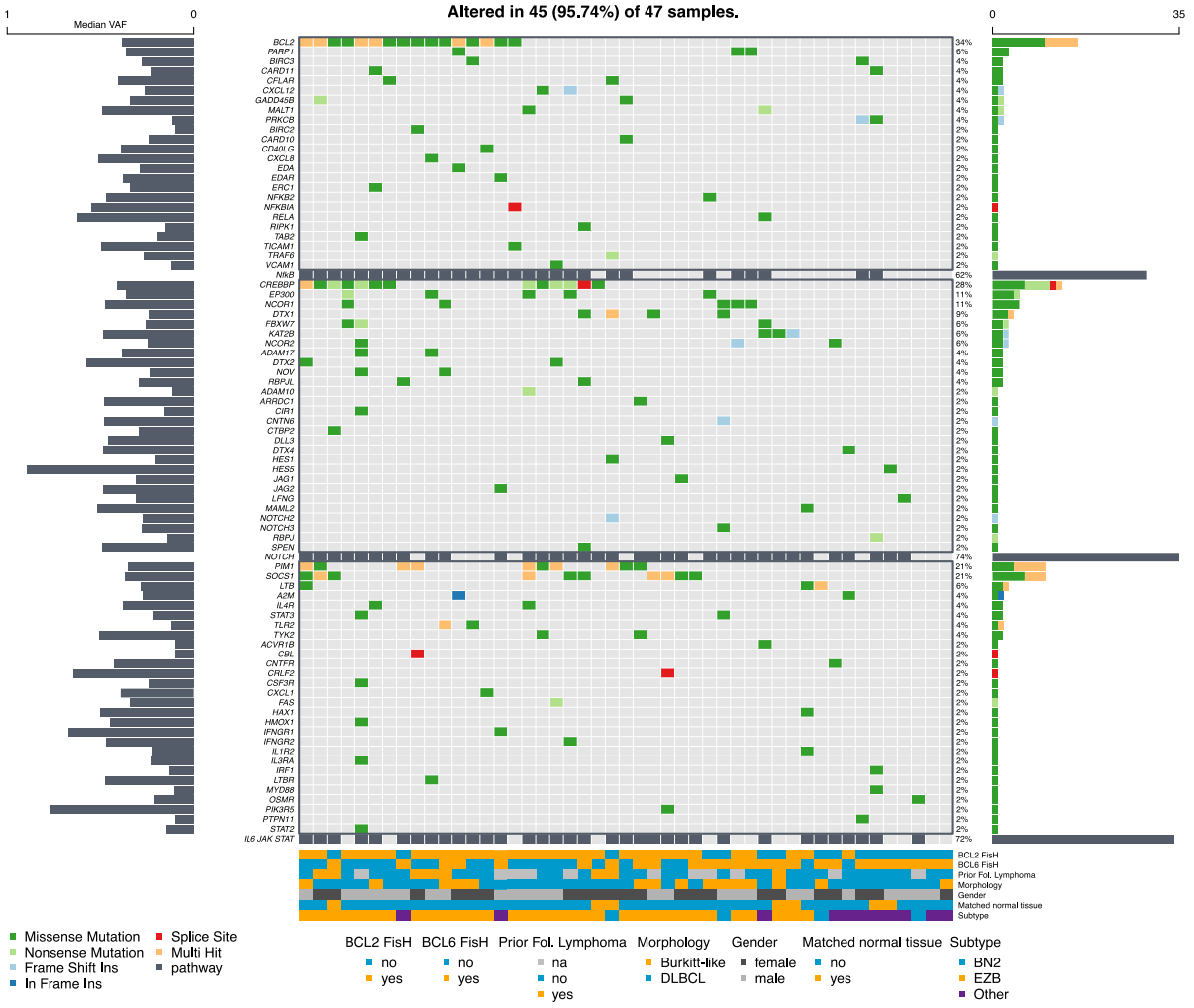
Supplementary Figure 3. Allocation of HGBL-DH/TH samples unto the molecular subgroups/clusters of DLBCL, according to Chapuy *et al.* based on their mutational signature(24). Additionally, covariates are shown above the plot for each sample.



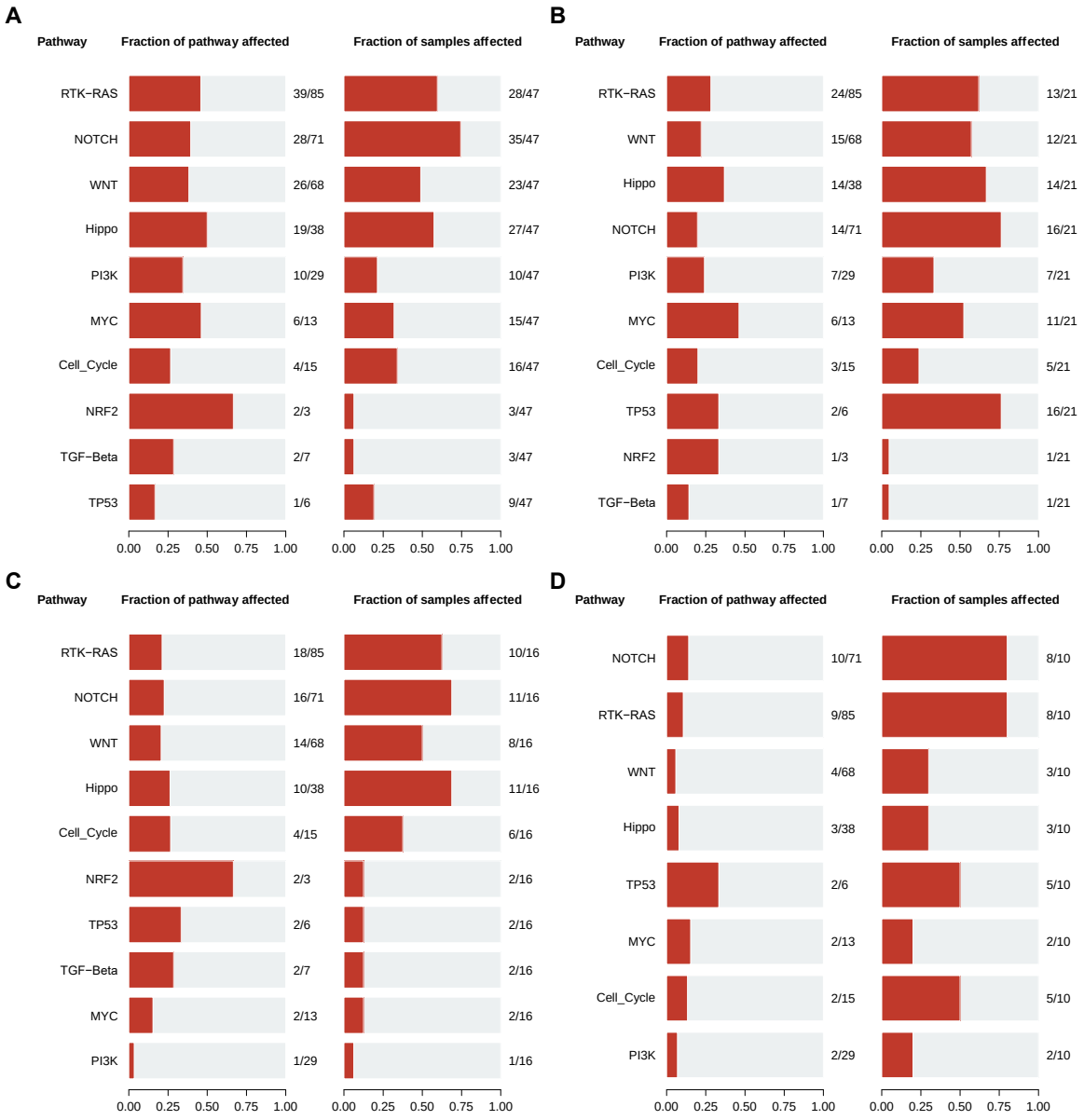
Supplementary Figure 4. Concordance of *BCL2* and *BCL6* rearranged cases with the C3 DLBCL mutational Cluster according to Chapuy *et al.* by means of a logistic regression model; * denotes p < 0.001 (see Supplementary Table 8 for details).**



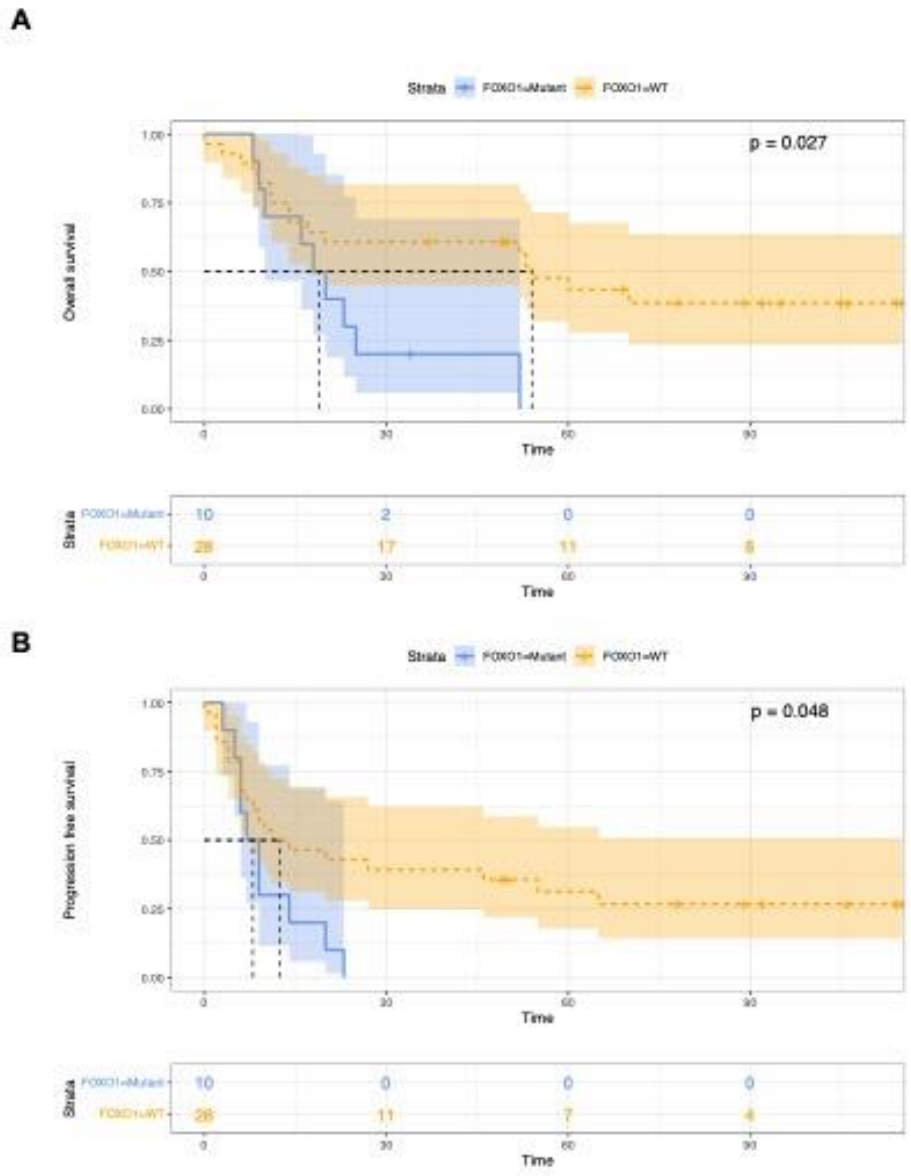
Supplementary Figure 5. Oncoplot depicting distribution of mutations unto the genes of NOTCH, IL6/JAK/STAT and NF-kB signaling pathways. Median of variant allele frequency per gene across mutated samples is shown in the left bar plot; the bar plot on the right-hand side shows the number of mutated samples per gene and pathway. Subtypes were inferred applying the LymphGen algorithm.



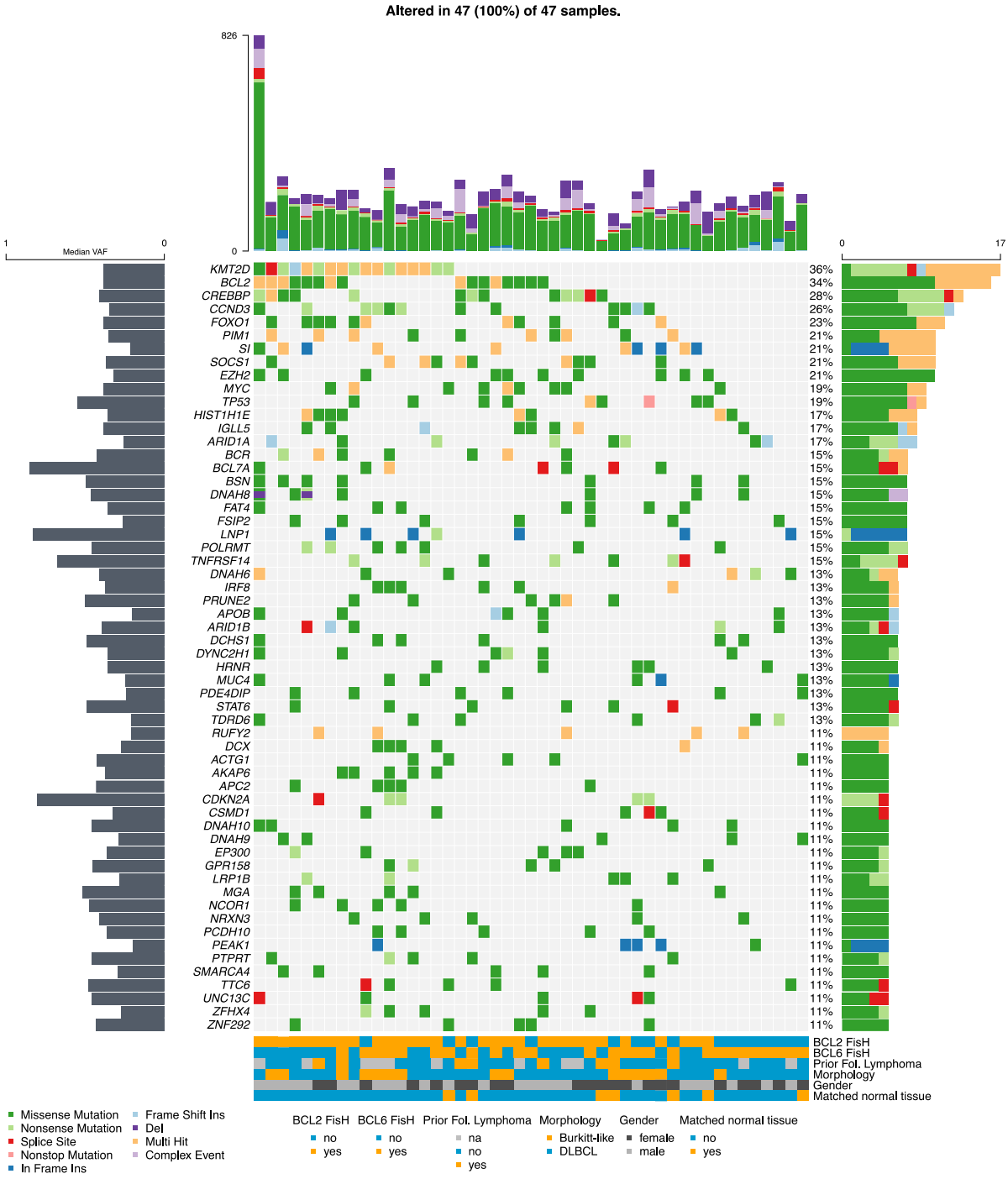
Supplementary Figure 6. Mutational distribution unto oncogenetic signaling pathways. Fraction of mutated genes and fraction of affected samples unto different oncogenetic signaling pathways for the HGBL-DH/TH data set (A), the *BCL2* subgroup (B), *BCL6* subgroup (C), and the triple hit subgroup (D). While RTK-RAS, NOTCH as well as WNT signaling appear to be ubiquitously predominant targets of mutations accross all cytogenetic subtypes, the *TP53* network is especially disrupted in the *BCL2* subgroup.



Supplementary Figure 7. Survival curves according to *FOXO1* mutational status. Overall (A) and progression-free survival (B) according to *FOXO1* mutational status.



Supplementary Figure 8. Oncoplot depicting the mutational spectrum encountered in HGBL-DH/TH prior to filtering through the MUTSIGCV algorithm.



References

1. Swerdlow SH, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood. 2016;127(20):2375-90.

2. Gebauer N, Bernard V, Gebauer W, et al. TP53 mutations are frequent events in double-hit B-cell lymphomas with MYC and BCL2 but not MYC and BCL6 translocations. Leuk Lymphoma. 2015;56(1):179-85.

3. Montes-Moreno S, Odqvist L, Diaz-Perez JA, et al. EBV-positive diffuse large B-cell lymphoma of the elderly is an aggressive post-germinal center B-cell neoplasm characterized by prominent nuclear factor-kB activation. Mod Pathol. 2012;25(7):968-82.

4. Rosenwald A, Bens S, Advani R, et al. Prognostic Significance of MYC Rearrangement and Translocation Partner in Diffuse Large B-Cell Lymphoma: A Study by the Lunenburg Lymphoma Biomarker Consortium. J Clin Oncol. 2019;37(35):3359-3368.

5. Sehn LH, Berry B, Chhanabhai M, et al. The revised International Prognostic Index (R-IPI) is a better predictor of outcome than the standard IPI for patients with diffuse large B-cell lymphoma treated with R-CHOP. Blood. 2007;109(5):1857-61.

6. Lister TA, Crowther D, Sutcliffe SB, et al. Report of a committee convened to discuss the evaluation and staging of patients with Hodgkin's disease: Cotswolds meeting. J Clin Oncol. 1989;7(11):1630-6.

7. O'Fallon BD, Wooderchak-Donahue W, Crockett DK. A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. Bioinformatics. 2013;29(11):1361-6.

8. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i890.

9. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. [cited 2020 07.06.2020]. Available from: https://arxiv.org/abs/1303.3997.

10. Institute B. Picard Toolkit. GitHub Repository: 2019. [07.06.2020]. Available from: http://broadinstitute.github.io/picard/.

11. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.

12. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213-9.

13. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

14. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8.

15. Shyr C, Tarailo-Graovac M, Gottlieb M, et al. FLAGS, frequently mutated genes in public exomes. BMC Med Genomics. 2014;7:64.

16. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-218.

17. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28(3):423-5.

18. Mermel CH, Schumacher SE, Hill B, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41.

19. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet. 2017;18(9):551-562.

20. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607-D613.

21. Picart-Armada S, Thompson WK, Buil A, Perera-Lluna A. diffuStats: an R package to compute diffusion-based scores on biological networks. Bioinformatics. 2018;34(3):533-534.

22. Ziemann MK, A. Multi-Contrast Gene Set Enrichment Analysis. 2019. [07.06.2020]. Available from: https://github.com/markziemann/mitch.

23. Cox J, Mann M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. BMC Bioinformatics. 2012;13 Suppl 16:S12.

24. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nat Med. 2018;24(5):679-690.

25. Wright GW, Huang DW, Phelan JD, et al. A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications. Cancer Cell. 2020;37(4):551-568 e14.

26. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. The Journal of Open Source Software. 2019;4(43).

27. Lawrence M, Huber W, Pages H, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.

28. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28(11):1747-1756.