Summary of Supplementary Materials:

Materials and Methods

Supplementary Figures:

Supplementary Materials and Methods

Patient Samples

Patient collection of diffuse large B-cell lymphoma (DLBCL) samples excluded transformed follicular lymphoma (tFL) cases except in one case where a co-diagnosisof tFL and DLBCL occurred and the included biopsy was identified by an expert hematopathologist (K.B.) as DLBCL without tFL. This case later showed no sign of FL. Information regarding clinical characteristics, overall survival (OS), PFS, and treatment regimens was obtained retrospectively from patient journals, in addition to the lymphoma registry at the Norwegian Radium Hospital. (Supplemental Table 1, supplemental Table 2, supplemental Table 3). Tumor cell content was measured by performing standard hematoxylin and eosin (H&E) staining and immunohistochemical analysis. Cell of origin (COO) was classified according to the Hans algorithm (1), and if necessary the Choi (2) algorithm. *C-MYC*, *BCL2*, and *BCL6* rearrangements were performed by standard FISH protocols performed at the pathology department at Oslo University Hospital.

For the validation cohort (supplemental Table 4), two cases (not used in results analyses) were duplicated from the exome sequencing cohort, representing spatial and fixation differences, and served as controls of the targeted sequencing performance. We confirmed, in these biopsies, between 81-90% of the originally identified coding variants.

DNA Extraction

For the whole exome sequencing cohort, genomic DNA was extracted from 52 fresh frozen lymphoma biopsies with >50% tumor content (mean 78%, median 80%) as determined by an expert hematopathologist (K.B.) using standard H&E staining and immunohistochemistry. For each biopsy, DNA was extracted from 10 sections with a size approximately 0.5 cm x 0.5 cm using the Promega Maxwell 16 instrumentation (Promega, Madison, WI). Briefly, tissue sections were pre-processed using the Fast-prep-24 homogenization with a ceramic bead for 30seconds at 4 m/s in 75µl of .5 M EDTA solution (Promega, Madison, WI). Samples were then incubated overnight with 300µl of nuclei lysis solution and proteinase K (Promega, Madison, WI). DNA was extracted from processed tissue via the Maxwell 16 DNA isolation automated magnetic bead instrument using the Tissue setting; 250µl of elution was eluted of beads using a magnet and stored at -80°C.

Genomic DNA from 36 cases of matched blood was extracted using a similar protocol. Briefly, 5 ml of EDTA-blood was pre-processed using centrifugation at 17000RCF with no brake at 4°C in a sucrose lysis solution and washed with PBS. Samples were then incubated overnight with 300µl of nuclei lysis solution and proteinase K (Promega, Madison, WI). DNA was extracted from processed tissue via the Maxwell 16 DNA isolation automated magnetic bead instrument using the Blood setting; 250µl of elution was removed of beads using a magnet and stored at -80°C.

4 biopsies and 2 matched blood samples (of which 2 biopsies and 1 blood case were extracted from an Oslo WES cohort case at a different spatial position and used as a sequencing and DNA extraction control) were collected and DNA extracted at Helsinki University Central Hospital Comprehensive Cancer Center. For tissue homogenization the Fast-prep-24 was utilized

and DNA extraction was completed using the Nucleospin TriPrep and Qiagen AllPrep (Qiagen, Venlo, Netherlands) kits per manufacturer's protocol.

For the validation cohort, DNA was extracted from 37 formalin-fixed, paraffin-embedded (FFPE) biopsies. For each biopsy, DNA was extracted using the Qiagen QIAamp DNA FFPE kit (No 56404) (Qiagen, Venlo, Netherlands) per manufacturer's instructions. 11 biopsies were collected and DNA extracted at Helsinki University Central Hospital Comprehensive Cancer Center.

Library Preparation and Whole Exome Sequencing (WES)

Library preparation was performed using SureSelectXT Human All Exon V5 (Agilent, Santa Clara, CA), following manufacturer's instructions. In brief, one microgram of genomic DNA was sheared using the Covaris E220 instrument to achieve fragment target size of 150-200 bp. Fragmented DNA was end-repaired, adenylated and adapters ligated before 6 PCR cycles of amplification. Seven hundred and fifty nanograms of amplified library was hybridized to exome capture probes for 16 hrs, before capturing with streptavidin-coated beads. Eluted exome libraries were further amplified by 10 PCR cycles to introduce post capture indexes. The libraries were quality controlled using Agilent Tape-station for size distribution, and quantified using Agilent qPCR kit for Illumina sequencing libraries (Agilent, Santa Clara, CA), before pooling and sequencing. Exome libraries were sequenced paired-end 2x100bp using sequencing by synthesis (SBS) chemistry v3 on an Illumina HighSeq 2500 (Illumina, San Diego, CA). Raw sequencing data was converted to FASTQ files and demultiplexed using the Illumina bcl2fastq v1.8 software.

NGS Variant Calling and Filtering

Reads of each sample were mapped lane-wise with BWA-mem (3) to the human reference genome (build b37 with an added decoy contig). Marking of duplicates was performed with Picard tools; GATK tools (4) were used for two-step local realignment around INDELS (each matched sample-pair was processed jointly), base quality recalibration and calculation of coverage statistics. Somatic SNV detection on the matching paired samples was performed with MuTect (5) and Strelka (6). Strelka alone was used for somatic INDEL detection. The variant calling pipeline was subject to benchmarking and validation through collaborations with the International Cancer Genome Consortium (7), where we experienced that the consensus SNV calls by MuTect and Strelka reduced the overall impact of false positives. Here, we used the same approach (*i.e.* relying primarily on consensus SNV calls), and additionally we implemented a custom filter to mitigate false positives among calls made by a single algorithm only. Calls identified in only one caller that had i) a sequencing depth less than 8 in the normal control and ii) less than 14 in the tumor, and iii) an alternate allele with an allelic fraction in the tumor of less than 0.15, and iv) allelic fraction in the normal greater than 0.05, were manually reviewed using IGV browser (v2.3) for inclusion. Using this approach, 71 calls out of a total of 120 were discarded as sequencing artefacts.

In order to understand the functional role of identified variants, all variants were subjected to a computational annotation workflow including a modified version of ANNOVAR (release 2015 Dec14, using RefSeq as the gene model), PFAM (protein domain information, v27.0), UniProt KB (functional protein properties, release 2015_08), and the Catalogue of Somatic Mutations in Cancer (COSMIC, release 78)(8-11). Variants were furthermore appended with annotations relating to clinical relevance and germline allele frequencies, using the database of curated mutations (DoCM, v3.2), dbSNP build 147/ClinVar (2016_10), 1000 Genomes Project phase 3, and ExAC (release 0.3) (12-15).

The standard processing pipeline (as described above) relies on the availability of matched control samples. A modified pipeline that employed an "artificial" control sample was therefore used for processing tumor samples that were missing a matched control. Each artificial control sample was created by altering the individual SAM/BAM alignment records of its tumor counterpart. While an artificial control alignment file preserves the important exome sequencing properties of the source tumor file (e.g., the characteristically uneven coverage distribution), its sequence information is derived from the reference genome.

The applied alignment record modifications: 1) the original read base sequence (SAM field "SEQ") was replaced by reference genome subsequence starting at the alignment's mapping position and having length equal to the original read's length. 2) the original value in the "CIGAR" SAM field was replaced by value "*x*M", where *x* equals the original read's length. 3) the optional MD tag was given a new value: "MD:Z:*x*", where *x* equals the original read's length. 4) the optional NM tag was given a new value: "NM:i:0". (Note: the artificial control samples were only used for somatic SNV and indel calling, not for somatic CNV or structural variation calling, for which they are not suitable in the described form.).

Variant calling in tumor samples paired with artificial control, including targeted sequencing samples, produced a mix of germline variants and somatic mutations. We therefore set up a set of filtering procedures, based on our annotations, to minimize the presence of known germline variants, enriching for somatic events. Specifically, we excluded all variants that overlapped with germline variants found in the 1000 Genomes Project (minor allele frequency > 1% in any population), and ExAC (minor allele frequency > 1% in any population). In addition, we excluded variants present in dbSNP that were not observed in COSMIC and that had no clinical associations (as given from ClinVar/DoCM cross-references). We also removed variants at a 2% frequency in the Broad panel of normal 9 with over 8,000 normals. We excluded any variant found in more than five cases. Finally, we restricted the variant set to coding variants (missense, stopgain/stoploss, frameshift/non-frameshift, splice site donor/acceptor).

Some of the sequenced tumor samples proved to be affected by guanine oxidation, a type of DNA damage which results in artifactual G>T/C>A variant calls (16). Additional filtering was applied in order to remove the artifactual calls. All G>T/C>A variant calls in the affected samples (overall OXIDATION_Q value as calculated by Picard's CollectOxoMetrics < 38)* were annotated with FoxoG values as defined in (16) (Strelka's internal realignment was not taken into account when considering the variant support for purpose of the FoxoG annotation). Individual FoxoG filter thresholds were selected for the affected samples, reflecting the varying extent of oxidation damage.**

* overall OXIDATION_Q values of the affected samples (calculated across all sequence context):
P23-2: ~27.6
P25-1: ~28.2
P25-2: ~37.5

* FoxoG thresholds applied (G>T/C>A variants satisfying the respective FoxoG threshold conditions were discarded):
P23-2: FoxoG >=0.8

P25-1: FoxoG >=0.8
P25-2: FoxoG =1.0

Targeted Next-Generation Sequencing

DNA from FFPE tumor samples was extracted using the Qiagen QIAamp DNA FFPE kit (No 56404, Qiagen, Germany). The custom Agilent SureSelect panel of 139 genes of interest (Supplemental Table 8) was designed based on the targeted custom exon capture SureSelect v6. The design utilized hg19 GRCh37 including exons and untranslated regions (UTRs) with an extension of 10 bases from each UTR. Library construction was performed as previously described but replacing the exome capture probe set for the custom probe set during hybridization. Libraries were sequenced paired-end 2x150bp using a High-output kit v2 on an Illumina NextSeq 500 instrument. Calling of variants did not include filtering of common FFPE induced artifacts.

Copy Number Analysis

Copy number aberration analysis was performed on the whole exome sequencing data utilizing FACETS(17) version 0.5.14. In brief, a snp pileup matrix against NCBI 00-All.vcf download 04-2019 was created for each biopsy. FACETS was run using a preprocessing critical-value of 150 or 300. For cases w/o matched normal FACETS was run against each sex-matched available normal and a normal with a comparable noise profile was selected for final analysis. SNP6.0 analysis of available biopsies, 36 of 37 WES with matched normal cohort cases and 4 of the 5 serial sampled cohort, was performed at AROS Applied Biotechnology A/S in Denmark.). SNP analysis was performed on the Affymetrix Genome Wide SNP6.0 Array platform (Gene Chip lot 4296976) using 0.25µg of DNA following the manufacturer's standard protocols. The resulting CEL-files were processed through the first step of the PennCNV-Affy pipeline to generate the Log R Ratio (LRR) and B-Allele Frequency (BAF) file (18). After allele-specific copy number aberrations were calculated and annotated from the Log R ratio and B allele frequency (BAF) values using the R-package ASCAT version 2.4 (19) with correction for GC content. The copy number and SNP probes were collapsed so that one gene had one copy number value per sample. The resulting copy number data were annotated with the annotation file version 35 retrieved from the Affymetrix website (GenomeWideSNP_6.na35.annot.csv). The resulting copy number states were integrated into clonality analyses described below.

Mutational Signatures Analysis

The weighted contributions of the known cancer associated mutational signatures based on nucleotide substitutions (20) was analyzed in house using Matlab v2017b (MathWorks, USA), the maximum number of mutational processes acting on a single tumor was set to 4. Input included all consensus calls, both coding and non-coding.

In addition, to delineate the weighted contributions of the known cancer associated mutational signatures the deconstructSigs (R package) was applied (20), the maximum number of mutational processes acting on a single tumor was set to 6, based on previous findings for DLBCL. Input included all consensus calls, both coding and non-coding. Two samples had too few calls (n < 50) for the analysis.

Significantly Mutated Genes

Significantly mutated genes were identified with MutSig2CV (Broad Institute, Cambridge, Massachusetts, USA) [12,25]. All variants in the WES with matched normal cohort, with serial biopsies combined in a union list, were used with default parameters.

Mutated driver genes were identified from the list of coding variants in the total cohort and from subsets of cases corresponding to GCB, ABC, PMBCL, deceased, alive cases and the various biopsy types. The OncodriveFM and OncodriveCLUST programs (version 2.4.1) [26] were used with the following settings; hg19 (GRCh37) assembly, number of mutations a gene must have for CLUST analysis >=2, 10% threshold for the minimum number of mutated samples for a gene for FM analysis and a default filter excluding genes not expressed across tumors from The Cancer Genome Atlas (TCGA) pan-cancer projects [27]. Genes with a q-value of less than 0.1 were selected as harboring potential driver mutations [25].

Clonality Analyses

For any variant in the serial samples in which a variant was not identified by the caller as present in either of the samples, we back tracked into the paired sample for potential reads confirming the variant. The clonal structure of these updated variant lists were inferred with sciClone (version 1.1;(21)) with default parameters, except for clustering method = BMM Gaussian for all samples and minDepth = 50 for cases P2, P13, P1824 and R13 due to too few mutations when using a minimum depth of 100. Single nucleotide variants (SNVs) in copy number altered regions or with evidence of complete or partial LOH (inferred by ASCAT) were reviewed and excluded. ASCAT infers tumor fraction in a given sample and this information was used to correct the variant allelic fractions (VAF) by dividing fraction by tumor percentage. The resulting cluster files from sciClone were given as input to the R package ClonEvol (https://github.com/hdng/clonevol) to build a tumor phylogeny. Clusters with too few mutations and clusters that did not make sense biologically (for example by including mutations present in both diagnostic and relapse, as well as mutations found exclusively in one of the biopsies) were removed, as recommended by the authors of ClonEvol. For each sample, the founding cluster was set as the cluster with largest fraction (around 50%) as well as largest number of mutations. Often a cluster below 10% fraction appeared with mutations found exclusively in each biopsy, and therefore was not considered a reliable estimate of a cluster (Fig. S6). The resulting phylogeny was illustrated using the R-package Fishplot (22) (R-version 3.4.1).

Deleterious Druggable Analysis

Targeted cancer drugs were identified through the Open Targets Platform (downloaded 09.19) [30], with additional filtering for drug association types (i.e. 'known_drug') and with disease phenotypes associated with cancers (expressed as Experimental Factor Ontology identifiers) collected from oncotree (http://oncotree.mskcc.org/#/home)(phenotypes available upon request). Coding variants, in serial samples, were further filtered based on prediction of deleterious effect using Combined Annotation Dependent Depletion (CADD) [31] phred score of over 10 (downloaded 15.10.19).

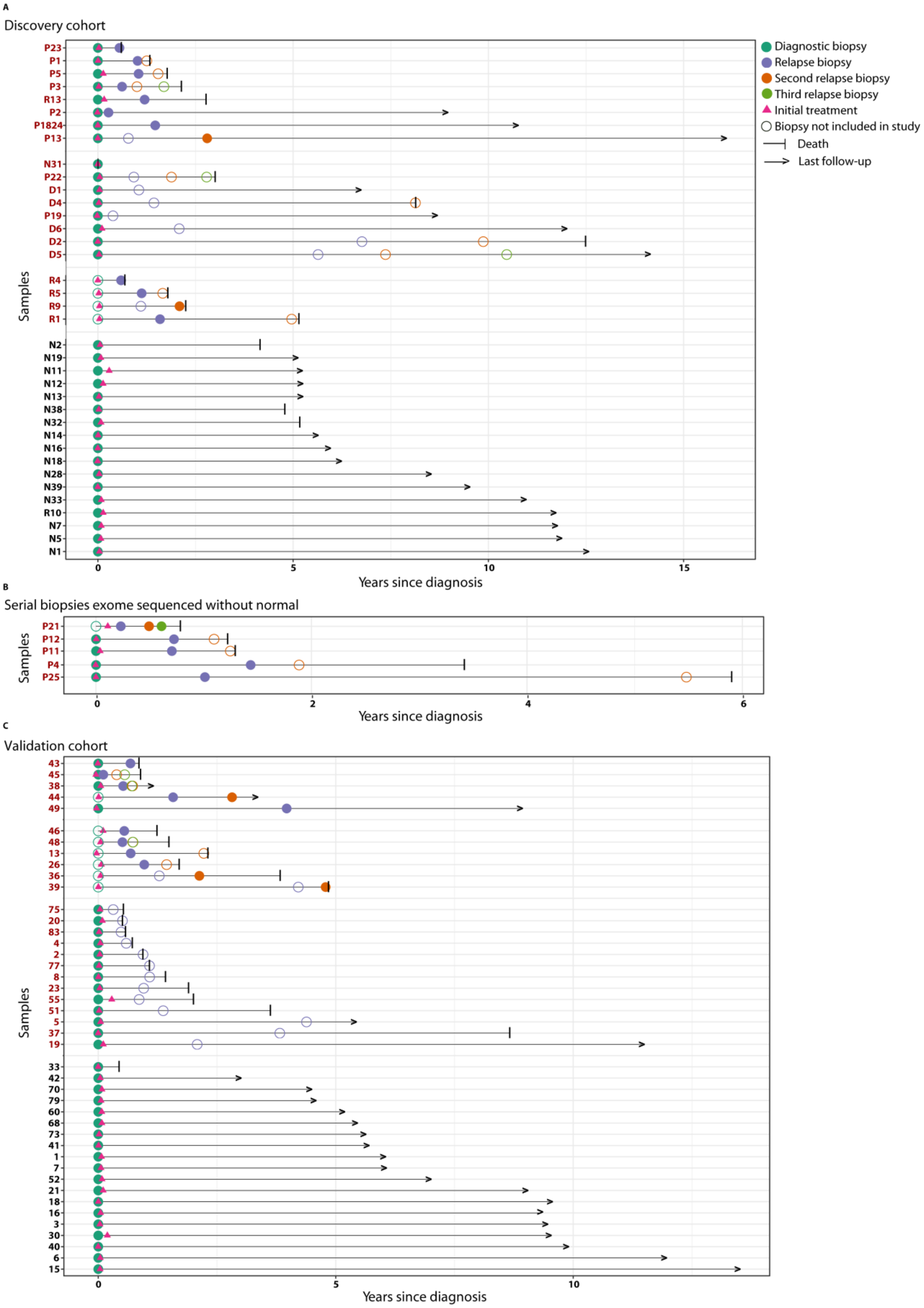
Immunohistochemical Staining

Three µm thick paraffin sections per case were exposed to Heat-induced Epitope Retrieval (HIER) using PT-Link product number PY11730 (Dako, CA) and low pH buffer for 20 minutes. Immunostaining was performed in a Dako Autostainer (Dako CA) using antibodies at concentrations of 1:200 EP1395Y (ab52922, Abcam, MA) and 2M2 (LS-B2200, Lifespan BioSciences, WA) against HLA-A and B2M, respectively, at 30 min at RT. Bound antibody was visualized applying the Dako EnVisionTM Flex+ System (K8012, Dako, CA).

Statistical Analysis

HLA-A coding variant burden at diagnosis was compared between cases that go on to relapse and those that do not. An unpaired two-tailed t-test with Welch's correction was applied using GraphPad Prism (Version 8.01) before and after POLYSOLVER calling and with and without 2 patients who did not receive RCHOP-like therapies. Each patient's fraction of shared mutations between initial and relapsed biopsies was calculated, and a linear regression on time-to-relapse was performed in R (version 3.4.1).

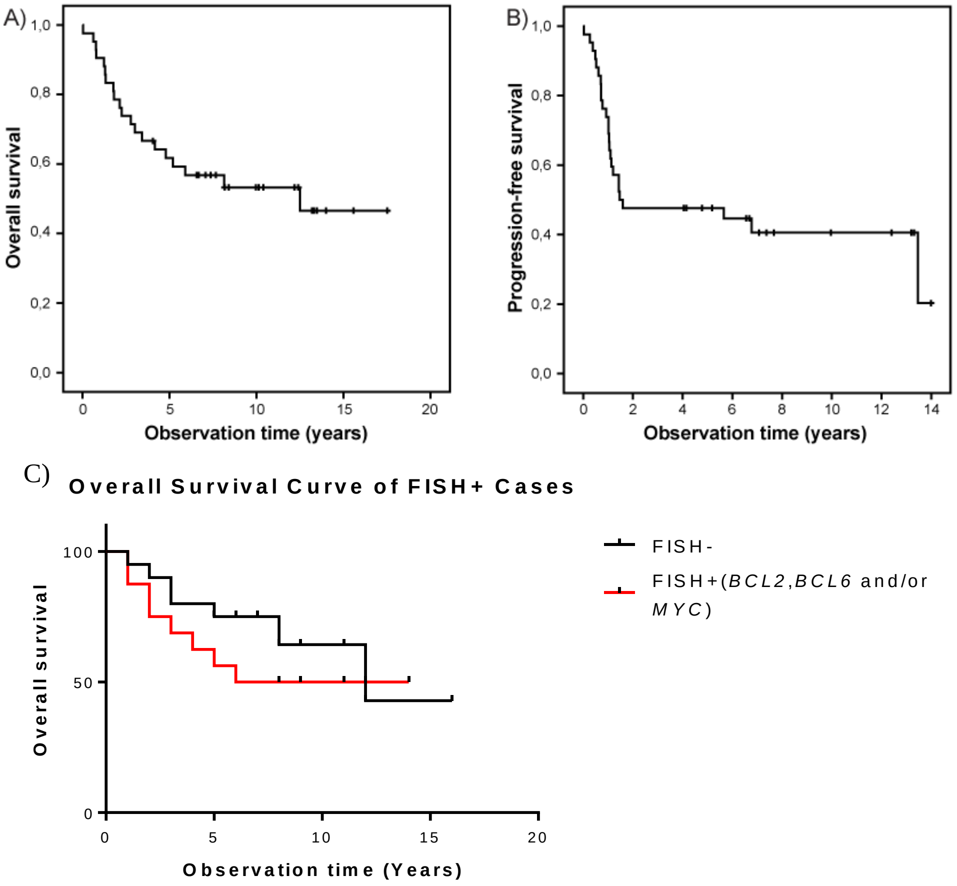
Supplemental Figure 1



Supplemental Figure 1. Time Course of the Nordic DLBCL Cohorts.

Each horizontal line represents the time line for a particular patient, the length indicating total observation time from diagnosis. Patients are ordered according to total observation time, with DLBCL biopsies from patients with relapsed or refractory disease (red labels) and DLBCL biopsies from patients without relapse (black labels) shown vertically. Black tick marks indicate time of death and black arrows representing last follow-up. Filled circles indicate DLBCL biopsies included in the study; empty circles indicate biopsies not available for sequencing. The circles colors represent the biopsy status (dark green = diagnostic biopsy, purple = first relapse, orange = second relapse, light green = third relapse). A pink triangle is used to represent the initial start of treatment; most patients received treatment on or immediately after diagnostic biopsy sampling. **A** represents the whole exome sequencing with matched normal (blood), **B** represents an additional cohort of serial rrDLBCL biopsies, which was exome sequenced without matched normal, **C** represents the validation cohort utilized for targeted sequencing with FFPE tissue.

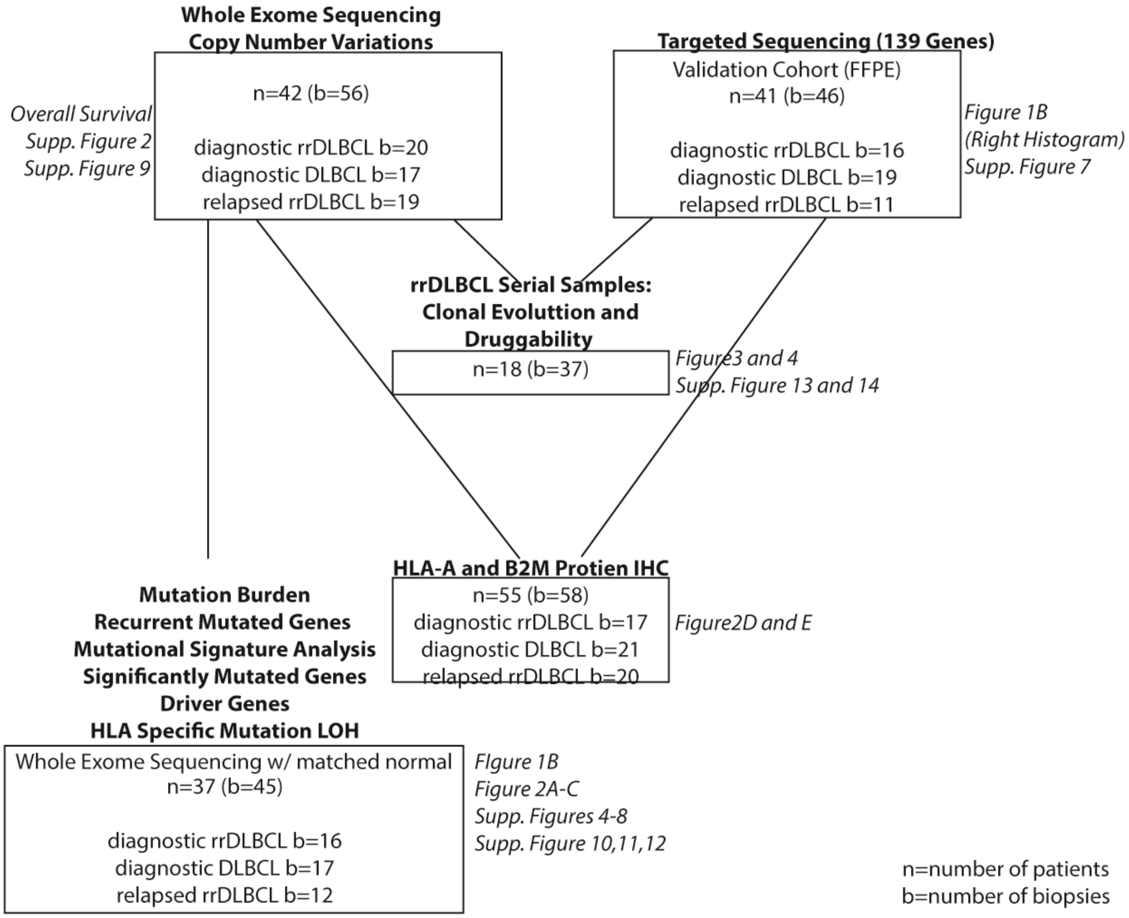
Supplemental Figure 2



Supplemental Figure 2. Survival Curves for Exome Sequenced Cohorts.
A Overall survival curves for the entire exome sequenced (including cases without matched blood) cohort plotted by survival proportion and observation time. **B** Progression-free survival curve for the entire exome sequenced (including cases without matched blood) cohort. **C** Overall

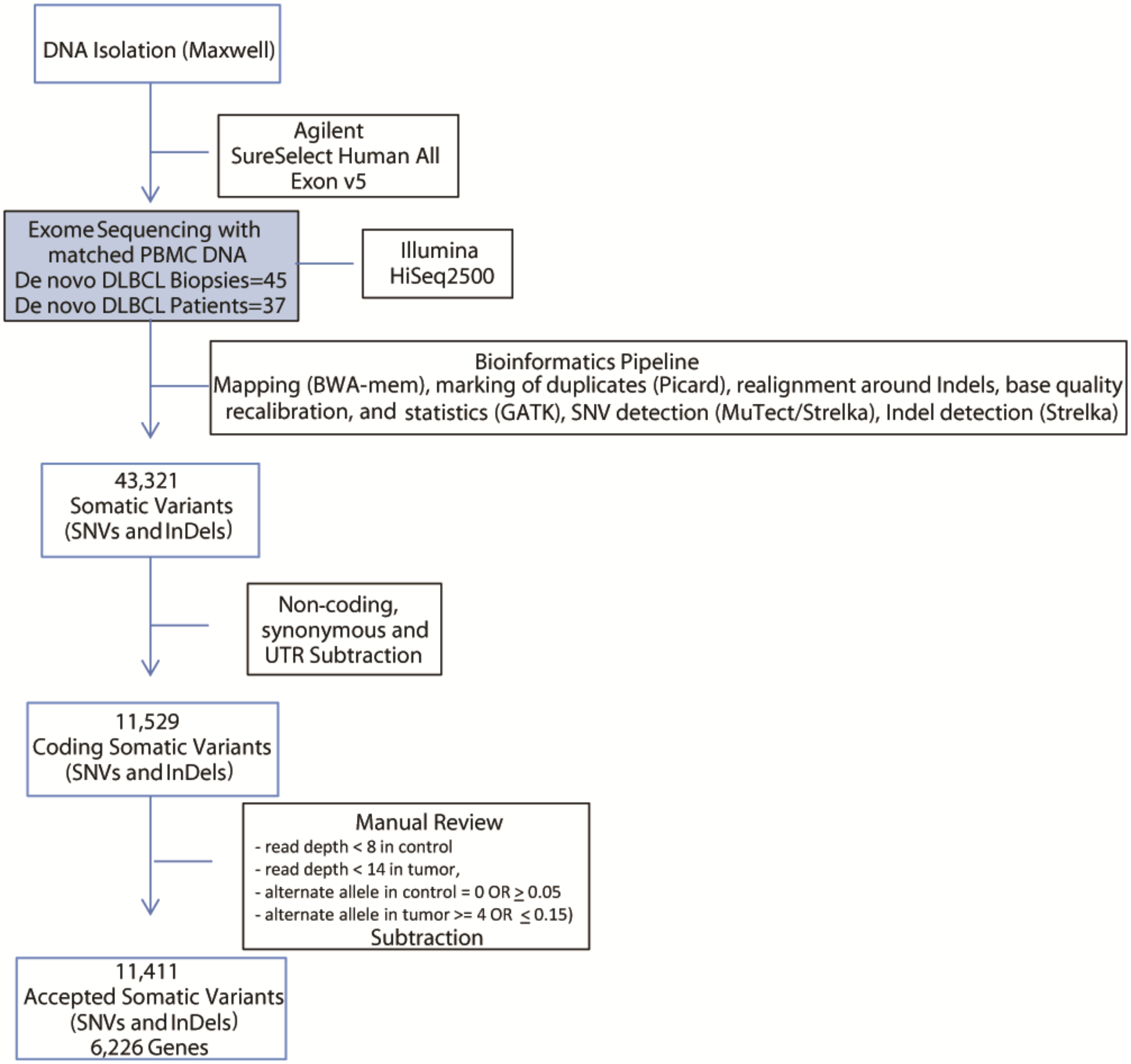
survival curves for the cohort divided by FISH translocation status plotted by survival proportion and observation time. (Log rank test, *p*=0.4580)

Supplemental Figure 3



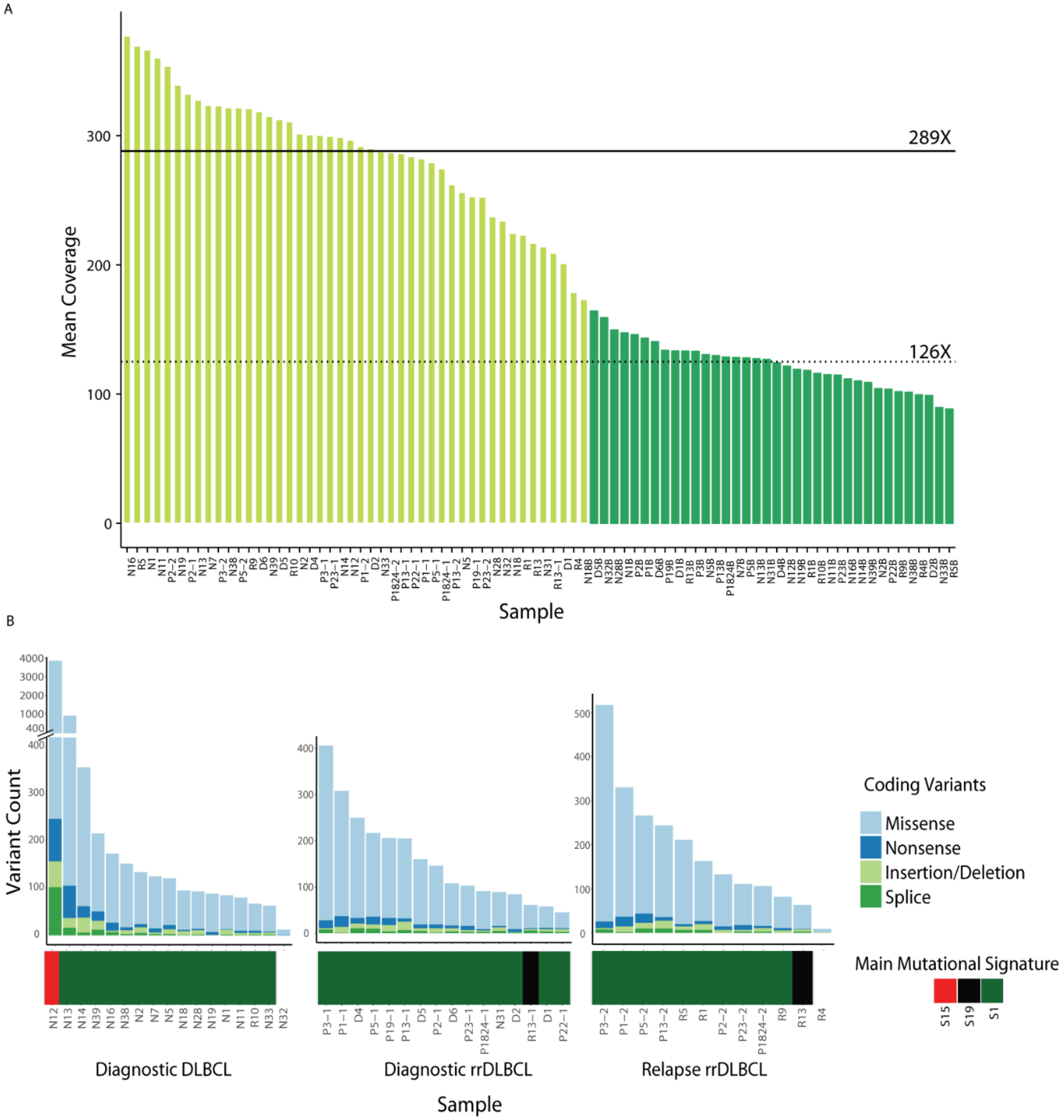
Supplemental Figure 3. Consort Diagram of DLBCL Biopsy Use.
Diagram of the number of cases (n) and which biopsies (b) and the relapse status of each type (within box) are used for each analysis (bolded experiments above each box). In italics to the right of each box is which figures are associated with the cases/biopsies.

Supplemental Figure 4


DNA Isolation (Maxwell)
Agilent SureSelect Human All Exon v5
Exome Sequencing with matched PBMC DNA
De novo DLBCL Biopsies=45
De novo DLBCL Patients=37
Illumina HiSeq2500
Bioinformatics Pipeline
Mapping (BWA-mem), marking of duplicates (Picard), realignment around Indels, base quality recalibration, and statistics (GATK), SNV detection (MuTect/Strelka), Indel detection (Strelka)
43,321 Somatic Variants (SNVs and InDels)
Non-coding, synonymous and UTR Subtraction
11,529 Coding Somatic Variants (SNVs and InDels)
Manual Review
- read depth < 8 in control
- read depth < 14 in tumor,
- alternate allele in control = 0 OR ≥ 0.05
- alternate allele in tumor >= 4 OR ≤ 0.15)
Subtraction
11,411 Accepted Somatic Variants (SNVs and InDels) 6,226 Genes

Supplemental Figure 5. Exome Sequencing Work Flow for WES with Matched Normal Cohort.
DNA was isolated from whole blood and frozen tumor tissue using the Promega automated Maxwell system. Library preparation was performed at the Oslo Genomics Core Facility using the Agilent Sure Select Human All Exon V5. An Illumina HighSeq2500 instrument was used for exome sequencing of samples from 45 biopsies of de novo DLBC. Reads of each sample were mapped lane-wise with BWA-mem to reference genome version GRCh37 with added decoy contig. Marking of duplicates was performed with Picard tools; GATK tools were used for two-step local realignment around indels, base-quality recalibration and calculation of coverage statistics. Somatic SNV detection was performed with MuTect and Strelka. Strelka alone was used for somatic InDel detection. Due to low coverage in various regions of the exome, low confidence calls in coding mutations were reviewed manually, 108 variants were discarded as sequencing errors.

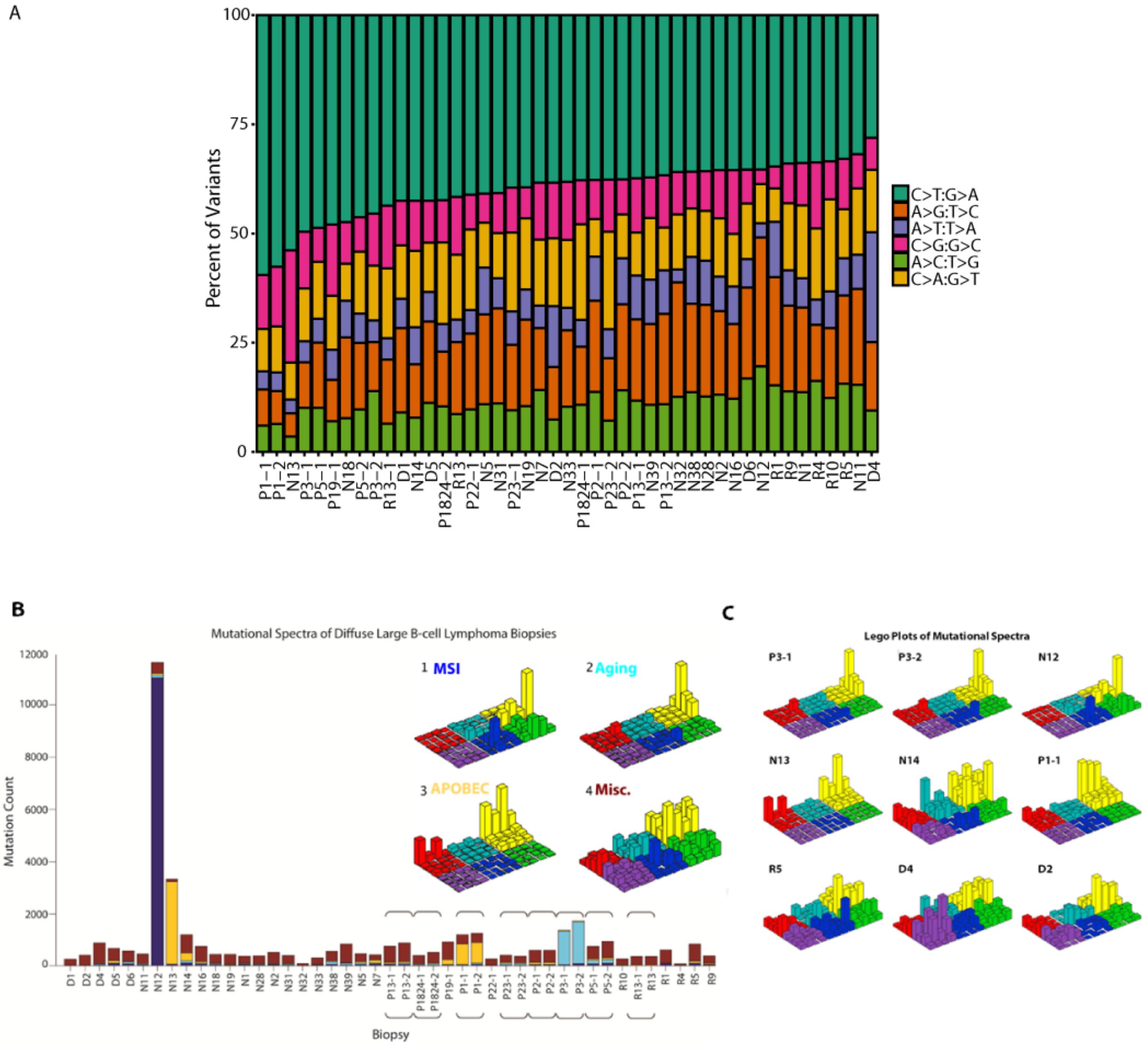
Supplemental Figure 5



Supplemental Figure 5. Exome Sequencing Coverage and Cosmic Mutational Spectra of WES with Matched Normal Cohort.

A Each vertical bar along the x-axis represents the mean coverage for a particular biopsy with the tumor tissue represented in light green and the matched normal (white blood cells isolated from whole blood) in green. Biopsies are ordered according to coverage calculated by GATK tools. The horizontal black line represents the average mean coverage for tumor tissue samples and the dotted black line representing the average mean coverage for matched blood. The average tumor tissue coverage was 289X and 126X for matched blood. **B** The individual biopsies are presented along the x-axis and grouped by type (diagnostic DLBCL (without relapse), diagnostic rrDLBCL, and relapsed rrDLBCL). Presented as a heatmap is the main mutational signature for each biopsy as calculated by deconstructSigs R package, based on work by Alexandrov and colleagues (23). S1 is an age related signature, S19 is an unknown aetiology, and S15 is associated with defective DNA mismatch repair. The coding variant counts per biopsy are displayed as bars, with light blue representing missense mutations, dark blue nonsense mutations, light green indels, and dark green splicing mutations. The average coding mutational burden is 125 variants per biopsy.

Supplemental Figure 6

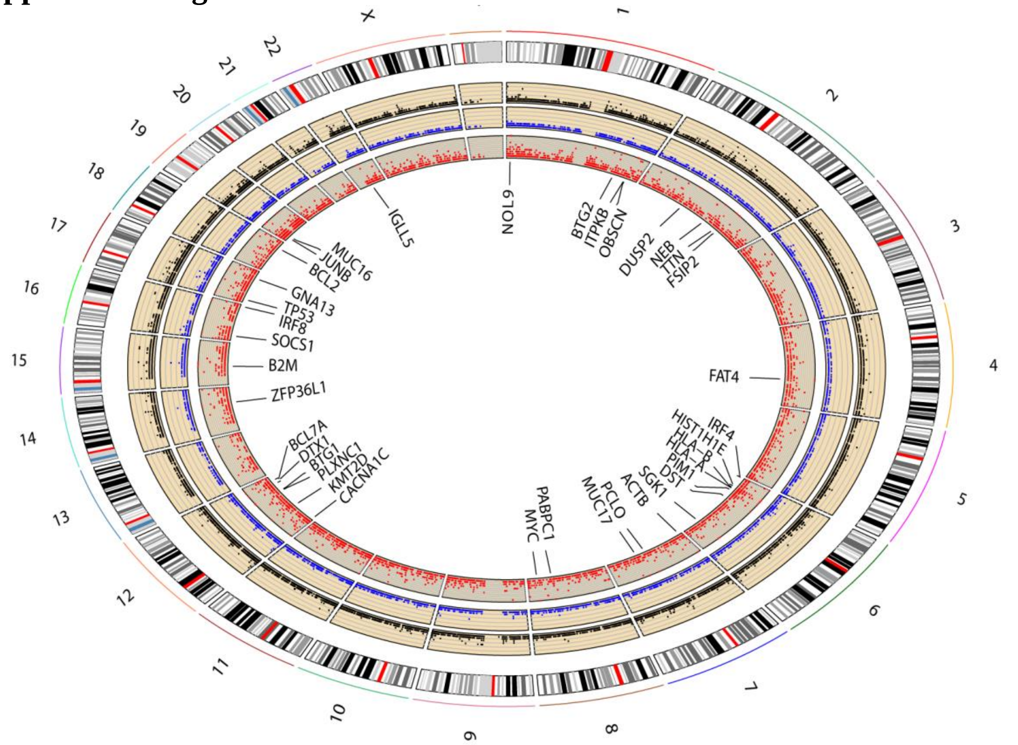


Supplemental Figure 6. NMF Mutational Spectra in the WES with Matched Normal Cohort.

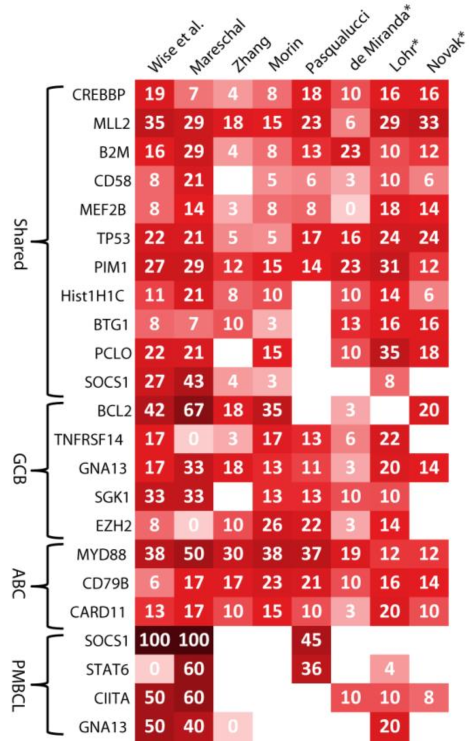
A The distribution of nucleotide substitutions types for each biopsy in the WES with matched normal cohort. As expected, all the tumors showed enrichment in C:G > T:A nucleotide changes. **B** The main mutational signatures for each biopsy as calculated by non-negative matrix factorization of the mutational spectra of total variants with a k = 4 per biopsy, based on work by Lawrence and colleagues(24). Inset are the lego plots (an organization of the 96 possible mutations into six blocks which are subdivided into possible pairs of 5' and 3' neighbors with height corresponding to the mutation frequency) four mutational spectra best fit to the DLBCL data: MSI, Aging, APOBEC and a miscellaneous flat signature that has a mixture of similarities to known mutational signatures. The colors of the mutational spectra labels correspond to the histogram coloring. Paired samples have brackets surrounding the columns and names for grouping. **C** Lego plots of the mutational spectra for selected biopsies with differing spectra based on NMF analysis.

Supplemental Figure 7

A

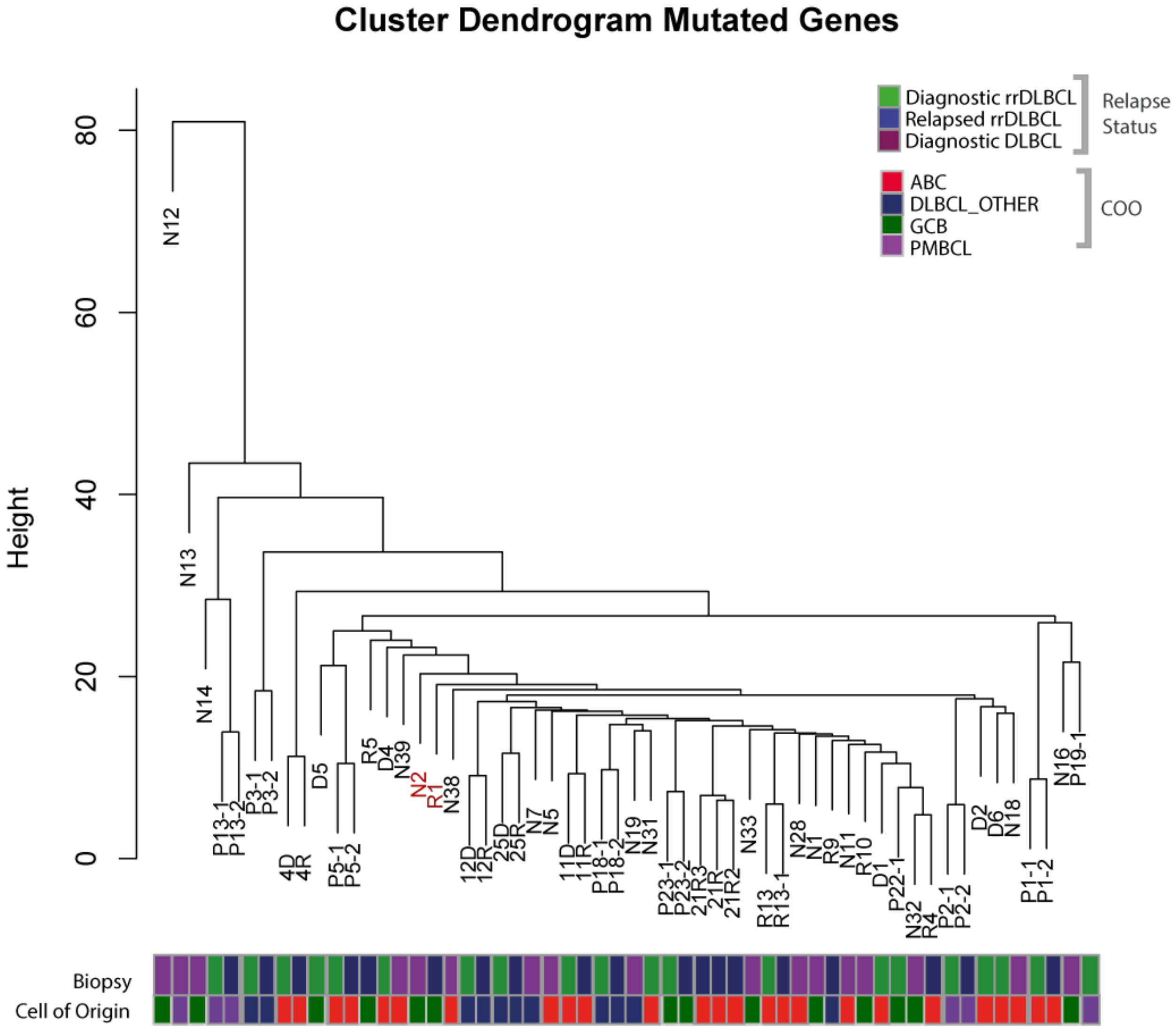


B



		Wise et al.	Mareschal	Zhang	Morin	Pasqualucci	de Miranda*	Lohr*	Novak*
Shared	CREBBP	19	7	4	8	18	10	16	16
	MLL2	35	29	18	15	23	6	29	33
	B2M	16	29	4	8	13	23	10	12
	CD58	8	21		5	6	3	10	6
	MEF2B	8	14	3	8	8	0	18	14
	TP53	22	21	5	5	17	16	24	24
	PIM1	27	29	12	15	14	23	31	12
	Hist1H1C	11	21	8	10		10	14	6
	BTG1	8	7	10	3		13	16	16
	PCLO	22	21		15		10	35	18
	SOCS1	27	43	4	3			8	
GCB	BCL2	42	67	18	35		3		20
	TNFRSF14	17	0	3	17	13	6	22	
	GNA13	17	33	18	13	11	3	20	14
	SGK1	33	33		13	13	10	10	
	EZH2	8	0	10	26	22	3	14	
ABC	MYD88	38	50	30	38	37	19	12	12
	CD79B	6	17	17	23	21	10	16	14
	CARD11	13	17	10	15	10	3	20	10
PMBCL	SOCS1	100	100			45			
	STAT6	0	60			36		4	
	CIITA	50	60				10	10	8
	GNA13	50	40	0				20	

C

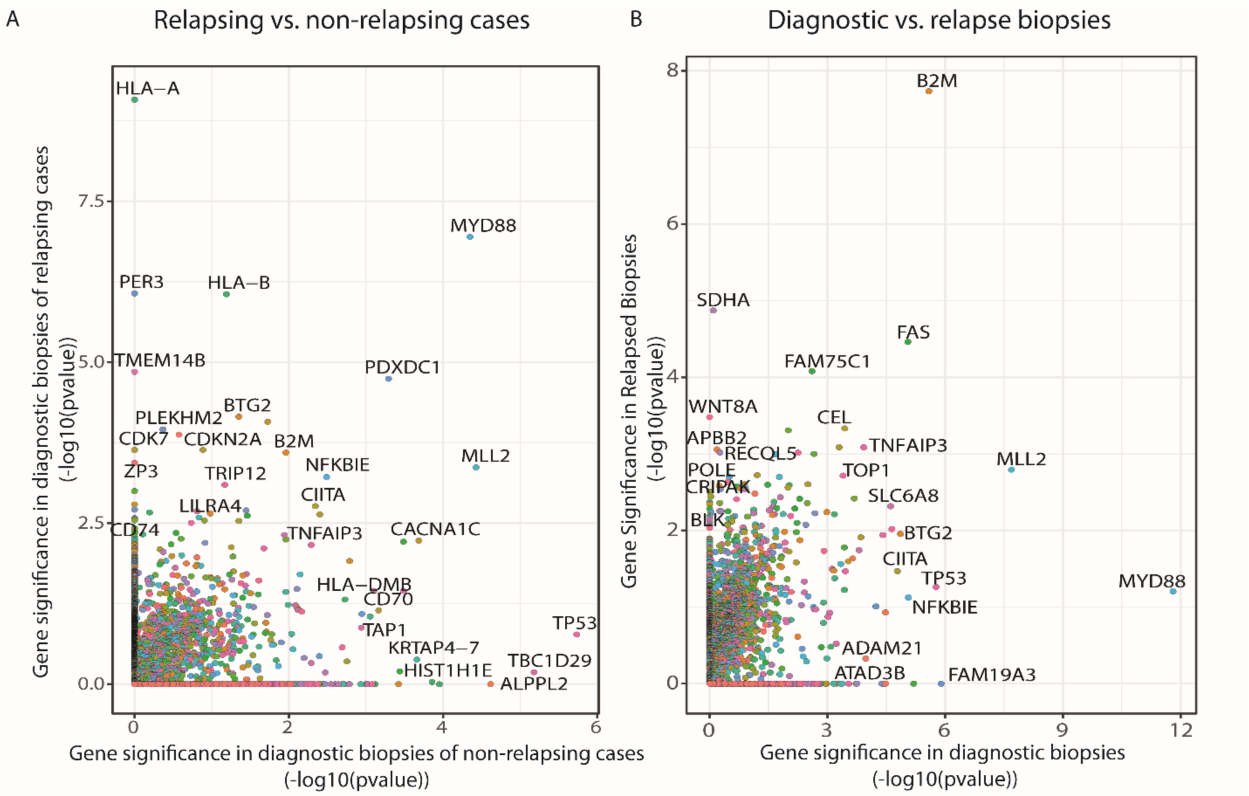


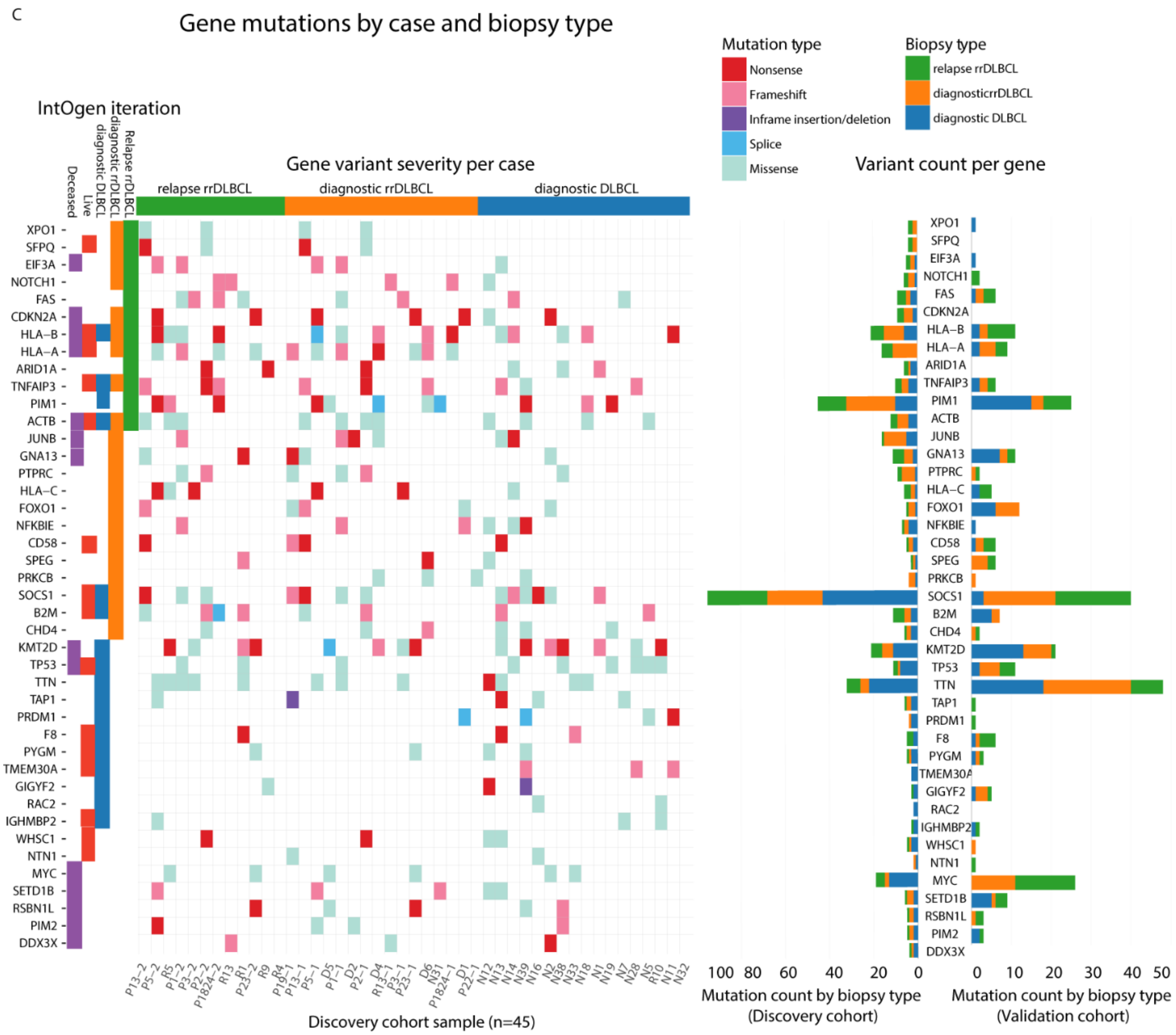
Supplemental Figure 7. Genetic Landscape of the Nordic DLBCL WES Cohort.

A A circos plot of the somatic variants found in the Nordic DLBCL WES with matched normal cohort with the most commonly mutated genes reported. The outer edge is a plot of the human chromosomes including a chromosomal ideogram. The total variant burden is represented by the black outer scatterplot, with the recurrence of each variant being displayed on the diameter axis. After filtering synonymous, UTR and non-coding mutations along with a manual review of low confidence mutation calls, the resulting coding variants are represented by the center scatter plot in blue. The inner red scatterplot represents somatic mutation recurrence of individual genes, with each gene having been assigned a unique chromosomal position. Any with over 10 coding variants in our cohort is listed in the inner circle. **B** A heatmap is displaying frequencies of commonly reported DLBCL variant genes in our WES with matched normal cohort in comparison to seven separate investigations (25-31) into DLBCL exome sequencing variants. *Three of the reports did not report frequency by subtype and thus for each gene the frequency is of the total reported cohort.

C A cluster dendogram of the mutation status (0 or 1) of the 6705 genes with coding variants in the cohort. Calculated with the complete linkage clustering based on Euclidean distance (R version 3.4.0). There is not a distinctive cluster of subtypes, however the two double hit lymphomas appear to cluster closely.

Supplemental Figure 8

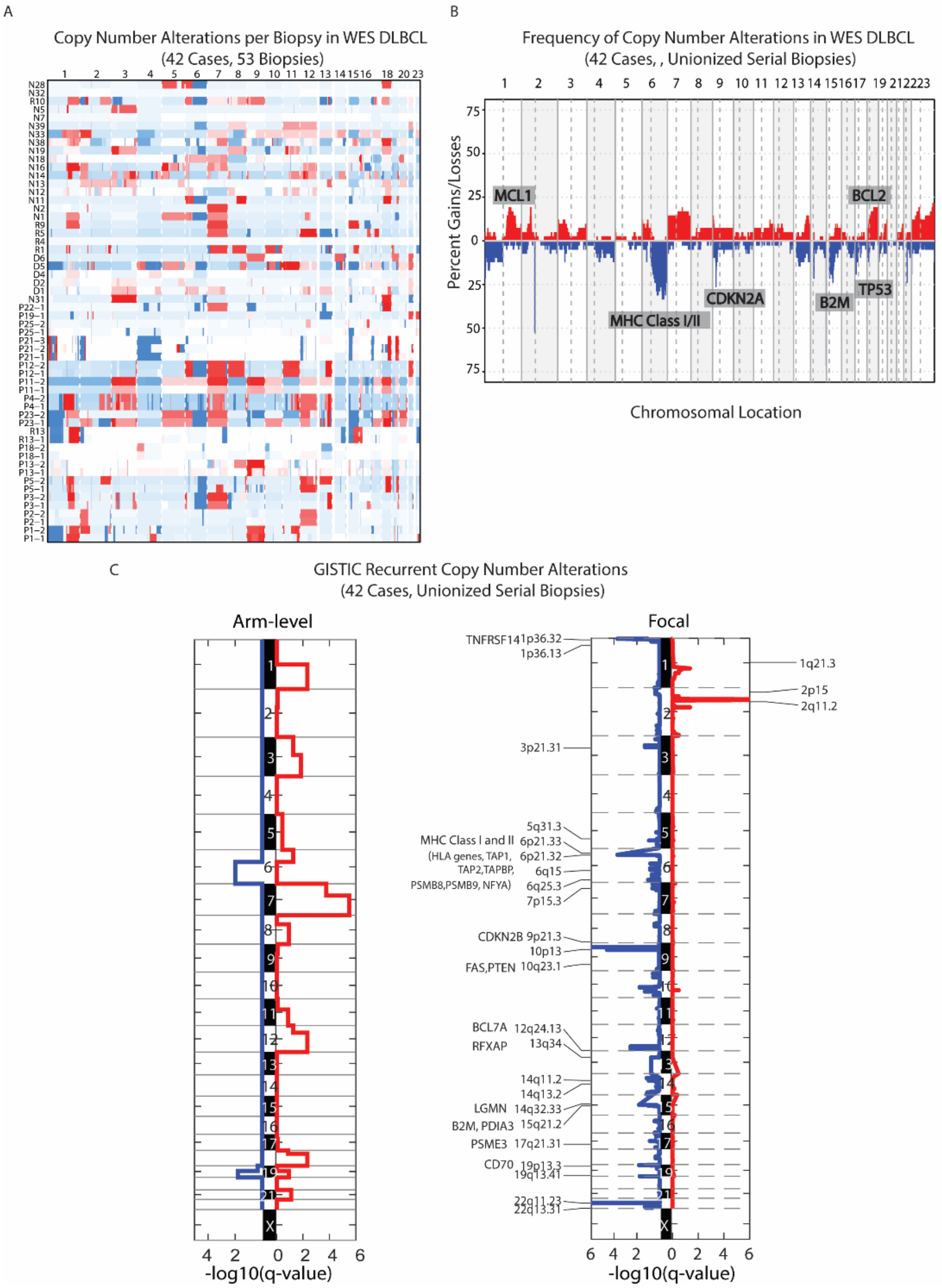




Supplemental Figure 8. Genes Significantly Mutated and Cancer Drivers in DLBCL by Relapse Status.

Significantly mutated genes were identified by MutSig2CV as mutated more often than expected by chance in the Nordic DLBCL WES with matched normal cohort. The software was run in multiple iterations on cohorts representing biopsy type (Supplemental Table 6). Gene significance was expressed as negative $\log_{10}$ of the p-value. Two comparisons are shown: **A** diagnostic biopsies of relapsing vs. non-relapsing cases; and **B** in relapsing cases, relapsed vs. diagnostic biopsies. **C** Driver genes enriched in coding mutations were identified using IntOGen's OncodriveCLUST and OncodriveFM software. Analysis was performed on nine biopsy subsets stratified by relapse status, resulting in 103 identified potential driver genes (Supplemental Table 7). Listed to the left are the 42 driver genes predicted by OncodriveFM in five of these cohorts: diagnostic biopsies of rrDLBCL, relapsed biopsies of rrDLBCL, non-relapsed diagnostic biopsies, live, and deceased patients. The first five columns indicate the cohort(s) in which the gene was found to be significantly mutated. The most severe variant is plotted per biopsy per gene. Biopsies are grouped by type. The histograms to the far right represent the number of variants per gene, separated by biopsy type (colors), and by WES with matched normal cohort (leftward bars) vs. validation cohort (rightward bars).

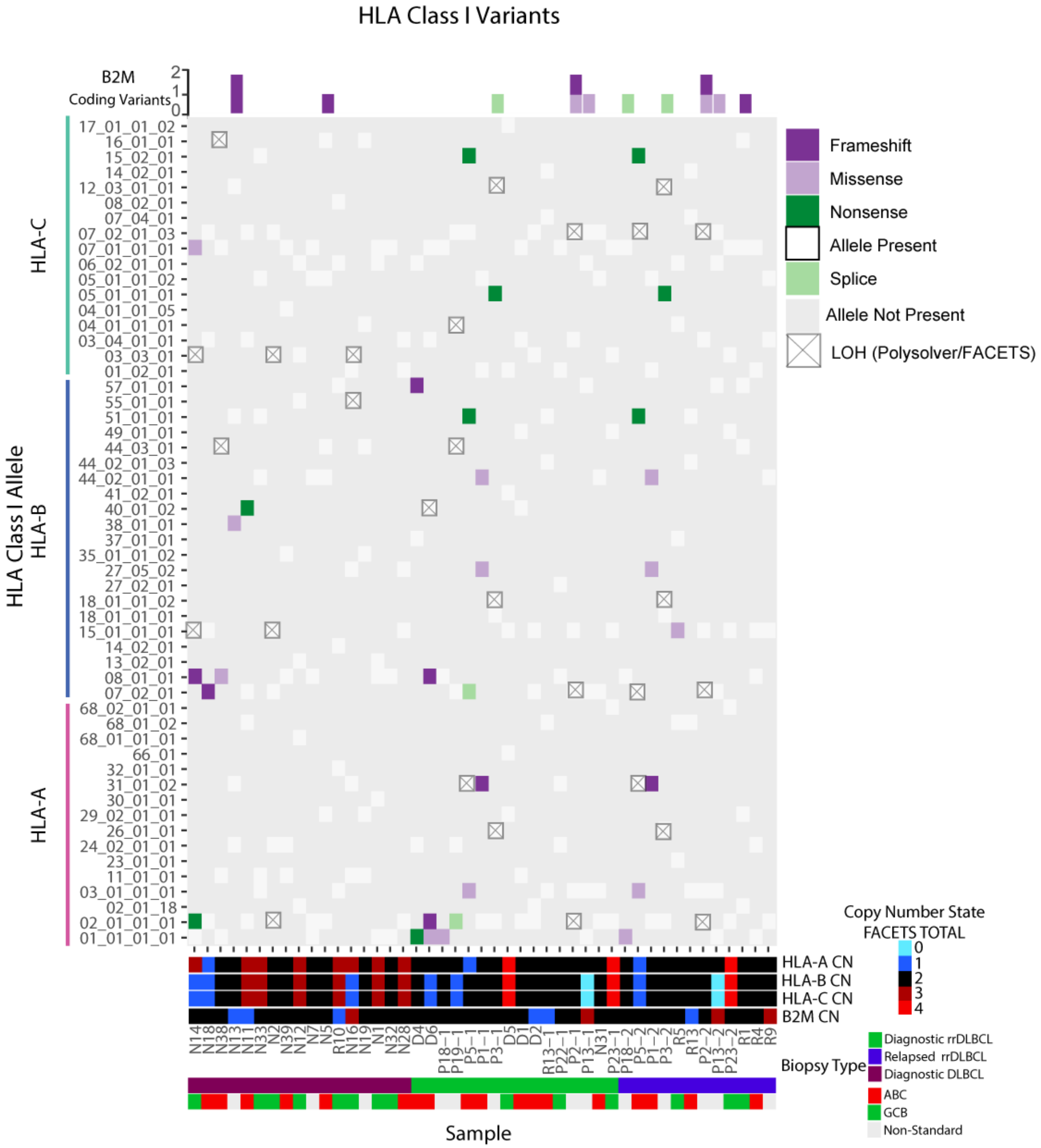
Supplemental Figure 9



Supplemental Figure 9. Copy Number Alterations in DLBCL.
A A heatmap of the copy number states of each segmented copy number file for each biopsy with each row representing a biopsy and the genome laid across the x-axis. Red represents gains over 0.5 copy number state while blue represents losses of less than -0.5 (logCNstates) **B** Copy Number Alteration Frequency plot of DLBCL cases. Cases with multiple biopsies were unionized (for regions with differing copy number states the greatest absolute value of the copy

number state per biopsy was chosen for the case copy number state). A change of above or below .4/-.4 is reported. The chromosomal location of the gain/loss is reported along the x-axis with chromosomes altering blue/white color and grey dotted lines representing the p/q arm distinction. **C** GISTIC2.0-defined recurrent arm level and focal level, respectively, losses (blue) and gains(red). Q-values are reported along the x-axis with respect to alterations for all markers over the entire region analyzed. Chromosomes label the center vertical axis. Focal alteractions identified as recurrent by a q-value of less than .25 are labeled with their cytoband/arm and genes of interest.

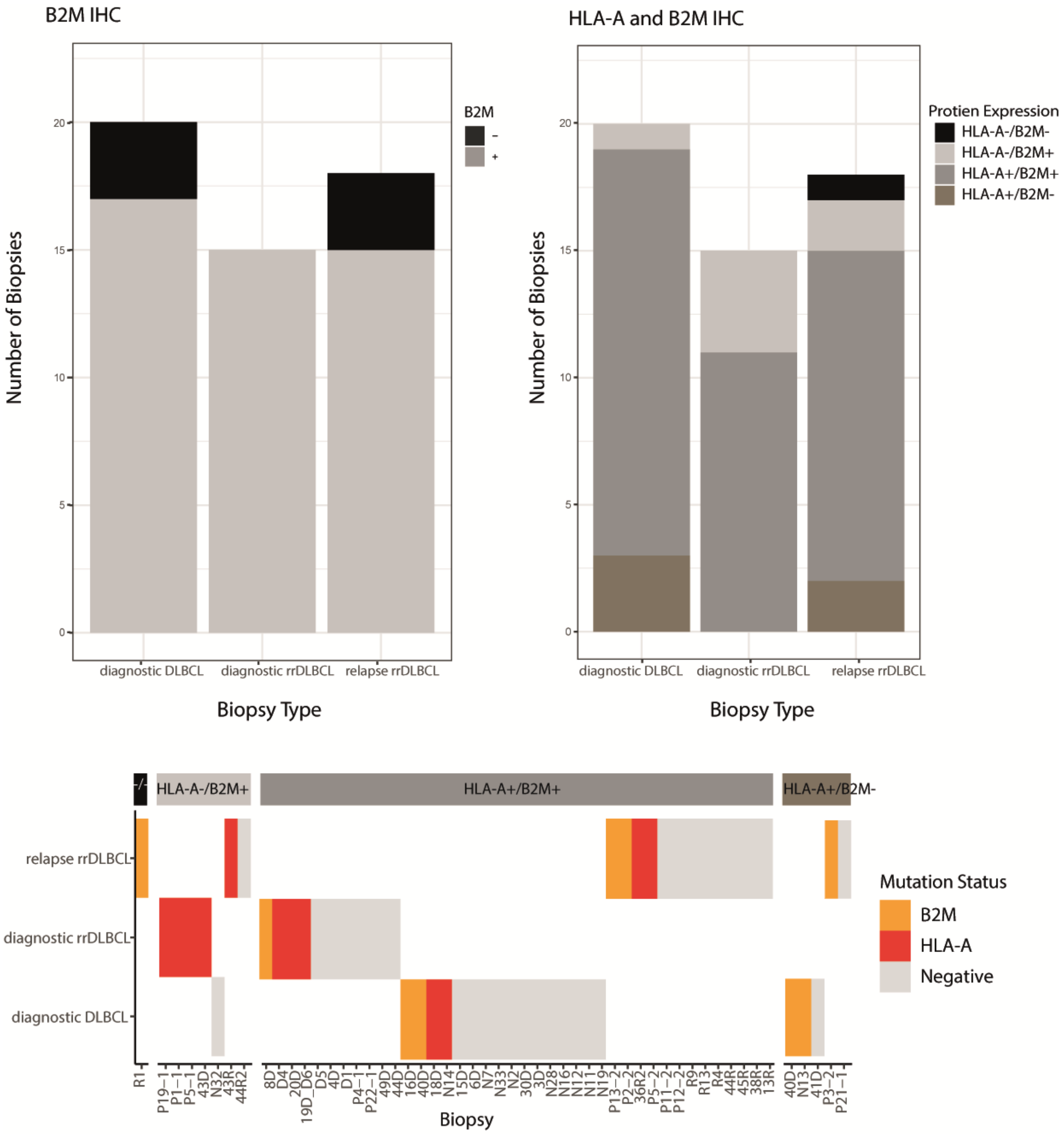
Supplemental Figure 10



Supplemental Figure 10. MHC Class I Genomic Alterations.
Distribution of HLA alleles and mutations across biopsies of DLBCL. The y-axis of the heatmap represents all inferred MHC Class I alleles in the DLBCL WES with matched normal cohort. A tile of white indicates the HLA allele was inferred for the respective biopsy along the x-axis. Mutated HLA alleles are represented by various colored tiles corresponding to the severity of the mutation. If FACETS copy minor copy number state was zero and POLYSOLVER was unable to identify the same HLA alleles in the normal and tumor than an X is marked for loss of heterozygosity in the biopsy. B2M variants per biopsy are reported as a histogram above the

main heatmap. Total copy number states for HLA-A, HLA-B, HLA-C and B2M as defined by FACETS is reported in heatmaps per biopsy below the main HLA inference heatmap. Biopsies are arranged along the x-axis by relapse status, with crimson representing cases that have a durable response to R-CHOP, green representing diagnostic biopsies of those that proceed to progress or relapse and blue representing post-treatment relapsed biopsies. ABC, GCB or other subtypes (PMBCL, PTPLD, EBV, etc.) are also depicted.

Supplemental Figure 11

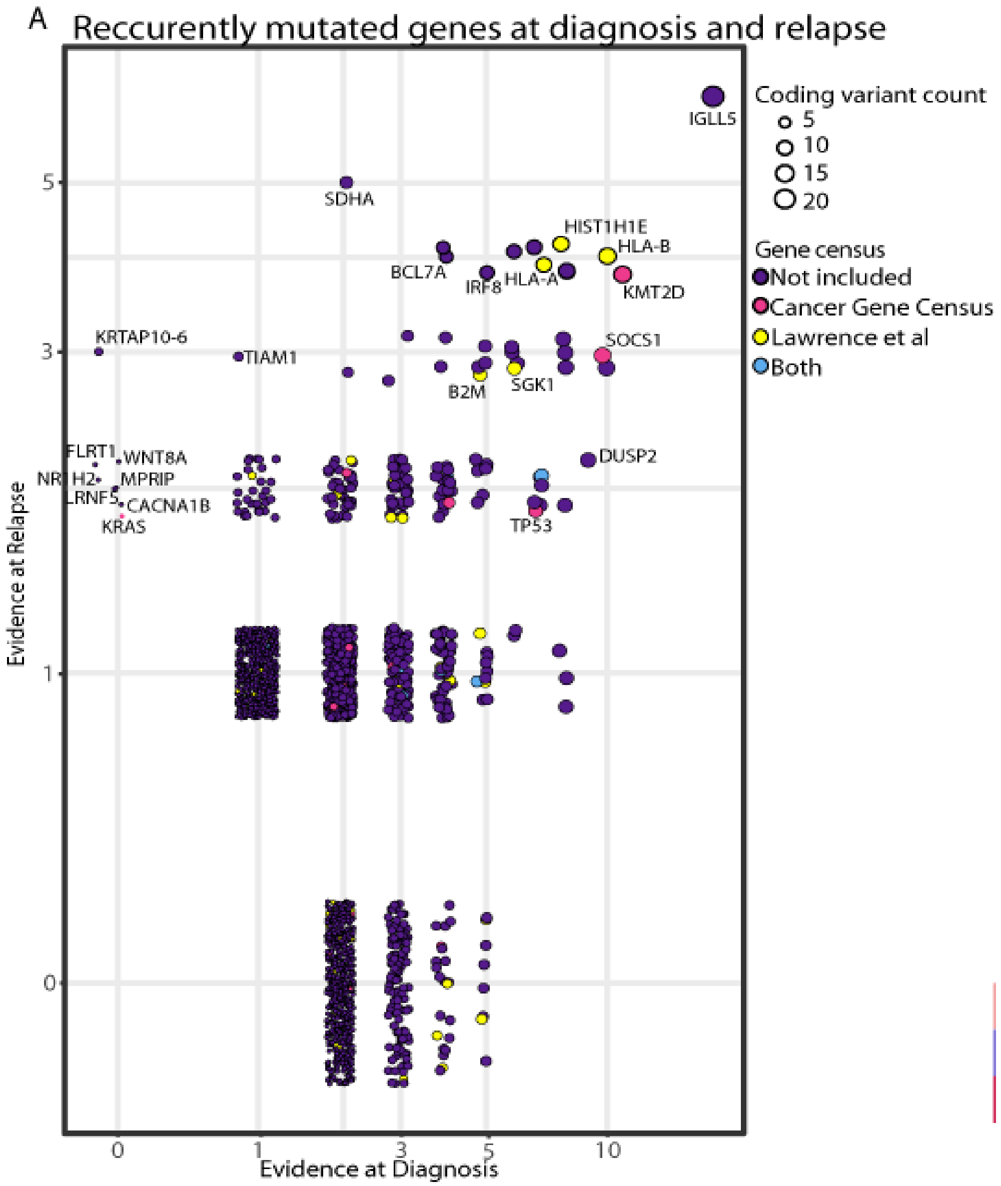


Supplemental Figure 11. B2M IHC and correlates to HLA-A alterations

Distribution of B2M protein staining across biopsy types (top left panel). Protein expression of HLA-A and B2M combinations across biopsy types (top right panel). The correlation of the four staining combination with mutation status of each biopsy is illustrated in the bottom panel.

Biopsy types are clustered and represented along the vertical axis. Each column represents a biopsy which are clustered by the four protein expression combinations. The histogram represents the HLA-A and B2M mutation status of each biopsy. The protein combination group with HLA-A positivity and B2M negativity frequently have B2M mutations.

Supplemental Figure 12

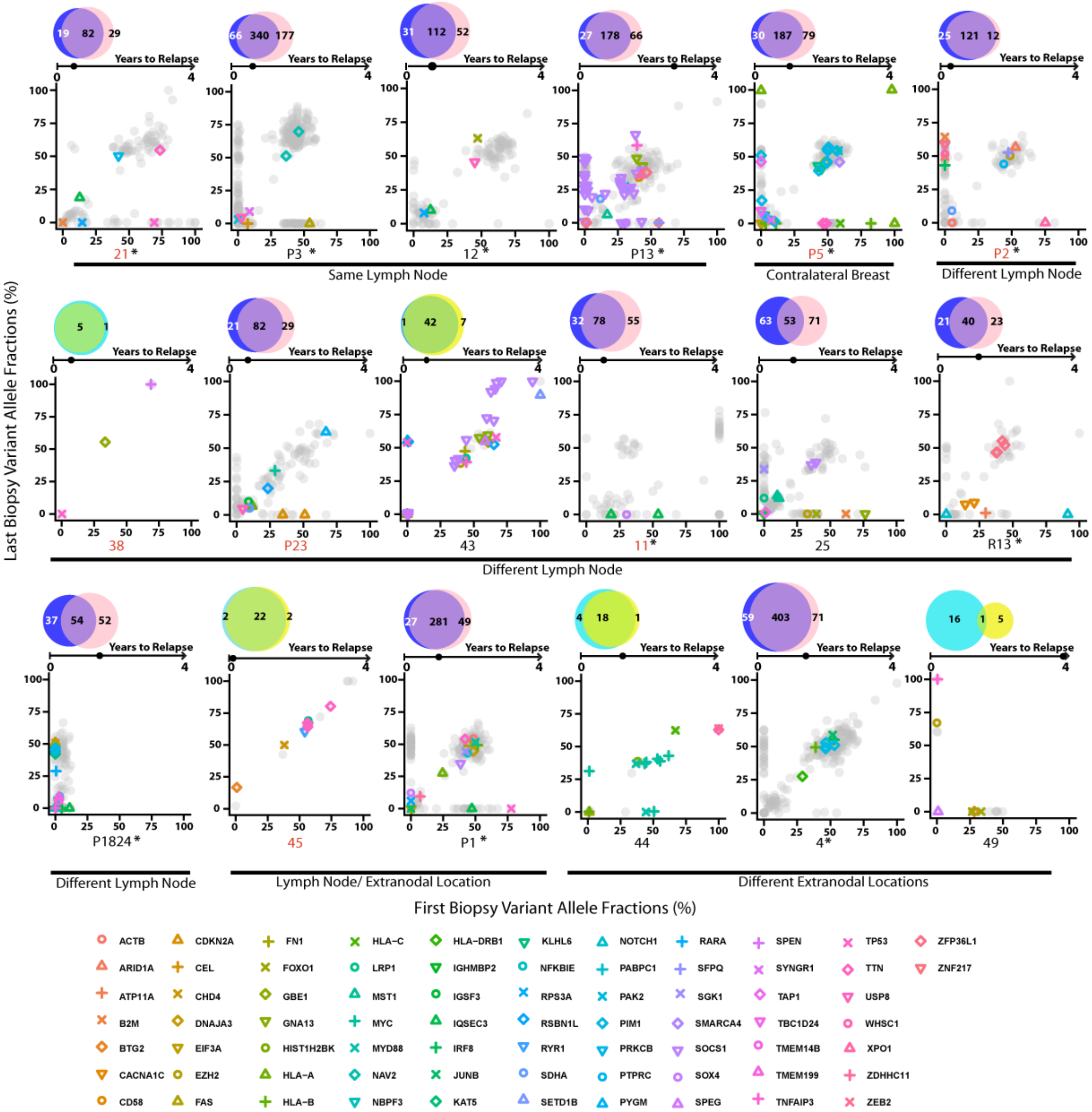


Supplemental Figure 12. Recurrent mutations associated with relapse status.

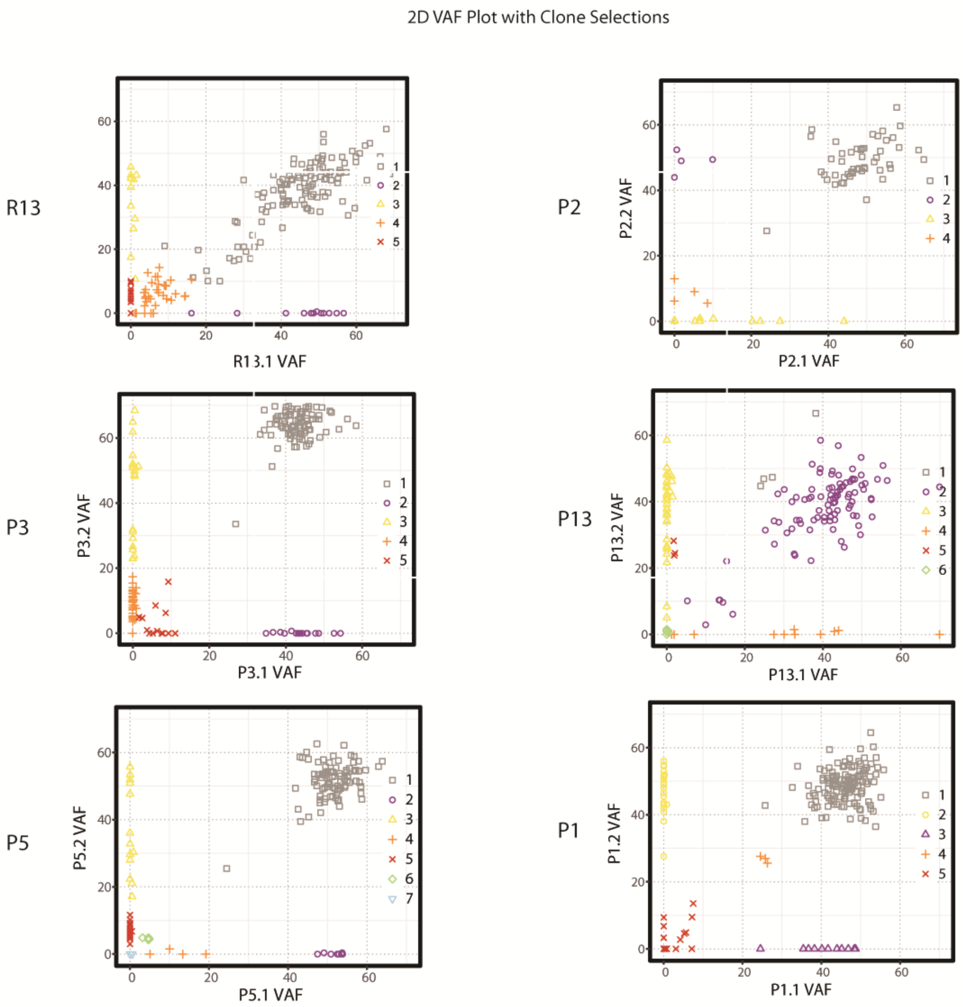
Genes with coding variants in at least two cases from the WES with matched normal cohort, are plotted by total coding variant count in diagnostic versus relapsed biopsies. Counts are reported by case (serial biopsies are unified). Point size indicates total number of coding variants. Point color indicates gene inclusion in the Cancer Gene Census (COSMIC) or Lawrence et al.(32)

Supplemental Figure 13

A



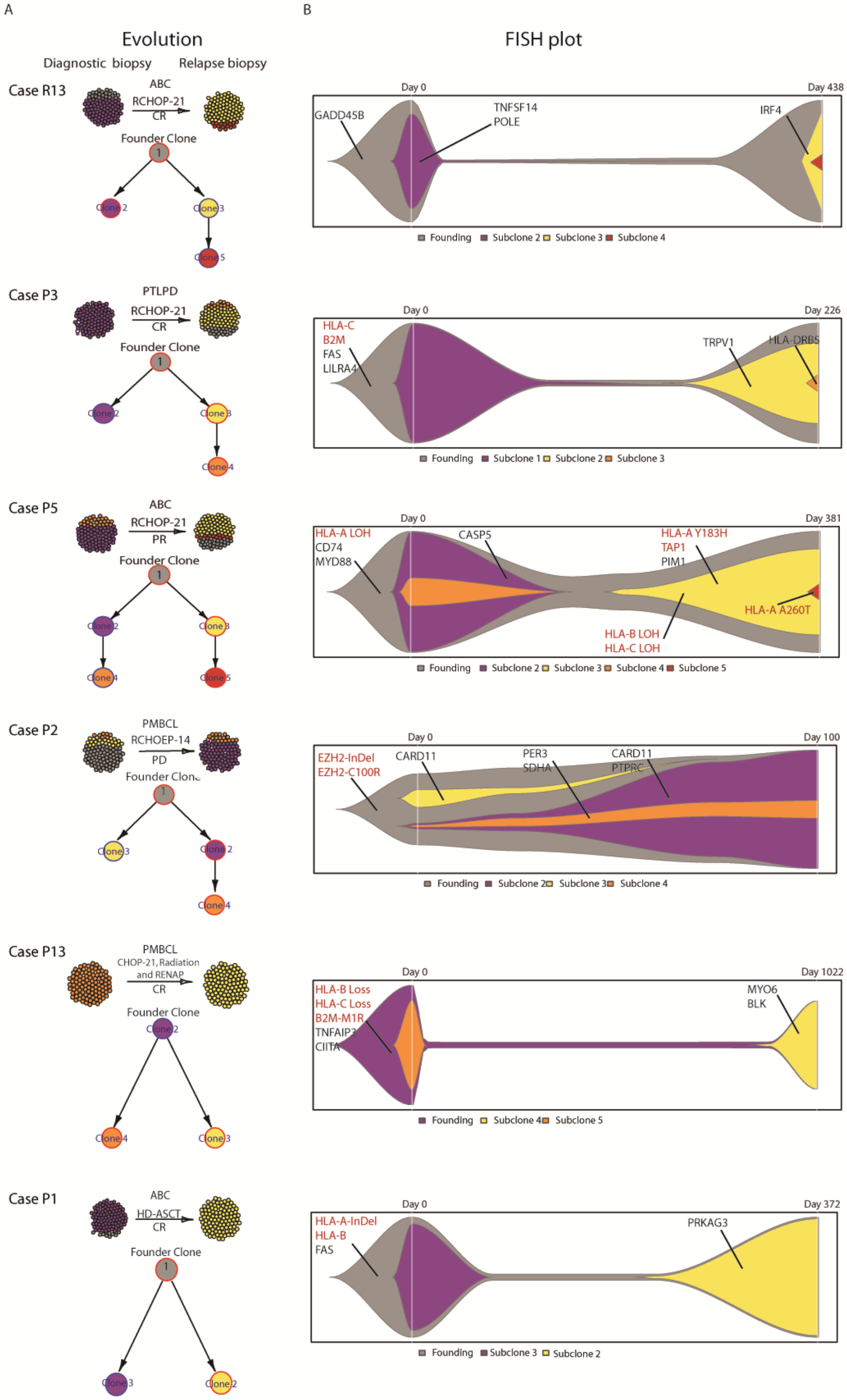
B



Supplemental Figure 13. 2-dimensional Clustering of VAFs and clonal dynamics tracked by somatic mutations in serial rrDLBCL biopsies.

A For each case of rrDLBCL, scatterplots show the normalized variant allele fractions (VAF) of all coding mutations. VAFs were adjusted for tumor purity (those calculated via SNP6.0 are indicated with an asterisk). Mutations were filtered to remove those in regions with copy number alterations (CNAs) and/or loss of heterozygosity (LOH). IntOGen-predicted driver gene mutations are highlighted with unique color/symbol combinations; the remaining mutations are displayed in gray. A timeline is displayed with a point indicating the time of relapse. Venn diagrams of mutational overlap between biopsies are shown (purple for initial and pink for relapse biopsy). Cases from the validation cohort are displayed in blue and yellow and cover genes of the validation panel. Case numbers displayed in black indicate complete response (CR) between biopsies, and those in red represent stable disease or partial response. Plots are divided by biopsy extraction site such as same or different lymph node site. **B** 2-dimensional clustering of VAFs was performed on serial biopsies using SciClone software. Manual review of clones was performed for input into evolutionary analysis.

Supplemental Figure 14



Supplemental Figure 14. Evolutionary progression of serial rrDLBCL cases.

Two-dimensional clustering of VAFs and evolutionary analysis was performed. (**A**) A schematic image of each biopsy is shown with inferred clonal population percentages and branching evolutionary tree. Clinical information including subtype and treatment are displayed (**B**) "Fish plots" illustrate the clonal architecture of the two tumor biopsies, inferred from clonal fractions of the clusters and response to treatment in clinic. Mutations in genes of interest are shown in clones where they occurred with genes in the antigen presentation pathway highlighted in red.

1. Hans CP, Weisenburger DD, Greiner TC, Gascoyne RD, Delabie J, Ott G, et al. Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. Blood. 2004;103(1):275-82.
2. Choi WW, Weisenburger DD, Greiner TC, Piris MA, Banham AH, Delabie J, et al. A new immunostain algorithm classifies diffuse large B-cell lymphoma into molecular subtypes with high accuracy. Clinical cancer research : an official journal of the American Association for Cancer Research. 2009;15(17):5494-502.
3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
4. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.
5. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology. 2013;31(3):213-9.
6. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14):1811-7.
7. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun. 2015;6:10001.
8. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
9. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004;32(Database issue):D115-9.
10. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011;39(Database issue):D945-50.
11. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):D222-30.
12. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862-8.
13. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.
14. Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, et al. DoCM: a database of curated mutations in cancer. Nature methods. 2016;13(10):806-7.
15. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
16. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 2013;41(6):e67.
17. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44(16):e131.
18. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome research. 2007;17(11):1665-74.
19. Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(39):16910-5.
20. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome biology. 2016;17:31.

21. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS computational biology. 2014;10(8):e1003665.
22. Miller CA, McMichael J, Dang HX, Maher CA, Ding L, Ley TJ, et al. Visualizing tumor evolution with the fishplot package for R. BMC genomics. 2016;17(1):880.
23. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-21.
24. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-8.
25. Pasqualucci L, Trifonov V, Fabbri G, Ma J, Rossi D, Chiarenza A, et al. Analysis of the coding genome of diffuse large B-cell lymphoma. Nature genetics. 2011;43(9):830-7.
26. Mareschal S, Dubois S, Viailly PJ, Bertrand P, Bohers E, Maingonnat C, et al. Whole exome sequencing of relapsed/refractory patients expands the repertoire of somatic mutations in diffuse large B-cell lymphoma. Genes Chromosomes Cancer.55(3):251-67.
27. Zhang J, Grubor V, Love CL, Banerjee A, Richards KL, Mieczkowski PA, et al. Genetic heterogeneity of diffuse large B-cell lymphoma. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(4):1398-403.
28. 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):1256-65.
29. de Miranda NF, Georgiou K, Chen L, Wu C, Gao Z, Zaravinos A, et al. Exome sequencing reveals novel mutation targets in diffuse large B-cell lymphomas derived from Chinese patients. Blood. 2014;124(16):2544-53.
30. 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 of the United States of America. 2012;109(10):3879-84.
31. Novak AJ, Asmann YW, Maurer MJ, Wang C, Slager SL, Hodge LS, et al. Whole-exome analysis reveals novel somatic genomic alterations associated with outcome in immunochemotherapy-treated diffuse large B-cell lymphoma. Blood Cancer J. 2015;5:e346.
32. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505(7484):495-501.