Molecular attributes underlying central nervous system and systemic relapse in diffuse large B-cell lymphoma

Keren Isaev,[1] Daisuke Ennishi,[2] Laura Hilton,[3] Brian Skinnider,[2] Karen L. Mungall,[4] Andrew J. Mungall,[4] Mehran Bakthiari,[1] Rosemarie Tremblay-LeMay,[5] Anjali Silva,[1] Susana Ben-Neriah,[2] Merrill Boyle,[2] Diego Villa,[2] Marco A. Marra,[4] Christian Steidl,[2] Randy D. Gascoyne,[2] Ryan D. Morin,[3] Kerry J. Savage,[2] David W. Scott[2#] and Robert Kridel[1#]

[1]Princess Margaret Cancer Center - University Health Network, Toronto, Ontario; [2]Center for Lymphoid Cancer, Britsh Columbia Cancer, Vancouver, British Columbia; [3]Simon Fraser University, Burnaby, British Columbia; [4]Canada's Michael Smith Genome Sciences Center, British Columbia Cancer, Vancouver, British Columbia and [5]Laboratory Medicine Program - University Health Network, Toronto, Ontario, Canada.

[#]DWS and RK contributed equally as co-senior authors.

Correspondence: ROBERT KRIDEL - robert.kridel@uhn.ca

doi:10.3324/haematol.2020.255950

Supplemental Methods

Exome sequencing, detection of single nucleotide variants and clonal analysis

Ten formalin-fixed and paraffin-embedded (FFPE) tissue DNA samples from five patients were submitted to the BC Genome Sciences Centre for library construction, Agilent SureSelect V6 exome capture and Illumina sequencing (Supplemental Table 1). Single nucleotide variants were called by Mutect2[1] (version 2.1) using tumor only mode. Hg19 coordinates in VCF files were left-normalized using bcftools (samtools v1.9).[2] Single nucleotide variants (SNVs) and indels were annotated by Annovar using the following datasets: ensGene, gnomad211, cosmic68 and avsnp142 (Version date 16 Apr 2018).[3] Variants marked as population variants (dbSNP) were used further to infer copy number status around these regions using CNVkit.[4] Additional filters were applied to obtain somatic variants. This included a minimum depth of 100 for all but one sample with low coverage (LY_CNSrel_003_T1, minimum depth of 30 was required for this sample), variant allele frequency greater than 0.1 as well as not being annotated in dbSNP or having a population allele frequency > 0.001. Variants mapping to sex chromosomes were excluded. CNVkit was used in tumor only mode with a flat reference to infer copy number changes in these ten samples using BAM files. Discrete copy number segments were first inferred using the haar algorithm.[5] VCF files with variants marked as potential germline SNPs were used in the "call" function to infer allele-specific copy number.

Copy number segments were merged with annotated somatic variants using Bedtools (v2.27.1)[6] to generate the input data for PyClone (v0.13.1)[7] that includes read counts, allele-specific copy number status and tumor content. Paired analysis of clonal evolution was performed using Pyclone's multisample mode. If a variant was identified as a somatic mutation in only one of the two samples, bam-readcount (https://github.com/genome/bam-readcount) was used to retrieve read counts for that position. Variants with no read coverage in one of the two paired samples were discarded for this analysis. Variants in the *IGH* locus were also removed and the final list included only those mutations in copy neutral regions (total copy number status = 2). PyClone was run on each pair using default parameters, in addition to tumor content, as determined by examination of corresponding H&E sections. Mutations were then clustered based on inferred cellular prevalence. To increase the confidence of clustering, we removed mutations that formed a cluster on their own, as well as mutations with a standard deviation of cellular prevalence > 0.15. PyClone was re-run using this new set of mutations. Citup[8] was used to infer clone phylogenies based on mutation cellular frequencies obtained from PyClone. The maximum number of nodes was set to the one obtained by the final number of clusters from PyClone. The R package *timescape* (https://github.com/maiasmith/timescape) was used to visualize clonal dynamics and clonal phylogenies.

Identification of gene mutations associated with CNS or systemic relapse

In order to identify associations between specific gene mutations and the three above-mentioned clinical risk groups, we compiled a partially overlapping dataset with mutation data of 45 genes in 223 diagnostic samples (n = 72 with CNS relapse, n = 62 with systemic relapse and n = 89 without relapse). This dataset was derived by combining our own mutation data[9] with two publicly available datasets (Reddy *et al.*[10] and Klanova *et al.*[11]). Enrichment across clinical groups was evaluated with a Bayesian implementation of the proportion test, using the BayesianFirstAid R package (v0.1).[12] To that effect, the bayes.prop.test function was run with default parameters.

Bayesian data analysis was used as it has several advantages over traditional null hypothesis significance testing, given that it allows to describe the uncertainty of parameter estimates and that it does not rely on arbitrary *P* value cut-offs.[13]

Case selection for gene expression profiling

Given the relatively infrequent occurrence of CNS relapse, we put emphasis on accruing a cohort of diagnostic samples that was enriched for outcome events. Samples from 222 patients were retrospectively identified from the population-based BC Cancer lymphoma biorepository. Events were considered when they occurred within 1 year following diagnosis, given that CNS relapse is typically reported to occur early.[14–16] Moreover, in this discovery study, we assumed that gene expression profiles were more likely to reveal predictive features of relapse if the time between diagnosis and relapse was short. Hence, patients were selected to fall into 3 different clinical groups: 1) Cases with documented CNS relapse (n=50) within 1 year of diagnosis of *de novo* DLBCL who were treated with R-CHOP (or R-CEOP) (n=39), or those who had concurrent CNS involvement at diagnosis (n=11); 2) Cases with refractory disease or systemic relapse (n=64) within the first year following diagnosis and prior treatment with R-CHOP, but who did not have evidence of CNS involvement at any point in time; and 3) Cases with neither CNS nor systemic relapse for at least 5 years following diagnosis and who had received R-CHOP (n=108). We balanced the 3 clinical groups with regards to cell-of-origin (COO), given that the ABC subtype has been reported as a risk factor for CNS and systemic relapse,[11,17] and is defined by a distinct transcriptional footprint.[18] Regarding CNS prophylaxis, prior to September 2002, intrathecal chemotherapy was recommended for all patients with involvement of specific extranodal sites (bone marrow if involvement by large cell lymphoma, epidural, testes or sinus). After September 2002, this guideline was restricted to only patients with sinus involvement. Beginning in 2013, high dose methotrexate (3.5g/m^2) was recommended for patients with testicular and kidney involvement and select other high-risk patients.

All pathological specimens were centrally reviewed by expert hematopathologists at BC Cancer. All cases were required to yield sufficient RNA from a diagnostic formalin-fixed paraffin-embedded (FFPE) tissue block. COO was determined using the Lymph2Cx assay, as previously described.[19,20] Up to five 10μm tissue sections were cut from FFPE tissue blocks and extracted using the Qiagen AllPrep DNA/RNA FFPE Kit. Gene expression profiling using Affymetrix Human Gene 2.0 ST arrays was performed by The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada.

Gene expression analysis

One sample (CNR7010T1) was an outlier based on average probe intensity and principal component analysis, but was retained in the study as its omission did not alter results in a significant way. We used the *oligo* R package (v1.46.0) to perform background correction and quantile normalization using the Robust Multi-array Average (RMA) method.[21] Multiple probes for a given gene were averaged to provide a single measurement per gene and per sample. The dataset was then filtered (genes retained if at least 5% of samples had an intensity of greater than 1.5, and if the coefficient of variation was between 0.15 and 5). Differential gene expression analysis was performed for each COO subtype (ABC and GCB) using the *limma* R package (v3.42.0).[22] In order to identify and remove latent variation, we applied Surrogate Variable

Analysis using the *sva* R package (v3.34.0), with the number of surrogate variables identified as 1 and 0 for ABC and GCB-DLBCL, respectively (parameters: method = leek, vfilter = 1,000).[23]

Double-hit signature

Calls for the recently published double hit gene expression signature were available for 74 of the 96 germinal centre B-cell-like (GCB) cases, based on the DLBCL90 NanoString assay presented in Ennishi *et al.*[9] For all 96 GCB cases, we also computed double hit signature calls using the Affymetrix dataset and the PRPS package (available from https://github.com/ajiangsfu/PRPS). When comparing to the NanoString-based assay, the accuracy for our calls was 86% and 93%, when including or excluding indeterminate cases, respectively.

Pathway enrichment analysis

For pathway enrichment analysis, we followed the recommendations outlined by Reimand *et al*.[24] For each of the three contrasts (CNS relapse vs. no relapse, systemic relapse vs. no relapse, CNS relapse vs. systemic relapse), we generated a gene list that was ranked by a descending moderated t-statistic and used as input for Gene Set Enrichment Analysis (GSEA, v4.0.3).[25] GSEA was run with default parameters (number of permutations 1000, set size 15-500, enrichment statistic weighted), using the following gene sets: hallmark (h.all.v7.1), curated canonical pathways (c2.cp.v7.1), GO biological processes (c5.bp.v7.1), oncogenic signatures (c6.all.v7.1) and immunologic signatures (c7.all.v7.1). In addition, we supplemented these gene sets with a publicly available list of signatures that are relevant in the lymphoid context.[26]

Data Sharing Statement

Mutation data can be found in Supplemental Table 2. Qualified researchers may obtain access to the original data used in this study. Microarray data have been deposited into the European Genome-phenome Archive (study accession EGAD00010001909). For exome sequencing data, please contact the senior authors of this study.

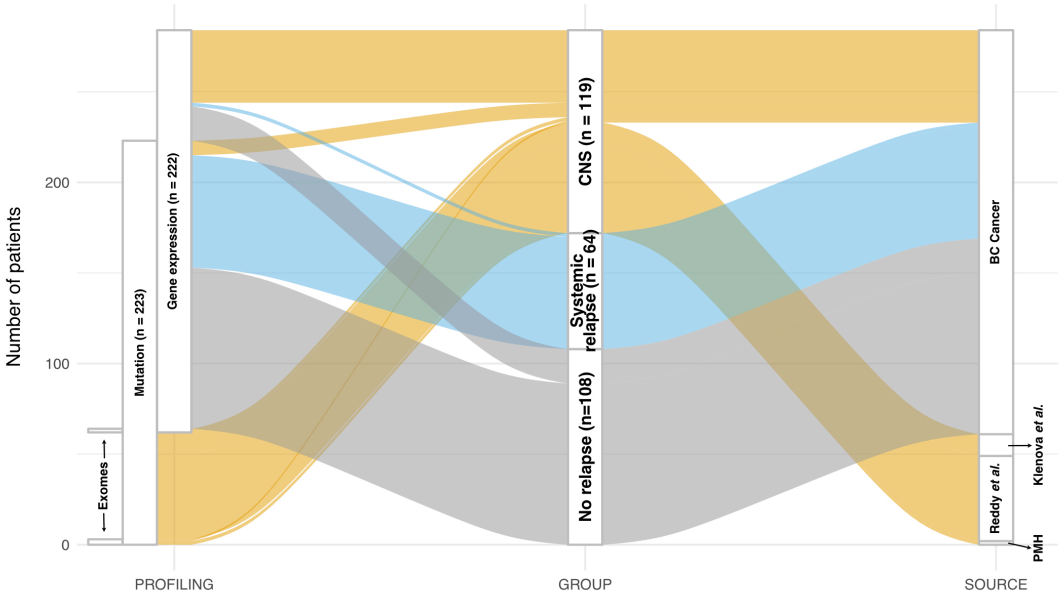
References

1. 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):213–219.
2. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079.
3. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):1–7.
4. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* 2016;12(4):e1004873.
5. Ben-Yaacov E, Eldar YC. A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*. 2008;24(16):139–145.
6. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.
7. Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*. 2014;11(4):396–8.
8. Malikic S, McPherson AW, Donmez N, Sahinalp CS. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*. 2015;31(9):1349–56.
9. Ennishi D, Jiang A, Boyle M, et al. Double-Hit Gene Expression Signature Defines a Distinct

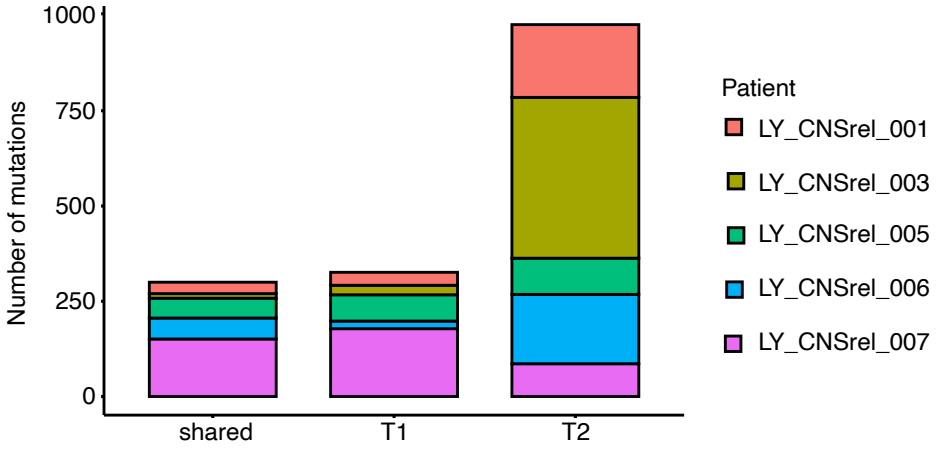
Subgroup of Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma. *J. Clin. Oncol.* 2019;37(3):190–201.
10. Reddy A, Zhang J, Davis NS, et al. Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell*. 2017;171(2):481-494.e15.
11. Klanova M, Sehn LH, Bence-Bruckler I, et al. Integration of cell of origin into the clinical CNS International Prognostic Index improves CNS relapse prediction in DLBCL. *Blood*. 2019;133(9):919–926.
12. Baath R. Bayesian First Aid: A Package that Implements Bayesian Alternatives to the Classical test Functions in R. *UseR 2014 - Int. R User Conf.* 2014;1(2):30–42.
13. Kruschke JK. Bayesian assessment of null values via parameter estimation and model comparison. *Perspect. Psychol. Sci.* 2011;6(3):299–312.
14. Boehme V, Zeynalova S, Kloess M, et al. Incidence and risk factors of central nervous system recurrence in aggressive lymphoma--a survey of 1693 patients treated in protocols of the German High-Grade Non-Hodgkin's Lymphoma Study Group (DSHNHL). *Ann. Oncol.* 2007;18(1):149–57.
15. Boehme V, Schmitz N, Zeynalova S, Loeffler M, Pfreundschuh M. CNS events in elderly patients with aggressive lymphoma treated with modern chemotherapy (CHOP-14) with or without rituximab: an analysis of patients treated in the RICOVER-60 trial of the German High-Grade Non-Hodgkin Lymphoma Study Group (DSHNHL). *Blood*. 2009;113(17):3896–902.
16. Bernstein SH, Unger JM, LeBlanc M, et al. Natural history of cns relapse in patients with aggressive non-hodgkin's lymphoma: A 20-year follow-up analysis of swog 8516-the southwest oncology group. *J. Clin. Oncol.* 2009;27(1):114–119.
17. Savage KJ, Slack GW, Mottok A, et al. Impact of dual expression of MYC and BCL2 by immunohistochemistry on the risk of CNS relapse in DLBCL. *Blood*. 2016;127(18):2182–8.
18. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403(6769):503–11.
19. Scott DW, Wright GW, Williams PM, et al. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood*. 2014;123(8):1214–7.
20. Kridel R, Mottok A, Farinha P, et al. Cell of origin of transformed follicular lymphoma. *Blood*. 2015;126(18):2118–2127.
21. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*. 2010;26(19):2363–7.
22. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.
23. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882–883.
24. Reimand J, Isserlin R, Voisin V, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* 2019;14(2):482–517.
25. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S.*

A. 2005;102(43):15545–50.
26. Schmitz R, Wright GW, Huang DW, et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* 2018;378(15):1396–1407.

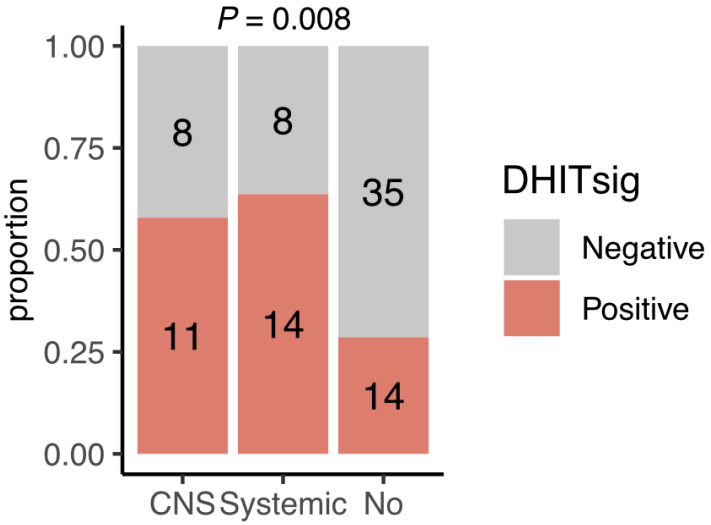
Supplemental Figure 1: Flow diagram illustrating all patients described in this study and their relationship with profiling techniques used, clinical risk groups and the source of samples and/or data.



Supplemental Figure 2: Number of mutations found to be shared, exclusively seen in T1 (i.e. extra-CNS biopsy) or T2 (CNS biopsy).



Supplemental Figure 3: Proportion of cases positive (in salmon color) for the double-hit signature, by clinical risk group.



Supplemental Table 1: Cases used for exome sequencing.

Sample	Source	Site timepoint 1	Site timepoint 2	Treatment between biopsies	Time to CNS relapse (years)
LY_CNSrel_001	UHN	Breast	Brain	R-CHOP	2.78
LY_CNSrel_003	UHN	Lymph node	Brain	CHOP or CHOP-like	5.37
LY_CNSrel_005	BC Cancer	Stomach	Brain	R-CHOP	0.38
LY_CNSrel_006	BC Cancer	Testis	Brain	CHOP or CHOP-like	0.72
LY_CNSrel_007	BC Cancer	Testis	Brain	CNS involvement at diagnosis	0.08

Supplemental Table 2: Mutation data from 223 diagnostic DLBCL samples, by clinical group and source.
(see separate Excel file)

Supplemental Table 3: Patient included in gene expression study, characteristics by clinical group.

	Clinical Group			***P* value**		
	CNS relapse (n=50)	Systemic relapse (n=64)	No relapse (n=108)	CNS vs. systemic relapse	CNS vs. no relapse	Systemic vs. no relapse
Age ≤ 60	19 (38%)	21 (33%)	40 (37%)			
Age > 60	31 (62%)	43 (67%)	68 (63%)	0.693	1.000	0.623
ECOG ≤ 1	12 (24%)	25 (39%)	65 (61%)			
ECOG > 1	38 (76%)	39 (61%)	42 (39%)	0.108	**<0.001**	**0.007**
Stage I-II	8 (16%)	12 (19%)	47 (43.5%)			
Stage III-IV	42 (84%)	52 (81.2%)	61 (56.5%)	0.806	**0.001**	**0.001**
LDH normal	10 (21%)	14 (23%)	55 (56%)			
LDH elevated	37 (79%)	47 (77%)	44 (44%)	1.000	**<0.001**	**<0.001**
Nb extranodal sites 0-1	22 (44%)	48 (75%)	89 (82%)			
Nb extranodal sites > 1	28 (56%)	16 (25%)	19 (18%)	**0.001**	**<0.001**	0.248
IPI low	2 (4%)	7 (11%)	33 (31%)			
IPI intermediate	19 (38%)	27 (42%)	56 (52%)			
IPI high	29 (58%)	30 (47%)	19 (18%)	0.288	**<0.001**	**<0.001**
CNS-IPI low	2 (4%)	6 (10%)	22 (22%)			
CNS-IPI intermediate	16 (34%)	25 (41%)	57 (58%)			
CNS-IPI high	29 (62%)	30 (49%)	20 (20%)	0.352	**<0.001**	**0.001**
Non-double hit	35 (88%)	48 (80%)	95 (91%)			
Double hit	5 (12%)	12 (20%)	9 (9%)	0.420	0.534	0.051
Cell-of-origin ABC	19 (40%)	31 (48%)	36 (33%)			
Cell-of-origin GCB	19 (40%)	25 (39%)	52 (48%)			
Cell-of-origin Unclassified	10 (21%)	8 (12%)	20 (19%)	0.458	0.617	0.148
R-CHOP	34 (68%)	64 (100%)	108 (100%)			
R-CEOP	5 (10%)	0 (0%)	0 (0%)			
CNS invasion at diagnosis	11 (22%)	0 (0%)	0 (0%)	**<0.001**	**<0.001**	1.000

ABC, activated B-cell subtype; CNS-IPI, Central Nervous System - International Prognostic Index; ECOG, Eastern Cooperative Oncology Group; GCB, germinal centre B-cell subtype; LDH, lactate dehydrogenase; IPI, International Prognostic Index.

Supplemental Table 4: Results from gene set enrichment analysis.
(see separate Excel file)