RESEARCH ARTICLE

Patterns of Oncogene Coexpression at Single-Cell Resolution Influence Survival in Lymphoma

Michal Marek Hoppe[1], Patrick Jaynes[1], Fan Shuangyi[2], Yanfen Peng[1], Shruti Sridhar[1], Phuong Mai Hoang[1], Clementine Xin Liu[3], Sanjay De Mel[3,4], Limei Poon[3,4], Esther Hian Li Chan[3,4], Joanne Lee[3,4], Choon Kiat Ong[5], Tiffany Tang[6], Soon Thye Lim[6], Chandramouli Nagarajan[7], Nicholas F. Grigoropoulos[7], Soo-Yong Tan[2,4], Susan Swee-Shan Hue[2,4], Sheng-Tsung Chang[8], Shih-Sung Chuang[8], Shaoying Li[9], Joseph D. Khoury[10], Hyungwon Choi[11], Carl Harris III[12], Alessia Bottos[12], Laura J. Gay[13], Hendrik F.P. Runge[13], Ilias Moutsopoulos[13], Irina Mohorianu[13], Daniel J. Hodson[13], Pedro Farinha[14], Anja Mottok[14], David W. Scott[14], Jason J. Pitt[1,4,15], Jinmiao Chen[16], Gayatri Kumar[17], Kasthuri Kannan[17], Wee Joo Chng[1,3,4,11], Yen Lin Chee[3,4], Siok-Bian Ng[1,2,4], Claudio Tripodo[18,19], and Anand D. Jeyasekharan[1,3,4,11]

Illustration by Bianca Dunn

ABSTRACT Cancers often overexpress multiple clinically relevant oncogenes, but it is not known if combinations of oncogenes in cellular subpopulations within a cancer influence clinical outcomes. Using quantitative multispectral imaging of the prognostically relevant oncogenes *MYC*, *BCL2*, and *BCL6* in diffuse large B-cell lymphoma (DLBCL), we show that the percentage of cells with a unique combination *MYC*+*BCL2*+*BCL6*– (M+2+6–) consistently predicts survival across four independent cohorts ($n = 449$), an effect not observed with other combinations including M+2+6+. We show that the M+2+6– percentage can be mathematically derived from quantitative measurements of the individual oncogenes and correlates with survival in IHC ($n = 316$) and gene expression ($n = 2{,}521$) datasets. Comparative bulk/single-cell transcriptomic analyses of DLBCL samples and *MYC*/*BCL2*/*BCL6*-transformed primary B cells identify molecular features, including cyclin D2 and PI3K/AKT as candidate regulators of M+2+6– unfavorable biology. Similar analyses evaluating oncogenic combinations at single-cell resolution in other cancers may facilitate an understanding of cancer evolution and therapy resistance.

SIGNIFICANCE: Using single-cell–resolved multiplexed imaging, we show that selected subpopulations of cells expressing specific combinations of oncogenes influence clinical outcomes in lymphoma. We describe a probabilistic metric for the estimation of cellular oncogenic coexpression from IHC or bulk transcriptomes, with possible implications for prognostication and therapeutic target discovery in cancer.

INTRODUCTION

Oncogene overexpression is common in cancer. The concomitant increase in oncogenic proteins (oncoproteins) influences both prognosis and treatment (1). Notable examples routinely assessed in clinical practice include HER2 in breast cancer and ALK in lung cancer. However, cancers often overexpress more than one oncogene. Whether multiple oncogenes interact at the single-cell level to influence clinical outcome remains an important unresolved question. This is particularly relevant because cancers are a heterogeneous mosaic of tumor cell subpopulations (2), and oncogenes show clinically significant intratumor heterogeneity (ITH) in expression (1). Clinical techniques for estimating oncogene overexpression in cancer (such as IHC) study them in isolation, and do not provide information on coexpression in subsets of cells within a tumor. It is therefore still not known if subsets of cells within a given cancer expressing specific combinations of oncogenes drive clinical phenotypes.

We aimed to address this question using multiplexed fluorescent IHC (mfIHC), a technique that can simultaneously and quantitatively evaluate a set of proteins with single-cell resolution. This allows measurement of single-cell oncogene coexpression from sufficient samples for robust multivariate correlations with clinical outcomes. We chose diffuse large B-cell lymphoma (DLBCL) as a model to evaluate the clinical impact of ITH in oncogene coexpression. DLBCL is the most common aggressive lymphoma worldwide (3), and overexpression of the oncogenes *MYC*, *BCL2*, and *BCL6* (4–6) influences pathogenesis and prognosis (7–9). However, there is significant variability among studies regarding the prognostic significance of these oncogenes, with debate on appropriate cutoff thresholds to define "positivity." These considerations offer an ideal scenario to evaluate whether these oncogenes show differential coexpression at the single-cell level in DLBCL, and to investigate how they cooperate or influence each other at the cellular level to affect survival.

[1]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore. [2]Department of Pathology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [3]Department of Haematology-Oncology, National University Health System, Singapore, Singapore. [4]NUS Centre for Cancer Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [5]Division of Cellular and Molecular Research, National Cancer Centre Singapore, Singapore, Singapore. [6]Division of Medical Oncology, National Cancer Centre Singapore, Singapore, Singapore. [7]Department of Haematology, Singapore General Hospital, Singapore, Singapore. [8]Department of Pathology, Chi-Mei Medical Center, Tainan City, Taiwan. [9]Department of Hematopathology, Division of Pathology and Laboratory Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas. [10]Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, Nebraska. [11]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [12]F. Hoffmann-La Roche Ltd, Basel, Switzerland. [13]Wellcome MRC Cambridge Stem Cell Institute, Cambridge, United Kingdom. [14]BC Cancer Research Centre, Vancouver, Canada. [15]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore. [16]Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore. [17]Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, Texas. [18]Tumor Immunology Unit, University of Palermo, Palermo, Italy. [19]IFOM ETS – The AIRC Institute of Molecular Oncology, Milan, Italy.

Corresponding Authors: Anand D. Jeyasekharan, Cancer Science Institute of Singapore, Centre for Translational Medicine (MD6) #13-01G, 14 Medical Drive, Singapore 117599, Singapore. Phone: 65-6516-5094; E-mail: csiadj@nus.edu.sg; and Claudio Tripodo, University of Palermo School of Medicine, Istituto di Patologia Generale, Corso Tukory 211, 90134, Palermo, Italy. Phone: 39-091-2389-6211; E-mail: claudio.tripodo@unipa.it

Cancer Discov 2023;13:1144–63

doi: 10.1158/2159-8290.CD-22-0998

RESULTS

Physiologic Patterns of MYC, BCL2, and BCL6 Coexpression Are Disrupted in DLBCL

We first compared the coexpression of the oncogenes *MYC*, *BCL2*, and *BCL6* between reactive lymphoid tissue and DLBCL by mfIHC (Fig. 1A). Consistent with known physiology, BCL6 and BCL2 expression was restricted to reactive lymphoid germinal centers (GC) and extra-GC regions, respectively, whereas MYC showed sparse positivity in the GC CD20$^+$ cells (ref. 10; Fig. 1B; Supplementary Fig. S1A; Supplementary Table S1). A binary ± map for each oncogene facilitates the quantitation of subpopulations of cells based on MYC BCL2 and BCL6 coexpression (Supplementary Fig. S1B). The repertoire of subpopulations defined by MYC/BCL2/BCL6 permutations in reactive lymphoid B cells was limited, with the single-positive M-2–6+ subpopulation being dominant within the GC and predominantly driving proliferative capacity (Supplementary Fig. S1C), and M-2+6– being dominant outside the GC (Fig. 1C). Hardly any cells displayed coexpression of all three oncogenes MYC, BCL2, and BCL6 in either compartment, consistent across several reactive lymphoid tissues analyzed (Fig. 1D; Supplementary Table S1).

In contrast, DLBCL cells frequently coexpress these three oncogenes (Fig. 1E and F; Supplementary Fig. S2; Supplementary Table S2). However, even within cases characterized by high overall levels of MYC, BCL2, and BCL6 expression, these three oncogenes were not always found in the same cells, underscoring ITH in DLBCL. The percentage of each subpopulation was variable between patients but was remarkably similar in overall distribution and clustering among four cohorts (Fig. 1F; Supplementary Fig. S3). The percentages of subpopulations were also stable across different tumor cores from the same patient, indicative of patient-specific subpopulation profiles (Fig. 1G; Supplementary Fig. S4A and S4B). These subpopulations were not consistently associated with clinicopathologic features such as age, gender, and International Prognostic Index (IPI) Risk Group, nor were they associated with MYC/BCL2/BCL6 translocation status (Supplementary Table S3; Fig. 1F), confirming previous observations that translocations do not account for the majority of MYC, BCL2, and BCL6 overexpression in DLBCL (11). Only subpopulations with BCL6 expression (irrespective of the coexpression of other oncogenes) showed Ki-67 expression in two DLBCL cohorts (Fig. 1H). This association was also observed in the context of reactive tonsil tissue, consistent with the role of BCL6 in B-cell proliferation (Supplementary Fig. S1C).

Spatial Interaction of Oncogenic B-cell Subpopulations Is Clustered and Nonrandom

Single-cell–resolved image data with spatial coordinates enable the assessment of spatial interaction patterns of the eight MYC, BCL2, and BCL6 subpopulations. Analyzing spatial subpopulation data of the Singapore General Hospital (SGH) and MD Anderson (MDA) cohorts, we first applied a pair correlation function (PCF; refs. 12, 13), which quantifies how a point (cell) of interest is surrounded by other cells and can investigate whether each subpopulation tends to cluster or show a random (Poisson) distribution (Supplementary Fig. S5A). In immediate neighborhoods—defined here as a range from 0- to 250-μm radius of a given cell—the PCF graphs demonstrate that for both cohorts each subpopulation deviates from Poisson spatial patterning (PCF = 1; Supplementary Fig. S5B). For each subpopulation, PCF is high at small radii, i.e., 10 to 20 μm, indicative of a clustered cell pattern among immediate neighbors. These patterns taper off as the radii increase, i.e., when more cells are considered across wider regions of the tumor. Supplementary Fig. S5C illustrates this visually for a single patient: each subpopulation shows a tendency to group in space within the tissue and does not display a random spatial Poisson distribution (as per random simulation).

To further quantify spatial heterogeneity between subpopulations, we calculated for each cell the percentage deviation (Δ%) of the observed from the expected subpopulation extent (as quantified across whole-tissue available) within the cell's local neighborhood (20 cells). In other words, if cells were distributed randomly in space, the observed abundance of a particular subpopulation in the neighborhood of any given cell would match the overall subpopulation extents measured across a tumor. However, if an over- or underrepresentation of a particular subpopulation occurs in the topological neighborhood of a given cell, this deviation provides a quantitative depiction of local interactions for that cell. Supplementary Fig. S5D demonstrates that each subpopulation (defined here by MYC, BCL2, and BCL6) has a unique pattern of co-occurrence with other subpopulations in terms of the range of Δ% scores in their local neighborhood. This empirical measurement suggests that typically cells of a particular subpopulation cluster with the same cell type (as shown in Supplementary Fig. S5B). There are patterns of heterotypic interaction with one another (Supplementary Fig. S5D, top, e.g., M+2+6– with M+2–6–), or heterotypic segregation (Supplementary Fig. S5D, top, e.g., M+2+6+ with M–2–6+ in the example tumor sample). Such interactions can be empirically established only through spatial investigation and provide a novel and independent feature of tumor heterogeneity that is patient-specific. These interaction patterns can be stable across different regions of the tumor for the same patient, or more rarely, heterogeneous with spatially varying interaction patterns in different tumor regions (Supplementary Figs. S5E and S6). We conclude from these investigations that B-cell subpopulations of different oncogenic coexpression aggregate spatially in a nonrandom manner (likely reflecting clustering due to parent cell–daughter cell relationships or, alternatively, embedding within local microenvironment milieus).

Cells Coexpressing MYC and BCL2 without BCL6 Confer Poor Survival in DLBCL

We next evaluated the relationship between MYC/BCL2/BCL6 subpopulations and prognosis, using pretreatment biopsies of R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone)–treated DLBCL patients, with clinical data available from three

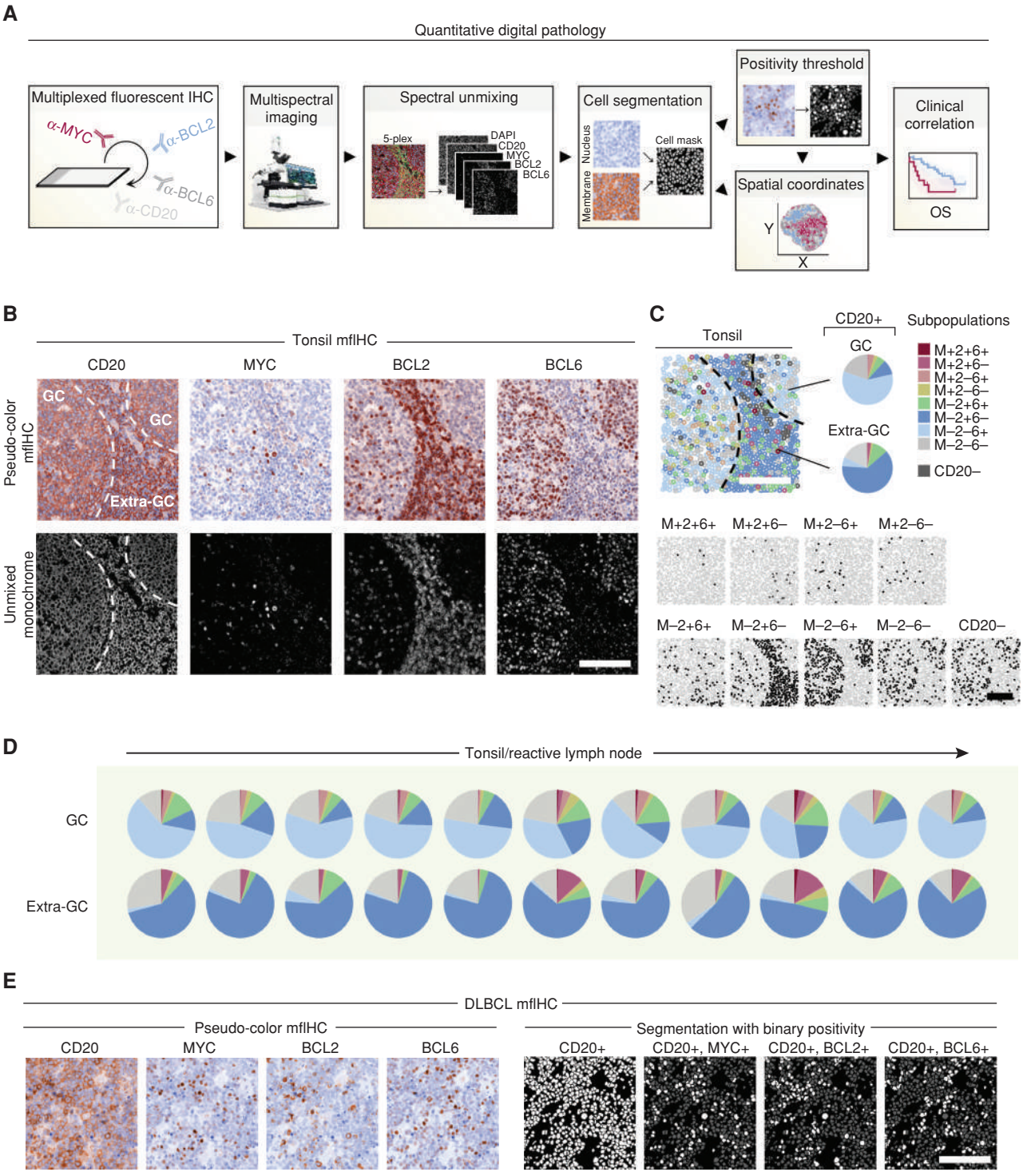


Figure 1. Quantitative single-cell analysis of MYC, BCL2, and BCL6 protein expression in B cells in nonmalignant tissues and diffuse large B-cell lymphoma. **A,** Schematic workflow of a quantitative digital pathology experiment. **B,** Spectrally unmixed multiplexed fluorescent images for CD20, MYC, BCL2, and BCL6 and nuclear counterstaining in tonsil tissue. The germinal center (GC) and extragerminal center (extra-GC) zones are indicated. **C,** Spatial map of MYC/BCL2/BCL6 subpopulations, i.e., possible permutations of MYC/BCL2/BCL6-positivity and -negativity within the CD20-positive cell population in a tonsil image. **D,** Quantitation of subpopulation extent within CD20-positive cells in tonsils and reactive lymph nodes resolved spatially between the GC and extra-GC zones. **E,** Example of pseudocolored MYC/BCL2/BCL6/CD20 mfIHC staining in diffuse large B-cell lymphoma (DLBCL; left). Cell segmentation and single oncogene positivity masks are shown within the CD20-positive cell population (right). (*continued on next page*)

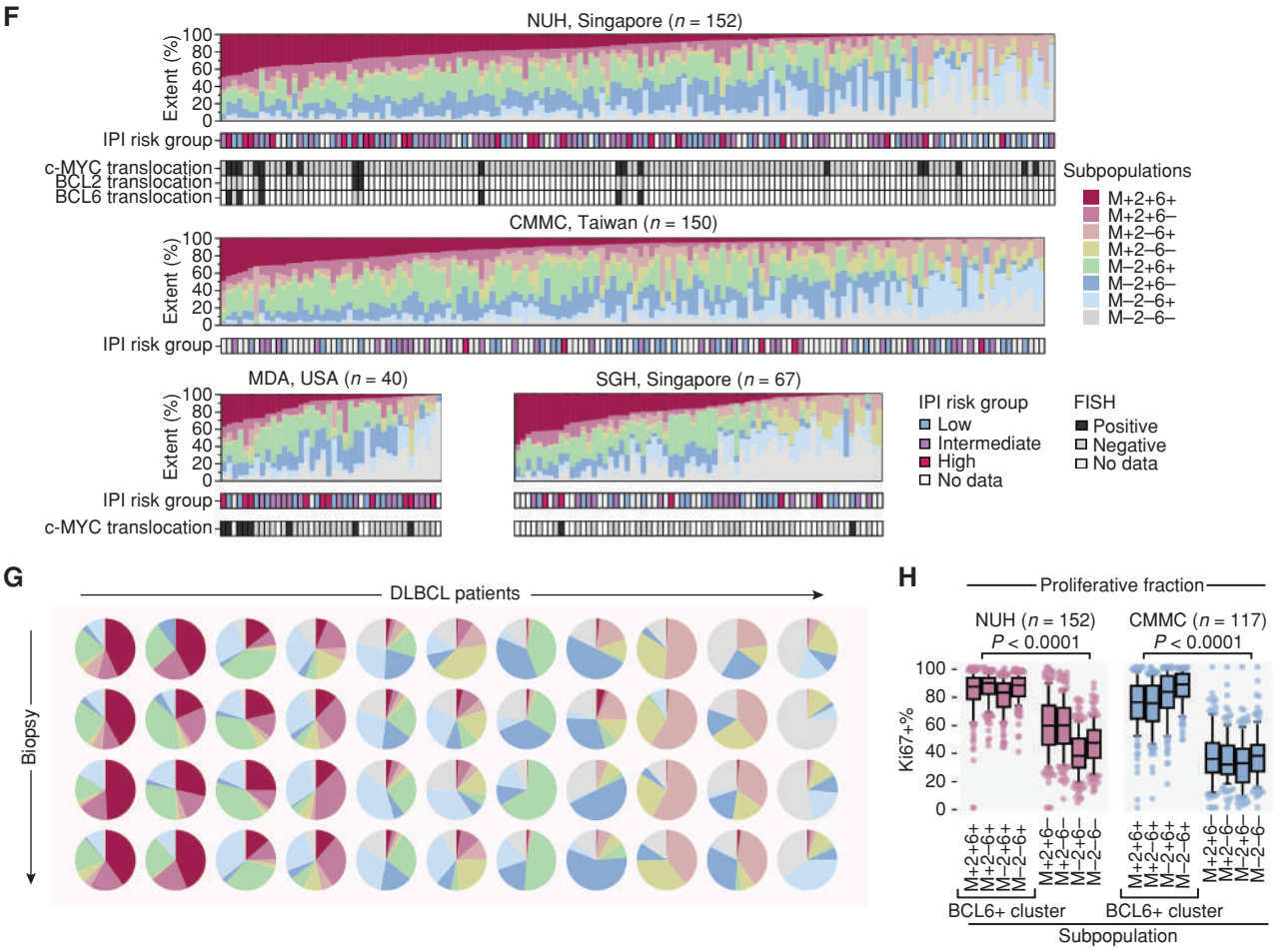


Figure 1. ***(Continued)*** **F,** Summary of percentage extent of subpopulations across patients from National University Hospital (NUH), Chi-Mei Medical Center (CMMC), MD Anderson (MDA), and Singapore General Hospital (SGH). Relevant clinicopathologic features are indicated; see also Supplementary Fig. S3. Patients were ordered arbitrarily according to extent of the triple-positive M+2+6+ subpopulation extent. IPI Risk Group, International Prognostic Index Risk Group; FISH, fluorescence *in situ* hybridization. **G,** Intrapatient spatial stability of subpopulations – proportion of subpopulations measured across four spatially distinct biopsies from the same patient (rows). Biopsy comparison overview is shown across 11 representative example DLBCL patients (columns). See also Supplementary Fig. S4A and S4B for a correlation analysis for all patients with multiple biopsies available. **H,** Proliferation analysis (i.e., Ki-67-positivity) among subpopulations in DLBCL samples. Proliferative BCL6-positive subpopulations are grouped. Median with interquartile range, whiskers denote 10th and 90th percentile. Mann–Whitney test (BCL6-positive vs. -negative subpopulations). All scale bars, 100 μm.

cohorts—National University Hospital, Singapore (NUH, $n = 98$), SGH ($n = 41$), and MDA ($n = 36$). To avoid arbitrary cutoffs, we initially evaluated the percentage of cells with oncogenic combinations as a continuous variable in a univariate Cox proportional hazards (Cox PH) analysis for overall survival (OS). Despite expected variability between cohorts in the prognostic impact of MYC, BCL2, and BCL6 as individual oncogenes (14), the percentage of M+2+6– cells stood out as a consistently poor prognostic variable (Fig. 2A). In this context, we define consistency as when hazard ratios (HR) for death, including 95% confidence intervals, for all cohorts have the same directionality (either consistently greater than 1 or less than 1). The M+2+6– subpopulation showed the greatest effect size and exclusively remained highly statistically significant in a pooled analysis across cohorts (Fig. 2A; Supplementary Table S4). This prognostic association is also illustrated in a dichotomized Kaplan–Meier survival analysis (Fig. 2B). Higher M+2+6– percentage remained statistically significant for poor OS in a multivariate Cox PH model adjusted for clinically relevant DLBCL clinicopathologic parameters of IPI Risk Group and MYC fluorescence *in situ* hybridization (FISH) status (Table 1; Supplementary Table S5). These results suggest that the prognostic impact of these oncogenes in DLBCL is driven by a unique subpopulation of cells expressing MYC and BCL2 without BCL6.

A Probabilistic Metric Accurately Predicts MYC, BCL2, and BCL6 Coexpression in DLBCL Cells

As the percentage of MYC+BCL2+BCL6– (M+2+6–) cells correlates with poor survival, we wanted to check if this

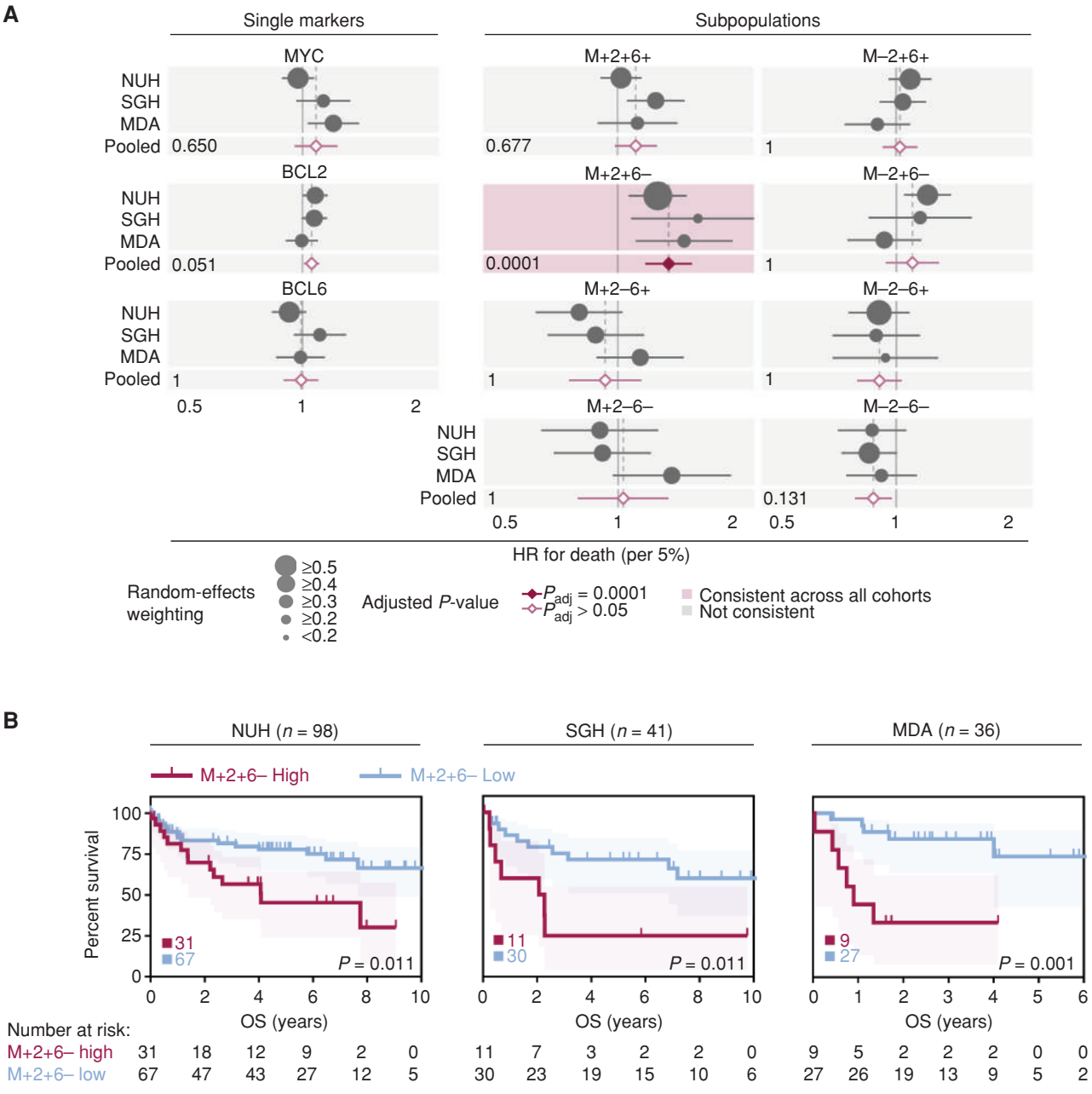


Figure 2. Prognostic significance of subpopulations after R-CHOP therapy. **A,** Pooled univariate Cox PH model analysis for MYC, BCL2, and BCL6 single oncogene and subpopulations percentage extent as predictors of OS across multiple DLBCL cohorts. Percentage extent was used as a continuous variable in the model at 5% increments (see Survival Analysis) for an unbiased comparison between the variables. Pooled *P* values were Bonferroni corrected for single oncogenes and subpopulations independently to adjust for multiple testing and are shown for each variable. Hazard ratio (HR) with 95% confidence interval (CI) per 5%-positivity increment is shown (see also Supplementary Table S4). **B,** Kaplan–Meier OS analysis of dichotomized into M+2+6– high and low groups. Log-rank test, shading denotes 95% CI. An optimal dichotomization cutoff was used for stratification; total patient numbers in each group are shown.

percentage could be inferred from knowledge of the individual oncogene components. For this, we first describe three possible "scenarios" of oncogene coexpression within a tumor: interdependent expression of each oncogene resulting in overlapping distribution patterns in a population of cells; independent/stochastic expression of each oncogene resulting in random distribution patterns across cells; mutually exclusive expression of each oncogene resulting in spatially excluded distribution patterns (depicted schematically in Fig. 3A). In terms of percentage extent between two oncogenes within a tumor, an interdependent expression would result in a strong positive correlation, an independent/

Table 1. Multivariate analysis of continuous M+2+6− percentage extent at 5% increments as a predictor of OS in the NUH, SGH, and MDA cohorts of DLBCL (Cox proportional hazards model)

	NUH Total cases ($n = 87$) missing values ($n = 11$)		SGH Total cases ($n = 37$) missing values ($n = 4$)		MDA Total cases ($n = 34$) missing values ($n = 2$)	
	HR (95% CI)	*P* value	HR (95% CI)	*P* value	HR (95% CI)	*P* value
Subpopulation						
M+2+6− (continuous, per 5% of extent)	1.3 (1.1–1.6)	0.004	1.6 (1.1–2.3)	0.026	1.6 (1.1–2.3)	0.010
IPI Risk Group		0.402		0.015		0.195
Low	Ref.		Ref.		Ref.	
Intermediate	1.6 (0.63–4.3)	0.314	4.4 (0.88–21.7)	0.299	1.3 (0.19–9.2)	0.780
High	2.0 (0.72–5.6)	0.187	12.6 (2.3–70.1)	0.005	4.5 (0.75–27.1)	0.099
c-MYC translocation status		0.503	—	—		0.536
Negative	Ref.		—	—	Ref.	
Positive	0.61 (0.14–2.6)		—	—	1.6 (0.35–7.7)	

Abbreviations: IPI Risk Group, International Prognostic Index Risk Group; 95% CI, 95% confidence interval; Ref., reference group.

stochastic expression would result in no correlation, and a mutually exclusive expression would result in a strong negative correlation. In contrast to the mutually exclusive expression pattern observed within specific topological compartments in reactive lymphoid tissues, MYC, BCL2, and BCL6 did not show strong correlation or anticorrelation with each other in DLBCL (Fig. 3B).

These results suggest that independent gene regulatory mechanisms drive the expression of *MYC*, *BCL2*, and *BCL6* in DLBCL, and that single-cell coexpression of these oncogenes is largely stochastic. This implies that the percentage of any oncogenic coexpression subpopulation can be inferred by a simple probabilistic metric based on the percentage extent of each component oncogene. If the overall percentages of each component oncogene are known, such a metric describing the percentage of any given subpopulation is derived by multiplying proportions for the presence or absence of each individual oncogene comprising the subpopulation (see Methods). We validated this hypothesis using our single-cell–resolved mfIHC data, observing a highly concordant correlation between observed and predicted percentages for each subpopulation (Fig. 3C; Supplementary Fig. S7).

An extension of this hypothesis is that any quantitative data of MYC, BCL2, and BCL6 allow estimation of the percentage of their coexpressed subpopulations. One such semiquantitative data source of clinical interest is the visual scoring of MYC, BCL2, and BCL6 percentage on chromogenic IHC, which remains the reference method for the assessment of these oncogenes. We checked if our metric could estimate prognostic MYC, BCL2, and BCL6 subpopulations from clinical-grade pathologist scores for chromogenic IHC in a well-characterized cohort of DLBCL from the British Columbia Cancer Agency (BCA; ref. 15). We first performed mfIHC on the BCA cohort to obtain empiric values of the MYC/BCL2/BCL6 coexpressing subpopulations (Supplementary Table S2). M+2+6− percentage extents measured by mfIHC were used to determine an optimal clinically relevant cutoff to classify a patient as a high M+2+6− expressor and therefore likely to have a poor outcome. A dichotomized cutoff of 15% of the M+2+6− subpopulation percentage extent produces the greatest effect size of OS stratification as determined by the Cox PH model between high and low groups in this cohort (Fig. 3D). We then calculated the inferred M+2+6− metric from retrospective semiquantitative chromogenic IHC values. A Kaplan–Meier analysis of the cohort dichotomized into high (≥15% M+2+6− metric) and low (<15%) demonstrated the poor survival of the high metric group (Fig. 3E), confirming the applicability of this probabilistic metric to clinical IHC scoring in DLBCL.

One key consideration for applicability of this metric as a biomarker would be the size of the region to be sampled for adequate representation. As our cohorts were studied in tissue microarray format with small (1 mm) cores, we also evaluated our multiplexed analysis on whole-tissue DLBCL tumor sections ($n = 8$; UP; Supplementary Table S2). The variance in M+2+6− percentage extent in different tissue regions/image fields was low (Supplementary Fig. S8A). Importantly, with the low variance, sampling just two or more high-power diagnostic fields is generally reliable to determine M+2+6− high vs. low samples using a single threshold cutoff of 15% (Supplementary Fig. S8B). Overall, these findings speak to the possible clinical applicability of the M+2+6− metric for pathologist scored chromogenic IHC, which requires validation in future prospective studies.

Estimation of MYC, BCL2, and BCL6 Coexpressing Subpopulations Can Be Extended to Gene-Expression Data

We hypothesized that if the percentage of MYC/BCL2/BCL6 subpopulations could be inferred from individual oncogene components on IHC, then our metric could

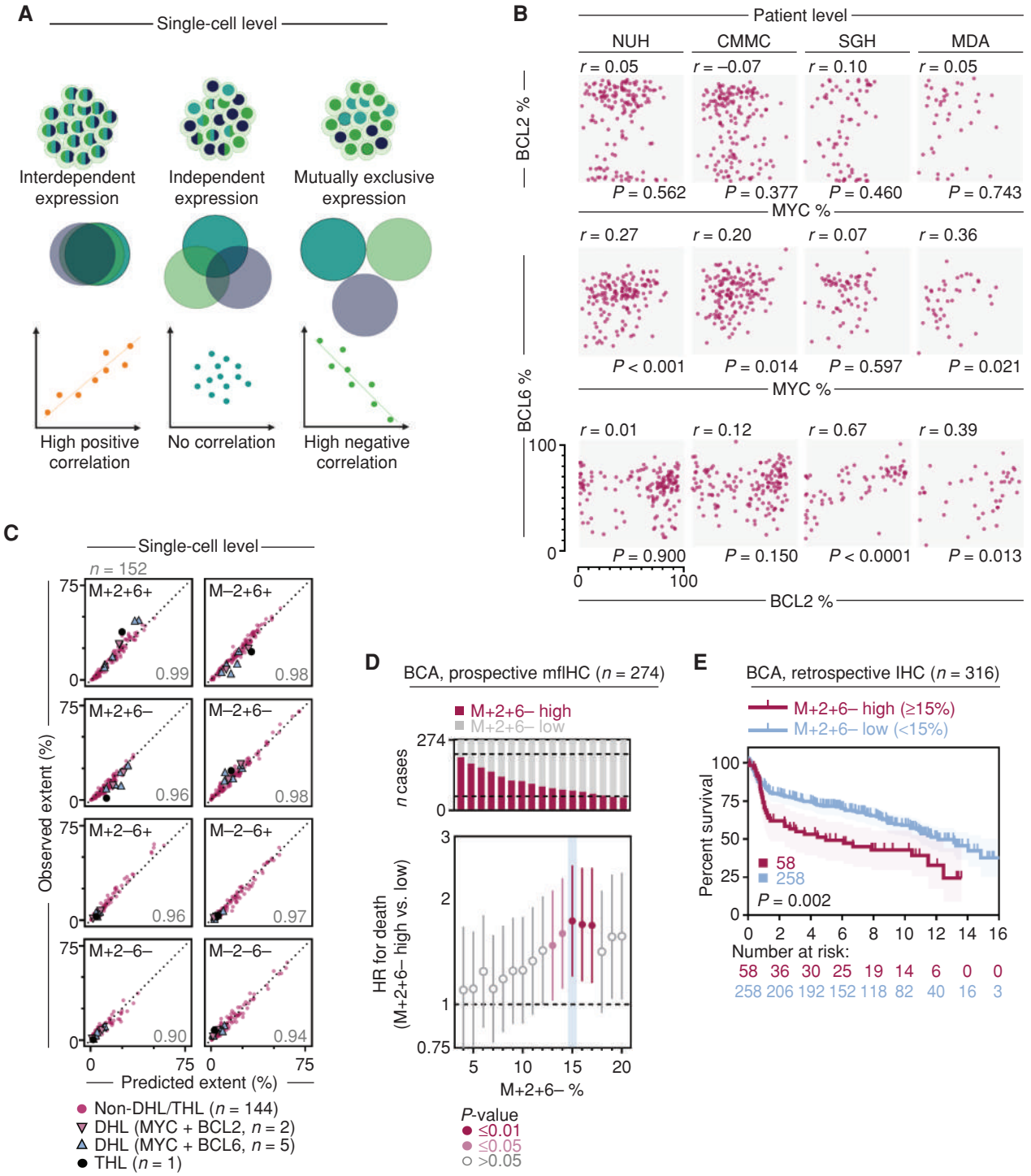


Figure 3. MYC, BCL2, and BCL6 protein coexpression in DLBCL can be inferred from individual marker data. **A,** Schematic of possible relationships between expression of three oncogenes in a population of cells. The distribution of these oncogenes can either reflect interdependent expression, independent/stochastic expression, or mutually exclusive expression. These relationships result in the percentage extent of oncogenes in the tumor being either strongly positively correlated, not correlated, or strongly negatively correlated, respectively. Created with BioRender.com. **B,** Correlation of MYC, BCL2, and BCL6 percentage extent across patients in DLBCL cohorts. Spearman correlation; axes are equivalent in all panels. **C,** Good correlation between probabilistically predicted subpopulation percentage extent based on single oncogene positivity and observed percentage extent in the NUH cohort. Cases of double-hit lymphoma (DHL, MYC+BCL2+ translocations or MYC+BCL6+ translocations) or triple-hit lymphoma (THL) are highlighted. Spearman rho for each correlation is shown; axes are equivalent in all panels. Correlation for other cohorts can be found in Supplementary Fig. S7. **D,** Prospective evaluation of an optimal dichotomization cutoff for M+2+6– percentage extent in the BCA cohort. Univariate Cox PH model at 1% extent positivity increment, HR for death with 95% CI. HR scale is exponential. Optimal dichotomization cutoff is highlighted in blue. **E,** Kaplan–Meier OS analysis of the chromogenic IHC BCA cohort evaluation stratified into M+2+6– metric high and low groups across an absolute value of 15% M+2+6–metric. Log-rank test, shading denotes 95% confidence interval. CMMC, Chi-Mei Medical Center.

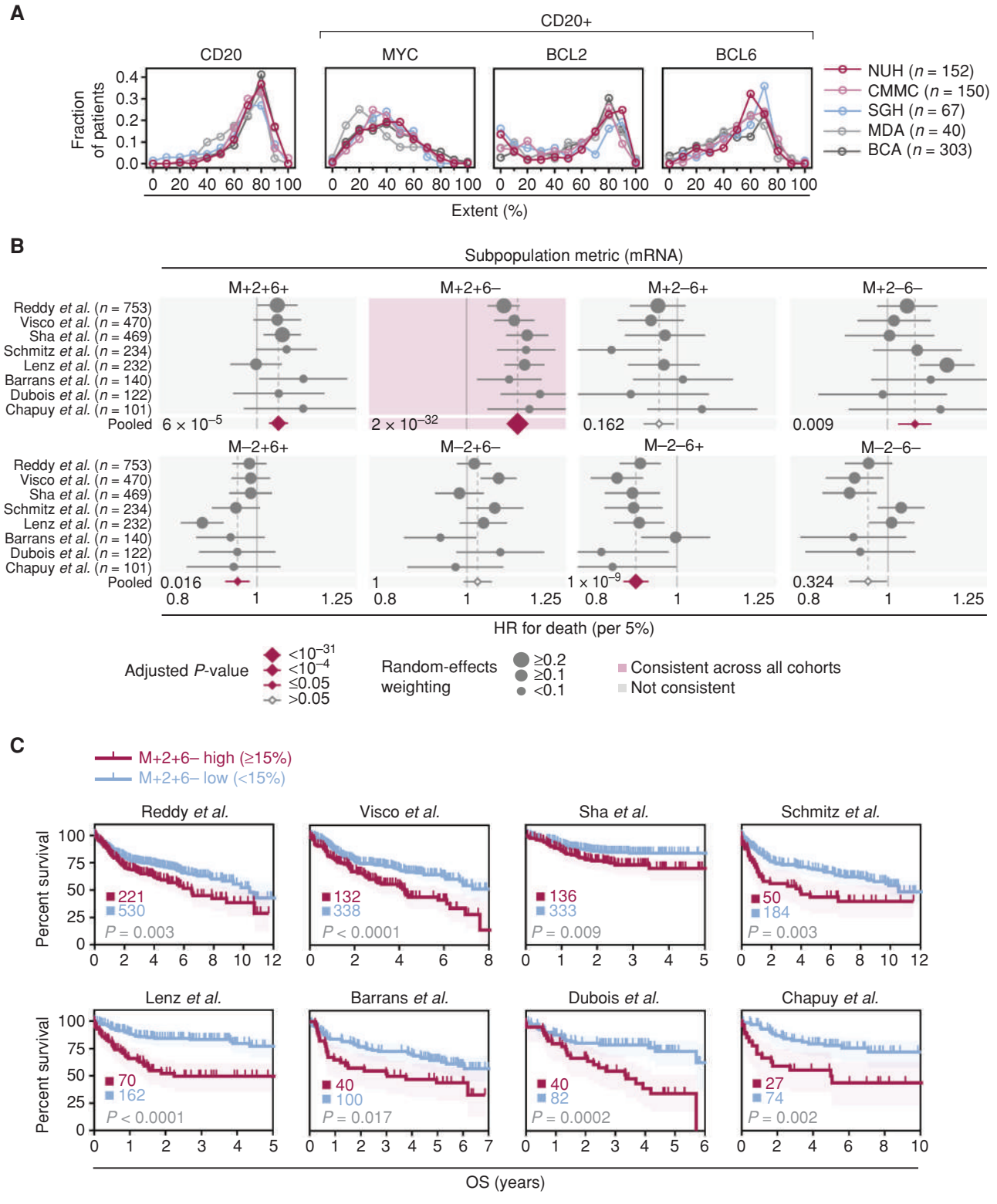


Figure 4. Validation of prognostic significance of the M+2+6– subpopulation metric in gene-expression datasets. **A,** Distribution of single oncogene positivity in DLBCL cohorts as assessed by mfIHC (see Supplementary Fig. S9A and S9B). **B,** Impact of subpopulation metrics in GEP datasets on OS. Pooled univariate Cox PH model analysis; metric was used as a continuous variable in the model at 5% increments. HR per 5% increment with 95% CI is shown; CI are proportional on both tails but are capped at the graph's edges. Pooled *P* values were Bonferroni corrected to adjust for multiple testing and are shown for each subpopulation. **C,** Kaplan–Meier OS analysis of GEP cohorts stratified uniformly across absolute 15% M+2+6– metric into -high and -low groups. Log-rank test, shading denotes 95% CI. Total patient numbers in each group are shown. CMMC, Chi-Mei Medical Center.

also be computed from other quantitative data measuring MYC/BCL2/BCL6 expression. This could also extend to gene-expression profiling (GEP) with the assumption of positive mRNA-protein correlation (which has been reported for MYC and BCL2 expression in DLBCL; ref. 16). To transform gene-expression data into predicted percentage extents, we first established an empirical cumulative distribution function (eCDF) for each individual protein marker (MYC/BCL2/BCL6 percentage extents) across five mfIHC cohorts (n = 712). Importantly, the eCDFs of single-marker protein and subpopulation percentages are similar across all five mfIHC cohorts of patients (Fig. 4A), allowing compilation of an aggregated consensus protein distribution for each oncogene (Supplementary Fig. S9A and S9B). We then perform eCDF mapping (matching percentile points in the eCDF of mRNA measurements to those in the eCDF of the corresponding protein scores) and convert the oncogene's quantitative mRNA score into an inferred single-marker percentage extent (see Methods; Supplementary Fig. S9C). These inferred percentage extents from mRNA data could be used to generate our aforementioned metric, estimating proportions of subpopulations (Supplementary Table S6).

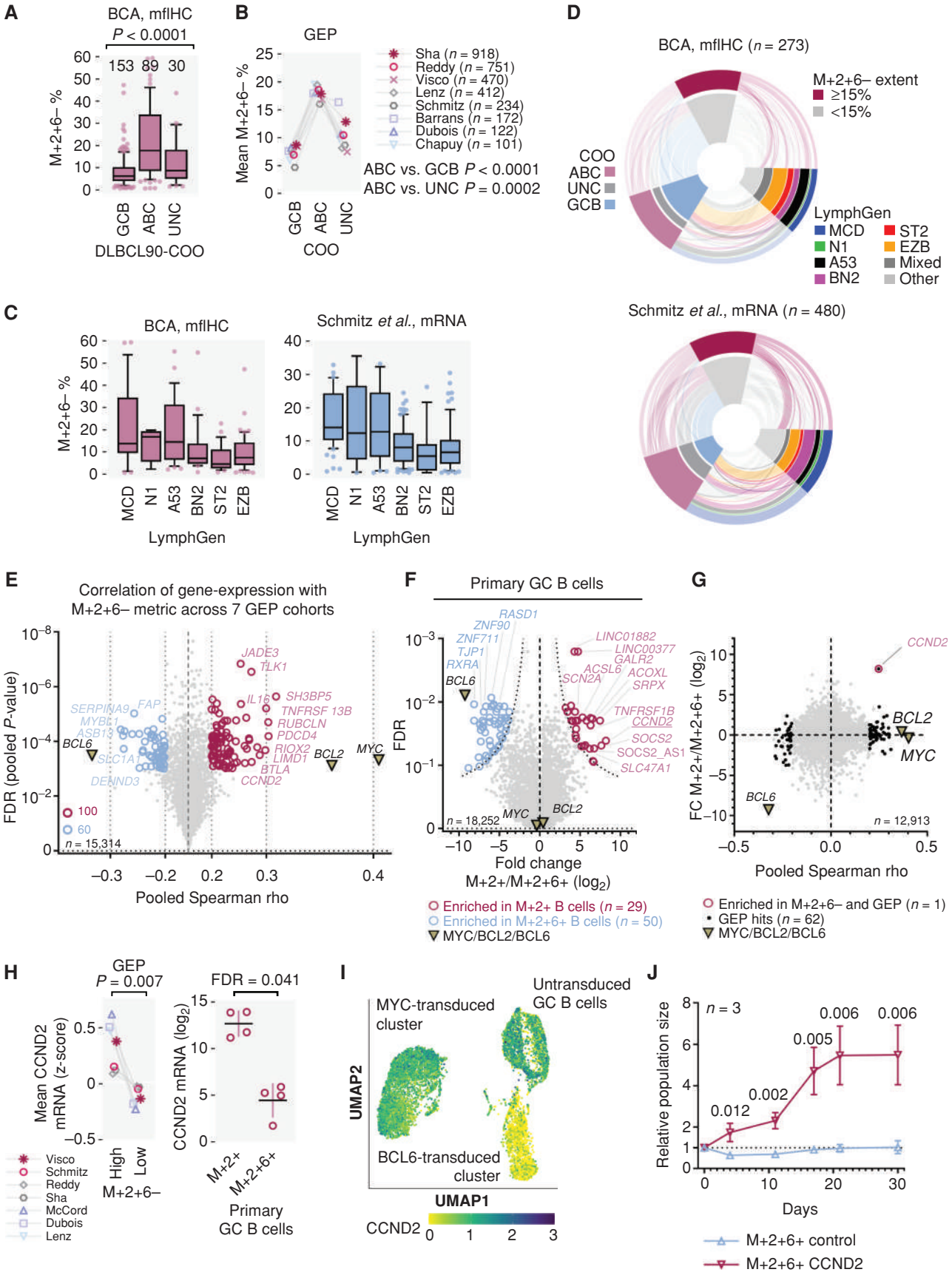
We then utilized GEP cohorts with available survival data after R-CHOP treatment (8 cohorts, n = 2,521) to evaluate the prognostic impact of the RNA-based metric for M+2+6− prediction. The M+2+6− metric remained the only RNA-inferred subpopulation consistently associated with poor survival across eight distinct cohorts of DLBCL patients (Fig. 4B; Supplementary Table S7). As with the mfIHC-based results, the pooled analysis revealed that the M+2+6− metric had the greatest effect size and most statistically significant P value with respect to HR for death. In the GEP analysis, metrics representing other subpopulations did show occasional statistically significant survival associations—but these were not consistent, of a smaller effect size and by many orders of magnitude less statistically significant compared with the M+2+6− metric. The M+2+6− metric was consistently prognostic in both microarray gene-expression–based cohorts (17–22) and RNA sequencing (RNA-seq)–based cohorts (23, 24), attesting to its validity for mRNA quantified from varying platforms. The significance of the M+2+6− metric was further corroborated in a multivariate Cox PH analysis correcting for IPI Risk Group and cell-of-origin (COO) gene-expression signature, where M+2+6− remained a statistically significant predictor of poor survival in seven out of eight cohorts as a continuous variable (Supplementary Table S8).

Finally, we performed an independent study on the mRNA-based M+2+6− metric in samples with biomarker data available from the GOYA clinical trial cohort (25). GOYA was a randomized phase III trial (NCT01287741) comparing two different anti-CD20 antibodies (rituximab and obinutuzumab) in combination with CHOP chemotherapy. Although the trial did not show differences in survival between the two arms, it remains a valuable source of evaluating molecular determinants of survival in chemoimmunotherapy-treated DLBCL. Although BCL6 IHC was not available for the GOYA samples, MYC and BCL2 IHC scores showed statistically significant correlations with mRNA levels of MYC and BCL2, respectively, supporting the rationale for the extension of our metric from protein to mRNA (Supplementary Fig. S10A and S10B). The M+2+6− metric was associated with progression-free survival and OS in this dataset, in both univariate and multivariate analyses (Supplementary Table S9). Finally, Kaplan–Meier OS analysis on GOYA as well as other publicly available GEP datasets confirmed that our previously established 15% threshold cutoff was relevant for prognostic stratification (Fig. 4C; Supplementary Fig. S10C; Supplementary Table S10).

Molecular Characteristics of M+2+6− High DLBCL

Inference of oncogene coexpression from GEP datasets allows an extended avenue for comparative analysis with other molecular characteristics in DLBCL, which can be utilized to describe molecular features characterizing the M+2+6− subpopulation. We first investigated the relationship between (mfIHC-generated) M+2+6− percentage extent and GEP-determined COO data available for the BCA cohort. M+2+6− percentage extent was associated with the ABC COO subtype (Fig. 5A), and this association with ABC COO was consistent for the inferred M+2+6− metric across the GEP datasets (Fig. 5B). The M+2+6− subpopulation and M+2+6− metric also was associated with the unfavorable MCD and A53 genetic subtypes of DLBCL (Fig. 5C; ref. 26). These relationships are depicted categorically in an integrated fashion in Fig. 5D.

To derive single-gene associations with the M+2+6− subpopulation on the bulk level, we correlated the M+2+6− metric with gene expression across seven GEP cohorts (n = 3,180 samples, Fig. 5E; Supplementary Table S11). One-hundred sixty genes consistently correlated either positively or negatively with the M+2+6− metric (Fig. 5E). To narrow down those of key biological significance in the first instance, we leveraged on the observation that M+2+6− percentages are strongly correlated with a poor prognosis, whereas the survival association with M+2+6+ is much weaker. We compared gene expression of DLBCL with gene expression of primary human tonsil-derived GC B cells immortalized by either the overexpression of *MYC* and *BCL2* (M+2+6−) or *MYC*, *BCL2*, and *BCL6* (M+2+6+; ref. 27). Because BCL6 is a transcriptional repressor (28), we hypothesized that the absence of BCL6 could influence the transcriptional profile of the M+2+6− subpopulation (Fig. 5F; Supplementary Table S12). Cross-comparing genes enriched in the M+2+6− GC B cells with genes correlated with the M+2+6− population from bulk clinical GEP datasets, we found that *CCND2* (which codes for cyclin D2) was highly enriched in both groups (Fig. 5G and H). Furthermore, single-cell RNA-seq (scRNA-seq) of a tonsil-derived GC B-cell sample clearly demonstrated an inverse correlation between *CCND2* and *BCL6* expression (Fig. 5I), confirming prior observations that *CCND2* is transcriptionally repressed by BCL6 (29, 30). We then transduced *CCND2* in M+2+6+ immortalized B cells, which were characterized by low background levels of cyclin D2 expression (Supplementary Fig. S11).


A
BCA, mflHC
P < 0.0001
153 89 30
M+2+6– %
GCB ABC UNC
DLBCL90-COO
B
GEP
Mean M+2+6– %
Sha (n = 918)
Reddy (n = 751)
Visco (n = 470)
Lenz (n = 412)
Schmitz (n = 234)
Barrans (n = 172)
Dubois (n = 122)
Chapuy (n = 101)
ABC vs. GCB P < 0.0001
ABC vs. UNC P = 0.0002
COO
D
BCA, mflHC (n = 273)
M+2+6– extent
≥15%
<15%
COO
ABC
UNC
GCB
LymphGen
MCD ST2
N1 EZB
A53 Mixed
BN2 Other
C
BCA, mflHC
Schmitz et al., mRNA
M+2+6– %
MCD N1 A53 BN2 ST2 EZB
LymphGen
Schmitz et al., mRNA (n = 480)
E
Correlation of gene-expression with
M+2+6– metric across 7 GEP cohorts
FDR (pooled P-value)
JADE3
TLK1
SERPINA9
FAP
MYBL1
ASB13
IL16
SH3BP5
TNFRSF 13B
RUBCLN
PDCD4
BCL6
SLC1A1
RIOX2
BCL2
MYC
LIMD1
BTLA
DENND3
CCND2
100
60
n = 15,314
Pooled Spearman rho
F
Primary GC B cells
RASD1
ZNF90
ZNF711
TJP1
RXRA
BCL6
LINC01882
LINC00377
GALR2
ACSL6
ACOXL
SCN2A
SRPX
TNFRSF1B
CCND2
SOCS2
SOCS2_AS1
SLC47A1
FDR
n = 18,252
MYC
BCL2
Fold change
M+2+/M+2+6+ (log2)
Enriched in M+2+ B cells (n = 29)
Enriched in M+2+6+ B cells (n = 50)
MYC/BCL2/BCL6
G
CCND2
BCL2
MYC
BCL6
FC M+2+/M+2+6+ (log2)
n = 12,913
Pooled Spearman rho
Enriched in M+2+6– and GEP (n = 1)
GEP hits (n = 62)
MYC/BCL2/BCL6
H
GEP
P = 0.007
FDR = 0.041
Mean CCND2 mRNA (z-score)
CCND2 mRNA (log2)
High Low
M+2+6–
M+2+ M+2+6+
Primary GC B cells
Visco
Schmitz
Reddy
Sha
McCord
Dubois
Lenz
I
MYC-transduced cluster
Untransduced GC B cells
BCL6-transduced cluster
UMAP2
UMAP1
CCND2 0 1 2 3
J
n = 3
Relative population size
0.012
0.002
0.005
0.006
0.006
Days
M+2+6+ control
M+2+6+ CCND2

The transduced M+2+6+/*CCND2*^{High} population started as a relatively small fraction, but rapidly expanded over time eventually outgrowing the M+2+6+/*CCND2*^{Low} population (Fig. 5J; Supplementary Fig. S11), confirming that increased cyclin D2 expression can confer a fitness advantage to cells with MYC and BCL2. Cyclin D2 expression has been reported as a marker for an adverse outcome in DLBCL (31, 32).

Single-Cell Transcriptomic Analyses of M+2+6− Cells in DLBCL

To further understand other molecular determinants underlying poor prognosis in cases with high numbers of M+2+6− cells, we leveraged on scRNA-seq datasets to profile the transcriptomic characteristics of M+2+6− malignant B cells within DLBCL samples. We first harmonized single-cell transcriptomic data from 6 DLBCL patient samples from two independent datasets (refs. 33, 34; Fig. 6A). Figure 6B demonstrates that the M+2+6− subpopulation is well represented in all samples. We identified genes associated with the M+2+6− B-cell subpopulation (Fig. 6C) and confirmed *CCND2* expression being more abundant in M+2+6− B cells compared with all other malignant B cells. In total, 13 concordant genes were enriched in both the scRNA-seq data and the bulk RNA-seq data (Supplementary Table S13). These include ABC-DLBCL–related genes such as the ROCK1 target PES1, which intersects MYD88 and NF-κB signaling (35), the PIM2 kinase whose overexpression has been associated with unfavorable DLBCL biology (36), and the IRF4 interactor BATF (37). Finally, Fig. 6D depicts a pathway analysis on this single-cell–resolved transcriptomic data revealing that the PI3K–AKT pathway, immune responses (including the complement pathway), as well as G-protein receptor–coupled signaling, were among the significantly enriched terms in M+2+6− B cells compared with all other malignant B cells (Fig. 6D; Supplementary Table S14). The enrichment of PI3K–AKT signaling signatures is particularly intriguing, as inhibitors of this pathway (e.g., copanlisib) are clinically applicable, suggesting a possible therapeutic strategy for unfavorable M+2+6− high tumors. Additional mechanistic studies will be needed to understand the relative significance and interplay between these genes and pathways in conferring poor outcome in M+2+6−. Of general significance, however, these results illustrate how the estimation of oncogene coexpression phenotypes through gene-expression data, coupled with single-cell resolved transcriptomic data, may uncover novel biological insight.

DISCUSSION

In this paper, we show for the first time that subpopulations of tumor cells expressing combinations of oncogenes at the single-cell level influence patient prognosis. We also show that (under conditions of independently regulated expression), these subpopulations can be inferred from quantitative single oncogene expression data, generating a metric that has a remarkable concordance to actual observed single-cell coexpression on multiplex IHC. We show two applications of predicting oncogenic subpopulations in the setting of DLBCL. First, the M+2+6− metric can be generated from diagnostic IHC scores, offering a refined method for utilizing MYC, BCL2, and BCL6 expression for prognostic use in DLBCL. Secondly, by estimating subpopulation percentages from GEP datasets, we demonstrate the feasibility of identifying molecular features associated with a poor prognostic oncogene combination from the plethora of gene-expression studies available for a given disease. Such features could identify therapeutic targets or offer biological insight–as with our demonstration that the cell-cycle regulator cyclin D2 (*CCND2*) may play a role in the aggressive phenotype of M+2+6− cells. Cyclin D2 promotes the G_1–S transition of hematopoietic cells (38), enhances cytokine induced-proliferation (39), and is stabilized by EBV infection (40), highlighting the rationale for further studies of *CCND2* in DLBCL pathogenesis and evolution.

Our single-cell–resolved quantitative imaging confirms that ITH in coexpression of MYC, BCL2, and BCL6 exists in almost every case of DLBCL. This coexpression shows distinct spatial organization with nonrandom clustered patterns, supporting the concept that forces beyond genetic heterogeneity shape DLBCL evolution. These findings also suggest that quantitative assessment of the M+2+6− subpopulation potentially refines the MYC-BCL2 "double expressor lymphoma" (DEL), a term used to describe DLBCL with overexpression of MYC and BCL2 protein in the absence of underlying genetic rearrangements (7, 41, 42). DEL is typically defined as >40% MYC-positive cells and >50% BCL2-positive cells (measured independently). As these DEL classifications do not take DLBCL ITH (43,

Figure 5. Transcriptomic analysis of M+2+6− high cases and potential role of *CCND2*. **A,** Correlation of observed M+2+6− percentage extent in the BCA cohort with the cell-of-origin (COO) DLBCL90-COO signature. Bonferroni corrected Kruskal–Wallis test for ABC vs. others. **B,** Correlation of M+2+6− metric in GEP cohorts with COO signatures. Mean M+2+6− metric value per group per cohort is shown. Bonferroni corrected paired-samples *t* test. **C,** Correlation of the M+2+6− percentage extent and metric evaluated by mfIHC and mRNA inference, respectively, with genetic subtypes (LymphGen classification). **D,** Sankey plot of M+2+6− dichotomized grouping matched with molecular features. **E,** Volcano plot of pooled direct correlation of gene mRNA expression and M+2+6− metric across seven GEP cohorts. Genes highly correlated with M+2+6− metric across datasets at absolute Spearman rho ≥0.2 and FDR≤0.001 are shown (see also Supplementary Table S11). The abscissa is scaled exponentially. **F,** Differential gene expression between primary GC B cells overexpressing M+2+ and M+2+6+ (see also Supplementary Table S12). Analysis is generated from 4 biological replicates from each condition, from cells of independent donors. **G,** Genes highly enriched in M+2+6− cells: correlation of results from **E** and **F**. **H,** *CCDN2* gene expression in GEP cohorts in patients dichotomized by M+2+6− 15% metric (left) and in primary B cells (right). Paired *t* test (left); mean with standard deviation and FDR (FDR as per Supplementary Table S12) for *t* test (right). **I,** Single-cell RNA-seq of GC primary B cells transduced either with BCL2 and MYC (MYC-transduced) or BCL2 and BCL6 (BCL6-transduced). Untransduced GC primary B cells are also included. Expression of *CCND2* is indicated in color. **J,** Proliferation analysis of M+2+6+ primary GC B cells overexpressing cyclin D2 (*CCND2*) compared with M+2+6+ primary GC B cells transduced with an empty vector (EV). Analysis performed with 3 biological replicates for each condition, using cells from 3 independent patients; mean with standard deviation; *t* test. UMAP, Uniform Manifold Approximation and Projection.

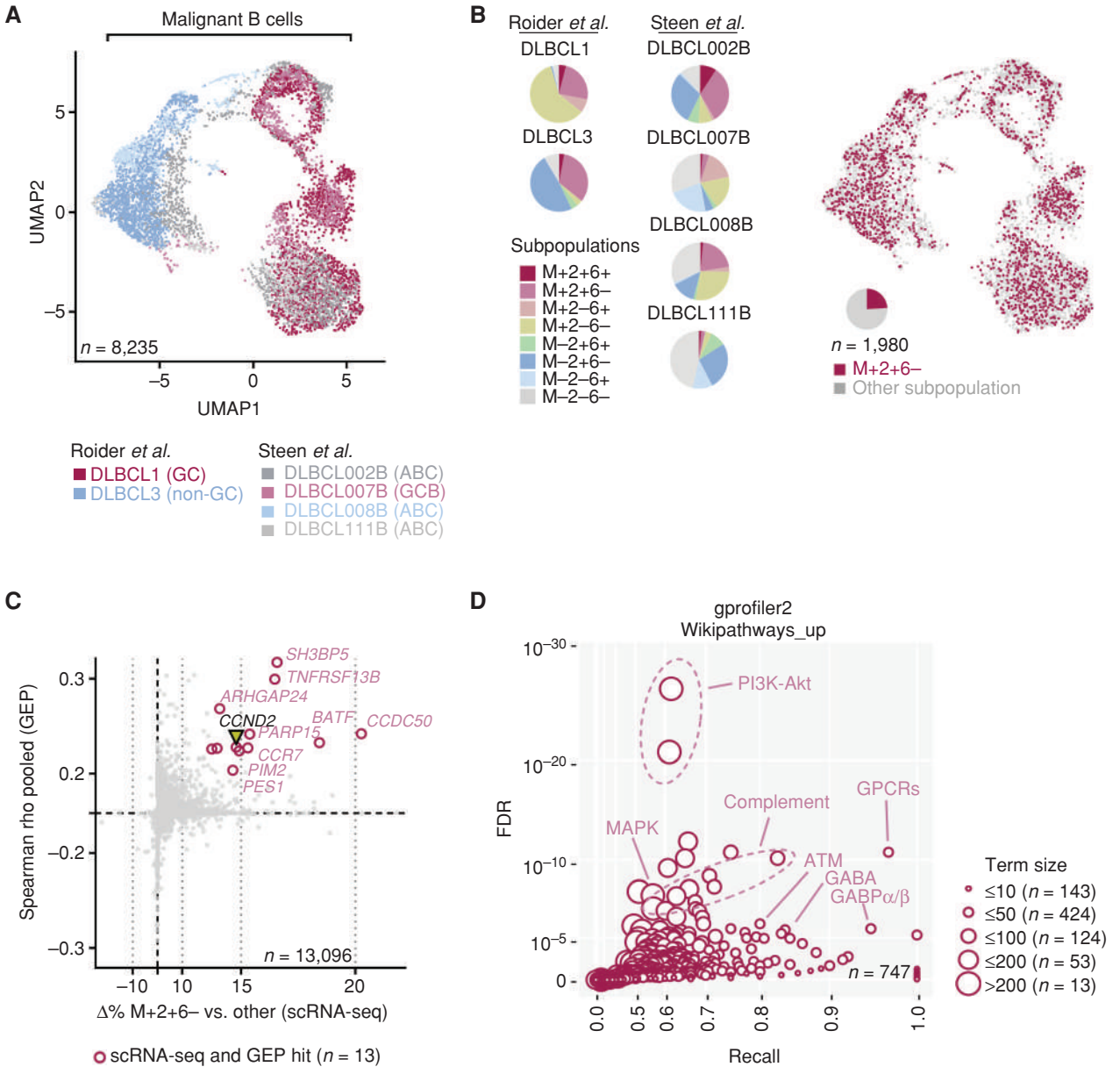


Figure 6. Evaluation of M+2+6– cells in scRNA-seq datasets of DLBCL. **A,** Uniform Manifold Approximation and Projection (UMAP) of malignant B cells from the Roider and Steen cohorts. **B,** Proportion of subpopulations across samples and annotation of M+2+6– cells in UMAP. **C,** Correlation of genes enriched in the M+2+6– subpopulation as evaluated by scRNA-seq with hits from the bulk GEP cohorts (Fig. 5E). *CCND2* is highlighted and is among the concordant hits (see also Supplementary Table S13). **D,** WikiPathways terms enrichment among genes positively associated with M+2+6– cells. Both axes in **C** and **D** are scaled exponentially for clarity (see also Supplementary Table S14).

44) into account, it was not known if DELs represent two distinct and coexisting clonal phenotypes within a lymphoma—one expressing MYC and the other BCL2. Nor was it understood why the poor outcome of DEL is exacerbated when BCL6 expression is absent (9, 45). These issues are addressed by the description of the M+2+6– subpopulation, which describes the phenomenon at single-cell resolution. DEL remains relevant in the era of genetic DLBCL classification (46) and novel targeted therapies. For example, patients with the DEL phenotype show improved survival on the polatuzumab arm in comparison with the R-CHOP arm of the POLARIX trial (47). Additional studies are required to clarify the relevance of the M+2+6– percentage extent in this setting.

Lymphomas that harbor translocations in MYC, BCL2, and/or BCL6, termed double-hit or triple-hit lymphomas (DHL/THL; ref. 48), have a particularly poor outcome (49–52). Recently, prognostic gene-expression signatures have been developed that accurately classify such DHL or THL cases: double-hit signature (DHITsig; ref. 49) and

molecular high-grade (MHG; ref. 17). Of DLBCL that are DHITsig positive, the majority fall within the EZB genetic subtype (26). Using our single-cell–resolved approach to DEL, we saw that cases with higher M+2+6− metrics were typically assigned to either the A53 or MCD genetic subtype (Fig. 5C). These results are consistent with double-hit (and DHL-like/MHG) lymphoma being biologically distinct from DEL, though both types display poor prognoses. Figure 5 demonstrates an association between the M+2+6− subpopulation and the ABC COO subtype as well as the MCD genetic subtype, consistent with observations that the MCD subtype is almost exclusively ABC (26). However, the M+2+6− subpopulation also shows enrichment in both A53 and "other" unclassified genetic subsets, and we posit that distinct genetic backgrounds can converge on this final phenotype through distinct mechanisms.

This is a proof-of-concept study with some limitations. First, the prognostic impact of M+2+6− percentages derive from retrospective analyses and need prospective validation. In this article, we have, where possible, presented HRs as a continuous variable, ascribing a risk score per unit of measure (per 5% of M+2+6− percentage extent). This is an unbiased method by which to assess the risk of the M+2+6− subpopulation, however, a more pragmatic approach for clinical biomarker use is to develop a standardized cutoff for the M+2+6− percentage. Our initial analyses from the BCA cohort and GEP datasets (including the GOYA trial) suggest that ≥15% M+2+6− percentage extent may be a suitable starting point for such prospective validation studies. Second, the prevalence of staining artifacts within FFPE tissue samples required us to use a semiautomated method (manual checking of intensity threshold per image), which is not optimal for upscaling of this approach. The development of deep learning approaches refined for evaluating marker positivity based on fluorescence intensity but also considering other background/morphologic features would be key toward implementing this diagnostic method into the clinic. Finally, although we use threshold-based positivity scoring in this study, per-cell quantitative proteomic data (ideally obtained through an amplification-free method such as imaging mass spectrometry) is an understudied area that may yield even greater resolution toward assessing prognostic outcomes.

The probabilistic metric we describe, which predicts oncogenic coexpression, holds true only when the expression of the oncogenes is independently regulated and will need validation in the setting of other oncogenes/cancers. Nonetheless, our demonstration that the M+2+6− phenotype confers poor survival in four empirically evaluated cohorts (mfIHC) and nine inferred cohorts (GEP) of DLBCL underscores the clinical importance of evaluating the ITH generated by the coexpression of oncogenes and suggests that similar studies in other cancer types will be informative. Oncogene ITH occurs at multiple molecular levels in cancer: genetic (53, 54), epigenetic, transcriptomic, and proteomic (55), and affects clinical phenotypes (1, 56). Single-cell approaches to evaluate genetic (57), transcriptomic (58, 59), and proteomic ITH have provided valuable insight into microevolutionary processes operating in cancer (60, 61). However, due to high experimental costs, the number of patients represented in scRNA-seq and mass cytometry datasets are invariably small, thus precluding clinically meaningful multivariate analyses. Here we demonstrate that multiplexed microscopy through mfIHC, though limited in multiplexing potential compared with scRNA-seq, is well suited to measure the clinical impact of single-cell–resolved ITH in clinically annotated patient cohorts.

METHODS

Samples and Datasets

Tonsils ($n = 15$) from patients diagnosed with chronic tonsillitis, reactive lymph nodes ($n = 2$), and DLBCL [$n = 152$, tissue microarray (TMA) format] were obtained from the NUH [approved by the Singapore NHG Domain Specific Review Board B study protocol (2015/00176)]. Additional DLBCL TMAs for quantitative mfIHC analyses were from the CMMC cohort ($n = 150$), the SGH cohort ($n = 67$), and the MDA cohort ($n = 40$). Pretreatment biopsies of the NUH, SGH, and MDA cohorts were used for survival analysis following standard first-line R-CHOP-like therapy. A TMA from the BCA cohort ($n = 274$) with first-line R-CHOP-like follow-up data was used as a validation cohort (49). Eight whole-slide DLBCL sections retrieved from the archives of the Tumor Immunology Laboratory of the University of Palermo were included in the study as approved by the University of Palermo Institutional Review Board (IRB) 09/2018. Full patient characteristics for all the above cohorts are provided in Supplementary Table S15. Samples from all institutions were obtained through IRB-approved ethics protocols, with written informed consent from the patients, or with IRB-approved waivers of consent where applicable in accordance with the ethical guidelines of the Declaration of Helsinki. Material transfer agreements from all providing institutions were incorporated into the framework of an NUS IRB-approved translational study (H-19-055E). Preprocessed gene-expression data were obtained from Gene Expression Omnibus (GEO; RRID: SCR_005012; ref. 62) for datasets GSE117556 ($n = 928$; ref. 17), GSE125966 ($n = 553$; ref. 63), GSE31312 ($n = 498$; ref. 18), GSE10846 ($n = 420$; ref. 19), GSE87371 ($n = 221$; ref. 20), GSE32918 ($n = 172$; ref. 21), and GSE98588 ($n = 137$; ref. 22). Raw gene-expression data for Reddy and colleagues ($n = 775$; ref. 23) were obtained through The European Genome-phenome Archive (EGA; https://ega-archive.org/) at the European Bioinformatics Institute, Study ID: EGAS00001002606. Raw gene-expression data from Schmitz and colleagues ($n = 481$; ref. 24) were obtained from the NIH database of Genotypes and Phenotypes (dbGaP; RRID:SCR_002709), accession number: phs001444.v2.p1; The Genomic Variation in Diffuse Large B-Cell Lymphomas study was supported by the Intramural Research Program of the National Cancer Institute, NIH, Department of Health and Human Services. Clinical data associated with the GOYA dataset (GSE125966) were analyzed in collaboration with F. Hoffmann-La Roche Ltd.

Quantitative mfIHC and Scoring

Quantitative mfIHC was performed using sequential OPAL-TSA staining as described in detail previously (ref. 64; Supplementary Tables S16 and S17). Images were acquired using the Vectra 2 imager and analyzed using inForm2.4.8 (RRID: SCR_019155). DAPI nuclear staining and CD20 membrane staining were used to segment cells. The mean membrane intensity per cell was captured for CD20; the mean cytoplasm intensity per cell for BCL2; and the mean nuclear intensity per cell for both MYC and BCL6. For each

image, cells with a marker intensity above a given intensity threshold for that image were declared to be positive for that marker. A pathologist manually inspected each image to determine a reliable threshold for each marker that resulted in minimal false-positive and false-negative assignments. Images were examined in pseudocolor brightfield. All cohorts were evaluated in a tissue microarray (TMA) format; depending on tissue availability, between 1 and 9 high-power 700 × 500 μm imaging fields were captured and evaluated per patient in the NUH cohort, and two 1,400 × 1,000 μm fields were evaluated in the CMMC, SGH, MDA, and BCA cohorts. For each of the DLBCL whole-tissue sections, 5 to 8 700 × 500 μm imaging fields were evaluated. Once positivity thresholds were set for each marker per image, the quantitative image data (mean intensity per cell and the intensity threshold per marker for each image) were exported to calculate per-cell marker positivity and coexpression status. Subsequently, the percentage of cells within the CD20$^+$ B-cell compartment that were ascribed a given subpopulation (M+2+6+, M+2+6–, M+2–6+, M+2–6–, M-2+6+, M–2+6–, M–2–6+, and M–2–6–) were calculated for each patient. Scores from patients with multiple cores were a mean across all cores, weighted on cell number per core.

Survival Analysis

For unbiased survival associations, subpopulations percentage extents were evaluated as continuous variables in Cox PH models at 5% unit increments (albeit 0%–1% compromising the first unit, followed by 1%–5%, 5%–10%, etc.). HRs are displayed per unit (of 5% extent). Variables satisfied proportional hazards assumptions. To leverage on the multicohort design of this study, associations of each subpopulation extent were evaluated in a univariate model in individual cohorts, which was followed by effect size and *P* value pooling. Effect sizes were pooled by a random-effects model to mitigate interstudy heterogeneity, and Paule–Mandel heterogeneity variance estimator was applied due to the small number of cohorts. The pooled *P* values were adjusted for multiple hypothesis testing using Bonferroni correction (8 hypotheses). Tests were performed using the R "survival" package (RRID: SCR_021137) and pooled using the R "poolr" package. Multivariate Cox PH models were performed in SPSS 23 (RRID: SCR_002865). For Kaplan–Meier analyses, a log-rank test was performed using GraphPad Prism 9 (RRID: SCR_000306). Cohorts were dichotomized at an optimal cutoff in exploratory analyses, and subsequently, analyses were dichotomized at an established positivity threshold of ≥15% of M+2+6– extent (actual or using the metric). Statistical tests were two-sided and $P \le 0.05$ was considered statistically significant.

Probabilistic Inference of Colocalization

Assuming the independent distribution of positivity between MYC, BCL2, and BCL6, a probability-based algorithm using single oncogene scores was derived to predict the percentage extent of subpopulations, i.e., permutations of MYC, BCL2, BCL6-positivity and -negativity in CD20$^+$ cells in DLBCL samples:

$$P(A=a,B=b,C=c)=(P_A)^a(1-P_A)^{1-a}(P_B)^b(1-P_B)^{1-b}(P_C)^c(1-P_c)^{1-c}$$

$$a,b,c \in \{0,1\}$$

$$0 \le P_A \le 1$$

$$0 \le P_B \le 1$$

$$0 \le P_C \le 1$$

$$P-extent\,(\cdot 100\%), A-MYC, B-BLC2, C-BCL6$$

Mapping of mRNA Expression Data into Percentage Extent

Transformation of MYC, BCL2, and BCL6 mRNA levels into predicted percentage extent values requires an initial transformation step to map percentile points from RNA data onto protein data distributions, similar in principle to QQ plots where data are transformed into equivalent Gaussian data. For each protein marker aggregated across five mfIHC cohorts ($n = 712$), the empirical cumulative distribution function (eCDF) of mfIHC-based MYC/BCL2/BCL6 percentage extents was estimated as the benchmark distribution. With a biological assumption that mRNA expression is correlated with protein percentage scores for these oncogenes (16), we perform eCDF mapping (matching percentile points in the eCDF of mRNA measurements to those in the eCDF of the corresponding protein percentage scores), and then convert the quantitative mRNA score into an inferred single-marker percentage extent from the mRNA mapped CDF (mCDF). Subpopulation extents are subsequently inferred from the mapped single-marker mRNA values using the probabilistic cellular co-occurrence assumption.

To this end, eCDF of the corresponding mRNA expression levels was obtained from the individual subjects in the external datasets. Both eCDFs, mfIHC and mRNA, were smoothed by a Gaussian kernel smoother with a bandwidth parameter set at 1% of the entire range, and the percentile points on the smoothed CDFs were matched between the two datasets. Because eCDF is a monotone increasing function, this operation guarantees one-to-one mapping between the two eCDFs, and we used this map to translate the mRNA measurements into the approximate protein percentages across the individual subjects in the external dataset. The mapping procedure is performed independently for each mRNA dataset with the consensus protein eCDF to mitigate batch effects between datasets that are created through different technologies, and also thus retaining the mRNA cohorts as independent datasets.

The mapping source code for this approximation is available at GitHub (RRID: SCR_002630): https://github.com/MichalMarekHoppe/Patterns-of-oncogene-co-expression-at-single-cell-resolution-influence-survival-in-lymphoma.

Correlation of Gene Expression and M+2+6– Metric

Preprocessed gene-expression matrices submitted by the original authors were used for microarray-based datasets (Sha and colleagues, ref. 17; McCord and colleagues, ref. 63; Visco and colleagues, ref. 18; Lenz and colleagues, ref. 19; and Dubois and colleagues, ref. 20) and RNA-seq datasets were processed in-house (Reddy and colleagues, ref. 23; Schmitz and colleagues, ref. 24). Analyses were done for a consensus of 15,314 genes annotated in-house. Barrans and colleagues (21) and Chapuy and colleagues (22) data were not evaluated in exploratory analyses due to a lower number of genes in the original mapping and lower number of samples. Standardized gene expression was correlated with the M+2+6– metric (Spearman correlation) – Spearman rho with 95% confidence intervals was obtained using the R "DescTools" package and results were pooled using the R "poolr" package. Genes with a pooled Spearman rho value of ≥0.2 and a false discovery rate (FDR) ≤0.001 were considered hits in this analysis.

Differential Gene Expression in Primary GC B Cells

Total RNA was isolated from primary transduced B cells using the TRIzol extraction method. Insert cDNA library creation (250–300 bp eukaryotic mRNA) and standard polyA paired-end sequencing on Illumina Hiseq-4000 (RRID: SCR_016386) PE150 was performed by NovogeneAIT. Raw sequencing files were processed using standard pipelines available publicly on the CSI NGS Portal (65). Gene

expression of 18,252 genes was compared between four transduced GC B cells samples of M+2+6+ and four M+2+6– samples from independent donors (see "Generation of immortalized patient-derived GC B cells, and *CCDN2* analysis" section for details on GC B-cell transduction and refs. 27 and 66). FDR of a two-sided *t* test was used to define differently expressed genes. The R package "stats" was used to perform the *t* test. Hits were defined by meeting an arbitrary dynamic threshold criterion defined by the rational function (see the dashed line Fig. 5F):

$$-\log_{10}(\mathrm{FDR}_{\mathrm{gene}}) \geq \frac{-\log_{10}(\mathrm{FDR}_{\mathrm{sig}})^{\frac{5}{|\mathrm{FC}|}}}{\sqrt{\frac{|\mathrm{FC}|}{5}}}$$

where $\mathrm{FDR}_{\mathrm{gene}}$ is the FDR value of the gene tested, $\mathrm{FDR}_{\mathrm{sig}}$ is an arbitrary threshold of significance of 0.05, and |FC| is the absolute value of $\log_2$ fold change difference between mean expression in M+2+6– and M+2+6+ samples.

Generation of Immortalized Patient-Derived GC B Cells, and CCDN2 Analysis

Discarded tonsil tissue was collected after tonsillectomy at Addenbrooke's ENT Department, Cambridge, UK, with written informed consent from the patient's parent/guardian. Ethical approval for human tissue use was granted by the Health Research Authority Cambridgeshire Research Ethics Committee (REC no. 07/MRE05/44). Human primary GC B cells were isolated from fresh tonsils with Human B-cell Negative Selection Isolation Kit II (MACS, Miltenyi Biotec, cat. no. 130-091-151) supplemented with anti-IgD and anti-CD44 antibodies as described previously (27, 66). GC B cells were frozen immediately after isolation. Tissue from two female and one male donor ages 4 to 5 years was collected in September 2018. As these were primary cells, authentication and *Mycoplasma* testing were not performed.

After thawing, cells were cultured *in vitro* on irradiated YK6-CD40lg-IL21 follicular dendritic feeder cells in Advanced RPMI-1640 (Invitrogen, cat. no. 12633020) supplemented with 20% Gibco FCS (Thermo Fisher Scientific, cat. no. 10270-106) and 1× Gibco penicillin–streptomycin–glutamine (from 100×, Thermo Fisher Scientific, cat. no. 10378016) as previously described (27, 66). GC B cells were passaged 1 to 3 times before they were stably transduced with BCL6-T2A-BCL2 (27) and MYC-IRES-GFP retrovirus according to the protocol described previously (27, 66). *MYC* was cloned into the MSCV-IRES-GFP plasmid (RRID: Addgene_20672) to create the MSCV-MYC-IRES-GFP construct expressing MYC and GFP. The pBMN-IRES-Lyt2 (EV) retroviral vector was a kind gift from Dr. Louis Staudt, National Cancer Institute, USA; *CCND2* was cloned into pBMN-IRES-Lyt2 to create the pBMN-CCND2-IRES-Lyt2 construct. Subsequently, cells were stably transduced with either EV (pBMN-IRES-Lyt2) or CCND2-IRES-Lyt2 lentivirus. The live-cell fractions of EV or *CCND2*-transduced cells were assessed by flow cytometry after staining for Lyt2 with antimurine CD8a-APC antibody (Miltenyi Biotec; cat. # 130-117-776, RRID: AB_2728039) and observed for 30 days. GFP protein was quantified by flow cytometry as a proxy for MYC expression.

Analysis of CCND2 in scRNA-seq Data

For scRNA-seq experiments, primary human GC B cells from a single donor were transduced with BCL2 and BCL6, or BCL2 and MYC (see "Generation of immortalized patient-derived GC B cells, and *CCDN2* analysis" section for transduction details and refs. 27 and 66). Seven days after transduction, cells were pooled and subjected to scRNA-seq using the 10X Genomics platform. Fresh transduced GC B cells from the same donor were spiked into the sequencing reaction. Raw fastq files were processed using cellranger (v3.1.0); the alignment was performed against the GRCh38-3.0.0 version of the *Homo sapiens* reference genome; the quantification and filtering of cells were done using default parameters.

Further filtering applied on the expression matrix was based on upper and lower bounds on the distributions of counts and features, and on the proportions of reads incident to mitochondrial DNA (mt%) and ribosomal genes (rp%). Cells with values outside these ranges (counts per cell/sequencing depth >5,000, number of features <2,000 or >8,000, mt%>15% rp%>50%) were considered outliers and excluded from downstream analyses. Post filtering, mitochondrial and ribosomal genes were excluded from the expression matrix. The expression matrix was log-normalized using the NormalizeData function in the Seurat package (v3.2.2; ref. 67).

Dimensionality reductions [PCA followed by Uniform Manifold Approximation and Projection (UMAP)], as well as clustering (Louvain method) were conducted in Seurat; the optimal number of clusters was selected based on default clustering parameters. Following an assessment of the stability of clustering results, for the subsequent steps, we focused on the 2,000 most abundant genes, determined across all cells in the dataset. Marker genes, determined for each cluster against all other genes, were identified based on differential expression tests (in Seurat) i.e., genes with $\log_2$(FC) >0.25, and adjusted *P* values, under a Benjamini–Hochberg multiple testing correction, less than 0.05. The data were also made available as a Shiny app (RRID: SCR_022756; ref. 68) at the link: https://bioinf.stemcells.cam.ac.uk/shiny/hodson/MYC-BCL2- BCL6_project.

Reprocessing of scRNA-seq Datasets

In this study, six DLBCL samples from two publicly available DLBCL scRNA-seq datasets were utilized. Dataset GSE182434 (33), containing sample pairs of B cells and non B cells for 3 ABC-DLBCL tumors and 1 GCB DLBCL tumor, was downloaded from the GEO database. B-cell samples were provided with annotations containing cell type and condition (i.e., tumor or normal). Only cells annotated as tumor and B cells were used for the analysis. DLBCL scRNA-seq dataset generated by Roider and colleagues (34) was downloaded from the heiDATA database (https://heidata.uni-heidelberg.de) from the link https://doi.org/10.11588/data/VRJUNV. The dataset contained four GC-derived DLBCLs, two of which were transformed follicular lymph nodes, which were excluded from the analysis and one nongerminal center-derived DLBCL. Upon further examination of the shared nearest neighbor clusters of the samples (original paper, Fig. 3B; ref. 34), one of the GC-derived DLBCL samples clustered closely with the transformed follicular lymph node cluster and was hence excluded from the analysis. Samples were provided with cell annotations denoting malignant B cells, healthy B cells, and myeloid cells. Only cells annotated as malignant B cells were used for the analysis. In total, 8,235 cells were used for subsequent analysis.

Seurat (v4.3.0; ref. 69) was used for the analysis of the single-cell datasets. All functions were run with default parameters unless specified otherwise. Low-quality cells, defined by <200 genes per cell and >10% mitochondrial genes, were excluded from the analysis. Genes expressed in less than 3 cells were excluded from the analysis. The two datasets were integrated using the Seurat Integration protocol for data normalized with the "sctransform" method (RRID: SCR_022146; ref. 70); https://satijalab.org/seurat/articles/integration_introduction.html#performing-integration-on-datasets-normalized-with-sctransform-1. The data were integrated with each study and treated as a batch. Default parameters were used with 2,000 genes being used for the SelectIntegrationFeatures() function. Following this, based on the count data, each cell was assigned an

expression status with double expressors being defined as below and the rest assigned as others.

$$\text{double expressors } \{[\text{MYC}] > 0 \text{ and } [\text{BCL2}] > 0 \text{ and } [\text{BCL6}] = 0\}$$
$$\text{with } [\text{MYC}], [\text{BCL2}], [\text{BCL6}] \text{ denoting count values}$$

Differential gene-expression analysis was conducted using the FindMarkers() function in Seurat with ident.1 being M+2+6− cells and ident.2 being others. Nonparametric Wilcoxon rank sum test was used for the FindMarkers function. Next, functional enrichment was conducted using the gProfiler2 (v0.2.1; RRID: SCR_018190; ref. 71) package utilizing the WikiPathways (RRID:SCR_002134; ref. 72) database as source via the gost() function. FDR was the correction method used for multiple testing and all enriched pathways survived an FDR of 5%. Upregulated genes (defined by avg_log2FC >0 and $P < 0.05$) from the differential expression analysis were interrogated in gprofiler2.

Software and Statistical Analysis

Graphical representations of data were generated in either GraphPad Prism 9 (RRID: SCR_000306) or R (RRID: SCR_001905). All relevant statistical tests were performed in GraphPad Prism 9 unless indicated otherwise. For Supplementary Table S3, Mann–Whitney and Kruskal–Wallis tests were performed using the R "stats" package; pooling of *P* values was performed using the R "poolr" package.

Supplementary Methods

Additional details on IHC, spatial analysis, and unsupervised clustering can be found in the Supplementary Methods section within the Supplementary Appendix.

Data Sharing Statement

- RNA-seq data generated from GC B cells overexpressing either MYC and BCL2 or MYC, BCL2, and BCL6 are deposited to GEO (RRID: SCR_005012; ref. 62) under the accession number: GSE203446.
- scRNA-seq data of GC primary B cells transduced either with BCL2 and MYC (MYC-transduced) or BCL2 and BCL6 (BCL6-transduced), or untransduced, have been made available as a Shiny app (68): https://bioinf.stemcells.cam.ac.uk/shiny/hodson/MYC-BCL2-BCL6_project.
- Source code for custom pipelines has been deposited to GitHub at the link: https://github.com/MichalMarekHoppe/Patterns-of-oncogene-co-expression-at-single-cell-resolution-influence-survival-in-lymphoma, and includes:
  - "Mapping of MYC, BCL2, BCL6 mRNA DLBCL cohort expression data"
  - "Pair correlation function clustering"
  - "Spatial analysis – calculating delta between cellular phenotypes"
  - "DLBCL scRNA-seq re-analysis"
- Publicly available gene-expression data for Sha and colleagues (17), GSE117556; McCord and colleagues (63), GSE125966; Visco and colleagues (18), GSE31312; Lenz and colleagues (19), GSE10846; Dubois and colleagues (20), GSE87371; Barrans and colleagues (21), GSE32918; and Chapuy and colleagues (22), GSE98588; can be found on GEO under the indicated accession numbers.
- Restricted access gene-expression data can be found at the following repositories under the respective accession numbers:
  - Reddy and colleagues (23); The EGA (https://ega-archive.org/), Study ID: EGAS00001002606
  - Schmitz and colleagues (24); NIH database of Genotypes and Phenotypes (dbGaP; RRID:SCR_002709), accession number: phs001444.v2.p1
- GEP-derived COO scores and LymphGen genetic subtype classifications for the BCA cohort and for Schmitz and colleagues were obtained from (49) and (24), respectively.
- Publicly available scRNA-seq data for Steen and colleagues (33) can be accessed through GEO under the accession number GSE182434; for Roider and colleagues (34) can be accessed through the heiDATA database (https://heidata.uni-heidelberg.de) from https://doi.org/10.11588/data/VRJUNV.

Authors' Disclosures

N.F. Grigoropoulos reports other support from Roche Therapeutics and AstraZeneca outside the submitted work. A. Bottos is an employee of and has equity ownership interests in F. Hoffmann-La Roche Ltd. H.F.P. Runge reports personal fees from Cambridge Nucleomics outside the submitted work. D.J. Hodson reports grants from AstraZeneca during the conduct of the study. D.W. Scott reports personal fees from AbbVie, AstraZeneca, and Incyte, and grants from Janssen and Roche outside the submitted work, as well as a patent for determining molecular subtypes of aggressive B-cell lymphomas using gene expression issued and licensed to NanoString. A.D. Jeyasekharan reports grants from the National Medical Research Council, the Ministry of Education, and the National Research Foundation Singapore and personal fees from Roche during the conduct of the study, as well as personal fees from Gilead, MSD, IQVIA, Turbine, and Antengene, and personal fees and other support from AstraZeneca and Janssen outside the submitted work. No disclosures were reported by the other authors.

Authors' Contributions

M.M. Hoppe: Conceptualization, formal analysis, methodology, writing–original draft, writing–review and editing, underlying data verification, interpretation of findings. **P. Jaynes:** Writing–original draft, writing–review and editing. **F. Shuangyi:** Formal analysis, investigation, methodology. **Y. Peng:** Formal analysis, investigation. **S. Sridhar:** Formal analysis. **P.M. Hoang:** Investigation. **C.X. Liu:** Data curation, project administration. **S. De Mel:** Data curation, project administration. **L. Poon:** Data curation, project administration. **E.H.L. Chan:** Data curation. **J. Lee:** Data curation. **C.K. Ong:** Data curation. **T. Tang:** Data curation. **S.T. Lim:** Data curation. **C. Nagarajan:** Data curation. **N.F. Grigoropoulos:** Data curation. **S.-Y. Tan:** Formal analysis. **S. Hue:** Formal analysis. **S. Chang:** Formal analysis. **S. Chuang:** Formal analysis. **S. Li:** Formal analysis. **J.D. Khoury:** Data curation. **H. Choi:** Formal analysis, methodology. **C. Harris:** Formal analysis. **A. Bottos:** Project administration. **L.J. Gay:** Formal analysis. **H.F.P. Runge:** Formal analysis. **I. Moutsopoulos:** Formal analysis. **I. Mohorianu:** Formal analysis. **D.J. Hodson:** Formal analysis, interpretation of the findings. **P. Farinha:** Formal analysis. **A. Mottok:** Formal analysis. **D.W. Scott:** Data curation. **J.J. Pitt:** Data curation, writing–review and editing. **J. Chen:** Formal analysis. **G. Kumar:** Formal analysis. **K. Kannan:** Formal analysis. **W.J. Chng:** Methodology, writing–review and editing. **Y. Chee:** Data curation. **S.-B. Ng:** Methodology, writing–review and editing, interpretation of findings. **C. Tripodo:** Methodology, interpretation of findings. **A.D. Jeyasekharan:** Conceptualization, formal analysis, methodology, writing–original draft, writing–review and editing, interpretation of findings, underlying data verification

Acknowledgments

The authors acknowledge a Yong Siew Yoon Research Grant to A.D. Jeyasekharan from the National University Cancer Institute, Singapore toward the purchase of the Vectra 2 multispectral imaging system microscope. A.D. Jeyasekharan was supported by the Singapore Ministry of Health's National Medical Research Council Clinician Scientist Award (MOH-000715-00). Work in A.D. Jeyasekharan's laboratory is funded by a core grant from the Cancer Science Institute of Singapore, National University of Singapore through the National Research Foundation Singapore, and the

Singapore Ministry of Education under its Research Centres of Excellence initiative. S.-B. Ng was supported by the National Medical Research Council Senior Investigator Clinician Scientist Award (MOH-001104). Work by W.J. Chng, S.-B. Ng, S.-Y. Tan, T. Tang, C.K. Ong, S.T. Lim, N.F. Grigoropoulos, and A.D. Jeyasekharan was also supported by the National Medical Research Council Open Fund Large Collaborative Grant (SYMPHONY; NMRC OF-LCG-18May-0028). The work in D.J. Hodson's lab was supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. D.J. Hodson was supported by a fellowship from Cancer Research UK (CRUK; RCCFEL/100072) and received core funding from Wellcome (203151/Z/16/Z) to the Wellcome-MRC Cambridge Stem Cell Institute and from the CRUK Cambridge Centre (A25117). For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. C. Tripodo was supported by the Italian Foundation for Cancer Research (AIRC) Investigator Grant IG ID.22145; 5 × 1000 Grant ID.22759.

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

Note

Supplementary data for this article are available at Cancer Discovery Online (http://cancerdiscovery.aacrjournals.org/).

Received September 7, 2022; revised December 29, 2022; accepted February 13, 2023; published first April 18, 2023.

REFERENCES

1. Marusyk A, Janiszewska M, Polyak K. Intratumor heterogeneity: The rosetta stone of therapy resistance. Cancer Cell 2020;37:471–84.
2. Maley CC, Aktipis A, Graham TA, Sottoriva A, Boddy AM, Janiszewska M, et al. Classifying the evolutionary and ecological features of neoplasms. Nat Rev Cancer 2017;17:605–19.
3. Miao Y, Medeiros LJ, Li Y, Li J, Young KH. Genetic alterations and their clinical implications in DLBCL. Nat Rev Clin Oncol 2019;16:634–52.
4. Casey SC, Baylot V, Felsher DW. The MYC oncogene is a global regulator of the immune response. Blood 2018;131:2007–15.
5. Vaux DL, Cory S, Adams JM. Bcl-2 gene promotes haemopoietic cell survival and cooperates with c-myc to immortalize pre-B cells. Nature 1988;335:440–2.
6. Cattoretti G, Pasqualucci L, Ballon G, Tam W, Nandula SV, Shen Q, et al. Deregulated BCL6 expression recapitulates the pathogenesis of human diffuse large B cell lymphomas in mice. Cancer Cell 2005;7:445–55.
7. Johnson NA, Slack GW, Savage KJ, Connors JM, Ben-Neriah S, Rogic S, et al. Concurrent expression of MYC and BCL2 in diffuse large B-cell lymphoma treated with rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone. J Clin Oncol 2012;30:3452–9.
8. Hu S, Xu-Monette ZY, Tzankov A, Green T, Wu L, Balasubramanyam A, et al. MYC/BCL2 protein coexpression contributes to the inferior survival of activated B-cell subtype of diffuse large B-cell lymphoma and demonstrates high-risk gene expression signatures: a report from the international DLBCL rituximab-CHOP consortium program. Blood 2013;121:4021–31.
9. Horn H, Ziepert M, Becher C, Barth TFE, Bernd H-W, Feller AC, et al. MYC status in concert with BCL2 and BCL6 expression predicts outcome in diffuse large B-cell lymphoma. Blood 2013;121:2253–63.
10. Dominguez-Sola D, Victora GD, Ying CY, Phan RT, Saito M, Nussenzweig MC, et al. The proto-oncogene MYC is required for selection in the germinal center and cyclic reentry. Nat Immunol 2012;13:1083–91.
11. Meriranta L, Pasanen A, Alkodsi A, Haukka J, Karjalainen-Lindsberg M-L, Leppä S. Molecular background delineates outcome of double protein expressor diffuse large B-cell lymphoma. Blood Adv 2020;4:3742–53.
12. Gavagnin E, Owen JP, Yates CA. Pair correlation functions for identifying spatial correlation in discrete domains. Phys Rev E 2018;97:062104.
13. Kaufmann J, Biscio CAN, Bankhead P, Zimmer S, Schmidberger H, Rubak E, et al. Using the R package spatstat to assess inhibitory effects of microregional hypoxia on the infiltration of cancers of the head and neck region by cytotoxic T lymphocytes. Cancers (Basel) 2021;13(8). doi 10.3390/cancers13081924.
14. Li L, Li Y, Que X, Gao X, Gao Q, Yu M, et al. Prognostic significances of overexpression MYC and/or BCL2 in R-CHOP-treated diffuse large B-cell lymphoma: a systematic review and meta-analysis. Sci Rep 2018;8:6267.
15. Ennishi D, Mottok A, Ben-Neriah S, Shulha HP, Farinha P, Chan FC, et al. Genetic profiling of MYC and BCL2 in diffuse large B-cell lymphoma determines cell-of-origin–specific clinical impact. Blood 2017;129:2760–70.
16. Xia B, Zhang L, Guo SQ, Li XW, Qu FL, Zhao HF, et al. Coexpression of MYC and BCL-2 predicts prognosis in primary gastrointestinal diffuse large B-cell lymphoma. World J Gastroenterol 2015;21:2433–42.
17. Sha C, Barrans S, Cucco F, Bentley MA, Care MA, Cummin T, et al. Molecular high-grade B-cell lymphoma: defining a poor-risk group that requires different approaches to therapy. J Clin Oncol 2019;37:202–12.
18. Visco C, Li Y, Xu-Monette ZY, Miranda RN, Green TM, Li Y, et al. Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: a report from the international DLBCL rituximab-CHOP consortium program study. Leukemia 2012;26:2103–13.
19. Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, et al. Stromal gene signatures in large-B-cell lymphomas. N Engl J Med 2008;359:2313–23.
20. Dubois S, Viailly PJ, Bohers E, Bertrand P, Ruminy P, Marchand V, et al. Biological and clinical relevance of associated genomic alterations in MYD88 L265P and non-L265P-mutated diffuse large B-cell lymphoma: analysis of 361 cases. Clin Cancer Res 2017;23:2232–44.
21. Barrans SL, Crouch S, Care MA, Worrillow L, Smith A, Patmore R, et al. Whole genome expression profiling based on paraffin embedded tissue can be used to classify diffuse large B-cell lymphoma and predict clinical outcome. Br J Haematol 2012;159:441–53.
22. Chapuy B, Stewart C, Dunford AJ, Kim J, Kamburov A, Redd RA, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nat Med 2018;24:679–90.
23. Reddy A, Zhang J, Davis NS, Moffitt AB, Love CL, Waldrop A, et al. Genetic and functional drivers of diffuse large B cell lymphoma. Cell 2017;171:481–94.
24. Schmitz R, Wright GW, Huang DW, Johnson CA, Phelan JD, Wang JQ, et al. Genetics and pathogenesis of diffuse large B-cell lymphoma. N Engl J Med 2018;378:1396–407.
25. Vitolo U, Trněný M, Belada D, Burke JM, Carella AM, Chua N, et al. Obinutuzumab or rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone in previously untreated diffuse large B-cell lymphoma. J Clin Oncol 2017;35:3529–37.
26. Wright GW, Huang DW, Phelan JD, Coulibaly ZA, Roulland S, Young RM, et al. A probabilistic classification tool for genetic subtypes of diffuse large B cell lymphoma with therapeutic implications. Cancer Cell 2020;37:551–68.
27. Caeser R, Di Re M, Krupka JA, Gao J, Lara-Chica M, Dias JML, et al. Genetic modification of primary human B cells to model high-grade lymphoma. Nat Commun 2019;10:4543.

28. Cattoretti G, Chang CC, Cechova K, Zhang J, Ye BH, Falini B, et al. BCL-6 protein is expressed in germinal-center B cells. Blood 1995;86:45–53.
29. Hirata Y, Ogasawara N, Sasaki M, Mizushima T, Shimura T, Mizoshita T, et al. BCL6 degradation caused by the interaction with the C-terminus of pro-HB-EGF induces cyclin D2 expression in gastric cancers. Br J Cancer 2009;100:1320–9.
30. Shaffer AL, Yu X, He Y, Boldrick J, Chan EP, Staudt LM. BCL-6 represses genes that function in lymphocyte differentiation, inflammation, and cell cycle control. Immunity 2000;13:199–212.
31. Hans CP, Weisenburger DD, Greiner TC, Chan WC, Aoun P, Cochran GT, et al. Expression of PKC-beta or cyclin D2 predicts for inferior survival in diffuse large B-cell lymphoma. Mod Pathol 2005;18:1377–84.
32. 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;103:275–82.
33. Steen CB, Luca BA, Esfahani MS, Azizi A, Sworder BJ, Nabet BY, et al. The landscape of tumor cell states and ecosystems in diffuse large B cell lymphoma. Cancer Cell 2021;39:1422–37.
34. Roider T, Seufert J, Uvarovskii A, Frauhammer F, Bordas M, Abedpour N, et al. Dissecting intratumour heterogeneity of nodal B-cell lymphomas at the transcriptional, genetic and drug-response levels. Nat Cell Biol 2020;22:896–906.
35. Dekker JD, Park D, Shaffer AL, Kohlhammer H, Deng W, Lee B-K, et al. Subtype-specific addiction of the activated B-cell subset of diffuse large B-cell lymphoma to FOXP1. Proc Natl Acad Sci 2016;113:E577–E86.
36. Gómez-Abad C, Pisonero H, Blanco-Aparicio C, Roncador G, González-Menchén A, Martinez-Climent JA, et al. PIM2 inhibition as a rational therapeutic approach in B-cell lymphoma. Blood 2011;118:5517–27.
37. Care MA, Cocco M, Laye JP, Barnes N, Huang Y, Wang M, et al. SPIB and BATF provide alternate determinants of IRF4 occupancy in diffuse large B-cell lymphoma linked to disease heterogeneity. Nucleic Acids Res 2014;42:7591–610.
38. Ando K, Ajchenbaum-Cymbalista F, Griffin JD. Regulation of G1/S transition by cyclins D2 and D3 in hematopoietic cells. Proc Natl Acad Sci 1993;90:9571–5.
39. Sasaki Y, Jensen CT, Karlsson S, Jacobsen SEW. Enforced expression of cyclin D2 enhances the proliferative potential of myeloid progenitors, accelerates in vivo myeloid reconstitution, and promotes rescue of mice from lethal myeloablation. Blood 2004;104:986–92.
40. Pei Y, Singh RK, Shukla SK, Lang F, Zhang S, Robertson ES. Epstein-barr virus nuclear antigen 3C facilitates cell proliferation by regulating cyclin D2. J Virol 2018;92:e00663–18.
41. Herrera AF, Mei M, Low L, Kim HT, Griffin GK, Song JY, et al. Relapsed or refractory double-expressor and double-hit lymphomas have inferior progression-free survival after autologous stem-cell transplantation. J Clin Oncol 2017;35:24–31.
42. Green TM, Young KH, Visco C, Xu-Monette ZY, Orazi A, Go RS, et al. Immunohistochemical double-hit score is a strong predictor of outcome in patients with diffuse large B-cell lymphoma treated with rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone. J Clin Oncol 2012;30:3460–7.
43. Wang Y, Feng W, Liu P. Genomic pattern of intratumor heterogeneity predicts the risk of progression in early stage diffuse large B-cell lymphoma. Carcinogenesis 2019;40:1427–34.
44. Ye X, Wang L, Nie M, Wang Y, Dong S, Ren W, et al. A single-cell atlas of diffuse large B cell lymphoma. Cell Rep 2022;39:110713.
45. Iqbal J, Greiner TC, Patel K, Dave BJ, Smith L, Ji J, et al. Distinctive patterns of BCL6 molecular alterations and their functional consequences in different subgroups of diffuse large B-cell lymphoma. Leukemia 2007;21:2332–43.
46. Xu-Monette ZY, Wei L, Fang X, Au Q, Nunns H, Nagy M, et al. Genetic subtyping and phenotypic characterization of the immune microenvironment and MYC/BCL2 double expression reveal heterogeneity in diffuse large B-cell lymphoma. Clin Cancer Res 2022;28:972–83.
47. Tilly H, Morschhauser F, Sehn LH, Friedberg JW, Trněný M, Sharman JP, et al. Polatuzumab vedotin in previously untreated diffuse large B-cell lymphoma. N Engl J Med 2021;386:351–63.
48. Rosenthal A, Younes A. High grade B-cell lymphoma with rearrangements of MYC and BCL2 and/or BCL6: double hit and triple hit lymphomas and double expressing lymphoma. Blood Rev 2017;31:37–42.
49. Ennishi D, Jiang A, Boyle M, Collinge B, Grande BM, Ben-Neriah S, et al. Double-hit gene expression signature defines a distinct subgroup of germinal center B-cell-like diffuse large B-cell lymphoma. J Clin Oncol 2019;37:190–201.
50. Ott G, Rosenwald A, Campo E. Understanding MYC-driven aggressive B-cell lymphomas: pathogenesis and classification. Blood 2013;122:3884–91.
51. Sarkozy C, Traverse-Glehen A, Coiffier B. Double-hit and double-protein-expression lymphomas: aggressive and refractory lymphomas. Lancet Oncol 2015;16:e555–e67.
52. Johnson NA, Savage KJ, Ludkovski O, Ben-Neriah S, Woods R, Steidl C, et al. Lymphomas with concurrent BCL2 and MYC translocations: the critical factors associated with survival. Blood 2009;114:2273–9.
53. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the evolution of non–small-cell lung cancer. N Engl J Med 2017;376:2109–21.
54. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 2012;366:883–92.
55. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. Cell 2017;168:613–28.
56. Aparicio S, Caldas C. The implications of clonal genome evolution for cancer medicine. N Engl J Med 2013;368:842–51.
57. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell 2012;148:886–95.
58. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med 2018;24:1277–89.
59. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell 2018;174:1293–308.
60. Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. Cell 2021;184:2239–54.
61. Iacobuzio-Donahue CA, Litchfield K, Swanton C. Intratumor heterogeneity reflects clinical disease course. Nat Cancer 2020;1:3–6.
62. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10.
63. McCord R, Bolen CR, Koeppen H, Kadel EE 3rd, Oestergaard MZ, Nielsen T, et al. PD-L1 and tumor-associated macrophages in de novo DLBCL. Blood Adv 2019;3:531–40.
64. Hoppe MM, Jaynes P, Wardyn JD, Upadhyayula SS, Tan TZ, Lie S, et al. Quantitative imaging of RAD51 expression as a marker of platinum resistance in ovarian cancer. EMBO Mol Med 2021;13:e13366.
65. An O, Tan K-T, Li Y, Li J, Wu C-S, Zhang B, et al. CSI NGS portal: an online platform for automated NGS data analysis and sharing. Int J Mol Sci 2020;21:3828.
66. Caeser R, Gao J, Di Re M, Gong C, Hodson DJ. Genetic manipulation and immortalized culture of ex vivo primary human germinal center B cells. Nat Protoc 2021;16:2499–519.
67. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888–902.
68. Ouyang JF, Kamaraj US, Cao EY, Rackham OJL. ShinyCell: simple and sharable visualization of single-cell gene expression data. Bioinformatics 2021;37:3374–6.

69. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184:3573–87.
70. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 2019;20:296.
71. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. gprofiler2 – an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler [version 2; peer review: 2 approved]. F1000Research 2020;9:ELIXIR–709.
72. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res 2017;46:D661–D7.