

## ARTICLE

# Single-cell analysis of germinal-center B cells informs on lymphoma cell of origin and outcome

Antony B. Holmes<sup>1\*</sup>, Clarissa Corinaldesi<sup>1\*</sup>, Qiong Shen<sup>1</sup>, Rahul Kumar<sup>1</sup>, Nicolo Compagno<sup>1</sup>, Zhong Wang<sup>2</sup>, Mor Nitzan<sup>7</sup>, Eli Grunstein<sup>3</sup>, Laura Pasqualucci<sup>1,2,6</sup>, Riccardo Dalla-Favera<sup>1,2,4,5,6</sup>, and Katia Basso<sup>1,2</sup>

**In response to T cell-dependent antigens, mature B cells are stimulated to form germinal centers (GCs), the sites of B cell affinity maturation and the cell of origin (COO) of most B cell lymphomas. To explore the dynamics of GC B cell development beyond the known dark zone and light zone compartments, we performed single-cell (sc) transcriptomic analysis on human GC B cells and identified multiple functionally linked subpopulations, including the distinct precursors of memory B cells and plasma cells. The gene expression signatures associated with these GC subpopulations were effective in providing a sc-COO for ~80% of diffuse large B cell lymphomas (DLBCLs) and identified novel prognostic subgroups of DLBCL.**

## Introduction

Germinal centers (GCs) are histological structures that form in the secondary lymphoid organs in response to engagement of mature naive B cells by the antigen. They represent the sites of antibody affinity maturation, a process based on multiple rounds of B cell receptor (BCR) editing by somatic hypermutation (SHM) followed by affinity-driven selection. In the GC, B cells undergo class switch recombination, although this process can also occur outside the GC (Roco et al., 2019). The GC includes two functionally distinct compartments: the dark zone (DZ), representing the site of intense B cell proliferation and SHM, and the light zone (LZ), where affinity selection occurs (Cyster and Allen, 2019; De Silva and Klein, 2015; Mesin et al., 2016; Victora and Nussenzweig, 2012). GC B cells recirculate between the DZ and LZ compartments, undergoing multiple rounds of BCR editing and affinity selection (Calado et al., 2012; Dominguez-Sola et al., 2012; Ersching et al., 2017; Finkin et al., 2019; Victora et al., 2010). GC B cell selection in the LZ requires BCR engagement and B-T cell interaction that trigger the signaling pathways driving the DZ reentry or differentiation toward post-GC memory B cells and plasma cells. Multiple lines of evidence support the notion that plasma cell differentiation is favored by the acquisition of high-affinity BCRs, while memory B cells rise from lower-affinity B cells (Phan et al., 2006; Shinnakasu et al., 2016). However, the molecular mechanisms driving differentiation into memory B cells or plasma cells are not fully understood.

GC B cells also represent the normal counterpart of most B cell non-Hodgkin lymphomas, including Burkitt lymphoma, follicular lymphoma, and diffuse large B cell lymphoma (DLBCL). Although they share a common GC B cell precursor, these lymphomas appear to originate from cells at different stages of the GC reaction and develop through distinct pathogenetic mechanisms (Basso and Dalla-Favera, 2015; Pasqualucci and Dalla-Favera, 2018; Shaffer et al., 2012). The cell-of-origin (COO) classification has identified two subtypes of DLBCL: the GC B cell-like (GCB) and the activated B cell-like (ABC) DLBCL, which appear to originate from LZ B cells and cells committed to plasmablast (PBL) differentiation, respectively (Alizadeh et al., 2000). The COO classification is relevant for its association with distinct clinical responses to therapy, with GCB cases being, in general, less aggressive than ABC cases (Alizadeh et al., 2000; Wright et al., 2003).

Although DZ and LZ compartmentalization has been instrumental in understanding GC biology, it most likely represents an oversimplification of the complex dynamics of proliferation, trafficking, and differentiation involving B cells within the GC. Toward a better understanding of the GC reaction, here we applied genome-wide single-cell (sc) RNA profiling to further dissect the heterogeneity of GC B cells. We identified multiple functionally linked subpopulations, as well as the precursors of memory B cells and plasma cells. Then, the gene signatures associated with these GC B cell subpopulations were used to

<sup>1</sup>Institute for Cancer Genetics, Columbia University, New York, NY; <sup>2</sup>Department of Pathology and Cell Biology, Columbia University, New York, NY; <sup>3</sup>Department of Otolaryngology Head and Neck Surgery, Columbia University, New York, NY; <sup>4</sup>Department of Microbiology and Immunology, Columbia University, New York, NY; <sup>5</sup>Department of Genetics and Development, Columbia University, New York, NY; <sup>6</sup>The Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY; <sup>7</sup>Harvard University, Cambridge, MA.

\*A.B. Holmes and C. Corinaldesi contributed equally to this paper; Correspondence to Katia Basso: [kb451@cumc.columbia.edu](mailto:kb451@cumc.columbia.edu); Riccardo Dalla-Favera: [rd10@cumc.columbia.edu](mailto:rd10@cumc.columbia.edu).

© 2020 Holmes et al. This article is distributed under the terms of an Attribution–Noncommercial–Share Alike–No Mirror Sites license for the first six months after the publication date (see <http://www.rupress.org/terms/>). After six months it is available under a Creative Commons License (Attribution–Noncommercial–Share Alike 4.0 International license, as described at <https://creativecommons.org/licenses/by-nc-sa/4.0/>).

further dissect the COO of DLBCL and to interrogate the prognostic significance of the newly identified sc-COO subgroups.

## Results

### Identification of GC B cell subpopulations by sc-transcriptomic analysis

Human GC B cells were isolated from tonsil tissue of two donors by cell sorting CD3<sup>-</sup>/IgD<sup>-</sup>/CD38<sup>+</sup> cells, while DZ and LZ B cells were purified based on the expression of the CXCR4 and CD83 markers, as previously reported (Allen et al., 2004; Caron et al., 2009; Victora et al., 2012, 2010; Fig. S1 A). sc-RNA sequencing (sc-RNAseq) analysis was performed by 10x Genomics technology on ~4,500 cells from each of two donors. RNAseq libraries were analyzed by the 10x Genomics computational pipeline, dimensional reduction was performed using the Uniform Manifold Approximation and Projection (UMAP) algorithm (Becht et al., 2018), and clusters were identified by the PhenoGraph algorithm (Levine et al., 2015; Fig. S1 B).

The independent analysis of GC B cells from the two donors showed that they displayed remarkably similar structures and shared the gene signatures associated with distinct subpopulations (Fig. S1, C and D). Given the consistency between donors, we applied batch correction using the Seurat package (Butler et al., 2018) and merged the cells of both donors into a single dataset. The analysis of the merged dataset, including 8,368 GC B cells, identified 13 clusters, which were annotated based on their gene expression signatures (Fig. 1 A). An inverted expression pattern of CXCR4 and CD83 characterized clusters representative of DZ and LZ B cells (Fig. 1 B; Allen et al., 2004; Caron et al., 2009; Victora et al., 2012, 2010). Several clusters displayed intermediate expression of DZ/LZ markers and may represent intermediate stages in the GC differentiation process (Fig. 1 C; Milpied et al., 2018). Two additional discrete clusters were preliminarily identified as precursors of plasma cells and memory B cells based on their gene signatures, including expression of *PRDMI* (Angelin-Duclos et al., 2000; Falini et al., 2000) and *CCR6* (Suan et al., 2017), respectively (Fig. 1, A–C; see dedicated paragraphs for additional details).

In addition, we performed sc assessment of a few surface protein markers in parallel with RNAseq using the cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) technology in an independent donor. Most subpopulations identified by the sc-RNAseq analysis were detected using the transcriptional profiles of the 8,871 cells analyzed by CITE-seq (Fig. S1 E). The results showed that, as expected, the clusters identified as DZ and LZ cells expressed both RNA and surface protein for CXCR4 and CD83, respectively. In the intermediate clusters, however, the CXCR4 and CD83 protein expressions were less pronounced than those of their respective transcripts, consistent with various levels of post-transcriptional regulation (Fig. S1 E). Surface markers associated with PBL (CD9) and precursor memory B cells (PreM; *CCR6*) were also detected in the respective populations (Fig. S1 E; see dedicated paragraph for additional details). These data show that the populations identified based on transcriptional profiles display consistent protein expression of known markers.

Overall, these results show that sc-RNAseq analysis can capture all known GC B cell subpopulations and also inform on additional GC heterogeneity.

### Proliferative compartments within the GC

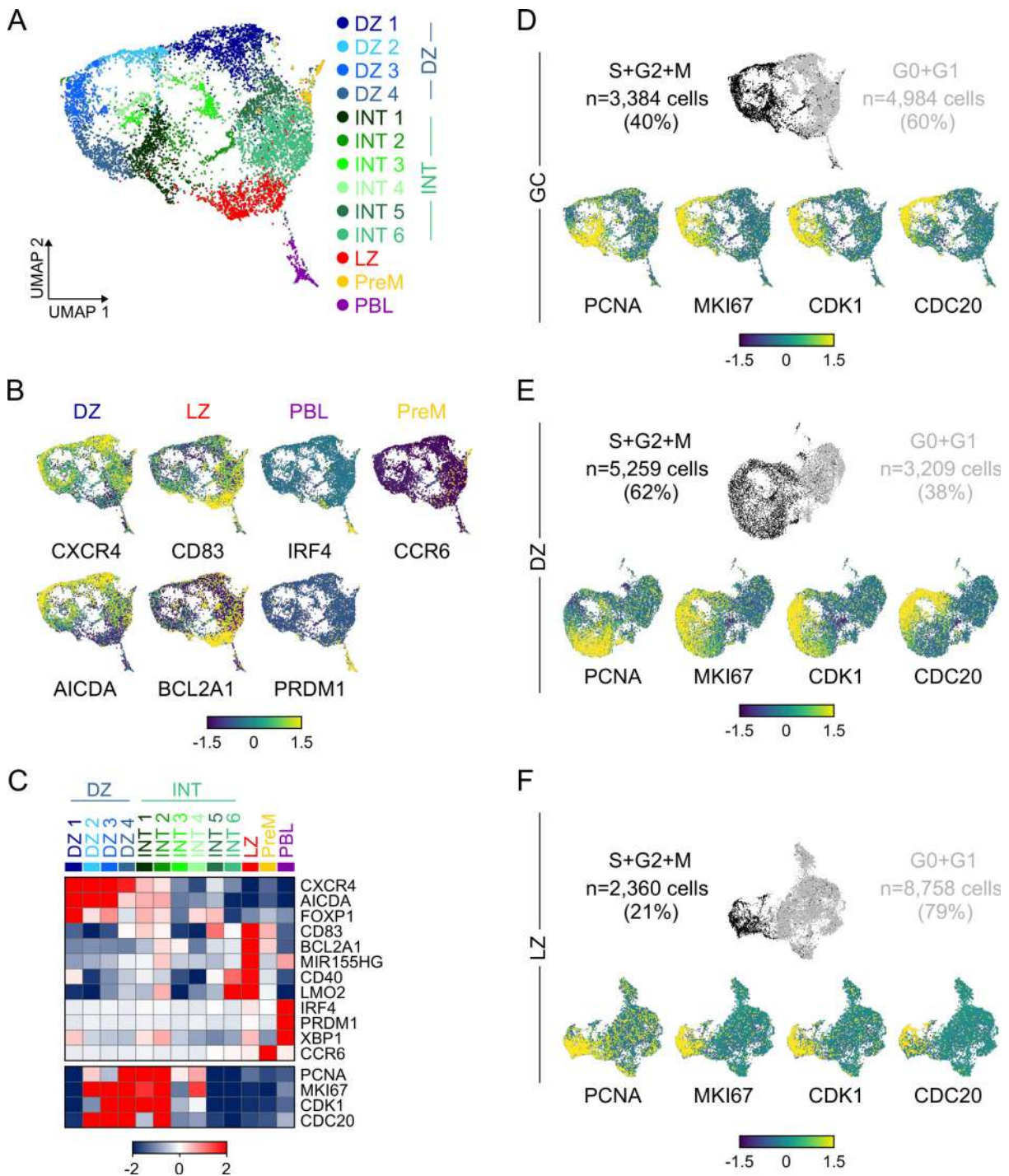
About 40% of the analyzed GC B cells belonged to clusters that displayed high average expression of genes associated with the S-G2-M stages of the cell cycle, including *PCNA*, *MKI67*, *CDKI*, and *CDC20*, and mainly included cells expressing markers of DZ or intermediate phenotypes (Fig. 1, C and D). Consistently, sc-gene expression profiling of 8,468 sorted DZ B cells showed that >60% of the cells displayed a transcriptional profile associated with S-G2-M (Fig. 1 E), while only ~20% of the 11,118 sorted LZ B cells expressed S-G2-M genes (Fig. 1 F). These results showed that although proliferation is a feature mostly associated with DZ GC B cells, both proliferating and resting cells are found in the DZ and LZ compartments.

### GC B cell developmental stages

The dominant cell cycle signatures associated with transition through the S-G2-M stages impair discriminating phenotypic subpopulations. To capture the functional and temporal dynamics among the GC B cell subpopulations, we thus focused our analysis on the GC B cell clusters (4,984 cells) with no or low expression of genes associated with the S, G2, or M phase of the cell cycle (Fig. 1 D). This analysis identified distinct subclusters previously hidden by the overwhelming “cell cycle signature” (Fig. 2 A). We assumed that given the proliferative nature of GC B cells, the analysis of this cell subset would be representative of all subpopulations just transitioning at a specific stage of the cell cycle. The specific contributions of cells that are in S-G2-M were investigated separately (see the analysis of DZ and LZ compartments below).

Genes differentially expressed in each cluster were identified by supervised analysis using model-based analysis of single-cell transcriptomics (MAST; Finak et al., 2015; Fig. 2 B and Table S1). Each cluster was labeled as DZ, intermediate, LZ, PBL, or PreM cells based on the expression of previously known marker genes, including CXCR4 and *AICDA* (DZ), CD83 and *BCL2A1* (LZ), *PRDMI* and *IRF4* (PBL), and *CCR6* (PreM; Fig. S2 A and Table S1). These assignments were confirmed by Gene Set Enrichment Analysis (GSEA), which showed that the DZ signatures were significantly enriched in purified DZ bulk populations, while the LZ signatures were significantly enriched in purified LZ bulk populations (Fig. S2 B and Table S2). The clusters labeled as intermediate displayed transitional profiles with some features alternatively related to DZ (INT-a) or LZ (INT-b, -c, -d, and -e; Fig. S2 C and Table S2).

Two very well-characterized master regulators of the GC reaction, *BCL6* and *FOXO1* (Basso and Dalla-Favera, 2012; Dominguez-Sola et al., 2015; Sander et al., 2015), displayed low levels of expression in the sc-RNAseq profiles, and they did not score among the top differentially expressed genes. Nonetheless, they showed the expected pattern of expression, with significant *BCL6* down-regulation in cells committing to post-GC differentiation, and induction of *FOXO1* in DZ GC B cells (Table S1). In addition, enrichment analysis confirmed that *BCL6* targets were significantly overrepresented among genes down-regulated in the DZ, in a subset of intermediate, and in the early LZ clusters,



**Figure 1. Identification of GC B cell subpopulations by sc-transcriptomic analysis. (A)** UMAP projection of sc-RNAseq profiles of 8,368 GC B cells ( $CD3^+$ ,  $IgD^+$ ,  $CD38^+$ ) isolated from two donors. sc-RNAseq data were merged upon batch effect correction performed by the Seurat package. Clusters in the UMAP plots were identified by PhenoGraph and color coded according to different cell stages: DZ, LZ, intermediate (INT), PreM, and PBL. **(B)** UMAP projections of the z-scored expression of selected marker genes associated with distinct stages. **(C)** Heat map displaying the relative expression, as fold change ( $\log_2$ ), of selected genes in the GC B cell clusters identified in A. **(D–F)** Top panels: UMAP projections of GC (D), DZ (E), and LZ (F) B cells color coded based on the expression (black dots) or not (gray dots) of genes associated with the S-G2-M stages of the cell cycle. Bottom panels: the UMAP projections were colored based on the z-scored expression of four representative genes associated with the S-G2-M stages of the cell cycle. See also Fig. S1.

all consistent with the pattern of expression of BCL6. Consistent with FOXO1 specific expression and required role in DZ development (Dominguez-Sola et al., 2015; Sander et al., 2015), FOXO1 targets were enriched in the signatures of the DZ compartments.

Two clusters clearly branched out from the continuous UMAP structure, composed by DZ, intermediate, and LZ GC B cells: one displayed elevated expression of plasma cell hallmarks such as PRDM1, XBP1, and IRF4 and was preliminarily



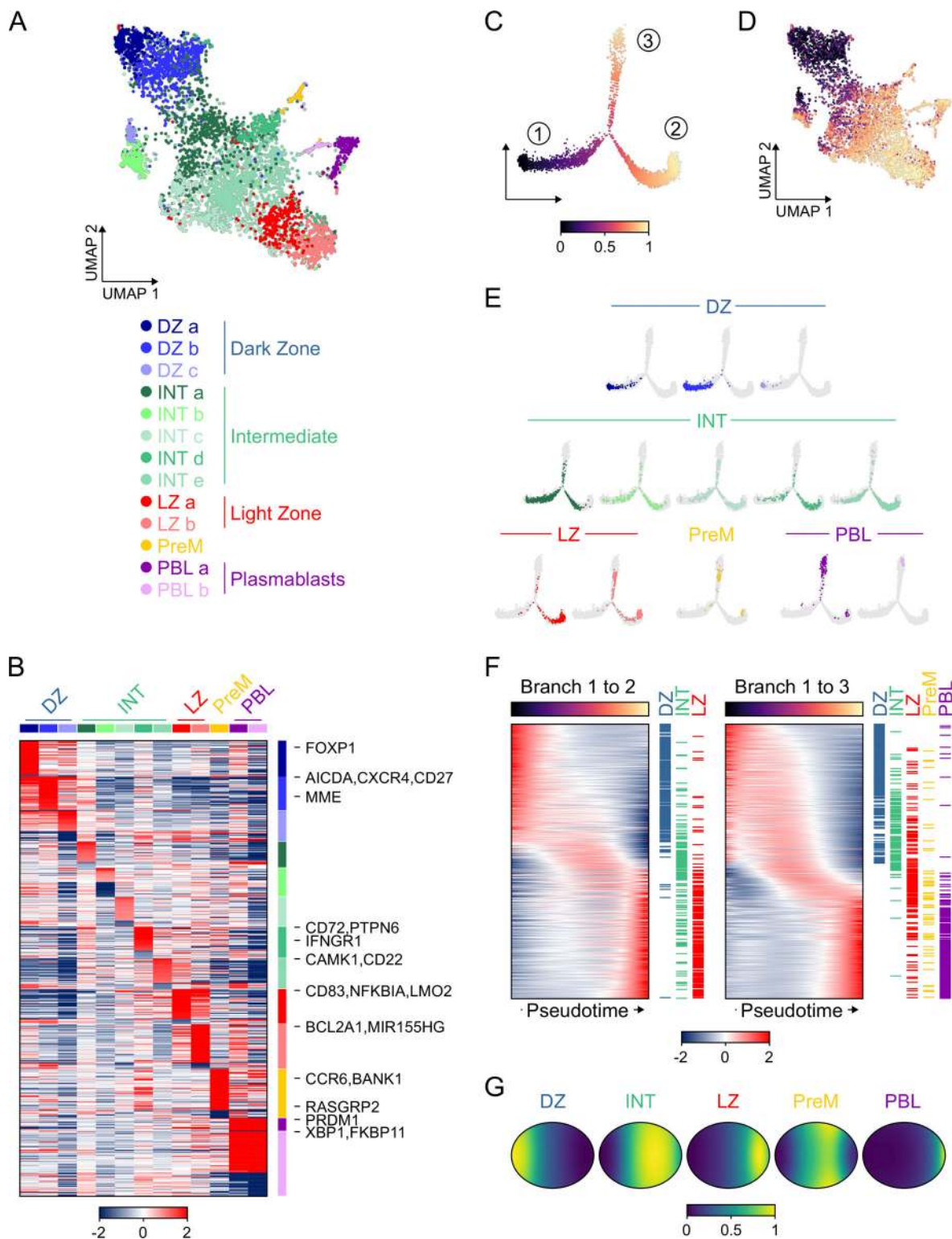


Figure 2. **GC B cell developmental stages.** **(A)** UMAP projection and cluster identification from sc-RNaseq profiles of 4,984 GC B cells in G0-G1 stages of the cell cycle. **(B)** The heat map shows the top 50 up- and down-regulated genes for each cluster. Genes that are significantly differentially expressed in more than one cluster are displayed in association with the cluster in which they show the best fold change. The heat map is colored by the  $\log_2$  fold change of the average expression. **(C)** Pseudo-time analysis by Monocle of the sc-RNaseq profiles displayed in A. The projection is colored by normalized pseudo-time. Numbers in the circle identify distinct branches. **(D)** UMAP projection colored by normalized pseudo-time analysis. **(E)** Mapping of the PhenoGraph clusters on the pseudo-time plot. **(F)** Heat map of pseudo-time gene expression changes. On the right, genes are labeled based on the phenotype in which they display the largest  $\log_2$  fold change. **(G)** Spatial reconstruction of the inferred locations of GC subpopulations in an idealized GC structure using novoSpaRc. The most likely inferred spatial position of a given cell type is denoted by its normalized score ranging from high (1, yellow) to low (0, blue) confidence values. See also Fig. S2, Table S1, and Table S2.

identified as PBL, while the other showed expression of the marker of PreM *CCR6* (Suan et al., 2017) and was thus labeled as PreM (Fig. 2 A, Fig. S2 A, and Table S1). These annotations were confirmed by GSEA showing that the PreM and PBL signatures displayed significant enrichment in the profiles of bulk memory B cells or plasma cells, respectively, compared with GC B cells (Fig. S2 D and Table S2; see paragraphs dedicated to PBL and PreM for additional details).

To investigate the temporal relationships among the identified clusters, we applied Monocle, a trajectory inference algorithm (Trapnell et al., 2014), to our dataset of 4,984 GC B cells. Monocle ordered the cells in a pseudo-temporal trajectory with three branches. The main trajectory (branch 1 to 2 in Fig. 2 C) ordered the DZ to LZ transition, while the differentiation toward PreM or PBL was inferred as a distinct branch (branch 3 in Fig. 2 C). The UMAP projection was color coded based on the pseudo-time inferred by Monocle and showed that the ordered stage progression could be identified also in the UMAP structure (Fig. 2 D). The pseudo-time ordering suggested that PreM B cells were generated earlier than the PBLs, although both of them appeared to be related to LZ B cells (Fig. 2 E). As expected, the genes informing the pseudo-time order displayed a pattern of expression mimicking the transitions through the GC stages (Fig. 2 F). Of note, the pseudo-time inference is linear and not directional; therefore, it cannot model the DZ reentry.

The sc-expression profile analysis requires disaggregation of the tissue into an sc suspension, leading to the loss of spatial information, which is critical for understanding of GC dynamics. Thus, we applied novoSpaRc (Nitzan et al., 2019), which allows for de novo spatial reconstruction of sc-gene expression data, to infer the spatial organization of GC B cells starting from their sc-RNAseq expression profiles. Consistent with well-characterized DZ and LZ compartmentalization, the results showed that the DZ and LZ cells were properly assigned at the opposite sides of a GC-resembling physical space (Fig. 2 G). The PBL population was closely related to an end stage of the LZ compartment, while the PreM population appeared closer to an intermediate GC stage, suggesting that, in agreement with the pseudo-temporal analysis (Fig. 2, C and E), this population may originate from cells in a stage preceding that associated with PBL commitment (Fig. 2 G).

Taken together, these results identify an ordered progression from DZ to LZ and suggest a temporally and spatially distinct process of differentiation into post-GC memory B cells and plasma cells.

### DZ B cell subpopulations

To investigate B cells in the DZ compartment, we analyzed 8,468 purified DZ B cells from two donors. This analysis revealed the presence of multiple cell clusters, most of which were associated with distinct stages of the cell cycle and with the majority of cells (62%) expressing S-G2-M genes (Fig. 3 A). We then analyzed separately the cells expressing or not expressing genes associated with the S-G2-M phases of the cell cycle.

In the proliferating compartment, we identified a population that was transitioning from G1 to S and displayed expression of *MYC* (Fig. 3, B and C). Based on the known *MYC* expression pattern and role in the GC (Calado et al., 2012; Dominguez-Sola

et al., 2012), this population may represent DZ B cells just entering the GC and/or reentering the DZ after selection in the LZ. Proliferating DZ cells also showed induction of numerous genes involved in DNA damage sensing and repair (Fig. 3 C and Table S3). Pathway enrichment analysis confirmed that cells entering the S phase displayed the most significant involvement of DNA repair pathways, including mismatch repair as expected for its role in SHM (Fig. 3 D and Table S3). Although expressed in most DZ clusters, several genes, including *AICDA* and *EZH2*, showed transcriptional modulation during the cell cycle. Specifically, *AICDA* was strongly induced in cells in G2-M, and *EZH2* displayed a peak of expression in cells in the S phase following the induction of *E2F1*, which is known to transcriptionally induce *EZH2* (Béguelin et al., 2017; Fig. 3 C).

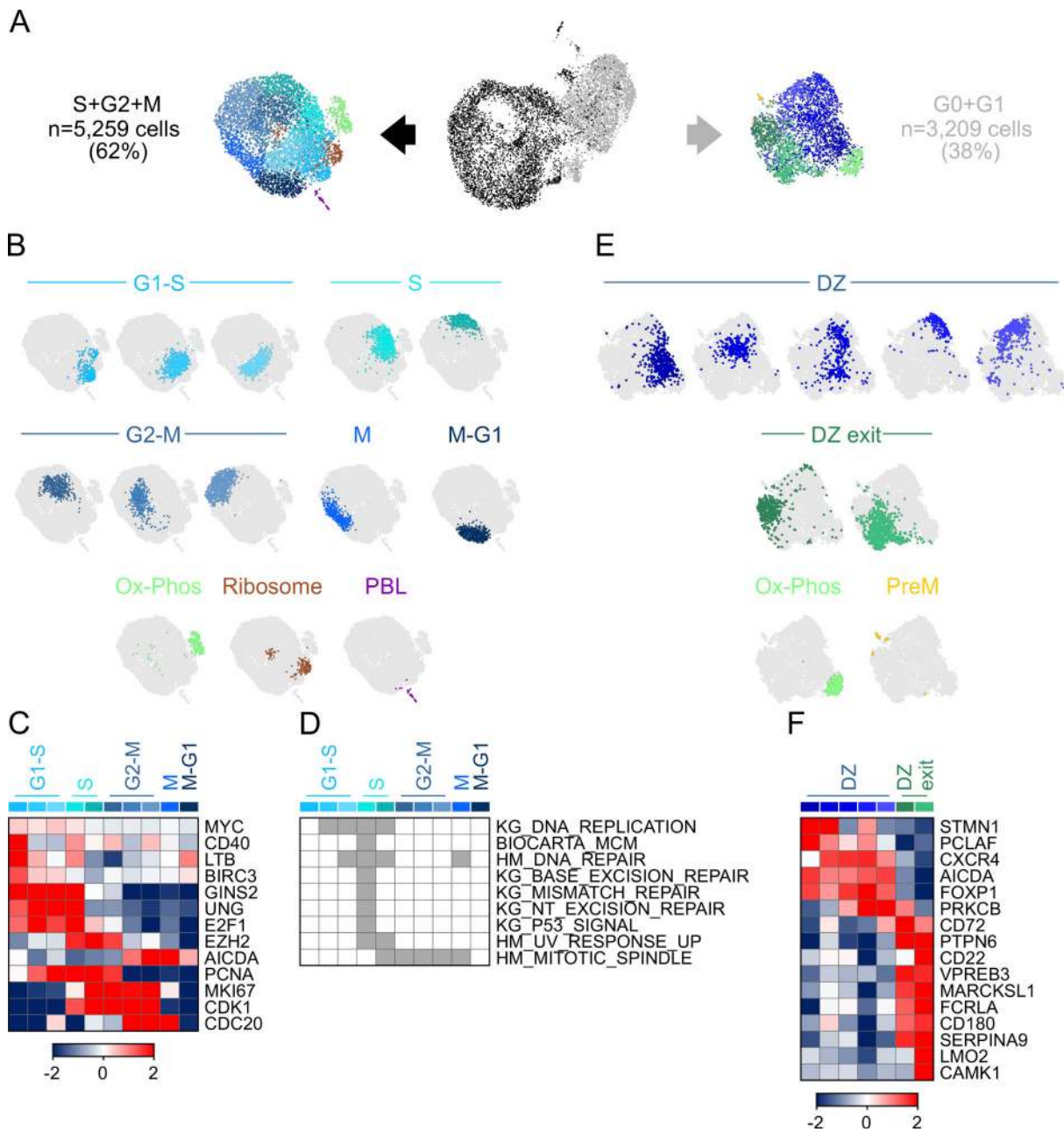
Among the cells in G0-G1, we identified a subset of DZ cells that appeared to down-regulate DZ markers, including *CXCR4*, *AICDA*, and *FOXPI*, and to up-regulate genes involved in calcium signaling (*PRKCB* and *CAMKI*) and BCR modulation (*CD72*, *PTPN6*, and *CD22*), a signature that is then further induced in the intermediate GC stages (DZ-exit in Fig. 3, E and F). We suggest that these cells are preparing to exit the DZ compartment and move forward to the LZ through multiple intermediate stages (Table S3).

In addition, we identified clusters of cells that were not associated with a specific cell cycle stage or phenotype but displayed higher expression of several genes involved in the oxidative-phosphorylation response and/or ribosome and RNA processing (Fig. 3, B and E; and Table S3). Two additional clusters carried the signatures of PreM and PBL, suggesting that small subpopulations of these precursors were sorted together with DZ cells, most likely based on their expression of *CXCR4* (Fig. 3, B and E).

Taken together, these data confirmed the overall proliferative nature of DZ B cells but were also able to capture the presence of quiescent cells, a subset of which displayed transcriptional changes consistent with DZ exit.

### LZ B cell subpopulations and fates

To investigate the heterogeneity of the LZ compartment, we interrogated 11,118 sorted LZ GC B cells from two donors. As mentioned above, most LZ GC B cells (~80%) did not display expression of genes associated with the S-G2-M phases of the cell cycle (Fig. 4 A). The analysis of these nonproliferating LZ GC B cells identified 12 distinct clusters that were annotated based on their expression profiles (Fig. 4 B and Table S4). Consistent with an initial stage of activation in the LZ, we annotated as “LZ entry” a cluster of cells that retained modest expression of DZ markers (*CXCR4*, *CD27*, and *FOXPI*; Fig. 4, B and C). This cluster showed up-regulation of genes in the BCR (including *BTK*, *BLK*, and *BLNK*) but not in the CD40 signaling pathway (Fig. 4, B and C; and Table S4), suggesting that these cells may have just entered the LZ and may be ready to interact with the antigen, but not with T cells. An additional group of clusters displayed up-regulation of genes associated with antigen presentation (including genes in the MHC class II and CD74 receptors) and BCR signaling, but they lacked expression of DZ markers. These clusters may represent subsequent LZ stages in which B cells



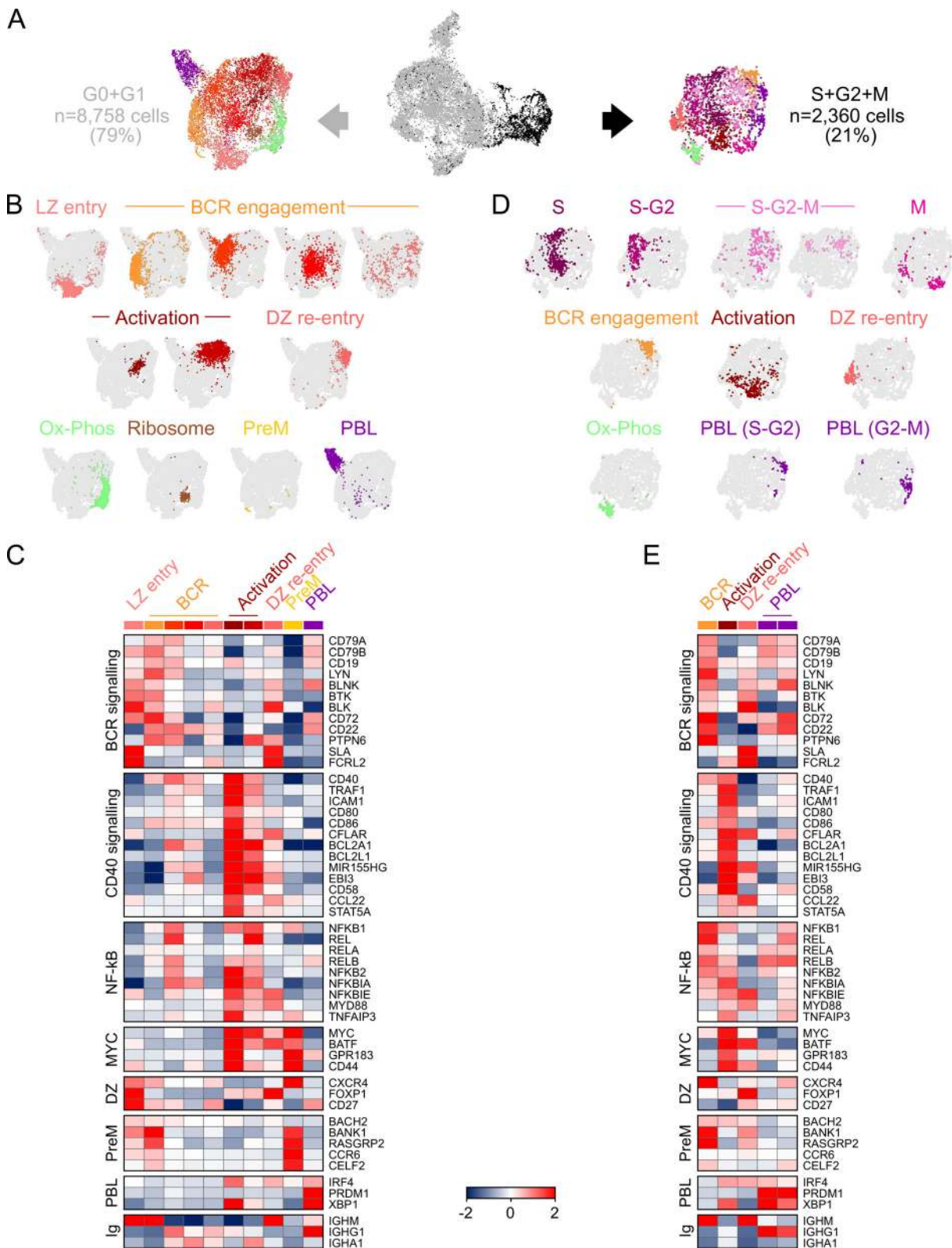
**Figure 3. DZ B cell subpopulations. (A)** UMAP projection and cluster identification from sc-RNAseq profiles of 8,468 DZ B cells ( $CD3^{-}$ ,  $IgD^{-}$ ,  $CD38^{+}$ ,  $CXCR4^{hi}$ ,  $CD83^{lo}$ ). Black and gray dots mark B cells that respectively express (left) or do not express (right) genes associated with the S-G2-M stages of the cell cycle. UMAP projections and cluster identification were performed independently for cells in the two cell cycle groups. **(B)** UMAP projection and cluster identification by PhenoGraph of DZ GC B cells that express genes associated with the S-G2-M stages of the cell cycle. **(C)** Heat map displaying the relative expression fold change ( $\log_2$ ) of selected genes in the DZ B cell clusters identified in B. **(D)** Pathway enrichment analysis for the gene signatures associated with the clusters identified in B. Pathways from KEGG database (KG), Hallmark database (HM), and Biocarta database that are significantly enriched (hypergeometric test with Benjamini-Hochberg correction,  $q < 0.05$ ) are shown in gray. **(E)** UMAP projection and cluster identification by PhenoGraph of DZ GC B cells that are at the G0-G1 stages of the cell cycle. **(F)** Heat map displaying the relative expression fold change ( $\log_2$ ) of selected genes in the DZ B cell clusters identified in E. See also Table S3. Ox-Phos, oxidative phosphorylation.

have captured the antigen presented by follicular dendritic cells and processed it for presentation by MHC class II complex and were thus labeled “BCR engagement” (Fig. 4, B and C; and Table S4). In addition to genes associated with BCR engagement, cells in these clusters displayed variable induction of genes in the CD40 signaling pathway (including *CD40*, *TRAF1*, and *ICAM1*) and

NF- $\kappa$ B activation (including *NFKB1*, *NFKB2*, *REL*, and *RELB*), suggesting that these cells were undergoing both antigen stimulation and T cell-mediated activation.

Of note, some of these clusters retained high expression of *IGHM*, while others showed expression of switched immunoglobulin receptors (*IGHG* and *IGHA*; Fig. 4 C).





**Figure 4. LZ B cell subpopulations. (A)** UMAP projection and cluster identification from sc-RNAseq profiles of 11,118 LZ B cells ( $CD3^+$ ,  $IgD^-$ ,  $CD38^+$ ,  $CXCR4^{lo}$ ,  $CD83^{hi}$ ). Black and gray dots mark B cells that respectively express (right) or do not express (left) genes associated with the S-G2-M stages of the cell cycle. UMAP projections and cluster identification were performed independently for cells in the two cell cycle groups. **(B)** UMAP projection and cluster identification by PhenoGraph of LZ GC B cells at the G0-G1 stages of the cell cycle. **(C)** Heat map displaying the relative expression fold change ( $\log_2$ ) of selected genes in LZ B cell clusters identified in B. **(D)** UMAP projection and cluster identification by PhenoGraph of LZ GC B cells that express genes associated with the S-G2-M stages of the cell cycle. **(E)** Heat map displaying the relative expression fold change ( $\log_2$ ) of selected genes in LZ B cell clusters identified in D. See also Table S4. Ox-Phos, oxidative phosphorylation.

The transcriptional effects of strong T cell-mediated activation, including a dominant signature of CD40 signaling and NF- $\kappa$ B activation, were detected in a small subpopulation (“Activation” in Fig. 4, B and C), which displayed high induction of B cell activation markers and *MYC* expression. These highly activated cells acquired transcriptional features, including enrichment for genes in the MTORC1 signaling pathway, suggestive of the metabolic burst shown in cells ready to reenter the DZ, and a subset of them consistently displayed induction of markers (*FOXPI* and *CD27*) associated with a DZ phenotype (Fig. 4 C and Table S4; Calado et al., 2012; Dominguez-Sola et al., 2012; Ersching et al., 2017; Finkin et al., 2019).

Transcriptional changes associated with BCR engagement, T cell-mediated activation, and LZ to DZ reentry were detected also among cells displaying expression of S-G2-M genes (Fig. 4, D and E; and Table S4). These results suggest that although representing a small fraction, proliferating cells are detectable in different LZ states. Additional clusters (PreM and PBL) showed features of commitment to post-GC differentiation (Fig. 4, B-E; see below for additional details). Overall, these results confirm the heterogeneity of LZ GC B cells and provide a transcriptional identity to the populations underlying it.

#### Identification of memory B cell precursors

The analysis of GC B cells identified a small (~2%) but clearly distinct cluster (named PreM) that shared a unique gene expression profile significantly related with that of previously reported CCR6<sup>+</sup> PreM B cells (Suan et al., 2017; Fig. 5 A). As shown by GSEA, the PreM sc-signature was significantly enriched in the transcriptome of bulk purified memory B cells compared with total or LZ GC B cells (Fig. 5 A and Fig. S2 D). In addition, although naive and memory B cells share a large portion of their transcriptional profiles, the genes up-regulated in the PreM cluster were enriched in the bulk expression profile of mature memory cells compared with naive B cells (Fig. 5 A). Finally, PreM cells did not display expression of S-G2-M genes, consistent with previous studies suggesting that GC B cells differentiating into memory B cells exit the cell cycle (Laidlaw et al., 2017; Suan et al., 2017; Wang et al., 2017). Overall, these observations support the identification of a distinct subpopulation composed of memory B cell precursors in the GC.

Although PreM cells represented a small fraction of the overall GC B cells, clusters expressing a subset of the PreM signature were detectable in both LZ (Fig. 4 C) and intermediate (INT-d (Fig. 5 B) GC stages, possibly representing cells at early stages of PreM commitment. To gain insights in the PreM developmental process, we then analyzed all clusters displaying enrichment of the PreM signature genes independent of GC stage (Fig. 5 B and Fig. S3, A and B). The analysis of this merged dataset, including 1,542 cells, suggests that commitment to the PreM starts in an LZ subpopulation expressing *BACH2*, consistent with previous observations (Shinnakasu et al., 2016; Fig. 5, C and D). These early precursors appear to transit from the LZ into an intermediate GC stage and then to acquire a fully established GC PreM identity characterized by high CCR6 expression (Fig. 5, C and D).

In addition to previously reported markers, including CCR6 and *CELF2* (Suan et al., 2017), the PreM cells displayed high

expression of *BANK1*, a scaffold protein involved in BCR-induced calcium mobilization and in the negative modulation of CD40-mediated AKT activation (Aiba et al., 2006; Yokoyama et al., 2002), and *RASGRP2*, a nucleotide exchange factor that can activate small GTPases including RAS and RAP1 (Ebinu et al., 1998). Based on our analysis, *BANK1* and *RASGRP2* appear to be the earliest specific markers induced in the GC PreM (Fig. 4 C and Fig. 5 B). We also detected induction of interferon-responsive genes (*IFITM1* and *IFITM2*), the interferon receptor *IFNGR1*, and the transcription factor *STAT1*, which is activated by IFN $\gamma$  signaling (Shuai et al., 1992), suggesting that interferon signaling may contribute to PreM differentiation (Fig. 5 D). Consistent with a program of GC-exit, several genes showed a significant down-regulation in PreM cells, including GC markers (*MEF2B* and *RSGI3*) and *SIPR2*, which is involved in GC confinement (Fig. 5 D).

Consistent with the RNA expression data, we confirmed by CITE-seq that surface protein expression of CCR6 was associated with the PreM cell cluster (Fig. S1 E). In addition, the activation markers CD44 and CD69 displayed the highest expression at both transcript and protein levels in the PreM cells, while the PBL marker CD9 was not associated with this population, suggesting that the combination of these four markers may be implemented in sorting strategies of PreM GC B cells (Fig. S1 E).

To validate these findings and spatially localize the PreM cells within the GC, we performed coimmunofluorescence analysis for *BANK1* and *CELF2* in tonsil tissue sections. As expected, *BANK1* was detected exclusively in B cells (naive and a small subset of GC cells) and *CELF2* mostly in T cells (Fig. S3 C). However, consistent with the sc-RNAseq data, *BANK1* and *CELF2* were coexpressed in a small fraction of cells ( $\leq 1\%$  of total GC B cells) that lacked the GC markers *MEF2B* and *BCL6*. In agreement with previous reports (Laidlaw et al., 2017; Wang et al., 2017), these cells localized mostly in the LZ compartment or close to the GC border (Fig. 5, E and F; and Fig. S3 D). In addition, we detected a subpopulation expressing *BANK1*, *MEF2B*, and *BCL6* in the absence of *CELF2*, supporting the observation that *BANK1* is an early marker of PreM commitment (Fig. 5, E and F). In conclusion, our analysis identified the transcriptional signature and the consistent expression of several surface protein markers of memory B cell precursors and traced their localization in the GC by immunofluorescence staining of *BANK1* and *CELF2*.

#### Discrete steps of PBL differentiation starting from a subset of LZ B cells

Clusters displaying expression of plasma cell hallmark genes, including *IRF4*, *PRDMI*, and *XBPI*, and significant enrichment by GSEA analysis in the bulk gene expression profiles of plasma cells compared with GC B cells were labeled as PBLs (Fig. S2 D). Of note, mature plasma cells were not included in the analyzed GC B cells because the sorting strategy excluded cells with high CD38 expression (Fig. S1 A). PBLs represented <3% of the overall GC B cells (Fig. 1 A) and ~8% of the purified LZ B cells, 22% of which were actively proliferating (Fig. 4).

To investigate the plasma cell differentiation process in the GC reaction, we merged the sc-RNAseq profiles of the 1,231 cells



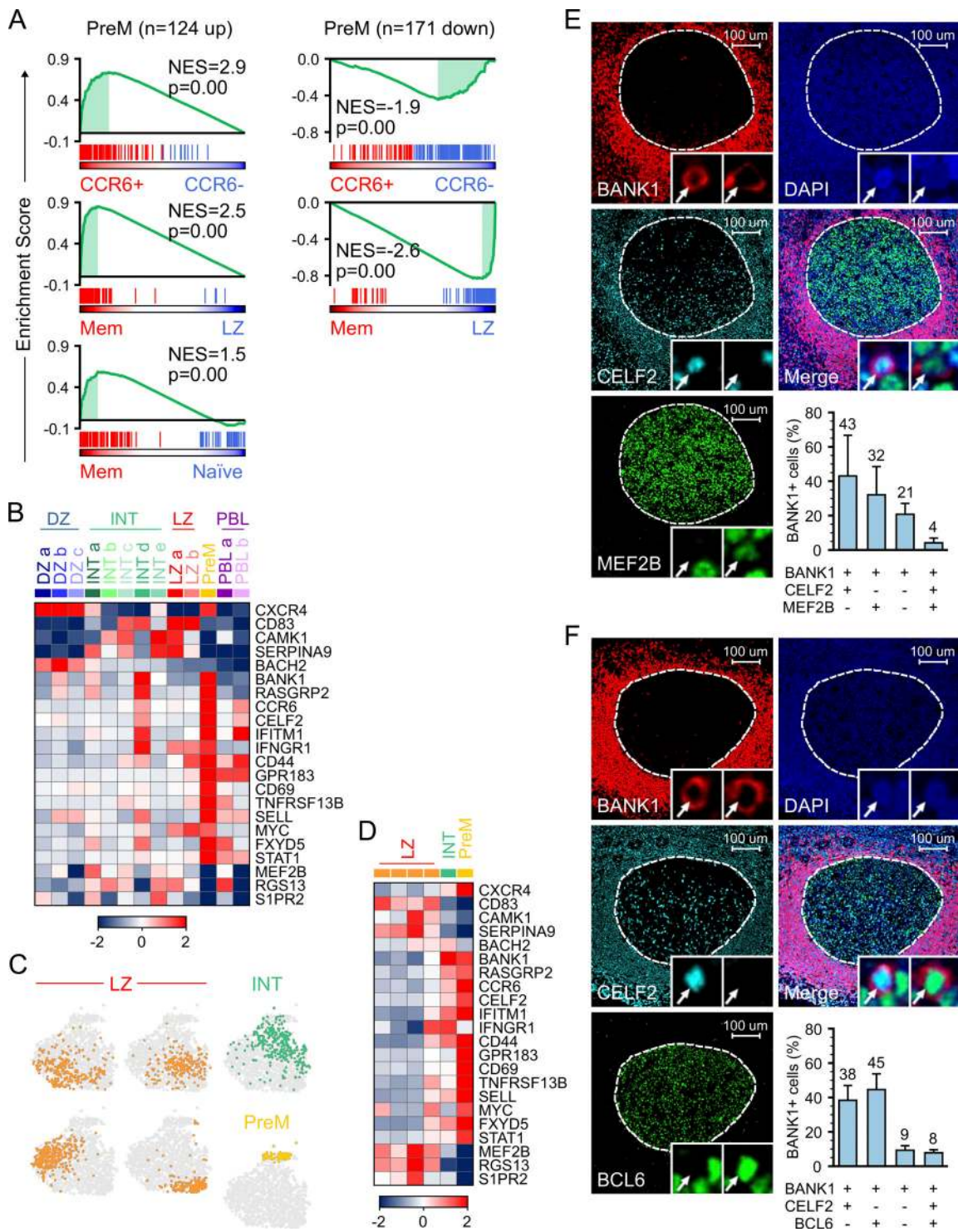
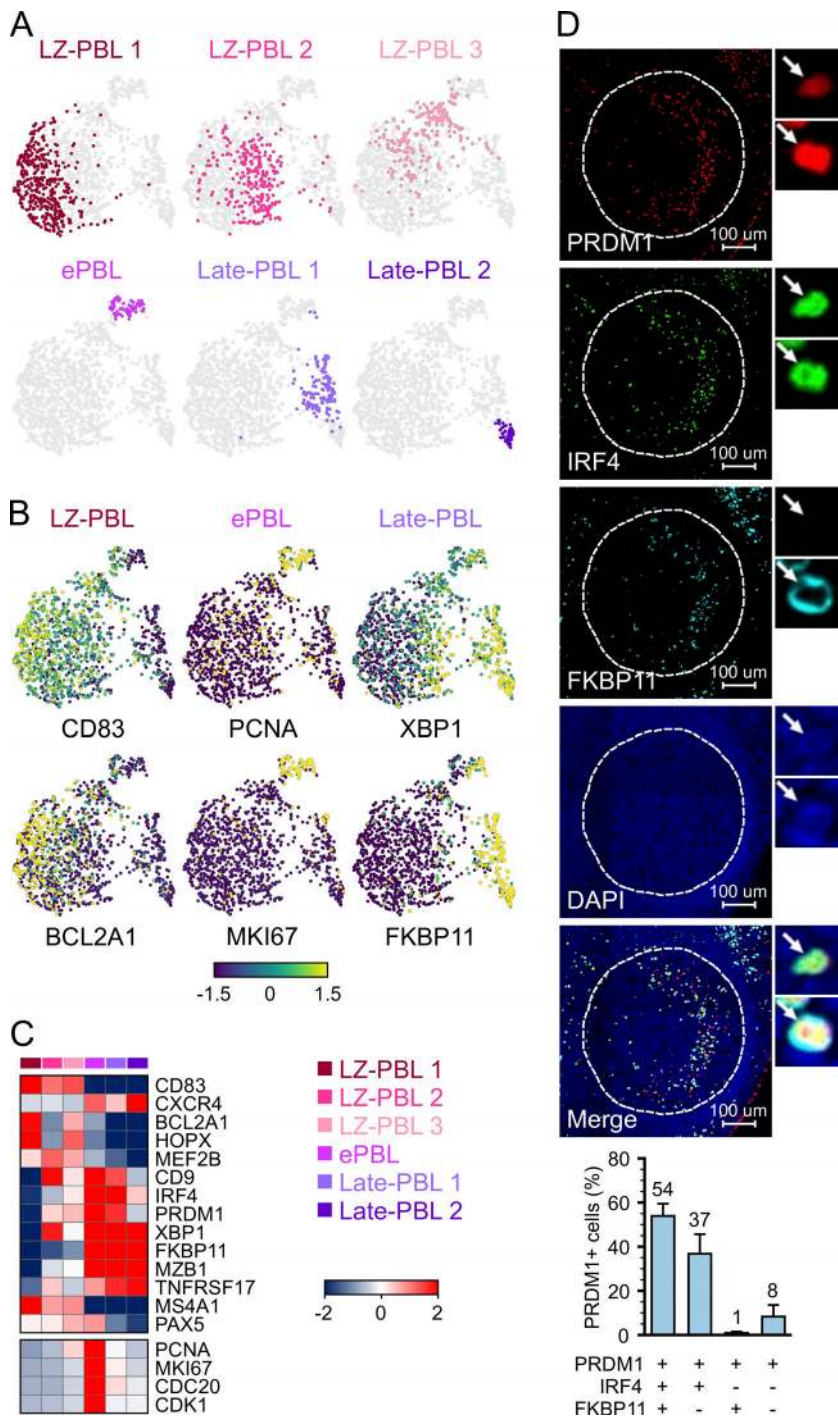


Figure 5. **Memory B cell precursors.** (A) GSEA of PreM sc-cluster signatures in the expression profiles of purified bulk populations of CCR6<sup>+</sup> versus CCR6<sup>-</sup> LZ GC B cells, memory (Mem) versus LZ GC B cells, and memory versus naive B cells. (B) Heat map displaying, in the GC B cell clusters identified in Fig. 2, the relative expression fold change (log<sub>2</sub>) of selected PreM signature genes. (C) UMAP projection and cluster identification of 1,542 GC B cells from clusters displaying enrichment for PreM signature genes. (D) Heat map displaying the relative expression fold change (log<sub>2</sub>) of selected PreM signature genes in the clusters identified in C. (E and F) Immunofluorescence analysis of BANK1, CELF2, and MEF2B (E) or BCL6 (F) in tonsil sections. Nuclei were stained with DAPI. In the high-magnification insets, arrows point to representative cells displaying marker coexpression. Bar plots display the percentage of BANK1<sup>+</sup> cells co-expressing the indicated markers. Each bar plot shows the average cell counts and standard deviations in three independent donors. In total, 373 GCs were analyzed for each staining combination. The scale bar corresponds to 100 μm. See also Fig. S3. NES, normalized enrichment score.



**Figure 6. Plasma cell precursors.** (A) UMAP projection and PhenoGraph clusters of 1,231 PBLs identified from the analysis of GC B cells. (B) UMAP projections colored based on the z-scored expression of genes associated with distinct stages of PBL development. (C) Heat map displaying the relative expression fold change ( $\log_2$ ) of selected genes differentially expressed across the clusters detected in A. (D) Immunofluorescence analysis of PRDM1, IRF4, and FKBP11 in tonsil sections. Nuclei were stained with DAPI. In the high-magnification insets, arrows point to representative cells displaying marker coexpression. The bar plot displays the percentage of PRDM1<sup>+</sup> cells coexpressing the indicated markers, and it shows the average cell counts and standard deviations in five donors. In total, 362 GCs were analyzed. The scale bar corresponds to 100  $\mu$ m. ePBL, early PBL.

identified as PBLs in either the total or the sorted GC subpopulations, as described above. This analysis identified six distinct clusters (Fig. 6 A), three of which we defined as LZ-PBLs because they mostly comprise sorted LZ cells characterized by decreasing expression of LZ markers, including CD83, and increasing expression of PBL markers, including CD9, PRDM1, and XBP1 (Fig. 6, B and C). We defined early PBLs, a cluster of actively proliferating cells that displayed further down-regulation of the LZ markers and enhanced expression of PBL markers (Fig. 6, B and C). Two additional clusters (late-PBL) displayed no expression of LZ markers, down-regulation of MS4A1 and PAX5, and

robust expression of plasma cell markers, including XBP1, MZB1, and TNFRSF17 (Fig. 6, B and C). Together with the absence of proliferation signatures, this pattern is highly consistent with a late stage of PBL differentiation, thus the name of late-PBL. Interrogation of these populations revealed that FKBP11, a peptidyl-prolyl cis/trans isomerase previously implicated in plasma cell differentiation (Ruer-Laventie et al., 2015), was highly induced in late stages of PBL differentiation (Fig. 6, B and C). Consistently, we validated FKBP11 protein expression in ~50% of GC B cells coexpressing IRF4 and PRDM1, suggesting that FKBP11 may represent a marker of late PBL differentiation (Fig. 6 D).



These data suggest a multistep process of PBL differentiation that starts in a subset of LZ cells and includes proliferation.

### GC B cell signatures define lymphoma subgroups

To investigate whether the sc-RNAseq-defined gene signatures of the GC subpopulations can further inform on the COO of lymphomas, we developed a classifier that allowed tracking of gene signatures in panels of gene expression profiles (Fig. S4 A). Toward this goal, we selected the top 50 up- and down-regulated genes from each sc-cluster signature in the GC dataset (Fig. 2) and first tested them in bulk RNAseq expression profiles obtained from purified normal B cell populations, including CD77<sup>+</sup> GC B cells, DZ, LZ, and memory B cells. The results showed that each population was properly classified using the sc-RNAseq signatures (Fig. S4 B).

We then applied the sc-based classifier on two published panels of DLBCL primary cases, including 481 (National Cancer Institute [NCI]-DLBCL; Schmitz et al., 2018) and 230 (British Columbia Cancer Agency [BCCA]-DLBCL; Arthur et al., 2018) bulk RNAseq expression profiles, respectively. These DLBCL panels were representative of the current COO groups (259 GCB, 333 ABC, and 119 unclassified). The two datasets were classified independently, and the results showed that in both, ~80% of the DLBCL cases could be assigned to at least one sc-RNAseq cluster with high confidence (P value <0.05; Fig. 7 A and Table S5). For cases that could be assigned to more than one class, we measured the consistency of multiple assignments displaying similar scores (variance 10%), and we found that 90% of cases were associated with one (75% of cases) or with contiguous classes and/or phenotypes (15% of cases). Each sample was annotated based on its best score and P value (Table S5; see Materials and methods for details). Of note, the sc-RNAseq class assignment showed a remarkably similar distribution in both datasets (Fig. 7 B).

Consistent with previous observations associating the COO of GCB DLBCL to the LZ (Ennishi et al., 2019; Victora et al., 2012), the majority of the sc-classified GCB DLBCLs was shown to be related with normal GC B cells that are in the LZ-like intermediate stages or early LZ (NCI: 77/108, 71%; BCCA: 52/85, 61%; overall: 129/193, 67%; Fig. 7 A and Fig. S5 A). In particular, clusters INT-b, INT-e, and LZ-a displayed significant enrichment for GCB DLBCL (Fig. 7 C).

The ABC-DLBCL cases were scattered across the sc-RNAseq classes, with a predominance in the late GC stages (Fig. 7 A). Significant enrichment was observed in the intermediate INT-d, LZ-b, and PreM clusters, which overall accounted for 60% of the classified ABC cases (NCI: 107/193, 55%; BCCA: 54/74, 73%; overall: 161/267, 60%; Fig. 7 C and Fig. S5 A). Of note, only 12% (NCI: 26/193, 13%; BCCA: 7/74, 9%; overall: 33/267, 12%) of the classified ABC cases were assigned to the PBL clusters (Fig. 7 A and Fig. S5 A). These data suggest that, contrary to current models, most ABC-DLBCLs may originate from cells that are not yet committed to PBL differentiation and that correspond in as many as one quarter of cases to PreM B cells (NCI: 38/193, 20%; BCCA: 28/74, 38%; overall: 66/267, 25%).

A large fraction of the COO-unclassified DLBCLs (NCI: 84/100, 84%; BCCA: 16/19, 84%; overall: 100/119, 84%) were assigned to an sc class, and 27% of them (27/100) were shown to

be related to PBLs, with a significant enrichment in the PBL-a cluster (Fig. 7 C and Fig. S5 A).

Finally, we identified a small group of DLBCLs (NCI: 46/385, 12%; BCCA: 22/175, 12%; overall: 68/560, 12%) that displayed DZ gene signatures and distributed across the ABC, GCB, and unclassified tumors (Fig. 7, A-C). Interestingly, these tumors were significantly enriched in double-hit (MYC/BCL2) DLBCL, as defined by a previously reported double-hit gene expression signature (DHITsig; Ennishi et al., 2019), supporting a distinct ontogenesis for this unfavorable prognostic category (Ennishi et al., 2019; Sha et al., 2019).

Recent progress in the genetic characterization of DLBCL led to the identification of genetic subgroups associated with distinct mechanisms of pathogenesis and displaying differences in outcome following chemoimmunotherapy (Chapuy et al., 2018; Schmitz et al., 2018). Comparison of the sc-COO classification with the genetic-based groups (see Materials and methods) showed that DLBCL classified as INT-b and INT-e were significantly associated with the EZB (based on EZH2 mutations and BCL2 translocations) and Cluster 3 (C3) genetic groups, while LZ-b DLBCL were enriched in the MCD (based on MYD88<sup>L265P</sup> and CD79B mutations) and Cluster 5 (C5) genetic groups (Fig. S5 B). As expected, these genetic classes were enriched in GCB and ABC cases, respectively. In addition, the N1 (enriched for mutations in NOTCH1) and the C5 genetic classes were significantly associated with the PreM compartment (Fig. S5 B).

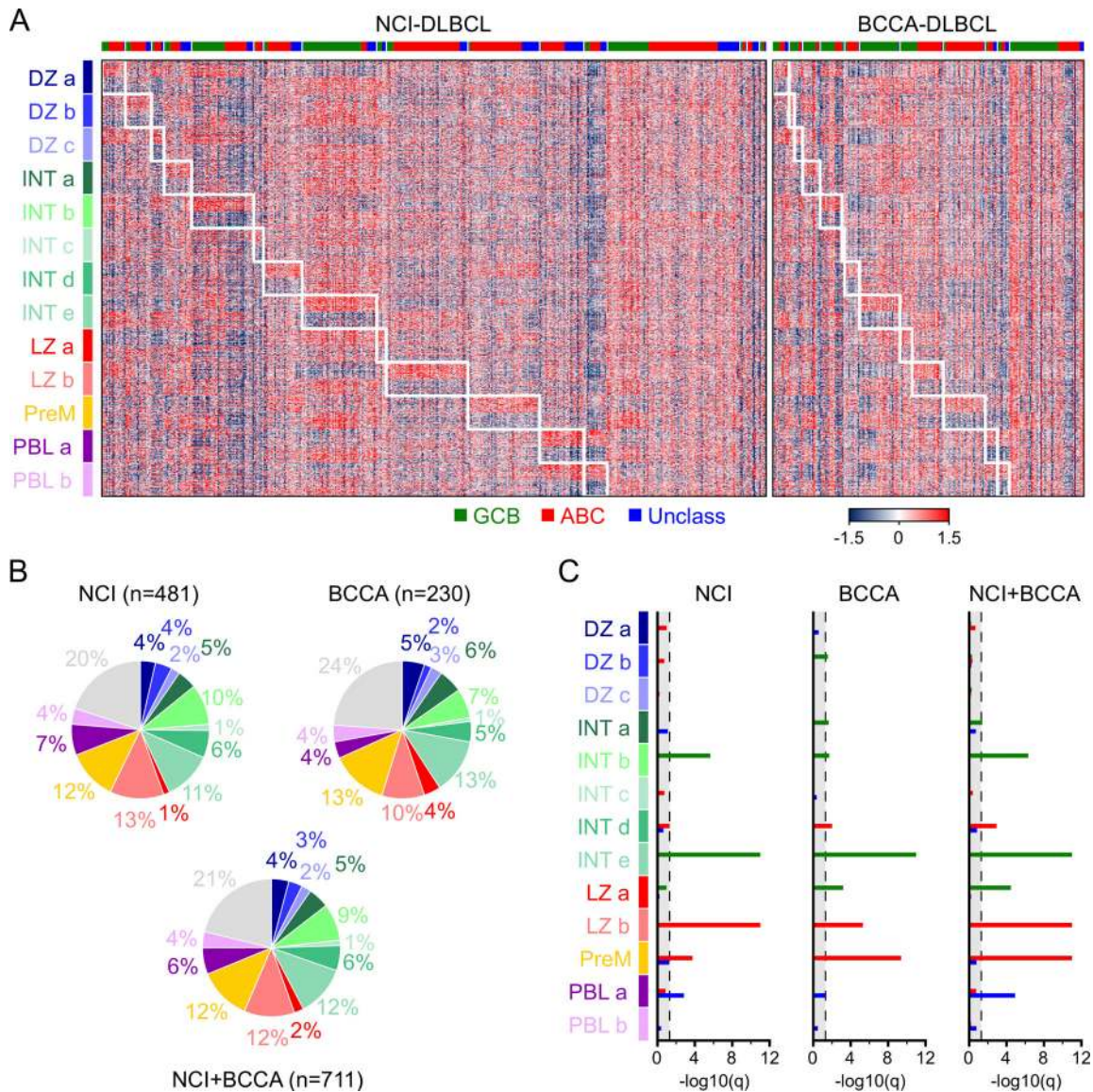
Taken together, these results provide a detailed classification of DLBCLs that pairs DLBCL subgroups to specific GC subpopulations.

### Prognostic value of the sc-COO classification

To test whether the newly identified subgroups carry an association with outcome, we used progression-free survival (PFS) data available for 356 (190 from the NCI and 166 from the BCCA datasets) of the 560 sc-COO-classified cases and for which, as expected, the COO classification retained its prognostic value (Fig. S5 C). First, we analyzed the NCI dataset classified according to the 13 sc-COO classes (Fig. S5 D). Then, in order to simplify the interpretation of the complex picture obtained from the analysis of 13 classes, we grouped contiguous developmental classes that displayed similar PFS. Following this approach, we could identify five contiguous categories (Groups I to V) that paired specific sc-COO classes to different survival outcomes (Fig. 8 A). Independent validation in the BCCA dataset provided remarkably similar survival trends, both in the analysis of the 13 sc-COO classes and of the merged groups (Fig. 8 B and Fig. S5 E).

Analysis of the combined datasets corroborated the strong association between the DZ signatures in Group II (DZ-b and DZ-c; NCI: 14/190, 7%; BCCA: 10/166, 6%; overall: 24/356 cases, 7%) and an extremely poor outcome (median PFS = 2 yr; Fig. 8 C and Fig. S5 F). Conversely, tumors displaying similarities with late LZ-like intermediate or early LZ clusters (INT-d, INT-e, and LZ-a; Group IV; NCI: 43/190, 23%; BCCA: 47/166, 28%; overall: 90/356 cases, 25%) were characterized by the best PFS (Fig. 8 C and Fig. S5 F). The remaining groups displayed an intermediate prognosis and were associated either with the signatures of DZ-a (Group I; NCI: 14/190, 7%; BCCA: 10/166, 6%; overall: 24/356 cases, 7%),





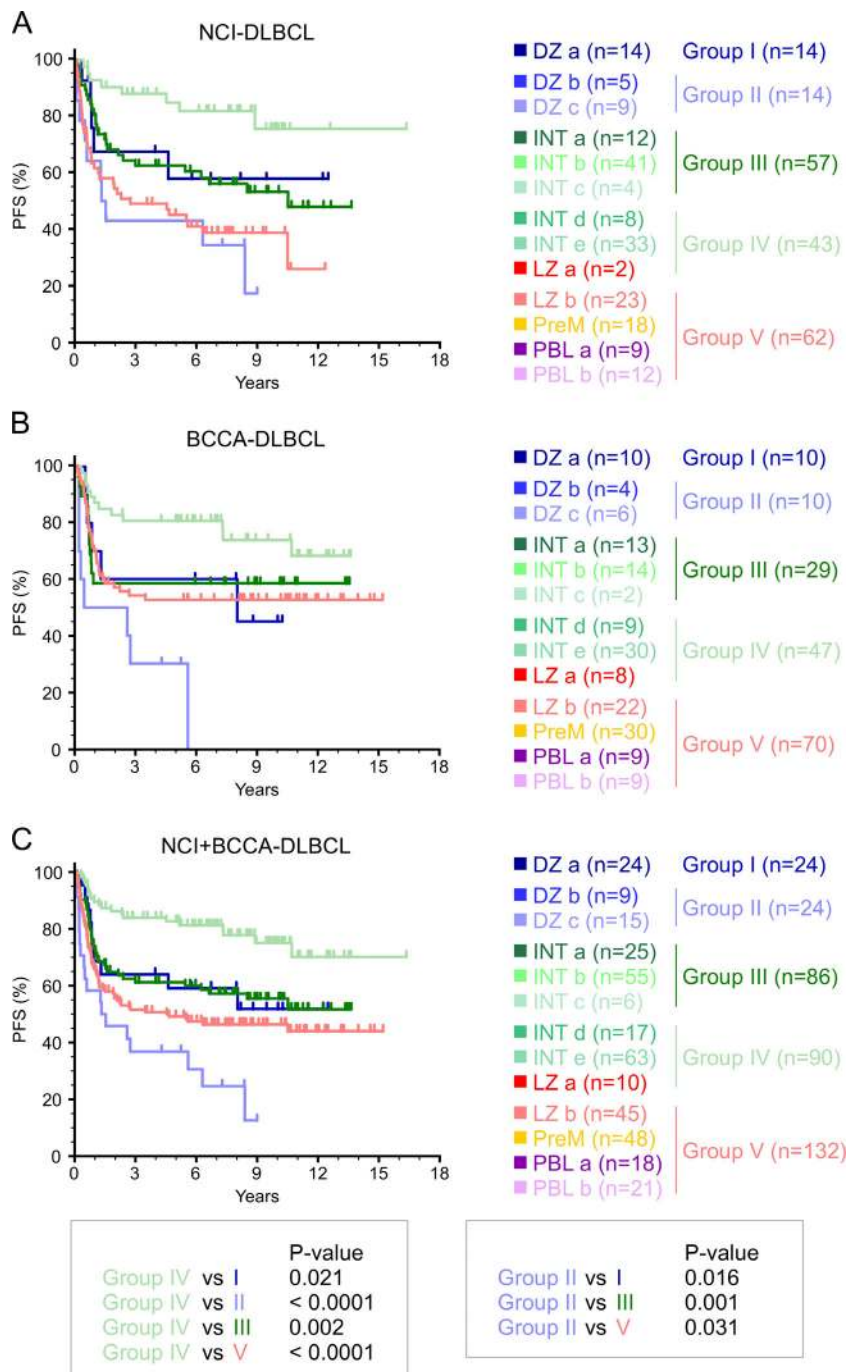
**Figure 7. GC B cell signatures define lymphoma subgroups. (A)** Heat map summarizing the sc-based classification of two panels of DLBCL primary cases for which bulk RNAseq profiles were previously reported. The datasets included 481 (NCI-DLBCL; Schmitz et al., 2018) and 230 (BCCA-DLBCL; Arthur et al., 2018) expression profiles and were representative of the current COO groups (259 GCB, 333 ABC, and 119 unclassified [Unclass]). Each column represents a DLBCL specimen, and each row displays the relative z-scored expression of an sc-classifier gene. The sc-RNAseq signatures include the top 50 up-regulated and the top 50 down-regulated genes in each GC cluster. The specific sc-clusters are labeled on the left of the heat map, while DLBCL cases are labeled based on the COO classification on the top. DLBCL samples are grouped based on their best classifier score. **(B)** Distribution of the DLBCL sc-cluster assignments in the NCI, BCCA, and merged datasets. Pies are color coded based on the sc-COO groups. Gray depicts the fraction of DLBCLs that remain unclassified by sc-COO classification. **(C)** Enrichment analysis in the sc-clusters of NCI, BCCA, or merged DLBCL cases based on their COO assignment (hypergeometric test with Benjamini-Hochberg correction,  $q < 0.05$ ). See also Fig. S4, Fig. S5, and Table S5.

intermediate (Group III; NCI: 57/190, 30%; BCCA: 29/166, 17%; overall: 86/356 cases, 24%), or late-GC stages, including late-LZ, PBL, and PreM (Group V; NCI: 62/190, 33%; BCCA: 70/166, 42%; overall: 132/356 cases, 37%; Fig. 8, A–C).

Notably, when applied separately to cases classified as GCB- or ABC-DLBCL, the sc-COO classification identified significantly favorable and unfavorable subsets of patients in both subgroups (Fig. 9, A and B). Specifically, patients classified as Group II (13/170, 8% GCB and 7/135, 5% ABC) displayed a particularly poor outcome independent of their COO subtype, while patients

classified as Group IV displayed very similar long-term PFS within both the GCB (76/170, 45%) and the ABC (10/135, 7%) subsets (Fig. 9, A and B).

A DHITsig (MYC/BCL2) was previously identified and associated with a DZ-like COO and poor outcome in GCB DLBCL (Ennishi et al., 2019), as confirmed here (Fig. S5, G and H). Notably, the sc classification identified distinct subgroups within the DHITsig-positive cases, which were associated with different PFS rates. In particular, the few cases belonging to Group IV showed a favorable outcome, with none of these



patients undergoing disease progression (Fig. S5 I). Conversely, DHITsig-positive cases displaying Group II signatures showed a significantly worse outcome (Fig. S5 I). Thus, the negative prognostic value of the DHITsig can be refined in combination with the sc-COO classification. These results indicate that the sc-COO classification can identify clinically relevant subgroups within the GCB- and ABC-DLBCLs, as well as DHITsig-positive cases.

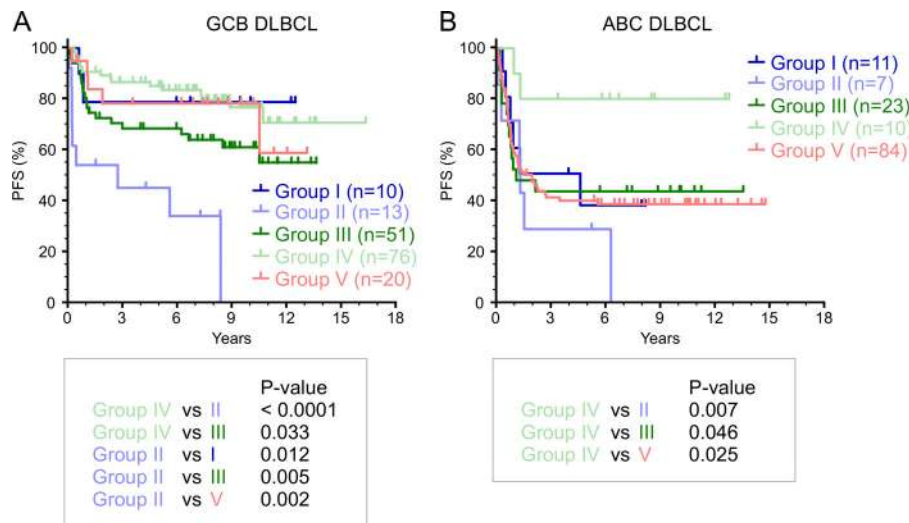
## Discussion

Transcriptomic analyses at the sc level in the context of development have been instrumental for the discovery and characterization

of novel populations, developmental stages, and transcriptional states. Our analysis of GC B cells allowed reconstructing their developmental stages and provided a granular reference to investigate the COO of lymphoma.

While the current view identifies the GC as a two-compartment structure (Mesin et al., 2016; Victora and Nussenzweig, 2012), our data expand on previous observations (Milpied et al., 2018) showing that a large fraction (in our analyses ~50%) of GC B cells do not express the classic immunophenotype of DZ or LZ cells. These cells display intermediate/low expression and/or coexpression of the CXCR4 and CD83 markers and represent transitional stages between





**Figure 9. Prognostic value of the sc-COO classification in the GCB- and ABC-DLBCL subtypes. (A and B)** Kaplan-Meier plots for PFS analysis in GCB-DLBCL ( $n = 170$ ; A) and ABC-DLBCL ( $n = 135$ ; B) cases as classified by the sc-COO classifier. P values were obtained by applying the Mantel-Cox test. See also Fig. S5.

the DZ and the LZ. Although previously published data (Milpied et al., 2018) reported the presence of two intermediate populations transitioning from the DZ to the LZ and vice versa, our observations expand on the heterogeneity of this compartment by identifying multiple distinct subpopulations that display different levels of relatedness to DZ and LZ GC B cells, as well as to memory B cell precursors. Of note, these intermediate subpopulations appear to represent the COO for a large fraction (~50%) of DLBCLs. Overall, our data confirm that the transition between DZ and LZ occurs through a gradient of transcriptional changes that define distinct intermediate phenotypes.

For a long time, cell division in the GC was considered a feature of DZ B cells (MacLennan, 1994). Later studies showed that cells could enter the S phase while in the LZ (Allen et al., 2007; Hauser et al., 2007), although progression into the G2/M stage of the cell cycle was not detected in this compartment (Victora et al., 2010). Our data on sorted LZ GC B cells clearly demonstrate that ~20% of LZ GC B cells express genes associated with all stages of the cell cycle. These data establish that a subset of phenotypically characterized LZ GC B cells is transcriptionally programmed to undergo complete cell division.

The distinct states characterizing LZ GC B cells appear to depend largely on the different stimuli that they receive. Our data are consistent with a model of progressive engagement of the BCR followed by T cell-mediated activation, which leads to DZ reentry or commitment to post-GC differentiation (Cyster and Allen, 2019; Mesin et al., 2016).

The signals and the transcription factors that control differentiation of GC B cells into memory or plasma cells are not fully understood. One notable finding of our analysis is that PreM cells and PBLs appear to originate from distinct phases of the GC reaction. In particular, PreM cells appear to start their commitment in the early stages of the LZ, while PBLs seem to originate from LZ cells that have fully transited through the LZ stages. These notions are consistent with previous observations showing that post-GC fate depends on the BCR affinity and strength of T cell-mediated activation and includes a temporal switch (Ise et al., 2018; Shinnakasu et al., 2016; Suan et al., 2017; Weisel et al., 2016).

Several studies reported on the identification of memory B cell precursors in the GC, most likely representing different stages in the memory B cell differentiation process (Laidlaw et al., 2017; Suan et al., 2017; Wang et al., 2017). We identified commitment to PreMs in an LZ population that displayed transcriptional evidence of BCR signaling and modest NF- $\kappa$ B activation and expressed *BACH2*. Previous work showed that *BACH2* promotes differentiation of GC B cells into memory B cells, and its expression inversely correlates with the strength of T cell help (Shinnakasu et al., 2016). In our study, PreM commitment in the LZ unfolds through an intermediate stage, leading to a population characterized by high expression of *CCR6*, previously shown to define memory B cell precursors in the GC (Suan et al., 2017), and of the *CD44* and *CD69* surface markers. The same population is also characterized by the expression of *BANK1* and *RASGRP2* transcripts. Although *BANK1*-deficient mice displayed enlarged GC (Aiba et al., 2006), a finding consistent with impaired post-GC differentiation, the role of *BANK1* and *RASGRP2* in memory B cell development remains to be elucidated. Finally, our analysis of PreMs revealed a potential role of interferon signaling in the commitment of GC B cells toward memory differentiation. Although previous studies established that, in humans, *STAT3* but not *STAT1* deficiency impairs memory B cell development (Avery et al., 2010; Domeier et al., 2016), we identified a significant transcriptional induction of *IFNGR1* and *STAT1*, but not *STAT3*, in the PreM population. Consistent with our results, *STAT1* and *IFNGR1*, but not *STAT3*, are transcriptionally induced in mature memory B cells compared with bulk GC B cells (data not shown). Further functional studies, including genetic-based approaches, are needed to address the role of these factors in memory B cell differentiation.

Our data provide transcriptional identities to PBL subpopulations and model a multistep process of PBL differentiation that starts in a subset of LZ cells displaying transcriptional changes associated with both BCR engagement and NF- $\kappa$ B activation. This process is consistent with the substantial body of data supporting the origin of plasma cells from GC cells expressing high affinity receptors and receiving T cell help (Ise



and Kurosaki, 2019; Nutt et al., 2015). During this differentiation process, PBLs remodel their transcriptional program by down-regulating GC markers and inducing plasma cell markers.

Gene expression-based classifications of DLBCL proved informative to dissect the heterogeneity of this malignancy and to identify subgroups associated with different outcomes (Alizadeh et al., 2000; Monti et al., 2005; Rosenwald et al., 2002). As of today, the COO classification represents the World Health Organization-approved reference to stratify patients into the clinically relevant GCB- or ABC-DLBCL subgroups. Using sc-based signatures, we revisited the COO of DLBCL, and we provide more detailed assignments for the putative normal counterparts of DLBCL. Notably, a significant fraction of COO-unclassified cases have found a clear address in various sc-defined clusters. In particular, and in contrast with current notions (Basso and Dalla-Favera, 2015; Pasqualucci and Dalla-Favera, 2018; Shaffer et al., 2012), our results indicate that a sizable (88%) fraction of ABC-DLBCLs are not yet committed to PBL differentiation and, in a significant subset of cases (25%), seem to derive from cells that are committing to PreM. Conversely, a clear PBL relationship emerged in 27% of the COO-unclassified cases. These conclusions appear robustly validated, since they consistently emerged from the analysis of two distinct panels of cases.

In general, our results indicate that the identification of additional GC subpopulations and corresponding sc-COO subgroups does lead to the recognition of additional prognostic subgroups. Despite the complexity of the emerging picture, it is notable that the sc-COO classification identifies dramatically distinct subgroups, both favorable and unfavorable, within GCB, ABC, and DHIT DLBCL. The gene signatures associated with these subgroups are ready for development in clinically established diagnostic platforms (Michaelsen et al., 2018) and eventually for further clinical validation.

## Materials and methods

### B cell isolation

Palatine tonsils were obtained at the Children's Hospital of Columbia-Presbyterian Medical Center as residual material from anonymous patients who had undergone elective tonsillectomy due to chronic tonsillitis in compliance with Regulatory Guideline 45 CFR 46.101 (b)(4) for Exempt Human Research Subjects of the US Department of Health and Human Services and according to protocols approved by the Columbia University Institutional Ethics Committee. Tonsil specimens were placed on ice immediately after surgical removal. Mononuclear cells were isolated from three donors by disaggregating tissues in RPMI 1640 medium (Gibco), followed by Ficoll-Isopaque (GE Healthcare) density centrifugation (Klein et al., 2003). Tonsillar mononucleated cells were stained using the following antibodies: anti-CD38-PE (clone HB7; BD Biosciences), anti-IgD-FITC (clone IA6-2; BD Biosciences), anti-CD3-FITC (clone UCHT1; Beckman Coulter), anti-CD184 (CXCR4)-Brilliant Violet 421 (clone 12G5; BioLegend), and anti-CD83-APC (clone HB15e; BioLegend). Total GC (CD38<sup>+</sup>/IgD<sup>-</sup>/CD3<sup>-</sup>), DZ (CD38<sup>+</sup>/IgD<sup>-</sup>/CD3<sup>-</sup>/CXCR4<sup>high</sup>/CD83<sup>low</sup>), and LZ (CD38<sup>+</sup>/IgD<sup>-</sup>/CD3<sup>-</sup>/CXCR4<sup>low</sup>/

CD83<sup>high</sup>) B cells were sorted using a FACSARIA flow cytometer (BD Biosciences). Data rendering was performed using FlowJo (TreeStar).

### Immunofluorescence analysis of paraffin-embedded lymphoid tissues

Human tonsil tissue was formalin fixed and paraffin embedded to prepare 3- $\mu$ m-thick slides to perform immunofluorescence analysis. Deparaffinization and rehydration of the tissue sections were performed before the heat-induced epitope retrieval (10 mM sodium citrate buffer, pH 6.0, or 10 mM Tris-Base, 1 mM EDTA, and 0.5% Tween20 buffer, pH 9.0). Endogenous peroxidases were quenched with 3% H<sub>2</sub>O<sub>2</sub> in 1 $\times$  PBS, and endogenous biotin was blocked with the Avidin-Biotin Blocking Kit (Vector Laboratories) according to manufacturer's instructions. All slides were blocked in 1 $\times$  PBS with 0.5% Tween20, 3% BSA (Sigma-Aldrich), and 5% serum of the species in which the secondary antibody was raised. Blocking was followed by overnight incubation at 4°C with primary antibody (Brescia et al., 2018). After repeated washes in 0.5% Tween20 in 1 $\times$  PBS, tissue sections were incubated at room temperature with HRP- or biotin-conjugated isotype-specific secondary antibodies. For detection of BCL6 (clone E5I8I; Cell Signaling Technology), CELF2 (Sigma-Aldrich), MEF2B (Brescia et al., 2018), PAX5 (NeoMarkers), and PRDM1 (clone EPRI6655; Abcam), a polymer-enhanced HRP-conjugated secondary antibody (EnVision<sup>+</sup> system; Agilent) was used, and immune-complexes were detected by tyramide-fluorescein isothiocyanate or tyramide-Cy3 amplification (in 1:1,000 dilution for 3 min; Perkin-Elmer). After tyramide detection, a heat-mediated antibody-stripping procedure, consisting of boiling the slides in citrate buffer, pH 6.0, was performed if primary antibodies from the identical host species were applied on the same slide (Tóth and Mezey, 2007). For detection of BANK1 (Sigma-Aldrich), FKBP11 (clone D-3; Santa Cruz Biotechnology), and IRF4 (clone M-17; Santa Cruz Biotechnology), we used the following biotin-conjugated secondary antibodies: Donkey anti-Goat IgG (Millipore), Donkey anti-Rabbit IgG (H+L; Jackson ImmunoResearch Labs), and Horse anti-Mouse IgG (Vector Laboratories). Streptavidin-fluorochrome Cy5-conjugated (Jackson ImmunoResearch Labs) or Cy3-conjugated (Molecular Probes) was added as a final step. After repeated washes in 0.5% Tween20 in 1 $\times$  PBS, tissue sections were mounted (VECTASHIELD Antifade Mounting Medium with DAPI; Vector Laboratories).

Immunofluorescence images were captured on the Aperio VERSA Digital Pathology Scanner (Leica Biosystems) using the 20 $\times$  objective using the Zyla 5.5 SCMOS camera (Andor). Image analysis was performed using the NIS-Elements Imaging Software (Nikon), ImageJ (Schindelin et al., 2015), the Aperio ImageScope (Leica Biosystems), and/or Halo (Indica Labs). GCs were identified by the expression of MEF2B or BCL6 and marked using the drawing tool in the Aperio ImageScope toolbar or the annotation tools in Halo. In the absence of GC markers, we annotated the GC using the image registration tool in Halo, which allows alignment and synchronization of two or more images from different serial sections of the same donor. The quantifications were restricted to the GC areas, as defined by expression of BCL6 or MEF2B. When analyzing expression of BANK1 (Fig. 5,

E and F; and Fig. S3 C), we restricted our analysis to the inner part of the GC (on average, three to five cell layers from the border of the GC) due to the inability to discriminate at the GC border between naive B cells and BANK1<sup>+</sup> GC B cells (Fig. S3 D). Cell counting on the acquired immunofluorescence images was performed manually using the ImageJ Cell Counter plugin (Schindelin et al., 2015). The results of manual counting were used as a reference to optimize the parameters of the automated analysis, which was performed using the Halo Image Analysis software. The immunofluorescence costainings were performed at least in duplicate in three to five independent donors, and one section for each donor was subjected to automated counting. Overall, in each section  $105 \pm 41$  total GCs were analyzed, corresponding to an average of  $264,484 \pm 112,297$  GC cells/section.

### Sc-gene expression profiling

Cell suspensions were diluted at the concentration of 1,000 cells/ $\mu$ l and analyzed using the Chromium Single Cell 3' Kit v2 and the Chromium Controller (10x Genomics). Sequencing of the libraries was performed on a NovaSeq6000 System (Illumina). The FASTQ files were aligned to the human GRCh38 reference genome using 10x Genomics Cell Ranger software v2.1.0 to create unique molecular identifier count tables of gene expression for each sample. Unique molecular identifier counts were normalized by library size. The six sequenced mRNA libraries displayed an average of  $6,347 \pm 1,006$  reads/cell and an average of  $1,893 \pm 187$  expressed genes/cell.

Counts from different patients were merged by log normalizing and scaling the data using the “NormalizeData” and “ScaleData” functions in the Seurat R package (Butler et al., 2018). The “vars.to.regress” feature of the ScaleData function was used to remove the batch effect between patients.

Individual samples or the batch-corrected data obtained from Seurat were reduced to their top 50 principle components using the Scikit-learn sklearn.decomposition.PCA Python package (parameters: `n_components = 50`, and `random_state = PCA_RANDOM_STATE`; Pedregosa et al., 2011).

UMAP 2D projections were created from the principal component analysis (PCA) matrix of the top 50 components using the UMAP Python package (parameters: `n_neighbors = 10`, `mist_dist = 0.1`, `spread = 1`, and `metric = “correlation”`; Becht et al., 2018). To display relative gene expression on the UMAP 2D projections, expression values were taken from the batch and scaled expression data and plotted as z-scores from  $-1.5$  to  $+1.5$ , such that they display low to high signal. The z-scored data do not indicate absolute zero (i.e., low expression [dark blue] represents the minimum expression of a gene, which is not necessarily zero).

The clustering of sc-expression profiles was performed using the PhenoGraph Python package on the normalized data for single samples or the batch-corrected data for merged samples using default parameters to build a k-nearest neighbor graph using the 20 nearest neighbors ( $k = 20$ ; Levine et al., 2015).

Once clusters were identified, we used the MAST R package to test for differentially expressed genes between individual clusters or pairs of clusters (Finak et al., 2015).

Separation of cells in G0-G1 and S-G2-M was performed at the level of clusters as defined by PhenoGraph. Each cluster was

assessed for expression of the reference genes PCNA, MKI67, CDK1, and CDC20 and then assigned to the G0-G1 or S-G2-M group. The two groups were reanalyzed independently by PCA, cluster identification (PhenoGraph), and differential gene expression (MAST). The new clusters in each group were then reassessed for the expression of the reference genes and, if necessary, reassigned to the correct G0-G1 or S-G2-M group. This iterative process was performed until all clusters in a group displayed a uniform G0-G1 or S-G2-M phenotype.

### CITE-seq

CITE-seq was performed as previously described (Stoeckius et al., 2017). After sorting, total GC B cells (CD3<sup>-</sup>/IgD<sup>-</sup>/CD38<sup>+</sup>) from one donor were incubated with TruStain FcX Fc blocking reagent (BioLegend) for 10 min on ice and then stained with the TotalSeq-A antibody cocktail ( $\sim 1 \mu$ g for each antibody) for 30 min on ice. The following antibodies were included in the cocktail: CXCR4 (clone 12G5), CD83 (clone HB15e), CD9 (clone HI9a; BioLegend), CCR6 (clone G034E3; BioLegend), CD69 (clone FN50; BioLegend), and CD44 (clone IM7; BioLegend). Each antibody was conjugated to a unique oligonucleotide bar code. The sc suspension (>95% viable) was diluted at the concentration of 1,000 cells/ $\mu$ l and analyzed using the Chromium Single Cell 3' Kit v3 and the Chromium Controller. The antibody-derived tags (ADTs) libraries were generated as detailed in the TotalSeq-A Antibodies Hashing with 10x Single Cell 3' Reagent Kit v3 protocol (BioLegend). Briefly, during the cDNA amplification step, the following ADT primer (0.2  $\mu$ M) was used: 5'-CCTTGGCAC CCGAGAATTCC-3'. After cDNA amplification ADT-derived (<180 bp) and mRNA-derived (>300 bp) cDNA fractions were size-selected using the SPRIselect Reagent (Beckman Coulter). The ADT libraries were amplified, cleaned up, and subjected to sequencing on a NovaSeq6000 System (Illumina).

The CITE-seq sequencing data, including mRNA and ADT libraries, were processed using the Feature Barcoding tool in Cell Ranger software v2.1.0 to produce gene and protein expression tables. The CITE-seq mRNA libraries showed on average  $10,764 \pm 7,769$  reads/cell and  $2,655 \pm 2,410$  genes/cell.

### Pseudo-temporal trajectory analysis

Pseudo-time analysis was performed using the Monocle 2.0 R package (Trapnell et al., 2014). Genes differentially expressed across PhenoGraph-identified clusters were used as an input for the Monocle analysis. For the heat map representation of pseudo-time genes, a time trace of each gene was taken using the “plot\_genes\_in\_pseudotime” function and dividing time into 100 equally sized bins. Time was measured by selecting the longest path through the trajectory plot going from  $t = 0$  to  $t = \max$ .

### Pseudo-spatial inference

We used novoSpaRc to reconstruct the spatial position of cells based on their sc-expression profiles (Nitzan et al., 2019). The probabilistic model was customized to assign cells to one of 326 zones in an oval-shaped virtual tissue representing a GC. The x, y, and z data were plotted using the contourf function in the Matplotlib library of Python3 (parameters: `levels = 100`, `antialiased = True`) to generate a smooth contour plot of the most

probable conformation of cells in the virtual tissue based on their phenotype (Hunter, 2007). The graphical representation of the inferred locations of GC subpopulations in an idealized GC structure displays normalized confidence scores ranging from 1 (high confidence) to 0.

### Gene set and pathway enrichment analyses

GSEA was performed using GSEA 2.2.0 with gene randomization (Subramanian et al., 2005). Gene sets were compiled from the significantly differentially expressed genes identified by comparing a given cluster with the remaining samples in the dataset. Significantly differentially expressed genes were filtered to have a  $\log_2$  expression fold change  $\geq 1.5$ . If these criteria yielded fewer than 50 genes, the top 50 differentially expressed genes by fold change were used as gene sets.

Pathway enrichment analysis was performed on the KEGG (c2.cp.kegg.v6.2), BioCarta (c2.cp.biocarta.v6.2), and Hallmark (h.all.v7.0) collections from the Molecular Signature Database v6.2 (<http://software.broadinstitute.org/gsea/msigdb/index.jsp>; Liberzon, 2014). A hypergeometric test assessing  $P(X \geq N)$  with a Benjamini-Hochberg false discovery rate correction was used to test enrichment in pathways using a background gene pool size of 45,956 to match the Molecular Signature Database.

### sc-COO tumor classifier

To classify DLBCL bulk RNAseq profiles, we developed a modified weighted-vote algorithm classifier (Golub et al., 1999). We used MAST to identify differentially expressed genes in each cluster identified in the sc-RNAseq data and divided them into up- and down-regulated subsets. We next took bulk gene expression profiles for a large panel of DLBCL and z-scored each gene per row. We used the relative expression of genes in a heterogeneous panel as a proxy for measuring similarity to a cluster. Using the sc-cluster signatures as reference, for each sample we assigned to each gene a score of either +1, when its expression trended in the same direction as the reference signature, or -1, when it did not. The classifier score for a cluster signature was the normalized sum of its gene scores, ranging from -1 to +1, with a positive score indicating similarity to the phenotype of interest. We used bootstrapping to randomize the classifier genes 1,000 times and to create a distribution to estimate the P value of the classifier scores. We set the significance level at 0.05 and only classified a sample as belonging to a cluster phenotype if its absolute score was  $\geq 0.3$  (outside the 95% confidence interval). If a sample was assigned to more than one class with a significant score, we used the best score (lowest P value). When a sample displayed the same best score for two classes, we used the call that was consistent in both classifications using the top 50 and the top 40 gene signatures. Contiguous classes and/or phenotypes were defined as follows: (1) contiguous class: when classes are next to each other in the provided order of the GC subpopulations; (2) same phenotype: when noncontiguous classes belong to the same phenotypic group (i.e., DZ-a and DZ-c); and (3) contiguous phenotype: when noncontiguous classes are developmentally linked (i.e., Int-d and PreM).

The scoring results were also summarized in a heat map format where samples were associated with the signature for

which they displayed the most significant (highest score, lowest P value) similarity.

The classifier was applied to two distinct datasets of bulk RNAseq expression profiles. The NCI-DLBCL dataset was downloaded from <https://portal.gdc.cancer.gov/projects/NCICCR-DLBCL> (Schmitz et al., 2018). These data were in the form of already mapped BAM files on GRCh38. The BCCA-DLBCL dataset was downloaded from <https://www.ebi.ac.uk/ega/datasets/EGAD00001003783> (Arthur et al., 2018; Ennishi et al., 2019). These data were in the form of already mapped BAM files on GRCh37. The alignment data from each dataset was converted to gene expression data using featureCounts 1.6.3 (Liao et al., 2014). The expression data were subsequently normalized to transcripts per million mapped reads and variance stabilizing transform using DESeq2 (Love et al., 2014). To remove outlier samples, we filtered out samples where 90% of the signal (sum of transcripts per million mapped reads for a sample) was contained in  $< 1,000$  genes.

### Genetic classification of DLBCL

The NCI-DLBCL dataset (Schmitz et al., 2018) was classified based on previously reported genetic classes (Chapuy et al., 2018) by performing nonnegative matrix factorization consensus clustering described in (Chapuy et al., 2018). First, to make the input matrix of the NCI-DLBCL samples for clustering, we downloaded the mutation, copy number alteration, and fusion gene datasets from the GDC Data Portal using gdc-client (accession ID: phs001444). Using these datasets, we prepared an input matrix of 153 genomic features, including mutations in 85 candidate cancer genes, 65 somatic copy number alterations, and three structural variants (available in the form of fusion genes). Of note, the original classifier used the structural variants of eight genes, out of which only MYC, BCL2, and BCL6 were available for the NCI-DLBCL dataset. We then used the computational scripts for clustering from GitHub ([https://github.com/broadinstitute/DLBCL\\_Nat\\_Med\\_April\\_2018](https://github.com/broadinstitute/DLBCL_Nat_Med_April_2018)) on the input matrix described above. The NCI-DLBCL dataset was classified in C1 to C5 using the default parameters as described by the developer (Chapuy et al., 2018).

### Statistical analysis

The MAST R package was used to identify differentially expressed genes in each cluster identified in the sc-RNAseq data. For the gene set and pathway enrichment analysis, a hypergeometric test with a Benjamini-Hochberg false discovery rate correction was used. GraphPad Prism v.6.0 was used to assess the significance of tumor survival using the Mantel-Cox and Gehan-Breslow-Wilcoxon tests.

Detailed information on the statistical test, number of replicates/samples (defined as  $n$ ) used in each experiment, and measurement precision is described in the Materials and methods and is reported in the figure legends. Significance was associated to  $P \leq 0.05$ .

### Data availability

The DZ and LZ bulk RNAseq gene expression data are available in the Gene Expression Omnibus database under accession no.



GSE139833. sc-gene expression data are available under accession no. GSE139891.

### Code availability

All analyses and visualizations were performed in R and Python with the following open-source algorithms and tools as described above: 10x Genomics Cell Ranger software v2.1.0 (<https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation>), Seurat (<https://satijalab.org/seurat/>), Phenograph (<https://github.com/jacoblevine/PhenoGraph>), UMAP (<https://github.com/lmcinnes/umap>), MAST (<https://bioconductor.org/packages/release/bioc/html/MAST.html>), DESeq2 (<https://bioconductor.org/packages/release/bioc/html/DESeq2.html>), Monocle (<http://cole-trapnell-lab.github.io/monocle-release/>), novoSpaRc (<https://github.com/rajewsky-lab/novosparc>), the Scikit-learn implementations of PCA, the Matplotlib contour function, featureCounts (<https://bioconductor.org/packages/release/bioc/html/Rsubread.html>), and GSEA (<https://www.gsea-msigdb.org/gsea/index.jsp>). Computer code is available upon reasonable request.

### Online supplemental material

**Fig. S1**, related to **Fig. 1**, identifies GC B cell subpopulations by sc-RNAseq and CITE-seq in distinct donors. **Fig. S2**, related to **Fig. 2**, shows GC B cell developmental stages. **Fig. S3**, related to **Fig. 5**, presents memory B cell precursors. **Fig. S4**, related to **Fig. 7**, shows development of a classifier based on the sc-RNAseq gene signatures. **Fig. S5**, related to **Figs. 7, 8, and 9**, shows how GC B cell signatures define lymphoma subgroups. Table S1, related to **Fig. 2**, shows cluster cell counts and results of differential expression analysis in GC B cells in G0-G1. Table S2, related to **Fig. 2**, lists differential expression genes obtained from the analysis of GC B cells in G0-G1 and used in the GSEA reported in **Fig. S2, B–D**. Table S3, related to **Fig. 3**, shows cluster cell counts, results of differential expression analysis in DZ GC B cells in S-G2-M and G0-G1, and their respective pathway enrichment analyses. Table S4, related to **Fig. 4**, shows cluster cell counts, results of differential expression analysis in LZ GC B cells in S-G2-M and G0-G1, and their respective pathway enrichment analyses. Table S5, related to **Fig. 7**, shows classification of DLBCLs according to the sc-COO classifier, as well as DLBCL identifiers, sc-COO class assignment, genetic subtypes, DHITsig groups, sc-COO classes, score and P value for the top sc-COO class, and sc-COO groups.

### Acknowledgments

We would like to thank Hongyan Tang for technical support, David Dominguez-Sola and Ulf Klein for critical reading of the manuscript, and Ryan D. Morin, Christian Steidl, and David Scott for expediting the access to their published data. This manuscript was prepared using a limited access dataset (EGAD00001003783) obtained from BC CANCER and does not necessarily reflect the opinions or views of BC CANCER.

This study was supported by National Institutes of Health grant R35CA-210105 (to R. Dalla-Favera) and funded in part through the NIH/NCI Cancer Center Support grant P30CA013696. This

research used the resources of the Cancer Center Flow Core Facility (Columbia University), the Genomics and High Throughput Screening Shared Resource (Columbia University), and the Digital and Computational Pathology Laboratory in the Department of Pathology and Cell Biology at Columbia University Irving Medical Center.

Author contributions: R. Dalla-Favera and K. Basso designed and supervised the study; C. Corinaldesi, Q. Shen, and K. Basso performed scRNAseq experiments; C. Corinaldesi, Q. Shen, and Z. Wang performed immunofluorescence analyses; N. Compagno and K. Basso generated RNAseq data of DZ and LZ cells; C. Corinaldesi and K. Basso performed survival analyses; A.B. Holmes and K. Basso analyzed scRNAseq data; A.B. Holmes, R. Kumar, and M. Nitzan performed computational analyses; E. Grunstein collected normal lymphoid tissue; L. Pasqualucci collected data and clinical information; A.B. Holmes, C. Corinaldesi, L. Pasqualucci, R. Dalla-Favera, and K. Basso wrote the paper. All authors reviewed and edited the final manuscript.

Disclosures: N. Compagno is currently employed at Novartis. No other disclosures were reported.

Submitted: 12 March 2020

Revised: 12 May 2020

Accepted: 15 May 2020

### References

- Aiba, Y., T. Yamazaki, T. Okada, K. Gotoh, H. Sanjo, M. Ogata, and T. Kurosaki. 2006. BANK negatively regulates Akt activation and subsequent B cell responses. *Immunity*. 24:259–268. <https://doi.org/10.1016/j.immuni.2006.01.002>
- Alizadeh, A.A., M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 403:503–511. <https://doi.org/10.1038/35000501>
- Allen, C.D., K.M. Ansel, C. Low, R. Lesley, H. Tamamura, N. Fujii, and J.G. Cyster. 2004. Germinal center dark and light zone organization is mediated by CXCR4 and CXCR5. *Nat. Immunol.* 5:943–952. <https://doi.org/10.1038/nri1100>
- Allen, C.D., T. Okada, H.L. Tang, and J.G. Cyster. 2007. Imaging of germinal center selection events during affinity maturation. *Science*. 315:528–531. <https://doi.org/10.1126/science.1136736>
- Angelini-Duclos, C., G. Cattoretti, K.I. Lin, and K. Calame. 2000. Commitment of B lymphocytes to a plasma cell fate is associated with Blimp-1 expression in vivo. *J. Immunol.* 165:5462–5471. <https://doi.org/10.4049/jimmunol.165.10.5462>
- Arthur, S.E., A. Jiang, B.M. Grande, M. Alcaide, R. Cojocaru, C.K. Rushton, A. Mottok, L.K. Hilton, P.K. Lat, E.Y. Zhao, et al. 2018. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nat. Commun.* 9:4001. <https://doi.org/10.1038/s41467-018-06354-3>
- Avery, D.T., E.K. Deenick, C.S. Ma, S. Suryani, N. Simpson, G.Y. Chew, T.D. Chan, U. Palendira, J. Bustamante, S. Boisson-Dupuis, et al. 2010. B cell-intrinsic signaling through IL-21 receptor and STAT3 is required for establishing long-lived antibody responses in humans. *J. Exp. Med.* 207:155–171. <https://doi.org/10.1084/jem.20091706>
- Basso, K., and R. Dalla-Favera. 2012. Roles of BCL6 in normal and transformed germinal center B cells. *Immunol. Rev.* 247:172–183. <https://doi.org/10.1111/j.1600-065X.2012.01112.x>
- Basso, K., and R. Dalla-Favera. 2015. Germinal centres and B cell lymphomagenesis. *Nat. Rev. Immunol.* 15:172–184. <https://doi.org/10.1038/nri3814>
- Becht, E., L. McInnes, J. Healy, C.A. Dutertre, I.W.H. Kwok, L.G. Ng, F. Ginhoux, and E.W. Newell. 2018. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* Available at: <https://doi.org/10.1038/nbt.4314>.

- Béguelin, W., M.A. Rivas, M.T. Calvo Fernández, M. Teater, A. Purwada, D. Redmond, H. Shen, M.F. Challman, O. Elemento, A. Singh, et al. 2017. EZH2 enables germinal center formation through epigenetic silencing of CDKN1A and an Rb-E2F1 feedback loop. *Nat. Commun.* 8:877. <https://doi.org/10.1038/s41467-017-01029-x>
- Brescia, P., C. Schneider, A.B. Holmes, Q. Shen, S. Hussein, L. Pasqualucci, K. Basso, and R. Dalla-Favera. 2018. MEF2B Instructs Germinal Center Development and Acts as an Oncogene in B Cell Lymphomagenesis. *Cancer Cell.* 34:453–465.e9. <https://doi.org/10.1016/j.ccell.2018.08.006>
- Butler, A., P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36:411–420. <https://doi.org/10.1038/nbt.4096>
- Calado, D.P., Y. Sasaki, S.A. Godinho, A. Pellerin, K. Köchert, B.P. Sleckman, I.M. de Alborán, M. Janz, S. Rodig, and K. Rajewsky. 2012. The cell-cycle regulator c-Myc is essential for the formation and maintenance of germinal centers. *Nat. Immunol.* 13:1092–1100. <https://doi.org/10.1038/ni.2418>
- Caron, G., S. Le Gallou, T. Lamy, K. Tarte, and T. Fest. 2009. CXCR4 expression functionally discriminates centroblasts versus centrocytes within human germinal center B cells. *J. Immunol.* 182:7595–7602. <https://doi.org/10.4049/jimmunol.0804272>
- Chapuy, B., C. Stewart, A.J. Dunford, J. Kim, A. Kamburov, R.A. Redd, M.S. Lawrence, M.G.M. Roemer, A.J. Li, M. Ziepert, et al. 2018. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat. Med.* 24:679–690. <https://doi.org/10.1038/s41591-018-0016-8>
- Cyster, J.G., and C.D.C. Allen. 2019. B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell.* 177:524–540. <https://doi.org/10.1016/j.cell.2019.03.016>
- De Silva, N.S., and U. Klein. 2015. Dynamics of the Germinal Center B-Cell Reaction. *Nat. Rev. Immunol.* 15:137–148. <https://doi.org/10.1038/nri3804>
- Domeier, P.P., S.B. Chodiseti, C. Soni, S.L. Schell, M.J. Elias, E.B. Wong, T.K. Cooper, D. Kitamura, and Z.S. Rahaman. 2016. IFN- $\gamma$  receptor and STAT1 signaling in B cells are central to spontaneous germinal center formation and autoimmunity. *J. Exp. Med.* 213:715–732. <https://doi.org/10.1084/jem.20151722>
- Dominguez-Sola, D., G.D. Victora, C.Y. Ying, R.T. Phan, M. Saito, M.C. Nussenzweig, and R. Dalla-Favera. 2012. The proto-oncogene MYC is required for selection in the germinal center and cyclic reentry. *Nat. Immunol.* 13:1083–1091. <https://doi.org/10.1038/ni.2428>
- Dominguez-Sola, D., J. Kung, A.B. Holmes, V.A. Wells, T. Mo, K. Basso, and R. Dalla-Favera. 2015. The FOXO1 Transcription Factor Instructs the Germinal Center Dark Zone Program. *Immunity.* 43:1064–1074. <https://doi.org/10.1016/j.immuni.2015.10.015>
- Ebinu, J.O., D.A. Bottorff, E.Y. Chan, S.L. Stang, R.J. Dunn, and J.C. Stone. 1998. RasGRP, a Ras guanyl nucleotide- releasing protein with calcium- and diacylglycerol-binding motifs. *Science.* 280:1082–1086. <https://doi.org/10.1126/science.280.5366.1082>
- Ennishi, D., A. Jiang, M. Boyle, B. Collinge, B.M. Grande, S. Ben-Neriah, C. Rushton, J. Tang, N. Thomas, G.W. Slack, et al. 2019. Double-Hit Gene Expression Signature Defines a Distinct Subgroup of Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma. *J. Clin. Oncol.* 37:190–201. <https://doi.org/10.1200/JCO.18.01583>
- Ersching, J., A. Efeyan, L. Mesin, J.T. Jacobsen, G. Pasqual, B.C. Grabner, D. Dominguez-Sola, D.M. Sabatini, and G.D. Victora. 2017. Germinal Center Selection and Affinity Maturation Require Dynamic Regulation of mTORC1 Kinase. *Immunity.* 46:1045–1058.e6. <https://doi.org/10.1016/j.immuni.2017.06.005>
- Falini, B., M. Fizzotti, A. Pucciarini, B. Bigerna, T. Marafioti, M. Gambacorta, R. Pacini, C. Alunni, L. Natali-Tanci, B. Ugolini, et al. 2000. A monoclonal antibody (MUM1p) detects expression of the MUM1/IRF4 protein in a subset of germinal center B cells, plasma cells, and activated T cells. *Blood.* 95:2084–2092. <https://doi.org/10.1182/blood.V95.6.2084>
- Finak, G., A. McDavid, M. Yajima, J. Deng, V. Gersuk, A.K. Shalek, C.K. Slichter, H.W. Miller, M.J. McElrath, M. Prlic, et al. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16:278. <https://doi.org/10.1186/s13059-015-0844-5>
- Finkin, S., H. Hartweg, T.Y. Oliveira, E.E. Kara, and M.C. Nussenzweig. 2019. Protein Amounts of the MYC Transcription Factor Determine Germinal Center B Cell Division Capacity. *Immunity.* 51:324–336.e5. <https://doi.org/10.1016/j.immuni.2019.06.013>
- Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 286:531–537. <https://doi.org/10.1126/science.286.5439.531>
- Hauser, A.E., T. Junt, T.R. Mempel, M.W. Sneddon, S.H. Kleinstein, S.E. Henrickson, U.H. von Andrian, M.J. Shlomchik, and A.M. Haberman. 2007. Definition of germinal-center B cell migration in vivo reveals predominant intrazonal circulation patterns. *Immunity.* 26:655–667. <https://doi.org/10.1016/j.immuni.2007.04.008>
- Hunter, J.D.. 2007. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* 9:90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Ise, W., and T. Kurosaki. 2019. Plasma cell differentiation during the germinal center reaction. *Immunol. Rev.* 288:64–74. <https://doi.org/10.1111/imr.12751>
- Ise, W., K. Fujii, K. Shiroguchi, A. Ito, K. Kometani, K. Takeda, E. Kawakami, K. Yamashita, K. Suzuki, T. Okada, et al. 2018. T Follicular Helper Cell-Germinal Center B Cell Interaction Strength Regulates Entry into Plasma Cell or Recycling Germinal Center Cell Fate. *Immunity.* 48:702–715.e4. <https://doi.org/10.1016/j.immuni.2018.03.027>
- Klein, U., Y. Tu, G.A. Stolovitzky, J.L. Keller, J. Haddad, Jr., V. Miljkovic, G. Cattoretti, A. Califano, and R. Dalla-Favera. 2003. Transcriptional analysis of the B cell germinal center reaction. *Proc. Natl. Acad. Sci. USA.* 100:2639–2644. <https://doi.org/10.1073/pnas.0437996100>
- Laidlaw, B.J., T.H. Schmidt, J.A. Green, C.D. Allen, T. Okada, and J.G. Cyster. 2017. The Eph-related tyrosine kinase ligand Ephrin-B1 marks germinal center and memory precursor B cells. *J. Exp. Med.* 214:639–649. <https://doi.org/10.1084/jem.20161461>
- Levine, J.H., E.F. Simonds, S.C. Bendall, K.L. Davis, A.D. Amir, M.D. Tadmor, O. Litvin, H.G. Fienberg, A. Jager, E.R. Zunder, et al. 2015. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-Like Cells that Correlate with Prognosis. *Cell.* 162:184–197. <https://doi.org/10.1016/j.cell.2015.05.047>
- Liao, Y., G.K. Smyth, and W. Shi. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 30:923–930. <https://doi.org/10.1093/bioinformatics/btt656>
- Liberzon, A.. 2014. A description of the Molecular Signatures Database (MSigDB) Web site. *Methods Mol. Biol.* 1150:153–160. [https://doi.org/10.1007/978-1-4939-0512-6\\_9](https://doi.org/10.1007/978-1-4939-0512-6_9)
- Love, M.I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. <https://doi.org/10.1186/s13059-014-0550-8>
- MacLennan, I.C.. 1994. Germinal centers. *Annu. Rev. Immunol.* 12:117–139. <https://doi.org/10.1146/annurev.iy.12.040194.001001>
- Mesin, L., J. Ersching, and G.D. Victora. 2016. Germinal Center B Cell Dynamics. *Immunity.* 45:471–482. <https://doi.org/10.1016/j.immuni.2016.09.001>
- Michaelsen, T.Y., J. Richter, R.F. Brøndum, W. Klapper, H.E. Johnsen, M. Albertsen, K. Dybkær, and M. Bøgsted. 2018. A B-cell-associated gene signature classification of diffuse large B-cell lymphoma by Nano-String technology. *Blood Adv.* 2:1542–1546. <https://doi.org/10.1182/bloodadvances.2018017988>
- Milpied, P., I. Cervera-Marzal, M.L. Mollicella, B. Tesson, G. Brisou, A. Traverse-Glehen, G. Salles, L. Spinelli, and B. Nadel. 2018. Human germinal center transcriptional programs are de-synchronized in B cell lymphoma. *Nat. Immunol.* 19:1013–1024. <https://doi.org/10.1038/s41590-018-0181-4>
- Monti, S., K.J. Savage, J.L. Kutok, F. Feuerhake, P. Kurtin, M. Mihm, B. Wu, L. Pasqualucci, D. Neuberger, R.C. Aguiar, et al. 2005. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood.* 105:1851–1861. <https://doi.org/10.1182/blood-2004-07-2947>
- Nitzan, M., N. Karaiskos, N. Friedman, and N. Rajewsky. 2019. Gene expression cartography. *Nature.* 576:132–137. <https://doi.org/10.1038/s41586-019-1773-3>
- Nutt, S.L., P.D. Hodgkin, D.M. Tarlinton, and L.M. Corcoran. 2015. The generation of antibody-secreting plasma cells. *Nat. Rev. Immunol.* 15:160–171. <https://doi.org/10.1038/nri3795>
- Pasqualucci, L., and R. Dalla-Favera. 2018. Genetics of diffuse large B-cell lymphoma. *Blood.* 131:2307–2319. <https://doi.org/10.1182/blood-2017-11-764332>
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12:2825–2830.
- Phan, T.G., D. Paus, T.D. Chan, M.L. Turner, S.L. Nutt, A. Basten, and R. Brink. 2006. High affinity germinal center B cells are actively selected into the plasma cell compartment. *J. Exp. Med.* 203:2419–2424. <https://doi.org/10.1084/jem.20061254>

- Roco, J.A., L. Mesin, S.C. Binder, C. Nefzger, P. Gonzalez-Figueroa, P.F. Canete, J. Ellyard, Q. Shen, P.A. Robert, J. Cappello, et al. 2019. Class-Switch Recombination Occurs Infrequently in Germinal Centers. *Immunity*. 51:337–350.e7. <https://doi.org/10.1016/j.immuni.2019.07.001>
- Rosenwald, A., G. Wright, W.C. Chan, J.M. Connors, E. Campo, R.I. Fisher, R.D. Gascoyne, H.K. Muller-Hermelink, E.B. Smeland, J.M. Giltnane, et al; Lymphoma/Leukemia Molecular Profiling Project. 2002. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* 346:1937–1947. <https://doi.org/10.1056/NEJMoa012914>
- Ruer-Laventie, J., L. Simoni, J.N. Schickel, A. Soley, M. Duval, A.M. Knapp, L. Marcellin, D. Lamon, A.S. Korganow, T. Martin, et al. 2015. Overexpression of Fkbp11, a feature of lupus B cells, leads to B cell tolerance breakdown and initiates plasma cell differentiation. *Immun. Inflamm. Dis.* 3:265–279. <https://doi.org/10.1002/iid3.65>
- Sander, S., V.T. Chu, T. Yasuda, A. Franklin, R. Graf, D.P. Calado, S. Li, K. Imami, M. Selbach, M. Di Virgilio, et al. 2015. PI3 Kinase and FOXO1 Transcription Factor Activity Differentially Control B Cells in the Germinal Center Light and Dark Zones. *Immunity*. 43:1075–1086. <https://doi.org/10.1016/j.immuni.2015.10.021>
- Schindelin, J., C.T. Rueden, M.C. Hiner, and K.W. Eliceiri. 2015. The ImageJ ecosystem: An open platform for biomedical image analysis. *Mol. Reprod. Dev.* 82:518–529. <https://doi.org/10.1002/mrd.22489>
- Schmitz, R., G.W. Wright, D.W. Huang, C.A. Johnson, J.D. Phelan, J.Q. Wang, S. Roulland, M. Kasbekar, R.M. Young, A.L. Shaffer, et al. 2018. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* 378:1396–1407. <https://doi.org/10.1056/NEJMoa1801445>
- Sha, C., S. Barrans, F. Cucco, M.A. Bentley, M.A. Care, T. Cummin, H. Kennedy, J.S. Thompson, R. Uddin, L. Worrillow, et al. 2019. Molecular High-Grade B-Cell Lymphoma: Defining a Poor-Risk Group That Requires Different Approaches to Therapy. *J. Clin. Oncol.* 37:202–212. <https://doi.org/10.1200/JCO.18.01314>
- Shaffer, A.L., III, R.M. Young, and L.M. Staudt. 2012. Pathogenesis of human B cell lymphomas. *Annu. Rev. Immunol.* 30:565–610. <https://doi.org/10.1146/annurev-immunol-020711-075027>
- Shinnakasu, R., T. Inoue, K. Kometani, S. Moriyama, Y. Adachi, M. Nakayama, Y. Takahashi, H. Fukuyama, T. Okada, and T. Kurosaki. 2016. Regulated selection of germinal-center cells into the memory B cell compartment. *Nat. Immunol.* 17:861–869. <https://doi.org/10.1038/ni.3460>
- Shuai, K., C. Schindler, V.R. Prezioso, and J.E. Darnell, Jr.. 1992. Activation of transcription by IFN-gamma: tyrosine phosphorylation of a 91-kD DNA binding protein. *Science*. 258:1808–1812. <https://doi.org/10.1126/science.1281555>
- Stoeckius, M., C. Hafemeister, W. Stephenson, B. Houck-Loomis, P.K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*. 14:865–868. <https://doi.org/10.1038/nmeth.4380>
- Suan, D., N.J. Kräutler, J.L.V. Maag, D. Butt, K. Bourne, J.R. Hermes, D.T. Avery, C. Young, A. Statham, M. Elliott, et al. 2017. CCR6 Defines Memory B Cell Precursors in Mouse and Human Germinal Centers, Revealing Light-Zone Location and Predominant Low Antigen Affinity. *Immunity*. 47:1142–1153.e4. <https://doi.org/10.1016/j.immuni.2017.11.022>
- Subramanian, A., P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*. 102:15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Tóth, Z.E., and E. Mezey. 2007. Simultaneous visualization of multiple antigens with tyramide signal amplification using antibodies from the same species. *J. Histochem. Cytochem.* 55:545–554. <https://doi.org/10.1369/jhc.6A7134.2007>
- Trapnell, C., D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N.J. Lennon, K.J. Livak, T.S. Mikkelsen, and J.L. Rinn. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32:381–386. <https://doi.org/10.1038/nbt.2859>
- Victoria, G.D., and M.C. Nussenzweig. 2012. Germinal centers. *Annu. Rev. Immunol.* 30:429–457. <https://doi.org/10.1146/annurev-immunol-020711-075032>
- Victoria, G.D., T.A. Schwickert, D.R. Fooksman, A.O. Kamphorst, M. Meyer-Hermann, M.L. Dustin, and M.C. Nussenzweig. 2010. Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell*. 143:592–605. <https://doi.org/10.1016/j.cell.2010.10.032>
- Victoria, G.D., D. Dominguez-Sola, A.B. Holmes, S. Deroubaix, R. Dalla-Favera, and M.C. Nussenzweig. 2012. Identification of human germinal center light and dark zone cells and their relationship to human B-cell lymphomas. *Blood*. 120:2240–2248. <https://doi.org/10.1182/blood-2012-03-415380>
- Wang, Y., J. Shi, J. Yan, Z. Xiao, X. Hou, P. Lu, S. Hou, T. Mao, W. Liu, Y. Ma, et al. 2017. Germinal-center development of memory B cells driven by IL-9 from follicular helper T cells. *Nat. Immunol.* 18:921–930. <https://doi.org/10.1038/ni.3788>
- Weisel, F.J., G.V. Zuccarino-Catania, M. Chikina, and M.J. Shlomchik. 2016. A Temporal Switch in the Germinal Center Determines Differential Output of Memory B and Plasma Cells. *Immunity*. 44:116–130. <https://doi.org/10.1016/j.immuni.2015.12.004>
- Wright, G., B. Tan, A. Rosenwald, E.H. Hurt, A. Wiestner, and L.M. Staudt. 2003. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc. Natl. Acad. Sci. USA*. 100:9991–9996. <https://doi.org/10.1073/pnas.1732008100>
- Yokoyama, K., I.H. Su Ih, T. Tezuka, T. Yasuda, K. Mikoshiba, A. Tarakhovskiy, and T. Yamamoto. 2002. BANK regulates BCR-induced calcium mobilization by promoting tyrosine phosphorylation of IP(3) receptor. *EMBO J.* 21:83–92. <https://doi.org/10.1093/emboj/21.1.83>



## Supplemental material

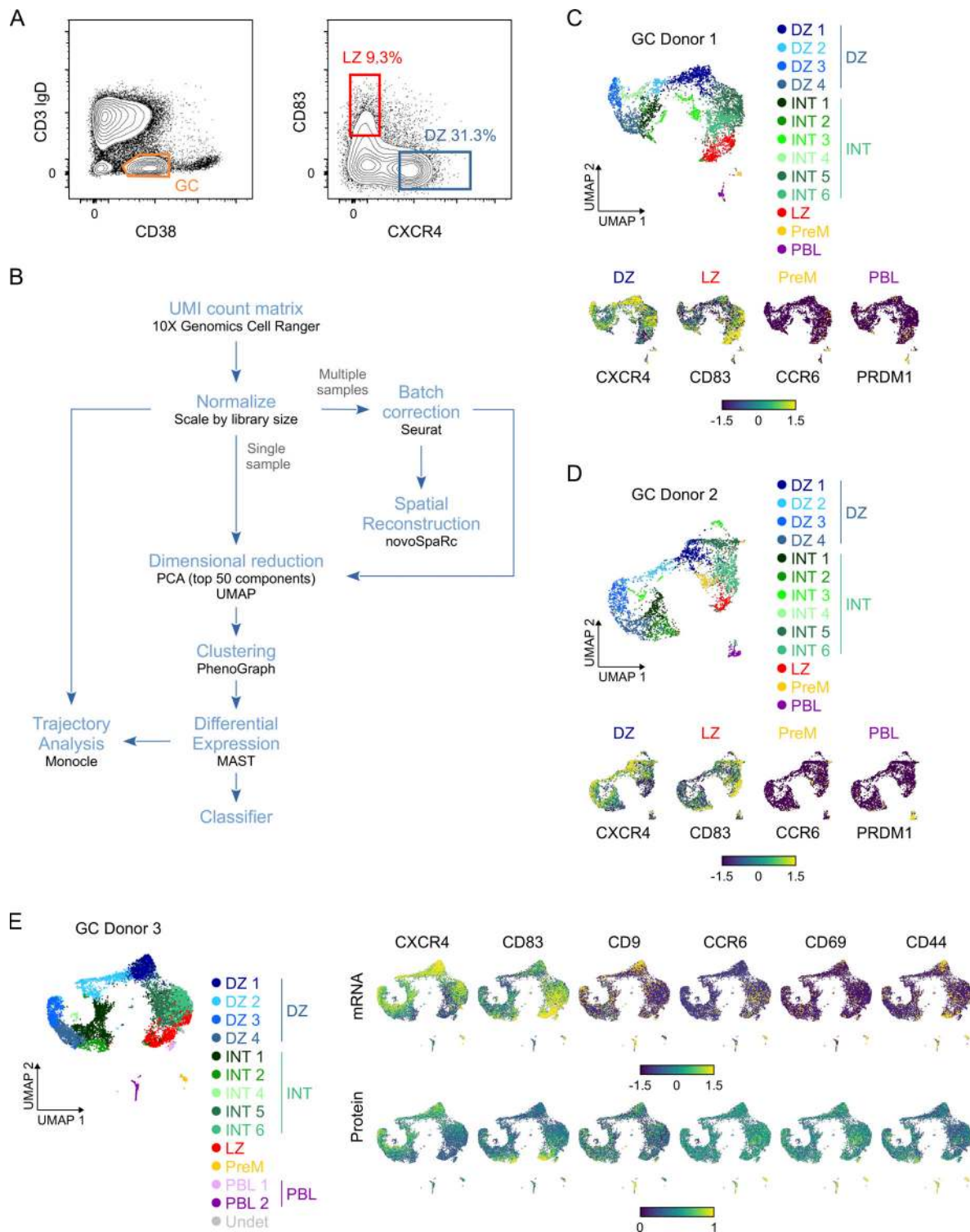


Figure S1. **Identification of GC B cell subpopulations by sc-RNAseq and CITE-seq in distinct donors.** Related to Fig. 1. **(A)** Representative counterplots and gating strategy from cytofluorimetric analysis of human GC ( $CD3^-/IgD^-/CD38^+$ ), DZ ( $CD3^-/IgD^-/CD38^+/CD83^{lo}/CXCR4^{hi}$ ), and LZ ( $CD3^-/IgD^-/CD38^+/CD83^{hi}/CXCR4^{lo}$ ) B cells isolated from tonsil tissue are shown. **(B)** Overview of the computational steps used to analyze sc-RNAseq data indicating the software tools used (black labels). **(C)** UMAP projection of 4,759 GC B cells ( $CD3^-/IgD^-/CD38^+$ ) isolated from Donor 1. Cells are colored by clusters identified by PhenoGraph and assigned to one of the following groups: DZ, intermediate (INT), LZ, PreM, or PBL. UMAP projections colored by the z-scored expression of CXCR4 (DZ), CD83 (LZ), CCR6 (PreM), and PRDM1 (PBL) are displayed below. **(D)** UMAP projection of 3,609 GC B cells ( $CD3^-/IgD^-/CD38^+$ ) isolated from Donor 2. Cells were clustered and colored using the same methods as for Donor 1. Below, the same gene expression markers described for Donor 1 are shown for Donor 2. **(E)** CITE-seq analysis of 8,871 GC B cells ( $CD3^-/IgD^-/CD38^+$ ) isolated from Donor 3. Left: UMAP projection based on mRNA data. Cells were clustered and colored using the same methods as for Donors 1 and 2. Right: UMAP projections colored by normalized z-scored mRNA expression (top) and center log ratio normalized protein expression (bottom) as measured by ADTs for selected markers (CXCR4, CD83, CD9, CCR6, CD69, and CD44).

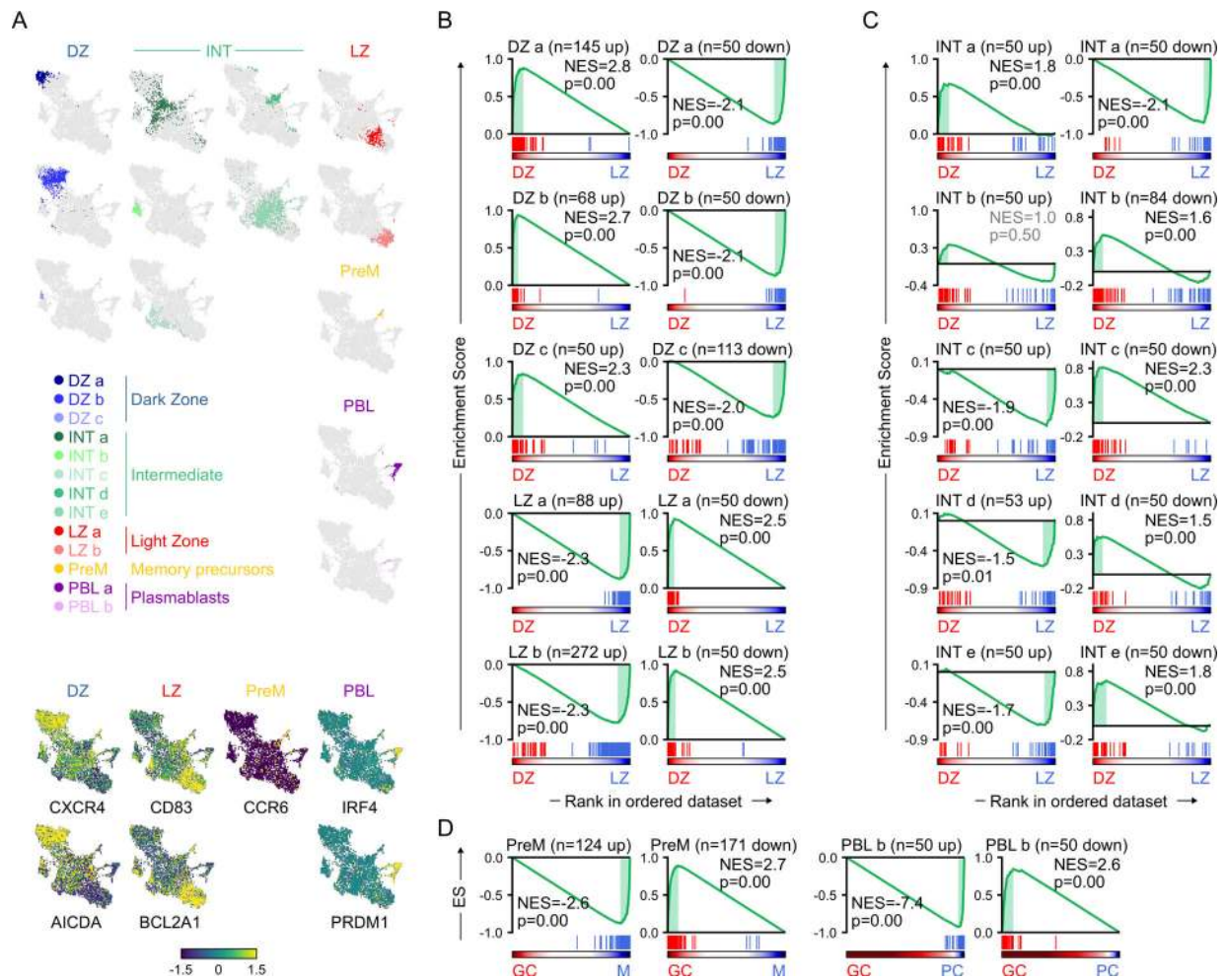


Figure S2. **GC B cell developmental stages.** Related to Fig. 2. **(A)** UMAP projection of the 4,984 GC B cells ( $CD3^+$ ,  $IgD^-$ ,  $CD38^+$ ) in G0-G1 phases of the cell cycle color coded to highlight the individual clusters identified by PhenoGraph (upper panels). The same UMAP projection showing the z-scored gene expression of selected markers (lower panels). **(B)** GSEA of sc-cluster gene signatures of DZ and LZ. **(C)** GSEA of intermediate (INT) subpopulations using the expression profile of bulk purified human DZ versus LZ RNAseq samples. **(D)** GSEA of sc-cluster PreM gene signatures using the expression profile of bulk purified human GC and memory RNAseq samples (left). GSEA of the sc-cluster PBL-b gene signatures using the expression profile of human GC versus tonsillar plasma cells RNAseq samples from Gene Expression Omnibus accession no. GSE114816 (right). The gene signatures include all genes that display a  $\log_2$  fold change  $>1.5$  or at least the top 50 differentially expressed genes. ES, enrichment score; NES, normalized enrichment score.



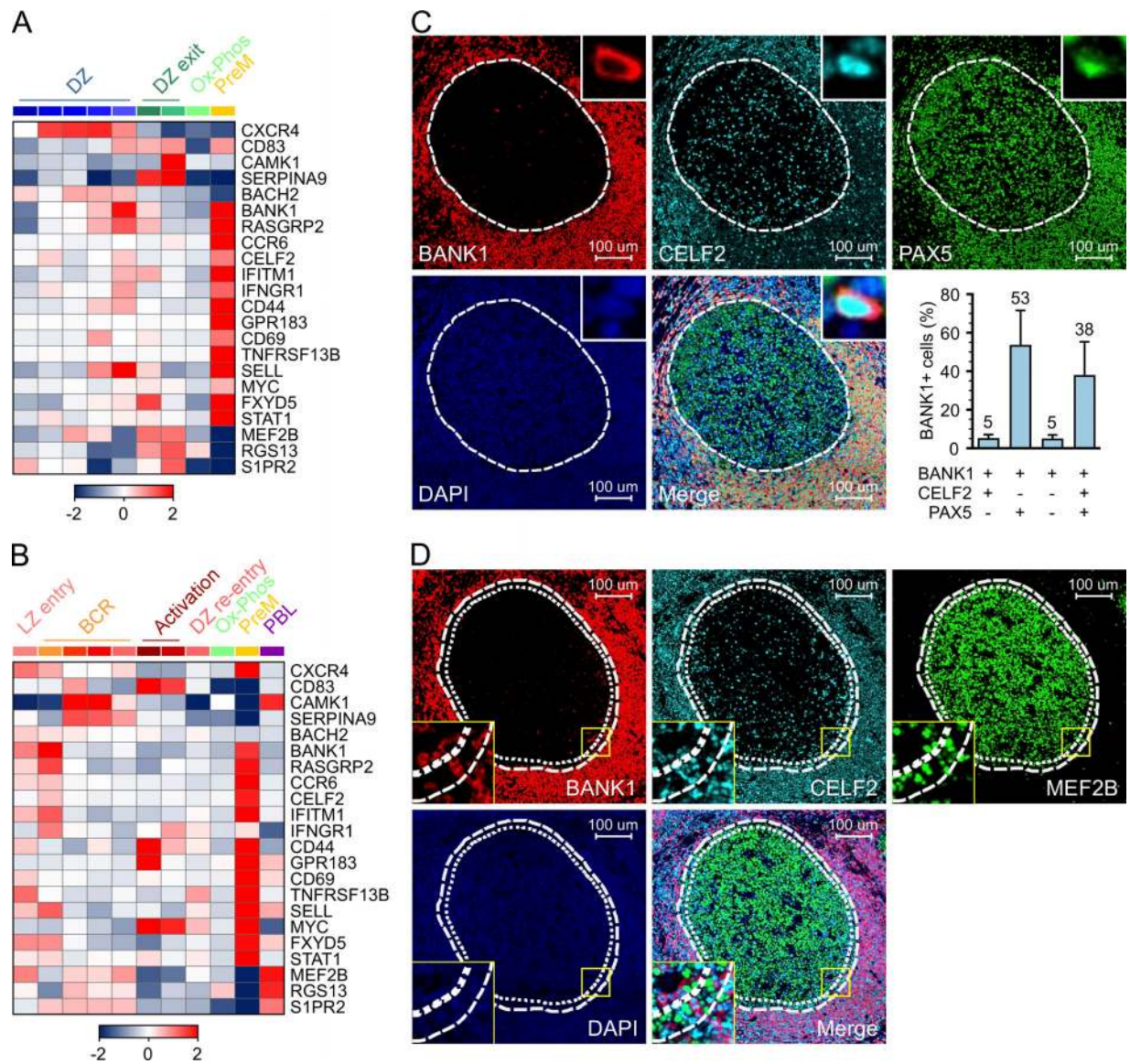


Figure S3. **Memory B cell precursors.** Related to Fig. 5. **(A)** Heat map showing the fold change ( $\log_2$ ) of selected genes in the PreM gene signature in the DZ clusters (cells in G0-G1 phases of the cell cycle). **(B)** Heat map showing the relative expression fold change ( $\log_2$ ) of selected genes in the PreM gene signature in the LZ clusters (cells in G0-G1 phases of the cell cycle). **(C)** Immunofluorescence analysis of BANK1 (red), CELF2 (cyan), and PAX5 (green) in tonsil tissue sections. High magnification at the sc level shows costaining of BANK1/CELF2/PAX5 (insets). Bar plot shows the percentage of BANK1<sup>+</sup> cells coexpressing the indicated markers. The average and standard deviation were calculated from the cell counts of 570 GCs in five donors. **(D)** Immunofluorescence analysis of BANK1 (red), CELF2 (cyan), and MEF2B (green) in tonsil tissue sections. High magnification (insets with yellow border) of the GC border shows the inner part of the GC, marked by the thick dotted line, which has been considered for the analysis. Scale bar = 100  $\mu$ m. Ox-Phos, oxidative phosphorylation.

Downloaded from [http://jupress.org/jem/article-pdf/217/10/e20200483/1046668/jem\\_20200483.pdf](http://jupress.org/jem/article-pdf/217/10/e20200483/1046668/jem_20200483.pdf) by guest on 03 November 2020

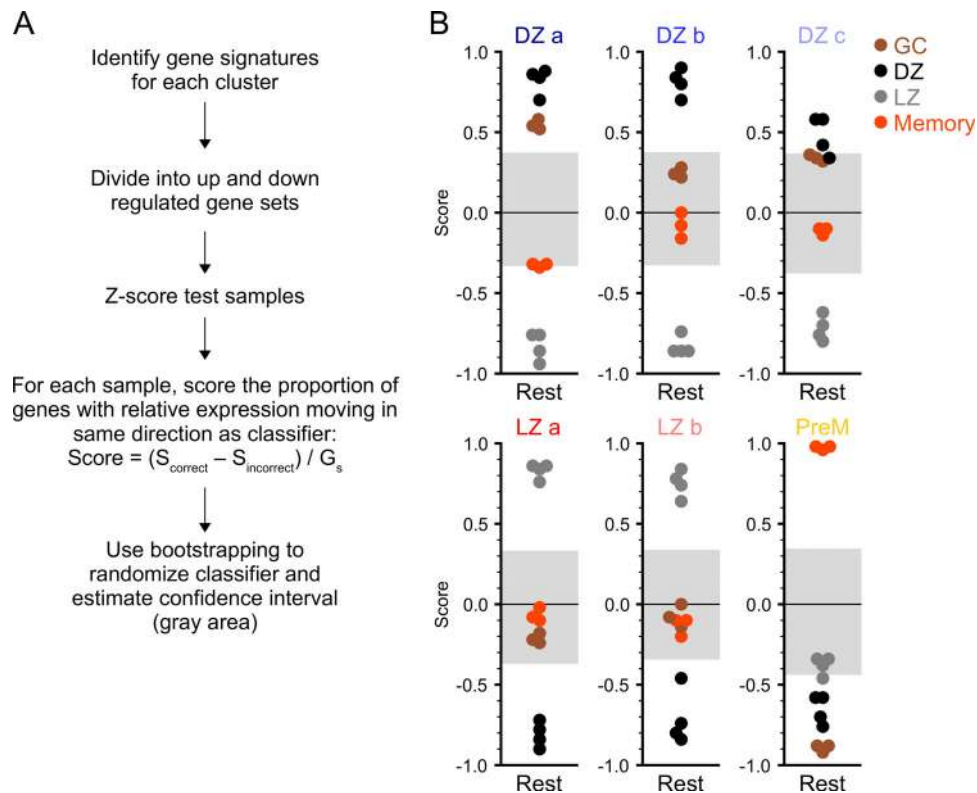
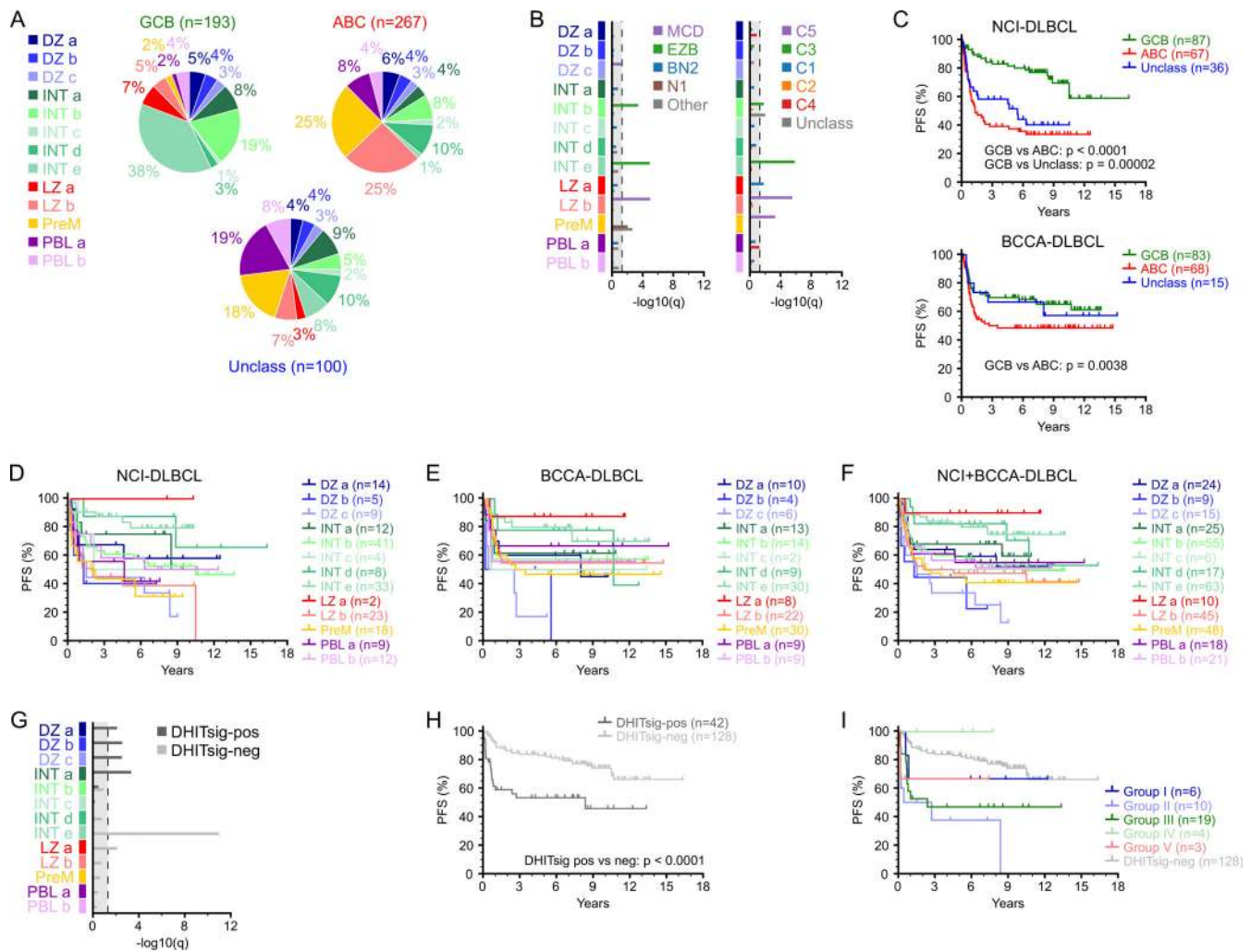


Figure S4. **Development of a classifier based on the sc-RNAseq gene signatures.** Related to Fig. 7. **(A)** Overview of sc-RNAseq classifier algorithm. **(B)** Classification of purified normal B cell populations (GC, DZ, LZ, and Memory) using bulk gene expression profiles and the GC B cell sc-cluster gene signatures. Each signature is identified by comparing a specific sc-cluster with the “Rest,” which refers to all other clusters. A normalized score from -1 (“Rest”) to 1 (cluster of interest), depending on the similarity to each phenotype, is assigned to each sample. The gray area indicates the region that is not statistically significant from a bootstrap test permuting gene in the signatures randomly 1,000 times. Samples with a more extreme absolute score outside the gray area were assigned to either the cluster phenotype (positive score) or the “Rest” (negative score).

Downloaded from [http://jupress.org/jem/article-pdf/217/10/e20200483/1046668/jem\\_20200483.pdf](http://jupress.org/jem/article-pdf/217/10/e20200483/1046668/jem_20200483.pdf) by guest on 03 November 2020



**Figure S5. GC B cell signatures define lymphoma subgroups.** Related to Figs 7, 8, and 9. **(A)** Distribution of the GCB, ABC, and unclassified (Unclass) DLBCL groups from the combined NCI and BCCA datasets within each sc-cluster. **(B)** Enrichment analysis in the GC clusters defined by the sc-signatures of the genetic classes reported in Schmitz et al., 2018 (left panel) or Chapuy et al., 2018 (right panel) within the NCI-DLBCL dataset. **(C)** Kaplan-Meier plot for PFS analysis of the COO subgroups in the NCI-DLBCL (upper panel) and BCCA-DLBCL (lower panel) datasets. P values were obtained using a Mantel-Cox test. **(D-F)** Kaplan-Meier plots for PFS analysis of NCI-DLBCL (D), BCCA-DLBCL (E), and merged NCI+BCCA-DLBCL (F) samples grouped by classifier assignment to one of the 13 sc-COO clusters. **(G)** Enrichment analysis in the sc-COO subgroups of the DHITsig cases (Ennishi et al., 2019) in the GCB NCI-DLBCL and BCCA-DLBCL. **(H)** Kaplan-Meier plot for PFS analysis of the DHITsig-positive (pos) and DHITsig-negative (neg) GCB NCI-DLBCL and BCCA-DLBCL cases. **(I)** DHITsig-positive cases are grouped based on their sc-COO assignment. Groups I to V are defined as follows: Group I (DZ-a); Group II (DZ-b, DZ-c); Group III (INT-a, INT-b, INT-c); Group IV (INT-d, INT-e, LZ-a); and Group V (LZ-b, PreM, PBL-a, PBL-b). EZB, DLBCL with EZH2 mutations and BCL2 translocations; MCD, DLBCL with co-occurrence of MYD88<sup>L265P</sup> and CD79B mutations; BN2, DLBCL with BCL6 fusions and NOTCH2 mutations; N1, DLBCL with NOTCH1 mutations.

Tables S1–S5 are provided as Excel files with multiple tabs. Table S1 (related to Fig. 2) shows cluster cell counts and results of the differential expression analysis in GC B cells in G0–G1. Table S2 (related to Fig. 2) shows the differential expression gene lists obtained from the analysis of GC B cells in G0–G1 and used in the GSEA reported in Fig. S2, B–D. Table S3 (related to Fig. 3) shows cluster cell counts, results from the differential expression analysis in DZ GC B cells in S–G2–M and G0–G1, and their respective pathway enrichment analyses. Table S4 (related to Fig. 4) shows cluster cell counts, results from the differential expression analysis in LZ GC B cells in S–G2–M and G0–G1, and their respective pathway enrichment analyses. Table S5 (related to Fig. 7) shows classification of DLBCL according to the sc-COO classifier, DLBCL identifiers, assignment to the COO subgroups, genetic classes, DHITsig groups, sc-COO classes, score and P value for the top scoring sc-COO class, and sc-COO groups.