## STAR Methods

## KEY RESOURCES TABLES

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| **Antibodies** | | |
| BCL-2 (clone EPR17509) | Fluidigm | Cat# 3146019D; RRID: AB_2811012 |
| BCL-6 (clone K112-91) | Fluidigm | Cat# 3147020D; RRID: AB_2811013 |
| CD134 (polyclonal) | Fluidigm | Cat# 3151024D; RRID: AB_2811014 |
| CD183 (clone G025H7) | Biolegend | Cat# 353733; RRID: AB_2563724 |
| CD194 (clone 205410) | Fluidigm | Cat# 3149003A; RRID: AB_2811015 |
| CD20 (clone H1) | Fluidigm | Cat# 3161029D; RRID: AB_2811016 |
| CD206 (clone 5C11) | Fluidigm | Cat# 3999999-2 |
| CD3 (Polyclonal) | Fluidigm | Cat# 3170019D; RRID: AB_2811048 |
| CD31 (clone C31.3) | Abcam | Cat# ab212709; RRID: AB_2811049 |
| CD34 (clone QBEnd/10) | Abcam | Cat# ab213054; RRID: AB_2811050 |
| CD4 (clone EPR6855) | Fluidigm | Cat# 3156033D; RRID: AB_2811051 |
| CD45RA (clone HI100) | Fluidigm | Cat# 3155011B; RRID: AB_2810246 |

|  |  |  |
|---|---|---|
| CD45RO (clone UCHL1) | Fluidigm | Cat# 3173016D; RRID: AB_2811052 |
| CD68 (clone KP1) | Fluidigm | Cat# 3159035D; RRID: AB_2810859 |
| CD8α | Fluidigm | Cat# 3162034D; RRID: AB_2811053 |
| c-Myc p67 (clone 9E10) | Fluidigm | Cat# 3164025D; RRID: AB_2811054 |
| Ephrin B2 | Abcam | Cat# ab233246; RRID: AB_2811055 |
| FoxP3 (clone 236A/E7) | Thermo Fischer | Cat# 14-4777-82; RRID: AB_467556 |
| Granzyme B (clone EPR20129-217) | Fluidigm | Cat# 3167021D; RRID: AB_2811057 |
| Histone 3 (clone D1H2) | Fluidigm | Cat# 3176023D; RRID: AB_2811058 |
| HLA-DR (clone YE2/36 HLK) | Fluidigm | Cat# 3174023D; RRID: AB_2811059 |
| ICOS (clone D1K2T) | Fluidigm | Cat# 3148021D; RRID: AB_2811060 |
| Ki67 (clone B56) | Fluidigm | Cat# 3168022D; RRID: AB_2811061 |
| LAG-3 (D2G40) | Fluidigm | Cat# 3153028D; RRID: AB_2811062 |
| PD-1 (clone NAT105) | Biolegend | Cat# 367402; RRID: AB_2565782 |
| PD-L1 (clone 28-8) | Abcam | Cat# ab209889; RRID: |

| | | |
|---|---|---|
| | | AB_2811063 |
| PD-L2 (clone 176611) | Fluidigm | Cat# 3172028D; RRID: AB_2811066 |
| pStat3 [Y705] (clone 4/P-Stat3) | Fluidigm | Cat# 3158005A; RRID: AB_2661827 |
| Tbet (clone D6N8B) | Fluidigm | Cat# 3145015D; RRID: AB_2811067 |
| TIM-3 (clone D5D5R) | Fluidigm | Cat# 3154024D; RRID: AB_2811068 |
| Vimentin (clone RV202) | Fluidigm | Cat# 3143029D; RRID: AB_2811069 |
| Vista (clone D1L2G) | Fluidigm | Cat# 3160025D; RRID: AB_2811070 |
| **Metal Isotopes** | | |
| 142Nd | Fluidigm | Cat# 201142A |
| 143Nd | Fluidigm | Cat# 201143A |
| 144Nd | Fluidigm | Cat# 201144A |
| 145Nd | Fluidigm | Cat# 201145A |
| 146Nd | Fluidigm | Cat# 201146A |
| 147Sm | Fluidigm | Cat# 201147A |
| 148Nd | Fluidigm | Cat# 201148A |
| 149Sm | Fluidigm | Cat# 201149A |
| 150Nd | Fluidigm | Cat# 201150A |
| 151Eu | Fluidigm | Cat# 201151A |
| 152Sm | Fluidigm | Cat# 201152A |

| | | |
|---|---|---|
| 153Eu | Fluidigm | Cat# 201153A |
| 154Sm | Fluidigm | Cat# 201154A |
| 155Gd | Fluidigm | Cat# 201155A |
| 156Gd | Fluidigm | Cat# 201156A |
| 158Gd | Fluidigm | Cat# 201158A |
| 159Tb | Fluidigm | Cat# 201159A |
| 160Gd | Fluidigm | Cat# 201160A |
| 161Gd | Fluidigm | Cat# 201161A |
| 162Dy | Fluidigm | Cat# 201162A |
| 163Dy | Fluidigm | Cat# 201163A |
| 164Dy | Fluidigm | Cat# 201164A |
| 166Er | Fluidigm | Cat# 201166A |
| 167Er | Fluidigm | Cat# 201167A |
| 168Er | Fluidigm | Cat# 201168A |
| 169Tm | Fluidigm | Cat# 201169A |
| 170Er | Fluidigm | Cat# 201170A |
| 172Yb | Fluidigm | Cat# 201172A |
| 172Yb | Fluidigm | Cat# 201142A |
| 173Yb | Fluidigm | Cat# 201173A |
| 174Yb | Fluidigm | Cat# 201174A |
| 175Lu | Fluidigm | Cat# 201175A |
| 176Yb | Fluidigm | Cat# 201176A |
| **Biological Samples** | | |

| | | |
|---|---|---|
| SignalSlide® PD-L1 control | Cell Signaling Technology | Cat# 13747 |
| **Chemicals, Peptides and Recombinant proteins** | | |
| m-Xylene anhydrous, >99% | Sigma Aldrich | Cat# 296325-2L |
| Ethanol | Sigma Aldrich | Cat# 459836-2L |
| PBS (Calcium & Magnesium free, pH7.4) | Genesee Scientific | Cat# 25-507 |
| Trizma base | Sigma Aldrich | Cat# T1503-250G |
| EDTA | Sigma Aldrich | Cat# E1644-250G |
| Tween 20 | Fischer Scientific | Cat# BP337-100 |
| BSA (DNase and Protease-free Powder) | Fischer Scientific | Cat# BP9706100 |
| Triton X-100 | Sigma Aldrich | Cat# T8787-100ML |
| Cell ID Intercalator (Iridium) | Fluidigm | Cat# 201192A |
| **Software and Algorithms** | | |
| CellProfiler (version 2.1.0) | | https://cellprofiler.org/previous_releases; RRID:SCR_007358 |
| Ilastik (version 1.3.0) | | https://www.ilastik.org/download.html; RRID:SCR_015246 |
| Rstudio (version 3.6.1) | | http://www.rstudio.com; RRID:SCR_000432 |

## LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for raw data, codes, and other resources should be directed to the Lead Contact, Akil Merchant (akil.merchant@cshs.org).

## SUBJECT DETAILS

This retrospective study included a subset of 33 patients from a previously studied cohort of 85 patients diagnosed with de novo DLBCL at Los Angeles County and University of Southern California (USC) Medical Centers between 2002 and 2012 [1]. The sub cohort was representative of the primary cohort and was not selected other than looking for samples with adequate remaining tissues for further analyses (Figure S1). This study was approved by the USC Health Sciences Institutional Review Board.

## METHODS DETAILS

### Tissue Microarray and Immunohistochemistry

Three of the six tissue microarray (TMA) blocks from the parent cohort with sufficient remaining tissues representative of viable tumor were obtained from the Pathology archive of Los Angeles County and USC Medical Centers. Immunohistochemistry (IHC) staining was performed on 4-μm tissue sections using DAKO ready-to-use antibodies with the EnVision FLEX and FLEX+ visualization systems (DAKO, Glostrup, Denmark) on an automated immunostainer (Autostainer Link 48, DAKO). Detailed IHC staining protocol and scoring methods have previously been published [1]. TMA cores were 2mm in diameter.

### Imaging Mass Cytometry staining

Three of the six TMAs with optimal quality of remaining tumor tissues from the larger cohort study were selected for this study. The TMAs contained 42 cores of FFPE DLBCL tissues from 33 patients and 2 cores from liver tissues. FFPE sections of 4-μm were baked at 60°C for 90 minutes on a hot plate, de-waxed for 20 minutes in xylene and rehydrated in a graded series of alcohol (100%, 95%, 80% and 70%) for 5 minutes each. Heat-induced antigen retrieval was conducted on a hot plate at 95°C in Tris-EDTA buffer at pH 9 for 30 minutes. After blocking with 3% BSA in PBS for 45 minutes, the sections were incubated overnight at 4°C with a cocktail of 32 antibodies tagged with rare lanthanide isotopes obtained from Fluidigm (Table S1). Titration for PD-1 antibody was performed on tonsil tissue (follicular T helper cells in the germinal center). PD-L1 titration was done on a commercial slide containing formalin-fixed paraffin-embedded cell pallets of HDLM-2 (PD-L1+) and PC-3 (PD-L1-) cell lines from Cell Signaling Technology (Key resources table). Validation and titration of all other markers were done on control tonsil tissue (Figure S3). HLA-DR, pSTAT3, PD-L2 and Ephrin-B2 antibodies showed unspecific staining patterns in the control tonsil tissue, however in our DLBCL tissues, Ephrin-B2 and pSTAT3 showed distinct cellular staining and was included in subsequent analyses. HLA-DR and PD-L2 omitted from functional association hypothesis testing.

### Tissue imaging and ablation

All cores were evaluated by two pathologists (I.S. and M.H.) to identify region of interest (ROI) on H&E. Slides were analyzed using the Fluidigm Hyperion Tissue Imager system that couples laser ablation with mass spectrometry [2]. Laser beam of 1-$\mu m^2$ spot size was used to ablate tissue area of 1000-$\mu m^2$ per core at a frequency of 200 Hz. The metal isotopes were simultaneously measured and indexed against the location of each spot to generate intensities and digital spatial maps of the ablated tissues. Detailed descriptions of the ablation techniques have been previously described [3,4].

### Image Analysis Pipeline

The ion counts for each metal-labeled antibody and slide location were compensated for the cross talk between channels then converted to OME-TIFF images [5,6] . Images for each antibody were scaled using the 95 percentiles of the cumulative signal to remove hot spot pixels and normalized across acquisitions [7].

### Image Segmentation

Channels representing distinct morphological features for cell nuclei (i.e. Ir193-DNA Intercalator, Histone H3, foxP3, Ki67) and membrane staining (i.e. CD8, CD68, CD45RA) were used for the Ilastik pixel classification training to predict nuclei, membrane/cytoplasm and background pixel class using cropped 2x scaled images [8,9]. The probability maps were segmented using CellProfiler by subtracting the membrane probability map from the nuclei and then expanding the nuclei by 4 pixels [7].

## QUANTIFICATION AND STATISTICAL ANALYSES

### Data transformation and normalization

The presented data used normalization such as the hyperbolic-arc-sine transformation, and Min/Max normalization at the 99th percentile [10,11] . The authors for Phenograph and t-SNE recommended using the 99th percentile to clip the data to [0,1] scale and remove outliers, which we followed. For the cluster expression estimates, and spatial expression heterogeneity modeling, we used the mixed-effects linear model with the single-cell marker expression intensities scaled to a standard normal distribution across each ROI. Thus, for linear modeling we ensure that for all tissues, each marker expression was on the standardized Gaussian distribution, and the marker expression was compared to the grand average using a generalized linear hypothesis test. For Cox proportional hazards estimates, the relative proportions of each cluster/sub-phenotype were used as features, with survival times (N=30, events=7), and for the pSTAT3/Ki67 the corresponding phenotype expression values per patient multiplied the corresponding patient phenotype proportion.

### Analysis workflow

The exploratory analysis used histoCAT , and downstream tSNE clustering using ‘Rtsne (v.0.15)’,‘lme4 (v.1.1.21)’ R (v.3.6.3) packages for clustering and mixed-effects linear models [12]. The neighborhood analysis used Bodemiller repository (https://github.com/BodenmillerGroup/neighbouRhood).
Codes for the analysis of this project are available here:
https://github.com/arcolombo/singleCell_DLBCL

### Clustering and metaclustering

### Phenotypic clustering

The images along with the masks were imported into histoCAT software for initial evaluation. Cell features were extracted and imported into R statistical software environment [12]. We hierarchically performed meta-clustering to identify cell “phenotypes”. The first step under-clustered the data using lineage related markers (BCL2, BCL6, CD20, CD206, CD3, CD30, CD31, CD4, CD45RA/RO, CD68, CD8, EphrinB2, FOXP3 and HLADR) clustering each ROI (nearest neighbor, $k_1$=45) and then clustering on the centroids (nearest neighbor, $k_2$=15) [10]. The major cell classification per case identified 14 meta-clusters, and each ROI had on average 1,145.174 cells per meta-cluster. Quality control analysis performed Phenograph (k=50) on each major cell component separately to ensure that the major cell expression was homogeneous on the corresponding marker. [Sup Figure S5, S6] Re-assignments of CD4 comprised of 1.54%, the initial tumor component reassigned 6.21% to CD8, CD4, MAC, and endothelial classes [ Sup Figure S5-S6]. Following the re-assignments, we identified minor clusters for each major cell component by re-performing meta-clustering ($k_1$=45, $k_2$=15) on each TME component separately, with the inclusion of inducible state markers (cMYC, CCR4, CXCR3, Granzyme, ICOS, Ki67, LAG-3, PD-1, PD-L1, PD-L2, TBET, TIM-3, Vimentin, VISTA, and pSTAT3) and morphological features (Area, eccentricity, solidity, perimeter, percent touching, number of neighbors).

## Assignment of phenotype sub-clusters to putative functional groups

Each phenotypic cluster expression profile corresponding to a Z-scaled marker intensity profile was modeled using a linear mixed-effects ('lme4 v. 1.1-27.1') model and the model estimates were contrasted against the overall cluster expression average using generalized hypothesis testing ('multcomp v. 1.4-17' across the 41 clusters), and the resultant differences from the grand average was used to group clusters into functional groups.

We grouped sub-clusters identified by phenograph with similar marker expression into functional groups and assigned labels based on expert review as follows: Tumor sub-clusters (6,8,10) that expressed high levels of chemokine receptors such as CXCR3 (p=1.9e-10, 2.2e-04, 0.039 respectively), CCR4 (p=7.7e-05, 8.6e-03, 0.089) and expression of PD-L1 (p=8.6e-14, 6.6e-06, 1.7e-07) were labeled as "inflammatory" while the remaining tumor sub-clusters (1-5,7,9) were "non-inflammatory."

CD4 T cells were grouped as baseline (4) b on below average expressed PD1 (p=0.018), and CXCR3 (p=1.2e-10), whereas CD4 cells (2, 3, 7) were labeled as activated/early exhaustion due to moderately expressed markers PD-1, and TIM-3. Late exhausted/inflammatory CD4 sub-types (1,5) were identified through highly expressed PD-1 (p=1.0e-05, 1.0e-06), TIM-3 (p=2.5e-06, 1.0e-06), LAG-3 (p=5.3e-08, <1.0e-05) and increased CXCR3 (p=0.80, 1.3e-07). Activated/proliferative CD4 cells (6) were identified through over-expression of Ki67 (p<1e-03). Additionally, activated/proliferative CD8 cells (1,4) were identified through highly expressed Ki67 (p<1e-03), whereas CD8 cells (2,6) with highly expressed CXCR3 (p=6.4e-03, 0.06), and PD-1 (p=1.7e-10, 1.0e-07) were subsequently labeled exhausted/inflammatory CD8 T cells. CD8 cells (3,5) with low CXCR3 expression, but highly expressed PD-1 (p=9.1e-03, 1.2-e04), LAG-3 (p=1.3e-15, 2.0e-15), and TIM-3 (p=0.11, 9.2e-06) were grouped as terminally exhausted.

$T_{REG}$ subtype (5) had uniquely high Ki67 (p<1e-03) were denoted as activated/proliferative. $T_{REG}$ cells (1,2,6) with above average/moderate co-expression of

CCR4(0.9, 1.3e-05, 0.42), PD-1(0.96, 5.6e-5, 0.4), TIM-3 (1), and LAG-3 (p=0.96, 1.8e-09, 0.9) were annotated as highly suppressive[13,14.] TREG cluster 1 estimates contrasted against the population average for TIM-3(estimate=0.07 (-0.08, 0.23); p=1), TREG cluster 2 differences from the population means TIM-3(estimate=0.31 (0.11, 0.0.49); p=0.069), and TREG cluster 6 estimated differences for TIM-3(estimate=0.61 (0.29, 0.92); p=5.5e-03). Whereas TREG cluster 3 and 4 estimates for TIM-3 were consistently below average albeit not significantly (-0.21 (-0.41,-0.01); p=0.82) and (-0.26 (-0.46, -0.05); p=0.44) respectively. Similarly, TREG cluster 1, cluster 2 and cluster 6 estimates for PD-1 was 0.08 (-0.007, 0.17; p=0.96), 0.26 (0.16, 0.37; p=5.6e-05) and 0.22 (0.05,0.40; p=0.38) respectively. The PD-1 cluster expression for TREG clusters 3 and 4 were 0.055 (-0.06, 0.17; p=1), and 0.14 (0.02, 0.25; p=0.58) respectively. Although the PD-1 profile was similar, important differences in LAG-3 were observed. TREG clusters 1, 2, and 6 had estimated mean differences 0.12 (-0.01, 0.25; p=0.97), 0.52 (0.37, 0.68; p=1.8e-09) and -0.0019 (-0.26, 0.25; p=1) which were generally average, or above the average LAG-3 expressions; importantly 3 and 4 were below the average -0.13 (-0.29, 0.03; p=0.99) and -0.06(-0.22, 0.11; p=1). CCR4 expression profile was another important contrast separating the ‘highly suppressive’ TREG family from ‘baseline‘ family such that TREG clusters 1 (0.06(-0.09, 0.21); p=1), 2(0.48(0.29, 0.66); p=1.29e-05) and 6(0.38(0.08, 0.69); p=0.42) all with generally moderate/above average estimated differences from the grand population mean, whereas clusters 3(-0.21(-0.41, -0.02); p=0.77) & 4(-0.1(-0.29, 0.097); p=1) were below it albeit non-significantly. Clusters 1, 2 and 6 had moderate/above average expression of TIM-3, CCR4, and LAG-3 were more similar compared to clusters 3 and 4. TREG cluster 1, had moderate CCR4, whereas clusters 3 & 4 were generally below average, additionally TREG 1 had approximately average LAG-3 expression whereas 3&4 were below the average; TREG 1 cluster had moderate average expression of TIM-3 (p=1), however the clusters 3 & 4 were clearly below the average. The significance (p-values) was not the sole determining factor for rarer sub-populations because the Bayesian model will penalize imprecise populations due to their smaller sample sizes, and place more weight onto the larger, robust, clusters. The functional annotation was reviewed for their estimated trends, and their similarity amongst the TREG family. Regulatory T cells (p=0.78, 0.9), TIM-3 (p= 0.8, 0.4), and LAG-3 (p=0.99, p=1) were grouped as baseline (3,4).

Macrophage cells with high CD206 (p= 1.0e-07, 3.01e-03, 4.9e-03) were classified as M2-MAC (1,5,6), and significant co-expression with PD-L1 (p<1.0e-04) were labeled PD-L1+M2-MAC (2), otherwise M1 labeling was used. Conversely, endothelial cells (4) with low TIM-3 (estimate= -0.24(-0.46,-0.02); p=0.78) expression and above average PD-L1 expression

(0.18 (0.03, 0.32); p=0.53) were denoted as PD-L1+ endothelial cells. Endothelial is a smaller sub-group, and had clear differences in its TIM-3/PD-L1 profile to justify its own annotation. Conversely, endothelial cells (4) with low TIM-3 expression and above average PD-L1 expression (p=0.53) were denoted as PD-L1+ endothelial cells. Lastly, significantly over-expressed endothelial cells (1) with high EphrinB2 (p=3.8e-04) were denoted as EphrinB2+, otherwise baseline (5) cells were used.

### Estimation of driver proteins on sub-cluster

We avoided calling a sub-cluster immune subset "positive" or "negative" by arbitrary thresholding but used a linear mixed-effects model to derive effect estimates of markers across all sub-clusters. We constructed the linear model using the ROI standardized single-cell expression by mapping all cells per ROI to a standard normal distribution. Hence all ROIs had standardized marker intensity, and for generalizability all estimates excluded non-treated subjects or subjects lost-to-follow-up. We provided point estimates using complete data only. The point estimates and 95% confidence intervals on key markers were derived using ('lme4 (v.1.1.21)', and 'multcomp (v.1.4.12)' R packages). The point estimates for "driver", or protein enrichments, on sub-clusters were performed using mixed-effects linear model which treated the case as a random effect, and the sub-cluster mean expression per ROI as the features. We created custom contrasts for all sub-cluster average expression and tested for enrichment by comparing to the global mean for that protein in a single model. The significance of a particular sub-cluster is interpreted as significantly higher/lower to the grand average of that protein (p<0.05), which determined "+/-" for that cluster. The confidence intervals were generated using general linear hypothesis test ('glht') function from the R package 'multcomp (v.1.4.12)' and derived from ROI standardized values which we hope increases reproducibility in this field. The R package 'lme4' (version 1.1.23) was used for mixed-effects model.

### TMA and replicate divergence analysis

The cohort comprised of 3 TMAs, and principal component analysis (PCA) was used to determine the presence of batch effects across TMAs. After annotating the 14 meta-clusters into major cell components, we performed PCA on the relative proportions per ROI and the PCA showed well-mixed visual representation of the variability of the data that did not have any TMA specific grouping. Case 17, 18, 21, 26, 27 and 31 had replicate ROIs taken and case 30 had triplicate ROIs. Using the R package 'entropy (v.1.2.1)', we computed the Kullback-Leibler (KL) divergence using a given replicate ROI relative frequencies per major cell component and compared a given replicate to the case relative average of the major components. The KL divergence scores per ROI were all below 0.10, however Case 26 had increased ROI heterogeneity (KL divergence of 0.26). The PCA analysis of the replicates showed minimal distance separations with the exception of Case 26. The replicates were mostly similar, with the exception of Case 26, and for the analysis of patient clinical variables we used the case averages across the ROIs.

### Association between genetic mutations, cell of origin and proportions of tumor-immune sub-phenotypes

The cluster analysis followed the guidelines from the authors, which used the 99th percentile normalization to remove outliers by scaling to relative 99th percentile of each marker. The heatmap visualization standardized each ROI to a standard normal distribution. The heatmap depicts the mean normalized intensities, which ensures that the tissues were standardized. The clustering of all the phenotypes present used bootstrapping (100) using ‘pvclust (v.2.2.0)‘ and hierarchically clustered using sub-cluster means (Euclidean distance, Ward’s method) ’dendsort (v.0.3.3)’ R package[15,16].

### Clinical association of sub-clusters to COO and mutation

In order to test for associations with clinical parameters, we used ‘diffcyt’ v.1.14.0 with a design matrix corresponding to Chapuy signatures, IPI, REF, COO, gender, double expressor, and patient replicates[17]. Chapuy signature contrasts used a “one-vs-all" linear contrast, which test a given Chapuy signature contrasted against all other signatures. Differential abundances and states used the edgeR/ limma-voom procedures with Benjamini-Hochberg multiple test corrections ($q<0.05$).

The international prognostic index scores [0-5] were measured, and a median cutoff (>3) was used to identify patients with high IPI scores.

BCL2 and MYC protein were measured by immune-histochemistry. BCL2 overexpression was measured using a 40% cutoff judged by IHC, and MYC over expression was determined as 70% threshold. Patients classified as double expressors were identified as BCL2 above the 40% and MYC above the 70% threshold.

### Phenotype association with mutation signatures

We obtained the molecular variant/mutations of a limited gene panel (Cancer genetics Inc.) in 22 subjects in this cohort and focused the mutation model on influential genes previously reported in lymphoma.[18] We used the ordered list of somatic copy number alterations (SCNA) reported by Chapuy et.al which identified the five coordinate DLBCL genetic signature clusters C1 (***BCL6***, ***BCL10***, ***TNFAIP3***, *UBE2A*, ***CD70***, ***B2M***, ***NOTCH2***, ***TMEM30A***, ***FAS***, *TP63, ZEB2, HLAB, SPEN, PDL1*) of which our targeted panel overlapped with 8 top ranked mutations in the C1 signature. Similarly, C2 (***TP53***), C3 (***BCL2***, ***CREBBP***, ***EZH2***, ***KMT2D***, ***TNFRSF14***, *HVCN1*, ***IRF8***, ***GNA13***, ***MEF2B***, ***PTEN***), of which the targeted panel overlapped with 9 of the 10 C3 signatures. C2 contained only 1 SCNA of which our panel contained. Our panel overlapped with C4 (***SGK1***, ***HIST1H1E***, *NFKBIE*, ***BRAF***, *CD83*, ***NFKBIA***, ***CD58***, ***HIST1H2BC***, ***STAT3***, ***HIST1H1C***, *ZFP36L1*, ***KLHL6***, *HIST1H1D*, ***HIST1H1B***, ***ETS1***, *TOX, HIST1H2AM, HIST1H2BK*, ***RHOA***, ***ACTB***, *LTB*, ***SF3B1***, ***CARD11***, *HIST1H2AC*), 15 of the top ranked C4 genetic signature. The C5 (***CD79B***, ***MYD88***, *ETV6*, ***PIM1***, ***TBLXR1***, *GRHPR, ZC3H12A, HLAA*, ***PRDM1***, ***BTG1***) genetic signature contained 6 top ranking genetic mutations.

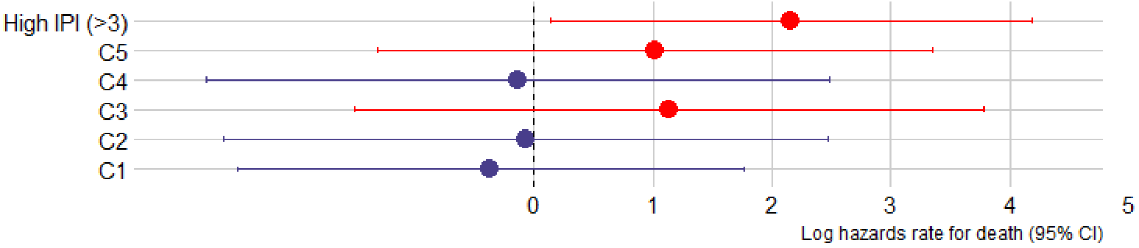
For each participant, we counted the total number of SCNA identified within a molecular signature class. If a participant had at least one mutation within a mutation signature list (Chapuy), then the molecular signature indicator variable was set to 1, otherwise that participant was set to 0 for that class. For hypothesis testing of differential abundances/states, the effects of a molecular signature was modeled as an independent explanatory variable but the hypothesis test used a linear contrast against the other classes defined as

$$\widehat{\beta_{C_1}} = \beta_{C_i} - \sum_{i \neq j}^{5} \frac{\beta_{C_j}}{4}, i = 1,2,3,4,5$$

The contrast identified molecular signature *enrichment* defined as the effect difference of a given molecular signature from the average effect of the other signature effects. Figure 2 reported the molecular signature enrichments (BH q<0.05) for C1-C5.

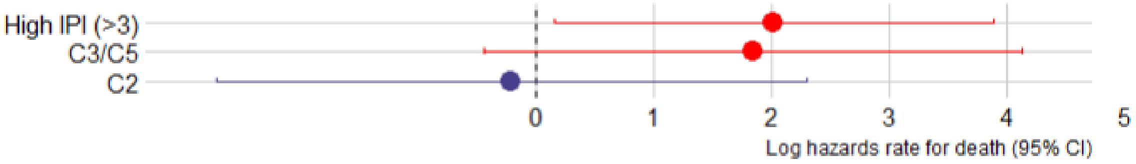
## Molecular assignment analysis

Each participant was assigned to a Chapuy molecular signature group based on the overlapping gene mutation between our targeted sequencing panel, and the Chapuy mutational list. Since not all participants had targeted sequencing performed they remained unassigned. In order to examine the concordance between this assignment and the original Chapuy signature paper, we performed a multiple Cox proportional hazards model of each signature (C1-C5) as an explanatory (indicator) variable, adjusted for high IPI (>3) as an additional indicator variable. The data in our cohort suggest that signatures comparing those in signatures C3 and C5 to those not in signature C3 or C5 or otherwise unassigned, the log-hazards rate for the event of death increased 1.14 and 1.02 times respectively (see below table). Comparing those in C1 and C4 to those not in C1 or C4 or otherwise unassigned, there was a protective trend such that the log-hazards rate marginally decreased 0.37 and 0.13 respectively. Participants in C2 had approximately a log-hazards rate very close to the null value estimated at –0.068, suggesting that the hazards rate for the event of death was proportional between those in C2 and those not assigned to C2. C2 was not concordant to the manuscript, which is likely due to using only SCNA for cluster assignment, which is not sufficient for this cluster.

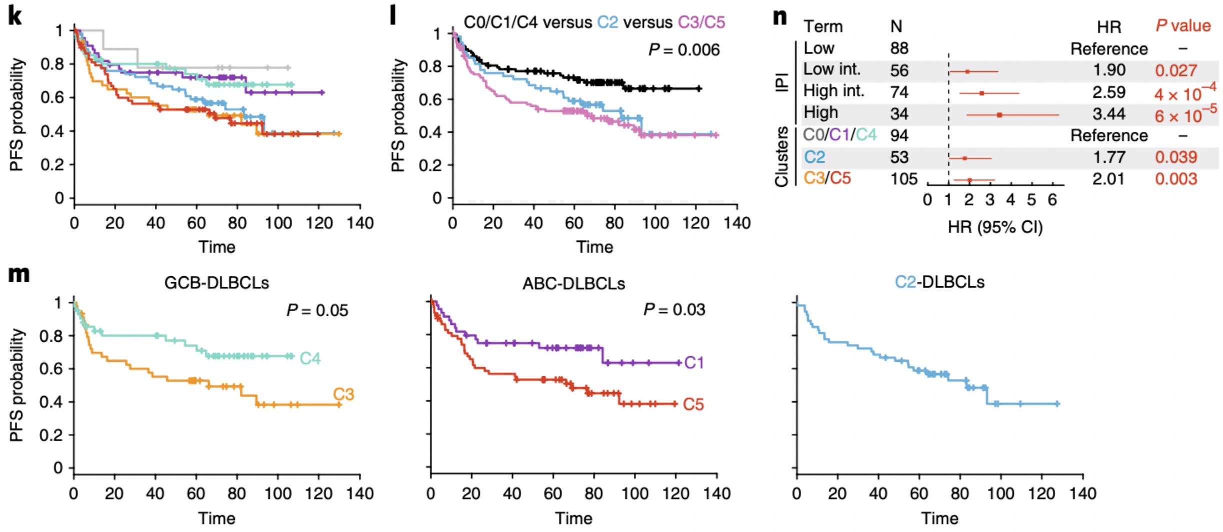


| Signature | Hazards (log) | 95% Confidence | p-value |
|---|---|---|---|
| C1 | -0.37 | (-2.49,1.76) | 0.73 |
| C2 | -0.068 | (-2.60,2.47) | 0.96 |
| C3 | 1.14 | (-1.51,3.78) | 0.39 |
| C4 | -0.13 | (-2.75,2.48) | 0.92 |
| C5 | 1.02 | (-1.31,3.34) | 0.39 |
| High IPI (>3) | 2.16 | (0.13,4.18) | 0.04 |

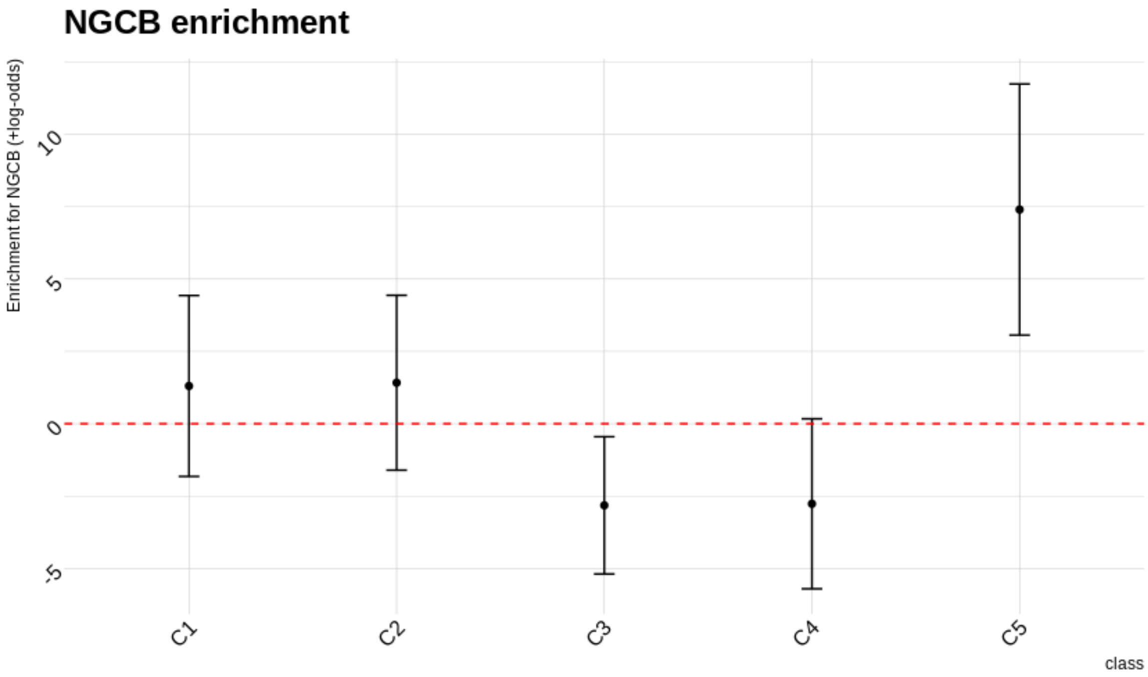
In order to best replicate the work by Chapuy et. Al, we fit a Cox hazards model using C1 and C4 signature as the reference, and we pooled C3 and C5 into a new grouping titled ‘C3/C5’ that indicates if a participant is either C3 or in C5, with 0 otherwise. Comparing C3/C5 to those in C1/C4 (reference category) the log-hazards rate was 1.85 (-0.43,4.12; p=0.11) times increased suggesting a concordant trend with the Chapuy reported model. Comparing those in C2 to those in C1/C4 (reference) there was a marginally decreased trend of –0.20 (-2.271,2.29; p=0.87) which was not concordant to the original model. The discordant hazard estimates in our model is likely due to the nature of C2 defined as 1 SNP on *TP53*, whereas C2 also includes numerous structural variants, of which our model is limited to a single *TP53* variant call. This suggests that this cohort does recapitulate the survival trends for C1, C3, C4 and C5, but C2 is dis-concordant. Below we include the log-hazards estimates in our cohort based on this patient assignment, and the original hazards estimate reported by Chapuy, respectively.



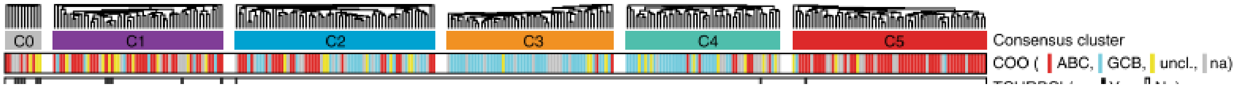
Key figure from Chapuy et.al manuscript. [19]



The logistic model of the molecular subtype assignment suggested that C5 had significantly increased log-odds for NGCB, which was aligned with the C5 definition, and C3, C4 had increased log-odds with GCB sub-types. C1 and C2 showed mixed odds with roughly equal likelihood for NGCB/GCB which was concordant with these subtype definitions. Below we show the log-odds for NGCB in our cohort, and show the COO distribution across molecular subtypes in the original manuscript.



Key figure from Chapuy et.al manuscript. [19]



### Multivariate clinical associations between sub-clusters, topological clusters using multivariate logistic regression

From Figure 2a, we applied a linear model using ‘diffcyt’.[17] The abundance associations with the clinical parameters used an ‘edgeR’ model, and the differential states were tested using an empirical Bayesian linear model (‘limma’ version 3.44.3). We applied the linear model and used Benjamini-Hochberg adjustment method (q threshold of 0.05).

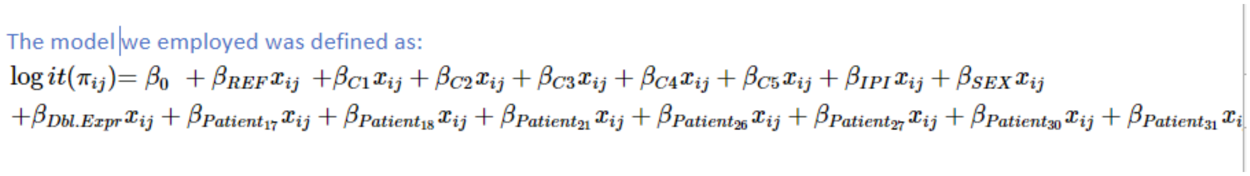
For the Chapuy signature enrichment effects (BH q<0.05) we report all the differences identified from the model after examining the boxplots of their abundances to visually inspect that the proportions of a given cluster were not driven by outliers.

In order to model ‘single cell data vs cohort/clinical (mutational and clinical)’ data we employed a linear model to detect differential cell population abundances shown in Figure 2a. In order to identify differences in proportions, we logit transformed the proportions and used them as a dependent variable in a linear model. The binomial distribution is an appropriate choice which accounts for the uncertainty of proportion estimates, and hence the logistic regression was used which inputs the cell counts as a relative proportion to the total sample count as input into the model.

We detected differential cell population abundances using the `diffcyt` package which tests for proportion differences of cell types across experimental conditions. [17] Given the cell counts for each sample relative to the total sample, the generalized linear model (GLM) can model differential abundance. Ideally, the generalized linear mixed model (GLMM) could be used to account for the overdispersion, by using a random intercept for each participant ID, however due to our limited sample size, we obtained convergence warning messages from the software, and we chose the fixed effect repeated measures model as the next best option.
The model relating cell type abundance to experimental conditions we assumed that for a given cell population, the cell counts Y_ij was modeled as a binomial distribution defined as

$$Y_{ij} \sim \text{Bin}(m_{ij},\ \pi_{ij})\text{Yij} \sim \text{Binmij}, \pi\text{ij}$$

where m_ij is the total number of cells in a given sample corresponding to participant i and condition j. Given an experimental design matrix, the x_ij is defined from the design matrix corresponding to sample i and condition j.



The model we employed was defined as:

$$\log it(\pi_{ij}) = \beta_0 \ + \beta_{REF} x_{ij} \ + \beta_{C1} x_{ij} + \beta_{C2} x_{ij} + \beta_{C3} x_{ij} + \beta_{C4} x_{ij} + \beta_{C5} x_{ij} + \beta_{IPI} x_{ij} + \beta_{SEX} x_{ij}$$
$$+ \beta_{Dbl.Expr} x_{ij} + \beta_{Patient_{17}} x_{ij} + \beta_{Patient_{18}} x_{ij} + \beta_{Patient_{21}} x_{ij} + \beta_{Patient_{26}} x_{ij} + \beta_{Patient_{27}} x_{ij} + \beta_{Patient_{30}} x_{ij} + \beta_{Patient_{31}} x_{i}$$

The linear model was defined by expert review of the clinical parameters including treatment response/refractory (REF/CR), molecular signatures (C1-C5), high IPI (>3), gender (M/F), double expressor (c-MYC pathological categorization (3-4) and high BCL2), and repeated measures corresponding to replicate samples per patient accounted as fixed effects.
The significance testing and multiple testing corrections for each phenotype cluster was performed via ‘edgeR’. This package contains functions utilizing the normalization of total cell counts per sample (library sizes) and total number of cells for a given phenotype sub-population scaled using the trimmed mean of M-values' (TMM) method which minimizes sample specific false positives. Sub-population differential testing required a minimum of 50 cells for a given cluster, required to be present in at least 7 samples.

The significant testing correction utilized was Benjamini-Hochberg (BH) corrections with an alpha level of 0.05. In summary, differential testing of phenotype sub-population abundances can model the association of single cell population relative counts with the clinically relative parameters relating to DLBCL, after accounting for IPI, sex, double expressor participant status, repeated measures, and multiple test correction procedures.

Using ‘diffcyt’, the differential analysis of markers expression utilized the linear model, where the marker expression (arcsinh-transformed, cofactor 5) was assumed to follow a Gaussian distribution. We used the same model formula previously described, to summarize the median protein marker cellular intensity in each phenotype population per sample.[17] The median intensity was assumed to follow a Gaussian distribution, and used in a linear model. The multiple test corrections and fitted model results was performed using ‘limma/voom’ over the previous design matrix. Threshold of 50 cells per a given phenotype population in at least 7 samples were required in order to identified stable intensity medians per cluster (BH q-value <0.05).

In summary, modeling differential abundance/states at the cellular level can be associated with an experimental design for hypothesis testing.

After fitting the model at the cluster level and identifying significant associations with the molecular subtypes, we examined the boxplots of the abundances to ensure that the differential clusters were not driven by influential samples/points; in our results Tumor 9 was significantly negatively associated with C4, but was driven by a single sample influencing the association, hence this result was ignored, the other cluster results had overall reasonable distributions.

### Cellular manifold approximation and projection (UMAP).

The Barnes-Hut t-stochastic neighborhood embedding (tSNE) was constructed using (‘Rtsne (v.0.15)’ Package), and UMAP (umap v0.2.2.1) were used. After performing the UMAP using all single-cells, manual annotations of the sub-clusters were performed by visual inspection. The tSNE parameters were initial dimensions: 50, perplexity: 30, theta: 15.

### Spatial classification and topology of DLBCL

The algorithm was motivated by Yuan et. Al study of infiltrating lymphocytes in breast cancer using H&E whole tumor slides [20]. Yuan defined infiltrating lymphocytes by using the centroid (the average of 5 nearest neighbors (NN) Euclidean distance to cancer cell) which identified a lymphocyte that resided inside a convex hull (domain) of 5 neighboring tumor cells. The 5-NN centroid provided additional information compared to the first nearest neighbor because if a T-cell had shorter distance to the centroid, compared to the 1-NN, this implied infiltration because the lymphocyte was in closer inside the tumor domain. We then sought to develop an algorithm which would classify the tumors by their proximity (5-NN centroids) to nearest immune cells and create a linear ordering by distance. Importantly, the tumor ordering in the context to TME proximity represented a *contour immunographic map*, where furthest distance to the nearest immune cell was analogous to low immune infiltration potential ("steep valley"), whereas tumors that were immediately co-localized to immune cells would have increased immune potential ("top-of-hill").

In order to develop the algorithm, we generated 4 synthetic point patterns using 'spatstat (v.1.59.0)' R package (Figure S9-S10). The synthetic image 1 and 2, had three pattern types, which simulated a germinal center (pattern 1), and the T-cell zone (pattern 2), we included a 'null' type (pattern 0) which was included in the distance algorithm, but a shape was left unassigned. The synthetic image 3 had four pattern types. Whereas synthetic image 4 had only two pattern types which demonstrated how the algorithm would order a pattern in terms of distance to the interface (pattern B). We further tested the distance classification algorithm on 6 reactive lymph nodes (Figure S10). From the raw ablation images (Figure S10C) we observed that the CD20 follicle, which captured light/dark zones, was subset (Figure S10A) into sub-types corresponding with nearest distance to PD-1+TFHs and other T-cells in the paracortex region.

The algorithm simply computed the average distance to the 5-NN (centroids) from each B-cell toward the other immune phenotype, and then used Phenograph to meta-cluster ($k_1$=45, $k_2$=15), the distances into classes which were then ordered[21]. The distance centroids for each tumor cell were used with Phenograph algorithm to classify the tumor cells into clusters dependent on their centroid distance to the immune cells, which provided an immune contour. We observed that the centroid distances were linear to 1-NN neighbors.

The spatial analysis (Figure 4) of the topological clusters (tumor a, b,c, … ,l) included all cases but excluded case 20 because this case had approximately 97% immune cells and less than 3% tumor cells relative proportions.

This case 20 was excluded because it did not contain enough tumor cells to cluster and model appropriately.

### Spatial organization of tumor topology in DLBCL comparing COO and reactive lymph node

The Clark-Evans aggregation index is used to measure spatial organization of a point pattern and was performed using the 'spatstat (v.1.59.0)' R package. We compared the Clark-Evans standardized indices at the ROI level in DLBCL and 6 lymph nodes and compared them using a Student's t-test (Tukey test for multiple comparisons). We did not include tumor_h nor tumor_e into the spatial organization model, because their abundances were highly associated with GCB ($p=0.057$) and NGBC ($p=0.026$). Tumor classes that were 10% prevalent were selected, and rare topography classes (tumor_a (0.93%), tumor_b (6.9%)) were excluded. By visual inspection, classes: "c" (15.2%), "d" (15.4%), "f" (21.1%), "g" (11%), and "i" (14.9%) were co-localized, however statistical significance was achieved after dropping tumor_d from the spatial organization model, however the trends were still observed.

### Multivariate logistic regression of topology class abundances in DLBCL

The multivariate logistic regression was performed on the topological relative case tumor proportions (%) that included the IPI scores. For the log-odds estimates in association with COO, NGCB/GCB was the response variable set to 1/0, and all the tumor topologies were fit into the multivariate model including IPI. For Chapuy log-odds estimation, for each C1-C5, the multivariate model included the case proportions of each tumor topology cluster, and IPI. In this multivariate model, without any subsetting/nested modeling, we reported any tumor topology in this multivariate model that had $p<0.05$.

### Neighborhood analysis

The tumor-centric neighborhood analysis depicted in Figure 4e/4f, Figure 5d, Figure S11a, S11b, S15d we used the algorithm deposited by Bodenmiller group [(https://github.com/BodenmillerGroup'neighbouRhood)], and significance was determined using 1,000 permutations and interaction distance of 15µm. Significant interactions used $p<0.01$ to determine a significant signed interaction. For Figure 2A, the neighborhood pairwise interactions were performed for all pairs, and the significant signed interactions were depicted as a heatmap along with the corresponding mean normalized intensity, association with clinical features, and mutations. Alternatively, Figure 4e-4f, summarized all the significant interactions from the tumor regions as the first label, and the major TME components as the second label. The cohort average signed interactions were computed in 4f (with 95% confidence interval). Figure 4g computed the significant interactions with between the first label as tumor_d and second label as the CD4 sub-types, and average the significant interactions ($p<0.01$) with the first label as CD4 phenotypes and second label as tumor_d. We average the total signed interactions ($p<0.01$) for both directions of spatial attractions.

Figure 5d used tertile gating separately for CCR4/PD1 on TREG in each cohort separately. The spatial interactions (with the previously described parameters) were computed in each cohort using the first label as the CCR4+PD1+ TREG and the second label as the PD1+/mid/- T cells. The total signed attractions/repulsions ($p<0.01$) were summed for each cohort, and divided by the total samples in each cohort separately. The interaction summary values ranged from [-2, +2] and the ANOVA was compared (BH $q<0.05$) to test for the frequency differences between the cohorts. Note that the gating and spatial attraction signed interactions were computed in a stratified manner.

For the neighborhood enrichment comparisons in DLBCL, we performed neighborhood analysis on the full cohort and identified significant interactions using 1,000 permutations which generated the null distribution and a significance threshold of 0.01. The comparison of neighborhoods between phenotypes used an ANOVA by summation of the total signed interactions which approximately ranged from [-5, +5]. The ANOVA was fit, residuals were examined and Benjamini-Hochberg corrections were used.

The topological association test for immune active vs. immune suppressive phenotypes present used the Fisher's Exact Test created a 2x2 contingency table for the total signed interactions relating to significant ($p<0.01$) suppressive/activated phenotypes in the mantle zone, and not in the mantle zone. Similar 2x2 table was created for the dispersed regions. Using 1,000 permutations with an alpha-level of 0.01, the suppressive phenotypes that was significantly attracted (using tumor-centric reference) included late exhausted/inflammatory CD4 (1,5), exhausted/inflam. CD8(2,6), terminally exhausted CD8(3,5), M2-MAC(1,5,6), PD-L1+M2-MAC(2), highly suppressive T-reg, EphrinB2+ endothelial(1), and PD-L1+ endothelial(4). Activated phenotypes included in this test were act./early exhausted CD4 (2,3,7), activated/proliferative CD4(6), activated/proliferative CD8(1,4), M1-MAC(3,4), activated/proliferative T-reg (5), proliferative endothelial(2), and activated endothelial(3,6).

### Spatial communities and TME quantification of penetration

The neighborhood analysis identified all pairwise interactions ($p<0.01$) and note that not all spatial interactions are symmetric. In order to describe the tumor topology neighborhoods, we summarized the TMEs from the topology classes, and toward the immune phenotypes. The total sum of the signed interactions in the outer/dispersed (topologies: b, e, g, and h), the tumor periphery (topologies: c, I, f) and tumor core (topology: d) were computed. We defined semi-penetrating as TMEs significantly attracted to the periphery zones and penetrating as TMEs significantly attracted to the tumor core. The 95% confidence intervals were computed using the signed interactions, and the Student t-distribution (d.f.=7).

### Cross-cohort analysis of DLBCL and Hodgkin lymphoma (HL)

To compare the TME in both diseases, we subset both experiments to include only CD4, CD8, Macrophages, and TREGs, and omitted all other phenotypes present in either experiment. The TME expression of integrated data including both DLBCL and HL identified primary phenotypes in the TME in both diseases. The HL data was obtained from our recent publication in Hodgkin's lymphoma that included 5 lymph nodes, 22 cases of HL (1 lymphocyte rich, 9 mixed cellularity, and 12 nodular sclerosis subtypes).[22] This experiment was performed with a 36-panel list, of which 21 overlapped with the DLBCL experiment (CCR4, CD206, CD20, CD30, CD3, CD45RA, CD45RO, CD4, CD68, CD8, CXCR3, EphrinB2, FOXP3, HLADR, ICOS, Ki67, LAG3, PD-1, PD-L1, TIM-3, and TBET). The meta-clustering ($k_1=45$, $k_2=15$) process identified 17 HL meta-clusters using Min/Max normalization as recommended by original authors.

To compare the TME expression in the 75 ROIs from both diseases, we subset both experiments to include only CD4, CD8, macrophages, and TREGs. Prior data integration, we first selected markers that had minimal differences per ROI across by visual inspection of the ROI interquartile ranges across both diseases. Markers such as CD4, CD8, CD68, FOXP3, CD206, DNA, PD-L1, PD-1, TIM-3, CCR4, CXCR3, ICOS, Ki67, and cell area were selected because they indicated similar inter-quartile ranges and most constant means across experiments. To integrate the single-cell data, we first standardized the expression to-standard normal, and then used 'limma (v.3.40.2)' to fit an empirical Bayes linear model and regress out the batch effects related to the 2 experiments at the single-cell level ('sva' (v.3.32.1) package).[23,24] To evaluate the batch effects, we used DNA, and the area of the cell objects as a positive control because we should expect that DNA content, and immune cell areas should be very similar across experiments. After data integration, we observed consistent area sizes across experiments and DNA expression.  In contrast, LAG-3 failed to integrate and had dissimilar dynamic range. and was excluded.  PCA, at both the case level and single cell level, along with uniform manifold mapping (uMAP) at the single cell level were performed to visually inspect the presence of batch across diseases.

After selecting markers and performing integration, we used the k-nearest neighbor batch-effect correction test ('kBET' R package (v.0.99.6)) to quantitatively measure the batch effect.  This test is a metric to measure the quality of the batch correction and performs a Pearson's $\chi^2$ test comparing a randomly sampled subset to measure the association to the levels of batch covariates[25] .  The Pearson's $\chi^2$ assumes that the data are inter-mixed, and the kBET returns a metric of the average rejection rate of the null hypothesis across the iterations (number of iterations= 100) of sampling.  Hence a low kBET score, corresponds with a low rejection of the null hypothesis in the $\chi^2$ test which assumes that the data are well inter-mixed. Therefore, low kBET scores corresponds with low batch effects. The kBET scores across all phenotypes yielded 0.2422 score, and CD4 had the lowest kBET score (kBET=0.1484, p-value=0.31), whereas TREGs had the highest batch effect score (kBET=0.37, p=0.08).  Note that we tested for batch on each TME component separately and did not adjust the p-values for multiple test corrections.   The input for the kBET metric where the 75 unique ROIs across both experiments (rows) and the columns were the Z-transformed selected protein measures previously described.

We examined the explained variance of the selected integrated expression markers by the phenotype levels (CD4, CD8, MAC, TREG), or the disease type (HL or DLBCL).  For a well-integrated experiment, the integrated expression should have a low explained variance associated with disease type.  We performed a Two-Way ANOVA regressing the protein expression onto categorical features such as phenotypes (4 levels) or disease type (2 levels). For PD-1, we observed that the phenotype categorical variable was a significant explanatory variable (F-statistic=15.38, p-value=0.025), whereas the experimental factor was not (F-statistic=1.32, p-value= 0.33). Similarly, PD-L1 the phenotype categorical variable was a marginal explanatory variable (F-statistic=4.56, p-value=0.12), whereas the experiment feature was not (F-statistic=0.02, p-value=0.88). For TIM-3, the phenotype feature explained more of the variability of TIM-3 (F-statistic=2.73, p-value=0.22) compared to the experimental factor (F-statistic=0.06, p-value=0.82).  The ANOVA indicated that the average expression had more association with the variability across phenotypes levels as opposed to experimental /disease type.

In order to compare proportions of the TME across disease types, the case relative proportions (%) were tested using beta regression, using an alpha threshold of 0.05 as the type-1 error rate.

The joint TME unsupervised clusters were annotated (Figure S13c) using threshold calls determined by a mixed-effects linear model ('lme4' version 1.1.23) which for each phenotype a contrast was developed which compared for significant differences of a given sub-cluster to the primary parent phenotype average and determined significant above average (+), average expression (mid), or significant below average (-) expression.

In order to further investigated the joint TME comparisons between DLBCL/HL/RLN we used unbiased tertile cutting of each normalized expression intensity relative to the same experiment and compared the relative case proportions to the reactive lymph node using beta regression modeling of the phenotype proportions (Figure S15).

The Figure S15d, and Figure 5d compared the neighborhood interactions with the reference as the $T_{REG}$ phenotypes (identified by unbiased tertile gating) and compared the relative frequencies of the total significant interactions ($p<0.01$) which ranged approximately from [-5, +5] relative to the total ROIs in the corresponding cohort (DLBCL/HL/RLN), and then compared the relative frequencies using an ANOVA (BH $q<0.05$). We checked the distribution of the frequencies which was approximately normal, and examined the residuals of the linear fit ensure adequacy.

## ACKNOWLEDGEMENTS/FUNDING

This project was supported by funding from the Ming Hsieh Institute at University of Southern California to A.M., J.H. and P.K. Additional support for this project was provided by K08 CA154975-01A1 career development award to A.M. from the National Cancer Institute, and a Paul Allen Distinguished Investigator award (Frontiers Group; C. Steidl). Access to the imaging mass cytometry system was obtained through the Fluidigm Early Access Program and support from the USC Michelson CSI-Cancer program. The content is solely the responsibility of the authors and does not necessarily

represent the official views of the National Cancer Institute or the National Institute of Health.

## BIBLIOGRAPHY

1. Xie Y, Bulbul MA, Ji L, et al. p53 Expression Is a Strong Marker of Inferior Survival in De Novo Diffuse Large B-Cell Lymphoma and May Have Enhanced Negative Effect With MYC Coexpression A Single Institutional Clinicopathologic Study. *Am J Clin Pathol*. 2014;141:593. doi:10.1309/AJCPPHMZ6VHF0WQV
2. Bandura DR, Baranov VI, Ornatsky OI, et al. Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry. *Analytical Chemistry*. 2009;81(16):6813-6822. doi:10.1021/ac901049w
3. Giesen C, Wang HAO, Schapiro D, et al. highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Articles nAture methods |*. 2014;11(4):417. doi:10.1038/nmeth.2869
4. Wang HAO, Grolimund D, Giesen C, et al. Fast Chemical Imaging at High Spatial Resolution by Laser Ablation Inductively Coupled Plasma Mass Spectrometry. *Analytical Chemistry*. 2013;85(21):10107-10116. doi:10.1021/ac400996x
5. Chevrier S, Crowell HL, Zanotelli VRT, Engler S, Robinson MD, Bodenmiller B. Compensation of Signal Spillover in Suspension and Imaging Mass Cytometry. *Cell Syst*. 2018;6(5):612-620.e5. doi:10.1016/j.cels.2018.02.010
6. Zanotelli V, Bodenmiller B. IMC Tools. https://github.com/BodenmillerGroup/imctools.
7. Jones TR, Kang IH, Wheeler DB, et al. CellProfiler Analyst: data exploration and analysis software for complex image-based screens. *BMC Bioinformatics*. 2008;9:482. doi:10.1186/1471-2105-9-482
8. Damond N, Engler S, Zanotelli VRT, et al. A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry. *Cell Metabolism*. 2019;29(3):755-768.e5. doi:10.1016/J.CMET.2018.11.014
9. Berg S, Kutra D, Kroeger T, et al. ilastik: interactive machine learning for (bio)image analysis. *Nature Methods 2019 16:12*. 2019;16(12):1226-1232. doi:10.1038/s41592-019-0582-9
10. Levine JH, Simonds EF, Bendall SC, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*. 2015;162(1):184-197. doi:10.1016/j.cell.2015.05.047
11. Jackson HW, Fischer JR, Zanotelli VRT, et al. The single-cell pathology landscape of breast cancer. *Nature*. 2020;578(7796):615-620. doi:10.1038/s41586-019-1876-x
12. Schapiro D, Jackson HW, Raghuraman S, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods*. 2017;14(9):873-876. doi:10.1038/nmeth.4391
13. Sakuishi K, Ngiow SF, Sullivan JM, et al. TIM3+FOXP3+ regulatory T cells are tissue-specific promoters of T-cell dysfunction in cancer. *Oncoimmunology*. 2013;2(4). doi:10.4161/ONCI.23849

14. Huang CT, Workman CJ, Flies D, et al. Role of LAG-3 in regulatory T cells. *Immunity*. 2004;21(4):503-513. doi:10.1016/J.IMMUNI.2004.08.010

15. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*. 2006;22(12):1540-1542. doi:10.1093/BIOINFORMATICS/BTL117

16. Sakai R, Winand R, Verbeiren T, Moere A vande, Aerts J. Dendsort: Modular leaf ordering methods for dendrogram representations in R. *F1000Res*. 2014;3. doi:10.12688/F1000RESEARCH.4784.1/DOI

17. LM W, M N, C S, MD R. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Commun Biol*. 2019;2(1). doi:10.1038/S42003-019-0415-5

18. Schmitz R, Wright GW, Huang DW, et al. Genetics and pathogenesis of diffuse large B-Cell lymphoma. *New England Journal of Medicine*. 2018;378(15):1396-1407. doi:10.1056/NEJMoa1801445

19. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nature Medicine*. 2018;24(5):679-690. doi:10.1038/s41591-018-0016-8

20. Yuan Y. Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. *Journal of the Royal Society Interface*. 2015;12(103). doi:10.1098/rsif.2014.1153

21. JH L, EF S, SC B, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*. 2015;162(1):184-197. doi:10.1016/J.CELL.2015.05.047

22. Aoki T, Chong LC, Takata K, et al. Single-Cell Transcriptome Analysis Reveals Disease-Defining T-cell Subsets in the Tumor Microenvironment of Classic Hodgkin Lymphoma. *Cancer Discov*. 2020;10(3):406-421. doi:10.1158/2159-8290.CD-19-0680

23. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/NAR/GKV007

24. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882-883. doi:10.1093/BIOINFORMATICS/BTS034

25. Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods*. 2019;16(1):43-49. doi:10.1038/S41592-018-0254-1