nature medicine

RESOURCE

https://doi.org/10.1038/s41591-018-0016-8

Corrected: Publisher Correction; Author Correction

Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes

Bjoern Chapuy[1,2,18], Chip Stewart[3,18], Andrew J. Dunford[3,18], Jaegil Kim[3], Atanas Kamburov[3], Robert A. Redd[4], Mike S. Lawrence[2,3,5], Margaretha G. M. Roemer[1], Amy J. Li[6], Marita Ziepert[7], Annette M. Staiger[8,9], Jeremiah A. Wala[3], Matthew D. Ducar[10], Ignaty Leshchiner[3], Ester Rheinbay[3], Amaro Taylor-Weiner[3], Caroline A. Coughlin[1], Julian M. Hess[3], Chandra S. Pedamallu[3], Dimitri Livitz[3], Daniel Rosebrock[3], Mara Rosenberg[3], Adam A. Tracy[3], Heike Horn[8], Paul van Hummelen[10], Andrew L. Feldman[11], Brian K. Link[12], Anne J. Novak[11], James R. Cerhan[11], Thomas M. Habermann[11], Reiner Siebert[13], Andreas Rosenwald[14], Aaron R. Thorner[10], Matthew L. Meyerson[2,3], Todd R. Golub[2,3], Rameen Beroukhim[2,3], Gerald G. Wulf[15], German Ott[9], Scott J. Rodig[2,16], Stefano Monti[6], Donna S. Neuberg[2,4], Markus Loeffler[7], Michael Pfreundschuh[17], Lorenz Trümper[15], Gad Getz[2,3,5,19*] and Margaret A. Shipp[1,2,19*]

Diffuse large B cell lymphoma (DLBCL), the most common lymphoid malignancy in adults, is a clinically and genetically heterogeneous disease that is further classified into transcriptionally defined activated B cell (ABC) and germinal center B cell (GCB) subtypes. We carried out a comprehensive genetic analysis of 304 primary DLBCLs and identified low-frequency alterations, captured recurrent mutations, somatic copy number alterations, and structural variants, and defined coordinate signatures in patients with available outcome data. We integrated these genetic drivers using consensus clustering and identified five robust DLBCL subsets, including a previously unrecognized group of low-risk ABC-DLBCLs of extrafollicular/marginal zone origin; two distinct subsets of GCB-DLBCLs with different outcomes and targetable alterations; and an ABC/GCB-independent group with biallelic inactivation of *TP53*, *CDKN2A* loss, and associated genomic instability. The genetic features of the newly characterized subsets, their mutational signatures, and the temporal ordering of identified alterations provide new insights into DLBCL pathogenesis. The coordinate genetic signatures also predict outcome independent of the clinical International Prognostic Index and suggest new combination treatment strategies. More broadly, our results provide a roadmap for an actionable DLBCL classification.

DLBCL is the most common lymphoid malignancy in adults, accounting for up to 35% of non-Hodgkin lymphomas. Although DLBCL is curable with combination therapy (R-CHOP) in over 60% of patients, the remainder develop recurrent or progressive disease that is often fatal. DLBCL is also a genetically heterogeneous disorder with multiple low-frequency mutations, somatic copy number alterations (SCNAs), and structural variants (SVs)[1–8]. These tumors are currently thought to arise from antigen-exposed B cells that transit through the germinal center (GC)[1]. Aspects of the GC environment, including the high proliferation rate, physiologic activation-induced cytidine deaminase (AID)-mediated immunoglobulin receptor editing, and aberrant somatic hypermutation (SHM) are conducive to malignant transformation[1].

The heterogeneity of DLBCL is reflected in transcriptionally defined subtypes that provide insights into disease pathogenesis and candidate treatment targets[9–14]. The cell-of-origin (COO) classification identifies ABC- and GCB-type DLBCLs[1,9]. ABC-DLBCLs are currently thought to be derived from B cells that have passaged through the GC and are committed to plasmablastic differentiation[1]. These tumors have increased NF-kB activity, and a subset exhibit

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Harvard Medical School, Boston, MA, USA. [3]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [5]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [6]Boston University School of Medicine, Section of Computational Biomedicine, Boston, MA, USA. [7]Institute for Medical Informatics, Statistics and Epidemiology, University Leipzig, Leipzig, Germany. [8]Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, and University of Tuebingen, Tuebingen, Germany. [9]Department of Clinical Pathology, Robert-Bosch Krankenhaus, Stuttgart, Germany. [10]Dana-Farber Cancer Institute, Center for Cancer Genome Discovery, Boston, MA, USA. [11]Mayo Clinic, Rochester, MN, USA. [12]University of Iowa, Iowa City, IA, USA. [13]Department for Human Genetics, University Ulm, Ulm, Germany. [14]Department of Pathology, University of Würzburg, Würzburg, Germany. [15]Department of Hematology and Oncology, Georg-August University Göttingen, Göttingen, Germany. [16]Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA. [17]Department of Medicine I, Saarland University, Homburg, Germany. [18]These authors contributed equally: Bjoern Chapuy, Chip Stewart, Andrew Dunford. [19]These authors jointly supervised this work: Gad Getz, Margaret A. Shipp. *e-mail: gadgetz@broadinstitute.org; margaret_shipp@dfci.harvard.edu

genetic alterations in NF-kB modifiers and proximal components of the B cell receptor (BCR) pathway and perturbed terminal B cell differentiation[1,11,13,15]. In contrast, GCB-DLBCLs are postulated to originate from light-zone GCBs[1]. A subset of these tumors have alterations in chromatin-modifying enzymes, PI3K signaling, and Gα-migration pathway components and frequent SVs of *BCL2* (refs [1,16–18]). Although patients with ABC-DLBCLs have been reported to have less favorable responses to standard therapy than those with GCB-DLBCLs[8,9,19], targeted analyses of select alterations have suggested that additional genetic complexity remains to be defined[2,11,18,20,21]. Despite the recognized clinical and molecular heterogeneity in DLBCL, previous genomic studies of this disease have largely focused on single types of alterations: mutations, SCNAs, or SVs.

To address these issues, we performed whole-exome sequencing (WES) with an expanded bait set to capture known SVs in 304 DLBCLs from newly diagnosed patients. Of the patients, 85% were uniformly treated with R-CHOP and had long-term follow-up; a subset of these patients were enrolled in the prospective multicenter RICOVER60 trial[22]. This representative and clinically annotated DLBCL cohort was used to comprehensively detect mutations, SCNAs, and SVs, and to identify five groups of patients with outcome-associated coordinate genetic signatures, three of which were previously undescribed.

Results

Significantly mutated driver genes. We detected mutations using WES data from 304 primary DLBCLs, 55% of which lacked patient-matched normal samples (Methods, Supplementary Fig. 1, and Supplementary Tables 1 and 2). To include all 304 samples in the discovery cohort for candidate cancer genes (CCGs), we developed new computational methods to filter germline variants and artifacts from tumor-only samples (Methods and Supplementary Figs. 2 and 3). After filtering, we found a median of 3.3 and 6.6 mutations/Mb in the paired and tumor-only samples, respectively, suggesting that an average of 3.3 germline variants per megabase persisted after filtering. Multiple lines of evidence indicate that these rare germline variants are spread throughout the genome and have minimal effect on the detection of CCGs (beta-binomial test, $P = 0.4$; Methods and Supplementary Figs. 2 and 3).

We applied MutSig2CV[23] to the 304 DLBCLs and detected 98 CCGs (q value < 0.1; Fig. 1 and Supplementary Table 3a). Our CCG list includes previously reported mutational drivers, including the tumor suppressor *TP53*; the chromatin modifiers *KMT2D*(*MLL2*), *CREBBP*, and *EP300*; the components of the BCR, Toll-like receptor (TLR), and NF-kB signaling pathways *CD79B*, *MYD88*, *CARD11*, and *TNFAIP3*(*A20*); certain components of the RAS pathway, *KRAS*, and *BRAF*; *NOTCH2* and the NOTCH signaling modifier *SPEN*; and the immunomodulatory pathway components, *B2M*, *CD58*, *CD70*, and *CIITA* (Fig. 1a)[3–8]. As a result of the improved methodology and increased sample size, we identified 40 additional previously undescribed CCGs in DLBCL[8], many of which have defined roles in other lymphoid malignancies or cancers (Supplementary Fig. 3r,s). These include the additional modifiers of the BCR and TLR signaling pathways *PTPN6*(*SHP1*), *LYN*, *HVCN1*, *PRKCB*, and *TLR2*; the histone genes *HIST1H1B*, *HIST1H1C*, *HIST1H1D*, *HIST1H2AC*, *HIST1H2AM*, *HIST1H2BK*, *HIST1H3B*, and *HIST2H2BE*; *BCL11A*, *IL6*, *CCL4* (*MIP-1β*), and the PD-1 ligand *CD274* (PD-L1) (Fig. 1a and Supplementary Fig. 4).

To identify genes with significant clustering in three-dimensional protein structures, we used CLUMPS[24], which revealed 22 CCGs (q value < 0.1). Notably, 7 of 22 CCGs were not captured by MutSig2CV, including an additional member of the KRAS-BRAF-MEK1 pathway, *MAP2K1*(*MEK1*) (Supplementary Fig. 5a–d and Supplementary Table 3b). CLUMPS also provided insights into the putative function of mutations: *TP53* alterations clustered into two distinct regions of the protein, the DNA-binding site and the Zn^{2+} atom coordinating residues required for p53 structural integrity (Fig. 1b); non-canonical *BRAF* mutations perturbed the autoinhibitory interaction of the P and activation loops (Fig. 1b and Supplementary Fig. 5e); and clustered mutations in *CREBBP*, *PTPN6*(*SHP1*), and *GNAI2* abolished polar interactions around the catalytic pocket (Fig. 1b and Supplementary Fig. 5). A second step in CLUMPS (called EMPRINT) identified enrichment of mutations at protein-protein interfaces. For example, *RHOA* mutations cluster at the binding interface with multiple Rho guanine nucleotide exchange factors (ARHGEFs), keeping RHOA in its inactive form and de-repressing PI3K signaling and Gα migration (Fig. 1c, Supplementary Fig. 6a–c, and Supplementary Table 3c)[1]. In addition, CLUMPS identified mutation clustering at the acceptor groove of *FBXW7* that limits CCNE1 recognition and CUL1/SKP1/FBXW7-mediated degradation, a previously reported tumor suppressor mechanism in other cancers (Supplementary Fig. 6d,e and Supplementary Table 3c).

Mutational processes. Mutational processes leave a characteristic imprint, a mutational signature, in the cancer genome that reflects both DNA damage and repair. We applied our SignatureAnalyzer tool[25], which uses both the three-base mutational sequence context and mutational clustering in genome coordinates, to discover four signatures (three signatures after removal of a single micosatelite instability case; Supplementary Methods, Fig. 2a, Supplementary Fig. 7a–c, and Supplementary Table 4). The predominant mutational signature, which explained 80% of all mutations, was a spontaneous deamination at CpG sites (C > T_CpG, hereafter referred to as aging; Fig. 2a,b and Supplementary Fig. 7). Consistent with the underlying etiology of this signature, older patients had more mutations driven by spontaneous deamination (Supplementary Fig. 7d). We also identified two AID-driven signatures, canonical AID (cAID) and AID2, that reflect different repair mechanisms following AID-induced deamination of cytosine to uracil. The cAID signature was characterized by increased C > T/G mutations at a known AID hotspot, the RCY-motif(R = A/G,Y = C/T)[25,26]. Consistently, cAID activity was enriched at sites of both physiologic and aberrant SHM (Fig. 2a,b, Supplementary Fig. 7e, and Supplementary Table 4)[27]. The AID2 signature was dominated by A > T/C/G mutations at WA(W = A/T) motifs and shared some properties of the COSMIC9/non-canonical AID signature[25,26].

Next, we determined the relative contributions of aging, cAID, and AID2 mutational processes to each CCG (Fig. 2c and and Supplementary Fig. 7f). Genes that are known targets of aberrant SHM, including *BCL2*, *SGK1*, *PIM1*, and *IGLL5* (Fig. 2c and Supplementary Table 4d,e)[25], had predominant AID signatures (cAID + AID2) comprised of mutations with the lowest ratio of non-silent to silent mutations (Fisher's exact test, $P = 1.97 \times 10^{-4}$) that clustered within 2 kb of the transcription start site (Fisher's exact test, $P = 2 \times 10^{-41}$), consistent with the AID mechanism. In contrast, genes including *MYD88*, *KMT2D*(*MLL2*), *EP300*, *TNFAIP3*(*A20*), *TP53 and PRDM1*(*BLIMP1*), had predominant aging mutational signatures (Fig. 2c, Supplementary Fig. 7f-g and Supplementary Table 4c).

Chromosomal rearrangements and SCNAs. We next assessed recurrent SVs using a previously described targeted sequencing approach[28] and a pipeline that included four different algorithms followed by a filtering and split-read validation step (Methods, Supplementary Figs. 1 and 8a–e, and Supplementary Table 5). We identified at least 1 SV in 64%(189 of 296) of tumors; translocations that juxtaposed genes to strong regulatory elements were the most common SVs (Fig. 3 and Supplementary Fig. 8d).

As expected[1,29,30], *IGH*, *BCL2*, *BCL6*, and *MYC* were the most frequently rearranged genes (40, 21, 19, and 8%, respectively)

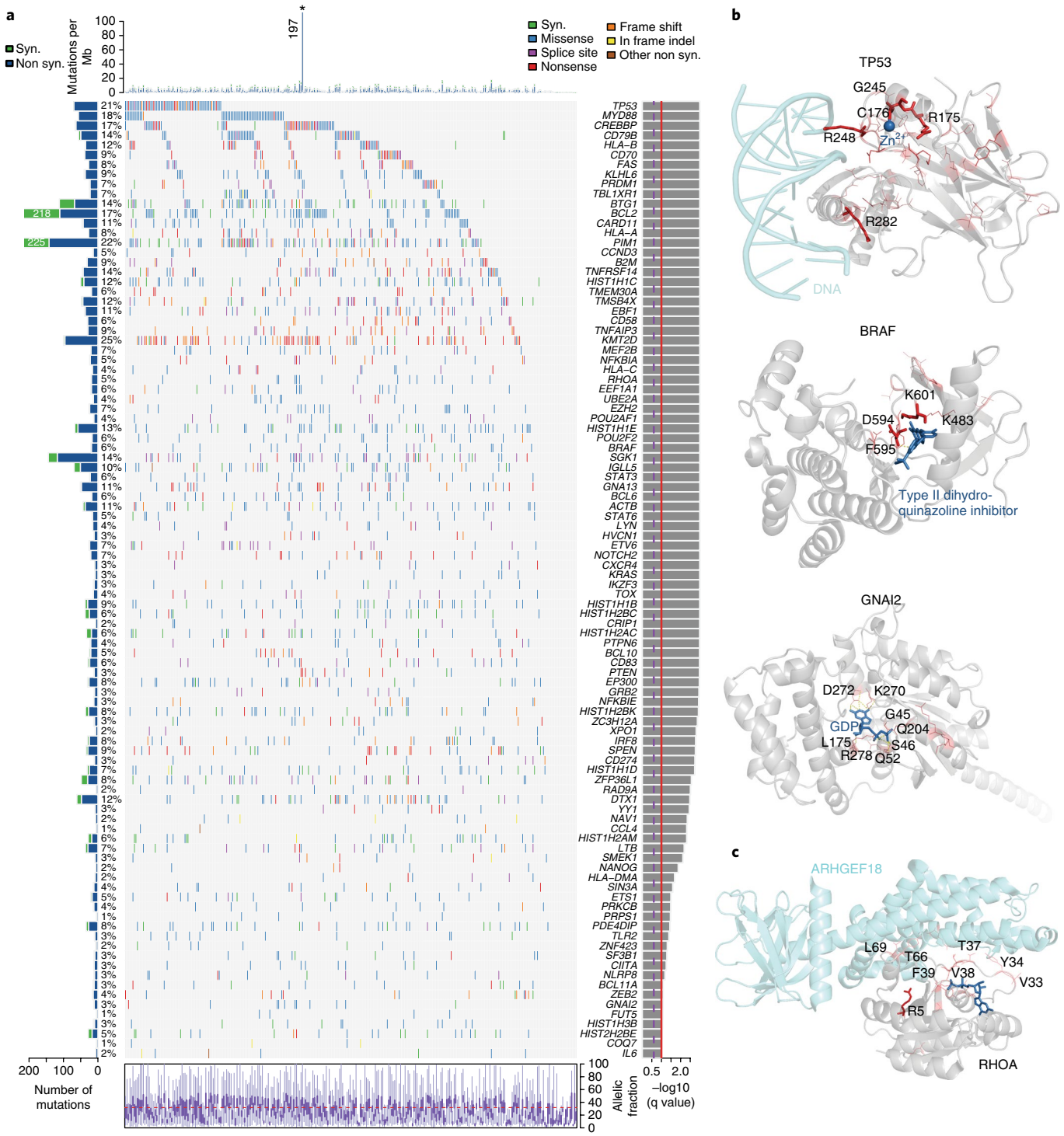

Fig. 1 | Recurrently mutated genes in 304 primary DLBCLs. a, Number and frequency of recurrent mutations (left), gene-sample matrix of recurrently mutated genes (color-coded by type, center), ranked by their significance (MutSig2CV q value, right). Total mutation density across the cohort is shown at the top, allelic fraction of mutations at the bottom. Asterisk indicates hypermutator case. **b**, Genes that were also identified by CLUMPS include: *TP53*, *CREBBP*, *KLHL6*, *BRAF*, *STAT6*, and *GNAI2*. Representative examples of genes with significant spacial clustering in protein structures (gray): TP53 (top; PDB:4MZR), BRAF (middle; PDB ID:4G9R), and GNAI2 (bottom; PDB:1AGR). Mutated residues are shown in red and color intensity scales with the number of mutations. Polar interactions are shown in dotted yellow lines. Frequently mutated residues are labeled in black. Co-crystalized proteins are shown in blue (Zn^+, Type II dihydroquinazoline inhibitor and GDP). **c**, Co-crystal structure of RHOA (gray) and ARHGEF18 (cyan; PDB:4D0N) highlights mutational clustering at the RHOA-ARHGEF interface. Residues at the interface are shown in black.

followed by the PD-1 ligands *PD-L1* and *PD-L2* (5%), and then *TBL1XR1* (4%), *TP63* (3%), *CIITA* (3%), and *ETV6* (2%) (Fig. 3a–g and Supplementary Figs. 8e and 9a–f). The *IgH* enhancer region was the predominant rearrangement partner (97%) of *BCL2*, and breakpoints were almost exclusively distal to the *BCL2* open reading frame (ORF) (Fig. 3a,d). Although *Ig* loci enhancers were the most common rearrangement partners for *BCL6* and *MYC* (57 and 58%, respectively), we identified multiple additional partners;

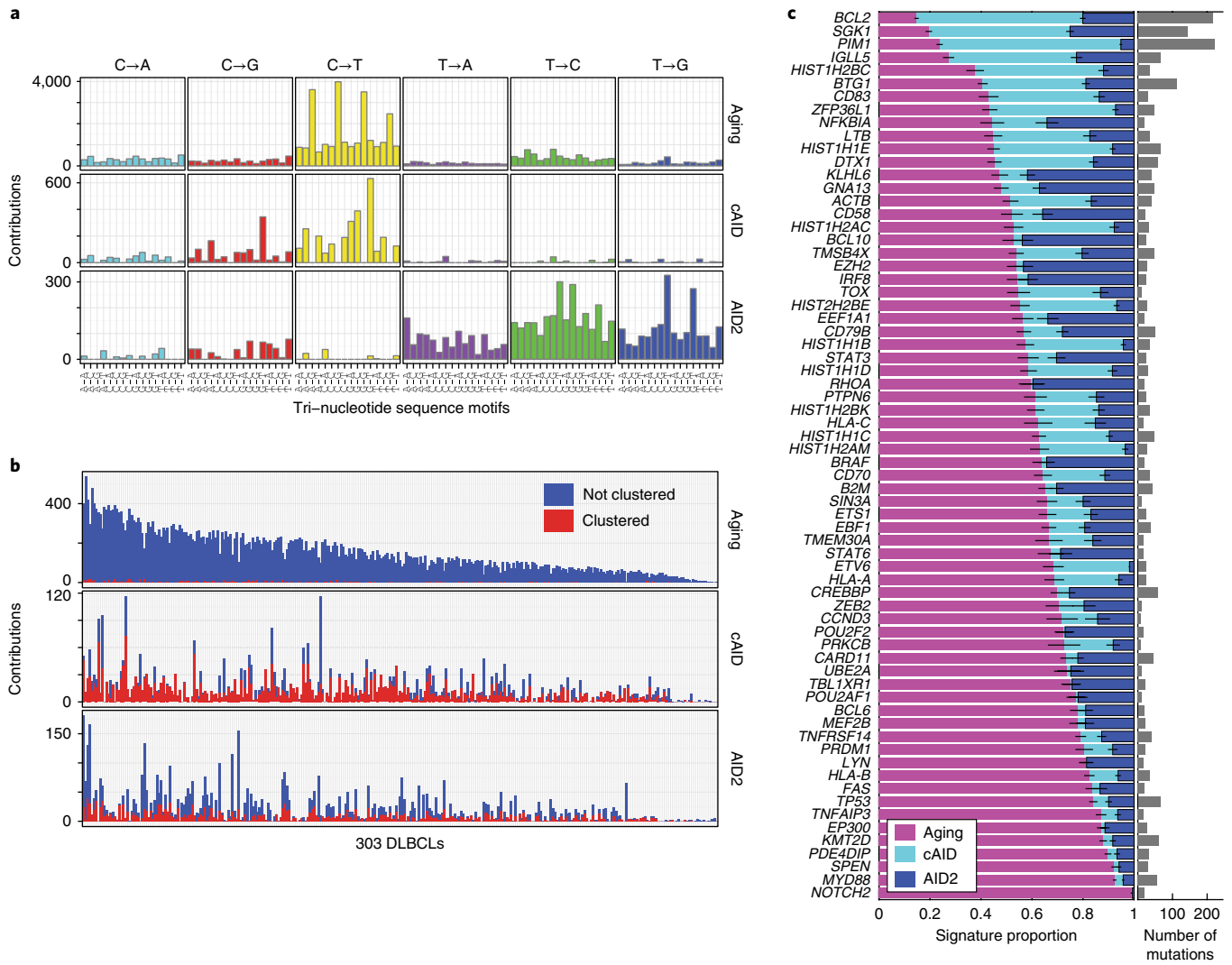

Fig. 2 | Mutational signatures operating in primary DLBCLs. a, Mutation signature analysis with the clustering information of mutations quantified by the nearest mutation distance (NMD) identified three mutational signatures: C > T mutations at CpG islands (C > T CpG, aging), cAID, and AID2 in 303 DLBCL samples. One sample with a predominant contribution of the MSI signature activity (SNVs > 5,000; Methods) was excluded. **b**, Signature activity (the number of mutations assigned to each signature) in each group of clustered (red; NMD ≤1kb) and non-clustered mutations (blue; NMD > 1kb) across 303 DLBCL samples sorted by decreasing mutation count. **c**, Relative enrichment of signature activities in significantly mutated genes with at least ten mutations. Number of mutations per gene are shown on the right. Genes are sorted by prevalence of the aging signature. Error bars show the s.e.m.

breakpoints in *BCL6* and *MYC* were predominantly proximal to the ORFs (Fig. 3b,c,e,f). *PD-L1* and *PD-L2* SVs involved multiple regulatory elements juxtaposed to intact ORFs with increased expression of the respective protein (Fig. 3g–i), as previously described[28]. Less frequently, *Ig* regulatory elements (*IgH*, *Igκ*, and *Igλ*) were juxtaposed to additional partners with known roles in GCBs (*BACH2*, *BCOR*, *FOXP1*, *miR-17-92*, *CCND1*, *CIITA*, *SOCS1*, and *NFKBIE*) (Supplementary Fig. 9a–g).

Next, we identified significantly recurrent SCNAs with the GISTIC2.0 program based on the WES data. We detected 18 arm-level and 18 focal regions of copy gain and 2 arm-level and 32 focal regions of copy loss (q value ≤ 0.1, frequency ≥ 3%; Fig. 4a). The frequencies of these SCNAs ranged between 5 and 32% and the number of genes in focal peaks varied from 4 (*2p16* gain) to 549 (*1q23.3* gain). We did not observe chromothripsis in our dataset (Supplementary Note).

To provide insights regarding candidate driver genes in SCNAs, we leveraged available gene expression data and performed an integrative analysis[2] (Supplementary Note and Supplementary Table 6). For each focal alteration, genes from the COSMIC Cancer Gene Census with a significant association between transcript abundance and SCNA were identified (Fig. 4a, Supplementary Table 6, and Supplementary Methods). In DLBCLs with focal *13q31.3* gain, the transcript with the highest fold change was miR-17-92 (Fig. 4a and Supplementary Table 6).

CCGs were significantly more likely to reside within focal SCNAs (Fisher exact test, $P = 10^{-44}$; Fig. 4a), suggesting that these driver genes were perturbed by multiple mechanisms. Significant genes altered by mutations, CN gain, and/or SVs included *NOTCH2*(*1q23.3*), *CCND3*(*6p21.1*), *PD-L1/PD-L2/JAK2*(*9p24.1*), and *BCL2*(*18q/18q21.33*); those perturbed by mutations and CN losses included *CD58*(*1q13.1*), *TNFAIP3*(*6q23.3*), *PRDM1*(*BLIMP1*; *6q21*), *B2M*(*15q15.3*), *PTEN/FAS*(*10q23.31*), *CD70*(*19p13.3*), *RHOA*(*3p21.31*), *TMEM30A*(*6q14.1*), and *TP63*(*3q28*). Of note, 74% of DLBCLs exhibited genetic bases of immune escape[7,28,31–33], including alterations of MHC class I loci,

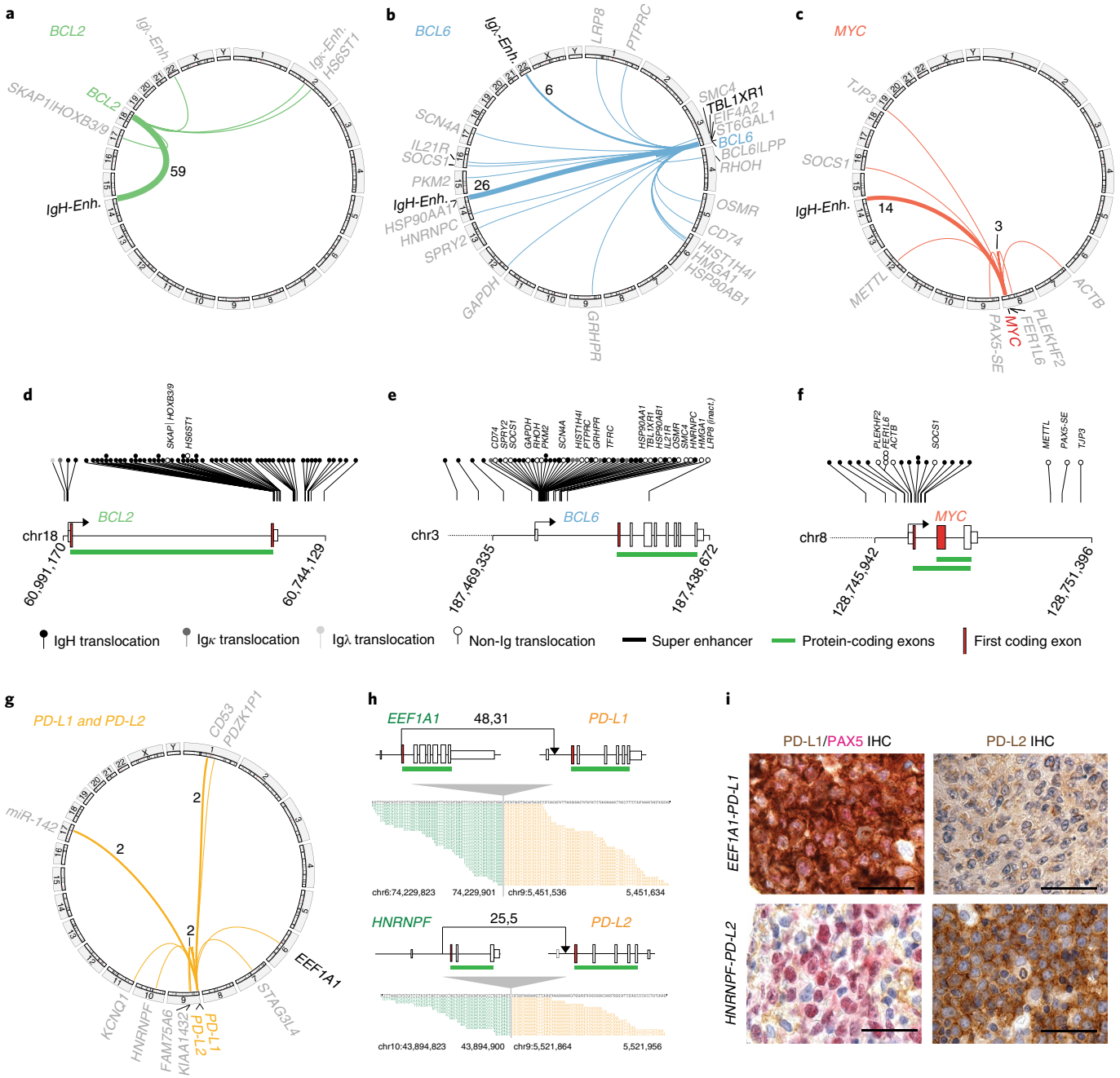

Fig. 3 | Chromosomal rearrangements in primary DLBCLs. a–**c**, SVs of *BCL2* (**a**, green), *BCL6* (**b**, blue), *MYC* (**c**, red), and partner genes (gray) are visualized as Circos plots. Genes also targeted by somatic mutations are highlighted in black. Thickness of partner linking lines indicates frequency (numbers indicate frequency > 1). **d**–**f**, Breakpoints in *BCL2* (**d**), *BCL6* (**e**), and *MYC* (**f**) are plotted in their indicated genomic context. Arrows indicate the transcription start site in the coding direction; boxes indicate exons including first coding exon (red); green bar below indicates which exons are protein coding. Translocation partners are indicated by the shading of the circle at the tip of each breakpoint (*IgH*, black; *Igκ*, dark gray; *Igλ*, light gray; non-Ig partners, white and name of partner gene above). **g**, Circos plots of chromosomal rearrangements involving the PD-1 ligand loci, *PD-L1*, and *PD-L2* (orange). Plots are labeled as in **a**–**d**. **h**, Stick figures for indicated translocations involving either *PD-L1* or *PD-L2*. See **h** for details. Raw reads counts are visualized below. Reads mapping to the first and second partner gene are highlighted in green and orange, respectively. **i**, PD-L1/PAX5 (left, PD-L1, brown; PAX5, pink) and PD-L2 (right; PD-L2, brown) immunohistochemical (IHC) analyses of the cases in **h**. IHC was repeated twice with similar results.

B2M, *CD70*, *CD58*, *CD274*(*PD-L1*), *PDCD1LG2*(*PD-L2*), and *CIITA* (Supplementary Fig. 9i).

Association of individual genetic features to outcome. Next, we assessed the prognostic value of our identified genetic drivers for progression-free survival (PFS) and overall survival (OS) in the subset of patients who were treated with R-CHOP-like therapy ($n = 259$, median follow-up 78.5 months). Loss of *1q42.12*, *MYC* SVs and gains of *18q21.33/BCL2*, *13q31.3/miR-17-92*, and *18p* were independently predictive of inferior PFS; all retained significance when added to IPI risk groups (Fig. 4b,c, Supplementary Fig. 10a, and Supplementary Table 7). *MYC* SVs, *13q31.3* gain, and *1q41.12* loss were also associated with shortened OS alone and when added to International Prognostic Index (IPI) risk groups

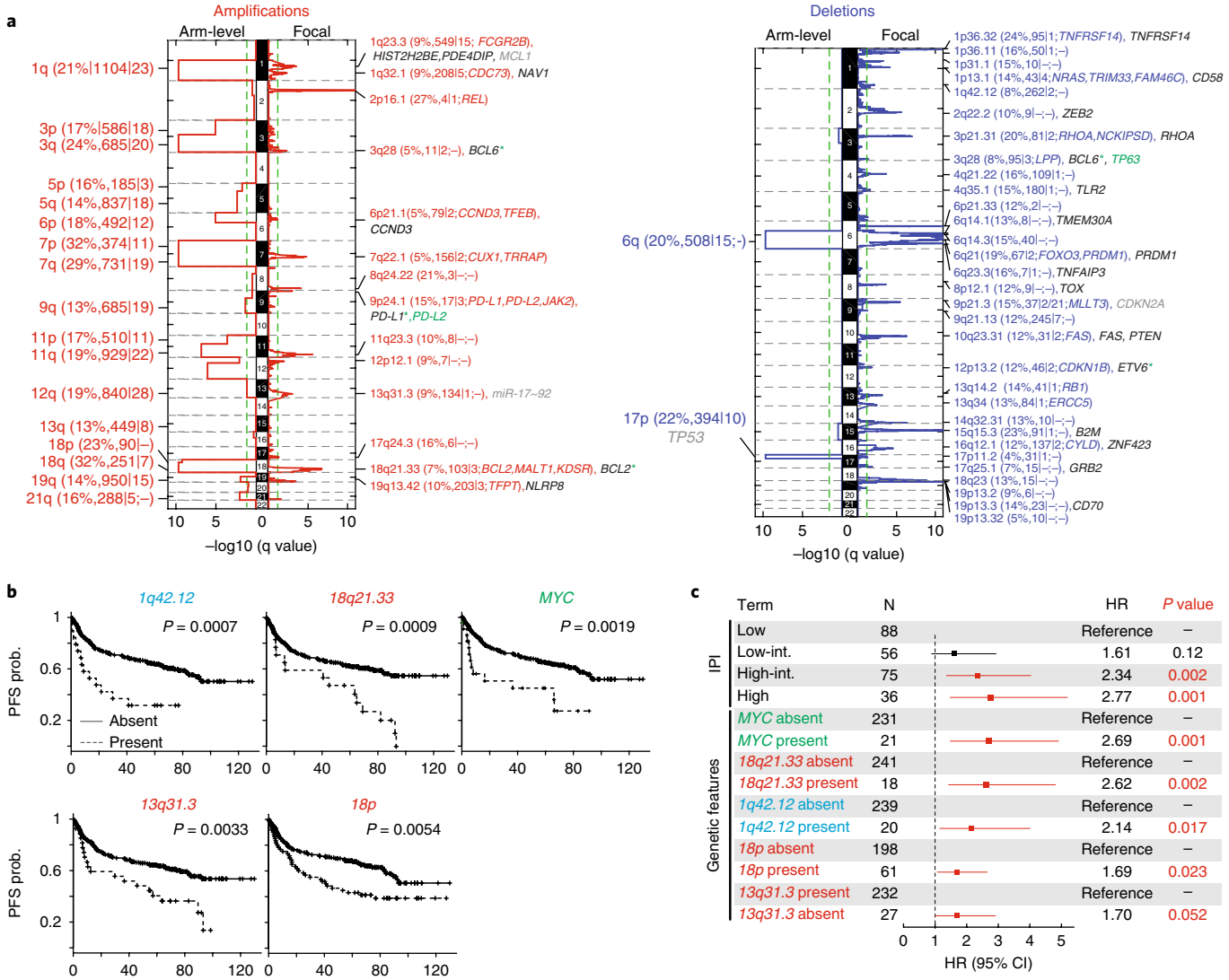

Fig. 4 | Recurrent SCNAs and outcome association of individual genetic factors. a, GISTIC2.0-defined recurrent copy number gains (red, left) and losses (blue, right) are visualized as mirror GISTIC plots, with arm-level events (left) and focal events (right). Chromosomes are shown on the vertical axis. Green line denotes q value of 0.1. SCNAs are labeled with their associated cytoband/arm followed in brackets by the frequency of the alteration, the number of total genes and COSMIC-defined cancer genes in GISTIC2.0-defined regions, respectively. For focal events, COSMIC cancer genes with a positive correlation to gene expression in our data (fold change > 1.2, q < 0.25) are indicated in the brackets. Genes that were also significantly mutated (in black) or subject to chromosomal rearrangement ($n = >2$, green) in our dataset are highlighted after the brackets. Other important drivers are labeled in gray. **b**, Kaplan Meier plots of individual genetic factors predictive for PFS in univariate and multivariate models of the R-CHOP treated cohort with PFS data ($n = 254$); alterations present, dashed line; *P* values were derived from log-rank test. **c**, Forest plots visualize the multivariate analysis of IPI risk groups and individual genetic factors for PFS in the R-CHOP treated cohort with PFS data ($n = 254$).

(Supplementary Fig. 10b–d and Supplementary Table 7). Notably, the prognostically significant individual alterations were SCNAs or SVs rather than mutations (Fisher's exact test; PFS, $P = 0.007$; OS, $P = 0.02$).

Coordinate genetic signatures capture biologic heterogeneity. DLBCLs in this series harbored a median of 17 (range: 0–48) genetic drivers, prompting additional analyses of co-occurring alterations. We applied non-negative matrix factorization (NMF) consensus clustering[34] to the 158 identified genetic driver alterations and discovered five robust subsets of tumors (clusters) with discrete genetic signatures (hereafter referred to as coordinate genetic signatures; C1–C5; 51–72 samples each) and an additional subset without detectable alterations (C0; 12 samples) (Methods, Fig. 5, Supplementary Figs. 11 and 12, and Supplementary Table 8).

Cluster 5. The 64 cluster 5 (C5) DLBCLs exhibited near-uniform *18q* gain, likely increasing expression of BCL2 and other candidate drivers such as MALT1 (refs.[21,35]). These tumors also had frequent mutations in *CD79B* (48%, 29 of 60) and *MYD88* (50%, 30 of 60), alterations previously associated with ABC-type DLBCLs[11,13,20]. *MYD88* mutations selectively involved L265P and often occurred in association with *CD79B* mutations (Fisher's exact test, $P = 0.036$; Figs. 5 and 6a,b and Supplementary Fig. 13a). Additional alterations linked to ABC-DLBCLs, including gains of *3q*, *19q13.42* and inactivation of *PRMD1*, were observed in this cluster[2,36], as were the prognostically significant *18p* copy gains (Figs. 4b and 5). In this cluster, 96% (45 of 47) of tumors with available COO designations were ABC-DLBCL type (Fisher exact test, $P < 0.001$).

Major components of the C5 signature, including frequent *BCL2* gain, concordant $MYD88^{L265P}$/*CD79B* mutations, and additional

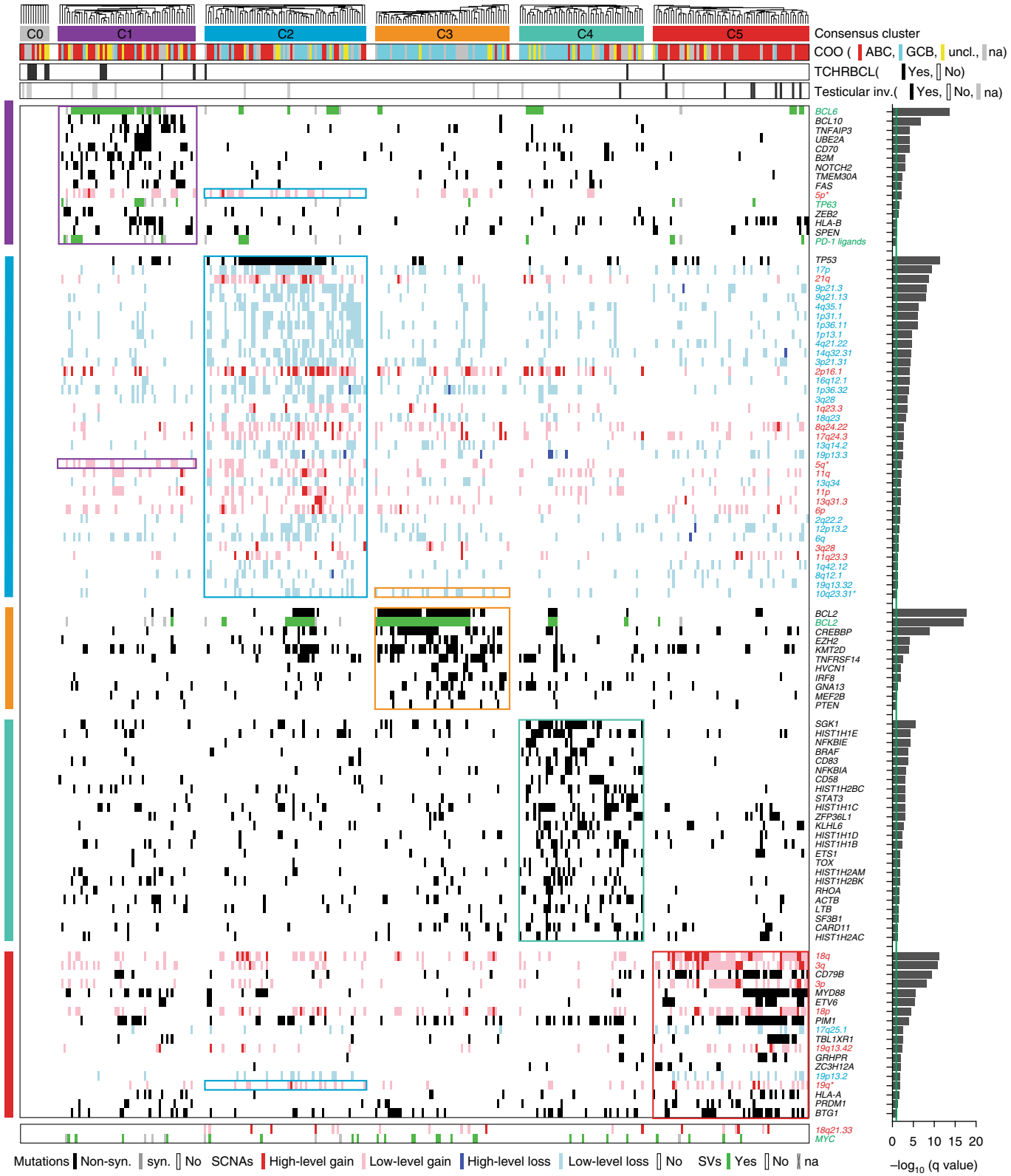

Fig. 5 | Identification of groups of tumors with coordinate genetic signatures. Non-negative matrix factorization consensus clustering was performed using all CCGs, SCNAs, and SVs in the 304 DLBCL samples (columns). Clusters C1–C5 with their associated landmark genetic alterations are visualized (boxed for each cluster). Samples without driver alterations are represented as cluster C0. Genetic alterations that were positively associated with each cluster were identified by a one-sided Fisher test and ranked by significance ($q < 0.1$, green line, bar graph to the right). Non-synonymous mutations, black; synonymous mutations, gray; single CN loss ($1.1 \leq CN \leq 1.6$ copies), cyan; double CN loss ($CN \leq 1.1$), blue; low-level CN gain (3.7 copies $\geq CN \geq 2.2$ copies), pink; high-grade CN gain ($CN \geq 3.7$ copies), red; chromosomal rearrangement, green; no alterations, white; gray crossed, not assessed. Header shows cluster association (C0, gray; C1, purple; C2, blue; C3, orange; C4, turquoise; C5, red), COO classification (ABC, red; GCB, cyan; unclassifiable, yellow; not assessed, gray), TCHRBCL cases (black, yes; white, no), and testicular involvement (black, yes; white, no; gray, na). Outcome-associated alterations that are not part of a specific cluster, SVs of *MYC*, and *18q21.33* copy gain are shown below.

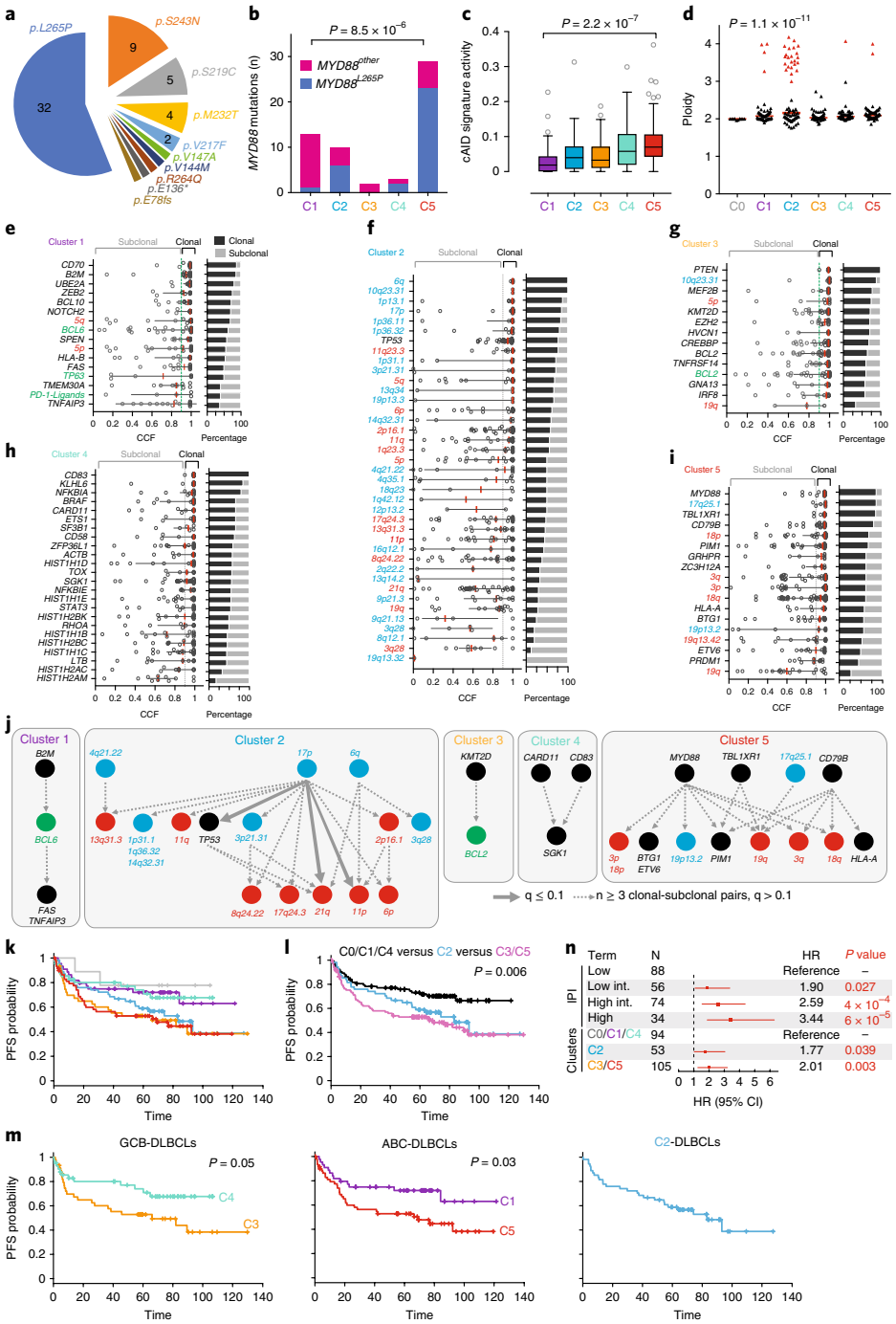

Fig. 6 | Type and incidence of *MYD88* mutations, cAID mutational signature activity, inferred timing of genetic drivers, and outcome association of DLBCL clusters. a, Type of *MYD88* mutations. **b**, Frequency of $MYD88^{L265P}$ and $MYD88^{other}$ mutations across clusters C1–C5 ($n=292$); *P* value by two-sided Fisher's exact test. **c**, Fraction of cAID mutational signature activity in clusters C1–C5 ($n=292$) as a Tukey boxplot (center, median; box, interquartile range (IQR); whiskers, 1.5 × IQR); *P* values by two-sided Mann-Whitney U test. **d**, Ploidy as inferred by ABSOLUTE in clusters C1–C5 ($n=292$) as scatter plot (red line, median). DLBCLs with genome doublings (an inferred ploidy ≥3) are indicated in red; *P* value by two-sided Fisher's exact test. **e–i**, CCFs of clusters C1–C5 (C1, $n=56$; C2, $n=66$; C3, $n=55$; C4, $n=51$; C5, $n=64$) are plotted and ranked by the fraction of clonal events of each landmark alteration (high to low, right). Median CCF in red bar, error bar represents the interquartile range. Mutations, black; CN gain, red; CN loss, blue; SVs, green. The threshold for assigning an alteration to be 'clonal' is a CCF of ≥0.9 (green dotted line). **j**, Timing of cluster-associated alterations is visualized with early events at top, late events at bottom. Color indicates alteration type as above. Arrows between two alterations were drawn when two drivers were found in one sample with an excess of clonal to subclonal events. Line type of arrows indicates significance derived from a binomial test (solid thick arrow, q value < 0.1; dotted line, too few clonal-subclonal pairs to formally test with binominal test). **k**, Kaplan Meier plots for PFS for all clusters, C0 (gray), C1 (purple), C2 (blue), C3 (orange), C4 (turquoise), C5 (red). **l**, KM plot for PFS for favorable DLBCL clusters (C0, C1, and C4) in black, C2-DLBCLs in blue and unfavorable DLBCLs (C3 and C5) in pink. The *P* value obtained using the log-rank test. **m**, KM plot for PFS for the genetically distinct GCB-DLBCL clusters (C3 and C4; left), the ABC-DLBCL clusters (C1 and C5; middle) and C2 DLBCLs. The *P* value obtained using the log-rank test. **n**, Forest plots visualize HR and *P* values obtained from the multivariate analysis of clusters and IPI for PFS. **k–n**, Analyses were performed in the R-CHOP treated cohort with PFS data ($n=254$).

mutations of *ETV6*, *PIM1*, *GRHPR*, *TBL1XR1*, and *BTG1* (Fig. 5), were similar to those recently described in primary CNS and testicular lymphoma[28]. Thus, we identified systemic DLBCLs with CNS or testicular involvement and found that eight of nine patients with testicular disease were in this cluster (Fisher's exact test, $P<0.001$), as was one of two patients with CNS involvement. These data suggest that the C5 genetic signature is associated with extranodal tropism and extend the findings of targeted sequencing studies linking $MYD88^{L265P}$ with extranodal disease[37,38]. C5 DLBCLs have the highest contribution of cAID and associated aberrant SHM indicative of tumors that have passaged through the GC (Fig. 6c)[1].

Cluster 1. The majority of the 56 cluster 1 (C1) DLBCLs exhibited *BCL6* SVs in combination with mutations of NOTCH2 signaling pathway components, predominantly activiating PEST-domain mutations of *NOTCH2* and truncating mutations of its negative regulator, *SPEN*. C1 DLBCLs also had increased transcriptional abundance of NOTCH2 and BCL6 target genes, as determined by gene set enrichment analysis (GSEA) (Supplementary Fig. 13f). In addition, these tumors harbored frequent mutations of the NF-kB pathway members *BCL10* and *TNFAIP3*(*A20*), and *FAS* (Fig. 5 and Supplementary Fig. 4). Alterations in NOTCH and NF-kB pathway components and *FAS* mutations were previously found in low-grade marginal zone lymphomas (MZLs)[39–42], and *BCL6* translocations were described in transformed MZLs[43].

C1 DLBCLs had no histologic features of MZLs, suggesting that these tumors were either occultly transformed before diagnosis or that they derived de novo from a common extrafollicular B cell precursor with shared genetic features. MZLs typically arise in a setting of chronic inflammation, often in response to pathogen-driven antigen stimulation[44]. Notably, C1 DLBCLs exhibited multiple genetic bases of immune escape, including inactivating mutations in *B2M*, *CD70*, *FAS*, and SVs of *PD-L1* and *PD-L2* (Fig. 5 and Supplementary Figs. 4 and 9i)[28,31].

The majority of C1 DLBCLs were classified as ABC-type tumors by transcriptional profiling (Fisher's exact test, $P=0.01$). Although 25% (13 of 51) of C1 DLBCLs exhibited *MYD88* mutations, these were almost exclusively $MYD88^{non\text{-}L265P}$, in contrast with the predominant $MYD88^{L265P}$ found in C5 ABC DLBCLs ($P<0.001$; Fig. 6a,b and Supplementary Fig. 13a). $MYD88^{L265}$ and $MYD88^{non\text{-}L265P}$ differ in their ability to coordinate IRAK1/IRAK4-containing signaling complexes and activate NF-kB[11]. C5 and C1 ABC-DLBCLs also differed in the contribution of cAID to their mutational spectrum (C1 versus C5, $P<0.001$; C1 versus rest, $P<0.001$; Fig. 6c and Supplementary Fig. 13d). In contrast with C5 tumors, C1 DLBCLs had low or absent cAID activity, providing additional evidence of an extrafollicular origin and a lower rate of SHM (Fig. 6c)[45].

Taken together, the coordinate genetic signatures of C1 and C5 ABC-type DLBCLs define subsets of tumors with distinct pathogenetic mechanisms. These findings (Figs. 5 and 6b) also suggest different targeted treatment strategies in the genetically distinct ABC-DLBCLs: inhibition of proximal BCR/TLR signaling and BCL2 in C5 and perturbation of NOTCH and BCL6 signaling and immune evasion mechanisms in C1.

Cluster 3. The majority of the 55 cluster 3 (C3) DLBCLs harbored *BCL2* mutations with concordant SVs that juxtaposed *BCL2* to the *IgH* enhancer (Fisher exact test, $P=3.3\times10^{-35}$; Fig. 5 and Supplementary Fig. 9h). C3 DLBCLs also exhibited frequent mutations in chromatin modifers, *KMT2D*, *CREBBP*, and *EZH2*, and increased transcriptional abundance of EZH2 targets by GSEA (Supplementary Fig. 13g). These tumors also had alterations of the B cell transcription factors *MEF2B* and *IRF8*, and indirect modifiers of BCR and PI3K signaling, *TNFSF14*(*HVEM*), *HCNV1*, and *GNA13* (Fig. 5 and Supplementary Fig. 4). In addition, these tumors had two alternative mechanisms of inactivating *PTEN*: focal *10q23.31/PTEN* loss and predominantly truncating *PTEN* mutations (Fig. 5). The two types of *PTEN* alterations are noteworthy because the PTEN N-terminal and C-terminal domains have distinct roles in antagonizing PI3K/AKT signaling, maintaining genomic stability, and inducing murine B cell lymphomas [18,46,47]. C3 genetic alterations have been described in follicular lymphoma (FL) and de novo GCB-type B cell lymphomas[4,16–18,36,48–53]. Consistent with this finding, 95% (38 of 40) of C3 DLBCLs with available COO designations were of the GCB type (Fig. 5).

Cluster 4. The 51 cluster 4 (C4) DLBCLs were characterized by mutations in four linker and four core histone genes, multiple immune evasion molecules (*CD83*, *CD58*, and *CD70*), BCR/Pi3K signaling intermediates (*RHOA*, *GNA13*, and *SGK1*), NF-kB modifiers (*CARD11*, *NFKBIE*, and *NFKBIA*), and RAS/JAK/STAT pathway members (*BRAF* and *STAT3*).

C4 DLBCLs were primarily GCB type (Fisher's exact test, $P=0.01$), suggesting that C4 and C3 DLBCLs represent genetically distinct subsets of GCB tumors (Fig. 5). Comparison of the C3 and C4 genetic signatures further revealed that these GCB-DLBCLs utilize distinct mechanisms to perturb common pathways such as PI3K signaling. In contrast with C3 DLBCLs, C4 tumors rarely exhibited *PTEN* alterations, but harbored more frequent *RHOA* mutations (Fig. 5). In addition, C4 DLBCLs rarely exhibited *BCL2* alterations.

Unlike C3 tumors, C4 DLBCLs largely lacked alterations in chromatin-modifying enzymes, but frequently exhibited mutations in H1 linker histones and additional core histones that have also been described in FL[52,54,55]. The identified mutations in the globular or C-terminal domains of H1 linker histones likely reduce their association with chromatin and/or perturb interactions with additional effector molecules (Supplementary Fig. 4)[54–56]. H1 linker and core histone alterations may increase mutation rates by opening chromatin and exposing DNA to ongoing AID activity; indeed, C4 tumors have a significantly higher mutational density ($P<0.0001$; Supplementary Fig. 13c).

The distinct genetic features of C3 and C4 GCB-DLBCLs also suggest specific targeted therapies, including inhibition of BCL2, PI3K, and the epigenetic modifiers EZH2 and CREBBP in C3 GCB tumors, and JAK/STAT and BRAF/MEK1 blockade in C4 GCB-DLBCLs.

Cluster 2. The 64 cluster 2 (C2) DLBCLs harbored frequent biallelic inactivation of *TP53* by mutations and *17p* copy loss (Fig. 5 and Supplementary Fig. 13e). In addition, C2 tumors often exhibited copy loss of *9p21.13/CDKN2A* and *13q14.2/RB1*, which perturb chromosomal stability and cell cycle[2]. Consistent with these findings, transcriptionally profiled C2 DLBCLs had decreased abundance of TP53 targets and increased levels of E2F targets, as determined by GSEA (Supplementary Fig. 13h). C2 tumors also had significantly more driver SCNAs ($P<0.0001$) and a higher proportion of genome doubling events ($P<0.001$; Fig. 6d and Supplementary Fig. 13b). This cluster included both GCB- and ABC-DLBCLs, as did prior DLBCL cohorts with *TP53* mutations in targeted analyses[57]. C2 DLBCLs shared features of previously described DLBCLs with *TP53* alterations and multiple SCNAs of p53/cell cycle modifiers[2]. These tumors also exhibited more frequent copy gains of *1q23.3/MCL1*. Prognostically significant SCNAs, including *13q31.31/miR-17-92* copy gain and *1q42.12* copy loss, were also more common in these DLBCLs (Fig. 5).

Cluster 0. A small subset of 12 DLBCLs lacked defining genetic drivers. Significance analyses (*MutSig2CV* and *GISTIC2.0*) restricted to C0 DLBCLs were also unrevealing. This group included increased numbers of T cell/histocyte-rich LBCLs (Fisher's exact test $P<0.001$), a morphologically defined subtype with a brisk inflammatory/immune cell infiltrate[10]. The absence of detectable drivers

in these DLBCLs may reflect lower tumor purity or different pathogenetic events.

BCL2 and MYC alterations. . Recently, subsets of tumors with co-occurring *BCL2* and *MYC* and/or *BCL6* SVs and/or increased protein expression have been described and associated with poor outcome ('double and triple hit' DLBCLs)[58]. Notably, we detected prognostically significant *MYC* SVs and focal *18q21.33/BCL2* gain (Fig. 5) and additional alterations that perturbed the expression of BCL2, BCL6, and MYC target genes in multiple clusters (*8q* gain, C5; *BCL2* SVs, C3; *13q14.2/miR-15/16* loss, C2; *BCL6* SVs, C1; *13q31.3/miR-17-92* gain[59], C2; Fig. 5). However, tumors with co-occurring *BCL2* and *MYC* SVs were significantly more frequent in C3 DLBCLs (Fisher's exact test, $P=0.003$). These findings identify multiple genetic bases of BCL2 and MYC deregulation and suggest that current definitions of double and triple hit DLBCLs are insufficiently precise.

Temporal ordering of genetic events in DLBCL clusters. We next determined the cancer cell fraction (CCF) for each genetic driver and used a CCF threshold of 0.9 to identify each alteration as being either clonal or subclonal; 74% of mutations, 49% of SCNAs, and 57% of SVs were clonal in this series (Supplementary Fig. 14, Supplementary Table 10a, and Methods). Each of the above-mentioned mutational signatures (Fig. 2) contributed to subclonal mutations, suggesting that all of the mutational processes were ongoing (Supplementary Fig. 14e). We also applied a method for mutation ordering[60] in tumors that harbored pairs of alterations that were clonal and subclonal. Pairs with an excess of clonal to subclonal events were identified and highly significant pairs were highlighted (q value < 0.1; Fig. 6j, Supplementary Fig. 15, Supplementary Table 10, and Methods). Given that clonal alterations occur before subclonal events[60], this method allowed us to order the timing of genetic alterations (Fig. 6e–j).

In C5 ABC-DLBCLs, defining mutations of *CD79B*, *MYD88*, and *TBL1XR1* were largely clonal, whereas additional genetic events, including *18q* copy gain and *PIM1*, *BTG1*, and *ETV6* mutations were more frequently subclonal (Fig. 6i,j). In C1 ABC-DLBCLs, mutations associated with MZL, *NOTCH2*, *SPEN*, and *BCL10*, and immune evasion, *CD70* and *B2M*, were largely clonal, whereas *FAS* and *TNFAIP3* mutations and *BCL6* and PD-1 ligand SVs were often subclonal (Fig. 6e). In informative tumors, the ordering of paired alterations supported the hypothesis that *BCL6* SVs were later, potentially transforming, events (Fig. 6j).

The alterations in C3 GCB-DLBCLs were largely clonal (Fig. 6g), although a subset of *BCL2* SVs were subclonal (Fig. 6g,j). In C4 primarily GCB-DLBCLs, defining alterations of immune evasion molecules, BCR/PI3K signaling intermediates, NF-kB modifiers, and RAS/JAK/STAT pathways members were largely clonal (Fig. 6h). In contrast, mutations of linker and core histone genes were variably clonal and subclonal (Fig. 6h), suggesting that at least some of these alterations were later events.

C2 DLBCLs were largely characterized by clonal loss of *17p*, followed by *TP53* mutations (Fig. 6f,j). Certain prognostically significant genetic alterations, *18q21.33* copy gain and *MYC* SVs, were often subclonal (Figs. 4 and 6, and Supplementary Fig. 14a–d).

Outcome associations of DLBCL clusters. We next assessed the prognostic significance of the newly defined coordinate genetic signatures and identified significant differences in PFS and OS (Fig. 6k,l and Supplementary Fig. 16a,b). Patients with C0, C1, and C4 DLBCLs had more favorable outcomes, whereas those with C3 and C5 tumors had less favorable outcomes (Fig. 6k,l and Supplementary Fig. 16a,b). Notably, in patient with C3 tumors, outcomes were not dependent on co-occurring *MYC/BCL2* SVs (Supplementary Fig. 16e). Patients with C2 DLBCLs had a distinct trajectory and a steady rate of progression over time (Fig. 6k,l and Supplementary Fig. 16a). The genetically distinct COO subtypes (C1 and C5 ABC-DLBCLs; C3 and C4 GCB-DLBCLs) had marked differences in PFS and OS, with more favorable outcomes occuring in the newly defined C1 ABC- and C4 GCB-DLBCLs (Fig. 6m and Supplementary Fig. 16d).

These findings likely explain the reported clinical and genetic heterogeneity in transcriptionally defined COO subsets[9,19–21]. For example, recent targeted studies have identified poor prognosis subsets of ABC DLBCLs with *BCL2* copy gain and GCB tumors with *BCL2* SVs, defining alterations of the genetically distinct C5 ABC and C3 GCB DLBCLs (Figs. 5 and 6m, and Supplementary Fig. 16d)[21].

We next constructed a multivariate model considering both IPI and genetic signatures as variables, with low-risk IPI and favorable (C0/C1/C4) genetic signatures as reference (PFS, Fig. 6n; OS, Supplementary Fig. 16d). For low-risk IPI patients, those with C5 features had a hazard ratio (HR) of 2.01 compared with patients with favorable genetic signatures (Fig. 6n). For patients with favorable genetic features, those with high-risk IPIs had a HR of 3.44 compared with those with low-risk IPIs (Fig. 6n). Patients with C5 features and high-risk IPI had a HR of 6.91 (3.44×2.01) compared with the reference group. Thus, the coordinate genetic signatures captured outcome differences that were independent of the IPI (Fig. 6n, Supplementary Fig. 16c, and Supplementary Table 11).

Discussion

We expanded the landscape of recurrent genetic drivers in DLBCL using increased sample size and technical innovations, including analyses of WES data in the absence of paired normal samples. We also temporally ordered these alterations, gained insight into biologic function of certain mutations by overlaying them onto three-dimensional protein structure and identified the dominant mutational processes in DLBCL exomes. Our results highlight the complexity of DLBCLs, which have a median of 17 different genetic alterations per tumor.

By integrating recurrent mutations, SCNAs, and SVs, we defined five distinct DLBCL subsets, including previously unappreciated favorable risk ABC-DLBCLs with genetic features of an extrafollicular, possibly marginal zone origin (C1); poor risk GCB-DLBCLs with *BCL2* SVs and alterations of *PTEN* and epigenetic enzymes (C3); a newly defined group of good-risk GCB-DLBCLs with distinct alterations in BCR/PI3K, JAK/STAT, and BRAF pathway components and multiple histones (C4); and a COO-independent group of tumors with biallelic inactivation of *TP53*, *9p21.3/CDKN2A* and associated genomic instability (C2). The key genetic features of these DLBCLs included mutations, SCNAs, and SVs, indicating that all three types of alterations are needed to capture disease heterogeneity and outcome differences. Moreover, DLBCL cluster-associated genes were perturbed by multiple mechanisms.

Our approach to define genetically distinct DLBCL subsets represents a framework for assessing previously unrecognized heterogeneity in transcriptionally defined subsets, linking mutational signatures with cluster-predominant pathogenetic mechanisms, assessing genetic bases of extranodal disease tropism, and developing faithful murine models of human tumors. Notably, the DLBCL outcome-associated genetic signatures will guide the development of rational single-agent and combination therapies in patients with the greatest need.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41591-018-0016-8.

Received: 23 February 2018; Accepted: 20 March 2018;
Published online: 30 April 2018

References

1. Basso, K. & Dalla-Favera, R. Germinal centres and B cell lymphomagenesis. *Nat. Rev. Immunol.* **15**, 172–184 (2015).
2. Monti, S. et al. Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer Cell* **22**, 359–372 (2012).
3. Pasqualucci, L. et al. Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat. Genet.* **43**, 830–837 (2011).
4. Morin, R. D. et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298–303 (2011).
5. Lohr, J. G. et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl. Acad. Sci. USA* **109**, 3879–3884 (2012).
6. Morin, R. D. et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood* **122**, 1256–1265 (2013).
7. de Miranda, N. F. et al. Exome sequencing reveals novel mutation targets in diffuse large B-cell lymphomas derived from Chinese patients. *Blood* **124**, 2544–2553 (2014).
8. Reddy, A. et al. Genetic and functional drivers of diffuse large B cell lymphoma. *Cell* **171**, 481–494.e15 (2017).
9. Rosenwald, A. et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346**, 1937–1947 (2002).
10. Monti, S. et al. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood* **105**, 1851–1861 (2005).
11. Ngo, V. N. et al. Oncogenically active MYD88 mutations in human lymphoma. *Nature* **470**, 115–119 (2011).
12. Caro, P. et al. Metabolic signatures uncover distinct targets in molecular subsets of diffuse large B cell lymphoma. *Cancer Cell* **22**, 547–560 (2012).
13. Davis, R. E. et al. Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. *Nature* **463**, 88–92 (2010).
14. Chen, L. et al. SYK inhibition modulates distinct PI3K/AKT- dependent survival pathways and cholesterol biosynthesis in diffuse large B cell lymphomas. *Cancer Cell* **23**, 826–838 (2013).
15. Lenz, G. et al. Oncogenic CARD11 mutations in human diffuse large B cell lymphoma. *Science* **319**, 1676–1679 (2008).
16. Muppidi, J. R. et al. Loss of signalling via Gα13 in germinal centre B-cell-derived lymphoma. *Nature* **516**, 254–258 (2014).
17. Morin, R. D. et al. Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nat. Genet.* **42**, 181–185 (2010).
18. Pfeifer, M. et al. PTEN loss defines a PI3K/AKT pathway–dependent germinal center subtype of diffuse large B-cell lymphoma. *Proc. Natl. Acad. Sci. USA* **110**, 12420–12425 (2013).
19. Lenz, G. et al. Stromal gene signatures in large-B-cell lymphomas. *N. Engl. J. Med.* **359**, 2313–2323 (2008).
20. Dubois, S. et al. Biological and clinical relevance of associated genomic alterations in MYD88 L265P and non-L265P-mutated diffuse large B-cell lymphoma: analysis of 361 cases. *Clin. Cancer Res.* **23**, 2232–2244 (2017).
21. Ennishi, D. et al. Genetic profiling of *MYC* and *BCL2* in diffuse large B-cell lymphoma determines cell-of-origin-specific clinical impact. *Blood* **129**, 2760–2770 (2017).
22. Pfreundschuh, M. et al. Six versus eight cycles of bi-weekly CHOP-14 with or without rituximab in elderly patients with aggressive CD20$^+$ B-cell lymphomas: a randomised controlled trial (RICOVER-60). *Lancet Oncol.* **9**, 105–116 (2008).
23. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
24. Kamburov, A. et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. USA* **112**, E5486–E5495 (2015).
25. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
26. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
27. Pasqualucci, L. et al. AID is required for germinal center-derived lymphomagenesis. *Nat. Genet.* **40**, 108–112 (2008).
28. Chapuy, B. et al. Targetable genetic features of primary testicular and primary central nervous system lymphomas. *Blood* **127**, 869–881 (2016).
29. Georgiou, K. et al. Genetic basis of PD-L1 overexpression in diffuse large B-cell lymphomas. *Blood* **127**, 3026–3034 (2016).
30. Scott, D. W. et al. TBL1XR1/TP63: a novel recurrent gene fusion in B-cell non-Hodgkin lymphoma. *Blood* **119**, 4949–4952 (2012).
31. Challa-Malladi, M. et al. Combined genetic inactivation of β2-Microglobulin and CD58 reveals frequent escape from immune recognition in diffuse large B cell lymphoma. *Cancer Cell* **20**, 728–740 (2011).
32. Green, M. R. et al. Integrative analysis reveals selective 9p24.1 amplification, increased PD-1 ligand expression, and further induction via JAK2 in nodular sclerosing Hodgkin lymphoma and primary mediastinal large B-cell lymphoma. *Blood* **116**, 3268–3277 (2010).
33. Steidl, C. et al. MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature* **471**, 377–381 (2011).
34. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* **101**, 4164–4169 (2004).
35. Dierlamm, J. et al. Gain of chromosome region 18q21 including the MALT1 gene is associated with the activated B-cell-like gene expression subtype and increased BCL2 gene dosage and protein expression in diffuse large B-cell lymphoma. *Haematologica* **93**, 688–696 (2008).
36. Lenz, G. et al. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc. Natl. Acad. Sci. USA* **105**, 13520–13525 (2008).
37. Pham-Ledard, A. et al. High frequency and clinical prognostic value of MYD88 L265P mutation in primary cutaneous diffuse large B-cell lymphoma, leg-type. *JAMA Dermatol.* **150**, 1173–1179 (2014).
38. Rovira, J. et al. MYD88 L265P mutations, but no other variants, identify a subpopulation of DLBCL patients of activated B-cell origin, extranodal involvement, and poor outcome. *Clin. Cancer Res.* **22**, 2755–2764 (2016).
39. Rossi, D. et al. The coding genome of splenic marginal zone lymphoma: activation of NOTCH2 and other pathways regulating marginal zone development. *J. Exp. Med.* **209**, 1537–1551 (2012).
40. Spina, V. et al. The genetics of nodal marginal zone lymphoma. *Blood* **128**, 1362–1373 (2016).
41. Zhang, Q. et al. Inactivating mutations and overexpression of BCL10, a caspase recruitment domain-containing gene, in MALT lymphoma with t(1;14)(p22; q32). *Nat. Genet.* **22**, 63–68 (1999).
42. Kiel, M. J. et al. Whole-genome sequencing identifies recurrent somatic NOTCH2 mutations in splenic marginal zone lymphoma. *J. Exp. Med.* **209**, 1553–1565 (2012).
43. Flossbach, L. et al. BCL6 gene rearrangement and protein expression are associated with large cell presentation of extranodal marginal zone B-cell lymphoma of mucosa-associated lymphoid tissue. *Int. J. Cancer* **129**, 70–77 (2011).
44. Zucca, E., Bertoni, F., Vannata, B. & Cavalli, F. Emerging role of infectious etiologies in the pathogenesis of marginal zone B-cell lymphomas. *Clin. Cancer Res.* **20**, 5207–5216 (2014).
45. MacLennan, I. C. et al. Extrafollicular antibody responses. *Immunol. Rev.* **194**, 8–18 (2003).
46. Erdmann, T. et al. Sensitivity to PI3K and AKT inhibitors is mediated by divergent molecular mechanisms in subtypes of DLBCL. *Blood* **130**, 310–322 (2017).
47. Sun, Z. et al. PTEN C-terminal deletion causes genomic instability and tumor development. *Cell Reports* **6**, 844–854 (2014).
48. Ortega-Molina, A. et al. The histone lysine methyltransferase KMT2D sustains a gene expression program that represses B cell lymphoma development. *Nat. Med.* **21**, 1199–1208 (2015).
49. Boice, M. et al. Loss of the HVEM tumor suppressor in lymphoma and restoration by modified CAR-T cells. *Cell* **167**, 405–418.e413 (2016).
50. Ying, C. Y. et al. MEF2B mutations lead to deregulated expression of the oncogene BCL6 in diffuse large B cell lymphoma. *Nat. Immunol.* **14**, 1084–1092 (2013).
51. Zhang, J. et al. The CREBBP acetyltransferase is a haploinsufficient tumor suppressor in B-cell lymphoma. *Cancer Discov.* **7**, 322–337 (2017).
52. Krysiak, K. et al. Recurrent somatic mutations affecting B-cell receptor signaling pathway genes in follicular lymphoma. *Blood* **129**, 473–483 (2017).
53. Béguelin, W. et al. EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. *Cancer Cell* **23**, 677–692 (2013).
54. Li, H. et al. Mutations in linker histone genes HIST1H1 B, C, D, and E; OCT2 (POU2F2); IRF8; and ARID1A underlying the pathogenesis of follicular lymphoma. *Blood* **123**, 1487–1498 (2014).
55. Okosun, J. et al. Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. *Nat. Genet.* **46**, 176–181 (2014).
56. Yang, S. M., Kim, B. J., Norwood Toro, L. & Skoultchi, A. I. H1 linker histone promotes epigenetic silencing by regulating both DNA methylation and histone H3 methylation. *Proc. Natl. Acad. Sci. USA* **110**, 1708–1713 (2013).
57. Xu-Monette, Z. Y. et al. Mutational profile and prognostic significance of TP53 in diffuse large B-cell lymphoma patients treated with R-CHOP: report from an International DLBCL Rituximab-CHOP Consortium Program Study. *Blood* **120**, 3986–3996 (2012).

58. Sesques, P. & Johnson, N. A. Approach to the diagnosis and treatment of high-grade B-cell lymphomas with MYC and BCL2 and/or BCL6 rearrangements. *Blood* **129**, 280–288 (2017).
59. Li, Y., Choi, P. S., Casey, S. C., Dill, D. L. & Felsher, D. W. MYC through miR-17-92 suppresses specific target genes to maintain survival, autonomous proliferation, and a neoplastic state. *Cancer Cell* **26**, 262–272 (2014).
60. Landau, D. A. et al. Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525–530 (2015).

Acknowledgements

We thank all of the members of the Broad Institute's Biological Samples Genetic Analysis Genome Sequencing Platforms. In addition, we thank all of the patients and their physicians for trial participation and donating the samples. This work was supported by a Claudia Adams Barr Program in Basic Cancer Research (B.C.), a Medical Oncology Translational Grant Program (B.C.), two LLS Translational Research Awards (M.A.S.), and the Lymphoma Target Testing Center (M.A.S.). The computational work for this study was supported by grants U54HG003067, P01CA163222, R01CA18246, U24CA143845, U24CA210999, and R01CA155010 from the National Cancer Institute and the National Human Genome Research Institute, as well as Leukemia & Lymphoma Society grant 0812-14. The Mayo group was supported by a grant from the US National Institutes of Health (P50 CA97274). R.S., M.L., and L.T. received Funding from BMBF (Federal Ministry of Research, Germany; Kennzeichen FZK 031A428B and FZK 031A428H). The Ricover60 Trial was supported by a research grant from Deutsche Krebshilfe (M.P.).

Author contributions

B.C., C.S., G.G., and M.A.S. conceived the project and provided leadership. B.C., C.S., A.D., J.K., A.K., R.R., M.L, A.J.L., G.G., and M.A.S analyzed the data. M.G.M.R., M.Z., A.M.S., J. W., M.D.D., I.L., E.R., A.T.-W, C.C., J.H., C.P., D.L., D.R., M.R., A.T., H.H., P.v.H., A.L.F., B.R.L., A.J.N., J.R.C., T.M.H., R.S., A.R., A.R.T., M.M., T.R.G., R.B., G.G.W., G.O., S.J.R., S.M., D.N., M.L., M.P., and L.T. contributed to the analysis and scientific discussions. B.C, C.S., A.D., G.G., and M.A.S. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0016-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to G.G. or M.A.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Patient samples. Our multi-institutional, international group assembled a cohort of 351 patient samples diagnosed with a previously untreated, primary DLBCL of which 304 passed all below described quality controls. This 304 sample dataset was obtained from 4 sources: 129 samples from patients enrolled in the prospective, randomized, multi-center RICOVER60 trial[22]; 103 samples from a DFCI/BWH cohort; 67 samples from the Mayo Clinic and University of Iowa Specialized Program of Research Excellence (SPORE) (51 previously reported WES analysis[5,61]); and 5 samples from the University of Göttingen, Germany. Forty-four percent (135 of 304) of samples had a paired normal specimen and 55% (168/304) of samples were obtained from formalin-fixed paraffin embedded (FFPE) tissue (Supplementary Fig. 1 and Supplementary Table 1). All patients had a diagnosed primary DLBCL per WHO criteria; this diagnosis was confirmed for all RICOVER60 samples by a central pathological review as previously described[22], and all DFCI/BWH and Mayo cases were confirmed by an expert hematopathologist (SJR). The patient characteristics are equally distributed across the different sources and summarized in Supplementary Table 2. A total of 85% (259/304) of patients were uniformly treated with state-of-the-art therapy (rituximab-containing CHOP-like regimen) and had long-term follow-up (median: 78.5 months). This study was approved by the institutional review board (IRB) of the Dana-Farber Cancer Institute and the IRBs of all other participating institutions. All relevant ethical regulations were followed. Informed conset was obtained from the human subjects on clinical trial. Per IRB protocol and approval, written human subject consent was waived for the additional samples.

Whole-exome sequencing. *DNA quality control.* Tumor and normal DNA was extracted as previously described from lymph node samples, blood and 31 B cell lymphoma cell lines, respectively[2,5]. DNA quality control was performed as previously described[62]. Briefly, genomic DNA was quantified using Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher Scientific) and identities of all tumor/normal DNA pairs were confirmed by mass spectrometric fingerprint genotyping of common SNPs.

Exome sequencing. Whole exome capture was performed using the Agilent SureSelect Human All Exon 44 Mb v2.0 bait set (Agilent Technologies) as previously described[28,63,64]. In summary, genomic DNA was sheared, end repaired, ligated with barcoded Illumina sequencing adapters, amplified, size selected and subjected to in solution hybrid capture using the Agilent SureSelect Human All Exon v2.0 bait set[63,64]. Resulting exome Illumina sequencing libraries were then qPCR quantified, pooled, and sequenced with 76 base paired-end reads using Illumina GAII or HiSeq 2000 sequencers (Illumina). In addition, raw sequencing reads of previously in house generated and published WES data for 49 DLBCL tumor/normal paired samples[5] were processed through identical pipelines as the newly generated WES data (Supplementary Figs. 2a and 8a). Exome sequencing of cell lines with the spiked-in bait set for SV detection was performed as previously described[28]. The new WES data has been deposited in the dbGAP database (www.ncbi.nlm.nih.gov/gap) with the accession number phs000450.v1.p1.

Alignment and quality control. To prepare read alignments for analysis, we processed all sequence data through the Broad Institute's data processing pipeline, Picard (http://picard.sourceforge.net/) as previously described[28]. For each sample, this pipeline combines data from multiple libraries and flow cell runs into a single BAM file. This file contains reads aligned to the human genome with quality scores recalibrated using the TableRecalibration tool from the Genome Analysis Toolkit (GATK)[65]. Reads were aligned to the Human Genome Reference Consortium build 37 (GRCh37) using BWA (version 0.5.9-tpx http://bio-bwa.sourceforge.net/). Variant detection and analysis of the BAM files were performed using the Broad Institute's Cancer Genome Analysis infrastructure program Firehose (http://archive.broadinstitute.org/cancer/cga/firehose). Firehose facilitates comparison of BAM files from matched tumor/normal pairs and coordinates the execution of specific modules including quality control, local realignment, mutation calling, small insertion and deletion identification, rearrangement detection, variant annotation, computation of mutation rates and calculation of sequencing metrics. Module versioning and logging of the specific analytical parameters is also tracked. The median sequencing depth of the exome region in the tumor samples meeting all quality control cut offs is 87.6× (range: 39–206.8). Additional quality control, see Supplementary Note.

Copy number analysis from WES data. Initial estimates of exome-wide copy number profiles were determined using ReCapSeg[66] which creates a copy number profile based on coverage across the exome and a panel of normals which obviates the need for a paired normal. The allele-specific copy number was determined using Allelic Capseg as previously described[67,68]. For paired samples, Allelic Capseg called heterozygous sites from the paired normal, while for the tumor-only samples heterozygous sites were called from the tumor itself. While this method has lower sensitivity for discovering sites with loss of heterozygosity (LOH) in the tumors, when paired samples are run with this method, they show high fidelity to the results when run with a paired normal (Supplementary Fig. 3g).

Significance analysis of recurrent SCNAs using GISTIC2.0. Arm-level and focal peaks of recurrent copy number alterations were identified from the results of *Allelic Capseg* using *GISTIC2.0* (version 129) as previously described[69]. Regions with germline copy number variants were excluded from the analysis. Events with a q-value of less than 0.1 were reported significant. We specified a 99% confidence interval to determine wide peak boundaries.

Mutation calling. Somatic SNVs and small insertions and deletions (Indels) were identified using MuTect (Firehose CallSomaticMutations v131 (ref.[70]), and Indelocator (Firehose *CallIndelsPipeline* v77 (ref.[62])), respectively. When a paired normal was not available, we chose a normal sample from our DLBCL cohort that showed no evidence of tumor in normal contamination and otherwise acceptable QC metrics to remove common germline and potentially remove artifacts resulting from batch effect. Mutations were annotated using the oncotator tool (v68)[71]. Of note, we detected a total of 67,518 unfiltered mutations in tumor samples with a paired normal and 364,692 in samples without a paired normal. Stringent filtering as described below reduced the numbers to 20,328 and 31,586 for samples with and without paired normal, respectively. All significant analyses (MutSig2CV, CLUMPS, SignatureAnalyser tool) were performed on the filtered MAF file. The True-Positive-Rate = Sensitivity (= detected true mutations / all true mutations) for MuTect in tumor/normal (TN) pairs is above 90% in a blind simulated competition among algorithms called Dream challenge 3 (https://www.synapse.org/#!Synapse:syn312572/wiki/63089). For our tumor-only pipeline, the sensitivity is higher than 90% relative to TN pair detection (Supplementary Fig. 2g).

Artifact filtering. OxoG-artifacts were filtered as previously described[72]. In brief, OxoG is an artifact signature results from oxidative damage to guanine during library preparation, which causes guanine to pair with adenine instead of cytosine, ultimately causing an observed G > T mutation. These artifacts will only occur on one strand whereas a somatic event will show the change on both strands of DNA, and this orientation bias is used to distinguish real events from artifacts. This cohort also had single nucleotide artifacts resulting from the use of FFPE samples, wherein formaldehyde causes deamination of cytosine resulting in C > T mutations similar to that of the aging signature but with the same orientation bias observed in OxoG events, allowing us to use the same algorithm for determining orientation bias which has previously been used on FFPE samples[73]. In addition to the canonical OxoG and FFPE artifacts, this cohort had an artifact characterized by recurrent mutations in repetitive regions that have many potential sites for mapping in the genome. To control for this, we first realigned SNV-containing regions with Novoalign v3.02.08 (http://novocraft.com) and preserved those variants that showed evidence in both sets of BAMs[74]. Subsequently, SNV- and Indel-containing regions were reassembled using an approach similar to that of Haplotype caller[65,75] (https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_hellbender_tools_walkers_mutect_Mutect2.php). We rejected variants in regions with sufficient coverage after reassembly that did not have evidence of an alternate allele.

Panel of normals (PoNs) filtering. To remove sequencing artifacts and frequent germline events (for tumor-only samples), SNVs and Indels were filtered using version 8 of the in-house PoNs, which includes 8,334 WES normals[74]. Briefly, the panel includes for each site eight values, which describe the percent of normals, different modes of artifact, and the likelihood that the event is a germline event at that site.

Estimation of purity, ploidy, and cancer cell fraction (CCF) using ABSOLUTE. For paired samples, purity, ploidy, and CCF estimates for mutations and copy number were determined applying the ABSOLUTE algorithm as previously described[76]. Candidate models were reviewed by three independent reviewers (B.C., A.J.D., and C.S.) and discordances in the solution picks were resolved by discussion. ABSOLUTE models based on AllelicCapseg results and mutation calls from tumor-only samples were similarly reviewable to those that came from paired samples. Due to the prevalence of heterozygous germline sites in the mutations going into ABSOLUTE, the solutions called were more driven by the ABSOLUTE copy number profile than the allele frequency distribution in tumor-only samples than for paired samples. However, when ABSOLUTE solutions were called, independently, on 147 available paired lymphoma samples and those same sample samples run without pairs, there was a high correlation in calls of ploidy and purity (Supplementary Fig. 3e,f).

Germline somatic logodds filter for tumor-only samples. For each event that passed all preceding filters (SNV or Indel), its CCF, purity, ploidy, and local copy number were used to determine the log ratio of the probability that its allele fraction is consistent with the allele fraction modeled for a hypothetical germline event and the probability it is consistent with a modeled somatic event. For additional details, see Supplementary Note.

ExAC filtering. After applying the Germline Somatic Log odds filter, we used the ExAC database as a final criterion for excluding potential germline events[77]. Using 147 paired non-hypermutator samples, we selected the allele frequency in

ExAC that yielded 98% sensitivity which cut out 50% of the remaining putative germline events.

Significance analysis of recurrently mutated genes (MutSig2CV). Significantly mutated genes were identified applying the MutSig2CV algorithm and genes with a q-value of less than 0.1 were reported as significant[23]. Notably, with the increased background mutation rate from 3.3/MB to 6.6/MB, the power to detect CCGs present in 10% of patients dropped from 100 to 98% in tumor-only samples.

Measuring the effect of remaining germline events on determination of significant mutated genes using the tumor-only pipeline. To evaluate the performance of the newly developed tumor-only pipeline, the paired normals of our DLBCL cohort were run as tumor-only samples through the tumor-only pipeline as a null model, using one of the paired normal as the 'normal' for the others, leaving us with a total of 134 samples run through this pipeline. Despite the size of the cohort, when running Mutsig2CV on these samples after all filtering 0 to 3 (assigning each normal its paired tumor's purity, assuming 10% or 90% purity) significant genes were found (Supplementary Fig. 3f), suggesting that any germline sites remaining after this pipeline are most likely randomly distributed throughout the genome and unlikely to affect the significantly mutated genes detected by Mutsig2CV[60]. In addition, we performed a beta binomial test to determine if the number of mutations from tumor-only samples occurring in SMGs was significantly overrepresented. The *P* value was calculated as

$$P = \sum_{MTO}^{MTO+MTN} \beta b(x, MTO + MTN, NTO + 1, NTN + 1) = 0.41$$

where βb is the beta-binomial probability density function, NTO is the number of tumor-only samples ($NTO = 169$), NTN is the number of tumor-normal paired samples ($NTN = 134$), MTO is the number of non-silent SMGs detected in tumor-only samples ($MTO = 1{,}516$), and MTN is the number of non-silent SMGs detected in tumor-normal paired samples ($MTN = 1{,}033$).

Targeted DNA-sequencing for the detection of chromosomal rearrangements. *Library construction, sequencing, and pre-analysis processing.* Targeted rearrangements (Supplementary Table 5a) were captured from either leftover uncaptured libraries from WES or genomic DNA, sequenced using an Illumina sequencing platform, de-multiplexed and aligned to the reference sequence b37 edition from the Human Genome Reference Consortium with bwa as described previously[28]. A total of 296 of 304 samples had a mean read depth is 221.4× and met all quality control checkpoints and 99% of samples had a power greater than 0.996 to detect chromosomal rearrangements.

Chromosomal rearrangement pipeline. Somatic rearrangements were detected using four different calling algorithms, BreaKmer[78], Lumpy[79], dRanger and SvABA[80], followed by Breakpointer validation, filtering and a CCF estimation module (Supplementary Fig. 8a), as described in Supplementary Note.

Consensus clustering of genetic alterations. *Generation of gene sample matrix.* All significant mutated genes (MutSig2CV and CLUMPS, q value ≤ 0.1 and frequency ≥ 3%), significant regions of SCNAs (GISTIC2.0, q value ≤ 0.1 and frequency ≥ 3%) and chromosomal rearrangements (frequency ≥ 3%) were assembled into a gene sample matrix (Supplementary Table 8a; non-synonymous mutations, 2; synonymous mutations, 1; no-mutation, 0; high-grade CN gain [CN ≥ 3.7 copies], 2; low-grade CN gain [3.7 copies ≥ CN ≥ 2.2 copies], 1; CN neutral, 0; low-grade CN loss [1.1 ≤ CN ≤ 1.6 copies], 1; high-grade CN loss [CN ≤ 1.1 copies], 2; chromosomal rearrangement present, 3; chromosomal rearrangement absent, 0; chromosomal rearrangements not assessed, na).

Assessing bias in individual genetic alterations due to remaining germline and FFPE artifacts. Fisher's exact test was applied to each putative genetic driver alteration in the gene sample matrix to determine if any of the putative genetic drivers occurred more than expected by random chance in tumor-only samples compared to patient-matched tumor-normal samples. This analysis revealed no outliers after FDR correction, suggesting that there is not a strong bias of remaining germline effect in the discovery of CCGs (Supplementary Table 3e and Supplementary Fig. 3g). The same Fisher's exact test was applied to assess if a putative driver is overrepresented in FFPE tissue compared to fresh-frozen tissue. After calculating the false discovery rate using the Benjamini-Hochberg, one focal amplification, 21q22.3, was highly significant and the 15 focal amplifications were exclusively found in FFPE samples (Supplementary Table 3f and Supplementary Fig. 3m). To further investigate the quality of this focal peak, the distribution of the difference in amplitude of adjacent targets as a noise measurement was plotted against other focal peaks (Supplementary Fig. 3o), where the distribution was found to be more irregular and to have the highest s.d. The higher noise level of the focal amplification 21q22.3, combined with the fact that it only appeared in FFPE samples and the event was exclusively subclonal served as justification for removal of the event as a likely FFPE artifact. After the removal of this event, no other genetic alterations were significantly overrepresented in FFPE after false discovery rate correction (Supplementary Table 3f and Supplementary Fig. 3n).

Non-negative matrix factorization consensus clustering. To robustly identify tumors with shared genetic features, we applied a non-negative matrix consensus clustering algorithm[34] with slight modifications. Briefly, we passed the gene sample matrix containing mutations, SCNAs and chromosomal rearrangements (Supplementary Table 8a) to the NMF consensus clustering algorithm (input parameters k = 4–10) bypassing the matrix normalization so that the cluster distance metric depended directly on the variant number in the gene-sample matrix. The NMF consensus clustering algorithm provided the cluster membership of each sample, the cophenetic coefficient for k = 4 to k = 10 clusters and silhouette values for the 'Best cluster' (k = 5) (Supplementary Table 8b). Samples without genetic drivers in the input matrix to the clustering were assigned to cluster C0. In addition, we identified marker genes associated with each cluster by applying a fisher test (2 × 2 table with variant present or absent as one dimension and within-cluster or outside-cluster the second dimension) and corrected the *P* values using the FDR procedure (Supplementary Table 8c). Features with a q value ≤ 0.1 were selected as cluster features (Supplementary Table 8c) and visualized as a color-coded heatmap using GENE-E (Fig. 5 and Supplementary Fig. 12; https://software.broadinstitute.org/GENE-E/)

Mutual exclusivity/co-occurrence estimations. For each gene of interest, the significance of the co-occurrence or mutual exclusivity for each pair of different events (mutations, amplification, deletion or structural variant) that affects that gene was calculated using a Fisher test, and then corrected for false discovery using the Benjamini-Hochberg method.

Inferred timing of genetic alterations. *CCF matrix of putative drivers.* First, we assembled for each of the 158 candidate driver events (for criteria, see generation of gene sample matrix above) the cancer cell fraction. When multiple events appeared in the same patient, the estimate based on the event with the highest coverage was used for mutations and SVs, while the one based on the longest segment was used for copy number alterations, as in each case this should represent the best-measured estimate (Supplementary Table 10a). In addition to the actual CCF value, for each genetic feature we added a binary distinction if this is clonal or subclonal alteration with 0.9 being the threshold.

Event ordering analysis. To infer the timing of genetic events in each cluster and the overall cohort, we applied the method previously described for mutation ordering[60]. Briefly, we first identified for all driver alterations event pairs where events occurred such that one event was subclonal and the other was clonal (Supplementary Table 10b). The 'effect-size' to quantify alteration pairs according to clonal and sub-clonal mixtures is simply the difference in counts of clonal and subclonal samples. Next, we assumed a null model in which the timing of genetic events was random, allowing us to perform a formal binomial test to determine if one event was more frequently clonal than the null model (Supplementary Table 10c). Of note, we restricted the test to those event pairs that were powered to achieve a significant result (q value ≤ 0.1) when occurring as maximal effect.

Clinical endpoint analyses. Statistical analyses were performed using R v3.3.2 with additional packages survival v2.41-2 for survival analyses, qvalue v2.6.0 for false discovery rate control, and knitr v1.15.1 for reproducible research.

OS was defined as time from treatment until death from any cause. Subjects not confirmed dead were censored at the time last known to be alive. PFS was defined as time from treatment until the earliest time of progression or death from any cause, and censored at time last known to be alive and free of progression.

Univariate and multivariable analyses of time-to-event endpoints were performed on the R-CHOP treated cohort ($n = 259$) using Cox regression. Genetic features had to be present in at least 3% of samples of the R-CHOP treated cohort to be tested for outcome associations. HRs with 95% confidence intervals (CI) and Wald *P* values were reported for model covariates; likelihood-ratio tests and *P* values were reported for multivariable models. Log-likelihoods of nested models were compared using a chi-square test to assess improvement in model fits. Median event times were estimated using the method of Kaplan and Meier (KM) and reported with 95% CIs; Greenwood's formula was used to approximate the variance of KM estimate, and 95% CIs were generated using the log-log transformation. Differences in survival curves were assessed using log-rank tests. Median follow-up time was estimated using the reverse KM method.

Fisher's exact test was used to test for association between categorical variables. ORs and 95% CIs were calculated for binary outcomes from contingency tables or logistic regression for continuous predictors. The Wilcoxon or Kruskal-Wallis rank-sum test was used to assess a location shift in the distribution of continuous variables between two or more than two groups, respectively. Descriptive statistics (proportions, medians, etc.) were reported with 95% exact binomial CIs or range. All *P* values were two-sided, and adjustments for multiple hypothesis testing was performed using the method of Benjamini and Hochberg; *P* and q value thresholds for significance were set at 0.05 and 0.2, respectively.

Additional methods. Additional quality control metrics, detailed descriptions of the estimation of and correction for tumor-in-normal content (deTiN), the germline somatic log odds filter for tumor-only samples, the clustering and

visualization of mutations in protein structures (CLUMPS) method, the correlation between driver genes and GISTIC2.0 peaks, the assessment of chromothripsis, the mutational signature analysis, the integrative analysis of gene expression and copy number data, the description of the chromosomal rearrangement pipeline, and the immunohistochemical staining protocol for PD-1 ligands are described in Supplementary Note.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary.

Code availability. Data processing was done in the Broad Firehose computing environment (http://archive.broadinstitute.org/cancer/cga/firehose). Code for modules from firehose as well as visualization and post-processing scripts are available upon request. The custom code for the NMF consensus clustering is available through GitHub at https://github.com/broadinstitute/DLBCL_Nat_Med_April_2018.

Data Availability. Sequence data that support the findings of this study is being deposited in the dbGAP database (www.ncbi.nlm.nih.gov/gap), accession number phs000450. Newly generated U133plus2 Affymetrix gene expression array data has been uploaded to GEO, accession number GSE98588. All the data are available within the article, supplementary information and supplementary data file or from the authors on request.

References

61. Novak, A. J. et al. Whole-exome analysis reveals novel somatic genomic alterations associated with outcome in immunochemotherapy-treated diffuse large B-cell lymphoma. *Blood Cancer J.* **5**, e346 (2015).
62. Chapman, M. A. et al. Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).
63. Fisher, S. et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
64. Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
65. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
66. Lichtenstein, L., Wood, B., MacBeth, A., Birsoy, O. & Lennon, N. ReCapSeg: Validation of somatic copy number alterations for CLIA whole exome sequencing. *Cancer Res.* **76** Supplement, abstr. 3641 (2016).
67. Giannikou, K. et al. Whole exome sequencing identifies TSC1/TSC2 biallelic loss as the primary and sufficient driver event for renal angiomyolipoma development. *PLoS Genet.* **12**, e1006242 (2016).
68. Burger, J. A. et al. Clonal evolution in patients with chronic lymphocytic leukaemia developing resistance to BTK inhibition. *Nat. Commun.* **7**, 11589 (2016).
69. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
70. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
71. Ramos, A. H. et al. Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).
72. Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
73. Giannakis, M. et al. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Reports* **17**, 1206 (2016).
74. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676–690 (2014).
75. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
76. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
77. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
78. Abo, R. P. et al. BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers. *Nucleic Acids Res.* **43**, e19 (2015).
79. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
80. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28,** 581–591 (2018).

nature research

Corresponding author(s): Margaret A Shipp
Gad Getz

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Data was collected and stored using MS Excel v16, R v3.3.2 with the knitr v1.15.1 package for reproducible research, Mathlab 2013a and the Firehose 2.19.3 workspace.
Data analysis	Detailed descriptions of our analytical pipelines are provided in Online Methods and Supplemental Information, listing for each step the version, algorithm and parameters used. A code availability statement has been added to the Online Methods and custom code has been made available through GitHub as indicated in the code availability statement. We used this software: Firehose 2.19.3 Matlab R2013a (8.1.0.604) BWA v0.5.9 ContEst Queue v1.4-437-g6b8a9e1 Coverage/Depth tool Firehose task GlobalCoverageByZone v23 TN swap tool Firehose task CrosscheckLaneFingerprintsPipeline v16 MuTect1 v1.1.6 MuTect2 v3.6-97-g881c5e9 Indellocator Firehose task CallIndelsPipeline v77 ReCapSeg Firehose tasks ReCapSegCoverage v20 and recapseg_tumor_pcov v34 AllelicCapseq Firehose task AllelicCapseg v22 OxoG Filter Firehose task oxoGFilter_v3 v62 FFPE Filter Firehose task OrientationBias_filter v1

PoN filtering Firehose task maf_pon_filter v23
ABSOLUTE v1.5
Logodds Tumor-only filter Firehose task Filter_For_Tumor_Only_Samples v10
deTin Firehose tasks TumorInNormalEst v85 and deTiN_allele_shift v43
Signature Analyzer v1.1
NMF consensus clustering custom script uploaded to GitHub
MutSig1 v.15
MutSig 2CV v2CV
GISTIC2.0 v2.0
CLUMPS v1
dRanger Firehose task dRanger v199
Breakpointer Firehose task BreakpointerFromBPFile v1
SVaBA Firehosetask Snowman v102
Lumpy v0.2.11
BreaKmer v0.0.6
MutationMapper v1.0.1
Pymol v1.8.0.5
IGV v2.4.9
Circos-069-6.tgz
GENE-E v3.0.215
Graph Pad Prism 7.0c
GSEA v3.0
R-Studio Version 1.0.153
R v3.3.2 with these packages:
survival v2.41-2
qvalue v2.6.0
knitr v1.15.1
bioconductor v3.6
limma v3.34.9
iEDGE
ggplot2 v2.2.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data Availability
Sequence data that support the findings of this study is being deposited in the dbGAP database (www.ncbi.nlm.nih.gov/gap), accession number phs000450.v1.p1. Newly generated U133plus2 Affymetrix gene expression array data has been uploaded to GEO, accession number GSE98588. All the data are available within the article, supplementary information and supplementary data file or from the authors on request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	1.) Composition and description of the cohort: Fig. S1 and Online Methods. 2.) With 304 tumors and the identified background mutation rate, we have >98% power to detect candidate cancer genes (CCGs) in at least 10% of patients (http://www.tumorportal.org/power).
Data exclusions	A total of 47 samples were omitted due to quality control concerns. Please see Supplemental Information page 52 for details.

Replication	Immunohistochemistry for indicated antibodies in Figure 3 are repeated twice with similar results.
Randomization	Training and test set allocation for evaluating the tumor-only filter was performed randomly.
Blinding	Detection of genetic clusters was performed independent of and blinded to clinical endpoints.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Research animals
☐	☒ Human research participants

Antibodies

Antibodies used	anti PD-L1, Cell Signaling, mAb #29122, clone 405.9A11 anti PD-L2, EMD Millipore, MABC1120, clone 366C.9E5 anti PAX5, BD Biosciences, 610863, clone 24/Pax-5
Validation	The indicated antibodies are all validated for use in human tissue for immunohistochemistry (see datasheets on the homepage of the respective manufactures). In addition, we have validated and used these antibodies in the past: anti-PD-L1 and anti-PAX5: -> Roemer et al., JCO 2016 Aug 10;34(23):2690-7. doi: 10.1200/JCO.2016.66.4482 and Roemer et al., JCO 2018 Feb 2:JCO2017773994. doi: 10.1200/JCO.2017.77.3994. anti PD-L2: -> Chapuy et al., Blood 2016 Feb 18;127(7):869-81. doi: 10.1182/blood-2015-10-673236.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Cell lines were obtained from indicated cell banks. For cell lines not available through these cell banks, the Shipp lab is one of the largest resources of STR profiled large B-cell lymphoma cell lines. ATCC (Pfeiffer,TOLEDO), DSMZ (DB,DHL10,DHL16,DHL4,DHL5,DHL6, DHL8,HT,K422,Ly19,Ly1,LY3,Ly7, WSU-DLCL2,WSU-NHL,DoHH2,SC1,WSU-FSCCL), JCRB (TK), RCB (CTB-1), Shipp laboratory (Balm3,DHL7,HBL1,Ly10,Ly18,Ly4, Ly8,TMD8,U2932).
Authentication	STR profiling.
Mycoplasma contamination	All cell lines used were negative for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	No

Human research participants

Policy information about studies involving human research participants

Population characteristics	Our multi-institutional, international group assembled a cohort of 351 patient samples diagnosed with a previously untreated, primary diffuse large B-cell lymphoma (DLBCL) of which 304 passed all below described quality controls. This 304 sample dataset was obtained from 4 sources: 129 samples from patients enrolled in the prospective, randomized, multi-center RICOVER60 trial22; 103 samples from a DFCI/BWH cohort; 67 samples from the Mayo Clinic and University of Iowa Specialized Program of Research Excellence (SPORE) (51 previously reported WES analysis5,61); and 5 samples from the University of Göttingen, Germany. Forty-four percent (135/304) of samples had a paired normal specimen and 55% (168/304) of samples were obtained from formalin-fixed paraffin embedded (FFPE) tissue (Supplementary Figure 1 and Supplementary Table 1). All patients had a diagnosed primary DLBCL per WHO criteria; this diagnosis was confirmed for all RICOVER60 samples by a central pathological review as previously described22, and all DFCI/BWH and Mayo cases were confirmed by an expert hematopathologist (SJR). The patient characteristics are equally distributed across the different sources and summarized in Supplementary Table 2. A total of 85% (259/304) of patients were uniformly treated with state-of-the-art therapy (rituximab-containing CHOP-like regimen) and had long-term follow-up (median: 78.5 months). This study was approved by the institutional review board (IRB) of the Dana-Farber Cancer Institute and the IRBs of all other participating institutions. All relevant ethical regulations were followed. Informed conset was obtained from the human subjects on clinical trial. Per IRB protocol and approval, written human subject cosents were waived for the additional samples.

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging