Resolving heterogeneity in Diffuse Large B-cell Lymphoma using a comprehensive modular expression map

Matthew A. Care[1,2*], Daniel Painter[1], Sharon Barrans[3], Chulin Sha[4], Peter Johnson[5], Andy Davies[5], Ming-Qing Du[6], Simon Crouch[1], Alex Smith[1,2], Eve Roman[1,2], Cathy Burton[3,2], Gina Doody[2,7], David Westhead[2,8], Ulf Klein2[2,7], Daniel J. Hodson[9], Reuben Tooze[2,3,7*]

[1]Epidemiology and Cancer Statistics Group, Department of Health Sciences, University of York, York YO10 5DD, UK;

[2]National Institute of Health Research Leeds Biomedical Research Centre

[3]Haematological Malignancy Diagnostic Service, St. James's Institute of Oncology, Leeds LS9 7TF, UK;

[4]Institute of Basic Medicine and Cancer of Chinese Academy of Sciences, Economic Technology Dev Zone, Hangzhou, 310000, Zhejiang province, China;

[5]University of Southampton, Southampton SO16 6YD, UK;

[6]Division of Cellular and Molecular Pathology, Department of Pathology, University of Cambridge, Cambridge CB2 0QQ, UK;

[7]Division of Haematology and Immunology, Leeds Institute of Medical Research, University of Leeds, Leeds, LS9 7TF, UK;

[8]Bioinformatics Group, School of Molecular and Cellular Biology, University of Leeds, Leeds LS2 9JT, UK;

[9]Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, Puddicombe Way, Cambridge CB2 0AW, UK;

Correspondence to:

*Reuben Tooze e-mail: r.tooze@leeds.ac.uk[i]

*Matthew Care e-mail: m.a.care@leeds.ac.uk

Lead contact: Reuben Tooze, Wellcome Trust Brenner Building, Leeds Institute of Medical Research, University of Leeds, Leeds, LS9 7TF, UK, tel: (44)-113-3438639, e-mail: r.tooze@leeds.ac.uk

Abstract

Diffuse large B-cell lymphoma (DLBCL) is characterised by pronounced genetic and biological heterogeneity. Several partially overlapping classification systems exist – developed from mutation, rearrangement or gene expression data. We apply a customised network analysis to nearly five thousand DLBCL cases to identify and quantify modules indicative of tumour biology. We demonstrate that network-level patterns of gene co-expression can enhance the separation of DLBCL cases. This allows the resolution of communities of related cases which correlate with genetic mutation and rearrangement status, supporting and extending existing concepts of disease biology and delivering insight into relationships between differentiation state, genetic subtypes, rearrangement status and response to therapeutic intervention. We demonstrate how the resulting fine-grained resolution of expression states is critical to accurately identify potential responses to treatment.

Significance statement: We demonstrate how exploiting data integration and network analysis of gene expression can enhance the segregation of diffuse large B-cell lymphoma, resolving pattens of disease biology and demonstrating how the resolution of heterogeneity can enhance the understanding of treatment response.

Introduction

Heterogeneity is a characteristic feature of diffuse large B-cell lymphomas (DLBCL). Consistent subtypes have been resolved based on expression state, mutational profiles and cytogenetic features. Amongst the most significant insights has been the association of DLBCL with either germinal centre (GCB) or post-germinal centre/activated B-cell (ABC) counterparts.[1] More recently analysis of mutation patterns in DLBCL have converged onto recurrent patterns of co-mutation that can be used to define genomic subtypes.[2-7] A third approach widely used in clinical practice is separation by gene rearrangement status with identification of high-risk DLBCL cases based on double-hit (DH) or triple-hit (TH) rearrangements of *MYC* and *BCL2* and/or *BCL6* genes.[8-11]

Expression-based classification of DLBCL is not restricted to identification of cell-of-origin (COO) classes, with the parallel consensus cluster classification (CCC) focusing on metabolic, signalling and host response features.[12] The latter have also been identified in separate stromal survival predictors.[13] Furthermore, high-risk DLBCL cases have been identified based on gene expression features, learned either from patterns in Burkitt lymphoma,[14] or based on direct similarity to cases with *MYC* and *BCL2* DH.[15] These approaches identify overlapping sets of cases and are enriched amongst DLBCL associated with mutations of *EZH2* and *BCL2*.[2,4,5]

Single-cell expression analysis of germinal centre (GC) B-cells has provided additional insight, separating features of the two main functional populations of dark and light zones, in which B-cells undergo proliferation, somatic hypermutation and T-cell mediated selection, along with intermediate populations including those transitioning to post-GC differentiation.[16-21] Additionally, the calculated contributions of multiple cell states, including both neoplastic B-cells and accompanying host response has been used to assign DLBCL to ecotypes;[22] and functional expression signatures have been used to define differences in lymphoma microenvironments across both conventional COO classes and genomic DLBCL categories.[23] However, an integrated picture unifying features of expression, mutation and rearrangement status is still developing. Here we address this issue using an approach which resolves the intrinsic structure of gene co-expression. Drawing on a broadly representative resource of close to 5000 DLBCL cases across prior studies, we demonstrate

how the resultant integrated view of DLBCL expression biology enhances the informed selection of features for expression-based classification.

Results

An integrated gene expression context for DLBCL

To provide a comprehensive exploration of DLBCL biology from the perspective of gene expression we employed parsimonious gene correlation network analysis (PGCNA) (Supplemental Figure 1).[24] This approach allows integration of multiple datasets across platforms and makes use of radical edge reduction to efficiently derive the configuration of large networks and optimize modularity.[24,25] In this approach the focus for each gene is on the mostly highly correlated gene partners, hub nodes/genes in the network emerge consequent to being highly correlated partners of many other genes and the primary determinant of resolved modules is the pattern and consistency of gene co-expression in DLBCL data.

We utilised a broadly representative sample of available DLBCL datasets divided into two components. For network discovery we used gene expression derived from 14 DLBCL datasets, restricting to datasets with >50 DLBCL cases, encompassing a total of 2,505 cases.[2,12,13,26-35] The resulting network reflects a representative sample of DLBCL expression data from multiple sources, mitigating against biases linked to case selection and platform features. Recent datasets that combine mutation and expression features, totalling 2,484 cases, were reserved for downstream analysis.[3,5,14,36,37]

The resulting DLBCL expression network resolved into 28 modules of gene co-expression, across 16,054 genes/nodes (Figure 1a, https://mcare.link/DLBCL2, Supplemental Table 1). A comprehensive gene ontology and signature enrichment was used to assess biological features associated with these modules. Representative terms were selected to identify notable expression patterns associated with each module (Figure 1b, Supplemental Figure 2, Supplemental Table 2).

Several features were evident. Classifier genes used for separation of ABC and GCB classes in the original COO classifier divided discretely and exclusively between two modules: M9 containing all GCB classifier genes, and M11 containing all ABC classifier genes (M11 ABC: p-value 6.95×10^{-19} and M9 GCB: 1.57×10^{-11}).[38] These two modules also capture the majority of

B-cell lineage gene expression as assessed by signature enrichment (e.g. SignatureDB pan-B or Resting Blood B-cell). Thus, the fundamental separation of GCB- and ABC-related expression patterns was rediscovered as a primary feature of gene co-expression patterns in the network structure. The more numerous CCC genes divided across several modules but with significant enrichments in specific modules (M9:B-cell receptor p-value $3.27x10^{-57}$, M3:Host response p-value $1.64x10^{-64}$ and M13:Oxphos p-value $1.38x10^{-05}$).[12] Additional modules separated features related to MYC function (M8), cell cycle (M7), glycolysis (M13), and sterol biosynthesis (M19). Host response/microenvironment was divided between modules related to stromal/angiogenesis (M4), T/NK-cells (M10), monocytes/macrophages (M3) and IFN responses (M20). Highly co-ordinated gene batteries were resolved for nucleosome components (M18), homeobox (M23), immediate early (M26) and metallothionein genes (M25). Several modules related to structural chromosomal regions including chr6 (M16), chr7 (M14), chr17p (M17), chrX (M13) and chrY (M24).

The network thus provides an integrated picture of gene expression in DLBCL encompassing details of lineage-specific gene expression set against a diverse backdrop of cell biology and microenvironment and resolving fine-grained patterns of gene co-expression. These are available as an extensive online resource covering gene correlation data, interactive network visualisations and downstream comparisons (https://mcare.link/DLBCL2).

Meta-hazard ratio analysis confirms associations between biology and outcome

A meta-hazard ratio analysis merging overall survival data from training datasets can illustrate the association of module and individual gene expression with outcome across the entire network. We limited analysis to cases treated with R-CHOP immunochemotherapy. The result included the segregation of good and adverse outcome between component genes of the GCB (M9) and ABC (M11) modules (Figure 1c and d), as well as adverse outcome with Mitochondrial and MYC overexpression module (M8) and Ribosome module (M7) and good outcome linked to expression of the Stromal/Angiogenesis module (M4) which encompasses features of known survival predictors.[13]

Resolving detailed differentiation states in GCB and ABC modules

Biological detail can be resolved at different levels of granularity in the network. An iterative analysis of gene correlation within modules can resolve the most highly related gene

neighbourhoods, and the extent to which underlying information is successfully separated by such neighbourhoods can be assessed with gene signature enrichment.[24]

Assessed systematically across the network, the ABC and GCB modules were amongst those with highest neighbourhood-level information content (Supplemental Figure 3, Supplemental Table 3 and 4). Given the central importance of B-cell differentiation to lymphomagenesis,[39,40] we focused on the relationships of the GCB and ABC neighbourhoods in more detail (Figure 2a and b, and https://mcare.link/DLBCL2). The GCB module (M9) resolved into 16 neighbourhoods (Figure 2b and c, Supplemental Table 4). These included neighbourhoods enriched for different GC B-cell subsets - LZ (M9_n2 and n6), DZ (M9_n5 and n15), CD40/NFκB responses linked to LZ B-cells (M9_n10), Intermediate-e/CCR6 memory B-cell precursor (M9_n3 and n4) and pre-memory/memory B-cells (M9_n14).[16] A similar informative separation was observed for the ABC module which divided into 17 neighbourhoods (Figure 2b and d, Supplemental Table 4). This separated expression related to CCR6+ memory B-cell precursors (M11_n5), pre-memory B-cells (M11_n5 and n7), and plasmablast/plasma cell (M11_n6). These cell state associations overlapped with targets of key transcriptional regulators IRF4 (M11_n1, n6 and n7), BLIMP1 (M11_n7) XBP1 (M11_n6), and NFκB-response genes specific to the ABC modules (M11_n4 and n8).

The COO gene expression classifier established the paradigm for the successful application of small numbers of representative genes to separate individual DLBCL cases into distinct expression states.[38] Genes that have a high degree of connectivity in the network may provide particularly useful information to summarise expression states of many correlated neighbours and overlapping information can be derived from different but correlated genes within network neighbourhoods/modules. The latter illustrates how different classifier gene sets can be selected to arrive at broadly similar answers for a given expression state. Many of the COO classifier genes emerged as hub nodes within the respective GCB and ABC neighbourhoods (Figure 2a). Considering the genes used in the original COO classifier, those used to classify GCB-DLBCL were all contained within the M9 module and segregated primarily between two neighbourhoods M9_n3 (*BCL6*, *MME* (*CD10*) and *SERPINA9*) and M9_n4 (*DENND3*, *ITPKB*, *LMO2* and NEK6) while *IRAG2* (*LRMP*) belonged to M9_n11. The COO ABC-classifier genes were distributed across 5 neighbourhoods in module M11 (M11_n1, n2, n3, n4 and n7), illustrating that these genes sample multiple aspects of the

ABC network module (Figure 2a and Supplemental Figure 3c). We conclude that neighbourhood-level patterns of gene expression can resolve details of biology related to differentiation state and can illustrate the inter-relationship of individual genes used in targeted expression-based classifiers.

Network gene expression patterns show consistent association with mutation state

HMRN, REMoDL-B and Reddy datasets include both gene expression and mutation data.[5,36,37] These studies were not used in network generation. We could therefore test whether patterns of gene co-expression in the network correlated with mutations in these independent datasets. Since many mutations in DLBCL are rare we considered both the correlation of expression and mutation within individual datasets and as a consensus meta-mutation correlation across the three datasets. We summarised module- or neighbourhood-level gene expression as metagenes (module or neighbourhood expression values, MEV or NEV) and assessed the correlation of these values with the presence or absence of gene mutation, reducing mutation data for each case into a mutated/unmutated binary call for each gene with annotated putative pathogenic mutations.

Metagene expression patterns were significantly associated with mutation state (Figure 3 and Supplemental Figure 4). At module level the primary separation was into two clusters corresponding to M9_GCB and M11_ABC mutation correlations. M9_GCB correlated significantly with a wide range of mutations consistent with known associations both of COO-defined GCB-DLBCL and follicular lymphoma (FL),[3,41] and anticorrelated with the primary mutations linked to COO-defined ABC-DLBCL. M11_ABC exhibited the reciprocal anticorrelation with mutations linked to COO-defined GCB-DLBCL and positive correlation with mutations characteristic of COO-defined ABC-DLBCL. Modules which shared mutation correlations with M9_GCB included: M13_ChrXGlycolysis, M23_Homeobox, M28_ITGB8, M6_ZincFinger, M6_Chr6; while modules with shared mutation correlations with M11_ABC included: M8_MitochondrionMYC, M18_Nucleosome, M7_CellCyle and M17_Chr17p. An additional cluster of modules focused on shared correlations with *CDKN2A* mutation including M14_Chr7 and M15_Ribosome.

Further subdivision was evident when modules were considered at neighbourhood-level. For instance, while most neighbourhoods in the GCB module (Figure 3 top left panel) shared correlation with a core set of gene mutations including *EZH2, BCL2* and *CREBBP* the

neighbourhood-level analysis could discriminate heterogeneity for other genes. For example, neighbourhoods discretely separated between those with *SOCS1, NFKBIA, SGK1, IRF8, TNFRSF14, S1PR2, CD83*, *STAT6* mutation association and those with *MYC*, *CDKN2A* mutation association. Additionally, mutation in several genes including *BTG2, DDX3X, FOXO1, RHOA, PAX5* or *STAT6* Y419 hotspot mutations associated with select patterns of GCB neighbourhood expression. Similarly, in the ABC module (Figure 3 top right panel) most neighbourhoods shared association with a common set of mutated genes including *MYD88, CD79B, PIM1, ETV6, TBL1XR1, BTG2*, *PRDM1*. However, more selective associations were evident for *BCL10, IRF4, CDKN2A, NOTCH2*, *TMEM30A*. Other M11_ABC neighbourhoods captured quite distinct mutation associations lacking positive correlation with the *MYD88, CD79B* mutation cluster and showing associations with *NFKBIA, SOCS1, SGK1*, *BTG1* or with *BCL2* mutation and features transitional to a GCB-like pattern.

Further examples of informative separation at neighbourhood-level are illustrated by the M3_ImmuneResponseMonocyteEnriched and M7_CellCycle neighbourhoods. For the M3_ImmuneResponseMonocyteEnriched module (Figure 3 bottom left panel) common association was observed for mutation in *CD58*. However, select neighbourhoods (M3_n6, M3_n8 and M3_n11) were also associated with mutation in either *TET2* or *DNMT3A*. For the M7_CellCycle module (Figure 3 bottom right panel), *TP53* mutation, was significantly associated with 12/15 of neighbourhoods which contrasted with its lack of significant association with any M8_GCB or M11_ABC neighbourhoods. Interestingly, 8/12 neighbourhoods linked to *TP53* mutation also share association with mutation of *TMEM30A*, recently identified as a tumour suppressor inactivated in DLBCL, but in contrast to *TP53* linked to a favourable response to R-CHOP therapy.[42] Multiple M7_CellCycle neighbourhoods shared association with mutation of genes such as *MYD88, CD79B, PIM1*, *ETV6* characteristically linked to the ABC-state, while from the perspective of GCB-linked associations select M7_CellCycle neighbourhoods were associated with mutation in *MYC* and *DDX3X*.

While this analysis is limited to the sets of genes tested in targeted mutation panels, the most important known mutational features of DLBCL are included. The results identify distinct and significant associations of network derived gene co-expression patterns with driver mutations and argue that the granularity provided by the network at module and neighbourhood-level is informative of underlying tumour biology.

238

239 **Network-level gene expression patterns enhance consistent segregation of DLBCL**

240 Given the consistent association between modules/neighbourhoods and mutation state we
241 reasoned that network-level information may contribute to enhanced expression-based
242 segregation of DLBCL cases. To evaluate this, we tested how recurrently discoverable the
243 co-segregation of DLBCL cases was with different selections of features and different
244 clustering methods. To select informative features for clustering we split all modules into
245 neighbourhoods and considered features either at the level of individual genes or collapsing
246 genes within neighbourhoods or modules into single values (MEV/NEVs). We then selected
247 the most informative features at either gene, NEV or MEV level using several different
248 approaches. The selected features were then used to cluster the DLBCL datasets (n=16, see
249 Supplemental Table 1, DatasetInfo) using either PGCNA or consensus clustering (CC).[43] The
250 resultant clusterings were compared based on the extent of recurrently discoverable co-
251 segregation of DLBCL cases. In total 33 attribute sets were examined (see Supplemental
252 Table 5). This showed (Figure 4a): i) that selections based on PGCNA network modules with
253 edge information (Net+Strct) outperformed ranking without edge information (Network); ii)
254 that having only M9/M11 at neighbourhood-level (NEV_M9M11) outperformed selections
255 using all neighbourhoods; iii) that collapsing genes per module/neighbourhood to
256 MEV/NEVs outperformed gene-level clusterings; and iv) that using network structure and
257 edge information (Net+Strct) to inform gene choice and collapsing genes to
258 module/neighbourhood values (MEV/NEVs) significantly outperformed gene selection based
259 on most variant expression (Top1000/5000). We conclude from this that use of network
260 information can enhance the selection of attributes for consistent clustering of DLBCL cases.

261

262 **Network information resolves communities of lymphoma cases with shared biology**

263 A common feature of classification based on multiple parameters is the identification of
264 consensus cases on which various classification tools can agree when using the same
265 features for classification, and edge cases where the classification is more ambiguous.[44]
266 Therefore, extending the concept of consensus or ensemble clustering,[12,45] we focused on
267 cases co-clustered by both CC and PGCNA methods. This identified 12 ensemble case
268 clusters (Figure 4b), which we refer to as Lymphoma Communities (LCs). These cases were
269 then used as a reference set to train a machine learning tool. The machine learning tool was

used to reclassify all cases in the HMRN (Figure 4b) and other datasets (see Supplemental Methods for details). A similar distribution of cases falling into these communities was recovered across a range of datasets (Figure 4c and Supplemental Figure 5 & 6). These communities were named according to patterns of module expression or associated pathological/molecular features that were identified in downstream analyses.

Amongst the 12 LCs (Figure 4b, Supplemental Figure 6) two communities identified cases dominated by ABC module gene expression which were distinguished by differences in immune response-related features as immune response poor (**$_{lc}$ABC**) or immune rich (**$_{lc}$ABC-IR**). Cases with plasmablastic features (**$_{lc}$PBL**) were distinguished by expression of XBP1 targets and cancer testis antigens. A distinct subset of cases with ABC-related features shared the expression of cancer testis antigens (**$_{lc}$CT**). Amongst cases with weak B-cell patterns, two communities were dominated by immune response modules (**$_{lc}$IR** and **$_{lc}$IR-TET2 -** distinguished by *TET2* mutation enrichment, see below). Six communities were characterised by expression of multiple GCB-related neighbourhoods. These included a community with GCB features and distinct stromal/EMT-related gene expression (**$_{lc}$GCB-EMT**) and a community with mixed GCB and ABC expression features along with cell cycle, MYC, and sterol biosynthesis modules, suggestive of transition from GCB to an ABC/post-GC state (**$_{lc}$GCB-Xit**, Supplemental Figure 6). Similar proliferation and growth-related module expression were combined with polarised high GC DZ and low GC LZ neighbourhood expression, along with low CD40/NFκB and MHC-II genes in a GCB DZ-like community (**$_{lc}$GCB-DZ**). The remaining three GCB communities were distinguished by different patterns of GCB neighbourhood expression, wider network features and other associations in downstream analyses (**$_{lc}$GCB** - most canonical GCB-like, **$_{lc}$GCB-FL** - underlying FL diagnosis, **$_{lc}$GCB-SOCS1 –** selective SOCS1 mutation**)**.

The combination of overview and gene-level granularity provided by the network afforded a further means to assess relationships between the resolved communities across all network genes or those specific to the GCB/ABC differentiation state (Supplemental Figure 7 and 8). This is exemplified at the level of GCB and ABC neighbourhood genes, where the $_{lc}$GCB-Xit community straddles both key GCB and ABC features and contrasts with the more discrete patterns of other communities such as $_{lc}$GCB-DZ and $_{lc}$ABC (Supplemental Figure 7). At the whole network level, wider differences in gene expression within modules are further illustrated as seen for the comparison between $_{lc}$GCB-FL and $_{lc}$GCB (Supplemental Figure 8).

Thus the 12 DLBCL LCs reinforced the significance of segregation into ABC and GCB expression patterns while distinguishing heterogeneity within these broad categories.

DLBCL communities link to mutation patterns

An important test of the DLBCL LCs was whether these also showed significant association with mutation state. To address this we analysed the enrichment of mutations in DLBCL communities by integrating enrichment p-values derived across the individual HMRN, REMoDL-B and Reddy datasets.[5,36,37] At a p-value threshold <0.01 (in ≥ 2 datasets) the communities showed significant and distinct associations with mutation patterns. These separated in concordance with the expression states between mutational features linked to ABC, GCB, and host/immune response characteristics (Figure 5a). Thus, $_{lc}$ABC and $_{lc}$ABC-IR shared association with *MYD88, CD79B, PIM1, PRDM1*, *CDKN2A* and differed in association of with *ETV6, TBL1XR1* and *IRF4* mutations. The ABC-related $_{lc}$CT community was selectively enriched for *MYD88* and *PIM1* mutations and was additionally enriched for *BCL10, TP53*, *MEF2B* mutations. These patterns separated most distinctly from $_{lc}$GCB and $_{lc}$GCB-DZ. These shared enrichment for *BCL2*, *EZH2*, *DDX3X*, *IRF8*, *PTEN, TNFRSF14*, *GNA13* mutation. GCB-DZ was additionally enriched for *CREBBP*, *KMT2D*, *BTK, POU2F2, S1PR2, FOXO1*, *MYC, MEF2B* mutation, while $_{lc}$GCB was enriched for *CARD11, B2M, EBF1, MSH6, RHOA*, *TET2*, *SGK1* mutation. $_{lc}$GCB-FL shared association with *CREBBP, KMT2D, TNFRSF14* but was distinguished by association with *STAT6* mutations including significant enrichment of the *STAT6* Y419 hotspot mutation. $_{lc}$GCB-EMT shared enrichment of *SGK1* and *TNFRSF14* with $_{lc}$GCB but otherwise lacked distinct associations. $_{lc}$GCB-SOCS1 and $_{lc}$GCB-Xit differed from other GCB-related expression groups. $_{lc}$GCB-SOCS1 associated with *SOCS1, NFKBIA, SGK1, BTG1, GNA13, CD83, NFKBIE*, *ZFP36L1* and *STAT6* but not the *STAT6* Y419 mutation hotspot, while $_{lc}$GCB-Xit showed a further distinctive pattern with selective enrichment for *CD70, CCND3*, *BTG2* mutations. While $_{lc}$IR only displayed anti-correlations with mutation state likely reflecting the diluting effect of host response components, $_{lc}$IR-TET2 cases were selectively associated with *TET2* mutations.

The observed associations between LCs and mutations resembled the patterns in recent mutation-based classifications of DLBCL. We therefore assessed the relationship between LCs and mutational classes assigned in HMRN data with either the LymphGen,[46] or the HMRN classification (Figure 5b).[5] These classifications though independently derived,[4,5] are

largely concordant,[46] and recapitulate the features of the C1-C5 Harvard classification.[2] The LymphGen MCD class was significantly associated with $_{lc}$ABC and $_{lc}$ABC-IR, while $_{lc}$CT was associated with cases with ambivalent MCD/ST2 calls. LymphGen N1 showed enrichment in $_{lc}$ABC while BN2 was significantly and selectively associated with $_{lc}$GCB-Xit. EZB was most significantly enriched amongst $_{lc}$GCB, $_{lc}$GCB-DZ and GCB-FL, while EZB-MYC+ was selectively enriched amongst $_{lc}$GCB-DZ. ST2 cases mapped selective on to $_{lc}$GCB-SOCS1. Similar patterns of association were evident when considering the associations of LCs with the independent but related HMRN mutational classification. We conclude that the mutation-based subdivision of DLBCL in LymphGen and related classifications overlaps significantly with LCs derived independently from gene co-expression patterns.

Lymphoma communities underline distinctions between double-hit lymphomas

In DLBCL the occurrence of double- or triple-hit with rearrangement of *MYC* and *BCL2* (MYC_BCL2_DH) or *MYC* and *BCL6* (MYC_BCL6_DH), or *MYC*, *BCL2* and *BCL6* (MYC_BCL2_BCL6_TH) identifies high-risk disease.[8-11] The relationships of these rearrangement-based categories to gene expression and mutation features is relatively clear-cut for MYC_BCL2_DH, but remains less well-defined for MYC_BCL6_DH.[47-49] We therefore tested the association of rearrangement status and LCs in the HMRN cohort. $_{lc}$GCB-DZ was highly enriched for MYC_BCL2_DH and MYC_BCL2_BCL6_TH cases (Figure 5b). In contrast, a more modest enrichment of MYC_BCL2_DH cases was evident in $_{lc}$GCB, while $_{lc}$GCB-FL was selectively enriched for BCL2_SH. $_{lc}$GCB-Xit was selectively enriched for MYC_BCL6_DH and BCL6_SH status but was not associated with any combination of *BCL2* rearrangement, while $_{lc}$ABC was enriched for cases with MYC_SH.

We further assessed whether the LCs were significantly associated with morphological diagnosis, COO class and molecular high grade (MHG) or double-hit signature (DHITsig) status.[14,15] Two communities, $_{lc}$GCB-FL and to a lesser extent $_{lc}$GCB-DZ were linked to concurrent or previous diagnosis of underlying FL supporting a primary separation of disease biology in transformed FL cases (Figure 5b). $_{lc}$PBL was selectively enriched for plasmablastic lymphoma morphological diagnosis, $_{lc}$IR and $_{lc}$IR-TET2 for T-cell histiocyte rich large B-cell lymphoma and $_{lc}$GCB-SOCS1 for primary mediastinal large B-cell lymphoma. At gene expression level the LC communities recovered appropriate enrichments of GCB, ABC and unclassified cases. $_{lc}$GCB-DZ was selectively enriched for cases designated as MHG and

cases that were DHITsig positive. Notably DHITsig differed from MHG in also showing significant enrichment amongst $_{lc}$PBL. We conclude that the DLBCL LCs separate meaningfully in relation to underlying rearrangement status, morphological diagnosis, previous expression-based classifiers of high-risk disease and the presence of underlying FL.

Lymphoma communities have prognostic significance in R-CHOP treated cases

Given the fact that LCs segregate cases with common pathological features, we assessed whether LCs showed significant and reproducible survival differences. We addressed this across 6 datasets encompassing sufficient cases treated with R-CHOP chemo-immunotherapy.[3,5,13,29,36,37] Across multiple datasets the DLBCL community structure separated risk (Figure 6) albeit with variation evident between cohorts most notably for $_{lc}$GCB-DZ and $_{lc}$CT. Most adverse risk, in terms of meta-HR across the 6 datasets, was observed for assignment of cases to $_{lc}$PBL, followed by $_{lc}$ABC-IR, $_{lc}$GCB-Xit, $_{lc}$ABC and $_{lc}$IR-TET2. Intermediate and variable risk was observed for $_{lc}$CT and $_{lc}$GCB-DZ. At the other end of the spectrum particularly good risk was associated with $_{lc}$GCB-FL, $_{lc}$GCB-SOCS1, $_{lc}$GCB and $_{lc}$GCB-EMT. We conclude that the refined break down of DLBCLs in the 12-fold LC structure has potential utility to refine the separation of risk groups, while remaining aligned to the concept of the COO differentiation state.

Lymphoma communities show distinct responses to RB-CHOP

The REMoDL-B trial recently reported on extended follow-up identifying a benefit for proteasome inhibitor bortezomib with R-CHOP in cases classified as ABC-DLBCL.[50] We were therefore interested to test whether LCs could provide further insight into responses to RB-CHOP. For overall (OS) and progression-free (PFS) survival the LCs were significantly separated overall in the R-CHOP arm of the trial. In the RB-CHOP arm the significance of separation for the LCs declined for both OS (R-CHOP p=0.0013, RB-CHOP p=0.11) and PFS (R-CHOP p=0.002, RB-CHOP p=0.02) (Figure 7). While there are inherent limitations in *post hoc* analysis and the impact of 12-fold classification on case numbers, we were interested to further compare the response for R-CHOP and RB-CHOP arms of the trial for each LC (Supplemental Figure 9 and 10). While some indication of improved responses for the RB-CHOP arm was evident, for example for $_{lc}$PBL case that achieved a 90%+ OS in the RB-CHOP

treated arm (OS p=0.04, PFS p=0.014), surprisingly, no difference in outcome was observed for $_{IC}$ABC or $_{IC}$ABC-IR between R-CHOP and RB-CHOP arms.

Bortezomib response in ABC cases with variant expression features

The absence of survival difference between the two arms of the trial for cases classified as $_{IC}$ABC contrasted with the apparent benefit of RB-CHOP amongst cases classified as ABC-DLBCL using the trial classifier ($_{T}$ABC). We therefore examined the relative distribution of $_{T}$ABC cases across the LC assignments in REMoDL-B (Figure 8A). Of the 249 $_{T}$ABC cases 120 were also assigned to $_{IC}$ABC and 47 were assigned to $_{IC}$ABC-IR (together 67% of $_{T}$ABC). These intersects included cases with most characteristic ABC expression patterns and showed no significant survival separation between RB-CHOP and R-CHOP arms of the trial. In contrast there was a suggestion of benefit for RB-CHOP over R-CHOP in $_{T}$ABC cases assigned to other LCs. Thus, we conclude that response to RB-CHOP rather than being associated with the most typical of ABC expression patterns is instead associated with cases that combine some ABC expression features with additional features that lead to other (non-ABC) assignments in the LC structure.

Discussion

Heterogeneity is a dominant feature of DLBCL biology. This has led to partially intersecting taxonomies. The most widely accepted of these are classification based on *MYC* and *BCL2* rearrangement status, COO classification based on gene expression patterns, and the recent mutational classifications defined in the LymphGen, Harvard and HMRN classifications.[1-4] However, there remains uncertainty about the inter-relationships between rearrangement, expression and mutation-based classifications.

Here we have used a framework of consistent patterns of gene expression and an inclusive, correlation-centred approach to address these questions from a different perspective. Our analysis learned from 2,500 cases drawing on broadly representative contributions of the DLBCL research community and was tested in nearly 2,500 additional cases from recent studies.[2,12,13,26-35] The network generated using PGCNA derives modules of co-expressed genes that reflect inherent features of DLBCL expression data and reinforces the central importance of the subdivision between ABC and GCB states. The analysis illustrates how

individual genes relate to each other within this context. Exemplifying this, the original COO classifier genes segregate discretely between the two primary modules linked to features of B-cell lineage, and many of the classifier genes and those in alternate COO classifiers, emerge as hub-nodes with high information content in the network.[34,38] The pattern of gene correlation at network level, illustrates how for any module or neighbourhood closely correlated genes could be selected to report similarly on the expression state.

When applying the network to expression-based classification, we have found that using correlation patterns captured in the network structure to select genes which are then collapsed into metagenes reflecting module- or neighbourhood-level expression features, as opposed to selecting highly variant genes and using these at gene-level, adds significant value to the reproducibility of classification. For DLBCL, while the prevailing models for COO classification enhanced with assessments such as MHG/DHITsig status or host response features already provide significant value to identify subsets of disease,[14,15,22,23,34,38,44] our analysis illustrates that the integration of network features can discriminate expression patterns that are not captured in previous approaches. In terms of clinical practice, the LC structure could potentially be developed for case-by-case application, but that has not been the intent of the current study.

Our approach is distinct from and complimentary to other recent studies of DLBCL expression biology such as the Ecotyper that highlight the importance of distinct patterns of host response in DLBCL biology.[22,23] These concepts build from earlier work identifying features of host response that are linked to good outcome.[12,13] The importance of immune and stromal response features is underlined by the contribution that the related network modules make to distinguishing subsets of cases. An interesting example is the link between specific patterns of the M3_ImmuneResponseMonocyteEnriched neighbourhoods and mutations in *TET2* and *DNMT3A*. Indeed, an association between COO unclassified DLBCL and *TET2* mutation enrichment has previously been identified,[3] and our analysis extends this to refine the subgroup of immune response-rich DLBCL with *TET2* mutation association. *TET2* inactivation can contribute to lymphomagenesis in a B-lineage intrinsic fashion in murine models. However, *TET2* inactivation contributes as an early event in haematopoiesis in these models and mutations in *TET2* shared between DLBCL and subsequent myelodysplasia/leukaemia in individual patients have been reported.[51-53] Similar to *TET2*, *DNMT3A* mutations are also common features of clonal haematopoiesis/clonal cytopenia of

undetermined significance.[54-56] In angioimmunoblastic T-cell lymphoma *TET2* and *DNMT3A* mutation have been identified as early events shared with both concurrent clonal haematopoiesis and subsequent DLBCL or other haematological malignancy.[57] It will therefore be interesting to establish to what extent the expression patterns of $_{lc}$IR-TET2 identifies patients with DLBCL who have associated clonal haematopoiesis and whether the mutations in such DLBCL derive exclusively from neoplastic B-cells, or potentially in some cases from components of the host response.

We demonstrate that the network-based approach can facilitate the recognition of groups of cases based on expression state that overlap with mutational classifications, rearrangement status and the presence of underlying FL. In the context of cases with underlying FL the cases separate between those with balanced GCB neighbourhood-level expression patterns and excellent prognosis on R-CHOP treatment and those with a dominant DZ-like expression profile with a poorer prognosis. The $_{lc}$GCB-DZ cases are also enriched for high-risk features such as *MYC* and *BCL2* rearrangement and MHG and DHITsig expression profiles.[8,10,14,49,58] In contrast the good risk $_{lc}$GCB-FL category is significantly enriched for *STAT6* mutations including the Y419F hotspot which has been characterised as a distinctive feature of a subset of FL.[41,59-62] The $_{lc}$GCB-FL category is in many ways consistent with recent studies showing that DLBCL with concurrent or underlying FL at diagnosis is not necessarily associated with adverse prognosis.[63,64] Together the $_{lc}$GCB-DZ and $_{lc}$GCB-FL categories argue for divergent patterns of FL evolution that are distinguished on the one hand by *MYC* rearrangement and related expression features for $_{lc}$GCB-DZ and on the other for $_{lc}$GCB-FL a link with underlying *STAT6* mutation.

The treatment landscape of DLBCL is rapidly evolving and introduction of polatuzumab vedoitin targeting CD79b is changing frontline therapy.[65] However, it remains of interest to explore refined separation of expression state in R-CHOP based trial data. Recently the longer-term follow-up in the REMoDL-B trial has been reported identifying a significant survival benefit in the trial-classified ABC or MHG cases.[50] Our analysis demonstrates that the apparent response to RB-CHOP in the ABC-DLBCL subset observed in the trial classifications is not associated with the set of cases that are identified both by the trial and the LC classification as ABC. This overlap includes cases with the most typical ABC expression patterns. Instead, the response appears to reside in cases with variant features related to host response or differentiation state. We acknowledge that the *post hoc*

subdivision of trial classified cases into multiple subgroups raises substantial caveats. However, a notable feature is that cases identified as plasmablastic ($_{IC}$PBL) and treated with RB-CHOP in the REMoDL-B trial show a favourable outcome. This is notable because $_{IC}$PBL cases in other case series have the worst outcome overall when treated with R-CHOP, which has until recently been widely used for such cases.[66] Our analysis of the REMoDL-B data supports the arguments put forward from other studies that bortezomib-containing regimens should be considered in DLBCL with plasmablastic features.[66-70]

We conclude that our network-based approach yields an encompassing map of DLBCL tumour biology and illustrates how data integration and network analysis can refine expression-based stratification of DLBCL. Our analysis supports the argument that such refined stratification is needed to accurately identify treatment responses even within existing molecular subtypes.

Methods

(See also Supplemental Methods)

Expression datasets

Thirteen expression datasets were used for PGCNA (GSE4475, GSE4732, GSE10846, GSE12195, GSE19246, GSE22470, GSE31312, GSE32918, GSE34171, GSE53786, GSE87371, GSE98588, Monti).[2,12,13,26-35] For validation 4 datasets were used: HMRN (GSE181063), NCI (NCICCR-DLBCL), REMoDLB (GSE117556) and Reddy (EGAS00001002606), (Supplemental Table 1).[3,5,13,29,36,37] For RNA-seq datasets (NCI, Reddy) count data was processed using DESeq2 v1.22.2 with VST-normalised data used for analysis. Probes were re-annotated (http://mygene.info) ambiguous mappings were manually assigned. Datasets were quantile normalised (Python qnorm) and probe sets merged (median value for probe sets with Pearson correlation ≥0.2 and maximum value for those with correlation <0.2).

Network generation

Discovery datasets (GSE10846 split into CHOP/RCHOP-treated) were processed using PGCNA2 (https://github.com/medmaca/PGCNA/tree/master/PGCNA2), retaining the top 70% most variant genes present in ≥50% of the datasets, carrying out 1000 Leidenalg clusterings and selecting the best using Scaled cluster enrichment scores. The resulting

network contained 16,054 genes (44,730 edges) split into 28 modules (Supplemental Table 1). The network was visualised using Gephi (version 0.9.2), and interactive HTML5 web visualisations exported using the sigma.js library. Interactive networks are at https://mcare.link/DLBCL2.

Network analysis

Clustering samples

See supplemental methods: Before clustering, we split each module into submodules/neighbourhoods and collapsed the genes within these to single values (MEV/NEVs). The most informative genes/MEV/NEVs were selected using several approaches, and used to cluster the DLBCL datasets using two different approaches PGCNA/ConsensusClustering (CC)[1]. We used cluster results to explore how recurrently discoverable the DLBCL communities were, allowing the selection of the most informative attributes for clustering. The best results for the HMRN dataset were combined between PGCNA/CC to generate ensemble Lymphoma Communities (LC). These were used to train a machine learning tool to recover the LCs in every dataset.

Neighbourhoods

Each of the 28 modules was sub-clustered to create neighbourhoods. The existing PGCNA edge file was split at the module level and then clustered 5,000 times using Leidenalg. The best clustering (based on modularity score) for each module was retained. Multiple different runs converged onto the same answer. The neighbourhoods are detailed in Extended Data Figure 3.

Module Expression Values and Lymphoma communities

Genes per module/neighbourhood were collapsed down to single values: within each dataset, which vary in available genes, the genes per module/neighbourhood were ranked by gene_strength (sum of genes edges/correlations within its module). Representative genes were selected and converted into a MEV or NEV by:

1. Per module/neighbourhood select top 10 genes based on ranks.
2. Per gene, standardize (z-score) the quantile-normalized $\log_2$ expression data.

3. Per sample (patient) calculate the median of the 10 z-scores to give a MEV/NEV.

Lymphoma Community machine learning tool

Clustering results for HMRN were merged by selecting significant overlaps (p-value < 0.0001) that form a community in either CC/PGCNA containing > 5 samples (Supplemental Figure 1 & Supplemental Methods). This generated a high confidence set of LCs (n=298) that formed a training dataset (Figure 4b top) that was used to train a machine learning (ML) tool to recover the LC in other datasets. The training dataset was split using the python Scikit-learn test_train_split function, stratifying on the LC class label, to give class-balanced randomised training (n=238) and validation (n=60) datasets. The validation dataset was set aside to test the final selected model. The training data was used to carry out stratified 5-fold cross-validation across 7 different machine learning methods. In total 4,802 parameters were tested across the ML tools, scoring with the Matthews correlation coefficient (MCC). The best 3 models, using their optimal parameters, were combined using soft voting to give a model with mean MCC of 0.89 across the 5-folds. This model was then tested on the unseen validation data (n=60) with MCC of 0.87. This final model, termed ML_LC, was retrained using all 298 training samples and was used for all subsequent classifications, including the HMRN dataset (Figure 4b bottom).

Survival analysis

Right-censored survival data, where available, was analysed using Survival library for R. The expression of each gene (as z-score) or the ML_LC *p*-value was used as a continuous variable in a Cox Proportional Hazards model and the ML_LC community for a Kaplan-Meier estimator (using merged survival data). Meta-analysis across datasets was conducted by fitting a fixed-effect model to hazard ratios, weighted by dataset size.

Mutation analysis

Analysis was carried out for HMRN (188 genes/431 samples), REMoDLB (70 genes/400 samples) and Reddy (150 genes/624 samples) datasets. Mutations were converted to a binary matrix for downstream analysis. For each dataset the point-biserial correlations were calculated between all pairs of mutated gene and MEV/NEV. The resulting correlation *p*-

values were converted to z-scores (python stats.norm.ppf) to convey the ±correlation along with its significance. Significance of overlap between mutations and LC was calculated using hypergeometric testing within each dataset. To generate meta results MEV/Community mutation analysis p-values were combined using the Stouffer's Z method.

Data Availability Statement

The underlying primary datasets are available at the indicated data source (see above under Expression Datasets). All resulting gene correlation data, module and neighbourhood gene lists, and signature/ontology enrichments are available at https://mcare.link/DLBCL2.

Code Availability Statement

The DLBCL LC classifier and networks are available at https://mcare.link/DLBCL2.

PGCNA is available at https://github.com/medmaca/PGCNA/tree/master/PGCNA2

All other code is available on request.

Acknowledgements

This work was supported by Cancer Research UK program grant (C7845/A17723 and C7845/A29212) (M.C, G.D., D.W, and R.T). HMRN is supported by Cancer Research UK program grant A29685 (D.P., S.C., A.S., E.R.). D.J.H. was supported by a fellowship from Cancer Research UK (CRUK) (RCCFEL\100072) and received core funding from Wellcome (203151/Z/16/Z) to the Wellcome-MRC Cambridge Stem Cell Institute and from the CRUK Cambridge Centre (A25117). D.J.H is supported by the National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014). R.T., G.D., E.R., A.S., D.P., D.W. are supported by the National Institute for Health and Care Research Leeds Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Alizadeh, A.A., *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511 (2000).
2. Chapuy, B., *et al.* Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat Med* **24**, 679-690 (2018).
3. Schmitz, R., *et al.* Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N Engl J Med* **378**, 1396-1407 (2018).
4. Wright, G.W., *et al.* A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications. *Cancer Cell* **37**, 551-568 e514 (2020).
5. Lacy, S.E., *et al.* Targeted sequencing in DLBCL, molecular subtypes, and outcomes: a Haematological Malignancy Research Network report. *Blood* **135**, 1759-1771 (2020).
6. Pedrosa, L., *et al.* Proposal and validation of a method to classify genetic subtypes of diffuse large B cell lymphoma. *Sci Rep* **11**, 1886 (2021).
7. Bolen, C.R., *et al.* Prognostic impact of somatic mutations in diffuse large B-cell lymphoma and relationship to cell-of-origin: data from the phase III GOYA study. *Haematologica* **105**, 2298-2307 (2020).
8. Aukema, S.M., *et al.* Biological characterization of adult MYC-translocation-positive mature B-cell lymphomas other than molecular Burkitt lymphoma. *Haematologica* **99**, 726-735 (2014).
9. Aukema, S.M., *et al.* Double-hit B-cell lymphomas. *Blood* **117**, 2319-2331 (2011).
10. Barrans, S., *et al.* Rearrangement of MYC is associated with poor prognosis in patients with diffuse large B-cell lymphoma treated in the era of rituximab. *J Clin Oncol* **28**, 3360-3365 (2010).
11. Li, S., *et al.* B-cell lymphomas with MYC/8q24 rearrangements and IGH@BCL2/t(14;18)(q32;q21): an aggressive disease with heterogeneous histology, germinal center B-cell immunophenotype and poor outcome. *Mod Pathol* **25**, 145-156 (2012).
12. 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).
13. Lenz, G., *et al.* Stromal gene signatures in large-B-cell lymphomas. *The New England journal of medicine* **359**, 2313-2323 (2008).
14. Sha, C., *et al.* Molecular High-Grade B-Cell Lymphoma: Defining a Poor-Risk Group That Requires Different Approaches to Therapy. *J Clin Oncol* **37**, 202-212 (2019).
15. Ennishi, D., *et al.* Double-Hit Gene Expression Signature Defines a Distinct Subgroup of Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma. *J Clin Oncol* **37**, 190-201 (2019).
16. Holmes, A.B., *et al.* Single-cell analysis of germinal-center B cells informs on lymphoma cell of origin and outcome. *J Exp Med* **217**(2020).
17. King, H.W., *et al.* Single-cell analysis of human B cell maturation predicts how antibody class switching shapes selection dynamics. *Sci Immunol* **6**(2021).
18. Milpied, P., *et al.* Human germinal center transcriptional programs are de-synchronized in B cell lymphoma. *Nat Immunol* **19**, 1013-1024 (2018).
19. Attaf, N., Baaklini, S., Binet, L. & Milpied, P. Heterogeneity of germinal center B cells: New insights from single-cell studies. *Eur J Immunol* **51**, 2555-2567 (2021).

20. Duan, L., *et al.* Follicular dendritic cells restrict interleukin-4 availability in germinal centers and foster memory B cell generation. *Immunity* **54**, 2256-2272 e2256 (2021).
21. Kennedy, D.E., *et al.* Novel specialized cell state and spatial compartments within the germinal center. *Nat Immunol* **21**, 660-670 (2020).
22. Steen, C.B., *et al.* The landscape of tumor cell states and ecosystems in diffuse large B cell lymphoma. *Cancer Cell* **39**, 1422-1437 e1410 (2021).
23. Kotlov, N., *et al.* Clinical and Biological Subtypes of B-cell Lymphoma Revealed by Microenvironmental Signatures. *Cancer discovery* **11**, 1468-1489 (2021).
24. Care, M.A., Westhead, D.R. & Tooze, R.M. Parsimonious Gene Correlation Network Analysis (PGCNA): a tool to define modular gene co-expression for refined molecular stratification in cancer. *NPJ Syst Biol Appl* **5**, 13 (2019).
25. Stephenson, S., *et al.* Growth Factor-like Gene Regulation Is Separable from Survival and Maturation in Antibody-Secreting Cells. *J Immunol* **202**, 1287-1300 (2019).
26. Compagno, M., *et al.* Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. *Nature* **459**, 717-721 (2009).
27. Williams, P.M., *et al.* A novel method of amplification of FFPET-derived RNA enables accurate disease classification with microarrays. *J Mol Diagn* **12**, 680-686 (2010).
28. Salaverria, I., *et al.* Translocations activating IRF4 identify a subtype of germinal center-derived B-cell lymphoma affecting predominantly children and young adults. *Blood* **118**, 139-147 (2011).
29. Frei, E., *et al.* Addition of rituximab to chemotherapy overcomes the negative prognostic impact of cyclin E expression in diffuse large B-cell lymphoma. *J Clin Pathol* **66**, 956-961 (2013).
30. Care, M.A., *et al.* SPIB and BATF provide alternate determinants of IRF4 occupancy in diffuse large B-cell lymphoma linked to disease heterogeneity. *Nucleic Acids Res* **42**, 7591-7610 (2014).
31. 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).
32. Hummel, M., *et al.* A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *The New England journal of medicine* **354**, 2419-2430 (2006).
33. Dave, S.S., *et al.* Molecular diagnosis of Burkitt's lymphoma. *The New England journal of medicine* **354**, 2431-2442 (2006).
34. Scott, D.W., *et al.* Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood* **123**, 1214-1217 (2014).
35. 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).
36. Reddy, A., *et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171**, 481-494 e415 (2017).
37. Davies, A., *et al.* Gene-expression profiling of bortezomib added to standard chemoimmunotherapy for diffuse large B-cell lymphoma (REMoDL-B): an open-label, randomised, phase 3 trial. *Lancet Oncol* **20**, 649-662 (2019).
38. Wright, G., *et al.* A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc Natl Acad Sci U S A* **100**, 9991-9996 (2003).

39. Basso, K. & Dalla-Favera, R. Germinal centres and B cell lymphomagenesis. *Nat Rev Immunol* **15**, 172-184 (2015).
40. Shaffer, A.L., 3rd, Young, R.M. & Staudt, L.M. Pathogenesis of human B cell lymphomas. *Annual review of immunology* **30**, 565-610 (2012).
41. Crouch, S.*, et al.* Molecular subclusters of follicular lymphoma: a report from the United Kingdom's Haematological Malignancy Research Network. *Blood Adv* **6**, 5716-5731 (2022).
42. Ennishi, D.*, et al.* TMEM30A loss-of-function mutations drive lymphomagenesis and confer therapeutically exploitable vulnerability in B-cell lymphoma. *Nat Med* **26**, 577-588 (2020).
43. Wilkerson, M.D. & Hayes, D.N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573 (2010).
44. Care, M.A.*, et al.* A Microarray Platform-Independent Classification Tool for Cell of Origin Class Allows Comparative Analysis of Gene Expression in Diffuse Large B-cell Lymphoma. *PLoS ONE* **8**, e55895 (2013).
45. Swift, S.*, et al.* Consensus clustering and functional interpretation of gene-expression data. *Genome biology* **5**, R94 (2004).
46. Runge, H.F.P.*, et al.* Application of the LymphGen classification tool to 928 clinically and genetically-characterised cases of diffuse large B cell lymphoma (DLBCL). *Br J Haematol* (2020).
47. Rosenwald, A.*, et al.* Prognostic Significance of MYC Rearrangement and Translocation Partner in Diffuse Large B-Cell Lymphoma: A Study by the Lunenburg Lymphoma Biomarker Consortium. *J Clin Oncol* **37**, 3359-3368 (2019).
48. Cucco, F.*, et al.* Distinct genetic changes reveal evolutionary history and heterogeneous molecular grade of DLBCL with MYC/BCL2 double-hit. *Leukemia* **34**, 1329-1341 (2020).
49. Scott, D.W.*, et al.* High-grade B-cell lymphoma with MYC and BCL2 and/or BCL6 rearrangements with diffuse large B-cell lymphoma morphology. *Blood* **131**, 2060-2064 (2018).
50. Davies, A.J.*, et al.* Differential Efficacy From the Addition of Bortezomib to R-CHOP in Diffuse Large B-Cell Lymphoma According to the Molecular Subgroup in the REMoDL-B Study With a 5-Year Follow-Up. *J Clin Oncol*, JCO2300033 (2023).
51. Quivoron, C.*, et al.* TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer Cell* **20**, 25-38 (2011).
52. Dominguez, P.M.*, et al.* TET2 Deficiency Causes Germinal Center Hyperplasia, Impairs Plasma Cell Differentiation, and Promotes B-cell Lymphomagenesis. *Cancer discovery* **8**, 1632-1653 (2018).
53. Mouly, E.*, et al.* B-cell tumor development in Tet2-deficient mice. *Blood Adv* **2**, 703-714 (2018).
54. Edwards, J.P., Thornton, A.M. & Shevach, E.M. Release of active TGF-beta1 from the latent TGF-beta1/GARP complex on T regulatory cells is mediated by integrin beta8. *J Immunol* **193**, 2843-2849 (2014).
55. Genovese, G.*, et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477-2487 (2014).
56. Jaiswal, S.*, et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488-2498 (2014).

57. Cheng, S., Zhang, W., Inghirami, G. & Tam, W. Mutation analysis links angioimmunoblastic T-cell lymphoma to clonal hematopoiesis and smoking. *eLife* **10**(2021).
58. Marshall, H.D.*, et al.* The transforming growth factor beta signaling pathway is critical for the formation of CD4 T follicular helper cells and isotype-switched antibody responses in the lung mucosa. *eLife* **4**, e04851 (2015).
59. Yildiz, M.*, et al.* Activating STAT6 mutations in follicular lymphoma. *Blood* **125**, 668-679 (2015).
60. Siddiqi, I.N.*, et al.* Characterization of a variant of t(14;18) negative nodal diffuse follicular lymphoma with CD23 expression, 1p36/TNFRSF14 abnormalities, and STAT6 mutations. *Mod Pathol* **29**, 570-581 (2016).
61. Xian, R.R.*, et al.* CREBBP and STAT6 co-mutation and 16p13 and 1p36 loss define the t(14;18)-negative diffuse variant of follicular lymphoma. *Blood Cancer J* **10**, 69 (2020).
62. Mentz, M.*, et al.* PARP14 is a novel target in STAT6 mutant follicular lymphoma. *Leukemia* **36**, 2281-2292 (2022).
63. Behdad, A.*, et al.* Survival outcomes of diffuse large B-cell lymphoma by association with concurrent or antecedent follicular lymphoma and double hit status. *Leuk Lymphoma* **60**, 3266-3271 (2019).
64. Wang, Y.*, et al.* Impact of concurrent indolent lymphoma on the clinical outcome of newly diagnosed diffuse large B-cell lymphoma. *Blood* **134**, 1289-1297 (2019).
65. Tilly, H.*, et al.* Polatuzumab Vedotin in Previously Untreated Diffuse Large B-Cell Lymphoma. *N Engl J Med* **386**, 351-363 (2022).
66. Makady, N.F., Ramzy, D., Ghaly, R., Abdel-Malek, R.R. & Shohdy, K.S. The Emerging Treatment Options of Plasmablastic Lymphoma: Analysis of 173 Individual Patient Outcomes. *Clin Lymphoma Myeloma Leuk* **21**, e255-e263 (2021).
67. Guerrero-Garcia, T.A., Mogollon, R.J. & Castillo, J.J. Bortezomib in plasmablastic lymphoma: A glimpse of hope for a hard-to-treat disease. *Leuk Res* **62**, 12-16 (2017).
68. Dittus, C., Grover, N., Ellsworth, S., Tan, X. & Park, S.I. Bortezomib in combination with dose-adjusted EPOCH (etoposide, prednisone, vincristine, cyclophosphamide, and doxorubicin) induces long-term survival in patients with plasmablastic lymphoma: a retrospective analysis. *Leuk Lymphoma* **59**, 2121-2127 (2018).
69. Castillo, J.J.*, et al.* Bortezomib plus EPOCH is effective as frontline treatment in patients with plasmablastic lymphoma. *Br J Haematol* **184**, 679-682 (2019).
70. Sabry, W., Wu, Y. & Kodad, S.G. Bortezomib, Lenalidomide and Dexamethasone Combination Induced Complete Remission in Relapsed/Refractory Plasmablastic Lymphoma: Case Report of a Potential Novel Treatment Approach. *Curr Oncol* **29**, 5042-5053 (2022).

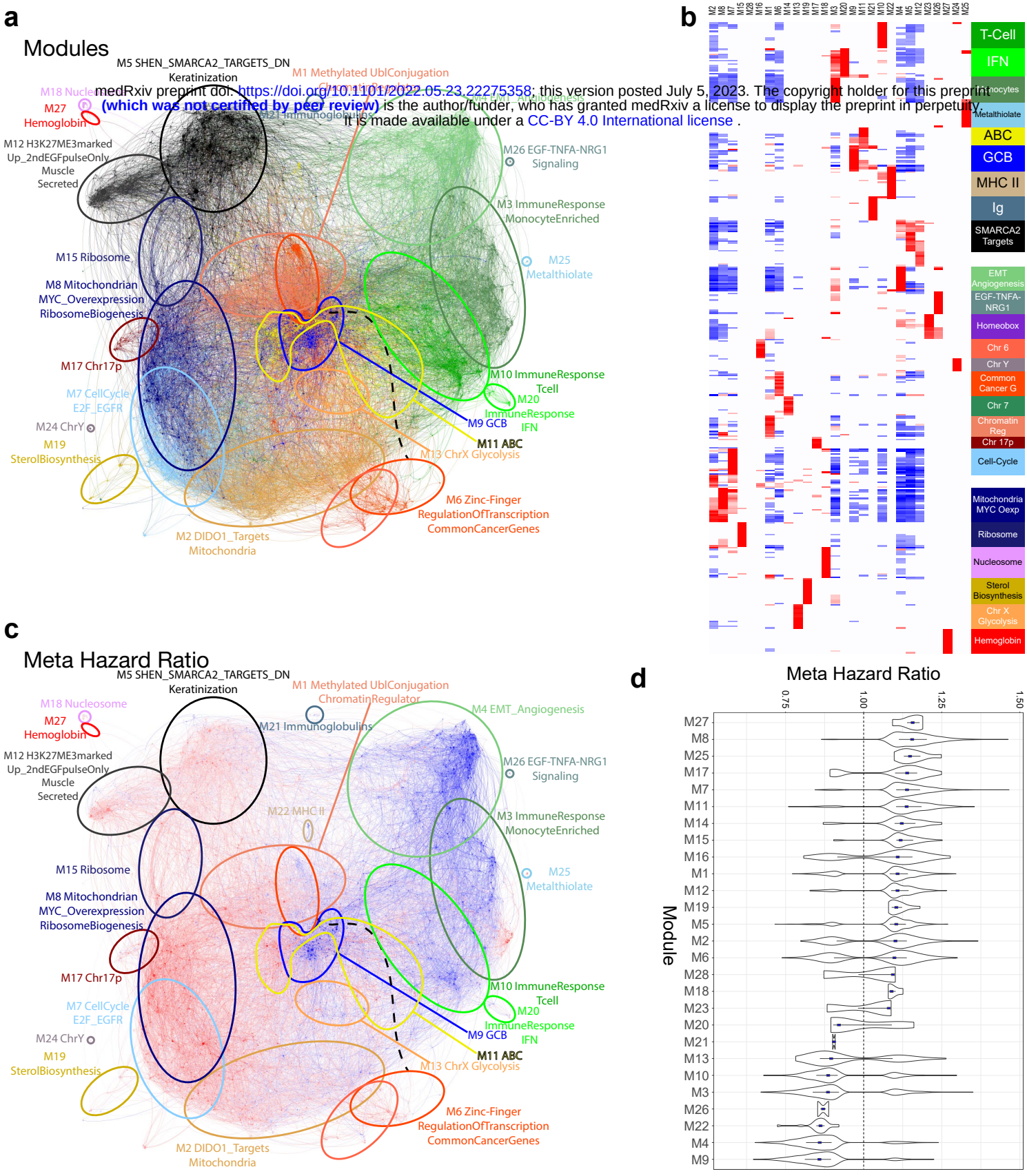

Figure 1. PGCNA network visualization for DLBCL
(**a**) DLBCL network with modules colour-coded, modules outlines are approximated with ellipses of same colour with module summary term indicated. Interactive versions are available at https://mcare.link/DLBCL for detailed exploration. (**b**) Separation of module signature and ontology associations is illustrated as a heatmap (filtered FDR <0.05 and ≥ 5 and ≤ 1000 genes; top 15 most significant signatures per module). Significant enrichment or depletion illustrated on red/blue scale, x-axis (modules) and y-axis (signatures). Hierarchical clustering according to gene signature enrichment. For high-resolution version and extended data see Extended Data Figure 2 and Supplemental Table 2. (**c**) Overlay of meta-hazard ratio (HR) of death across available meta-data in training datasets. Corresponding module outline approximations are illustrated as in (a). Colour scales: outcome blue (low HR - good outcome) to red (high HR - poor outcome). Interactive version available at https://mcare.link/DLBCL for detailed exploration, along with additional meta-data overlays. (**d**) Ranked module level association with HR of death. Distribution of HR associations for module genes with p-value < 0.05, along with median (blue square) and IQR.

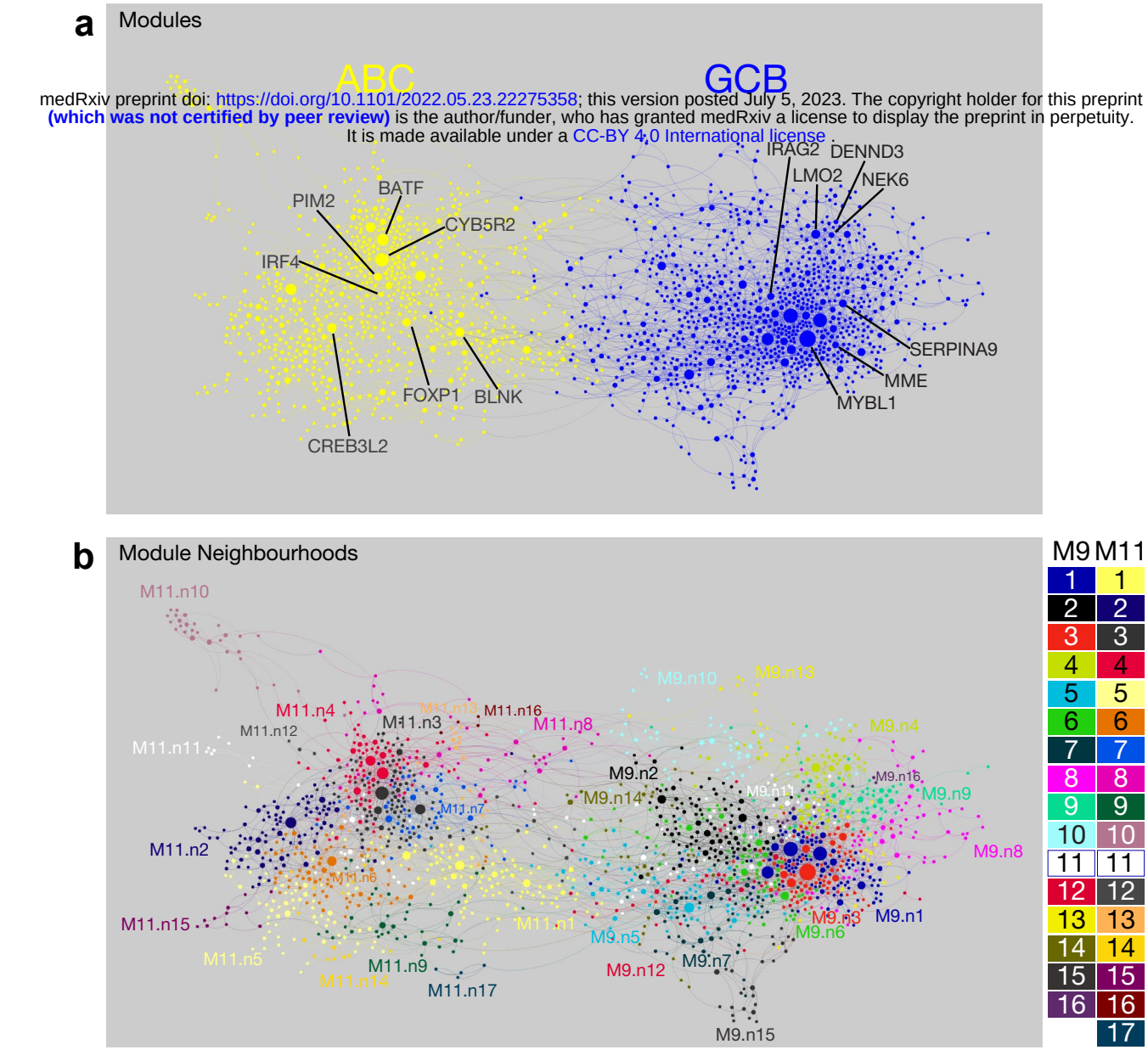

Figure 2. GCB and ABC module neighbourhoods provide fine-grained resolution
(a) Modules M9_GCB (blue) and M11_ABC (yellow) are displayed as a merged ABC/GCB specific network with node sizes reflecting degree. **(b)** The ABC/GCB specific network structure as in (a) colour coded according to gene neighbourhood membership, with colour key provided to the right. **(c)** and **(d)** Separation of neighbourhood signature and ontology associations are illustrated as a bubble plot for M9_GCB (c) and M11_ABC (d) (showing select signatures). Significant enrichment illustrated by colour and bubble size, x-axis (neighbourhoods) and y-axis (signatures) for relevant signatures. For high-resolution version and extended data see Extended Data Figure 3 and Supplemental Table 4.

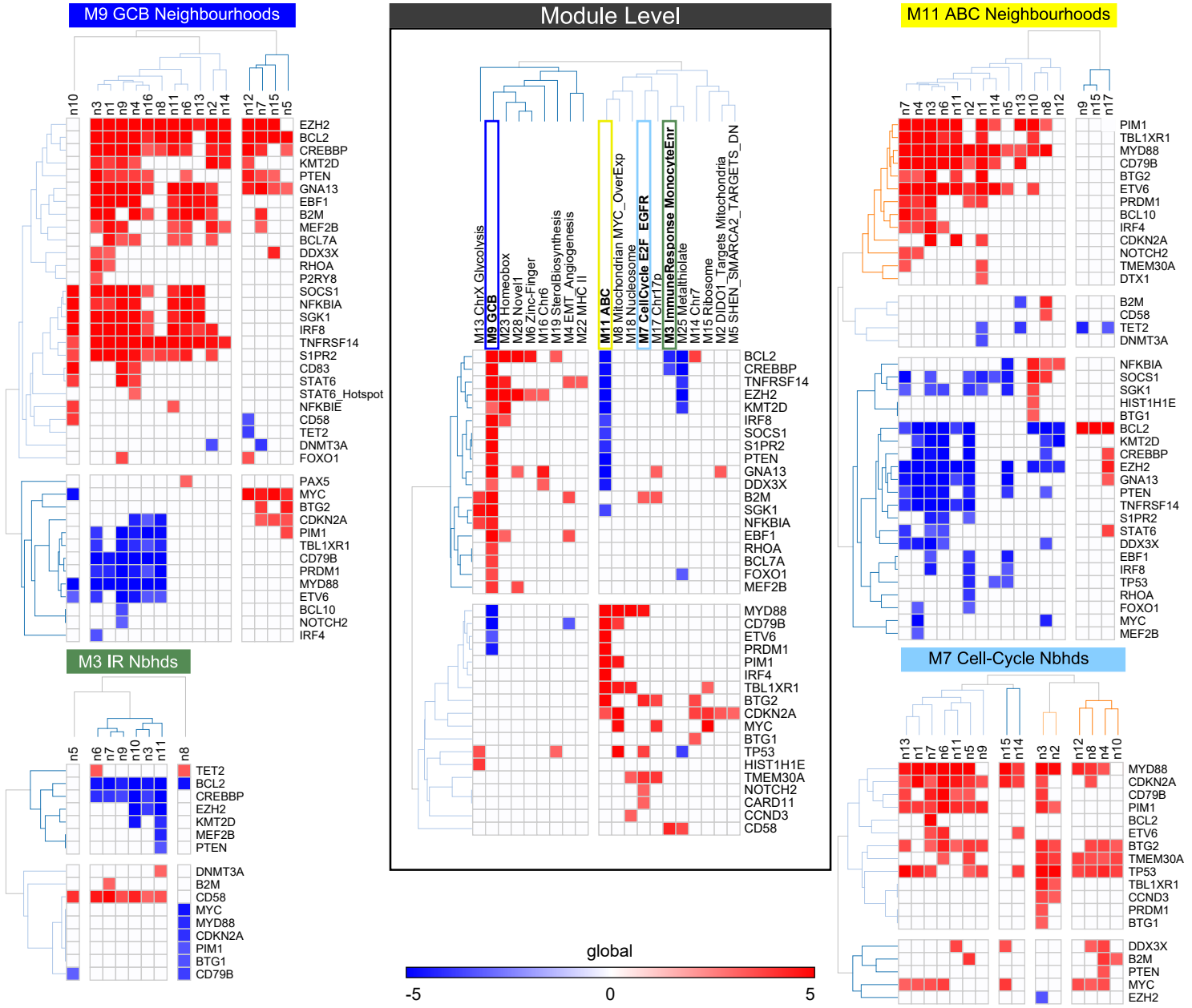

Figure 3. Relationship between MEV/NEVs and mutations

Association of Module/Neighbourhood Expression Values (MEV/NEV) with mutations. Shown are combined significance (Stouffer method) of p-value based on point-biserial correlations between binary mutation status and MEV/NEVs across 3 datasets (HMRN, Reddy & REMoDLB). Only mutations with p-value < 0.001, occurring in ≥ 2 datasets were retained, Z-scores for p-values > 0.001 were set to 0. Significance is shown as z-scores on a blue (significant depletion) to red (significant enrichment) scale (-5 to +5). Centre shows the relationship between MEVs and mutations. Outside shows the relationship between select NEVs and mutations.

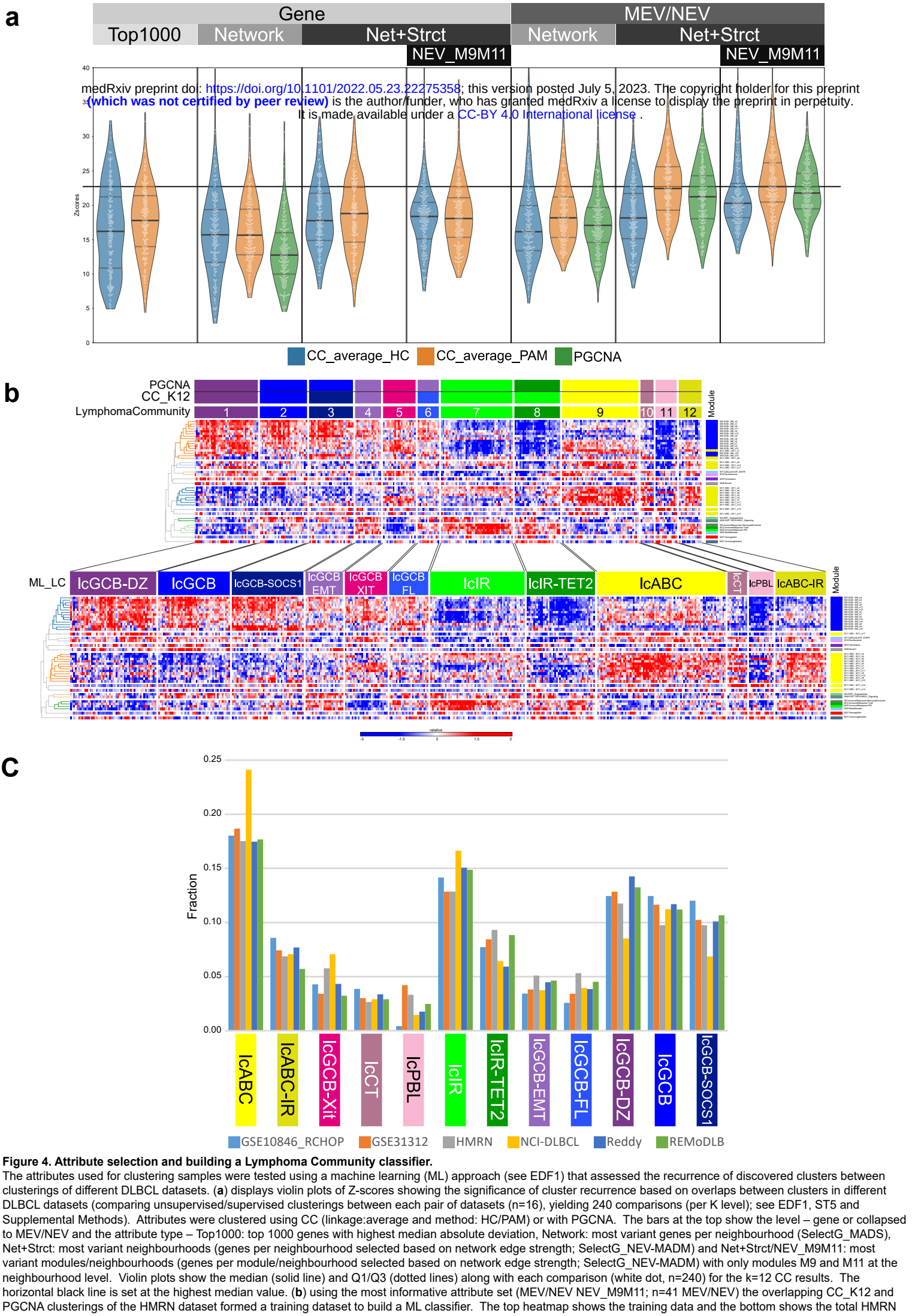

Figure 4. Attribute selection and building a Lymphoma Community classifier.
The attributes used for clustering samples were tested using a machine learning (ML) approach (see EDF1) that assessed the recurrence of discovered clusters between clusterings of different DLBCL datasets. (**a**) displays violin plots of Z-scores showing the significance of cluster recurrence based on overlaps between clusters in different DLBCL datasets (comparing unsupervised/supervised clusterings between each pair of datasets (n=16), yielding 240 comparisons (per K level); see EDF1, ST5 and Supplemental Methods). Attributes were clustered using CC (linkage:average and method: HC/PAM) or with PGCNA. The bars at the top show the level – gene or collapsed to MEV/NEV and the attribute type – Top1000: top 1000 genes with highest median absolute deviation, Network: most variant genes per neighbourhood (SelectG_MADS), Net+Strct: most variant neighbourhoods (genes per neighbourhood selected based on network edge strength; SelectG_NEV-MADM) and Net+Strct/NEV_M9M11: most variant modules/neighbourhoods (genes per module/neighbourhood selected based on network edge strength; SelectG_NEV-MADM) with only modules M9 and M11 at the neighbourhood level. Violin plots show the median (solid line) and Q1/Q3 (dotted lines) along with each comparison (white dot, n=240) for the k=12 CC results. The horizontal black line is set at the highest median value. (**b**) using the most informative attribute set (MEV/NEV NEV_M9M11; n=41 MEV/NEV) the overlapping CC_K12 and PGCNA clusterings of the HMRN dataset formed a training dataset to build a ML classifier. The top heatmap shows the training data and the bottom shows the total HMRN dataset reclassified using the trained ML tool (ML_LC). Each row shows the expression of the displayed MEV/NEV across the samples on a blue (low) to red (high) z-score colour scale. (**c**) the fraction of each LC across the datasets classified using ML_LC.

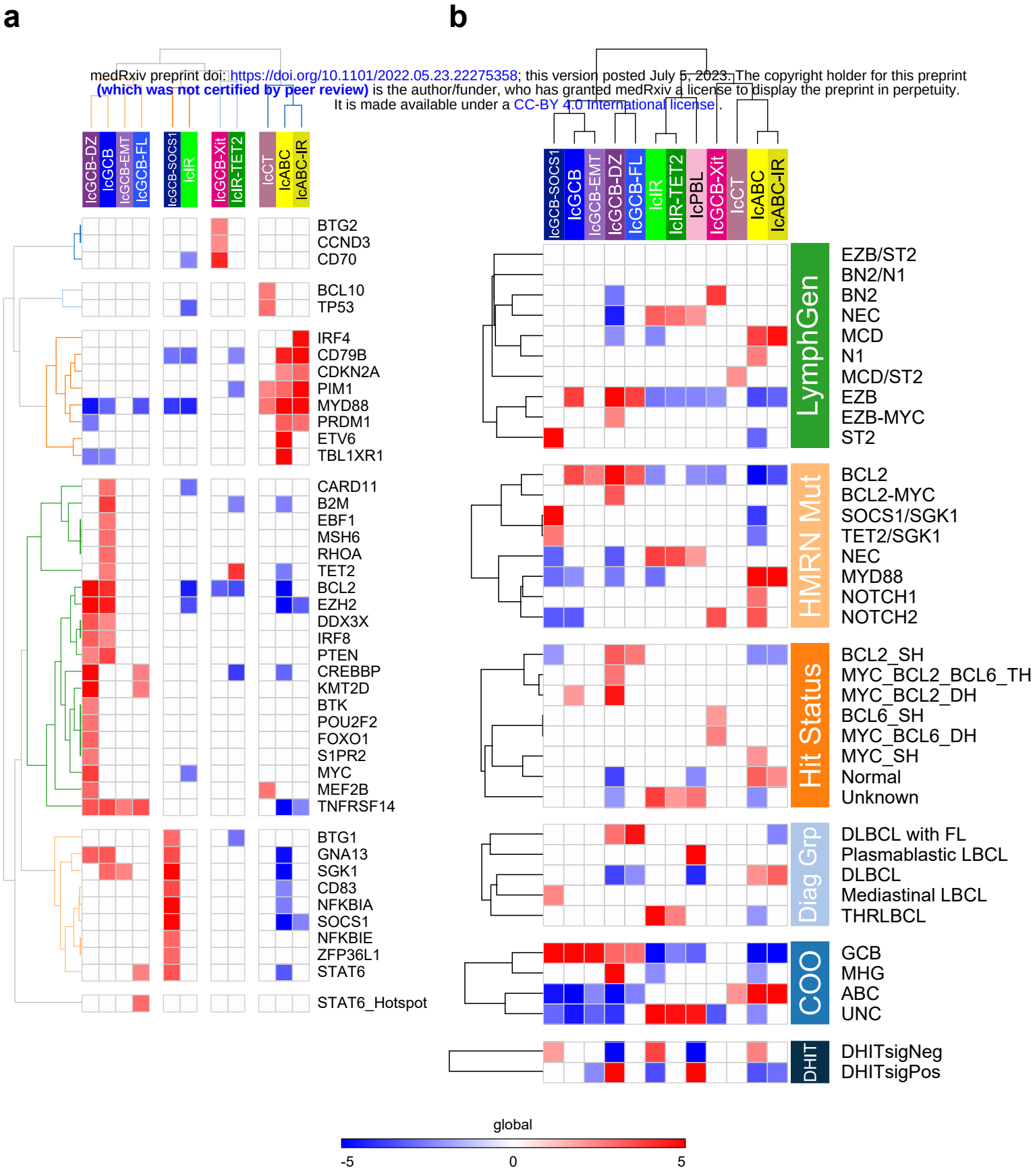

Figure 5. Lymphoma Communities (LC) have distinct mutational and rearrangement associations.
Association of LC with mutation and rearrangement status. (**a**) The differential enrichment of gene mutations across LC integrated across three datasets. Shown is combined significance (Stouffer method) of LC enrichment/depletion of mutations as z-scores on a blue (significant depletion) to red (significant enrichment) scale (-5 to +5). X-axis shows hierarchically clustered LC and y-axis gene symbols. Only mutations with p-value < 0.01, occurring in ≥ 2 datasets were retained, Z-scores for p-values > 0.01 were set to 0. (**b**) Significance of enrichment/depletion of LC with LymphGen, HMRN mutational group (PMID: 32187361), Hit-status (rearrangement status for MYC, BCL2 and BCL6 indicating single hit (SH), double hit MYC_BCL2-DH or MYC_BCL6-DH, or triple hit MYC_BCL2_BCL6-TH), Diagnostic-Group, cell-of-origin/MHG and DHITsig assignments in the HMRN dataset. X-axis hierarchically clustered LC, against y-axis clustered within each group. Z-scores with p-value > 0.05 were set to 0.

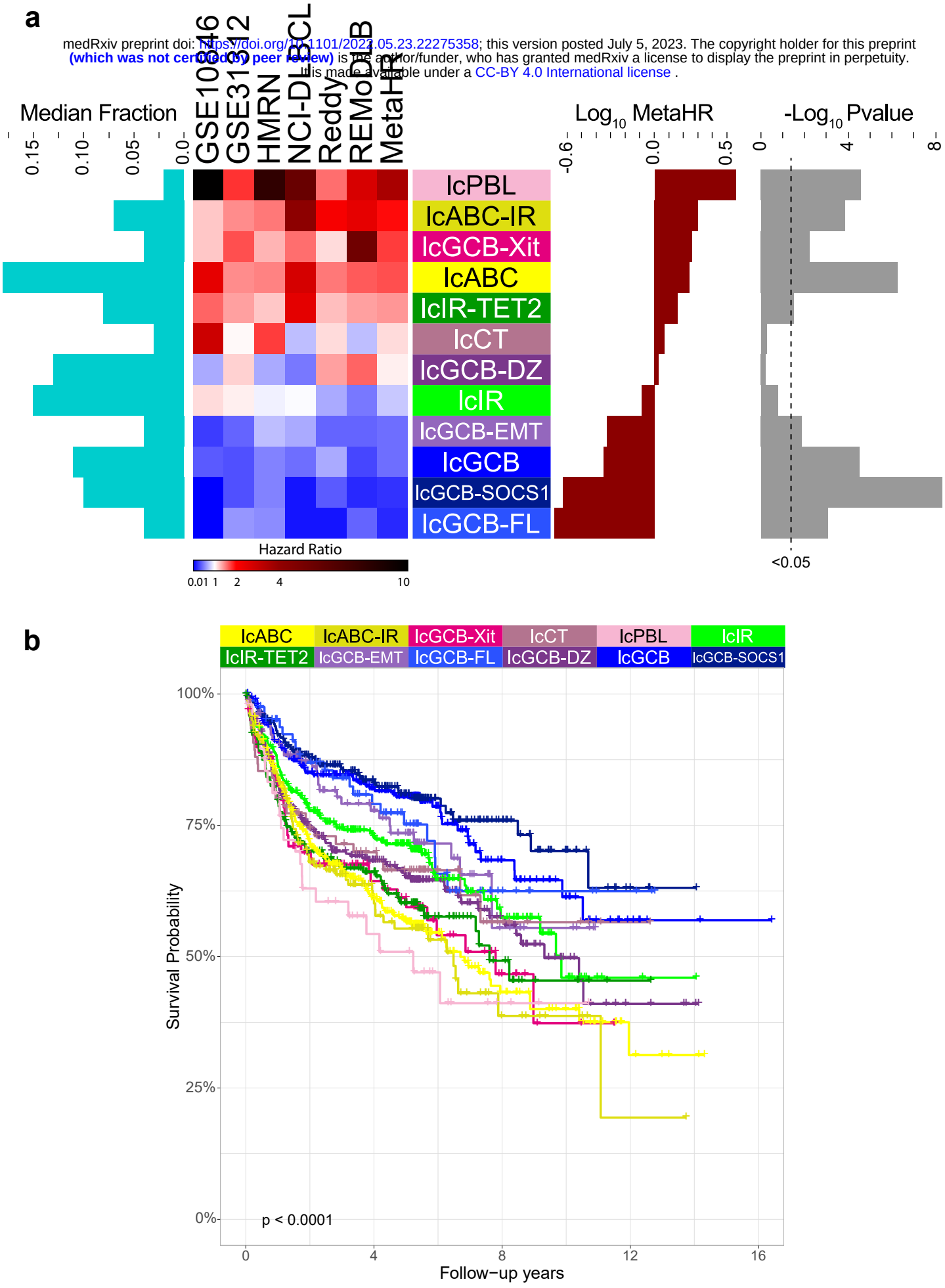

Figure 6. Lymphoma communities are significantly associated with overall survival
(**a**) Lymphoma Communities (LC) across the indicated datasets show consistent associations for hazard ratio (HR) of overall survival for R-CHOP treated patients. Heatmaps shows HR on a blue (good) to red (poor) colour scale across 6 datasets ordered by metaHR (from Cox proportional hazards regression; ML_LC LC p-value as the explanatory variable). The left chart shows the median fraction size of each LC across the 6 datasets. The right charts show the Log10 MetaHR (red) and -Log10 p-value (grey) with p<0.05 indicated by a dashed-line. (**b**) Kaplan-Meier plots of overall survival for R-CHOP treated patients across the 6 DLBCL datasets for shown LC; p-value from log-rank test.

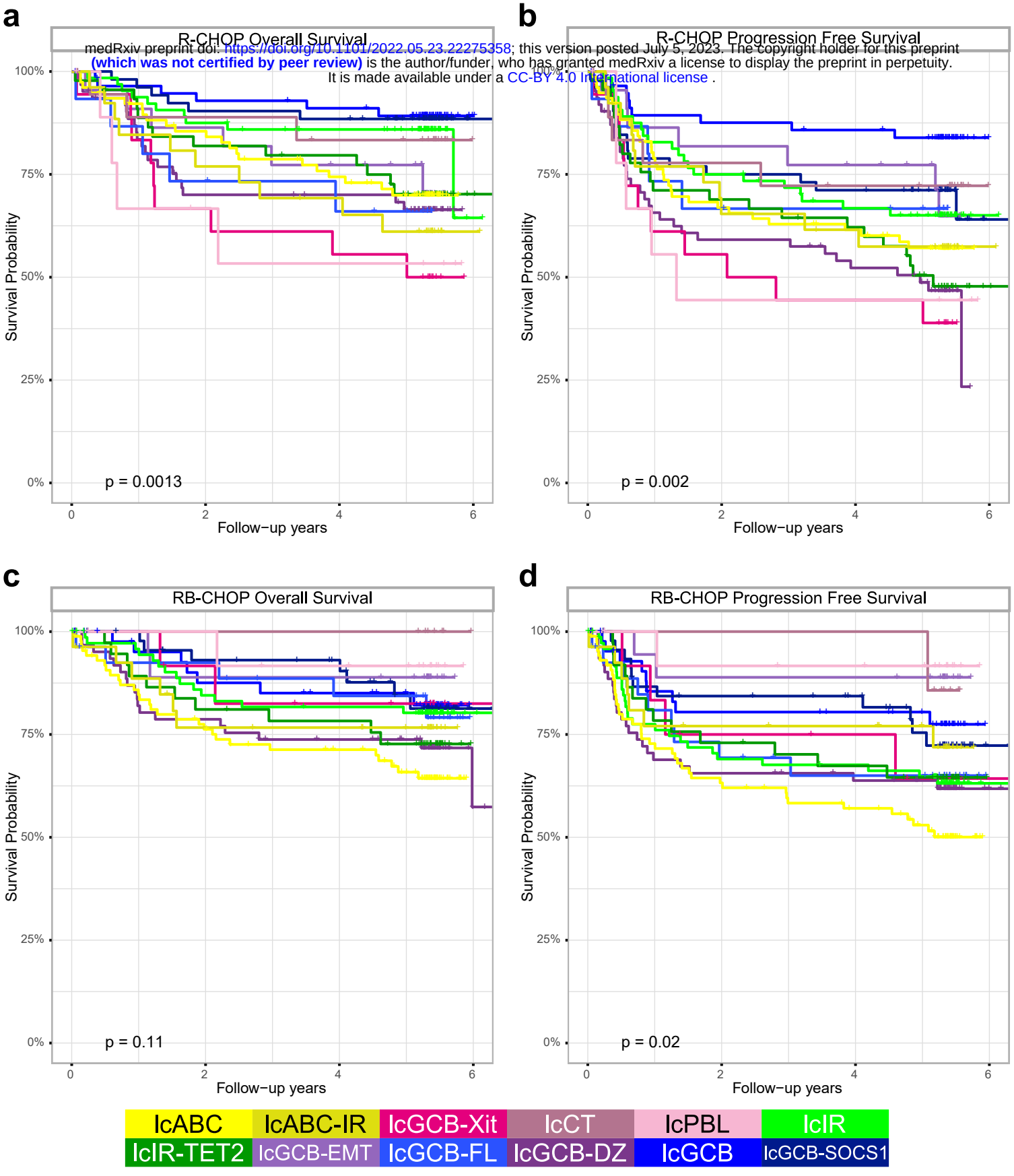

Figure 7. IcPBL and IcCT show improved survival with addition of bortezomib
Shows Kaplan-Meier plots of survival for the indicated Lymphoma Communities (LC) within the REMoDLB dataset split by treatment (R-CHOP or with addition of bortezomib; RB-CHOP) and survival (overall-survival:OS, progression-free-survival:PFS). (**a**) R-CHOP OS, (**b**) R-CHOP PFS, (**c**) RB-CHOP OS and (**d**) RB-CHOP PFS. P-value from log-rank test.

a

	lcABC	lcABC-IR	lcGCB-Xit	lcCT	lcPBL	lcIR	lcIR-TET2	lcGCB-EZMT	lcGCBFL	lcGCBDZ	lcGCB2	lcGCB-SCGS1	Total
TABC	120	47	6	14	7	35	16	0	1	1	0	2	249
TGCB	30	4	18	11	6	48	22	38	40	72	87	92	468
TMHG	2	0	5	0	7	0	1	0	0	49	16	3	83
TUNC	12	2	1	2	3	55	43	5	1	1	1	2	128
Total	164	53	30	27	23	138	82	43	42	123	104	99	928

b

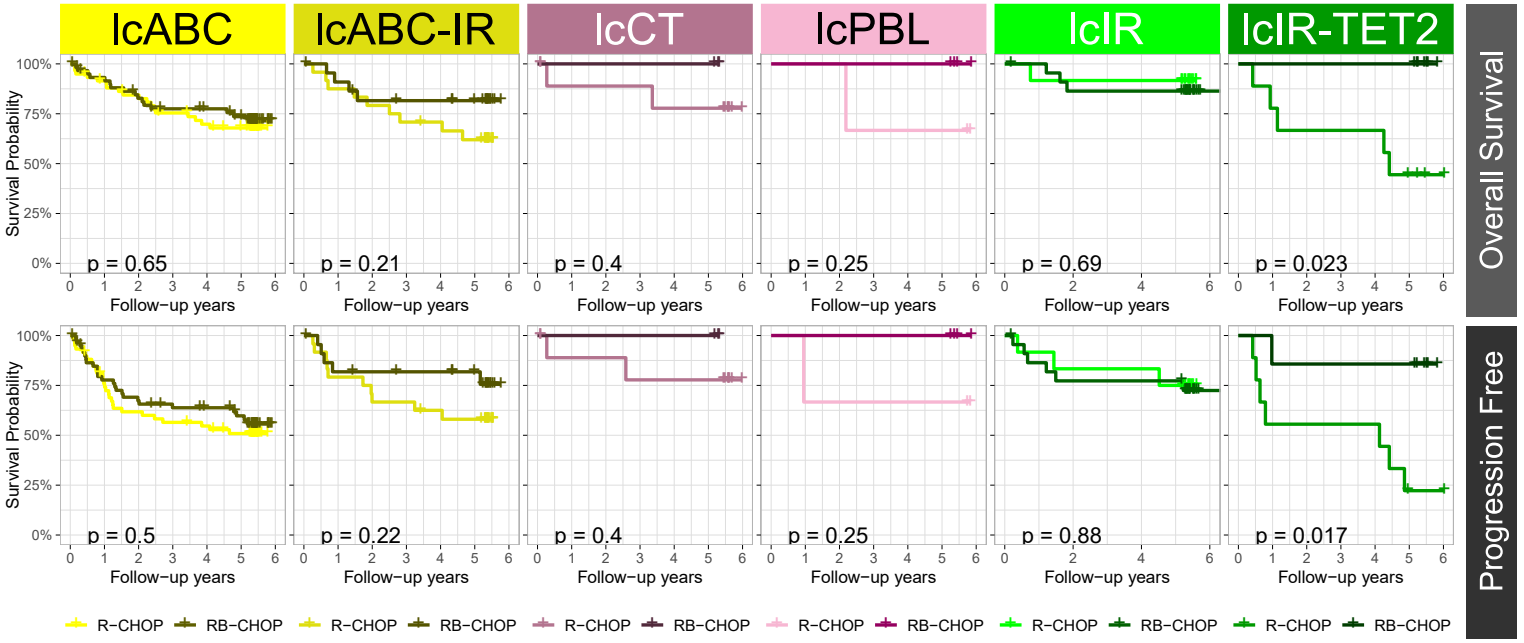

Figure 8. High confidence ABC does not show significant survival advantage with addition of bortezomib
(**a**) shows the overlap between REMoDLB trial COO assignments (rows) and Lymphoma Communities (columns). The samples used in part b have a black border. (**b**) Shows Kaplan-Meier plots of survival for Lymphoma Community overlap subsets of the TABC group. Each plot compares the treatments: R-CHOP (lighter-line) and RB-CHOP (darker-line) with overall-survival (top row) and progression-free-survival (bottom row). P-value from log-rank test.