A comprehensive modular map of Diffuse Large B-cell Lymphoma

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

[1]Division of Haematology and Immunology, Leeds Institute of Medical Research, University of Leeds, Leeds, LS9 7TF, UK;
[2]Bioinformatics Group, School of Molecular and Cellular Biology, University of Leeds, UK;
[3]Epidemiology and Cancer Statistics Group, Department of Health Sciences, University of York, York, UK;
[4]Haematological Malignancy Diagnostic Service, St. James's Institute of Oncology, Leeds, UK;
[5]Institute of Basic Medicine and Cancer of Chinese Academy of Sciences, Economic Technology Dev Zone, Hangzhou, 310000, Zhejiang province, China;
[6]University of Southampton, Southampton, UK;
[7]Division of Cellular and Molecular Pathology, Department of Pathology, University of Cambridge, Cambridge CB20QQ, UK;
[8]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 from single studies. We apply a customised network analysis to nearly five thousand DLBCL cases to identify and quantify modules of tumour biology. Shared patterns of activity in these tumour biology modules identify biologically similar DLBCL, which resolve into communities of related cases. These lymphoma communities 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 and rearrangement status. A notable example are communities linked to distinct germinal centre states that separate DLBCL respectively associated with *MYC/BCL2* and *MYC/BCL6* double hit rearrangement. By exploiting network analysis across a large number of well characterised cases, this study provides the first birds eye view of DLBCL tumour biology.

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, that sits alongside expression and mutation-based subdivisions, is separation by gene rearrangement status. Wherein the identification of high-risk DLBCL cases based on double hit (DH) or triple hit (TH) rearrangements of *MYC* and *BCL2* and/or *BCL6* genes has provided the central paradigm.[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] More recently high risk DLBCL cases related to those with rearrangement of *MYC* and *BCL2* 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* (EZB MYC+).[2,4,5]

Single cell expression analysis of germinal centre B-cells has provided additional insight, separating features of the two main functional populations of dark and light zones, in which B-cells respectively undergo proliferation, somatic hypermutation and T-cell mediated selection, along with intermediate populations including those transitioning to post-GC differentiation.[16-21] Such analyses have demonstrated that DLBCL can be subdivided based on similarity to physiological single cell expression patterns of GC B-cells.[16] However, single cell studies have also revealed that neoplastic GC B-cells exhibit aberrant patterns of expression and may not faithfully copy physiological counterparts.[18] Advanced deconvolution approaches starting from single cell derived expression profiles of multiple cell populations have further shown that DLBCL can be assigned to ecotypes reflecting the calculated contributions of multiple cell states, including both neoplastic B-cells and accompanying host response.[22] In a parallel approach, externally derived functional

79 expression signatures have been used to define lymphoma microenvironments which
80 identify heterogeneity of host response across both conventional COO classes and genomic
81 DLBCL categories.[23] These approaches conceptually share the use of externally derived
82 signatures to interrogate tumour expression profiles, and have particular value in exploring
83 tumour microenvironment. However, the alignment between expression and genomic
84 assessments is not substantially enhanced. Furthermore, a recurrent feature shared by
85 refined expression based and mutation classifications is a substantial fraction of cases that
86 are not expressly classified.
87 Consequently, an integrated picture unifying features of expression, mutation and
88 rearrangement status is still developing. Recent studies raise the question of whether the
89 inevitable conclusion of more refined expression analysis is a divergence from, rather than
90 convergence with, recurrent patterns of genomic features. Here we address this issue using
91 an approach anchored on expression correlation and drawing on the resource of close to
92 5000 cases analysed in DLBCL research studies to date. In contrast to previous studies, this
93 approach views expression patterns directly from the nature of intrinsic co-expression
94 structure eschewing the lens of externally derived signatures. The resultant integrated
95 overview allows separation of consistent DLBCL communities that significantly illuminate
96 relationships between tumour cell state, microenvironmental expressed biology, underlying
97 indolent lymphoma and genetic features of disease.
98

Methods

(See supplemental methods)

Expression datasets

Thirteen gene expression datasets were used for parsimonious gene correlation network analysis (PGCNA)(GSE4475, GSE4732, GSE10846, GSE12195, GSE19246, GSE22470, GSE31312, GSE32918, GSE34171, GSE53786, GSE87371, GSE98588, Monti).[2,12,13,24-33] For validation 4 additional datasets were used HMRN (GSE181063), NCI (NCICCR-DLBCL), REMoDLB (GSE117556) and Reddy (EGAS00001002606), (Supplemental Table 1).[3,5,13,27,34,35] For RNA-seq datasets (NCI, Reddy) the 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 2). 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/DLBCL.

Network analysis

Neighbourhoods

The PGCNA edge file was split at module level and clustered 5,000 times using Leidenalg. Convergence onto the same answer was observed across multiple runs, the best clustering (based on modularity score) was retained for each module.

Module Expression Values and Lymphoma communities

Module/Neighbourhood Expression Values (MEVs/NEVs) were derived using the median Z-score of the 10 most representative genes per module/neighbourhood. To derive DLBCL communities a combination of 53 MEV/NEVs was selected for downstream analysis (23 MEVs (excluding M9, M11, M21, M24, M27 & M28), M9 NEVs (n=16), M11 NEVs (n=15, excluding n16 & n17 < 10 genes). The resulting HMRN data transposed matrix (451 x 53) was analysed with pgcna2 (-f 1, -e 3 -n 10000) with 10,000 clusterings, selecting the best clustering based on modularity score. Multiple runs converged on the same answer, splitting HMRN data into 12 lymphoma communities (LC).

Lymphoma Community machine learning tool

A machine learning classifier was trained on the HMRN Lymphoma Communities using the 54 MEV/NEVs per sample as input (addition of M24; gender). The dataset was split (test_train_split function), stratifying on the community label, to give class balanced randomised training (n=267) and a validation (n=70) dataset which was held-out to test the final selected model. Stratified 5-fold cross-validation across 8 machine learning methods sampling 9,040 parameters, scoring with Matthews correlation coefficient (MCC) was performed on training data. The best 6 models were combined using soft voting to give the final ML_LC (machine learning lymphoma communities) assignments, with mean MCC of 0.89 across 5-folds of test dataset and held-out validation data. ML_LC assignments were used to subdivide validation datasets, generating p-values for each community per sample.

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. The correlation between mutation and MEV/NEVs was calculated for each dataset using Spearman's rank correlations, the p-values converted to Z-scores and filtered (<0.01) before visualisation. Significance of overlap between mutations and communities 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.

Results

An integrated gene expression context for DLBCL

To provide a comprehensive context in which to explore the biology of DLBCL from the perspective of gene expression, in a fashion which was data-led and independent of prior assumptions, we employed parsimonious gene correlation network analysis (PGCNA) (Supplemental Figure 1).[36] 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.[36,37] Since the primary determinant of resolved modules is the pattern and consistency of gene co-expression in DLBCL data our approach differs fundamentally from those that learn from external signatures or expression patterns to explore and impute cellular contributions and differentiation states.[22,23]

We divided the large body of existing DLBCL data sets into two components. For network discovery we used gene expression derived from 14 DLBCL datasets encompassing a total of 2,505 cases.[2,12,13,24-33] The network derived from these datasets benefits from accessing a highly representative sample of DLBCL expression data derived from multiple sources mitigating against biases linked to case selection and platform features. Recent datasets that combine mutation and expression features and include an additional 2,484 cases were held back, providing independent validation data.[3,5,14,34,35]

The resulting DLBCL expression network resolved into 28 modules of gene co-expression, across 16,054 genes/nodes, which can be visualised and interrogated interactively (Figure 1a, https://mcare.link/DLBCL, Supplemental Table 1). A comprehensive gene ontology and

190 signature enrichment was used to assess biological features associated with these modules.
191 Representative terms were selected to identify notable expression features associated with
192 each module (Figure 1b, Supplemental Figure 2, Supplemental Table 2).
193 A number of expectations were met including resolution of two modules encompassing
194 features of the ABC (M11) and GCB state (M9). Additional modules related to cell growth,
195 cell division and metabolism for example separated MYC function, cell cycle, glycolysis, and
196 sterol biosynthesis; host response/microenvironment was separated into modules related
197 to stromal/angiogenesis features, T/NK-cells, monocytes, IFN responses, Immunoglobulin
198 genes, and MHC class-II/antigen presentation; highly co-ordinated gene batteries were
199 resolved for nucleosome components, homeobox, immediate early and metallothionein
200 genes. Several modules related to structural chromosomal regions (e.g. chr6, chr7, chr17p,
201 chrX and chrY). Relationships of genes associated with COO and CCC were recovered. The
202 classifier genes used for separation of ABC and GCB class separation divided discretely
203 across the two respective modules (M11 ABC: p-value $6.95x10^{-19}$ and M9 GCB: $1.57x10^{-11}$),[38]
204 while the more numerous CCC genes divided across multiple modules but with highly
205 significant enrichments (M9:B-cell receptor p-value $3.27x10^{-57}$, M3:Host response p-value
206 $1.64x10^{-64}$ and M13:Oxphos p-value $1.38x10^{-05}$).[12] The network thus provides an integrated
207 picture of gene expression in DLBCL encompassing details of lineage specific gene
208 expression set against a diverse backdrop of cell biology and microenvironment. These are
209 integrated into an extensive online resource covering all underlying gene correlation data,
210 interactive network visualisations and downstream comparisons.
211
212 **Meta-hazard ratio analysis confirms consistent associations between biology and outcome**
213 Gene co-expression patterns resolved in the network showed consistent associations. A
214 meta-hazard ratio analysis merging overall survival data across the discovery datasets
215 illustrated the association of individual gene expression with outcome across the entire
216 network. This included the segregation of good and adverse outcome between component
217 genes of the GCB and ABC modules (Figure 1c and d), as well as adverse outcome with
218 Mitchondrial and MYC overexpression module (M8) and Ribosome module (M7) and good
219 outcome linked to expression of the Stromal/Angiogenesis module (M4) which
220 encompasses features of known survival predictors.[13]

221 While there was a clear bias for the association of the majority of module components with
222 a particular outcome, some modules exhibited greater diversity. For example, within the
223 Monocyte module (M3), adverse risk is associated with expression of *CD163* and
224 neighbours. These include *MS4A4A* a marker of alternative (IL4-driven) macrophage
225 activation,[39] which links *CD163* to *C1Q* complement genes (Figure 1c expanded box). This is
226 consistent with studies identifying adverse outcome in DLBCL linked to alternative
227 macrophage activation and CD163 expression.[40-42] C1Q has multiple effects on macrophage
228 function including on enhancing apoptotic clearance,[43,44] and may impact on cancer immune
229 responses.[45] The expression network thus resolved patterns of gene co-expression that
230 reveal differences in outcome related to B-cell state and microenvironment.
231
232 **Resolving detailed differentiation states in GCB and ABC modules**
233 Biological detail can be resolved at different levels of granularity within the network. An
234 iterative analysis of gene correlation within modules can resolve the most highly related
235 gene neighbourhoods.[36] Assessed systematically across the network, the ABC and GCB
236 modules were amongst those with highest granular information content (Supplemental
237 Figure 3, Supplemental Table 3 and 4). Given the central importance of B cell differentiation
238 to lymphomagenesis, a focused network visualisation was generated to explore the
239 relationships of these GCB and ABC neighbourhoods (Figure 2a and b, and
240 https://mcare.link/DLBCL).
241 The GCB module (M9) resolved into 16 neighbourhoods (Figure 2b and c, Supplemental
242 Table 4). These included neighbourhoods enriched for different GC B-cell subsets - LZ
243 (M9_n2 and n6), DZ (M9_n5 and n15), CD40/NFκB responses linked to LZ B-cells (M9_n10),
244 Intermediate-e/CCR6 memory B-cell precursor (M9_n3 and n4) and pre-memory/memory
245 B-cells (M9_n14).[16] Genes used to classify GCB DLBCL in the COO classifier segregated
246 between two neighbourhoods with one outlier. Thus the original COO classification samples
247 two primary components of GC related co-expression residing at the heart of a wider
248 network of interrelated gene expression (Supplemental Figure 3b).[38]
249 A similar informative separation was observed for the ABC module which divided into 17
250 neighbourhoods (Figure 2b and d, Supplemental Table 4). This separated expression related
251 to CCR6+ memory B-cell precursors (M11_n5), pre-memory B-cells (M11_n5 and n7),
252 plasmablast/plasma cell (M11_n6). These cell state associations overlapped with targets of

253 key transcriptional regulators IRF4 (M11_n1, n6 and n7), BLIMP1 (M11_n7) and XBP1
254 (M11_n6). A further set of NFκB-response genes were distinguished, specific to the ABC
255 module (M11_n8). The COO ABC-classifier genes were distributed across 5 neighbourhoods,
256 illustrating how these genes sample multiple aspects of the ABC state (Supplemental Figure
257 3c).
258
259 **DLBCL communities are characterised by module and neighbourhood expression patterns**
260 The set of modules and neighbourhoods encompass features of consistent gene co-
261 expression in DLBCL. Their derivation is independent of external signatures and provides a
262 robust reflection of co-expression derived from DLBCL data alone. The next step in our
263 correlation-centred approach was to assess whether reducing granularity from gene to
264 module/neighbourhood level using module and neighbourhood expression values
265 (MEV/NEV) would allow DLBCL cases to be separated into communities that shared related
266 biology. This approach resolved 12 communities of cases in the previously unseen HMRN
267 DLBCL data. A machine learning classifier trained on the HMRN data retrieved a similar
268 distribution of cases falling into these communities across a range of other data sets (Figure
269 3a and Supplemental Figure 4 & 5).
270 The 12 lymphoma communities (LC) aligned to existing paradigms (Figure 3b and c,
271 Supplemental Figure 4) and broadly reinforced the segregation of ABC and GCB expression
272 patterns while providing significantly enhanced and biologically meaningful granularity.
273 Two primary communities of ABC-DLBCL related cases were distinguished as immune
274 response poor ($_{lc}$**ABC**) or immune rich ($_{lc}$**ABC-IR**). A further ABC community was principally
275 linked to memory B-cell neighbourhoods ($_{lc}$**ABC-M**). Cases with plasmablastic features
276 ($_{lc}$**PBL**) were distinguished and were notable for expression of XBP1 targets and cancer testis
277 antigens, as also seen in myeloma. Amongst cases with weak B-cell patterns, two
278 communities were dominated by immune response modules ($_{lc}$**IR-1** and $_{lc}$**IR-2**) with B-cell
279 features including ABC memory B-cell neighbourhoods. An additional community ($_{lc}$**STR**)
280 combined strong Stromal/EMT/Angiogenesis and CD40/NFκB related expression with
281 modest GCB neighbourhood expression. Five communities were characterised by expression
282 of multiple GCB related neighbourhoods. These included a community with distinctively
283 mixed GCB and ABC expression features along with cell cycle, MYC, and sterol biosynthesis
284 modules. These features are in keeping with transition from GCB to an ABC/post-GC state

285 ($_{lc}$**GCB-Xit**, Figure 3d). Similar proliferation and growth-related module expression was
286 combined with polarised high GC DZ and low GC LZ neighbourhood expression, along with
287 low CD40/NFκB and MHC-II genes in an apparent GCB DZ-like community ($_{lc}$**GCB-DZ**). The
288 remaining three communities were distinguished based on the consistency of GCB
289 neighbourhood and host response modules as $_{lc}$**GCB-1**, $_{lc}$**GCB-2** and $_{lc}$**GCB-3**.
290

291 **Transition of expression states across DLBCL communities**

292 We next took advantage of the combination of overview and gene level granularity provided
293 by the network visualisation to assess the relationships between the resolved communities
294 across all network genes or those specific to the GCB/ABC differentiation state, generating a
295 "birds eye view" of expression differences. These analyses illustrate the shifting patterns of
296 gene co-expression separating the resolved DLBCL community structure (Figure 3d and
297 Supplemental Figure 6 and 7). This is particularly exemplified at the level of GCB and ABC
298 neighbourhood genes, where the $_{lc}$GCB-Xit community straddles both key GCB and ABC
299 features and contrasts with the more discrete patterns of other communities such as $_{lc}$GCB-
300 DZ and $_{lc}$ABC (Figure 3d). For example, the $_{lc}$GCB-Xit community distinctively co-expresses
301 features of both the GCB (e.g. *MYBL1*) and ABC (e.g. *BATF*, *CYB5R2*) states, which in the
302 wider context of DLBCL are strongly anti-correlated. This co-expression also extends to
303 primary COO classifier genes. Implicit in the binary separation of DLBCL into ABC and GCB
304 classes is the fact that some cases are difficult to classify.[38] Much of this challenge has been
305 attributed to cases dominated by host response features, where the B-cell derived signal is
306 essentially diluted by microenvironmental features.[46] These are captured in the community
307 structure here as $_{lc}$IR-1, $_{lc}$IR-2 or $_{lc}$STR. Here the identification of the $_{lc}$GCB-Xit community
308 establishes another important feature of DLBCL biology, an intrinsically transitional
309 expression state, as a biological feature of a subset of DLBCL.
310

311 **Lymphoma communities and ecosystems**

312 One of the most advanced methods for exploring cellular composition in DLBCL is the
313 recently proposed concept of ecotyping.[22] Signatures of specific cell states, defined from
314 single cell analysis, are used to deconvolute cellular composition for individual samples. The
315 imputed cell states are then analysed using a community detection approach to assemble
316 cases into ecotypes sharing patterns of cellular composition. We were therefore interested

317 to explore the relationship between lymphoma communities and ecotypes in the same data.
318 One difference lay in the fact that ecotype assignments were more restrictive with between
319 26-32% of cases unassigned across data sets (Figure 4a). The cases unassigned to ecotypes
320 were distributed across a range of communities but $_{lc}$ABC-1, $_{lc}$GCB-Xit and $_{lc}$GCB-DZ were
321 particularly enriched amongst the unassigned cases (Figure 4b; LE_UNC). Amongst cases
322 with lymphoma ecotype assignments (LE1-9) significant overlaps with communities were
323 observed, with LE1 enriched for $_{lc}$ABC-1 and LE8 enriched for $_{lc}$GCB-1 and -3 communities.
324 LE5 overlapped with $_{lc}$ABC-Mem indicating agreement for the detection of cases with
325 preferential memory B-cell differentiation features and ecotype cell states (Figure 4c)
326 emphasised the distinctive microenvironmental association in these cases. Agreement was
327 also observed for LE9 which primarily enriched for $_{lc}$STR. For immune response rich cases
328 those assigned to LE4 belonged variously to $_{lc}$IR-1, $_{lc}$IR-2 and $_{lc}$ABC-IR communities. Ecotype
329 LE3 was primarily enriched for $_{lc}$GCB-DZ, but much of $_{lc}$GCB-DZ along with $_{lc}$GCB-Xit were
330 unassigned to ecotype. Thus, deconvolution-centred and correlation-centred analytical
331 approaches converge onto several recurrent elements of DLBCL gene expression. However,
332 differences are apparent relating in particular to immune rich ABC cases ($_{lc}$ABC-IR), and
333 specific GCB states.
334
335 **Lymphoma communities have prognostic significance**
336 Gene expression and mutation-based classifications allow prognostic separation of DLBCL.
337 Therefore, in order to consider the lymphoma community structure as a valid framework
338 within which to understand existing classifications an expectation is that DLBCL
339 communities should show significant and reproducible survival differences. We addressed
340 this across the 4 representative validation datasets, also used in the ecotype comparisons,
341 restricting to cases treated with R-CHOP chemo-immunotherapy.[3,5,13,27,34,35] Across multiple
342 datasets the DLBCL community structure separated risk consistently (Figure 5a and b). Most
343 adverse risk, in terms of meta-HR, was observed for assignment of cases to $_{lc}$PBL, followed
344 by $_{lc}$ABC-IR, $_{lc}$ABC, $_{lc}$GCB-Xit and $_{lc}$GCB-DZ. Particularly good risk was associated with $_{lc}$STR,
345 $_{lc}$GCB-2 and to a lesser extent $_{lc}$GCB-1 and GCB-3.
346 At a gene level we identified alternative macrophage expression patterns as a prognostic
347 feature, we therefore asked whether the expression of such macrophage genes was a
348 specific feature of DLBCL communities. Indeed, alternative macrophage gene expression

349 was a prominent feature both of the poor risk $_{lc}$ABC-IR community and good/indeterminate
350 risk $_{lc}$IR-1/-2 communities (Supplemental Figure 8). While these communities are joined in
351 ecotype LE4, the lymphoma community analysis suggests that alternative macrophage
352 activation can be a feature of either very poor risk disease accompanying a detectable ABC
353 expression pattern, or relatively good/indeterminate risk DLBCL with dominant immune
354 response features. Hence, the ability to resolve associations between host response
355 features and outcome can be further informed by consideration of modular and
356 neighbourhood level gene expression patterns. We conclude that the DLBCL community
357 structure is associated with consistent prognostic differences in relation to current standard
358 of care and thus provides both a biologically and clinically meaningful framework.
359

360 **DLBCL communities link to mutation patterns**

361 The most significant recent advance in understanding DLBCL biology has come from analysis
362 of recurrent mutation patterns. Such studies have converged onto recurrent patterns across
363 multiple data sets.[2-5] However, the integration between mutation and gene expression
364 defined categories remains incomplete. An important test of the DLBCL communities was
365 therefore to assess to what extent these help to explain the relationship between mutation
366 and expression states. To address this we analysed the enrichment of mutations in DLBCL
367 communities by integrating enrichment p-values derived across the individual HMRN,
368 REMoDL-B and Reddy data-sets.[5,34,35] At a p-value threshold <0.01 (in ≥ 2 datasets) the
369 communities showed significant and distinct associations with mutation type (Figure 6a).
370 $_{lc}$ABC was enriched for a characteristic combination of *MYD88, CD79B, PIM1, ETV6,*
371 *TBL1XR1, BTG2*, *PRDM1*, *CDKN2A*, *BTG1*, *IRF4* (in order of significance). The two variant
372 communities $_{lc}$ABC-IR and $_{lc}$ABC-M showed selective associations with *MYD88* and either
373 *PIM1* or *CD79B* mutation respectively, while *MYD88* mutation was anti-correlated with all
374 GCB communities and $_{lc}$IR1/2. GCB communities shared enrichment for *BCL2* and *EZH2*
375 mutation while other specific associations separated $_{lc}$GCB-1 and $_{lc}$GCB-3 from $_{lc}$GCB-2 and
376 $_{lc}$GCB-DZ or $_{lc}$GCB-Xit. For example, $_{lc}$GCB-1 and -3 shared enrichment of *NFKBIA*, *SGK1*, and
377 *SOCS1* mutations but lacked enrichment for *KMT2D* or *CREBBP* mutations. While, $_{lc}$GCB-2
378 and $_{lc}$GCB-DZ shared the reciprocal pattern. And $_{lc}$GCB-DZ showed enrichment for mutation
379 in *FOXO1*, *MYC*, *DDX3X*, *PTEN, KRAS* and *TP53*, while $_{lc}$GCB-Xit showed selective enrichment
380 for *BCL10*, *CD70, TMEM30A* and *CCND3* mutations. Consistent with these community level

381 patterns, concordant associations could be resolved at the level of individual module and
382 neighbourhood correlations with mutation state (Supplemental Figure 9).[2,4,5]
383 The observed associations between communities and mutations resembled patterns
384 underlying current mutational classifications. To directly address this association we
385 assessed mutational classes assigned in HMRN data with either the LymphGen (Figure 6b),[47]
386 or the local HMRN classification (Supplemental Figure 10).[5] These classifications though
387 independently derived,[4,5] are largely concordant,[47] and recapitulate the features of the C1-
388 C5 Harvard classification.[2] The LymphGen MCD class showed the expected association with
389 $_{lc}$ABC-1 and $_{lc}$ABC-M, with $_{lc}$ABC-1 also including some cases belonging to the N1 category.
390 The similarly poor risk $_{lc}$ABC-IR community was notable for failing to be enriched for MCD
391 assignments but was enriched in the related HMRN MYD88 mutation class. This likely
392 reflects differing emphasis on the *MYD88* mutation state in particular, combined potentially
393 with the diluting effect of rich immune infiltrates on variant allele detection. In contrast,
394 LymphGen BN2 was significantly and selectively associated with $_{lc}$GCB-Xit. EZB was most
395 significantly enriched amongst $_{lc}$GCB-2 and was significantly depleted from $_{lc}$GCB-Xit. ST2
396 differed from EZB in showing selective enrichment amongst $_{lc}$GCB-1 but no other $_{lc}$GCB
397 community. EZB-MYC+ was selectively enriched amongst $_{lc}$GCB-DZ. Cases with an
398 ambiguous classification of EZB/ST2 were restricted in enrichment to the icSTR community.
399 Similar patterns of association were evident when considering the associations of DLBCL
400 communities with the independent but related HMRN mutational clustering of DLBCL cases
401 (Supplemental Figure 10). We conclude that the mutation-based subdivision of DLBCL in
402 LymphGen and related classifications, overlaps significantly with lymphoma communities
403 derived independently and *ab initio* from gene co-expression patterns.
404
405 **LC underlines distinctions between double hit lymphomas**
406 A further substantial body of work has focused on the separation of aggressive DLBCL based
407 on gene rearrangement status. Wherein, the occurrence of double or triple hit features with
408 rearrangement of *MYC* and *BCL2* (MYC_BCL2-DH) or *MYC* and *BCL6* (MYC_BCL6-DH), or
409 *MYC*, *BCL2* and *BCL6* (MYC_BCL2-TH) identifies high risk disease.[8-11] The relationships of
410 these rearrangement based categories to gene expression and mutation features is
411 relatively well defined for MYC_BCL2-DH, but remains ill-defined for MYC_BCL6-DH leaving
412 the latter category potentially adrift in classification schemes.[48-50] Of the HMRN cohort 107

cases had rearrangement of one or more of these oncogenes. $_{lc}$GCB-DZ and $_{lc}$GCB-Xit communities showed selective and reciprocal associations with rearrangement status (Figure 6b). $_{lc}$GCB-DZ was selectively enriched for MYC_BCL2-DH and MYC_BCL2_BCL6-TH cases but was not significantly enriched for other rearrangement states. In contrast, $_{lc}$GCB-Xit cases were selectively enriched for MYC_BCL6-DH and MYC single hit (SH) status but were not associated with any combination of *BCL2* rearrangement. This supports the concept that MYC_BCL6-DH, although often classified together with MYC_BCL2-DH lymphoma, in fact belong to a biologically distinct entity. BCL2-SH were enriched in $_{lc}$GCB-2 and $_{lc}$GCB-3 with the most significant enrichment being in $_{lc}$GCB-2 (p-value=$7.2x10^{-08}$).

Both *BCL2* rearrangement and EZB mutational features are features of follicular lymphoma (FL).[5,51] We therefore asked whether lymphoma communities were differentially linked to the diagnostic status of DLBCL with evidence of underlying FL, a feature which might be broadly associated with GCB-DLBCL as a whole. Strikingly of the GCB lymphoma communities $_{lc}$GCB-2 was highly significantly enriched for cases with a diagnosis of underlying FL (defined as either antecedent follicular lymphoma or composite nodes at diagnosis) (Figure 6b; p-value= $9.9x10^{-11}$). Indeed, this community was significantly depleted of cases diagnosed as DLBCL, not otherwise specified. As illustrated previously $_{lc}$GCB-2 had very good outcome in the available follow up across multiple datasets of R-CHOP treated patients (Figure 5b-c). The DLBCL lymphoma community thus links expression patterns to mutation state, rearrangement status and the presence of underlying indolent lymphoma.

Discussion

Heterogeneity is a dominant feature of DLBCL biology. This has led to a number of partially intersecting taxonomies; with little argument the most accepted of these are the cell of origin classification, and the recent mutational classifications defined in the LymphGen, Harvard and HMRN classifications.[1-4] However, despite these major advances and the profound insights these have produced, there remains uncertainty about the inter-relationships between these classifications and ambiguous patterns of mutation or expression, rearrangement status or even broad diagnostic category.

443 Here we have addressed this problem with the aim of presenting a common framework
444 within which to explore relationships between various features of DLBCL tumour biology.
445 The framework is gene expression-based, and uses an inclusive, correlation-centred
446 approach, to derive modules of co-expressed genes that reflect inherent features of the
447 DLBCL data. The resolution of these modules and the underlying neighbourhood
448 substructure is independent of externally derived signatures or prior knowledge regarding
449 patterns of gene co-expression. Signature and gene ontology databases are used to
450 understand the biology identified within the modules of gene co-expression observed in
451 DLBCL data, but the modules and neighbourhoods themselves are derived independently of
452 such prior knowledge. This approach differs fundamentally from the use of externally
453 derived signature or gene co-expression patterns as the primary tool with which to explore
454 and deconvolute gene expression in DLBCL. Our analysis learns from 2,500 cases drawing
455 on principle contributions of the DLBCL research community and validates in nearly 2,500
456 additional cases from recent studies with mutational and outcome data.[2,12,13,24-33] The
457 combination of this extensive combined sample number, and the inclusive "leave no case
458 behind" community clustering has allowed the identification of significant heterogeneity at
459 the expression level, which extends even beyond that of recent deconvolution-centred
460 single cell led approaches to illustrate key features of biology.
461 These features, summarised in Supplemental Figure 11 and 12, include the distinction of
462 immune rich communities of ABC DLBCL with atypical macrophage activation patterns, the
463 expression-based distinction between double hit lymphomas with *MYC* and *BCL2* or *MYC*
464 and *BCL6* translocation, the identification of a DLBCL community highly significantly linked
465 to underlying follicular lymphoma, and separation at expression level of communities
466 associated with mutational subtypes including distinction between BN2, EZB and ST2 types.
467 This community structure both reinforces and extends existing paradigms by demonstrating
468 inter-relationships and important differences at a modular level of tumour biology which
469 are linked to differentiation state and microenvironment. The community analysis agrees
470 with many assignments reached using the deconvolution-centred signature-led EcoTyper
471 approach, which focuses in particular on cell states in the microenvironment.[22] But the
472 community analysis and network structure reinforce the central importance of the ABC and
473 GCB cell state subdivision and refined features of these cell states. These refined cell state

features clarify the relationships of cases that are harder to assign in a binary classification and help to explain heterogeneity in clinical responses.

A particular example of the utility of the approach is provided by the identification of the poor risk $_{lc}$GCB-Xit community. This community of cases has a pattern of mixed ABC and GCB features, which we hypothesise is related to a poised GCB exit state. This is reflected in relatively low confidence of COO class assignment for many such cases, and a community of cases which is largely unassigned amongst ecotypes. Yet this community is distinctively enriched for MYC_BCL6-DH status and BN2 mutational features. It is also consistently associated with poor outcome amongst R-CHOP treated cases illustrating how the BN2 category might be valuably subdivided.

We conclude that our network-based approach yields an encompassing map of DLBCL tumour biology and provides an encompassing framework to support targeting of phenotypic vulnerabilities in clinical trials. Illustrating the power of data integration, the analysis illuminates the tumour expression landscape and identifies communities of cases that are linked by differentiation state, microenvironment, mutational pattern, rearrangement status and underlying indolent lymphoma association.

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 the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. D.J.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). 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).

543 17. King, H.W., *et al.* Single-cell analysis of human B cell maturation predicts how
544 antibody class switching shapes selection dynamics. *Sci Immunol* **6**(2021).
545 18. Milpied, P., *et al.* Human germinal center transcriptional programs are de-
546 synchronized in B cell lymphoma. *Nat Immunol* **19**, 1013-1024 (2018).
547 19. Attaf, N., Baaklini, S., Binet, L. & Milpied, P. Heterogeneity of germinal center B cells:
548 New insights from single-cell studies. *Eur J Immunol* **51**, 2555-2567 (2021).
549 20. Duan, L., *et al.* Follicular dendritic cells restrict interleukin-4 availability in germinal
550 centers and foster memory B cell generation. *Immunity* **54**, 2256-2272 e2256 (2021).
551 21. Kennedy, D.E., *et al.* Novel specialized cell state and spatial compartments within the
552 germinal center. *Nat Immunol* **21**, 660-670 (2020).
553 22. Steen, C.B., *et al.* The landscape of tumor cell states and ecosystems in diffuse large
554 B cell lymphoma. *Cancer Cell* **39**, 1422-1437 e1410 (2021).
555 23. Kotlov, N., *et al.* Clinical and Biological Subtypes of B-cell Lymphoma Revealed by
556 Microenvironmental Signatures. *Cancer discovery* **11**, 1468-1489 (2021).
557 24. Compagno, M., *et al.* Mutations of multiple genes cause deregulation of NF-kappaB
558 in diffuse large B-cell lymphoma. *Nature* **459**, 717-721 (2009).
559 25. Williams, P.M., *et al.* A novel method of amplification of FFPET-derived RNA enables
560 accurate disease classification with microarrays. *J Mol Diagn* **12**, 680-686 (2010).
561 26. Salaverria, I., *et al.* Translocations activating IRF4 identify a subtype of germinal
562 center-derived B-cell lymphoma affecting predominantly children and young adults.
563 *Blood* **118**, 139-147 (2011).
564 27. Frei, E., *et al.* Addition of rituximab to chemotherapy overcomes the negative
565 prognostic impact of cyclin E expression in diffuse large B-cell lymphoma. *J Clin*
566 *Pathol* **66**, 956-961 (2013).
567 28. Care, M.A., *et al.* SPIB and BATF provide alternate determinants of IRF4 occupancy in
568 diffuse large B-cell lymphoma linked to disease heterogeneity. *Nucleic Acids Res* **42**,
569 7591-7610 (2014).
570 29. Monti, S., *et al.* Integrative analysis reveals an outcome-associated and targetable
571 pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer*
572 *Cell* **22**, 359-372 (2012).
573 30. Hummel, M., *et al.* A biologic definition of Burkitt's lymphoma from transcriptional
574 and genomic profiling. *The New England journal of medicine* **354**, 2419-2430 (2006).
575 31. Dave, S.S., *et al.* Molecular diagnosis of Burkitt's lymphoma. *The New England*
576 *journal of medicine* **354**, 2431-2442 (2006).
577 32. Scott, D.W., *et al.* Determining cell-of-origin subtypes of diffuse large B-cell
578 lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood*
579 **123**, 1214-1217 (2014).
580 33. Dubois, S., *et al.* Biological and Clinical Relevance of Associated Genomic Alterations
581 in MYD88 L265P and non-L265P-Mutated Diffuse Large B-Cell Lymphoma: Analysis of
582 361 Cases. *Clin Cancer Res* **23**, 2232-2244 (2017).
583 34. Davies, A., *et al.* Gene-expression profiling of bortezomib added to standard
584 chemoimmunotherapy for diffuse large B-cell lymphoma (REMoDL-B): an open-label,
585 randomised, phase 3 trial. *Lancet Oncol* **20**, 649-662 (2019).
586 35. Reddy, A., *et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma.
587 *Cell* **171**, 481-494 e415 (2017).

588 36. Care, M.A., Westhead, D.R. & Tooze, R.M. Parsimonious Gene Correlation Network
589 Analysis (PGCNA): a tool to define modular gene co-expression for refined molecular
590 stratification in cancer. *NPJ Syst Biol Appl* **5**, 13 (2019).
591 37. Stephenson, S.*, et al.* Growth Factor-like Gene Regulation Is Separable from Survival
592 and Maturation in Antibody-Secreting Cells. *J Immunol* **202**, 1287-1300 (2019).
593 38. Wright, G.*, et al.* A gene expression-based method to diagnose clinically distinct
594 subgroups of diffuse large B cell lymphoma. *Proc Natl Acad Sci U S A* **100**, 9991-9996
595 (2003).
596 39. Sanyal, R.*, et al.* MS4A4A: a novel cell surface marker for M2 macrophages and
597 plasma cells. *Immunol Cell Biol* **95**, 611-619 (2017).
598 40. Marchesi, F.*, et al.* High density of CD68+/CD163+ tumour-associated macrophages
599 (M2-TAM) at diagnosis is significantly correlated to unfavorable prognostic factors
600 and to poor clinical outcomes in patients with diffuse large B-cell lymphoma.
601 *Hematol Oncol* **33**, 110-112 (2015).
602 41. Wada, N.*, et al.* Tumour-associated macrophages in diffuse large B-cell lymphoma: a
603 study of the Osaka Lymphoma Study Group. *Histopathology* **60**, 313-319 (2012).
604 42. Vajavaara, H.*, et al.* Prognostic impact of soluble CD163 in patients with diffuse large
605 B-cell lymphoma. *Haematologica* **106**, 2502-2506 (2021).
606 43. Thielens, N.M., Tedesco, F., Bohlson, S.S., Gaboriaud, C. & Tenner, A.J. C1q: A fresh
607 look upon an old molecule. *Mol Immunol* **89**, 73-83 (2017).
608 44. Bobak, D.A., Gaither, T.A., Frank, M.M. & Tenner, A.J. Modulation of FcR function by
609 complement: subcomponent C1q enhances the phagocytosis of IgG-opsonized
610 targets by human monocytes and culture-derived macrophages. *J Immunol* **138**,
611 1150-1156 (1987).
612 45. Roumenina, L.T.*, et al.* Tumor Cells Hijack Macrophage-Produced Complement C1q
613 to Promote Tumor Growth. *Cancer immunology research* **7**, 1091-1105 (2019).
614 46. Care, M.A., Westhead, D.R. & Tooze, R.M. Gene expression meta-analysis reveals
615 immune response convergence on the IFNgamma-STAT1-IRF1 axis and adaptive
616 immune resistance mechanisms in lymphoma. *Genome medicine* **7**, 96 (2015).
617 47. Runge, H.F.P.*, et al.* Application of the LymphGen classification tool to 928 clinically
618 and genetically-characterised cases of diffuse large B cell lymphoma (DLBCL). *Br J*
619 *Haematol* (2020).
620 48. Rosenwald, A.*, et al.* Prognostic Significance of MYC Rearrangement and
621 Translocation Partner in Diffuse Large B-Cell Lymphoma: A Study by the Lunenburg
622 Lymphoma Biomarker Consortium. *J Clin Oncol* **37**, 3359-3368 (2019).
623 49. Cucco, F.*, et al.* Distinct genetic changes reveal evolutionary history and
624 heterogeneous molecular grade of DLBCL with MYC/BCL2 double-hit. *Leukemia* **34**,
625 1329-1341 (2020).
626 50. Scott, D.W.*, et al.* High-grade B-cell lymphoma with MYC and BCL2 and/or BCL6
627 rearrangements with diffuse large B-cell lymphoma morphology. *Blood* **131**, 2060-
628 2064 (2018).
629 51. Okosun, J.*, et al.* Integrated genomic analysis identifies recurrent mutations and
630 evolution patterns driving the initiation and progression of follicular lymphoma. *Nat*
631 *Genet* **46**, 176-181 (2014).

Figure Legends

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 Supplemental 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. Highlighted are the correlated gene neighbours in the adverse risk macrophage component of module M3, illustrating the inter-relationships of genes identified in text. **(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. The highlighted macrophage genes (*C1QB*, *C1QC*, *CD163* and *MS4A4*) are shown as red circles for M3.

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 Supplemental Figure 3 and Supplemental Table 4.

Figure 3. Module and neighbourhood expression values and Lymphoma Communities (LC)

HMRN DLBCL dataset was used to derive consistent communities of DLBCL cases based on expression across network modules with GCB_M9 and ABC_M11 used at neighbourhood level granularity. Twelve consistent Lymphoma Communities (LC) were resolved and used to train a machine learning tool (ML_LC). (**a**) the fraction of each LC across the datasets classified using ML_LC. **(b)** HMRN cases hierarchically clustered within each LC group (ML_LC). MEV and NEV are shown on a blue (low) to red (high) z-score colour scale. Modules and neighbourhoods are separated across the y-axis and colour coded for GCB (blue) and ABC (yellow) neighbourhoods. For version that includes meta-data see high resolution version in Supplemental Figure 5. **(c)** Patterns of MEV and NEV (y-axis) across the 12 LC (x-axis) colour coded as in (b) with expression illustrated on z-score blue to red scale as an interquartile range plot, with dendrogram of MEV/NEV hierarchical clustering shown on left. **(d)** Relationship of select LC to the ABC/GCB specific network. Module and neighbourhood level gene distributions are illustrated in upper panel (as in Figure 2b). Beneath this the median expression of genes across the network is illustrated in a blue (low) to red (high) z-score scale for the indicated LC, lcGCB-DZ, lcGCB-Xit and lcABC. See Supplemental Figures 6 and 7 for high-resolution versions containing all LC.

Figure 4. Lymphoma Communities (LC) and Lymphoma EcoTypes (LE)

The Lymphoma EcoTyper classifier tool was used to separate the validation datasets into EcoTypes. **(a)** The fraction of each Lymphoma EcoType (LE) across the validation datasets (LE_UNC = unassigned cases). **(b)** The significance of EcoType enrichment/depletion within the Lymphoma communities (LC), z-scores on a blue (significant depletion) to red (significant enrichment) scale (-5 to +5). X-axis shows LC and y-axis LE within each dataset (p-value <0.05) and a combined significance (Meta; p-value <0.01, Stouffer's Z). **(c)** The significance of EcoTyper cell-state enrichment/depletion within the Lymphoma communities (LC), z-scores on a blue (significant depletion) to red (significant enrichment) scale (-5 to +5). X-axis shows LC and y-axis cell-states. Z-scores are Stouffer's Z across the 4 validation datasets (p-value <0.01).

Figure 5. 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 4 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 4 datasets. The right charts show the Log_{10} MetaHR (red) and $-Log_{10}$ 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 4 validation datasets for shown LC; p-value from log-rank test.

Figure 6. 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, 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 and cell-of-origin/MHG 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.

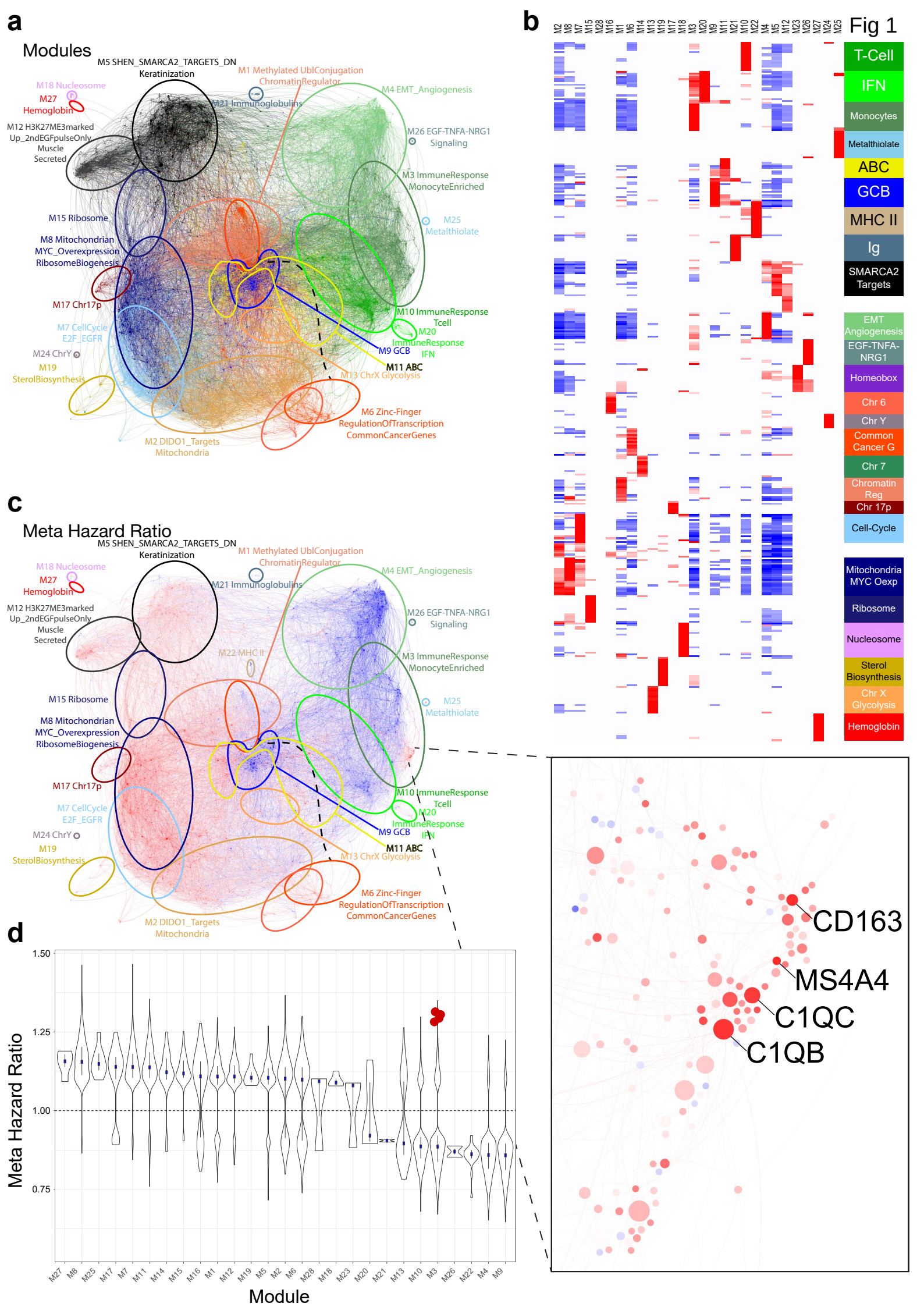
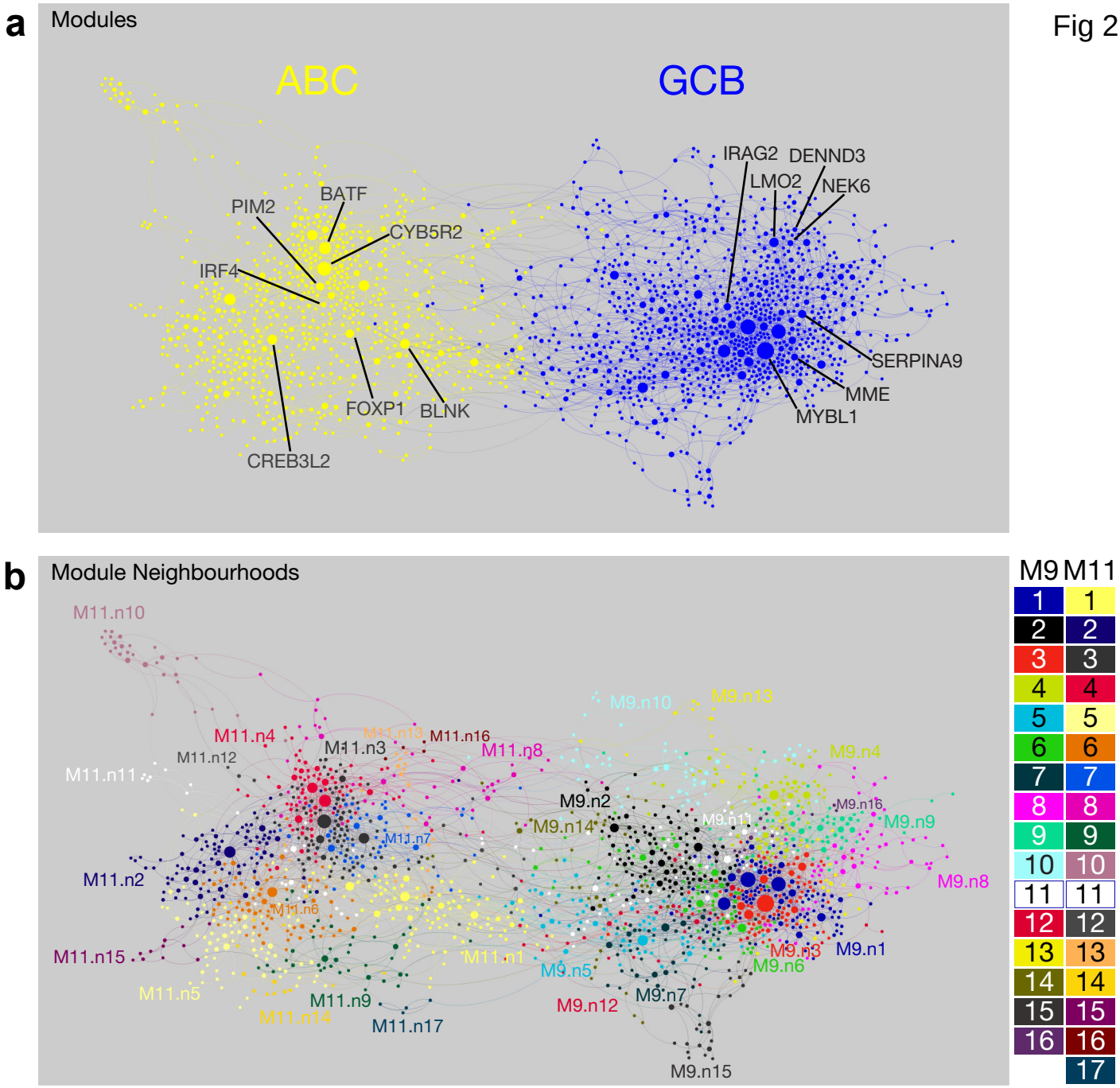
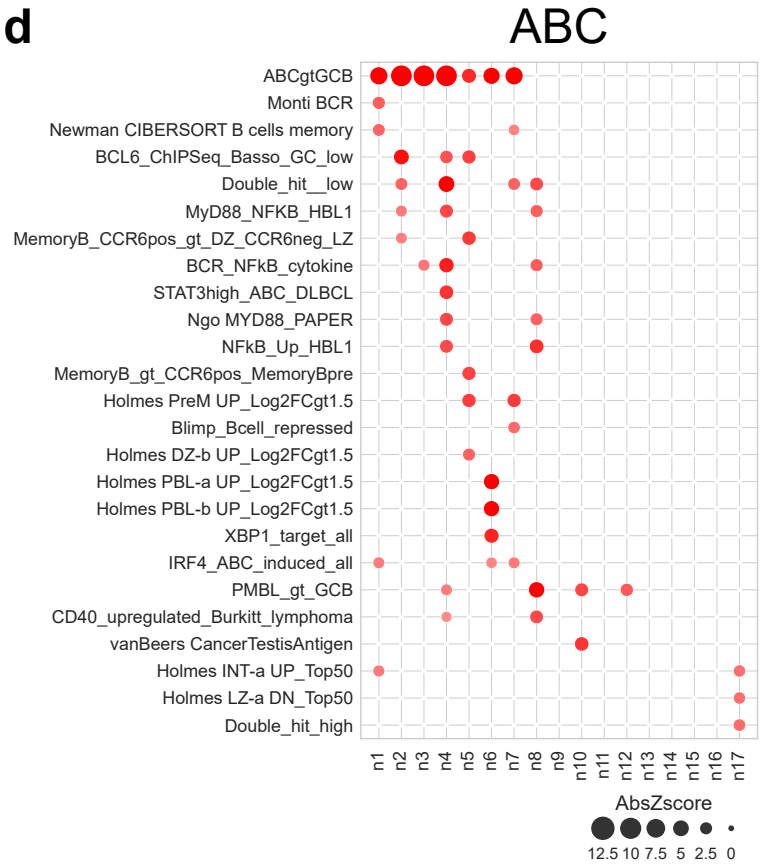

a
Modules
M5 SHEN_SMARCA2_TARGETS_DN
Keratinization
M1 Methylated UblConjugation
ChromatinRegulator
M18 Nucleosome
M27
Hemoglobin
M21 Immunoglobulins
M4 EMT_Angiogenesis
M12 H3K27ME3marked
Up_2ndEGFpulseOnly
Muscle
Secreted
M26 EGF-TNFA-NRG1
Signaling
M3 ImmuneResponse
MonocyteEnriched
M15 Ribosome
M25
Metalthiolate
M8 Mitochondrian
MYC_Overexpression
RibosomeBiogenesis
M17 Chr17p
M10 ImmuneResponse
Tcell
M7 CellCycle
E2F_EGFR
M20
ImmuneResponse
IFN
M24 ChrY
M9 GCB
M19
SterolBiosynthesis
M11 ABC
M13 ChrX Glycolysis
M6 Zinc-Finger
RegulationOfTranscription
CommonCancerGenes
M2 DIDO1_Targets
Mitochondria
b
Fig 1
M2 M8 M7 M15 M28 M16 M1 M6 M14 M13 M19 M17 M18 M3 M20 M9 M11 M21 M10 M22 M4 M5 M12 M23 M26 M27 M24 M25
T-Cell
IFN
Monocytes
Metalthiolate
ABC
GCB
MHC II
Ig
SMARCA2 Targets
EMT Angiogenesis
EGF-TNFA-NRG1
Homeobox
Chr 6
Chr Y
Common Cancer G
Chr 7
Chromatin Reg
Chr 17p
Cell-Cycle
Mitochondria MYC Oexp
Ribosome
Nucleosome
Sterol Biosynthesis
Chr X Glycolysis
Hemoglobin
c
Meta Hazard Ratio
M5 SHEN_SMARCA2_TARGETS_DN
Keratinization
M1 Methylated UblConjugation
ChromatinRegulator
M18 Nucleosome
M27
Hemoglobin
M21 Immunoglobulins
M4 EMT_Angiogenesis
M12 H3K27ME3marked
Up_2ndEGFpulseOnly
Muscle
Secreted
M26 EGF-TNFA-NRG1
Signaling
M22 MHC II
M3 ImmuneResponse
MonocyteEnriched
M15 Ribosome
M25
Metalthiolate
M8 Mitochondrian
MYC_Overexpression
RibosomeBiogenesis
M17 Chr17p
M10 ImmuneResponse
Tcell
M7 CellCycle
E2F_EGFR
M20
ImmuneResponse
IFN
M24 ChrY
M9 GCB
M19
SterolBiosynthesis
M11 ABC
M13 ChrX Glycolysis
M6 Zinc-Finger
RegulationOfTranscription
CommonCancerGenes
M2 DIDO1_Targets
Mitochondria
d
Meta Hazard Ratio
1.50
1.25
1.00
0.75
M27 M8 M25 M17 M7 M11 M14 M15 M16 M1 M12 M19 M5 M2 M6 M28 M18 M23 M20 M21 M13 M10 M3 M26 M22 M4 M9
Module
CD163
MS4A4
C1QC
C1QB

Fig 2


a
Modules
ABC
GCB
PIM2
BATF
CYB5R2
IRF4
FOXP1
BLNK
CREB3L2
IRAG2
DENND3
LMO2
NEK6
SERPINA9
MME
MYBL1
b
Module Neighbourhoods
M11.n10
M11.n4
M11.n12
M11.n3
M11.n13
M11.n16
M11.n8
M11.n11
M11.n7
M11.n2
M11.n6
M11.n15
M11.n5
M11.n14
M11.n9
M11.n17
M11.n1
M9.n10
M9.n13
M9.n4
M9.n2
M9.n14
M9.n11
M9.n16
M9.n9
M9.n8
M9.n5
M9.n12
M9.n7
M9.n3
M9.n6
M9.n1
M9.n15
M9 M11


d
ABC
ABCgtGCB
Monti BCR
Newman CIBERSORT B cells memory
BCL6_ChIPSeq_Basso_GC_low
Double_hit__low
MyD88_NFKB_HBL1
MemoryB_CCR6pos_gt_DZ_CCR6neg_LZ
BCR_NFkB_cytokine
STAT3high_ABC_DLBCL
Ngo MYD88_PAPER
NFkB_Up_HBL1
MemoryB_gt_CCR6pos_MemoryBpre
Holmes PreM UP_Log2FCgt1.5
Blimp_Bcell_repressed
Holmes DZ-b UP_Log2FCgt1.5
Holmes PBL-a UP_Log2FCgt1.5
Holmes PBL-b UP_Log2FCgt1.5
XBP1_target_all
IRF4_ABC_induced_all
PMBL_gt_GCB
CD40_upregulated_Burkitt_lymphoma
vanBeers CancerTestisAntigen
Holmes INT-a UP_Top50
Holmes LZ-a DN_Top50
Double_hit_high
n1 n2 n3 n4 n5 n6 n7 n8 n9 n10 n11 n12 n13 n14 n15 n16 n17
AbsZscore
12.5 10 7.5 5 2.5 0
Zscore
6 4.5 3 1.5 0

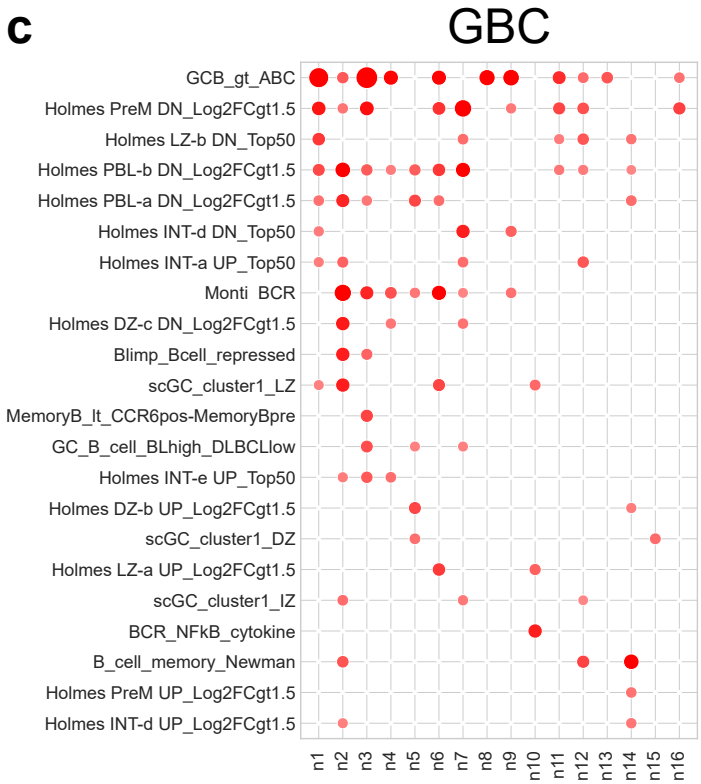

c
GBC
GCB_gt_ABC
Holmes PreM DN_Log2FCgt1.5
Holmes LZ-b DN_Top50
Holmes PBL-b DN_Log2FCgt1.5
Holmes PBL-a DN_Log2FCgt1.5
Holmes INT-d DN_Top50
Holmes INT-a UP_Top50
Monti BCR
Holmes DZ-c DN_Log2FCgt1.5
Blimp_Bcell_repressed
scGC_cluster1_LZ
MemoryB_lt_CCR6pos-MemoryBpre
GC_B_cell_BLhigh_DLBCLlow
Holmes INT-e UP_Top50
Holmes DZ-b UP_Log2FCgt1.5
scGC_cluster1_DZ
Holmes LZ-a UP_Log2FCgt1.5
scGC_cluster1_IZ
BCR_NFkB_cytokine
B_cell_memory_Newman
Holmes PreM UP_Log2FCgt1.5
Holmes INT-d UP_Log2FCgt1.5
n1 n2 n3 n4 n5 n6 n7 n8 n9 n10 n11 n12 n13 n14 n15 n16

Fig 3

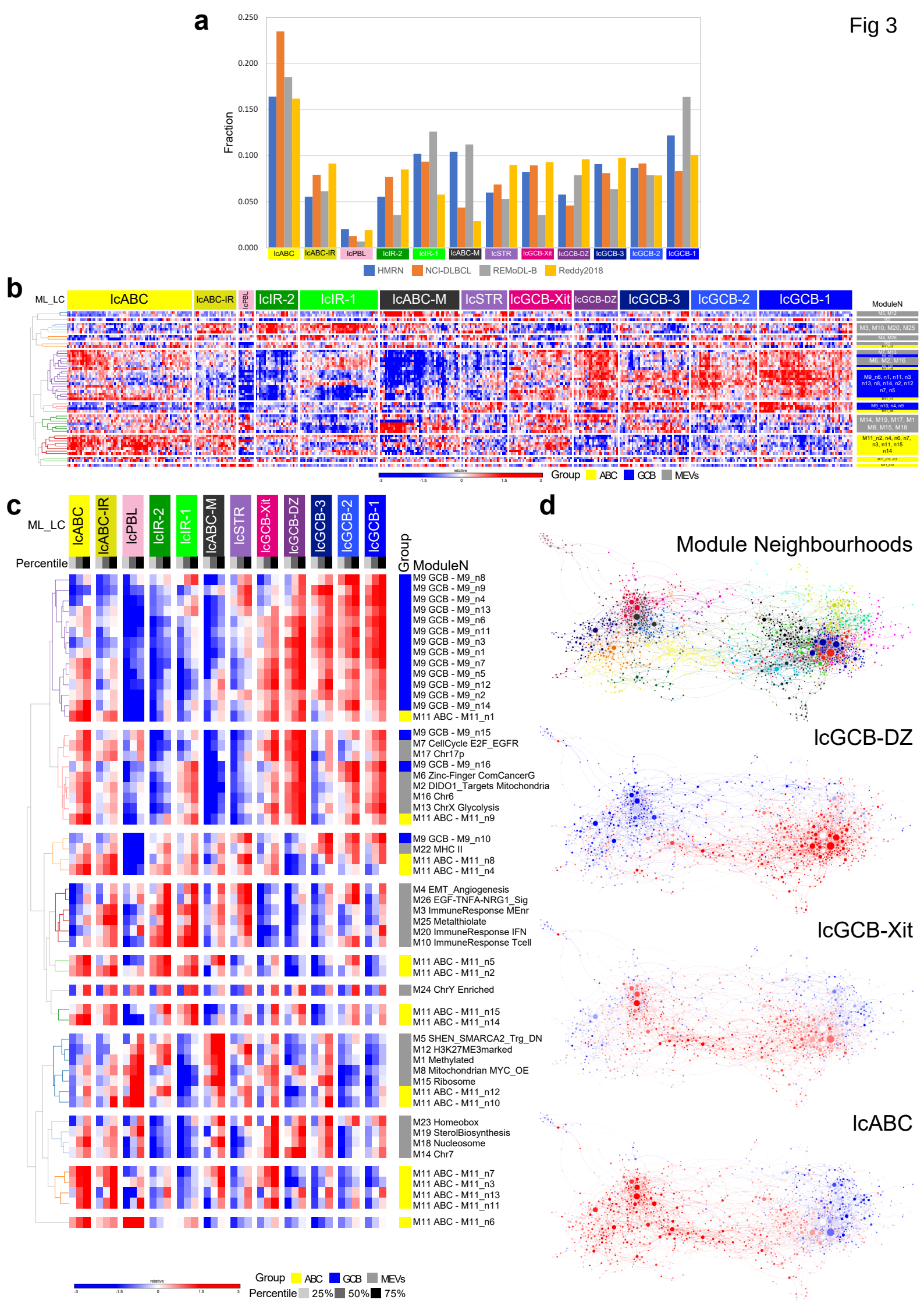
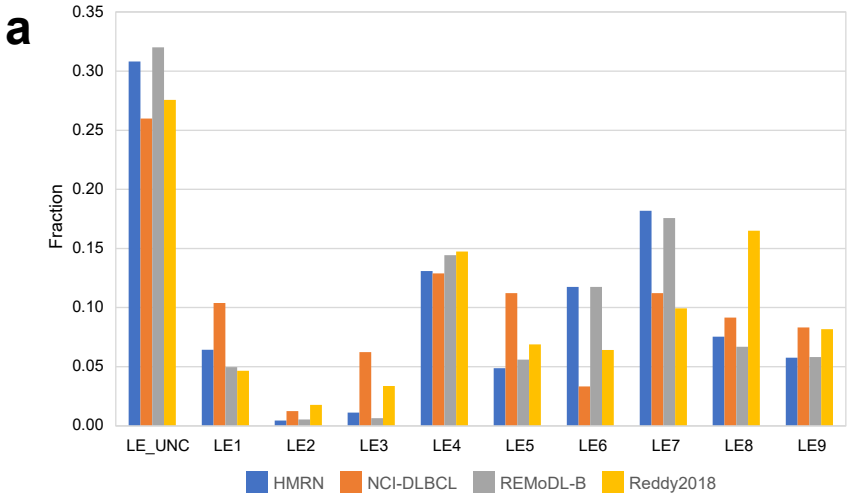
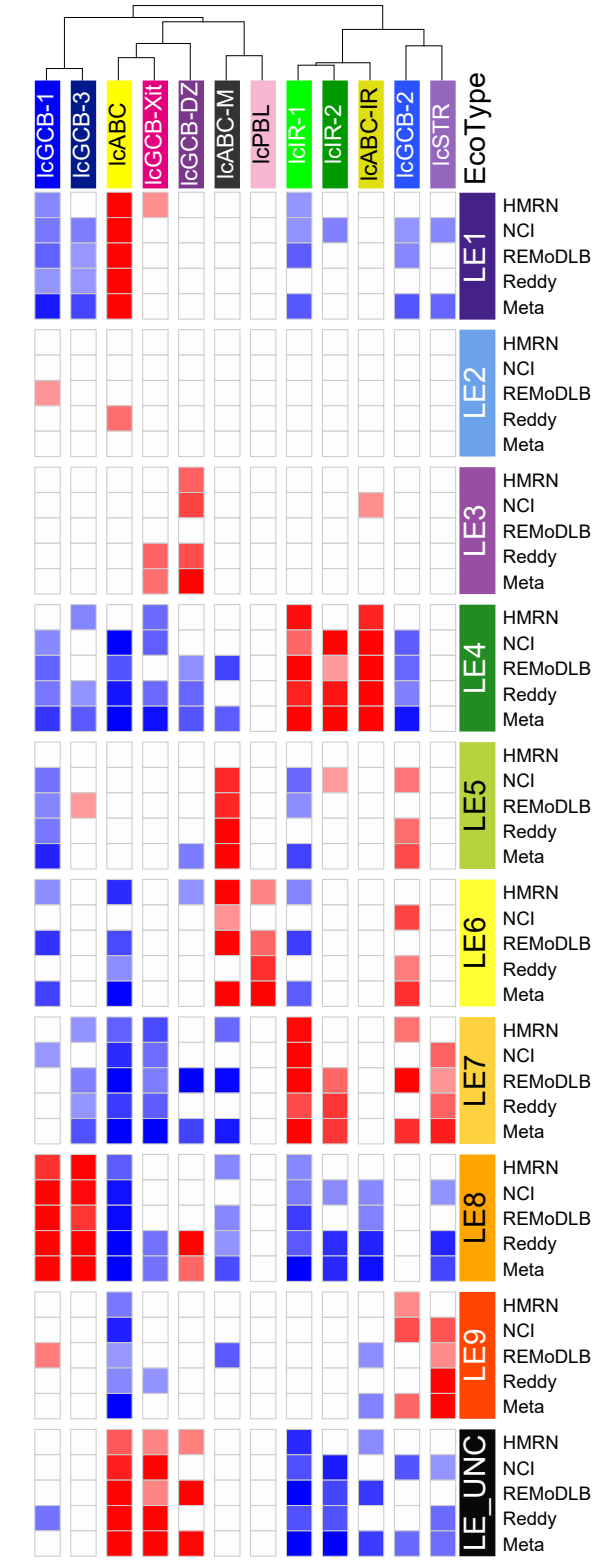

a
Fraction
0.250
0.200
0.150
0.100
0.050
0.000
lcABC
lcABC-IR
lcPBL
lcIR-2
lcIR-1
lcABC-M
lcSTR
lcGCB-Xit
lcGCB-DZ
lcGCB-3
lcGCB-2
lcGCB-1
HMRN
NCI-DLBCL
REMoDL-B
Reddy2018
b
ML_LC
ModuleN
M3, M10, M20, M25
M6, M2, M16
M14, M19, M17, M1 M8, M15, M18
Group
ABC
GCB
MEVs
c
ML_LC
Percentile
Group
ModuleN
M9 GCB - M9_n8
M9 GCB - M9_n9
M9 GCB - M9_n4
M9 GCB - M9_n13
M9 GCB - M9_n6
M9 GCB - M9_n11
M9 GCB - M9_n3
M9 GCB - M9_n1
M9 GCB - M9_n7
M9 GCB - M9_n5
M9 GCB - M9_n12
M9 GCB - M9_n2
M9 GCB - M9_n14
M11 ABC - M11_n1
M9 GCB - M9_n15
M7 CellCycle E2F_EGFR
M17 Chr17p
M9 GCB - M9_n16
M6 Zinc-Finger ComCancerG
M2 DIDO1_Targets Mitochondria
M16 Chr6
M13 ChrX Glycolysis
M11 ABC - M11_n9
M9 GCB - M9_n10
M22 MHC II
M11 ABC - M11_n8
M11 ABC - M11_n4
M4 EMT_Angiogenesis
M26 EGF-TNFA-NRG1_Sig
M3 ImmuneResponse MEnr
M25 Metalthiolate
M20 ImmuneResponse IFN
M10 ImmuneResponse Tcell
M11 ABC - M11_n5
M11 ABC - M11_n2
M24 ChrY Enriched
M11 ABC - M11_n15
M11 ABC - M11_n14
M5 SHEN_SMARCA2_Trg_DN
M12 H3K27ME3marked
M1 Methylated
M8 Mitochondrian MYC_OE
M15 Ribosome
M11 ABC - M11_n12
M11 ABC - M11_n10
M23 Homeobox
M19 SterolBiosynthesis
M18 Nucleosome
M14 Chr7
M11 ABC - M11_n7
M11 ABC - M11_n3
M11 ABC - M11_n13
M11 ABC - M11_n11
M11 ABC - M11_n6
Group
ABC
GCB
MEVs
Percentile
25%
50%
75%
d
Module Neighbourhoods
lcGCB-DZ
lcGCB-Xit
lcABC

Fig 4


a
Fraction
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
LE_UNC
LE1
LE2
LE3
LE4
LE5
LE6
LE7
LE8
LE9
HMRN
NCI-DLBCL
REMoDL-B
Reddy2018


b
IcGCB-1
IcGCB-3
IcABC
IcGCB-Xit
IcGCB-DZ
IcABC-M
IcPBL
IcIR-1
IcIR-2
IcABC-IR
IcGCB-2
IcSTR
EcoType
LE1
LE2
LE3
LE4
LE5
LE6
LE7
LE8
LE9
LE_UNC
HMRN
NCI
REMoDLB
Reddy
Meta

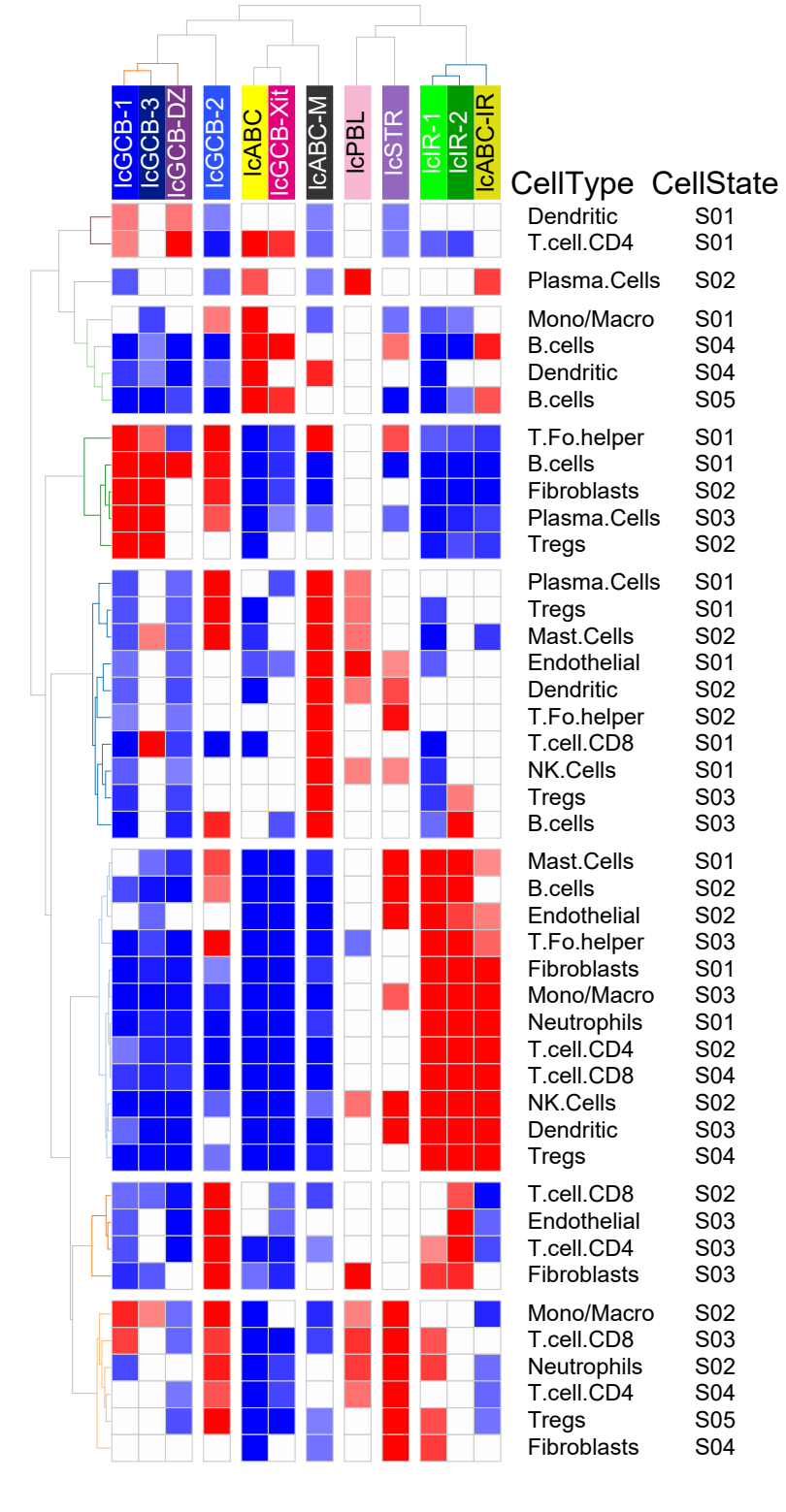

c
IcGCB-1
IcGCB-3
IcGCB-DZ
IcGCB-2
IcABC
IcGCB-Xit
IcABC-M
IcPBL
IcSTR
IcIR-1
IcIR-2
IcABC-IR
CellType
CellState
Dendritic S01
T.cell.CD4 S01
Plasma.Cells S02
Mono/Macro S01
B.cells S04
Dendritic S04
B.cells S05
T.Fo.helper S01
B.cells S01
Fibroblasts S02
Plasma.Cells S03
Tregs S02
Plasma.Cells S01
Tregs S01
Mast.Cells S02
Endothelial S01
Dendritic S02
T.Fo.helper S02
T.cell.CD8 S01
NK.Cells S01
Tregs S03
B.cells S03
Mast.Cells S01
B.cells S02
Endothelial S02
T.Fo.helper S03
Fibroblasts S01
Mono/Macro S03
Neutrophils S01
T.cell.CD4 S02
T.cell.CD8 S04
NK.Cells S02
Dendritic S03
Tregs S04
T.cell.CD8 S02
Endothelial S03
T.cell.CD4 S03
Fibroblasts S03
Mono/Macro S02
T.cell.CD8 S03
Neutrophils S02
T.cell.CD4 S04
Tregs S05
Fibroblasts S04
global
-5
0
5

Fig 5

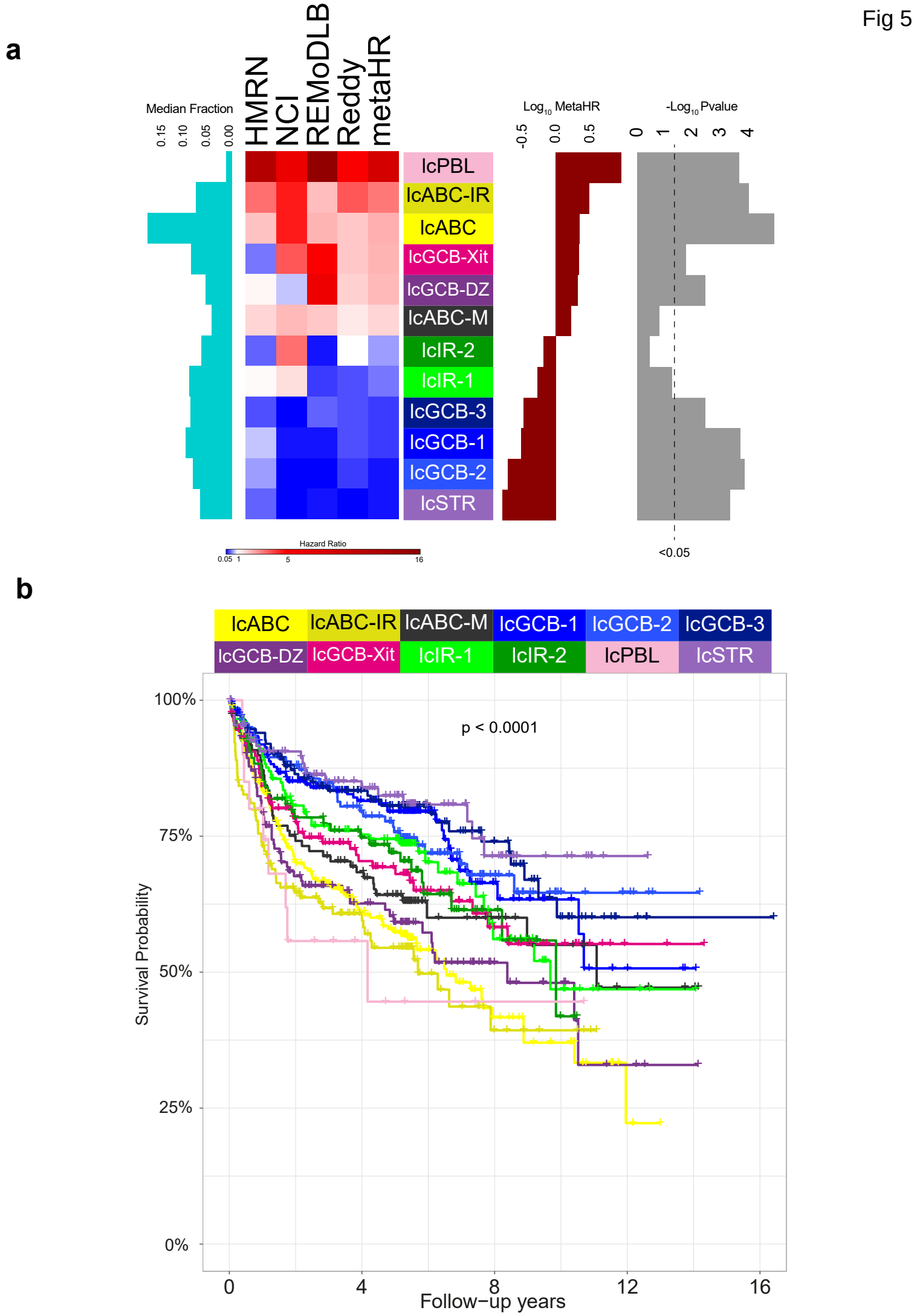
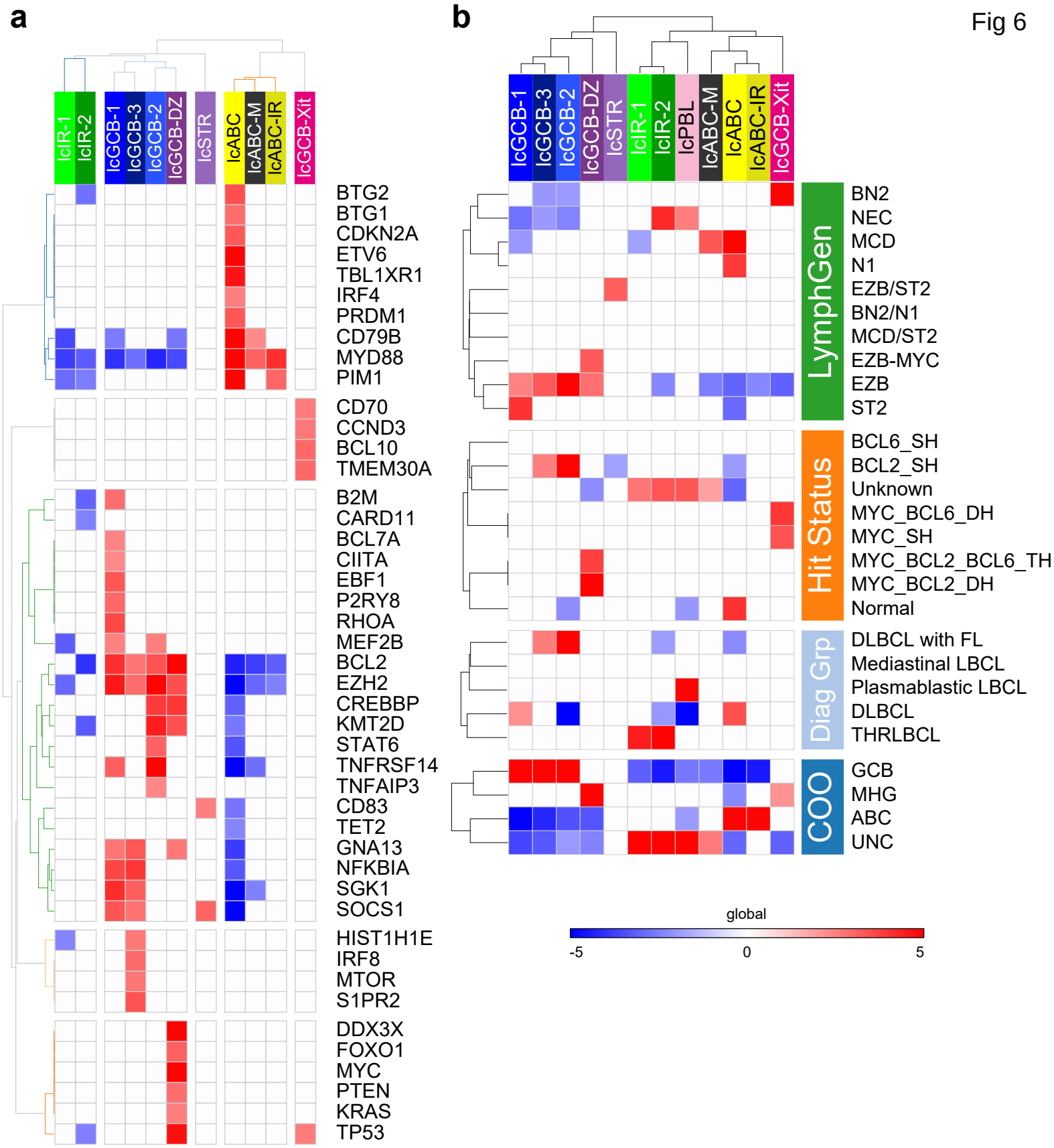

a
Median Fraction
0.15
0.10
0.05
0.00
HMRN
NCI
REMoDLB
Reddy
metaHR
lcPBL
lcABC-IR
lcABC
lcGCB-Xit
lcGCB-DZ
lcABC-M
lcIR-2
lcIR-1
lcGCB-3
lcGCB-1
lcGCB-2
lcSTR
Log10 MetaHR
-0.5
0.0
0.5
-Log10 Pvalue
0
1
2
3
4
Hazard Ratio
0.05
1
5
16
<0.05
b
lcABC
lcABC-IR
lcABC-M
lcGCB-1
lcGCB-2
lcGCB-3
lcGCB-DZ
lcGCB-Xit
lcIR-1
lcIR-2
lcPBL
lcSTR
p < 0.0001
100%
75%
50%
25%
0%
Survival Probability
0
4
8
12
16
Follow−up years

Fig 6


a
b
IcIR-1
IcIR-2
IcGCB-1
IcGCB-3
IcGCB-2
IcGCB-DZ
IcSTR
IcABC
IcABC-M
IcABC-IR
IcGCB-Xit
IcPBL
BTG2
BTG1
CDKN2A
ETV6
TBL1XR1
IRF4
PRDM1
CD79B
MYD88
PIM1
CD70
CCND3
BCL10
TMEM30A
B2M
CARD11
BCL7A
CIITA
EBF1
P2RY8
RHOA
MEF2B
BCL2
EZH2
CREBBP
KMT2D
STAT6
TNFRSF14
TNFAIP3
CD83
TET2
GNA13
NFKBIA
SGK1
SOCS1
HIST1H1E
IRF8
MTOR
S1PR2
DDX3X
FOXO1
MYC
PTEN
KRAS
TP53
LymphGen
BN2
NEC
MCD
N1
EZB/ST2
BN2/N1
MCD/ST2
EZB-MYC
EZB
ST2
Hit Status
BCL6_SH
BCL2_SH
Unknown
MYC_BCL6_DH
MYC_SH
MYC_BCL2_BCL6_TH
MYC_BCL2_DH
Normal
Diag Grp
DLBCL with FL
Mediastinal LBCL
Plasmablastic LBCL
DLBCL
THRLBCL
COO
GCB
MHG
ABC
UNC
global
-5
0
5