

1 Resolving heterogeneity in Diffuse Large B-cell Lymphoma 2 using a comprehensive modular expression map

3

4 Matthew A. Care^{1,2*}, Daniel Painter¹, Sharon Barrans³, Chulin Sha⁴, Peter Johnson⁵, Andy
5 Davies⁵, Ming-Qing Du⁶, Simon Crouch¹, Alex Smith^{1,2}, Eve Roman^{1,2}, Cathy Burton^{3,2}, Gina
6 Doody^{2,7}, David Westhead^{2,8}, Ulf Klein^{2,7}, Daniel J. Hodson⁹, Reuben Tooze^{2,3,7*}

7

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

10 ²National Institute of Health Research Leeds Biomedical Research Centre

11 ³Haematological Malignancy Diagnostic Service, St. James's Institute of Oncology, Leeds LS9
12 7TF, UK;

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

15 ⁵University of Southampton, Southampton SO16 6YD, UK;

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

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

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

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

24 Correspondence to:

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

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

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

30

31 Abstract

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

43

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

48

49 Introduction

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

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

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

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

81

82 Results

83 An integrated gene expression context for DLBCL

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

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

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

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

110 B-cell lineage gene expression as assessed by signature enrichment (e.g. SignatureDB pan-B
111 or Resting Blood B-cell). Thus, the fundamental separation of GCB- and ABC-related
112 expression patterns was rediscovered as a primary feature of gene co-expression patterns in
113 the network structure. The more numerous CCC genes divided across several modules but
114 with significant enrichments in specific modules (M9:B-cell receptor p-value 3.27×10^{-57} ,
115 M3:Host response p-value 1.64×10^{-64} and M13:Oxphos p-value 1.38×10^{-05}).¹² Additional
116 modules separated features related to MYC function (M8), cell cycle (M7), glycolysis (M13),
117 and sterol biosynthesis (M19). Host response/microenvironment was divided between
118 modules related to stromal/angiogenesis (M4), T/NK-cells (M10), monocytes/macrophages
119 (M3) and IFN responses (M20). Highly co-ordinated gene batteries were resolved for
120 nucleosome components (M18), homeobox (M23), immediate early (M26) and
121 metallothionein genes (M25). Several modules related to structural chromosomal regions
122 including chr6 (M16), chr7 (M14), chr17p (M17), chrX (M13) and chrY (M24).

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

128

129 **Meta-hazard ratio analysis confirms associations between biology and outcome**

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

138

139 **Resolving detailed differentiation states in GCB and ABC modules**

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

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

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

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

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

178

179 **Network gene expression patterns show consistent association with mutation state**

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

190 Metagene expression patterns were significantly associated with mutation state (Figure 3
191 and Supplemental Figure 4). At module level the primary separation was into two clusters
192 corresponding to M9_GCB and M11_ABC mutation correlations. M9_GCB correlated
193 significantly with a wide range of mutations consistent with known associations both of
194 COO-defined GCB-DLBCL and follicular lymphoma (FL),^{3,41} and anticorrelated with the
195 primary mutations linked to COO-defined ABC-DLBCL. M11_ABC exhibited the reciprocal
196 anticorrelation with mutations linked to COO-defined GCB-DLBCL and positive correlation
197 with mutations characteristic of COO-defined ABC-DLBCL. Modules which shared mutation
198 correlations with M9_GCB included: M13_ChroGlycolysis, M23_Homeobox, M28_ITGB8,
199 M6_ZincFinger, M6_Chrom6; while modules with shared mutation correlations with M11_ABC
200 included: M8_MitochondrionMYC, M18_Nucleosome, M7_CellCycle and M17_Chrom17p. An
201 additional cluster of modules focused on shared correlations with *CDKN2A* mutation
202 including M14_Chrom7 and M15_Ribosome.

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

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

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

233 While this analysis is limited to the sets of genes tested in targeted mutation panels, the
234 most important known mutational features of DLBCL are included. The results identify
235 distinct and significant associations of network derived gene co-expression patterns with
236 driver mutations and argue that the granularity provided by the network at module and
237 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

270 used to reclassify all cases in the HMRN (Figure 4b) and other datasets (see Supplemental
271 Methods for details). A similar distribution of cases falling into these communities was
272 recovered across a range of datasets (Figure 4c and Supplemental Figure 5 & 6). These
273 communities were named according to patterns of module expression or associated
274 pathological/molecular features that were identified in downstream analyses.
275 Amongst the 12 LCs (Figure 4b, Supplemental Figure 6) two communities identified cases
276 dominated by ABC module gene expression which were distinguished by differences in
277 immune response-related features as immune response poor (${}_{ic}ABC$) or immune rich (${}_{ic}ABC-$
278 **IR**). Cases with plasmablastic features (${}_{ic}PBL$) were distinguished by expression of XBP1
279 targets and cancer testis antigens. A distinct subset of cases with ABC-related features
280 shared the expression of cancer testis antigens (${}_{ic}CT$). Amongst cases with weak B-cell
281 patterns, two communities were dominated by immune response modules (${}_{ic}IR$ and ${}_{ic}IR-$
282 **TET2** - distinguished by *TET2* mutation enrichment, see below). Six communities were
283 characterised by expression of multiple GCB-related neighbourhoods. These included a
284 community with GCB features and distinct stromal/EMT-related gene expression (${}_{ic}GCB-$
285 **EMT**) and a community with mixed GCB and ABC expression features along with cell cycle,
286 MYC, and sterol biosynthesis modules, suggestive of transition from GCB to an ABC/post-GC
287 state (${}_{ic}GCB-Xit$, Supplemental Figure 6). Similar proliferation and growth-related module
288 expression were combined with polarised high GC DZ and low GC LZ neighbourhood
289 expression, along with low CD40/NF κ B and MHC-II genes in a GCB DZ-like community
290 (${}_{ic}GCB-DZ$). The remaining three GCB communities were distinguished by different patterns
291 of GCB neighbourhood expression, wider network features and other associations in
292 downstream analyses (${}_{ic}GCB$ - most canonical GCB-like, ${}_{ic}GCB-FL$ - underlying FL diagnosis,
293 ${}_{ic}GCB-SOCS1$ – selective SOCS1 mutation association).

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

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

304

305 **DLBCL communities link to mutation patterns**

306 An important test of the DLBCL LCs was whether these also showed significant association
307 with mutation state. To address this we analysed the enrichment of mutations in DLBCL
308 communities by integrating enrichment p-values derived across the individual HMRN,
309 REMoDL-B and Reddy datasets.^{5,36,37} At a p-value threshold <0.01 (in ≥ 2 datasets) the
310 communities showed significant and distinct associations with mutation patterns. These
311 separated in concordance with the expression states between mutational features linked to
312 ABC, GCB, and host/immune response characteristics (Figure 5a). Thus, ${}_{1c}ABC$ and ${}_{1c}ABC-IR$
313 shared association with *MYD88*, *CD79B*, *PIM1*, *PRDM1*, *CDKN2A* and differed in association
314 of with *ETV6*, *TBL1XR1* and *IRF4* mutations. The ABC-related ${}_{1c}CT$ community was selectively
315 enriched for *MYD88* and *PIM1* mutations and was additionally enriched for *BCL10*, *TP53*,
316 *MEF2B* mutations. These patterns separated most distinctly from ${}_{1c}GCB$ and ${}_{1c}GCB-DZ$. These
317 shared enrichment for *BCL2*, *EZH2*, *DDX3X*, *IRF8*, *PTEN*, *TNFRSF14*, *GNA13* mutation. GCB-DZ
318 was additionally enriched for *CREBBP*, *KMT2D*, *BTK*, *POU2F2*, *S1PR2*, *FOXO1*, *MYC*, *MEF2B*
319 mutation, while ${}_{1c}GCB$ was enriched for *CARD11*, *B2M*, *EBF1*, *MSH6*, *RHOA*, *TET2*, *SGK1*
320 mutation. ${}_{1c}GCB-FL$ shared association with *CREBBP*, *KMT2D*, *TNFRSF14* but was
321 distinguished by association with *STAT6* mutations including significant enrichment of the
322 *STAT6* Y419 hotspot mutation. ${}_{1c}GCB-EMT$ shared enrichment of *SGK1* and *TNFRSF14* with
323 ${}_{1c}GCB$ but otherwise lacked distinct associations. ${}_{1c}GCB-SOCS1$ and ${}_{1c}GCB-Xit$ differed from
324 other GCB-related expression groups. ${}_{1c}GCB-SOCS1$ associated with *SOCS1*, *NFKBIA*, *SGK1*,
325 *BTG1*, *GNA13*, *CD83*, *NFKBIE*, *ZFP36L1* and *STAT6* but not the *STAT6* Y419 mutation hotspot,
326 while ${}_{1c}GCB-Xit$ showed a further distinctive pattern with selective enrichment for *CD70*,
327 *CCND3*, *BTG2* mutations. While ${}_{1c}IR$ only displayed anti-correlations with mutation state
328 likely reflecting the diluting effect of host response components, ${}_{1c}IR-TET2$ cases were
329 selectively associated with *TET2* mutations.

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

334 largely concordant,⁴⁶ and recapitulate the features of the C1-C5 Harvard classification.² The
335 LymphGen MCD class was significantly associated with $_{1c}ABC$ and $_{1c}ABC-IR$, while $_{1c}CT$ was
336 associated with cases with ambivalent MCD/ST2 calls. LymphGen N1 showed enrichment in
337 $_{1c}ABC$ while BN2 was significantly and selectively associated with $_{1c}GCB-Xit$. EZB was most
338 significantly enriched amongst $_{1c}GCB$, $_{1c}GCB-DZ$ and $GCB-FL$, while EZB-MYC+ was selectively
339 enriched amongst $_{1c}GCB-DZ$. ST2 cases mapped selective on to $_{1c}GCB-SOCS1$. Similar patterns
340 of association were evident when considering the associations of LCs with the independent
341 but related HMRN mutational classification. We conclude that the mutation-based
342 subdivision of DLBCL in LymphGen and related classifications overlaps significantly with LCs
343 derived independently from gene co-expression patterns.

344

345 **Lymphoma communities underline distinctions between double-hit lymphomas**

346 In DLBCL the occurrence of double- or triple-hit with rearrangement of *MYC* and *BCL2*
347 (MYC_BCL2_DH) or *MYC* and *BCL6* (MYC_BCL6_DH), or *MYC*, *BCL2* and *BCL6*
348 ($MYC_BCL2_BCL6_TH$) identifies high-risk disease.⁸⁻¹¹ The relationships of these
349 rearrangement-based categories to gene expression and mutation features is relatively
350 clear-cut for MYC_BCL2_DH , but remains less well-defined for MYC_BCL6_DH .⁴⁷⁻⁴⁹ We
351 therefore tested the association of rearrangement status and LCs in the HMRN cohort.
352 $_{1c}GCB-DZ$ was highly enriched for MYC_BCL2_DH and $MYC_BCL2_BCL6_TH$ cases (Figure 5b).
353 In contrast, a more modest enrichment of MYC_BCL2_DH cases was evident in $_{1c}GCB$, while
354 $_{1c}GCB-FL$ was selectively enriched for $BCL2_SH$. $_{1c}GCB-Xit$ was selectively enriched for
355 MYC_BCL6_DH and $BCL6_SH$ status but was not associated with any combination of *BCL2*
356 rearrangement, while $_{1c}ABC$ was enriched for cases with MYC_SH .

357 We further assessed whether the LCs were significantly associated with morphological
358 diagnosis, COO class and molecular high grade (MHG) or double-hit signature (DHITsig)
359 status.^{14,15} Two communities, $_{1c}GCB-FL$ and to a lesser extent $_{1c}GCB-DZ$ were linked to
360 concurrent or previous diagnosis of underlying FL supporting a primary separation of
361 disease biology in transformed FL cases (Figure 5b). $_{1c}PBL$ was selectively enriched for
362 plasmablastic lymphoma morphological diagnosis, $_{1c}IR$ and $_{1c}IR-TET2$ for T-cell histiocyte rich
363 large B-cell lymphoma and $_{1c}GCB-SOCS1$ for primary mediastinal large B-cell lymphoma. At
364 gene expression level the LC communities recovered appropriate enrichments of *GCB*, *ABC*
365 and unclassified cases. $_{1c}GCB-DZ$ was selectively enriched for cases designated as MHG and

366 cases that were DHITsig positive. Notably DHITsig differed from MHG in also showing
367 significant enrichment amongst $_{1c}$ PBL. We conclude that the DLBCL LCs separate
368 meaningfully in relation to underlying rearrangement status, morphological diagnosis,
369 previous expression-based classifiers of high-risk disease and the presence of underlying FL.

370

371 **Lymphoma communities have prognostic significance in R-CHOP treated cases**

372 Given the fact that LCs segregate cases with common pathological features, we assessed
373 whether LCs showed significant and reproducible survival differences. We addressed this
374 across 6 datasets encompassing sufficient cases treated with R-CHOP chemo-
375 immunotherapy.^{3,5,13,29,36,37} Across multiple datasets the DLBCL community structure
376 separated risk (Figure 6) albeit with variation evident between cohorts most notably for
377 $_{1c}$ GCB-DZ and $_{1c}$ CT. Most adverse risk, in terms of meta-HR across the 6 datasets, was
378 observed for assignment of cases to $_{1c}$ PBL, followed by $_{1c}$ ABC-IR, $_{1c}$ GCB-Xit, $_{1c}$ ABC and $_{1c}$ IR-
379 TET2. Intermediate and variable risk was observed for $_{1c}$ CT and $_{1c}$ GCB-DZ. At the other end of
380 the spectrum particularly good risk was associated with $_{1c}$ GCB-FL, $_{1c}$ GCB-SOCS1, $_{1c}$ GCB and
381 $_{1c}$ GCB-EMT. We conclude that the refined break down of DLBCLs in the 12-fold LC structure
382 has potential utility to refine the separation of risk groups, while remaining aligned to the
383 concept of the COO differentiation state.

384

385 **Lymphoma communities show distinct responses to RB-CHOP**

386 The REMoDL-B trial recently reported on extended follow-up identifying a benefit for
387 proteasome inhibitor bortezomib with R-CHOP in cases classified as ABC-DLBCL.⁵⁰ We were
388 therefore interested to test whether LCs could provide further insight into responses to RB-
389 CHOP. For overall (OS) and progression-free (PFS) survival the LCs were significantly
390 separated overall in the R-CHOP arm of the trial. In the RB-CHOP arm the significance of
391 separation for the LCs declined for both OS (R-CHOP $p=0.0013$, RB-CHOP $p=0.11$) and PFS (R-
392 CHOP $p=0.002$, RB-CHOP $p=0.02$) (Figure 7). While there are inherent limitations in *post hoc*
393 analysis and the impact of 12-fold classification on case numbers, we were interested to
394 further compare the response for R-CHOP and RB-CHOP arms of the trial for each LC
395 (Supplemental Figure 9 and 10). While some indication of improved responses for the RB-
396 CHOP arm was evident, for example for $_{1c}$ PBL case that achieved a 90%+ OS in the RB-CHOP

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

399

400 **Bortezomib response in ABC cases with variant expression features**

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

413

414 Discussion

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

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

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

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

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

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

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

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

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

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

504

505 Methods

506 (See also Supplemental Methods)

507 Expression datasets

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

517

518 Network generation

519 Discovery datasets (GSE10846 split into CHOP/RCHOP-treated) were processed using
520 PGCNA2 (<https://github.com/medmaca/PGCNA/tree/master/PGCNA2>), retaining the top
521 70% most variant genes present in $\geq 50\%$ of the datasets, carrying out 1000 Leidenalg
522 clusterings and selecting the best using Scaled cluster enrichment scores. The resulting

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

527

528 Network analysis

529 Clustering samples

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

539

540 Neighbourhoods

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

546

547 Module Expression Values and Lymphoma communities

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

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

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

555

556 **Lymphoma Community machine learning tool**

557 Clustering results for HMRN were merged by selecting significant overlaps (p -value <
558 0.0001) that form a community in either CC/PGCNA containing > 5 samples (Supplemental
559 Figure 1 & Supplemental Methods). This generated a high confidence set of LCs ($n=298$)
560 that formed a training dataset (Figure 4b top) that was used to train a machine learning
561 (ML) tool to recover the LC in other datasets. The training dataset was split using the
562 python Scikit-learn `test_train_split` function, stratifying on the LC class label, to give class-
563 balanced randomised training ($n=238$) and validation ($n=60$) datasets. The validation
564 dataset was set aside to test the final selected model. The training data was used to carry
565 out stratified 5-fold cross-validation across 7 different machine learning methods. In total
566 4,802 parameters were tested across the ML tools, scoring with the Matthews correlation
567 coefficient (MCC). The best 3 models, using their optimal parameters, were combined using
568 soft voting to give a model with mean MCC of 0.89 across the 5-folds. This model was then
569 tested on the unseen validation data ($n=60$) with MCC of 0.87. This final model, termed
570 ML_LC, was retrained using all 298 training samples and was used for all subsequent
571 classifications, including the HMRN dataset (Figure 4b bottom).

572

573 **Survival analysis**

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

579

580 **Mutation analysis**

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

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

589

590 [Data Availability Statement](#)

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

594

595 [Code Availability Statement](#)

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

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

598 All other code is available on request.

599

600 [Acknowledgements](#)

601 This work was supported by Cancer Research UK program grant (C7845/A17723 and
602 C7845/A29212) (M.C, G.D., D.W, and R.T). HMRN is supported by Cancer Research UK
603 program grant A29685 (D.P., S.C., A.S., E.R.). D.J.H. was supported by a fellowship from
604 Cancer Research UK (CRUK) (RCCFEL\100072) and received core funding from Wellcome
605 (203151/Z/16/Z) to the Wellcome-MRC Cambridge Stem Cell Institute and from the CRUK
606 Cambridge Centre (A25117). D.J.H is supported by the National Institute for Health and Care
607 Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014). R.T., G.D., E.R.,
608 A.S., D.P., D.W. are supported by the National Institute for Health and Care Research Leeds
609 Biomedical Research Centre. The views expressed are those of the authors and not
610 necessarily those of the NIHR or the Department of Health and Social Care. For the purpose
611 of Open Access, the authors have applied a CC BY public copyright licence to any Author
612 Accepted Manuscript version arising from this submission.

613

614 References

- 615 1. Alizadeh, A.A., *et al.* Distinct types of diffuse large B-cell lymphoma identified by
616 gene expression profiling. *Nature* **403**, 503-511 (2000).
- 617 2. Chapuy, B., *et al.* Molecular subtypes of diffuse large B cell lymphoma are associated
618 with distinct pathogenic mechanisms and outcomes. *Nat Med* **24**, 679-690 (2018).
- 619 3. Schmitz, R., *et al.* Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N*
620 *Engl J Med* **378**, 1396-1407 (2018).
- 621 4. Wright, G.W., *et al.* A Probabilistic Classification Tool for Genetic Subtypes of Diffuse
622 Large B Cell Lymphoma with Therapeutic Implications. *Cancer Cell* **37**, 551-568 e514
623 (2020).
- 624 5. Lacy, S.E., *et al.* Targeted sequencing in DLBCL, molecular subtypes, and outcomes: a
625 Haematological Malignancy Research Network report. *Blood* **135**, 1759-1771 (2020).
- 626 6. Pedrosa, L., *et al.* Proposal and validation of a method to classify genetic subtypes of
627 diffuse large B cell lymphoma. *Sci Rep* **11**, 1886 (2021).
- 628 7. Bolen, C.R., *et al.* Prognostic impact of somatic mutations in diffuse large B-cell
629 lymphoma and relationship to cell-of-origin: data from the phase III GOYA study.
630 *Haematologica* **105**, 2298-2307 (2020).
- 631 8. Aukema, S.M., *et al.* Biological characterization of adult MYC-translocation-positive
632 mature B-cell lymphomas other than molecular Burkitt lymphoma. *Haematologica*
633 **99**, 726-735 (2014).
- 634 9. Aukema, S.M., *et al.* Double-hit B-cell lymphomas. *Blood* **117**, 2319-2331 (2011).
- 635 10. Barrans, S., *et al.* Rearrangement of MYC is associated with poor prognosis in
636 patients with diffuse large B-cell lymphoma treated in the era of rituximab. *J Clin*
637 *Oncol* **28**, 3360-3365 (2010).
- 638 11. Li, S., *et al.* B-cell lymphomas with MYC/8q24 rearrangements and
639 IGH@BCL2/t(14;18)(q32;q21): an aggressive disease with heterogeneous histology,
640 germinal center B-cell immunophenotype and poor outcome. *Mod Pathol* **25**, 145-
641 156 (2012).
- 642 12. Monti, S., *et al.* Molecular profiling of diffuse large B-cell lymphoma identifies robust
643 subtypes including one characterized by host inflammatory response. *Blood* **105**,
644 1851-1861 (2005).
- 645 13. Lenz, G., *et al.* Stromal gene signatures in large-B-cell lymphomas. *The New England*
646 *journal of medicine* **359**, 2313-2323 (2008).
- 647 14. Sha, C., *et al.* Molecular High-Grade B-Cell Lymphoma: Defining a Poor-Risk Group
648 That Requires Different Approaches to Therapy. *J Clin Oncol* **37**, 202-212 (2019).
- 649 15. Ennishi, D., *et al.* Double-Hit Gene Expression Signature Defines a Distinct Subgroup
650 of Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma. *J Clin Oncol* **37**, 190-
651 201 (2019).
- 652 16. Holmes, A.B., *et al.* Single-cell analysis of germinal-center B cells informs on
653 lymphoma cell of origin and outcome. *J Exp Med* **217**(2020).
- 654 17. King, H.W., *et al.* Single-cell analysis of human B cell maturation predicts how
655 antibody class switching shapes selection dynamics. *Sci Immunol* **6**(2021).
- 656 18. Milpied, P., *et al.* Human germinal center transcriptional programs are de-
657 synchronized in B cell lymphoma. *Nat Immunol* **19**, 1013-1024 (2018).
- 658 19. Attaf, N., Baaklini, S., Binet, L. & Milpied, P. Heterogeneity of germinal center B cells:
659 New insights from single-cell studies. *Eur J Immunol* **51**, 2555-2567 (2021).

- 660 20. Duan, L., *et al.* Follicular dendritic cells restrict interleukin-4 availability in germinal
661 centers and foster memory B cell generation. *Immunity* **54**, 2256-2272 e2256 (2021).
- 662 21. Kennedy, D.E., *et al.* Novel specialized cell state and spatial compartments within the
663 germinal center. *Nat Immunol* **21**, 660-670 (2020).
- 664 22. Steen, C.B., *et al.* The landscape of tumor cell states and ecosystems in diffuse large
665 B cell lymphoma. *Cancer Cell* **39**, 1422-1437 e1410 (2021).
- 666 23. Kotlov, N., *et al.* Clinical and Biological Subtypes of B-cell Lymphoma Revealed by
667 Microenvironmental Signatures. *Cancer discovery* **11**, 1468-1489 (2021).
- 668 24. Care, M.A., Westhead, D.R. & Tooze, R.M. Parsimonious Gene Correlation Network
669 Analysis (PGCNA): a tool to define modular gene co-expression for refined molecular
670 stratification in cancer. *NPJ Syst Biol Appl* **5**, 13 (2019).
- 671 25. Stephenson, S., *et al.* Growth Factor-like Gene Regulation Is Separable from Survival
672 and Maturation in Antibody-Secreting Cells. *J Immunol* **202**, 1287-1300 (2019).
- 673 26. Compagno, M., *et al.* Mutations of multiple genes cause deregulation of NF-kappaB
674 in diffuse large B-cell lymphoma. *Nature* **459**, 717-721 (2009).
- 675 27. Williams, P.M., *et al.* A novel method of amplification of FFPET-derived RNA enables
676 accurate disease classification with microarrays. *J Mol Diagn* **12**, 680-686 (2010).
- 677 28. Salaverria, I., *et al.* Translocations activating IRF4 identify a subtype of germinal
678 center-derived B-cell lymphoma affecting predominantly children and young adults.
679 *Blood* **118**, 139-147 (2011).
- 680 29. Frei, E., *et al.* Addition of rituximab to chemotherapy overcomes the negative
681 prognostic impact of cyclin E expression in diffuse large B-cell lymphoma. *J Clin*
682 *Pathol* **66**, 956-961 (2013).
- 683 30. Care, M.A., *et al.* SPIB and BATF provide alternate determinants of IRF4 occupancy in
684 diffuse large B-cell lymphoma linked to disease heterogeneity. *Nucleic Acids Res* **42**,
685 7591-7610 (2014).
- 686 31. Monti, S., *et al.* Integrative analysis reveals an outcome-associated and targetable
687 pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer*
688 *Cell* **22**, 359-372 (2012).
- 689 32. Hummel, M., *et al.* A biologic definition of Burkitt's lymphoma from transcriptional
690 and genomic profiling. *The New England journal of medicine* **354**, 2419-2430 (2006).
- 691 33. Dave, S.S., *et al.* Molecular diagnosis of Burkitt's lymphoma. *The New England*
692 *journal of medicine* **354**, 2431-2442 (2006).
- 693 34. Scott, D.W., *et al.* Determining cell-of-origin subtypes of diffuse large B-cell
694 lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood*
695 **123**, 1214-1217 (2014).
- 696 35. Dubois, S., *et al.* Biological and Clinical Relevance of Associated Genomic Alterations
697 in MYD88 L265P and non-L265P-Mutated Diffuse Large B-Cell Lymphoma: Analysis of
698 361 Cases. *Clin Cancer Res* **23**, 2232-2244 (2017).
- 699 36. Reddy, A., *et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma.
700 *Cell* **171**, 481-494 e415 (2017).
- 701 37. Davies, A., *et al.* Gene-expression profiling of bortezomib added to standard
702 chemoimmunotherapy for diffuse large B-cell lymphoma (REMoDL-B): an open-label,
703 randomised, phase 3 trial. *Lancet Oncol* **20**, 649-662 (2019).
- 704 38. Wright, G., *et al.* A gene expression-based method to diagnose clinically distinct
705 subgroups of diffuse large B cell lymphoma. *Proc Natl Acad Sci U S A* **100**, 9991-9996
706 (2003).

- 707 39. Basso, K. & Dalla-Favera, R. Germinal centres and B cell lymphomagenesis. *Nat Rev*
708 *Immunol* **15**, 172-184 (2015).
- 709 40. Shaffer, A.L., 3rd, Young, R.M. & Staudt, L.M. Pathogenesis of human B cell
710 lymphomas. *Annual review of immunology* **30**, 565-610 (2012).
- 711 41. Crouch, S., *et al.* Molecular subclusters of follicular lymphoma: a report from the
712 United Kingdom's Haematological Malignancy Research Network. *Blood Adv* **6**, 5716-
713 5731 (2022).
- 714 42. Ennishi, D., *et al.* TMEM30A loss-of-function mutations drive lymphomagenesis and
715 confer therapeutically exploitable vulnerability in B-cell lymphoma. *Nat Med* **26**, 577-
716 588 (2020).
- 717 43. Wilkerson, M.D. & Hayes, D.N. ConsensusClusterPlus: a class discovery tool with
718 confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573 (2010).
- 719 44. Care, M.A., *et al.* A Microarray Platform-Independent Classification Tool for Cell of
720 Origin Class Allows Comparative Analysis of Gene Expression in Diffuse Large B-cell
721 Lymphoma. *PLoS ONE* **8**, e55895 (2013).
- 722 45. Swift, S., *et al.* Consensus clustering and functional interpretation of gene-expression
723 data. *Genome biology* **5**, R94 (2004).
- 724 46. Runge, H.F.P., *et al.* Application of the LymphGen classification tool to 928 clinically
725 and genetically-characterised cases of diffuse large B cell lymphoma (DLBCL). *Br J*
726 *Haematol* (2020).
- 727 47. Rosenwald, A., *et al.* Prognostic Significance of MYC Rearrangement and
728 Translocation Partner in Diffuse Large B-Cell Lymphoma: A Study by the Lunenburg
729 Lymphoma Biomarker Consortium. *J Clin Oncol* **37**, 3359-3368 (2019).
- 730 48. Cucco, F., *et al.* Distinct genetic changes reveal evolutionary history and
731 heterogeneous molecular grade of DLBCL with MYC/BCL2 double-hit. *Leukemia* **34**,
732 1329-1341 (2020).
- 733 49. Scott, D.W., *et al.* High-grade B-cell lymphoma with MYC and BCL2 and/or BCL6
734 rearrangements with diffuse large B-cell lymphoma morphology. *Blood* **131**, 2060-
735 2064 (2018).
- 736 50. Davies, A.J., *et al.* Differential Efficacy From the Addition of Bortezomib to R-CHOP in
737 Diffuse Large B-Cell Lymphoma According to the Molecular Subgroup in the
738 REMoDL-B Study With a 5-Year Follow-Up. *J Clin Oncol*, JCO2300033 (2023).
- 739 51. Quivoron, C., *et al.* TET2 inactivation results in pleiotropic hematopoietic
740 abnormalities in mouse and is a recurrent event during human lymphomagenesis.
741 *Cancer Cell* **20**, 25-38 (2011).
- 742 52. Dominguez, P.M., *et al.* TET2 Deficiency Causes Germinal Center Hyperplasia, Impairs
743 Plasma Cell Differentiation, and Promotes B-cell Lymphomagenesis. *Cancer discovery*
744 **8**, 1632-1653 (2018).
- 745 53. Mouly, E., *et al.* B-cell tumor development in Tet2-deficient mice. *Blood Adv* **2**, 703-
746 714 (2018).
- 747 54. Edwards, J.P., Thornton, A.M. & Shevach, E.M. Release of active TGF-beta1 from the
748 latent TGF-beta1/GARP complex on T regulatory cells is mediated by integrin beta8. *J*
749 *Immunol* **193**, 2843-2849 (2014).
- 750 55. Genovese, G., *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood
751 DNA sequence. *N Engl J Med* **371**, 2477-2487 (2014).
- 752 56. Jaiswal, S., *et al.* Age-related clonal hematopoiesis associated with adverse
753 outcomes. *N Engl J Med* **371**, 2488-2498 (2014).

- 754 57. Cheng, S., Zhang, W., Inghirami, G. & Tam, W. Mutation analysis links
755 angioimmunoblastic T-cell lymphoma to clonal hematopoiesis and smoking. *eLife*
756 **10**(2021).
- 757 58. Marshall, H.D., *et al.* The transforming growth factor beta signaling pathway is
758 critical for the formation of CD4 T follicular helper cells and isotype-switched
759 antibody responses in the lung mucosa. *eLife* **4**, e04851 (2015).
- 760 59. Yildiz, M., *et al.* Activating STAT6 mutations in follicular lymphoma. *Blood* **125**, 668-
761 679 (2015).
- 762 60. Siddiqi, I.N., *et al.* Characterization of a variant of t(14;18) negative nodal diffuse
763 follicular lymphoma with CD23 expression, 1p36/TNFRSF14 abnormalities, and
764 STAT6 mutations. *Mod Pathol* **29**, 570-581 (2016).
- 765 61. Xian, R.R., *et al.* CREBBP and STAT6 co-mutation and 16p13 and 1p36 loss define the
766 t(14;18)-negative diffuse variant of follicular lymphoma. *Blood Cancer J* **10**, 69
767 (2020).
- 768 62. Mentz, M., *et al.* PARP14 is a novel target in STAT6 mutant follicular lymphoma.
769 *Leukemia* **36**, 2281-2292 (2022).
- 770 63. Behdad, A., *et al.* Survival outcomes of diffuse large B-cell lymphoma by association
771 with concurrent or antecedent follicular lymphoma and double hit status. *Leuk*
772 *Lymphoma* **60**, 3266-3271 (2019).
- 773 64. Wang, Y., *et al.* Impact of concurrent indolent lymphoma on the clinical outcome of
774 newly diagnosed diffuse large B-cell lymphoma. *Blood* **134**, 1289-1297 (2019).
- 775 65. Tilly, H., *et al.* Polatuzumab Vedotin in Previously Untreated Diffuse Large B-Cell
776 Lymphoma. *N Engl J Med* **386**, 351-363 (2022).
- 777 66. Makady, N.F., Ramzy, D., Ghaly, R., Abdel-Malek, R.R. & Shohdy, K.S. The Emerging
778 Treatment Options of Plasmablastic Lymphoma: Analysis of 173 Individual Patient
779 Outcomes. *Clin Lymphoma Myeloma Leuk* **21**, e255-e263 (2021).
- 780 67. Guerrero-Garcia, T.A., Mogollon, R.J. & Castillo, J.J. Bortezomib in plasmablastic
781 lymphoma: A glimpse of hope for a hard-to-treat disease. *Leuk Res* **62**, 12-16 (2017).
- 782 68. Dittus, C., Grover, N., Ellsworth, S., Tan, X. & Park, S.I. Bortezomib in combination
783 with dose-adjusted EPOCH (etoposide, prednisone, vincristine, cyclophosphamide,
784 and doxorubicin) induces long-term survival in patients with plasmablastic
785 lymphoma: a retrospective analysis. *Leuk Lymphoma* **59**, 2121-2127 (2018).
- 786 69. Castillo, J.J., *et al.* Bortezomib plus EPOCH is effective as frontline treatment in
787 patients with plasmablastic lymphoma. *Br J Haematol* **184**, 679-682 (2019).
- 788 70. Sabry, W., Wu, Y. & Kodad, S.G. Bortezomib, Lenalidomide and Dexamethasone
789 Combination Induced Complete Remission in Relapsed/Refractory Plasmablastic
790 Lymphoma: Case Report of a Potential Novel Treatment Approach. *Curr Oncol* **29**,
791 5042-5053 (2022).
- 792
-

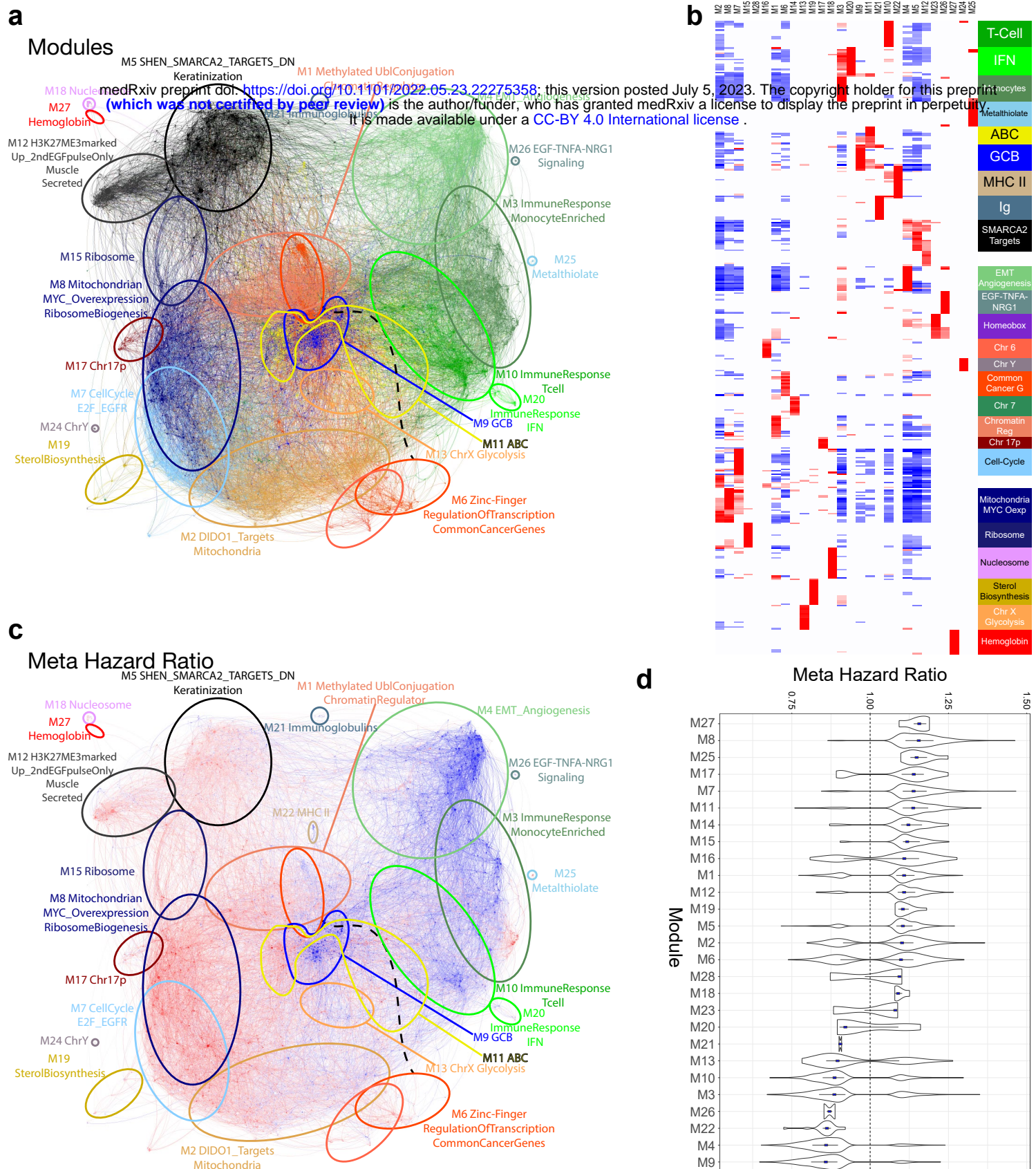
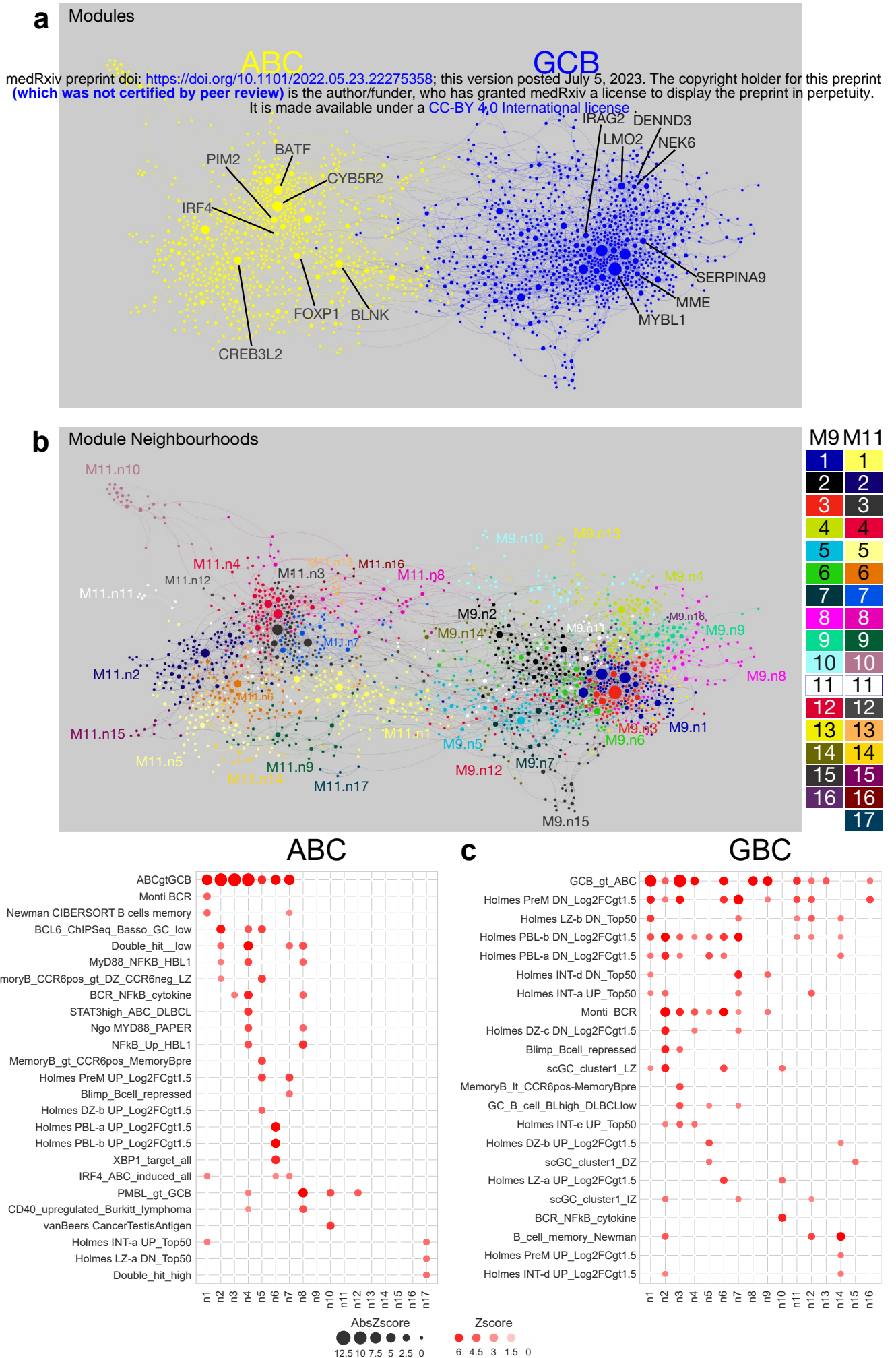


Figure 1. PGCNA network visualization for DLBCL

(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.



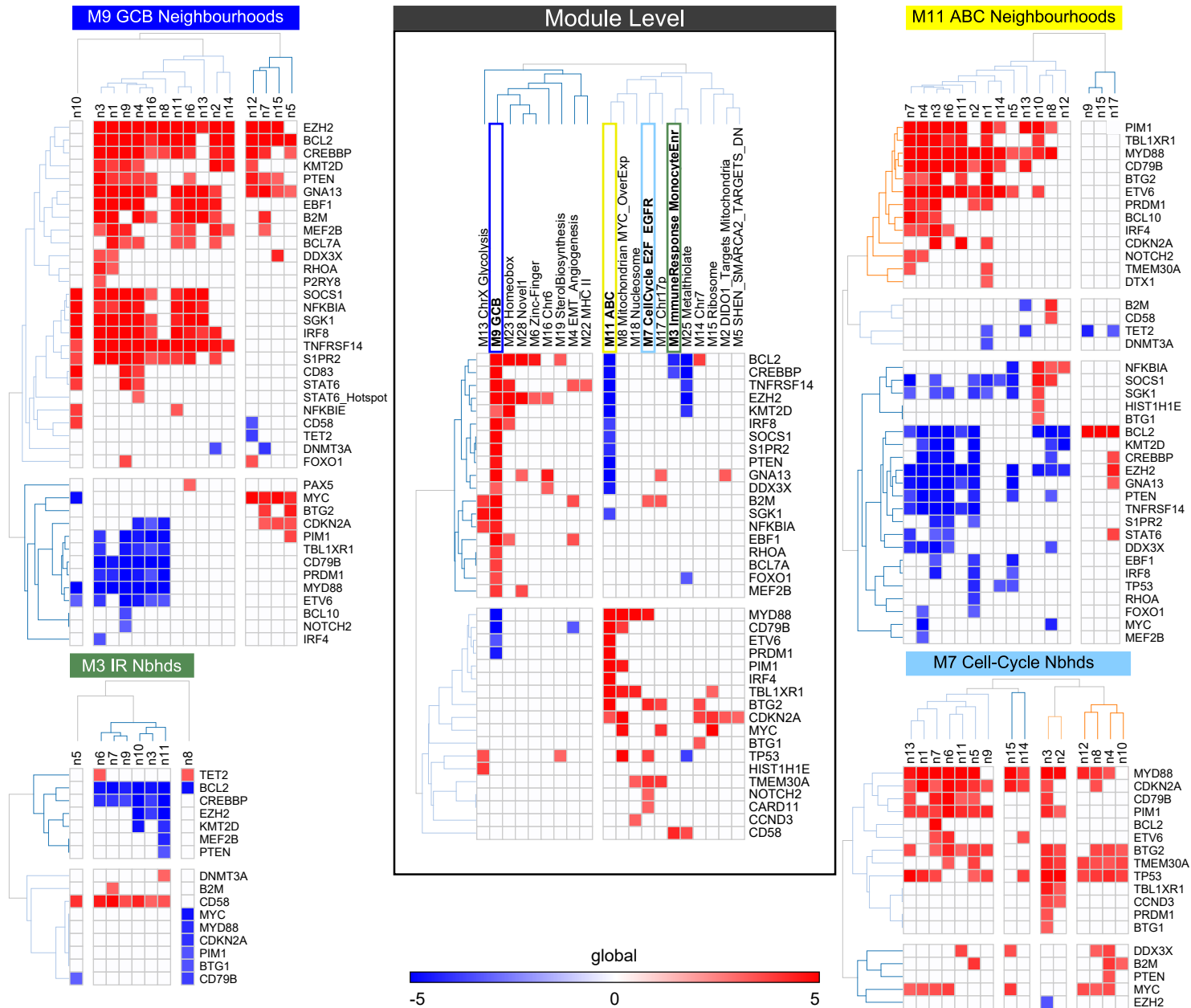


Figure 3. Relationship between MEV/NEVs and mutations

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.

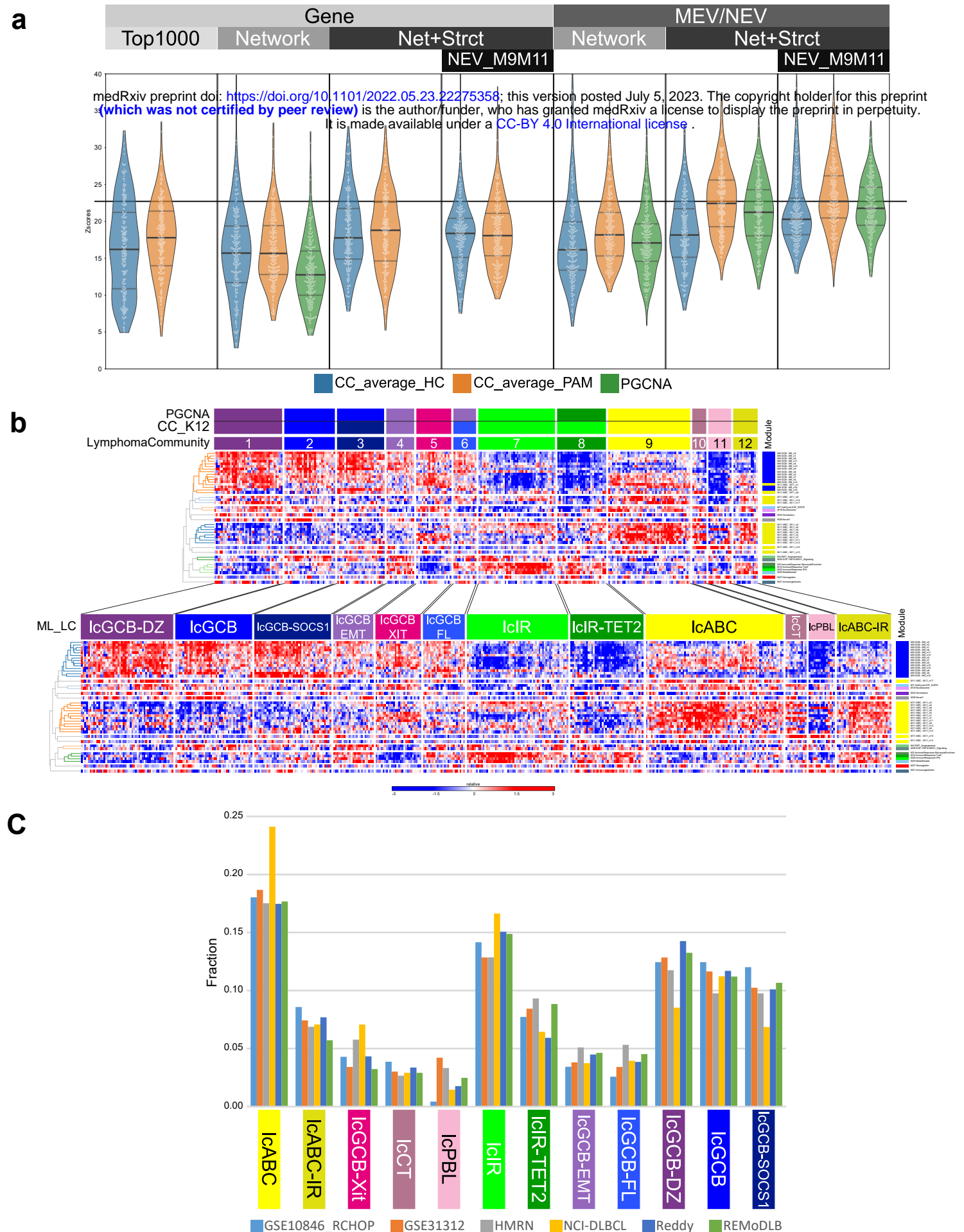
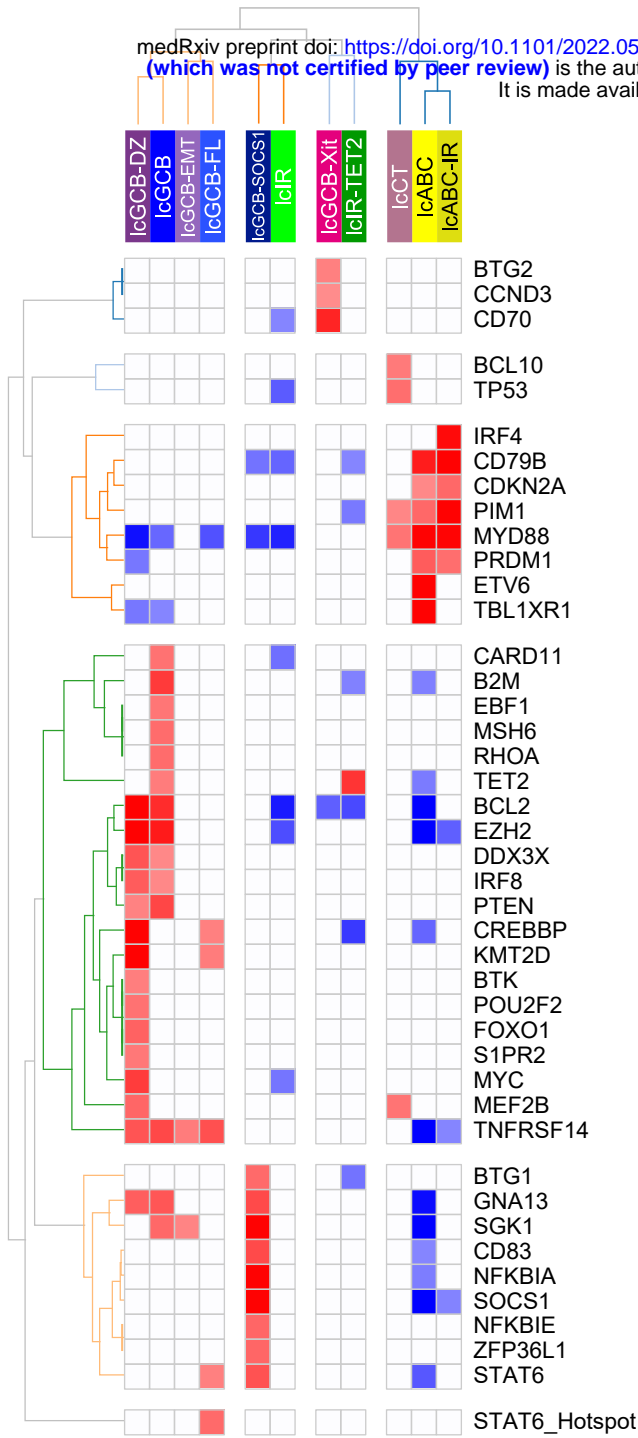


Figure 4. Attribute selection and building a Lymphoma Community classifier.

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.

a



b

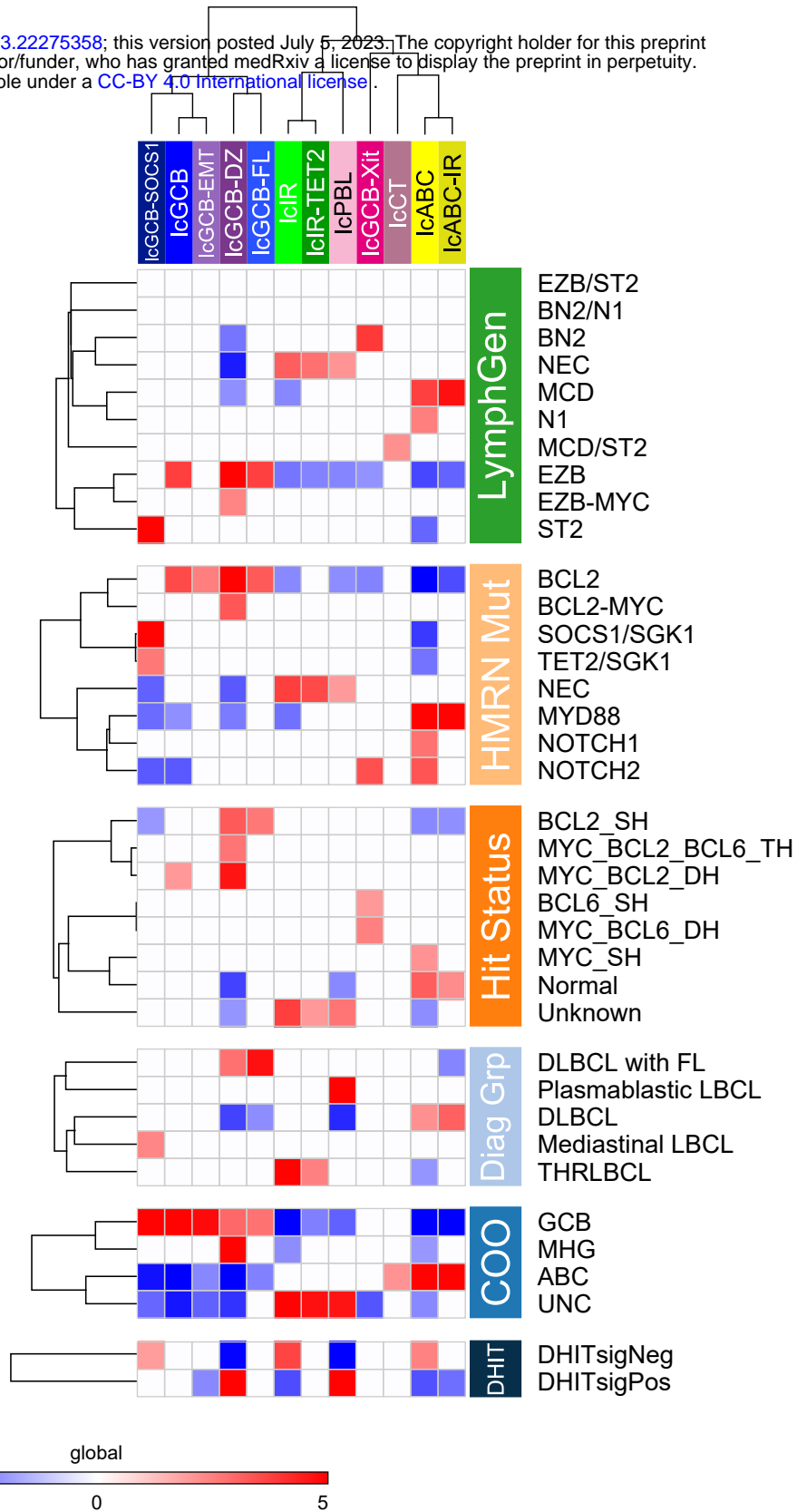


Figure 5. Lymphoma Communities (LC) have distinct mutational and rearrangement associations.

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.

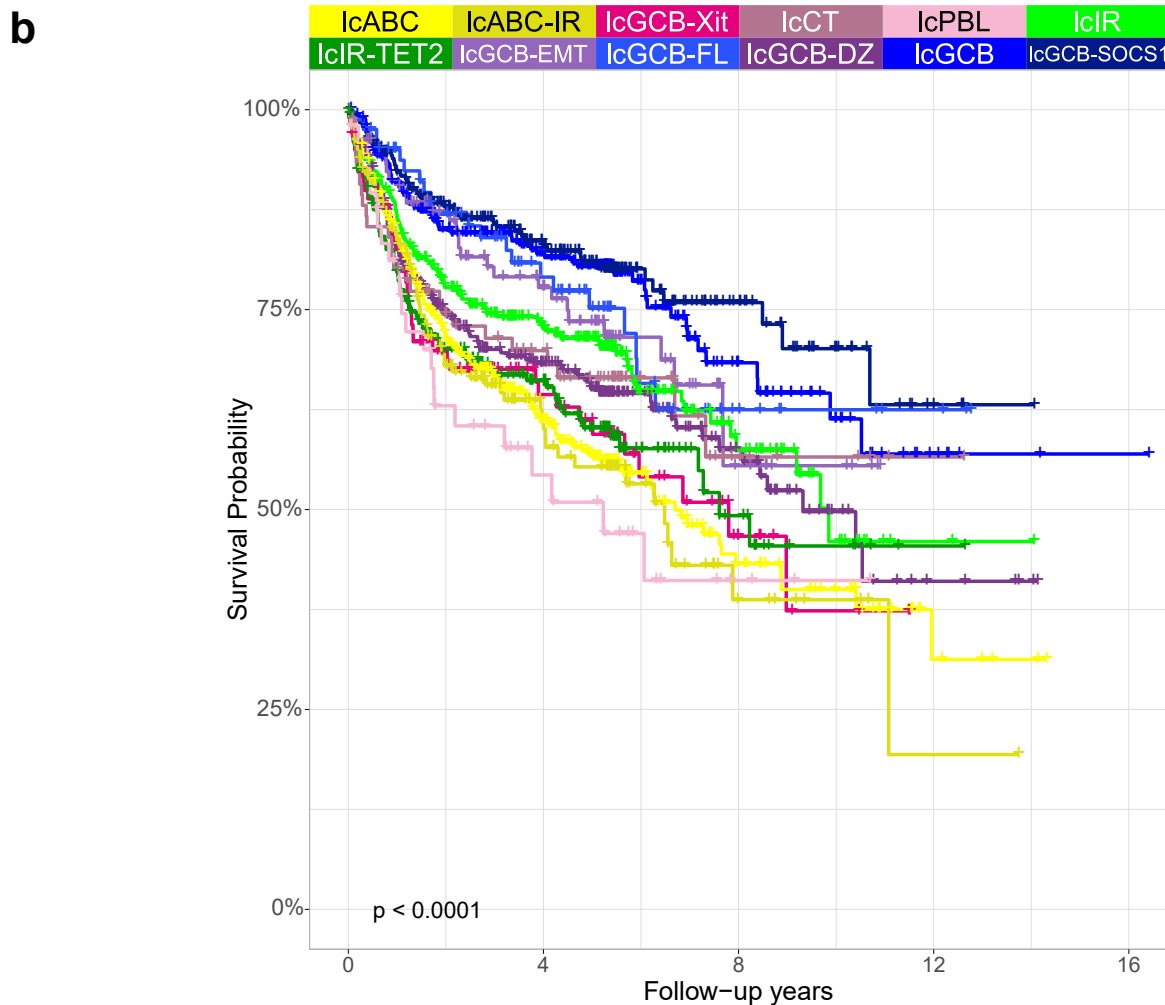
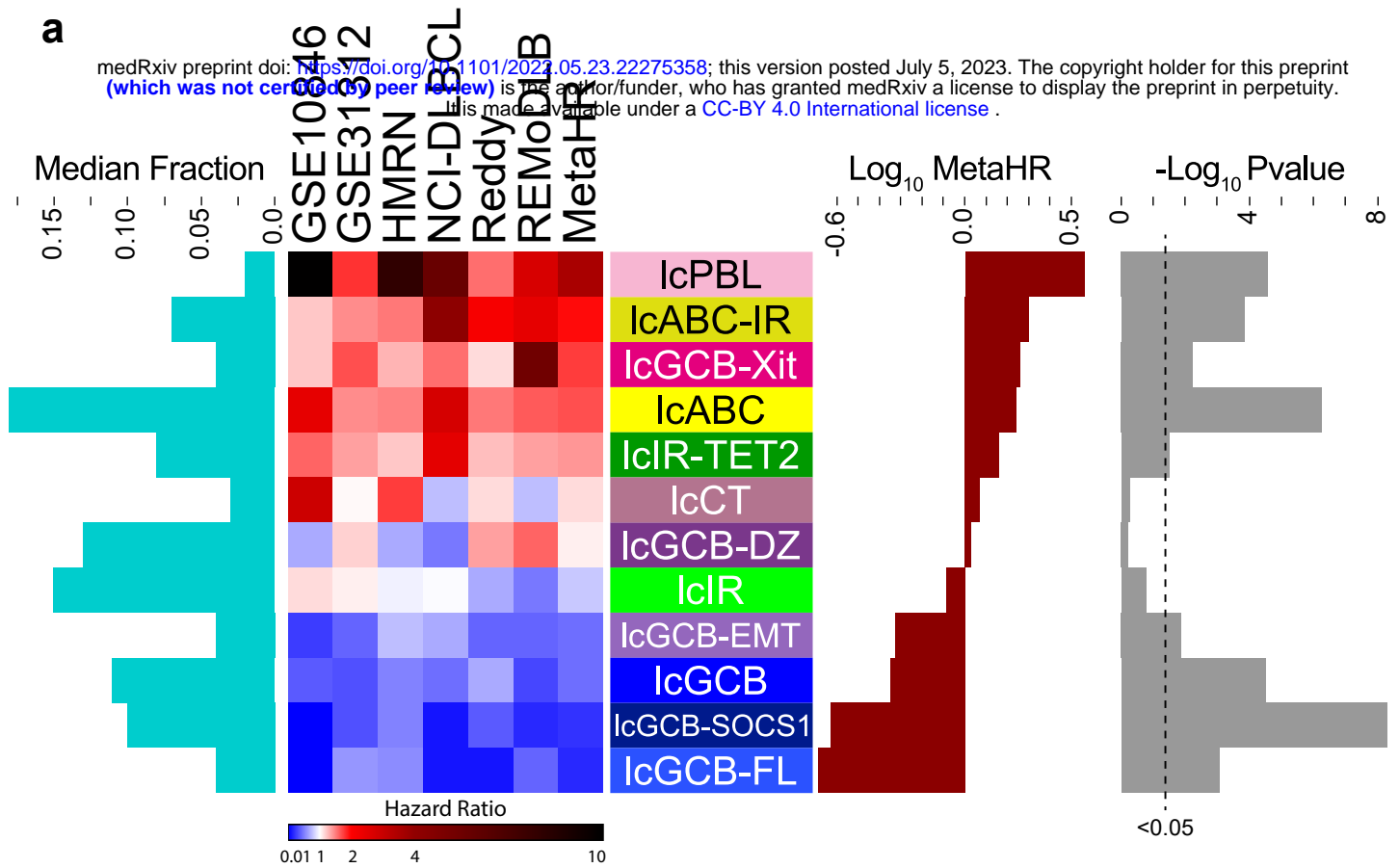


Figure 6. Lymphoma communities are significantly associated with overall survival

(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.

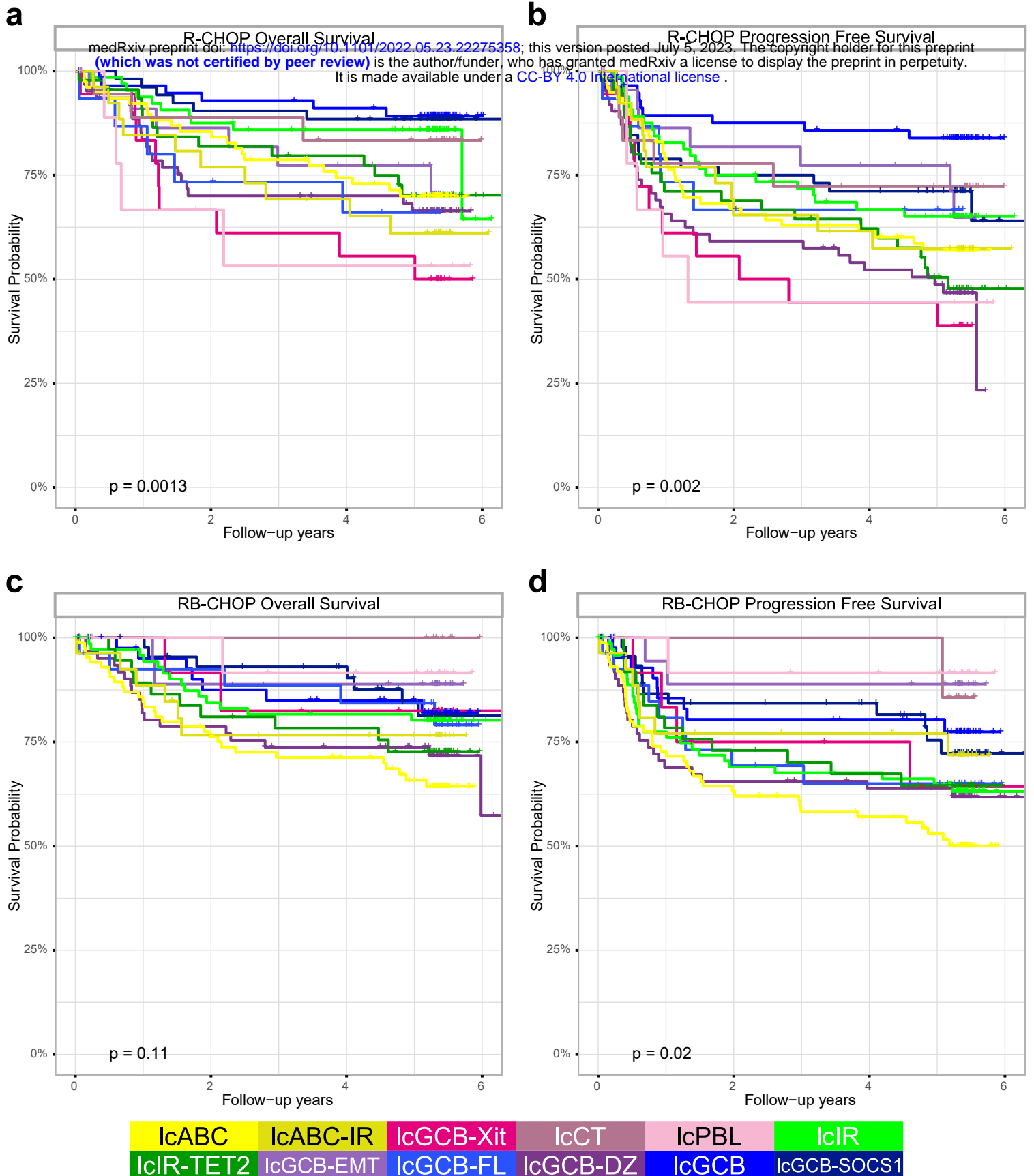


Figure 7. IcPBL and IcCT show improved survival with addition of bortezomib

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

medRxiv preprint doi: <https://doi.org/10.1101/2022.05.23.22275358>; this version posted July 5, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

	IcABC	IcABC-IR	IcGCB-rit	IcC	IcPBL	IcIR	IcIR-TET2	IcGCB-FMT	IcGCB-FEL	IcGCB-DZ	IcGCB	IcGCB-SOS1	Total
τ_{ABC}	120	47	6	14	7	35	16	0	1	1	0	2	249
τ_{GCB}	30	4	18	11	6	48	22	38	40	72	87	92	468
τ_{MHG}	2	0	5	0	7	0	1	0	0	49	16	3	83
τ_{UNC}	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

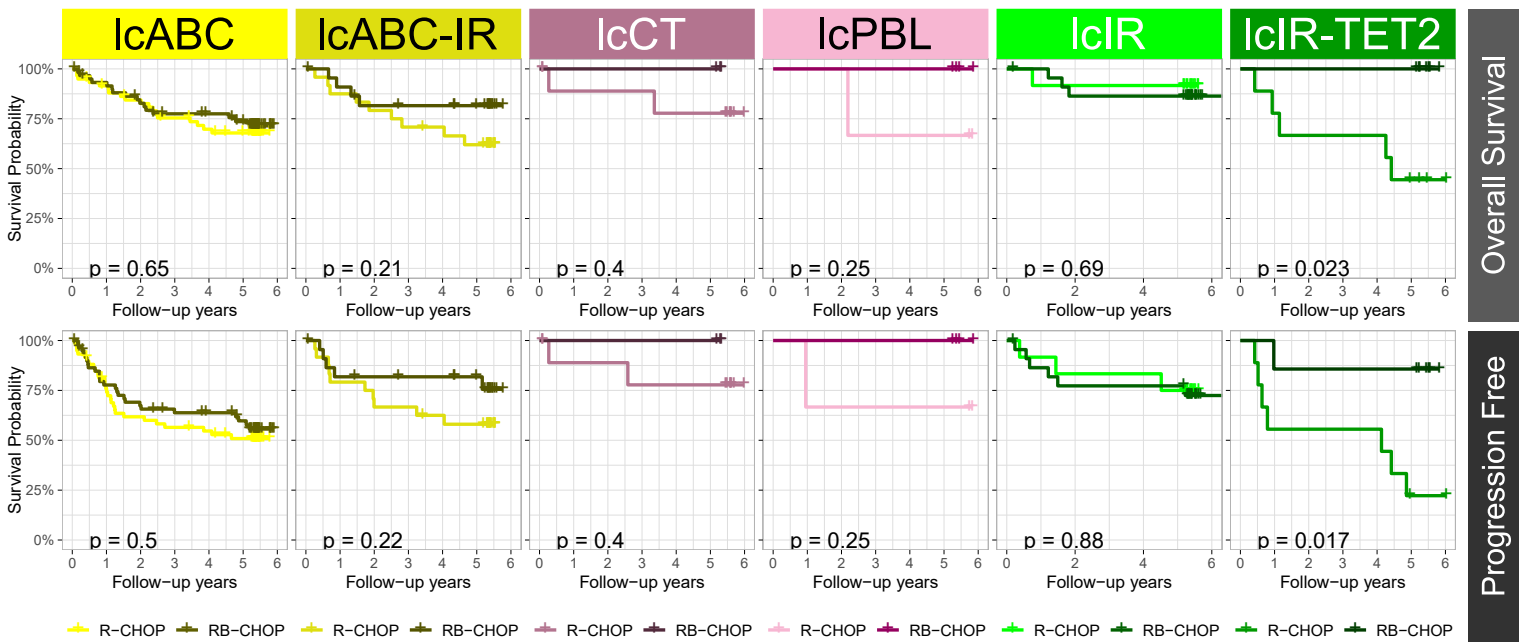


Figure 8. High confidence ABC does not show significant survival advantage with addition of bortezomib

(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 τ_{ABC} 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.