A comprehensive modular map of Diffuse Large B-cell Lymphoma

Matthew A. Care¹,²*, Daniel Painter³, Sharon Barrans⁴, Chulin Sha⁵, Peter Johnson⁶, Andy Davies⁶, Ming-Qing Du⁷, Simon Crouch², Alex Smith², Eve Roman², Cathy Burton⁴, Gina Doody¹, David Westhead², Ulf Klein¹, Daniel Hodson⁸, Reuben Tooze¹,⁴*

¹Division of Haematology and Immunology, Leeds Institute of Medical Research, University of Leeds, Leeds, LS9 7TF, UK;
²Bioinformatics Group, School of Molecular and Cellular Biology, University of Leeds, UK;
³Epidemiology and Cancer Statistics Group, Department of Health Sciences, University of York, York, UK;
⁴Haematological Malignancy Diagnostic Service, St. James's Institute of Oncology, Leeds, UK;
⁵Institute of Basic Medicine and Cancer of Chinese Academy of Sciences, Economic Technology Dev Zone, Hangzhou, 310000, Zhejiang province, China;
⁶University of Southampton, Southampton, UK;
⁷Division of Cellular and Molecular Pathology, Department of Pathology, University of Cambridge, Cambridge CB20QQ, UK;
⁸Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, Puddicombe Way, Cambridge CB20AW, UK;

Correspondence to:
*Reuben Tooze e-mail: r.tooze@leeds.ac.uk
*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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
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. More recently analysis of mutation patterns in DLBCL have converged onto recurrent patterns of co-mutation that can be used to define genomic subtypes. 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.

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. The latter have also been identified in separate stromal survival predictors. 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, or based on direct similarity to cases with MYC and BCL2 DH. These approaches identify overlapping sets of cases and are enriched amongst DLBCL associated with mutations of EZH2 and BCL2 (EZB MYC+).

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. Such analyses have demonstrated that DLBCL can be subdivided based on similarity to physiological single cell expression patterns of GC B-cells. However, single cell studies have also revealed that neoplastic GC B-cells exhibit aberrant patterns of expression and may not faithfully copy physiological counterparts. 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.
expression signatures have been used to define lymphoma microenvironments which identify heterogeneity of host response across both conventional COO classes and genomic DLBCL categories. These approaches conceptually share the use of externally derived signatures to interrogate tumour expression profiles, and have particular value in exploring tumour microenvironment. However, the alignment between expression and genomic assessments is not substantially enhanced. Furthermore, a recurrent feature shared by refined expression based and mutation classifications is a substantial fraction of cases that are not expressly classified.

Consequently, an integrated picture unifying features of expression, mutation and rearrangement status is still developing. Recent studies raise the question of whether the inevitable conclusion of more refined expression analysis is a divergence from, rather than convergence with, recurrent patterns of genomic features. Here we address this issue using an approach anchored on expression correlation and drawing on the resource of close to 5000 cases analysed in DLBCL research studies to date. In contrast to previous studies, this approach views expression patterns directly from the nature of intrinsic co-expression structure eschewing the lens of externally derived signatures. The resultant integrated overview allows separation of consistent DLBCL communities that significantly illuminate relationships between tumour cell state, microenvironmental expressed biology, underlying indolent lymphoma and genetic features of disease.
Methods

(See supplemental methods)

Expression datasets

Thirteen gene expression datasets were used for parsimonious gene correlation network analysis (PGCNA)\(\text{GSE4475, GSE4732, GSE10846, GSE12195, GSE19246, GSE22470, GSE31312, GSE32918, GSE34171, GSE53786, GSE87371, GSE98588, Monti}\)\(^2,12,13,24-33\). 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 \(\geq 0.2\) 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 \(\geq 50\%\) 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 MEVs/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). 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. 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.

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. 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.

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
signature enrichment was used to assess biological features associated with these modules. Representative terms were selected to identify notable expression features associated with each module (Figure 1b, Supplemental Figure 2, Supplemental Table 2).

A number of expectations were met including resolution of two modules encompassing features of the ABC (M11) and GCB state (M9). Additional modules related to cell growth, cell division and metabolism for example separated MYC function, cell cycle, glycolysis, and sterol biosynthesis; host response/microenvironment was separated into modules related to stromal/angiogenesis features, T/NK-cells, monocytes, IFN responses, Immunoglobulin genes, and MHC class-II/antigen presentation; highly co-ordinated gene batteries were resolved for nucleosome components, homeobox, immediate early and metallothionein genes. Several modules related to structural chromosomal regions (e.g. chr6, chr7, chr17p, chrX and chrY). Relationships of genes associated with COO and CCC were recovered. The classifier genes used for separation of ABC and GCB class separation divided discretely across the two respective modules (M11 ABC: p-value 6.95x10^-19 and M9 GCB: 1.57x10^-11), while the more numerous CCC genes divided across multiple modules but with highly significant enrichments (M9:B-cell receptor p-value 3.27x10^-57, M3:Host response p-value 1.64x10^-64 and M13:Oxphos p-value 1.38x10^-05). The network thus provides an integrated picture of gene expression in DLBCL encompassing details of lineage specific gene expression set against a diverse backdrop of cell biology and microenvironment. These are integrated into an extensive online resource covering all underlying gene correlation data, interactive network visualisations and downstream comparisons.

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

Gene co-expression patterns resolved in the network showed consistent associations. A meta-hazard ratio analysis merging overall survival data across the discovery datasets illustrated the association of individual gene expression with outcome across the entire network. This included the segregation of good and adverse outcome between component genes of the GCB and ABC modules (Figure 1c and d), as well as adverse outcome with Mitochondrial and MYC overexpression module (M8) and Ribosome module (M7) and good outcome linked to expression of the Stromal/Angiogenesis module (M4) which encompasses features of known survival predictors.
While there was a clear bias for the association of the majority of module components with a particular outcome, some modules exhibited greater diversity. For example, within the Monocyte module (M3), adverse risk is associated with expression of CD163 and neighbours. These include M544A, a marker of alternative (IL4-driven) macrophage activation, which links CD163 to C1Q complement genes (Figure 1c expanded box). This is consistent with studies identifying adverse outcome in DLBCL linked to alternative macrophage activation and CD163 expression. C1Q has multiple effects on macrophage function including on enhancing apoptotic clearance, and may impact on cancer immune responses. The expression network thus resolved patterns of gene co-expression that reveal differences in outcome related to B-cell state and microenvironment.

Resolving detailed differentiation states in GCB and ABC modules

Biological detail can be resolved at different levels of granularity within the network. An iterative analysis of gene correlation within modules can resolve the most highly related gene neighbourhoods. Assessed systematically across the network, the ABC and GCB modules were amongst those with highest granular information content (Supplemental Figure 3, Supplemental Table 3 and 4). Given the central importance of B cell differentiation to lymphomagenesis, a focused network visualisation was generated to explore the relationships of these GCB and ABC neighbourhoods (Figure 2a and b, and https://mcar.e.link/DLBCL).

The GCB module (M9) resolved into 16 neighbourhoods (Figure 2b and c, Supplemental Table 4). These included neighbourhoods enriched for different GC B-cell subsets - LZ (M9_n2 and n6), DZ (M9_n5 and n15), CD40/NFkB responses linked to LZ B-cells (M9_n10), Intermediate-e/CCR6 memory B-cell precursor (M9_n3 and n4) and pre-memory/memory B-cells (M9_n14). Genes used to classify GCB DLBCL in the COO classifier segregated between two neighbourhoods with one outlier. Thus the original COO classification samples two primary components of GC related co-expression residing at the heart of a wider network of interrelated gene expression (Supplemental Figure 3b).

A similar informative separation was observed for the ABC module which divided into 17 neighbourhoods (Figure 2b and d, Supplemental Table 4). This separated expression related to CCR6+ memory B-cell precursors (M11_n5), pre-memory B-cells (M11_n5 and n7), plasmablast/plasma cell (M11_n6). These cell state associations overlapped with targets of
key transcriptional regulators IRF4 (M11_n1, n6 and n7), BLIMP1 (M11_n7) and XBP1 (M11_n6). A further set of NFκB-response genes were distinguished, specific to the ABC module (M11_n8). The COO ABC-classifier genes were distributed across 5 neighbourhoods, illustrating how these genes sample multiple aspects of the ABC state (Supplemental Figure 3c).

**DLBCL communities are characterised by module and neighbourhood expression patterns**

The set of modules and neighbourhoods encompass features of consistent gene co-expression in DLBCL. Their derivation is independent of external signatures and provides a robust reflection of co-expression derived from DLBCL data alone. The next step in our correlation-centred approach was to assess whether reducing granularity from gene to module/neighbourhood level using module and neighbourhood expression values (MEV/NEV) would allow DLBCL cases to be separated into communities that shared related biology. This approach resolved 12 communities of cases in the previously unseen HMRN DLBCL data. A machine learning classifier trained on the HMRN data retrieved a similar distribution of cases falling into these communities across a range of other data sets (Figure 3a and Supplemental Figure 4 & 5).

The 12 lymphoma communities (LC) aligned to existing paradigms (Figure 3b and c, Supplemental Figure 4) and broadly reinforced the segregation of ABC and GCB expression patterns while providing significantly enhanced and biologically meaningful granularity. Two primary communities of ABC-DLBCL related cases were distinguished as immune response poor (lc\text{ABC}) or immune rich (lc\text{ABC-IR}). A further ABC community was principally linked to memory B-cell neighbourhoods (lc\text{ABC-M}). Cases with plasmablastic features (lc\text{PBL}) were distinguished and were notable for expression of XBP1 targets and cancer testis antigens, as also seen in myeloma. Amongst cases with weak B-cell patterns, two communities were dominated by immune response modules (lc\text{IR-1} and lc\text{IR-2}) with B-cell features including ABC memory B-cell neighbourhoods. An additional community (lc\text{STR}) combined strong Stromal/EMT/Angiogenesis and CD40/NFκB related expression with modest GCB neighbourhood expression. Five communities were characterised by expression of multiple GCB related neighbourhoods. These included a community with distinctively mixed GCB and ABC expression features along with cell cycle, MYC, and sterol biosynthesis modules. These features are in keeping with transition from GCB to an ABC/post-GC state...
Similar proliferation and growth-related module expression was combined with polarised high GC DZ and low GC LZ neighbourhood expression, along with low CD40/NFκB and MHC-II genes in an apparent GCB DZ-like community (γGCB-DZ). The remaining three communities were distinguished based on the consistency of GCB neighbourhood and host response modules as γGCB-1, γGCB-2 and γGCB-3.

**Transition of expression states across DLBCL communities**

We next took advantage of the combination of overview and gene level granularity provided by the network visualisation to assess the relationships between the resolved communities across all network genes or those specific to the GCB/ABC differentiation state, generating a “birds eye view” of expression differences. These analyses illustrate the shifting patterns of gene co-expression separating the resolved DLBCL community structure (Figure 3d and Supplemental Figure 6 and 7). This is particularly exemplified at the level of GCB and ABC neighbourhood genes, where the γGCB-Xit community straddles both key GCB and ABC features and contrasts with the more discrete patterns of other communities such as γGCB-DZ and γABC (Figure 3d). For example, the γGCB-Xit community distinctively co-expresses features of both the GCB (e.g. MYBL1) and ABC (e.g. BATF, CYB5R2) states, which in the wider context of DLBCL are strongly anti-correlated. This co-expression also extends to primary COO classifier genes. Implicit in the binary separation of DLBCL into ABC and GCB classes is the fact that some cases are difficult to classify. Much of this challenge has been attributed to cases dominated by host response features, where the B-cell derived signal is essentially diluted by microenvironmental features. These are captured in the community structure here as γIR-1, γIR-2 or γSTR. Here the identification of the γGCB-Xit community establishes another important feature of DLBCL biology, an intrinsically transitional expression state, as a biological feature of a subset of DLBCL.

**Lymphoma communities and ecosystems**

One of the most advanced methods for exploring cellular composition in DLBCL is the recently proposed concept of ecotyping. Signatures of specific cell states, defined from single cell analysis, are used to deconvolute cellular composition for individual samples. The imputed cell states are then analysed using a community detection approach to assemble cases into ecotypes sharing patterns of cellular composition. We were therefore interested...
to explore the relationship between lymphoma communities and ecotypes in the same data. One difference lay in the fact that ecotype assignments were more restrictive with between 26-32% of cases unassigned across data sets (Figure 4a). The cases unassigned to ecotypes were distributed across a range of communities but $\kappa$ABC-1, $\kappa$GCB-Xit and $\kappa$GCB-DZ were particularly enriched amongst the unassigned cases (Figure 4b; LE_UNC). Amongst cases with lymphoma ecotype assignments (LE1-9) significant overlaps with communities were observed, with LE1 enriched for $\kappa$ABC-1 and LE8 enriched for $\kappa$GCB-1 and -3 communities. LE5 overlapped with $\kappa$ABC-Mem indicating agreement for the detection of cases with preferential memory B-cell differentiation features and ecotype cell states (Figure 4c) emphasised the distinctive microenvironmental association in these cases. Agreement was also observed for LE9 which primarily enriched for $\kappa$STR. For immune response rich cases those assigned to LE4 belonged variously to $\kappa$IR-1, $\kappa$IR-2 and $\kappa$ABC-IR communities. Ecotype LE3 was primarily enriched for $\kappa$GCB-DZ, but much of $\kappa$GCB-DZ along with $\kappa$GCB-Xit were unassigned to ecotype. Thus, deconvolution-centred and correlation-centred analytical approaches converge onto several recurrent elements of DLBCL gene expression. However, differences are apparent relating in particular to immune rich ABC cases ($\kappa$ABC-IR), and specific GCB states.

**Lymphoma communities have prognostic significance**

Gene expression and mutation-based classifications allow prognostic separation of DLBCL. Therefore, in order to consider the lymphoma community structure as a valid framework within which to understand existing classifications an expectation is that DLBCL communities should show significant and reproducible survival differences. We addressed this across the 4 representative validation datasets, also used in the ecotype comparisons, restricting to cases treated with R-CHOP chemo-immunotherapy.\textsuperscript{3,5,13,27,34,35} Across multiple datasets the DLBCL community structure separated risk consistently (Figure 5a and b). Most adverse risk, in terms of meta-HR, was observed for assignment of cases to $\kappa$PBL, followed by $\kappa$ABC-IR, $\kappa$ABC, $\kappa$GCB-Xit and $\kappa$GCB-DZ. Particularly good risk was associated with $\kappa$STR, $\kappa$GCB-2 and to a lesser extent $\kappa$GCB-1 and GCB-3.

At a gene level we identified alternative macrophage expression patterns as a prognostic feature, we therefore asked whether the expression of such macrophage genes was a specific feature of DLBCL communities. Indeed, alternative macrophage gene expression...
was a prominent feature both of the poor risk $\lambda$ABC-IR community and good/indeterminate risk $\lambda$IR-1/-2 communities (Supplemental Figure 8). While these communities are joined in ecotype LE4, the lymphoma community analysis suggests that alternative macrophage activation can be a feature of either very poor risk disease accompanying a detectable ABC expression pattern, or relatively good/indeterminate risk DLBCL with dominant immune response features. Hence, the ability to resolve associations between host response features and outcome can be further informed by consideration of modular and neighbourhood level gene expression patterns. We conclude that the DLBCL community structure is associated with consistent prognostic differences in relation to current standard of care and thus provides both a biologically and clinically meaningful framework.

**DLBCL communities link to mutation patterns**

The most significant recent advance in understanding DLBCL biology has come from analysis of recurrent mutation patterns. Such studies have converged onto recurrent patterns across multiple data sets.\(^2-^5\) However, the integration between mutation and gene expression defined categories remains incomplete. An important test of the DLBCL communities was therefore to assess to what extent these help to explain the relationship between mutation and expression states. To address this we analysed the enrichment of mutations in DLBCL communities by integrating enrichment $p$-values derived across the individual HMRRN, REMoDL-B and Reddy data-sets.\(^5,^34,^35\) At a $p$-value threshold $<0.01$ (in $\geq 2$ datasets) the communities showed significant and distinct associations with mutation type (Figure 6a). $\lambda$ABC was enriched for a characteristic combination of *MYD88*, *CD79B*, *PIM1*, *ETV6*, *TBL1XR1*, *BTG2*, *PRDM1*, *CDKN2A*, *BTG1*, *IRF4* (in order of significance). The two variant communities $\lambda$ABC-IR and $\lambda$ABC-M showed selective associations with *MYD88* and either *PIM1* or *CD79B* mutation respectively, while *MYD88* mutation was anti-correlated with all GCB communities and $\lambda$IR1/2. GCB communities shared enrichment for *BCL2* and *EZH2* mutation while other specific associations separated $\lambda$GCB-1 and $\lambda$GCB-3 from $\lambda$GCB-2 and $\lambda$GCB-DZ or $\lambda$GCB-Xit. For example, $\lambda$GCB-1 and -3 shared enrichment of *NFKBIA*, *SGK1*, and SOCS1 mutations but lacked enrichment for *KMT2D* or *CREBBP* mutations. While, $\lambda$GCB-2 and $\lambda$GCB-DZ shared the reciprocal pattern. And $\lambda$GCB-DZ showed enrichment for mutation in *FOXO1*, *MYC*, *DDX3X*, *PTEN*, *KRAS* and *TP53*, while $\lambda$GCB-Xit showed selective enrichment for *BCL10*, *CD70*, *TMEM30A* and *CCND3* mutations. Consistent with these community level
patterns, concordant associations could be resolved at the level of individual module and
neighbourhood correlations with mutation state (Supplemental Figure 9). The observed associations between communities and mutations resembled patterns underlying current mutational classifications. To directly address this association we assessed mutational classes assigned in HMRN data with either the LymphGen (Figure 6b), or the local HMRN classification (Supplemental Figure 10). These classifications though independently derived are largely concordant and recapitulate the features of the C1-C5 Harvard classification. The LymphGen MCD class showed the expected association with \( \mu \text{ABC-1} \) and \( \mu \text{ABC-M} \), with \( \mu \text{ABC-1} \) also including some cases belonging to the N1 category. The similarly poor risk \( \mu \text{ABC-IR} \) community was notable for failing to be enriched for MCD assignments but was enriched in the related HMRN MYD88 mutation class. This likely reflects differing emphasis on the MYD88 mutation state in particular, combined potentially with the diluting effect of rich immune infiltrates on variant allele detection. In contrast, LymphGen BN2 was significantly and selectively associated with \( \mu \text{GCB-Xit} \). EZB was most significantly enriched amongst \( \mu \text{GCB-2} \) and was significantly depleted from \( \mu \text{GCB-Xit} \). ST2 differed from EZB in showing selective enrichment amongst \( \mu \text{GCB-1} \) but no other \( \mu \text{GCB} \) community. EZB-MYC+ was selectively enriched amongst \( \mu \text{GCB-DZ} \). Cases with an ambiguous classification of EZB/ST2 were restricted in enrichment to the icSTR community. Similar patterns of association were evident when considering the associations of DLBCL communities with the independent but related HMRN mutational clustering of DLBCL cases (Supplemental Figure 10). We conclude that the mutation-based subdivision of DLBCL in LymphGen and related classifications, overlaps significantly with lymphoma communities derived independently and ab initio from gene co-expression patterns.

**LC underlines distinctions between double hit lymphomas**

A further substantial body of work has focused on the separation of aggressive DLBCL based on gene rearrangement status. Wherein, the occurrence of double or triple hit features with rearrangement of \( \text{MYC} \) and \( \text{BCL2} \) (MYC_BCL2-DH) or \( \text{MYC} \) and \( \text{BCL6} \) (MYC_BCL6-DH), or \( \text{MYC}, \text{BCL2} \) and \( \text{BCL6} \) (MYC_BCL2-TH) identifies high risk disease. The relationships of these rearrangement based categories to gene expression and mutation features is relatively well defined for MYC_BCL2-DH, but remains ill-defined for MYC_BCL6-DH leaving the latter category potentially adrift in classification schemes. Of the HMRN cohort 107
cases had rearrangement of one or more of these oncogenes. \( \mu \)GCB-DZ and \( \mu \)GCB-Xit communities showed selective and reciprocal associations with rearrangement status (Figure 6b). \( \mu \)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, \( \mu \)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 \( \mu \)GCB-2 and \( \mu \)GCB-3 with the most significant enrichment being in \( \mu \)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 \( \mu \)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 \( \mu \)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.
Here we have addressed this problem with the aim of presenting a common framework within which to explore relationships between various features of DLBCL tumour biology. The framework is gene expression-based, and uses an inclusive, correlation-centred approach, to derive modules of co-expressed genes that reflect inherent features of the DLBCL data. The resolution of these modules and the underlying neighbourhood substructure is independent of externally derived signatures or prior knowledge regarding patterns of gene co-expression. Signature and gene ontology databases are used to understand the biology identified within the modules of gene co-expression observed in DLBCL data, but the modules and neighbourhoods themselves are derived independently of such prior knowledge. This approach differs fundamentally from the use of externally derived signature or gene co-expression patterns as the primary tool with which to explore and deconvolute gene expression in DLBCL. Our analysis learns from 2,500 cases drawing on principle contributions of the DLBCL research community and validates in nearly 2,500 additional cases from recent studies with mutational and outcome data.²¹²,¹³,²⁴⁻³³ The combination of this extensive combined sample number, and the inclusive “leave no case behind” community clustering has allowed the identification of significant heterogeneity at the expression level, which extends even beyond that of recent deconvolution-centred single cell led approaches to illustrate key features of biology.

These features, summarised in Supplemental Figure 11 and 12, include the distinction of immune rich communities of ABC DLBCL with atypical macrophage activation patterns, the expression-based distinction between double hit lymphomas with MYC and BCL2 or MYC and BCL6 translocation, the identification of a DLBCL community highly significantly linked to underlying follicular lymphoma, and separation at expression level of communities associated with mutational subtypes including distinction between BN2, EZB and ST2 types. This community structure both reinforces and extends existing paradigms by demonstrating inter-relationships and important differences at a modular level of tumour biology which are linked to differentiation state and microenvironment. The community analysis agrees with many assignments reached using the deconvolution-centred signature-led EcoTyper approach, which focuses in particular on cell states in the microenvironment.²² But the community analysis and network structure reinforce the central importance of the ABC and 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 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


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)

HMHRN 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) HMHRN 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 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.
Fig 3

Module Neighbourhoods

HMRN
NCI-DLBCL
REMoDL-B
Reddy2018

Fig 3
**Fig 5**

(a) Hazard Ratio and Median Fraction

- Log10 MetaHR
- Log10 P-value
- Hazard Ratio
- Median Fraction

(b) Survival Probability

- Follow-up years
- Survival Probability

p < 0.0001