



Identification of gene modules associated with survival of diffuse large B-cell lymphoma treated with CHOP-based chemotherapy

YongChao Gao^{1,2,3,4} · Bao Sun^{1,2,3,4} · JingLei Hu^{1,2,3,4} · Huan Ren^{1,2,3,4} · HongHao Zhou^{1,2,3,4} · Ling Chen⁵ · Rong Liu^{1,2,3,4} · Wei Zhang^{1,2,3,4}

Received: 2 June 2019 / Revised: 24 January 2020 / Accepted: 29 January 2020 / Published online: 11 February 2020
© The Author(s), under exclusive licence to Springer Nature Limited 2020

Abstract

Diffuse Large B-cell Lymphoma (DLBCL), a heterogeneous disease, is influenced by complex network of gene interactions. Most previous studies focused on individual genes, but ignored the importance of intergenic correlations. In current study, we aimed to explore the association between gene networks and overall survival (OS) of DLBCL patients treated with CHOP-based chemotherapy (cyclophosphamide combination with doxorubicin, vincristine and prednisone). Weighted gene co-expression network analysis was conducted to obtain insights into the molecular characteristics of DLBCL. Ten co-expression gene networks (modules) were identified in training dataset ($n = 470$), and their associations with patients' OS after chemotherapy were tested. The results were validated in four independent datasets ($n = 802$). Gene ontology (GO) biological function enrichment analysis was conducted with Metascape. Three modules (purple, brown and red), which were enriched in T-cell immune, cell–cell adhesion and extracellular matrix (ECM), respectively, were found to be related to longer OS. Higher expression of several hub genes within these three co-expression modules, for example, *LCP2* (HR = 0.77, $p = 5.40 \times 10^{-2}$), *CD2* (HR = 0.87, $p = 6.31 \times 10^{-2}$), *CD3D* (HR = 0.83, $p = 6.94 \times 10^{-3}$), *FYB* (HR = 0.82, $p = 1.40 \times 10^{-2}$), *GZMK* (HR = 0.92, $p = 1.19 \times 10^{-1}$), *FNI* (HR = 0.88, $p = 7.06 \times 10^{-2}$), *SPARC* (HR = 0.82, $p = 2.06 \times 10^{-2}$), were found to be associated with favourable survival. Moreover, the associations of the modules and hub genes with OS in different molecular subtypes and different chemotherapy groups were also revealed. In general, our research revealed the key gene modules and several hub genes were upregulated correlated with good survival of DLBCL patients, which might provide potential therapeutic targets for future clinical research.

Introduction

Diffuse Large B-cell Lymphoma (DLBCL) was recognized as a neoplasm, which has a larger nuclear size than the benign

histiocytes in the same tissue, with large B-cells arranged in a diffuse pattern [1]. As the most common type of adult non-Hodgkin lymphoma, DLBCL accounts for about 30–40% of lymphoid neoplasms patients around the world [2]. Approximately 25,380 new cases were diagnosed in the United State in 2016, and the incidence was as high as 6.3% [3]. Gastrointestinal tract is the most usual site of the DLBCL, although it may occur in any tissue organ. Thus far, the standard regimen

Supplementary information The online version of this article (<https://doi.org/10.1038/s41397-020-0161-6>) contains supplementary material, which is available to authorized users.

- ✉ Ling Chen
chendoc@126.com
- ✉ Rong Liu
liuronghyw@csu.edu.cn
- ✉ Wei Zhang
csuzhangwei@csu.edu.cn

Hunan Key Laboratory of Pharmacogenetics, 110 Xiangya Road, Changsha 410078, PR China

- ³ Engineering Research Center of Applied Technology of Pharmacogenomics, Ministry of Education, 110 Xiangya Road, Changsha 410078, PR China
- ⁴ National Clinical Research Center for Geriatric Disorders, 87 Xiangya Road, Changsha 410008 Hunan, PR China
- ⁵ Department of Gastrointestinal Surgery, Xiangya Hospital, Central South University, 87 Xiangya Road, Changsha 410008, PR China

¹ Department of Clinical Pharmacology, Xiangya Hospital, Central South University, 87 Xiangya Road, Changsha 410008, PR China

² Institute of Clinical Pharmacology, Central South University,

for these patients is cyclophosphamide combination with doxorubicin, vincristine and prednisone (CHOP) without or with rituximab (R-CHOP). Approximately 60–70% of the patients were sensitive to these treatments, but a proportion of the patients were still refractory to this initial therapy regimen [4]. Thus, it is increasingly imperative to find effective prognostic and therapeutic biomarkers for DLBCL.

The accumulation of DLBCL datasets in the Gene Expression Omnibus (GEO) of NCBI, which generated from different researches and platforms, provides an opportunity to reanalysis these genome wide gene mRNA expression data. Integrated information from various studies could highly lessen the heterogeneity of DLBCL and the variability of the different microarray techniques.

Weighted gene co-expression network analysis (WGCNA), a powerful method of multigene analysis, was widely applied to detect the association between networks/genes and clinical traits. In WGCNA, gene modules were constructed with mRNA profiles by using unsupervised hierarchical clustering, which was different from anteriorly defined genes sets method. The genes, with a consistent expression profile and concordant biological role, were assigned to the same module [5]. WGCNA also has been used to detect co-expression gene modules and biomarkers, which usually be applied to cancer diagnosis [6–9], prognosis [10] and treatment [11, 12]. Similarly, it has been successfully applied in other diseases to identify the genetic determinants, such as osteoporosis [13, 14], obesity [14] and neurodegenerative disease [15]. As for DLBCL, WGCNA has been used in a single dataset contained 93 samples, identifying a class of Cation-transporting ATPase activity related genes for cancer recognition [16], including five hub genes: *BCRP*, *VAOE2*, *CTDS1*, *PHX2B*, *1aMX1A* and *AT10D*. However, the researcher did not confirm these findings in independent dataset.

In our study, WGCNA was applied to analyse a dataset comprising 470 patients to detect gene modules and hub genes that facilitate DLBCL prognosis and treatment. Subsequently, the results we found were verified in four independent datasets ($n = 412, 68, 155, 167$, respectively). The association relationships of the modules and hub genes with overall survival (OS) were tested in different molecular subtypes and within DLBCL patients treated with CHOP or R-CHOP chemotherapy. GO enrichment analysis was conducted in the modules to explore the biological processes that they might be involved.

Materials and methods

Data acquisition

The raw gene expression data were downloaded from the GEO data repository (<http://www.ncbi.nlm.nih.gov/geo/>)

according to the accession numbers GSE34171, GSE10846, GSE31312, GSE4475, GSE32918 and GSE56315. The information of each dataset involved in our study were shown in Table S1. After filtering the samples without survival information, the basic characteristics of these datasets including age, gender, molecular subtype and survival time were summarized in Table 1. The training dataset GSE31312 consists of gene expression of 470 patients, was used to construct the co-expression gene module. GSE34171, GSE10846, GSE4475 and GSE32918 were taken as independent validation datasets. Meanwhile, the dataset GSE56315 was applied to drug sensitivity test. These datasets were labelled with their corresponding GEO accession numbers.

The raw gene expression data of the six datasets that we used were pre-processed with the RMA algorithm by utilizing the ‘affy’ R package (<http://www.bioconductor.org/packages/release/bioc/html/affy.html>). Before processing WGCNA, we filtered out the probes that cannot be annotated as known gene symbols, and the probe-level expressions of these datasets were converted to gene-base expression profiles by using the `collapseRows` R function [17].

Construction of the co-expression gene modules

WGCNA analysis was conducted with the training dataset (GSE31312) by using the R package of ‘wgcna’ [18]. Firstly, we selected the top 15,000 varying genes from the 470 patients in the training dataset after their standard deviations were sorted in an ascending order. Then, we chose the top 10,000 genes based on intramodular connectivity (k_{in}), which were closely related to the gene modules to construct the co-expression gene modules.

To conduct the gene co-expression networks of the selected genes, an unsupervised co-expression relationship was initially built on the basis of the adjacency matrix of connection strengths by using Pearson’s correlation coefficients for gene pairs. This matrix was increased to $\beta = 4$ as based on the scale-free topology criterion (Fig. S1) to lessen the weaker connections between gens and strengthen the strong connections. The network connectivity (k_{total}) of a specific gene was defined as the sum of its weighted adjacency with all the other ones in the integrated network. Meanwhile, the intramodular connectivity (k_{in}) was defined as the standard of the network connectivity of a specific gene and the module it belongs. The topological overlap matrix (TOM) was further used to calculate the expression similarity of each pair of genes. Modules were defined as gene clusters with a high topologic overlap by using the hybrid dynamic tree-cutting method with a minimum gene number of 30 in each module and a dynamic cut height.

The module eigengenes (MEs) that represent the expression level of each module were calculated (supplementary

Table 1 Basic characteristics of datasets.

Characteristic	Training dataset (GSE31312)	Validating dataset 1 (GSE10846)	Validating dataset 2 (GSE34171)	Validating dataset 3 (GSE32918)	Validating dataset 4 (GSE4475)
Sample size	470	412	68	167	87
Age, years mean (sd)	61.83 (14.91)	61.14 (15.45)	–	69.06 (12.90)	59.13 (16.29)
Gender					
Female	199	172	0	62	39
Male	271	222	0	105	48
Unknown	0	18	68	0	0
Stage					
1	124	66	0	0	13
2	96	122	0	0	22
3	101	96	0	0	31
4	128	121	0	0	18
Unknown	21	7	68	167	3
Microarray subtype					
GCB	227	182	–	81	41
ABC	199	167	–	52	29
Unclassified subtype	44	63	–	34	17
Overall survival					
Time, years mean(sd)	2.11 (3.22)	3.18 (3.12)	3.63 (2.19)	3.47 (2.50)	3.21 (3.77)
Death	170	163	17	88	43
Chemotherapy					
CHOP	0	180	0	0	87
R-CHOP	470	232	68	167	0

GCB Germinal Centre B-Cell-like Subtype, *ABC* activated B-Cell-like subtype, *CHOP* cyclophosphamide, doxorubicin, vincristine and prednisone, *R-CHOP* rituximab, cyclophosphamide, doxorubicin, vincristine and prednisone.

file: Dataset 1). Then, the relationship between gene modules and OS was assessed. Meanwhile, module membership assignment (kME) was defined as Pearson's correlation coefficient between gene expression values and MEs, which were defined as the first principal component. Finally, gene significance (GS) of gene in the co-expression module, which represented the biological significance of the gene for clinical traits, was calculated as minus log 10 of the p values by survival analysis [18].

The detailed information of the WGCNA algorithm could be found in the publication of Zhang Bin et al. [19].

DLBCL molecular subtypes

The DLBCL patients that involved in our study were segmented into two molecular subtypes: germinal center B-cell-like subtype (GCB) and activated B-cell-like subtype (ABC). This classification has been proved to be of great significance in understanding the pathogenesis of this disease [20].

Hub genes definition

Hub genes were defined as a such of a class of genes that were strongly associated with clinical traits and displayed a

high network connectivity. The hub genes were identified through GS together k.in. In our study, hub genes were selected by the following norm: (1) top 15 genes with the largest k.in in the module they belonging and (2) GS is larger than 2 (the p value of the association test < 0.01) [6].

Survival analysis

Associations between modules/hub genes and survival were conducted by 'survival' R package. The hazard ratio (HR) and its corresponding 95% confidence interval (CI) were calculated via a Cox regression model. OS was used as the survival endpoints. Survival curves were plotted by utilizing the Kaplan–Meier estimates. For module associations, each ME was divided into high and low expression group by its median value. For single-gene association analysis, the samples were segmented into high and low expression according to the median value of that gene.

These datasets (including GSE31312, GSE10846, GSE34171), from the same platform (Affymetrix U133Plus 2.0 chips, GPL570), were used to performed the survival analysis in different molecular subtypes and treatment

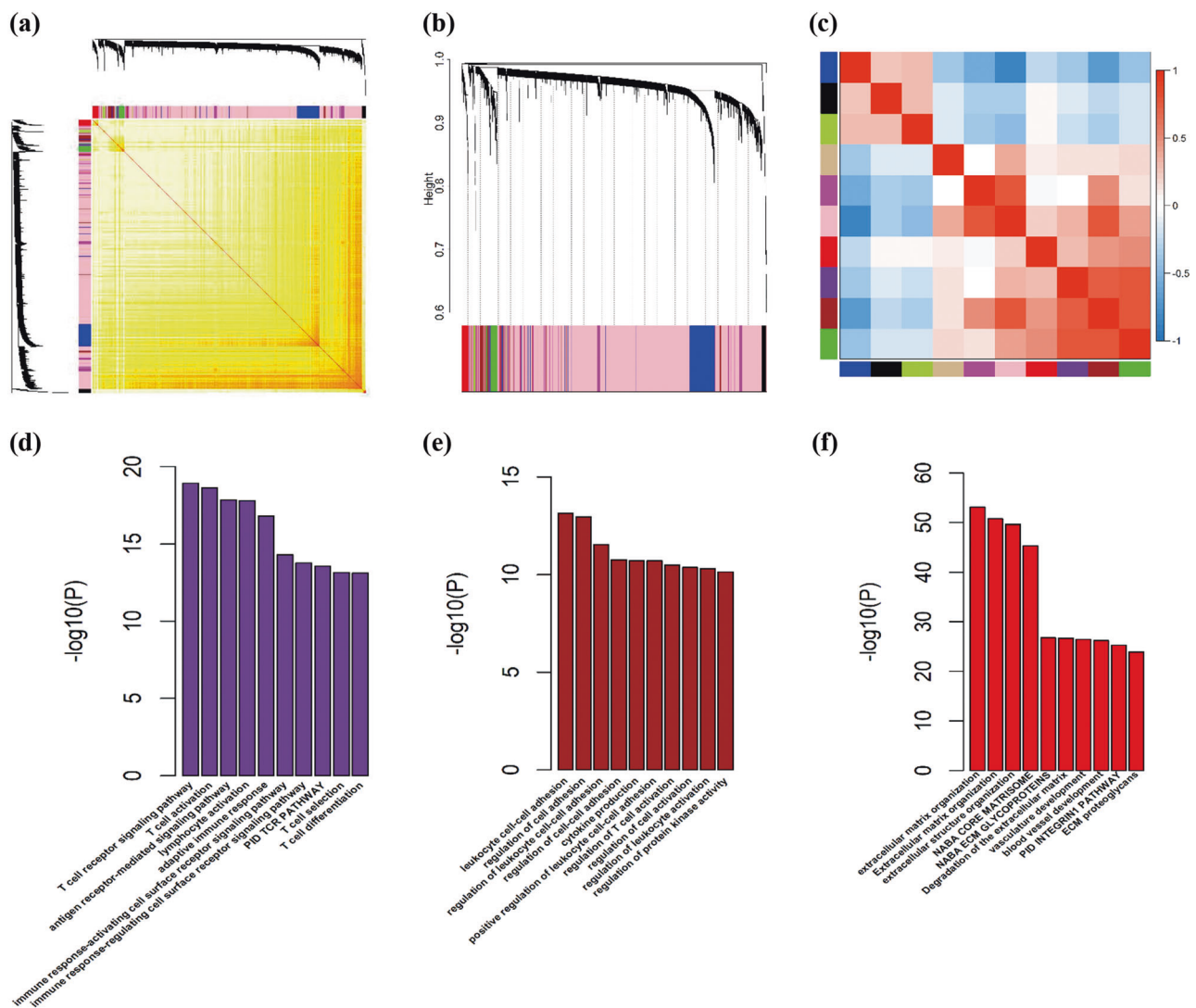


Fig. 1 Identification of co-expression modules with WGCNA. a Heatmap representation the Topological Overlap Matrix (TOM) among all 10,000 genes in the analysis. Degree of overlap is represented by the colour shade; lighter colour represents lower overlap and darker colour represents higher overlap. **b** Hierarchical cluster analysis dendrogram was applied to identify co-expression clusters with the gene profiler from the dataset with 470 patients. Branches are presented as the co-expression modules with highly interconnected genes with different colours to indicate module

subgroups. The ComBat algorithm was utilized to adjust the data for potential batch effects [21].

Functional annotation of gene clusters

Gene ontology (GO) biological function enrichment analysis of the co-expression gene modules was conducted by using Metascape (<https://omictools.com/metascape-tool>). All of the functional annotation clustering option were used for each analysis, and the enrichment score was calculated of the GO biological function. Top ten enriched biological

assignment. Ten modules ranging from 53 to 7154 genes in size were identified (c) PCC matrix between MES. The PCC values range from -1 (blue) to $+1$ (red) depending on the strength of the relationship. GO enrichment analysis for the top 10 ranked genes comprising the purple (d), brown (e) and red (f) modules presented multiple processes. The raw p values reported by Metascape for GO biological processes were transformed to “ $-\log(p \text{ value})$ ” for plotting (supplementary file: Dataset 2–4).

processes of each module in GO analysis were presented in the figure.

Results

Construction of the co-expression gene modules

10,000 genes were derived from 470 DLBCL patients from training dataset to investigate the functional organisation of the DLBCL transcriptome, and construct the co-expression

Table 2 Association between modules and overall survival in the training and validation datasets.

Modules	Gene count	Training dataset (GSE31312)		Validating dataset 1 (GSE10846)		Validating dataset 2 (GSE34171)		Validating dataset 3 (GSE32918)		Validating dataset 4 (GSE4475)	
		HR (95% CI)	p value	HR (95% CI)	p value	HR (95% CI)	p value	HR (95% CI)	p value	HR (95% CI)	p value
Black	147	1.54 (1.13–2.10)	6.66×10^{-3}	1.46 (1.06–2.00)	1.93×10^{-2}	1.74 (0.64–4.70)	2.76×10^{-1}	0.93 (0.62–1.39)	7.09×10^{-1}	1.67 (0.90–3.10)	1.02×10^{-1}
Blue	959	1.09 (0.80–1.48)	5.83×10^{-1}	1.67 (1.21–2.30)	1.92×10^{-3}	2.52 (0.89–7.15)	8.26×10^{-2}	0.85 (0.57–1.28)	4.49×10^{-1}	1.69 (0.91–3.13)	9.65×10^{-2}
Brown	588	0.62 (0.45–0.85)	3.18×10^{-3}	0.54 (0.39–0.76)	2.74×10^{-4}	0.19 (0.05–0.65)	8.32×10^{-3}	1.24 (0.83–1.87)	2.96×10^{-1}	0.37 (0.20–0.69)	1.95×10^{-3}
Green	250	0.78 (0.57–1.05)	1.04×10^{-1}	0.96 (0.70–1.30)	7.78×10^{-1}	0.28 (0.09–0.85)	2.50×10^{-2}	0.94 (0.62–1.40)	7.47×10^{-1}	0.39 (0.21–0.73)	3.32×10^{-3}
Green-yellow	75	1.84 (1.34–2.53)	1.60×10^{-4}	1.56 (1.13–2.14)	6.15×10^{-3}	2.53 (0.89–7.19)	8.07×10^{-2}	0.94 (0.63–1.43)	7.79×10^{-1}	1.62 (0.87–3.00)	1.26×10^{-1}
Magenta	482	0.88 (0.65–1.20)	4.13×10^{-1}	0.72 (0.53–0.99)	4.32×10^{-2}	4.99 (1.43–17.37)	1.15×10^{-2}	1.65 (1.09–2.49)	1.75×10^{-2}	1.00 (0.55–1.83)	9.93×10^{-1}
Pink	7154	0.70 (0.51–0.95)	2.30×10^{-2}	0.55 (0.40–0.77)	4.36×10^{-4}	1.85 (0.69–5.01)	2.24×10^{-1}	1.35 (0.90–2.03)	1.50×10^{-1}	0.83 (0.45–1.51)	5.35×10^{-1}
Purple	81	0.70 (0.52–0.95)	2.37×10^{-2}	0.64 (0.47–0.87)	4.81×10^{-3}	0.18 (0.05–0.62)	6.63×10^{-3}	0.60 (0.40–0.90)	1.37×10^{-2}	0.59 (0.32–1.08)	8.91×10^{-2}
Red	211	0.72 (0.53–0.97)	3.36×10^{-2}	0.55 (0.40–0.76)	2.28×10^{-4}	0.11 (0.03–0.48)	3.44×10^{-3}	1.02 (0.68–1.53)	9.16×10^{-1}	0.48 (0.26–0.89)	2.01×10^{-2}
Tan	53	1.26 (0.92–1.71)	1.45×10^{-1}	1.65 (1.19–2.28)	2.88×10^{-3}	0.73 (0.28–1.93)	5.29×10^{-1}	0.77 (0.51–1.15)	2.00×10^{-1}	2.17 (1.17–4.04)	1.43×10^{-2}

*Significant p values (p value < 0.05) of modules purple, brown and red are given in bold.

gene modules by WGCNA. A total of ten gene modules were obtained (Fig. 1a, b) with gene number from 53 to 7154 (Table 2). Survival analysis were conducted to identify the relationship between these modules and OS as a whole, as well as with different molecular subtypes and different treatment groups. The MEs, generated through principal component analysis, show a general measure of the overexpression information in each module. The module membership between each of the 10,000 genes and the modules of these genes belonging (kMEs) was also calculated. The complete information of the module membership and the network metrics (MEs and kMEs) of each gene was shown in supplementary file: Dataset 1. The correlativity between the ten modules was shown (Fig. 1c).

Gene modules associated with overall survival

The HRs and p values of the MEs were calculated through Cox regression to investigate the relationship between expression of these co-expression gene modules and OS (Table 2). We found that the black, brown, pink, green-yellow, purple and red modules were significantly associated with OS in the training dataset. According to the results of survival analysis in the independent verification datasets, module purple, brown and red were selected for Kaplan–Meier estimates of OS (Figs. 2, S2, S4 and S6). Nevertheless, only the association between the purple module and OS was confirmed across all the validating datasets. The increased expression of genes in the purple module revealed well OS (HR = 0.70, 95% CI = 0.52–0.95, $p = 2.37 \times 10^{-2}$ in the training dataset, Fig. 2a, d, HR = 0.64, 95% CI = 0.47–0.87, $p = 4.81 \times 10^{-3}$ in the validating dataset 1, Fig. S2a, e, HR = 0.17, 95% CI = 0.05–0.62, $p = 6.63 \times 10^{-3}$ in the validating datasets 2, Fig. S2b, f, HR = 0.60, 95% CI = 0.40–0.90, $p = 1.37 \times 10^{-2}$ in validating dataset 3, Fig. S2c, g, HR = 0.59, 95% CI = 0.32–1.08, $p = 8.91 \times 10^{-2}$ in validating dataset 4, Fig. S2d, h). From the results of the other two modules, similar tendencies were shown in multiple databases (Table 2, Figs. S4 and S6), while no significant associations were found in the validating dataset 3. The genes included in the purple module were mapped to T-cell related pathways (Fig. 1d, supplementary file: Dataset 2). GO results showed that cell–cell adhesion pathways were enriched in the brown module (Fig. 1e, supplementary file: Dataset 3), while module red was enriched in extracellular matrix (ECM) related pathways (Fig. 1f, supplementary file: Dataset 4).

Hub genes are associated with OS

Survival analysis were performed for a total of 10,000 genes to test the validity of WGCNA in the finding of novel hub genes that can be used as potential therapeutic markers. The

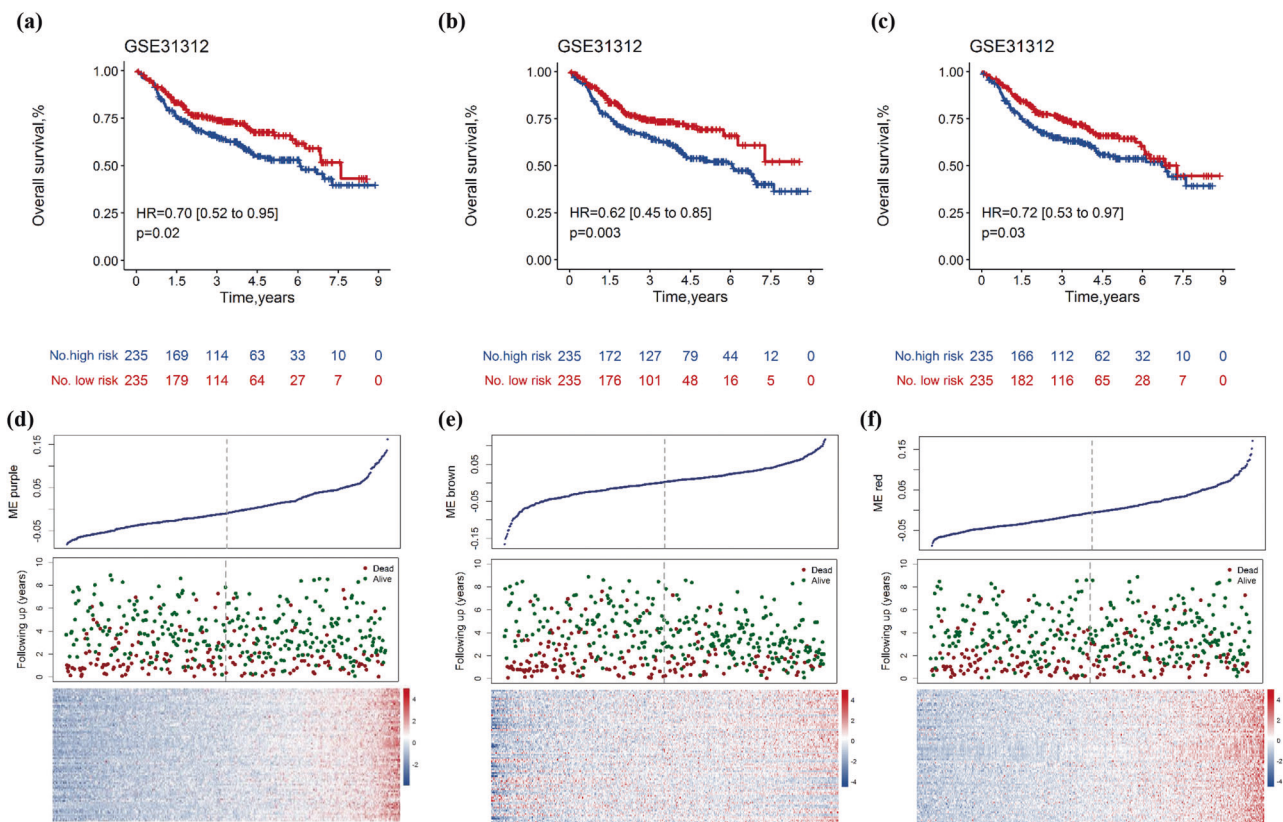


Fig. 2 Association between purple, brown and red modules and overall survival. Kaplan–Meier survival plots for OS were shown (DLBCL patients were divided into two groups based on the median expression of module MEs). Increased expression (red) of the purple module (a, d), brown module (b, e) and red module (c, f) are all associated with good OS in training dataset.

HRs and the corresponding p values of these genes were presented in supplementary file: Dataset 1.

The increased expression of the purple co-expression module containing 81 genes shown a favourable OS outcome (Table 2). In the single-gene survival analysis against OS, 5 hub genes (Table S2) including *CD2*, *CD3D*, *FYB*, *GZMK* and *LCP2* were demonstrated to have striking correlation with good survival outcomes in the purple module. *LCP2* was found to be closely related to OS, not only in the training dataset (HR = 0.56, 95% CI = 0.40–0.78, $p = 5.63 \times 10^{-4}$) but also in the validating datasets (HR = 0.68, 95% CI = 0.53–0.87, $p = 2.52 \times 10^{-3}$ in the validating dataset 1, HR = 0.41, 95% CI = 0.21–0.77, $p = 5.67 \times 10^{-3}$ in the validating dataset 2, HR = 0.61, 95% CI = 0.41–0.90, $p = 1.32 \times 10^{-2}$ in the validating dataset 3, HR = 0.58, 95% CI = 0.38–0.89, $p = 1.20 \times 10^{-2}$ in the validating dataset 4, Table S2). The hub gene *CD3D*, which is closely related to the immune biological processes, has also been identified to be significantly associated with OS (HR = 0.69, 95% CI = 0.58–0.83, $p = 7.30 \times 10^{-5}$ in the training dataset), and the similar trend was found in the validating datasets. Similarly, *CD2* (HR = 0.74, 95% CI = 0.59–0.92, $p = 7.30 \times 10^{-3}$), *FYB* (HR = 0.74, 95% CI = 0.62–0.90, $p = 2.40 \times 10^{-3}$) and *GZMK* (HR = 0.85, 95% CI = 0.75–0.96, $p = 9.44 \times 10^{-3}$)

were all detected to have strong correlation with OS in training dataset (Table S2), and the similar OS-related trends were also found in the validating datasets. Module brown, including 588 genes, showed a well OS with the increased gene expression (Table 2). In this module, 11 hub genes (*ANXA7*, *CAP1*, *ERAP1*, *HIPK3*, *IQGAP1*, *ITM2B*, *PTPRC*, *RAB10*, *RAC1*, *SNX6*, *TRAM1*, Table S2) were found have significant association with better OS through the single-genes survival analysis in the training dataset, and OS-related analysis were conducted in all databases (Table S2). Meanwhile, 211 genes were identified in the red module, which was accompanied with a good OS outcome with the up-regulation of module expression (Table 2). 4 hub genes (*CALD1*, *FNI*, *SPARC* and *SULF1*) in it were positively correlated with favourable OS in the training data base, and the same tendencies were found in the validation databases (Table S2).

Gene modules and hub genes are significantly associated with OS within molecular subtype

The DLBCL patients can be divided into two different molecular subtypes, namely ABC and GCB subtype (Table 1). Based on this, the exploration of the associations

Table 3 Relationship between modules and OS within DLBCL molecular subtypes and different treatment groups.

Modules	Molecular subtype				Chemotherapy			
	ABC subtype (<i>N</i> = 366)		GCB subtype (<i>N</i> = 410)		CHOP like (<i>N</i> = 180)		R-CHOP like (<i>N</i> = 771)	
	HR (95% CI)	<i>p</i> value	HR (95% CI)	<i>p</i> value	HR (95% CI)	<i>p</i> value	HR (95% CI)	<i>p</i> value
Purple	0.81 (0.61–1.10)	1.8×10^{-1}	0.53 (0.36–0.78)	9.17×10^{-4}	0.81 (0.51–1.11)	1.5×10^{-1}	0.63 (0.49–0.82)	4.71×10^{-4}
Brown	0.64 (0.47–0.86)	4.0×10^{-3}	0.56 (0.38–0.83)	4.0×10^{-3}	0.67 (0.45–0.98)	4.0×10^{-2}	0.58 (0.45–0.76)	5.0×10^{-5}
Red	0.79 (0.59–1.06)	1.0×10^{-1}	0.67 (0.46–0.97)	3.0×10^{-2}	0.62 (0.42–0.92)	2.0×10^{-2}	0.63 (0.49–0.81)	3.0×10^{-4}

Hazard ratios (HRs), 95% confidence intervals (CI), and *p* values were calculated using Cox proportional hazard regression analysis, the MEs were dichotomized around their median expressions.

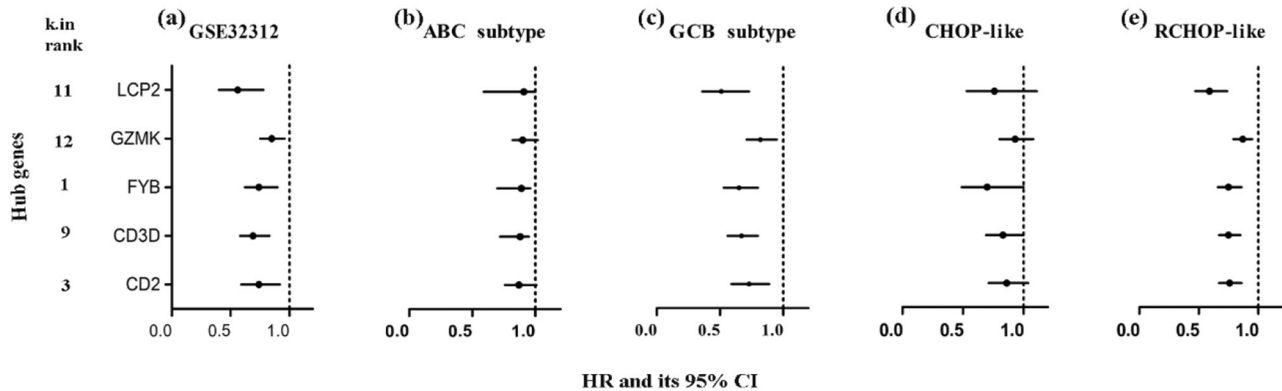


Fig. 3 Forest plots for hub genes of module purple. The HRs (and 95% CI) of the univariate survival analyses in the training dataset (a), within different molecular subtypes (b, c) and different treatment groups (d, e) in the global combined cohort. The ME of hub gene was dichotomized around its median expression.

of gene modules and hub genes with OS in different molecular subtypes was performed by utilizing the merged dataset (including GSE31312, GSE10846 and GSE34171). The increased expression of the purple module was significantly associated with good OS outcomes in the GCB subtype (HR = 0.53, 95% CI = 0.36–0.78, $p = 9.17 \times 10^{-4}$, Table 3, Fig. S3b, f), and the similar trend was shown within the ABC subtype (HR = 0.81, 95% CI = 0.61–1.10, $p = 0.18$, Fig. S3a, e). Meanwhile, the results revealed that the associations between the upregulated expression of hub gene *CD3D* and *FYB* and OS were all reached to statistic difference, in both the ABC and GCB subtype (Table S3, Fig. 3b, c). The other hub genes also have similar trends in both molecular subtypes. In module brown, the significant associations between increased gene and favourable OS outcomes were found in both the ABC subtype (HR = 0.64, 95% CI = 0.47–0.86, $p = 4.0 \times 10^{-3}$, Table 3, Fig. S5a, e) and the GCB subtype (HR = 0.56, 95% CI = 0.38–0.83, $p = 4.0 \times 10^{-3}$, Table 3, Fig. S5b, f). The analysis of hub genes showed that the elevated expression of nine genes, except for *HIPK3* and *RAC1*, were significantly associated with good OS, in the ABC subtype and GCB subtype (Table S3, Fig. 4b, c). The noteworthy association between increased module expression and longer OS has been identified in red module, in the GCB subtype (Table 3,

Fig. S7b, f), but not in the ABC subtype (Table 3, Fig. S7a, e). The similar tendencies were found in two different molecular subtypes by hub genes analysis (Table S3, Fig. 5b, c).

Associations of gene modules and hub genes with OS in different DLBCL chemotherapy regimens

Based on the treatment information of the patients involved in our study, two groups were defined, CHOP-like chemotherapy group and R-CHOP like chemotherapy group (Table 1). The relationships of gene modules and hub genes with OS in different chemotherapy regimens were conducted, and the noteworthy positive association was identified between increased purple module gene expression and good OS in the R-CHOP like chemotherapy group (HR = 0.63, 95% CI = 0.49–0.82, $p = 4.71 \times 10^{-4}$, Table 3, Fig. S3d, h), and the similar tendency was also be found in the CHOP-like chemotherapy group (HR = 0.81, 95% CI = 0.51–1.11, $p = 0.15$, Table 3, Fig. S3c, g). Similarly, the analysis between hub genes and OS showed that DLBCL patients with higher hub genes expression have a good OS, regardless of the chemotherapy regimens (Table S3). Nevertheless, the relationships between hub genes and OS were all have statistical significance only in

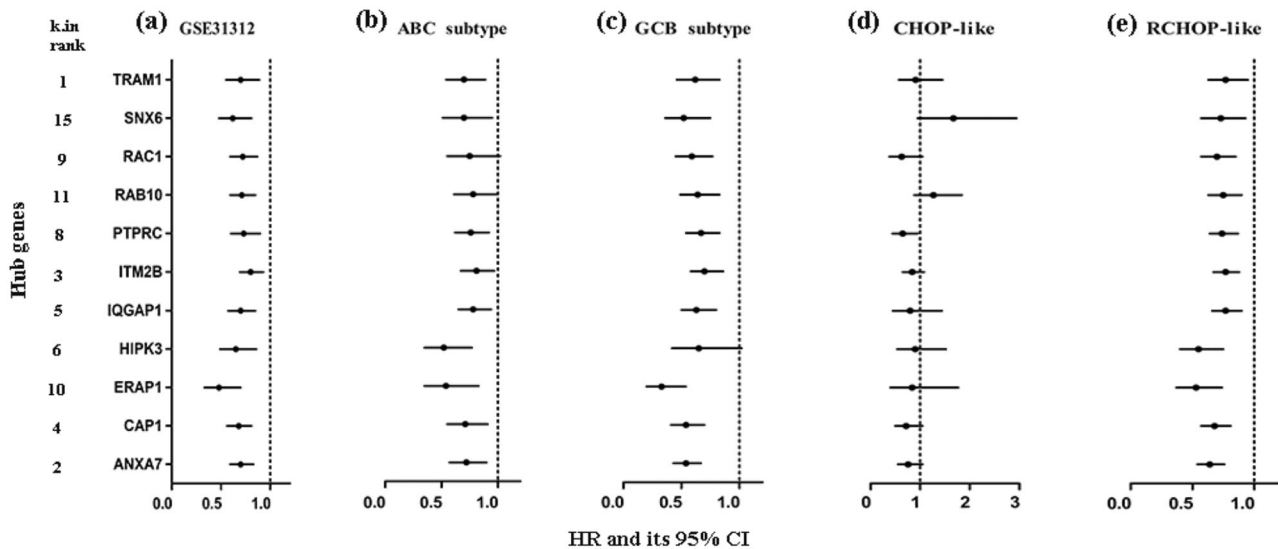


Fig. 4 Forest plots for hub genes of module brown. The HRs (and 95% CI) of the univariate survival analyses in the training dataset (a), within different molecular subtypes (b, c) and different treatment

groups (d, e) in the global combined cohort. The ME of hub gene was dichotomized around its median expression.

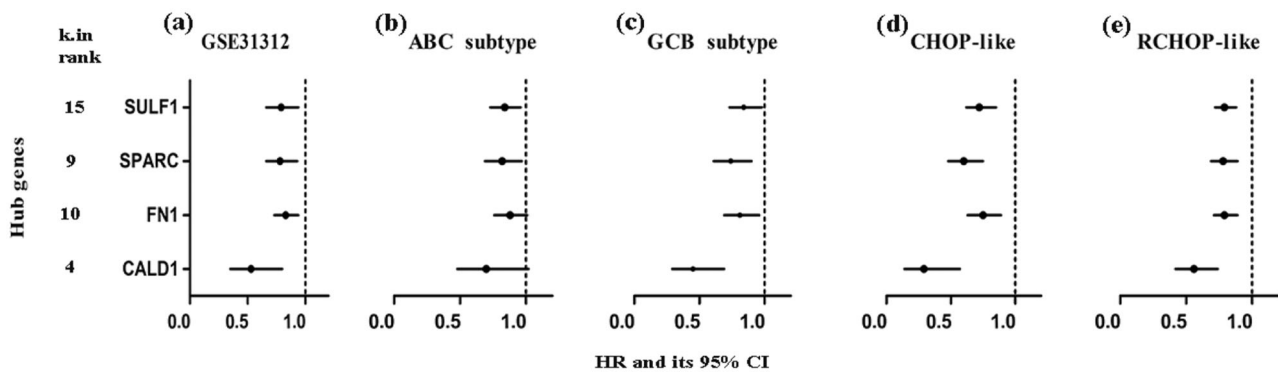


Fig. 5 Forest plots for hub genes of module red. The HRs (and 95% CI) of the univariate survival analyses in the training dataset (a), within different molecular subtypes (b, c) and different treatment groups (d, e) in the global combined cohort. The ME of hub gene was dichotomized around its median expression.

the R-CHOP like chemotherapy group. The results of the other two gene modules showed that DLBCL patients have a well OS with the elevated module expression, in both two chemotherapy groups (Table 3, Figs. S5 and S7). Meanwhile, analysis revealed that DLBCL patients, who have an upregulated hub genes expression under the R-CHOP like chemotherapy treatment, had a longer OS, and the results were statistically significant differences (Table S3, Figs. 4e and 5e). And the results in the CHOP like chemotherapy group showed analogous associations between most hub genes and OS (Table S3, Figs. 4d and 5d).

Correlation between modules expression with drug resistance probability

The patients of GSE56315 were applied to the analysis of drug sensitivity. Interestingly, with an increasing expression

of purple or red or brown gene module, patients were less likely to develop resistance to cyclophosphamide (Fig. 6a–c), doxorubicin (Fig. 6d–f), and vincristine (Fig. 6g–i). The enhanced immune function with the increasing expression of the modules may be a main contributor for these results.

Discussion

In this study, a total of 10 co-expression gene networks were constructed with 10,000 genes from 470 samples by WGCNA, which was applied to detect the relationship between DLBCL transcriptome and OS after chemotherapy treatment. We then confirmed our findings in four independently validating datasets. WGCNA has many obvious advantages over other methods, because the analysis focus on the relationship between co-expression gene modules

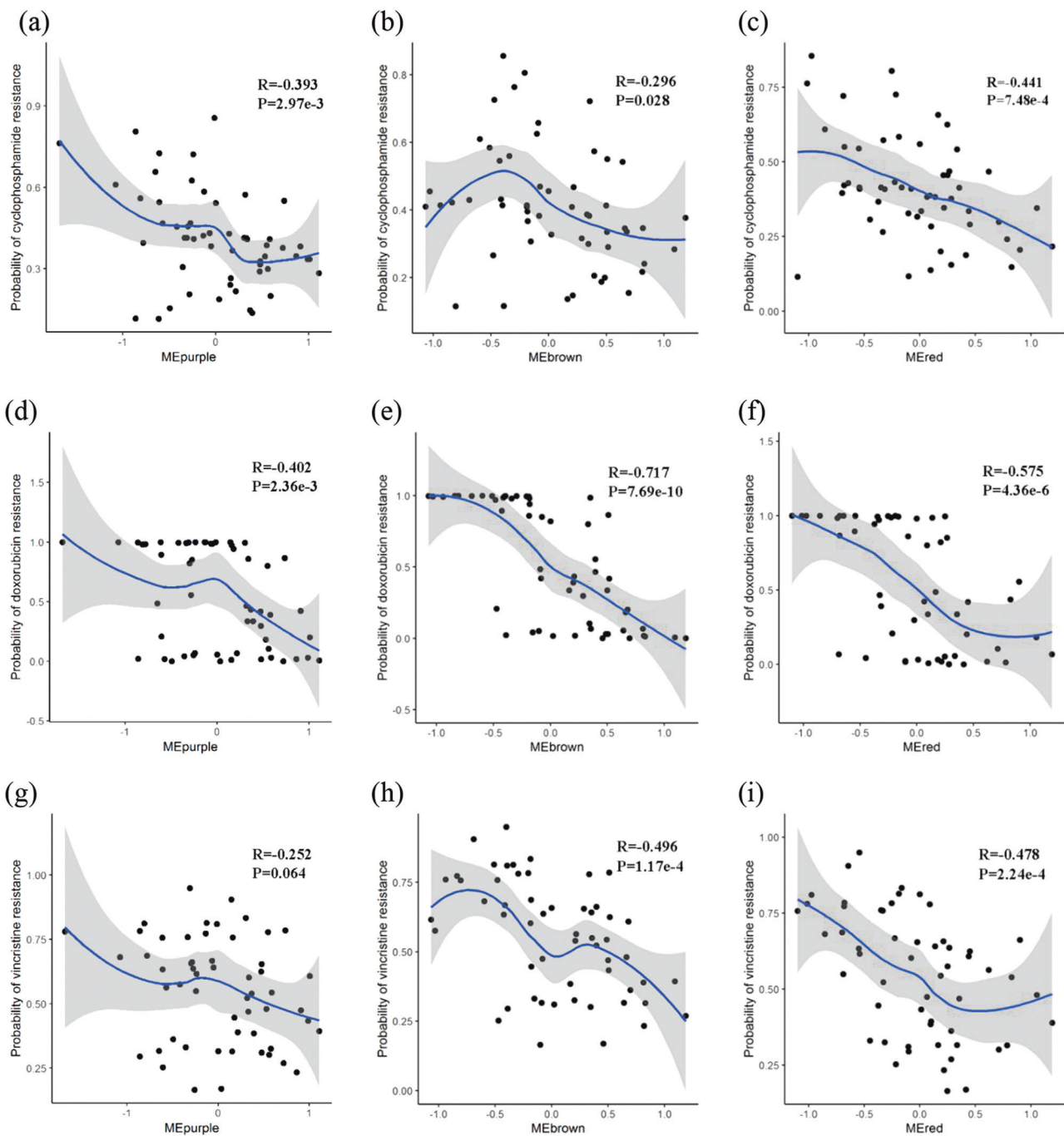


Fig. 6 Association between modules and probability resistance of different chemotherapeutic drugs. Scatter plots for probability chemotherapy resistance of cyclophosphamide (a–c), doxorubicin (d–f), vincristine (g–i) in the module purple, brown and red were shown.

and clinical outcomes, and the results have much higher dependability and biological significance [22], moreover, the consistency among all of samples could be maintained by this method [6]. The genes that are related in function were assigned into the same module. Thus, the analysis was allowed to identified biologically-relevant modules and hub genes, which may eventually serve as biomarkers for diagnosis or treatment.

Among the three modules we selected, module purple was taken as the main discussing object based on the results in all datasets. The elevated expression of purple module comprising 81 genes mostly related to the T-cell immune respond was correlated with positive OS outcome. Based on size and precision of the effect rather than just the *p* value [23], we came to the conclusion that with an increased expression of purple module, the survival time was

significantly lengthened. The results of HR and its corresponding 95% CI of the purple module in the validation datasets demonstrated that the relationship between module purple and OS of DLBCL patients was clinically significant. Similar associations between gene expression and OS were found in two other modules, although the opposite was detected in the validation dataset 3 with no statistically significant difference.

Recently, a profound understanding of the interaction between the tumour microenvironment and neoplastic cells, and the immune escape of cancer, has induced the identification of novel targets in oncotherapy. Based on the idea of improving host immune system to combat cancer, more and more studies focusing on the enhancement of T-lymphocyte function were conducted in various cancers, including DLBCL [24]. The most popular salvage regimens that targeted inhibition of CTLA-4 and PD-1/L1 pathways have been or are being conducted in DLBCL, such as Ipilimumab [25], Nivolumab [26] and Pembrolizumab [27]. Enhancing T-cell response may be an effective way to improve the response to immunotherapy, which is characterised by low efficiency.

Through GO analysis, we found most of the hub genes in purple module are involved in T-cell immune processes. Hub gene *LCP2*, encoding a 533-amino-acid protein contains a single carboxy-terminal SH2 domain, has been demonstrated to be involved in T-cell activation and increasing the activity of the promoter for the *IL-2* gene by its transient overexpression [28]. In children with acute lymphoblastic leukaemia (ALL), *LCP2* has been found to be overexpressed, which may be involved in the pathogenesis and prognosis [29]. The analysis of the public datasets found that the overexpression of *LCP2* may be correlated with metastasis and progression in colon cancer [30] and glioblastoma [31]. However, a study of DLBCL has shown that there was a negative coefficient β between the expression of *LCP2* and survival ($\beta < 0$, that is, the risk ratio < 1 , which indicates that the gene is a protective factor) [32]. These differences may be caused by tumour heterogeneity. Likewise, it shown that the hub gene *CD3D*, an important gene involved in T-cell immune related processes, played an important role in predicting chemotherapy response in the previous report [33]. Most importantly, as one of the major histocompatibility complex class I molecules, the overexpression of *CD3D* indicated an enhanced immune response in DLBCL [34]. Studies in other tumours have also found that a longer OS was associated with an upregulated expression of *CD3D* [35, 36]. The protein encoded by human *CD2* gene was found on all peripheral blood T cells, and played an important role in immune recognition [37]. Study in HER2-positive breast cancer has found a long OS with the increased *CD2* expression [38]. The product of *GZMK* is a member of a group of related serine proteases derives from the cytoplasmic granules

of cytotoxic lymphocytes. The important role of *GZMK* in immune activation and maintenance of immunity has been shown in multiple studies [39, 40]. As the adaptor of FYN protein and LCP2 signals cascades in T cells, the protein encoded by *FYB* is involved in the regulation of interleukin-2 (*IL-2*) expression [41]. Several researches have demonstrated the positive association between *FYB* and T-cell activation [42, 43]. Favourable prognoses have also been found with the elevated expression of *GZMK* and *FYB* in lung cancer [44]. DLBCL studies indicated that higher expression of both *FNI* and *SPARC* correlated with longer survival [45, 46]. However, there are some discrepancies between different cancer studies on these genes, and further investigations are needed to validate the role of these genes in DLBCL.

In our study, the positive correlation between purple module genes expression and OS was found in the GCB subtype ($p = 9.17 \times 10^{-4}$) and the R-CHOP like chemotherapy group ($p = 4.71 \times 10^{-4}$), but the associations in the ACB subtype and the CHOP like chemotherapy group had no statistical difference. These distinct differences may be induced by the diverse mechanisms of the different molecular subtypes [20]. But these associations have no differences between brown module expression and OS in different molecular subtypes as well as chemotherapy cohorts. Meanwhile, less drug resistance probability of three chemotherapeutic drugs (cyclophosphamide, doxorubicin, vincristine) was found with the increased expression of hub genes. Studies have found that effector T cells can abrogate stromal-mediated chemoresistance, which is vital for cancer cell resistance [47], by altering glutathione and cystine metabolism in fibroblasts [48], and the activation of T-cell immune response could reverse chemoresistance [49] in ovarian cancer. Similarly, melanoma related research has reported that chemosensitivity would be increased by activated $CD4^+$ T cells [50]. Activated immunity with the upregulated expression of co-expression module genes in our study may be the main inducement factor of decreased DLBCL chemoresistance, which remains to be further explored.

As a retrospective study, there is no denying that there are limitations. Firstly, a major deficiency of our study was that the lack of further experimental verification for hub genes such as *CD3D*. Secondly, although the tendencies of the relationship between the modules and OS in multiple validating datasets were similar to that in the training dataset, some p values showed only marginal significance. Thirdly, although these hub genes were found to positively correlated with patients' survival, there were still some discrepancies with other tumour related researches that have not been elucidated. Therefore, the significance of gene modules and hub genes should be further verified in clinical practice.

In conclusion, WGCNA was applied to construct gene co-expression network, identifying and validating network

hub genes associated with the survival of DLBCL patients. Three modules and multiple hub genes were identified to play key roles in patients' survival, most of which were new-found to be implicated in DLBCL, and further molecular experiments were required to elucidate detail mechanisms. All in all, our work may provide clues into means of therapeutically regulating the host immunity for promoting immune recognition and immunotherapy of DLBCL.

Acknowledgements The authors appreciate Dr Yong Zhou, The Third Xiangya Hospital of Central South University, for his help.

Funding This study was supported by the National Natural Science Foundation of China (No. 81573511, 81874329, 81522048 and 31801121), the National Key Research and Development Program (No. 2016YFC0905000, 2016YFC0905001), and Scientific Foundation of Xiang Ya hospital (2016Q04).

Author contributions RL, LC and WZ designed the study, RL analysed the data and made the tables and figures, YCG drafted the manuscript; BS, HR, RL, JLH and WZ revised the whole paper; LC polished the whole manuscript; WZ approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

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

References

1. Swerdlow SH, Campo E, Pileri SA, Harris NL, Stein H, Siebert R, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*. 2016;127:2375–90.
2. Swerdlow SH. WHO classification of tumours of haematopoietic and lymphoid tissues. In: IARC World Health Organization classification of tumours. (Lyon, France: WHO, 2008).
3. Teras LR, DeSantis CE, Cerhan JR, Morton LM, Jemal A, Flowers CR. US lymphoid malignancy statistics by World Health Organization subtypes. *CA Cancer J Clin*. 2016;66:443–59.
4. Sehn LH, Gascoyne RD. Diffuse large B-cell lymphoma: optimizing outcome in the context of clinical and biologic heterogeneity. *Blood*. 2015;125:22–32.
5. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science*. 2003;302:249–55.
6. Liu R, Zhang W, Liu ZQ, Zhou HH. Associating transcriptional modules with colon cancer survival through weighted gene co-expression network analysis. *BMC Genomics*. 2017;18:361.
7. Liu R, Lv QL, Yu J, Hu L, Zhang LH, Cheng Y, et al. Correlating transcriptional networks with pathological complete response following neoadjuvant chemotherapy for breast cancer. *Breast Cancer Res Treat*. 2015;151:607–18.
8. Liu R, Cheng Y, Yu J, Lv QL, Zhou HH. Identification and validation of gene module associated with lung cancer through coexpression network analysis. *Gene*. 2015;563:56–62.
9. Wang T, Zheng X, Li R, Liu X, Wu J, Zhong X, et al. Integrated bioinformatic analysis reveals YWHAB as a novel diagnostic biomarker for idiopathic pulmonary arterial hypertension. *J Cell Physiol*. 2019;234:6449–62.
10. Liu R, Guo CX, Zhou HH. Network-based approach to identify prognostic biomarkers for estrogen receptor-positive breast cancer treatment with tamoxifen. *Cancer Biol Ther*. 2015;16:317–24.
11. Zhang L, Kang W, Lu X, Ma S, Dong L, Zou B. Weighted gene co-expression network analysis and connectivity map identifies lovastatin as a treatment option of gastric cancer by inhibiting HDAC2. *Gene*. 2019;681:15–25.
12. Ge Y, He Z, Xiang Y, Wang D, Yang Y, Qiu J, et al. The identification of key genes in nasopharyngeal carcinoma by bioinformatics analysis of high-throughput data. *Mol Biol Rep*. 2019;46:2829–40.
13. Chen YC, Guo YF, He H, Lin X, Wang XF, Zhou R, et al. Integrative analysis of genomics and transcriptome data to identify potential functional genes of BMDs in females. *J Bone Min Res*. 2016;31:1041–9.
14. Hu Y, Tan LJ, Chen XD, Liu Z, Min SS, Zeng Q, et al. Identification of novel potentially pleiotropic variants associated with osteoporosis and obesity using the cFDR method. *J Clin Endocrinol Metab*. 2018;103:125–38.
15. Rangaraju S, Dammer EB, Raza SA, Rathakrishnan P, Xiao H, Gao T, et al. Identification and therapeutic modulation of a pro-inflammatory subset of disease-associated-microglia in Alzheimer's disease. *Mol Neurodegener*. 2018;13:24.
16. Lu X, Deng Y, Huang L, Feng B, Liao B. A co-expression modules based gene selection for cancer recognition. *J Theor Biol*. 2014;362:75–82.
17. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, et al. Strategies for aggregating gene expression data: the collapseRows R function. *BMC Bioinform*. 2011;12:322.
18. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform*. 2008;9:559.
19. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005;4: Article17.
20. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl J Med*. 2002;346:1937–47.
21. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8:118–27.
22. Chou WC, Cheng AL, Brotto M, Chuang CY. Visual gene-network analysis reveals the cancer gene co-expression in human endometrial cancer. *BMC Genomics*. 2014;15:300.
23. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. *Nat Methods*. 2015;12:179–85.
24. Schuster SJ, Bishop MR, Tam CS, Waller EK, Borchmann P, McGuirk JP, et al. Tisagenlecleucel in adult relapsed or refractory diffuse large B-cell lymphoma. *N. Engl J Med*. 2019;380:45–56.
25. Ansell SM, Hurvitz SA, Koenig PA, LaPlant BR, Kabat BF, Fernando D, et al. Phase I study of ipilimumab, an anti-CTLA-4 monoclonal antibody, in patients with relapsed and refractory B-cell non-Hodgkin lymphoma. *Clin Cancer Res*. 2009;15:6446–53.
26. Lesokhin AM, Ansell SM, Armand P, Scott EC, Halwani A, Gutierrez M, et al. Nivolumab in patients with relapsed or refractory hematologic malignancy: preliminary results of a phase Ib study. *J Clin Oncol*. 2016;34:2698–704.
27. Zinzani PL, Ribrag V, Moskowitz CH, Michot JM, Kuruvilla J, Balakumaran A, et al. Safety and tolerability of pembrolizumab in patients with relapsed/refractory primary mediastinal large B-cell lymphoma. *Blood*. 2017;130:267–70.
28. Motto DG, Ross SE, Wu J, Hendricks-Taylor LR, Koretzky GA. Implication of the GRB2-associated phosphoprotein SLP-76 in T

- cell receptor-mediated interleukin 2 production. *J Exp Med*. 1996;183:1937–43.
29. Niini T, Vetteranta K, Hollmen J, Larramendy ML, Aalto Y, Wikman H, et al. Expression of myeloid-specific genes in childhood acute lymphoblastic leukemia - a cDNA array study. *Leukemia*. 2002;16:2213–21.
 30. Chu S, Wang H, Yu M. A putative molecular network associated with colon cancer metastasis constructed from microarray data. *World J Surg Oncol*. 2017;15:115.
 31. He WQ, Gu JW, Li CY, Kuang YQ, Kong B, Cheng L, et al. The PPI network and clusters analysis in glioblastoma. *Eur Rev Med Pharm Sci*. 2015;19:4784–90.
 32. Li C, Zhu B, Chen J, Huang X. Novel prognostic genes of diffuse large B-cell lymphoma revealed by survival analysis of gene expression data. *Onco Targets Ther*. 2015;8:3407–13.
 33. Denkert C, Loibl S, Noske A, Roller M, Muller BM, Komor M, et al. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. *J Clin Oncol*. 2010;28:105–13.
 34. Linderoth J, Eden P, Ehinger M, Valcich J, Jerkeman M, Bendahl PO, et al. Genes associated with the tumour microenvironment are differentially expressed in cured versus primary chemotherapy-refractory diffuse large B-cell lymphoma. *Br J Haematol*. 2008;141:423–32.
 35. Györfy B, Bottai G, Fleischer T, Munkacsy G, Budczies J, Paladini L, et al. Aberrant DNA methylation impacts gene expression and prognosis in breast cancer subtypes. *Int J Cancer*. 2016;138:87–97.
 36. Shi MJ, Meng XY, Wu QJ, Zhou XH. High CD3D/CD4 ratio predicts better survival in muscle-invasive bladder cancer. *Cancer Manag Res*. 2019;11:2987–95.
 37. Tibaldi EV, Salgia R, Reinherz EL. CD2 molecules redistribute to the uropod during T cell scanning: implications for cellular activation and immune surveillance. *Proc Natl Acad Sci USA*. 2002;99:7582–7.
 38. Han J, Choi YL, Kim H, Choi JY, Lee SK, Lee JE, et al. MMP11 and CD2 as novel prognostic factors in hormone receptor-negative, HER2-positive breast cancer. *Breast Cancer Res Treat*. 2017;164:41–56.
 39. Jiang W, Chai NR, Maric D, Bielekova B. Unexpected role for granzyme K in CD56bright NK cell-mediated immunoregulation of multiple sclerosis. *J Immunol*. 2011;187:781–90.
 40. Wensink AC, Kemp V, Fermie J, Garcia Laorden MI, van der Poll T, Hack CE, et al. Granzyme K synergistically potentiates LPS-induced cytokine responses in human monocytes. *Proc Natl Acad Sci USA*. 2014;111:5974–9.
 41. Veale M, Raab M, Li Z, da Silva AJ, Kraeft SK, Weremowicz S, et al. Novel isoform of lymphoid adaptor FYN-T-binding protein (FYB-130) interacts with SLP-76 and up-regulates interleukin 2 production. *J Biol Chem*. 1999;274:28427–35.
 42. Griffiths EK, Krawczyk C, Kong YY, Raab M, Hyduk SJ, Bouchard D, et al. Positive regulation of T cell activation and integrin adhesion by the adapter Fyb/Slap. *Science*. 2001;293:2260–3.
 43. Medeiros RB, Burbach BJ, Mueller KL, Srivastava R, Moon JJ, Highfill S, et al. Regulation of NF-kappaB activation in T cells via association of the adapter proteins ADAP and CARMA1. *Science*. 2007;316:754–8.
 44. Weng TY, Wang CY, Hung YH, Chen WC, Chen YL, Lai MD. Differential expression pattern of THBS1 and THBS2 in lung cancer: clinical outcome and a systematic-analysis of microarray databases. *PLoS ONE*. 2016;11:e0161007.
 45. Lossos IS, Czerwinski DK, Alizadeh AA, Wechsler MA, Tibshirani R, Botstein D, et al. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N Engl J Med*. 2004;350:1828–37.
 46. Meyer PN, Fu K, Greiner T, Smith L, Delabie J, Gascoyne R, et al. The stromal cell marker SPARC predicts for survival in patients with diffuse large B-cell lymphoma treated with rituximab. *Am J Clin Pathol*. 2011;135:54–61.
 47. Senthebane DA, Rowe A, Thomford NE, Shipanga H, Munro D, Mazeedi M, et al. The role of tumor microenvironment in chemoresistance: to survive, keep your enemies closer. *Int J Mol Sci*. 2017;18:1586–616.
 48. Wang W, Kryczek I, Dostal L, Lin H, Tan L, Zhao L, et al. Effector T cells abrogate stroma-mediated chemoresistance in ovarian cancer. *Cell*. 2016;165:1092–105.
 49. Xu S, Tao Z, Hai B, Liang H, Shi Y, Wang T, et al. miR-424(322) reverses chemoresistance via T-cell immune response activation by blocking the PD-L1 immune checkpoint. *Nat Commun*. 2016;7:11406.
 50. Radfar S, Wang Y, Khong HT. Activated CD4+ T cells dramatically enhance chemotherapeutic tumor responses in vitro and in vivo. *J Immunol*. 2009;183:6800–7.