Tumor Purity-Related Genes for Predicting the Prognosis and Drug Sensitivity of DLBCL Patients

Zhenbang Ye[1#], Ning Huang[2#], Yongliang Fu[1], Rongle Tian[1], Liming Wang[2*], Wenting Huang[1,3*]

[1]Department of Pathology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China;

[2]Department of Hepatobiliary Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China

[3]Department of Pathology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, 518116, China.

#These authors contributed equally to this work.

*Corresponding author

Wenting Huang, Ph.D.

Department of Pathology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China; No. 17 Panjiayuan Nanli, Chaoyang Dist., Beijing, PRC.

Department of Pathology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, 518116, China; No. 113 Baohe Ave. Longgang Dist. Shenzhen, PRC.

E-mail: huangwt@cicams.ac.cn

Liming Wang, M.D.

Department of Hepatobiliary Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, No. 17 PanjiayuanNanli Road, Chaoyang District, Beijing 100021, China

Email: stewen_wang@sina.com

Abstract

Background: Diffuse large B-cell lymphoma (DLBCL) is the predominant type of malignant B-cell lymphoma. Although various treatments have been developed, the limited efficacy calls for more and further exploration of its characteristics.

Methods: Datasets from Gene Expression Omnibus (GEO) database were used for identifying the tumor purity of DLBCL. Survival analysis was employed for analyzing the prognosis of DLBCL patients. Immunohistochemistry was conducted to detect the important factor that influenced the prognosis. Drug sensitive prediction was performed to evaluate the value of the constructed model.

Results: VCAN, CD3G and C1QB were identified as three key genes that impacted the outcome of DLBCL patients both in GEO datasets and samples from our center. Among them, VCAN and CD3G+ T cells were correlated with favorable prognosis, and C1QB was correlated with worse prognosis. The ratio of CD68+ macrophages and CD8+ T cells was associated with better prognosis. In addition, CD3G+ T cells ratio was significantly correlated with CD68+ macrophages, CD4+ T cells and CD8+ T cells ratio, indicating it could play an important role in the anti-tumor immunity in DLBCL. The riskScore model constructed based on the RNASeq data of VCAN, C1QB and CD3G work well in predicting the prognosis and drug sensitivity.

Conclusion: VCAN, CD3G and C1QB were three key genes that influenced the tumor purity of DLBCL, and could also exert certain impact on drug sensitivity and prognosis of DLBCL patients.

Keywords: DLBCL, tumor purity, VCAN, CD3G, C1QB

Introduction

The latest refined classification by the World Health Organization (WHO) categorizes large B-cell lymphoma as a heterogeneous group of B-cell lymphomas[1]. Diffuse large B-cell lymphoma (DLBCL) is the most prevalent type among them, accounting for around 30% of all non-Hodgkin lymphomas. DLBCL can be classified into three subtypes based on its immunohistochemical expression patterns: germinal center B-cell-like (GCB), activated B-cell-like (ABC), and unclassified[2]. Moreover, an additional classification is established by evaluating the immunohistochemical expression patterns according to Hans algorithm, leading to two subtypes:

GCB and non-GCB[3]. After undergoing R-CHOP chemotherapy, about 60% of patients achieve long-term remission; however, approximately 30% of patients experience relapse, resulting in poor prognosis and a considerable number of deaths from refractory lymphoma[4]. Consequently, to explore the characteristics of DLBCL in detail is urgently needed for developing more effective therapy.

Solid tumor tissue comprises tumor cells and the surrounding stroma, which encompasses diverse types of matrix cells, immune cells, endothelial cells[5], etc. The tumor microenvironment (TME) is a complex and dynamic system that consists of the extracellular matrix and a variety of cellular components. Recent studies have unveiled multiple subgroups of immune cells within the microenvironment of DLBCL, including T cells, B cells, NK cells, monocytes/macrophages, dendritic cells, as well as the distribution of stromal cell components like fibroblasts and endothelial cells[6, 7]. Despite the relatively limited composition of the TME in DLBCL, its role in tumor proliferation and evasion of the immune system should not be disregarded. The interaction between tumors and the microenvironment is a vital factor that impacts the development and prognosis of B-cell lymphoma[8]. Nevertheless, the existing research on the influence of the TME on the prognosis of DLBCL patients is limited and lacks a consensus.

Moreover, the comprehensive investigation of non-immune cell components in the TME is still lacking. Previous research on stroma in DLBCL has predominantly indicated that a higher quantity of extracellular matrix is associated with a more favorable prognosis, while increased vascular density is associated with poorer prognosis[9]. Furthermore, higher stromal scores have been associated with an improved prognosis in DLBCL patients[10]. Additionally, a fibrotic tumor microenvironment has been correlated with a better prognosis after DLBCL chemotherapy and immunotherapy[11]. These research findings stem from computational analysis of stromal and immune scoring in gene databases and have not been experimentally validated as of yet.

Tumor purity quantifies the relative ratio of tumor cells to the surrounding stromal components in solid tumors, elucidating the dynamics between tumor cells and their microenvironment[12]. It can partly reflect the characteristics of TME, namely, a higher tumor purity indicates a lower abundance of stromal components in TME. Tumor purity is associated with patient prognosis, and the strength of this association varies across different tumor types[13-15]. Therefore, when investigating the influence of TME on the prognosis of DLBCL, it is crucial to analyze not only

the immune cell components but also the significance of non-immune cell components.

This study utilized bioinformatic analysis to establish the relationship between immune and stromal components and the prognostic outcomes of DLBCL patients. We developed a novel immunohistochemical panel to assess prognostic outcomes and treatment sensitivity by detecting the expression of VCAN, CD3G, C1QB, CD68, CD4 and CD8 in both the TME and tumor cells of 190 DLBCL patients. We then explored their relationship with DLBCL clinicopathological features as well as overall survival (OS).

Materials and Methods

Data collection and tumor purity-related genes (TPGs) selection

The RNA-Sequence and clinical data of GSE53786 and GSE32918 datasets were download from Gene Expression Omnibus (GEO) database. The first gene symbols of GSE53786 datasets were retained when one probe detected multiple genes. Average expression value of genes in each dataset were calculated and used when one gene was detected by multiple probes. Tumor purity was assessed by ESTIMATE (Estimation of Stromal and Immune cells in Malignant Tumor tissues using Expression data) algorithm[16] and the then its correlation with genes expression was analyzed. The genes with $|r| \geq 0.5$ and p value < 0.05 was defined as the TPGs.

TPGs function analysis

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were executed to analyze the biological processes, cellular components, molecular functions and pathways related to the TPGs. The statistical significance was considered as p.adjust < 0.05.

The protein-protein interactions (PPI) analysis was utilized to investigate the interaction among TPGs, and those with interactive confidence greater than 0.90 on the STRING platform (version 11.5) were selected to establish an interaction network with Cytoscape software (version 3.8.2).

Prognostic model

The prognostic model was constructed with “survival” package in R (version 4.1.3). The genes enrolled in this model was selected among the prognostic and PPI hub TPGs by function “step” in “survival” package, which can optimize the model. The prognostic model was represented by

$\text{riskScore} = \sum_{i=1}^{n} \text{gene expression}_i \times \text{coef}_i$.

Clinical specimens and follow-up

190 patients from Cancer hospital Chinese Academy of Medical Sciences, the CHCAMS cohort, were enrolled in this study (**Supplementary table 1**). All patients received surgery or biopsy during September, 2010 and September, 2020, and then standard follow-ups were carried out until March, 2023. The overall survival (OS) was defined as the interval between the operation and death or the last follow-up. The specimens from the CHCAMS cohort were used for immunohistochemistry assay. The study was designed according to the Declaration of Helsinki and approved by the institutional ethics committee of Cancer Hospital Chinese Academy of Medical Sciences. Informed consent was taken from all the patients.

IHC

Paraffin embedded DLBCL tissues of CHCAMS Cohort were used for immunohistochemistry (IHC). After de-paraffinization and hydration, heat-induced method was performed for antigen retrieval. Primary antibody of VCAN (AB177480, 1:100, Abcam, USA), CD3G (AB134096, 1:1000, Abcam, USA), C1QB (AB92508, 1:50, Abcam, USA), CD68 (303565, 1:1000, Abcam, USA), CD4 (ZM-0418, ZSGB-BIO, China), CD8 (ZA-0508, ZSGB-BIO, China), CD206(24595S, 1:400, CST, USA), and CD32(15625-1-AP, 1:1000, proteintech, China) was incubated at 4°C overnight. Sections were washed with TBS-T buffer, and then incubated with secondary antibody, and finally stained with DAB. The quantitative analysis of the slices was conducted by QuPath-0.4.3. VCAN and C1QB were assessed by H-score, and CD3G, CD68, CD4, CD8, CD206, CD32 were assessed as the ratio of the corresponding positive cells among all cells.

Drug sensitivity prediction

Drug sensitivity prediction was conducted utilizing “oncoPredict” packages in R 4.1.3. The drug sensitivity data was collected from Genomics of Drug Sensitivity in Cancer (GDSC). The drugs that was analyzed in this study was selected according to clinical practice or clinical trials searched in Pubmed.

Statistical analysis

Data in this study was shown in the form of mean ± SEM. Correlation between two variates was determined with Spearman analysis. Kaplan–Meier (K–M) curve and Log rank test were used for survival analysis. The cut-offs of survival analysis were provided by X-tile. The independent risk factor analysis was performed with Cox regression analysis. Receiver operating characteristic (ROC) curve was used for test the efficacy of prognostic model. The clinicopathological characteristics difference analysis was conducted with χ^2 test, Fisher's exact test or Wilcoxon rank sum test. The drug sensitivity scores were compared with Wilcoxon rank sum test. In this study, $p < 0.05$ were considered statistically significant.

Results

Tumor purity related genes were correlated with extracellular matrix organization and immune response

Based on GSE53786 dataset, we first assessed the tumor purity of DLBCL, which ranged from 17.2% to 67.4% (**Figure 1A**). In order to screen out the TPGs, we then analyzed the correlation between genes expression and tumor purity. According to the thresholds mentioned above, 642 genes were identified as TPGs, among which 31 genes were positively correlated with tumor purity, while 611 genes were negatively correlated with it (**Figure 1B**). In addition, tumor purity did have influence on the prognosis of DLBCL patients, which showed that patients with high tumor purity had lower OS rate than those with low tumor purity (**Figure 1C,** $p = 0.025$).

Next, we performed GO and KEGG enrichment analyses to explore the functions and signaling pathways in which these TPGs were involved. It turned out that the TPGs were mainly associated with extracellular matrix organization and immune response (**Figure 1C, 1D**). Not only did the enrichment results confirm that these genes were reliable to be related with the tumor purity, but it also laid solid foundations for the sequent analyses.

A prognostic model was constructed with three TPGs

With the 642 TPGs, we exerted PPI analysis to investigate their interaction and the hub genes (**Figure 2A**). The TPGs who had five or more interactive genes were shown in **Figure 2B,** and defined as hub genes. Then, we performed univariate Cox regression analysis to figure out the

TPGs that were associated with the prognosis of DLBCL patients, and 103 genes were identified (**Figure 2C**). Interestingly, most of the TPGs were correlated with good outcome (with HR < 1), and only six genes were associated with poor outcome (with HR > 1). Through conducting intersection analysis, we found nine genes (LUM, VCAN, YAP1, COL5A2, SDC2, TWIST1, CD3G, C1QB and C3) were intersection genes, indicating that they had an active effect in modulating the tumor purity, as well as influencing the prognosis of DLBCL patients.

After ascertaining the key genes, we tried to construct a prognostic model with them. The model was constructed by Cox regression, and the three selected genes (VCAN, CD3G, C1QB) and their parameters like coefficient, HR and 95%CI of HR, were shown in **Figure 3A**. It showed that VCAN and CD3G were correlated with good prognosis and C1QB was correlated with poor prognosis. All patients were divided into high and low-risk group according to the median value of riskScore (**Figure 3B**). As expected, the high-risk group has worse prognosis than the low-risk group (**Figure 3C**). In addition, the three genes were differentially expressed between high and low-risk group, with VCAN and CD3G showing high expression level in low-risk group, and C1QB showing high expression level in high-risk group, which was consistent with the coefficient (**Figure 3D**). To appraise the efficacy of these prognostic model, we conducted survival analysis and ROC analysis. High-risk group had lower OS rate than low-risk group (**Figure 3E**, $p < 0.001$), and the areas under curve (AUC) for 1-year, 3-year and 5-year ROC were 0.73, 0.77 and 0.77 respectively (**Figure 3F**). Just similar to tumor purity, the high riskScore indicated bad outcome, which was consistent with the positive correlation between tumor purity and riskScore (**Figure 3G**). This TPGs signature prognostic model manifested satisfying prognostic efficacy.

When we applied this model to GSE32918 dataset, it still did excellently and the results were in accordance with that in GSE53786 dataset (**Figure 4A, 4B; Supplementary Figure 2**). Next, we analyzed the relationship between riskScore and some clinicopathological characteristics provided in GSE53786 dataset. The results showed that high-risk group had more ABC type DLBCL, while low-risk group had more the GCB type DLBCL (**Figure 4C**). Besides, high-risk group displayed higher lactic dehydrogenase (LDH) ratio (**Figure 4F**). However, the Eastern Cooperative Oncology Group (ECOG) performance and stage was not associated with the riskScore (**Figure 4D, 4E**). Still, the high-risk group has more Stage III and Stage IV patients, but less Stage I patients than low-risk group. Finally, we employed the univariate and multivariate analysis to

explore whether riskScore was an independent prognostic factor for DLBCL patients. As expected, the riskScore was associated with the poor prognosis (**Figure 4G**, $p < 0.001$, HR = 1.545, 95%CI 1.284–1.861) and was an independent prognostic factor (**Figure 4H**, $p = 0.002$, HR = 1.474, 95%CI 1.156–1.879).

The prognostic value of VCAN, CD3G and C1QB were validated by IHC assay

With the purpose of the further validation of the prognostic value of VCAN, CD3G and C1QB, we detected the expression of these genes in CHCAMS cohort by IHC. For VCAN, the patients were divided into high and low group according to the cut-off of the H-score (275.42) provided by X-tile. The survival analysis showed that patients with high expression of VCAN had higher OS rate (**Figure 5A**, $p = 0.003$). For CD3G, previous study revealed that it was a component of T cell receptor complex, for which it could be regarded as a marker of T cells[17]. Therefore, we assessed the expression level of CD3G by counting the CD3G+ T cells ratio, and divided patients by the cut-off (2.5%). The survival analysis revealed that patients with high CD3G+ T cells infiltration showed favorable prognosis (**Figure 5B**, $p < 0.001$). For C1QB, the patients in high expression group (cut-off = 82.41) showed adverse prognosis (**Figure 5C**, $p = 0.015$). Although the detection of protein level was not convenient to build a prognostic model for the difference of assessment methods and the lake of coefficient, these results were in accordance with those of GEO datasets, which successfully proved the prognostic value of VCAN, CD3G and C1QB.

Given that these genes could potentially influence the tumor purity of DLBCL, we then analyzed the relationship between them and CD68+ macrophages, CD4+ T cells and CD8+ T cells. As was shown in **Supplementary figure 3A**, CD68+ macrophages [(17.75±1.05) %] account for more ratio than CD4+ T cells [(0.68±0.20) %] and CD8+ T cells [(6.69±0.56) %] ($p < 0.001$, Kruskal-Wallis Test and Dunn's Test). In the survival analysis of these three types of immune cells, we found that CD68+ macrophages, CD8+ T cells and CD4+ T cells were associated with better prognosis (**Figure 5D–F**, $p = 0.029$, $p = 0.002$, $p = 0.053$). And XCELL and QUANTISEQ algorithm revealed that M1 macrophages accounted for more proportion than M2 macrophages in GSE53786 and GSE32918 (**Supplementary figure 3L–O**), which was confirmed in the CHCAMS cohort (**Supplementary figure 3P**). Besides, the ratio of CD3G+ T cells was positively correlated with that of CD68+ macrophages, CD8+ T cells and CD4+ T cells, C1QB expression

level was positively correlated with CD8+ T cells, and VCAN expression level was positively correlated with CD8+ T cells ratio (**Figure 5G)**. GSEA analysis based on the differentially expressed genes between high-risk and low-risk group in the GEO datasets above revealed that the cellular adhesion, extracellular structures and immune-related processes could result in the different outcome (**Figure 5H–I**)

In addition to the above analyses, we also explored the relationship between these three genes and location of DLBCL. It turned out that CD3G+ T cells ratio was higher in DLBCL originated from groin and testis, and VCAN featured higher expression in lymph node originated DLBCL (**Figure 5J–K,** $p < 0.05$, $p < 0.01$**, Supplementary figure 3B−K**).

These results showed that VCAN, CD3G and C1QB played important roles in the microenvironment of DLBCL, possibly regulating the immune infiltration via modulating the extracellular organization and cellular interaction.

The TPGs signature model could also predict the drug sensitivity of DLBCL patients

In order to learn about the ability of the previously mentioned model to predict drug sensitivity, we performed the prediction with "oncoPredict" package in R. Fifteen drugs **(Supplementary table 2)** included in the GDSC and used in clinical practice or under clinical trials (searched on Pubmed) were enrolled in this prediction analysis.

As is shown in **Figure 6A** (prediction of GSE53786), patients in high-risk group could be sensitive to Carmustine, Cytarabine, Oxaliplatin, Vincristine, Vorinostat, and Bortezomib, but no drug could work better in low-risk group. And in GSE32918 (**Figure 6B**), Carmustine, Cytarabine, Oxaliplatin, Vorinostat, Afuresertib, Bortezomib, Ibrutinib and Tamoxifen could work better in high-risk group, and Vincristine (sensitivity score: low-risk vs high-risk = 0.219±0.026 vs 0.223±0.031) could work better in low-risk group. The discrepancy between the prediction in two datasets might be due to the samples and sequencing platforms. However, the intersection analysis of the drugs to which the high-risk patients in both datasets could be sensitive revealed that Carmustine, Cytarabine, Oxaliplatin, Vorinostat and Bortezomib could be reliable candidates for treating high-risk patients based on the three TPGs signature prognostic model (**Supplementary table 2**).

Discussion

In this study, bioinformatics techniques were employed to identify three genes (VCAN, CD3G, C1QB) that exhibit associations with prognosis in both immune and stromal environments, thereby revealing their relationship with the prognosis of DLBCL patients. The findings indicate that higher expression of VCAN, increased infiltration of CD3G+ T cells, and decreased expression of C1QB are correlated with favorable prognostic outcomes. Conversely, a lower infiltration of CD68+ macrophages and lower infiltration of CD8+ T cells are associated with poorer prognosis. Furthermore, we investigated the relationship between risk genes related to tumor purity and treatment sensitivity and established a list of possible drugs that might be helpful for enhancing outcomes.

Previous studies have extensively investigated the VCAN gene in relation to tumorigenesis and metastasis[18]. VCAN, also known as versican, is a crucial component of extracellular matrix[19], and exists in several isoforms[20]. Research has shown that VCAN plays a multifaceted role in TME depending on the cell type expressing it. When expressed by myeloid cells, VCAN induces an anti-inflammatory and immunosuppressive microenvironment. Conversely, its expression by stromal cells typically leads to a pro-inflammatory response[21]. In gastric cancer, high expression of VCAN has been associated with increased infiltration of fibroblasts, significant enrichment of stromal-associated signaling pathways and poor prognosis[22]. In hepatocellular carcinoma, VCAN exhibits a strong association with immune checkpoint gene expression[23]. Despite these findings in other tumor types, the role of VCAN in DLBCL has not been explored yet. Our study reveals that high expression of VCAN is actually associated with a more favorable prognosis. This suggests that VCAN may have different functions in different tumor types. One possible mechanism through which VCAN influences prognosis is that VCAN overexpression in DLBCL may also impact tumor cell proliferation. A study has shown that overexpression of VCAN V1 has an inhibitory effect on cell proliferation, partly due to its promotion of activation-induced cell death in lymphoid cell lines[20]. Hence, the high expression of VCAN in DLBCL could impact not only the TME but also tumor cell proliferation, suggesting a potential mechanism for the observed preferable prognosis.

C1q is synthesized in the tumor microenvironment and functions as an extracellular matrix protein, and C1QB is a component of C1q[24]. Previous studies have provided insights into the diverse

roles of C1q in cancer progression. However, the majority of these results, as observed in non-small cell lung carcinoma and gastric cancer, indicate that high C1q expression in TME is associated with a poor prognosis [25-27]. Additionally, C1QB has been found to exert an impact on the TME and is positively associated with infiltration levels of CD8+ T cell, as well as with M1 and M2 macrophages in osteosarcoma[28]. Moreover, C1QB expression shows a positive correlation with predictive biomarkers for immunotherapy, such as PD-L1 expression and CD8+ T cell infiltration[27]. Furthermore, in malignant melanoma, C1QB promotes proliferation, migration and invasion, while inhibiting cell apoptosis[29], and the high-expression group exhibits significant enrichment of genes related to immune and apoptosis[24]. In our study, we found that high expression of C1QB in DLBCL was associated with a worse prognosis and positively correlated with CD8+ T cells infiltration. Based on these findings, we propose that C1QB in DLBCL might share similarities with its functions in other tumor types, particularly regarding the promotion of recruitment and subsequent deactivation of CD8+ T cells within the TME through the induction of immune checkpoint effects. These results shed light on the intricate role of C1QB in TME and its potential significance as a prognostic marker in DLBCL.

CD3G is a member of the TCR/CD3 complex primarily expressed in lymphocytes subgroups. It plays a crucial role in initiating the activation of T cells[17]. It is also involved in coupling antigen recognition[30]. It is reported to associate with long-term OS and good prognosis in breast invasive carcinoma[31] as well as in head and neck squamous cell carcinoma[32]. However, its role in DLBCL has not been fully explored. In our study, we revealed that high infiltration of CD3G+ T cells is correlated with good prognosis. The infiltration of CD3G+ T cells was found to be positively related to the infiltration of CD8+, CD4+ and CD68+ cells. This indicates that CD3G+ T cells in DLBCL may enhance the tumor antigen recognition process and stimulate the infiltration of immune cells, leading to an increased abundance of immune cell infiltration in the TME. The presence of CD3G+ T cells in the TME may contribute to a favorable prognosis by facilitating the activation of immune responses against tumor cells.

Macrophages play a crucial role in TME, and CD68 is a surface marker specific to macrophages. Macrophages can be roughly classified into two types based on their functional features: M1 or M2. M1 macrophages exert anti-tumor effects, whereas M2 macrophages promote tumor growth and progression in TME[33]. A previous study found that low infiltration of CD68+ macrophages

was associated with an inferior prognosis[34]. Similarly, our study has yielded similar results, revealing a noteworthy correlation between a high proportion of CD68+ macrophages in the TME and improved prognosis. Additionally, by analyzing the datasets, we observed a higher proportion of M1 macrophages infiltration compared to M2 macrophages. This suggests that, within our DLBCL cohort, these macrophages may also exhibit the M1 phenotype and consequently play a protective role against tumor progression.

CD8 is widely recognized as a marker of CD8+ T cells, also known as cytotoxic T cells[35]. These cells are crucial for the immune response against tumors. However, in DLBCL, CD8+ T cells exhibits elevated levels of inhibitory molecules on their surface, such as PD-1, PD-L1, TIM3. High expression of TIM3, an inhibitory immune checkpoint receptor, on CD8+ T cells has been associated with tumor progression and poor outcomes[36, 37]. These inhibitory molecules may impair the function of CD8+ T cells and hinder their anti-tumor activity. Surprisingly, our study demonstrates a correlation between the infiltration of CD8+ T cells and favorable prognosis in DLBCL. Here, we propose a hypothesis that in our study, the observed high expression of VCAN might create a suppressive environment for PD-1+ CD8+ T cells[21, 38]. Intriguingly, our study revealed a statistically significant correlation between VCAN expression, C1QB expression and CD8+ T cell infiltration. VCAN has the potential to modulate immune infiltration by reducing the immunosuppressive phenotype of immune cells[39], thus enabling a more efficient anti-tumor response. This aspect is still worth of consideration.

Taken together, our findings underscore the significant roles of VCAN, CD3G, C1QB, which influence both the TME and the behavior of tumor cells. The interaction between each component and the TME is rather complicated. To fully comprehend the underlying mechanisms and identify potential therapeutic targets in DLBCL, further investigation is required.

However, this study still has several limitations that should be addressed. Firstly, the patients included in this study were form a single center, which may introduce biases into the results. Although we made efforts to minimize these biases, it is inevitable that some may persist. Secondly, we hypothesized that VCAN, CD3G and C1QB could serve as continuous prognostic parameters, thereby eliminating the need for a cut-off. However, the methodology used in this study, which utilized IHC staining to assess the protein expression levels, may have potential limitations. While IHC is a widely used technique, additional validation is needed to confirm the

prognostic value of VCAN, CD3G and C1QB in DLBCL. Furthermore, due to the potential variability in interpreting IHC results across different centers, a standardized coefficient and formula have not been established to calculate the final prognostic index for patients with DLBCL. Developing a standardized approach would be beneficial in ensuring consistent and accurate interpretation of IHC results. To address these limitations and expand upon our findings, future studies should strive to incorporate a diverse range of patients from multiple centers. Additionally, it is crucial to employ rigorous experimental techniques to authenticate the prognostic significance of VCAN, CD3G, and C1QB in DLBCL.

Author contributions

Conceptualization: Wenting Huang, Ning Huang, Zhenbang Ye, Data Curation: Zhenbang Ye, Ning Huang, Formal Analysis: Zhenbang Ye, Ning Huang, Liming, Wang, Investigation: Zhenbang Ye, Methodology: Zhenbang Ye, Ning Huang, Wenting Huang, Liming Wang, Resources: Wenting Huang, Yongliang Fu, Liming Wang, Supervision: Wenting Huang, Liming Wang, Visualization: Zhenbang Ye, Ning Huang, Validation: Ning Huang, Writing-Original Draft: Zhenbang Ye, Ning Huang, Writing-Review & Editing: Wengting Huang, Yongliang Fu, Rongle Tian.

Funding

This work is supported by Shenzhen High-level Hospital Construction Fund and CAMS Innovation Fund for Medical Sciences (CIFMS) (2022-I2M-C&T-B-062). The funders had no role in study design, data collection and analysis, interpretation of data, or preparation of the manuscript.

Acknowledgement

Not applicable.

Conflicts of interest

The authors made no disclosures.

Figure 1 TPGs were screened out with GSE53786 dataset.

(A) The range (17.2 –67.4%) of tumor purity of samples in GSE53786. (B) The heatmap showing genes defined as TPGs. (C) The K–M curve showed high tumor purity was correlated with poor prognosis in DLBCL patients in GSE53678 dataset (Patients were divided into two groups according to the best-cutoff provided by "survminer" package in R). (D) GO analysis of TPGs. (E) KEGG analysis of TPGs. TPGs, tumor purity-related genes; K–M, Kaplan-Meier; DLBCL, diffuse large B cell lymphoma; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

Figure 2 The key gene candidates used for constructing prognostic model were selected.

(A) The PPI network of TPGs (orange nodes representing genes positively correlated with tumor purity, and green nodes representing genes negatively correlated with tumor purity). (B) The barplot showing hub genes with five or more interactive genes. (C) The forest plot showing prognostic TPGs of DLBCL patients in GSE53786. (D) The venn plot showing intersection genes of PPI hub gene and prognostic TPGs. PPI, protein-protein interaction.

Figure 3 TPGs signature prognostic model was constructed.

(A) Three genes enrolled in the prognostic model. (B) The patients in GSE53786 dataset were divided into high and low-risk group according to the median riskScore based on the prognostic model. (C) High-risk group had worse prognosis than low-risk group. (D) The heapmap showing expression discrepancy of the three genes. (E) Survival analysis revealed that high-risk group had poor prognosis in GSE53786 dataset. (F) The ROC curve showed that the prognostic model performed well in predicting 1-year, 3-year and 5-year prognosis in GSE53786 dataset. (G) Tumor purity was positively correlated with riskScore in GSE53786 dataset. ROC, receiver operating characteristic.

Figure 4 The riskScore of three TPGs signature prognostic model was an independent prognostic factor in DLBCL patients.

(A) Survival analysis results in GSE32918 dataset was consistent to that of GSE53786 dataset. (B) The prognostic model also did well in GSE32918 dataset. (C) High-risk group in GSE53786 dataset contained more ABC type DLBCL, while low-risk group contained more GCB type DLBCL. (D) The ECOG performance of two groups (GSE53786 dataset) showed no statistical difference. (E) More patients in high-risk group were at Stage III or Stage IV,

and less patients were at Stage I, compared with low-risk group, although no statistical significance was shown (GSE53786 dataset). (F) High-risk group had higher LDH ratio (GSE53786 dataset). (G) The riskScore was associated with poor prognosis of DLBCL patients in GSE53786 dataset. (H) The riskSocre was an independent prognostic factor for DLBCL patient. * $p < 0.05$, ** $p < 0.01$, ns, not significant. ABC, activated B cell; GCB, germinal center B cell; ECOG, Eastern Cooperative Oncology Group; LDH, lactic dehydrogenase.

Figure 5 The analysis of CHCAMS cohort.

(A) The representative image of VCAN staining of high and low expression groups and the survival analysis based on VCAN expression. (B) The representative image of CD3G staining of high and low CD3G+ T cells ratio groups and the survival analysis based on CD3G+ T cells ratio. (C) The representative image of C1QB staining of high and low expression group and the survival analysis based on C1QB expression. (D) The representative image of CD68 staining of high and low CD68+ macrophages ratio groups and the survival analysis based on CD68+ macrophages ratio. (E) The representative image of CD8 staining of high and low CD8+ T cells ratio groups and the survival analysis based on CD8+ T cells ratio. (F) The representative image of CD4 staining of high and low CD4+ T cells ratio groups and the survival analysis based on CD4+ T cells ratio. (G) The correlation between VCAN, CD3G+ T cells ration, C1QB and CD68+ macrophages, CD8+ T cells and CD4+ T cells ratio. "×" means no statistical significance. (H–I) GSEA analysis based on the differentially expressed genes between high-risk and low-risk group in GSE53786 and GSE32918. (J) The CD3G+ T cells infiltration varied from colon to testis originating DLBCL in male. (K) The VCAN expression level was different between intra- and extra-lymph node DLBCL.

Figure 6 Drug sensitivity prediction revealed therapeutic candidates for high-risk group.

(A) Drug sensitivity prediction results with statistical significance in GSE53786 dataset. (B) Drug sensitivity prediction results with statistical significance in GSE53786 dataset. According to "oncoPredict" algorithm, sensitivity score indicates IC50 of drugs, with higher sensitivity score indicating lower sensitivity.

Supplementary Figure 1 Flow chart and study design of this research.

Supplementary Figure 2 The correlation between tumor purity and prognosis and

riskScore in GSE32918 dataset.

(A) High ESTIMATE-Score group had better prognosis. (B) ESTIMATE-Score was negatively associated with riskScore. Note: ESTIMATE-Score manifests the tumor purity, with high ESTIMATE-Score indicating low tumor purity and high immune and stromal components. Because the GSE32918 datasets were sequenced with illumina platform, the ESTIMATE algorithm only provided ESTIMATE Score to represent the tumor purity.

Supplementary Figure 3 The immunoenvironment analysis and clinicopathological analysis of CHCAMS cohort and GEO datasets.

(A) The ratio of CD68+ macrophages, CD8+ T cells and CD4+ T cells in CHCAMS cohort. (B–D) The correlation between VCAN, CD3G+ T cells, C1QB and types of DLBCL in CHCAMS cohort. (E–K) The correlation between VCAN, CD3G+ T cells, C1QB and DCBCL origination in CHCAMS cohort. “Location to LN” means the relative location of the tumor to lymph node, that is, intra-lymph node or extra-lymph node. (L–M) The predicted M1 and M2 macrophages proportion in GSE53786 dataset. (N–O) The predicted M1 and M2 macrophages proportion in GSE32918 dataset. (P) The validation of M1 (CD32+ cells) and M2(CD206+ cells) macrophages infiltration in DLBCL in CHCAMS cohort. **$p < 0.01$, ***$p < 0.001$.

References:

1. Alaggio, R., et al., *The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms.* Leukemia, 2022. **36**(7): p. 1720-1748.
2. Sehn, L.H. and G. Salles, *Diffuse Large B-Cell Lymphoma.* N Engl J Med, 2021. **384**(9): p. 842-858.
3. Hans, C.P., et al., *Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray.* Blood, 2004. **103**(1): p. 275-82.
4. Autio, M., et al., *Immune cell constitution in the tumor microenvironment predicts the outcome in diffuse large B-cell lymphoma.* Haematologica, 2021. **106**(3): p. 718-729.
5. Joyce, J.A. and J.W. Pollard, *Microenvironmental regulation of metastasis.* Nat Rev Cancer, 2009. **9**(4): p. 239-52.
6. Steen, C.B., et al., *The landscape of tumor cell states and ecosystems in diffuse large B cell lymphoma.* Cancer Cell, 2021. **39**(10): p. 1422-1437.e10.
7. Ciavarella, S., et al., *Dissection of DLBCL microenvironment provides a gene expression-based predictor of survival applicable to formalin-fixed paraffin-embedded tissue.* Ann Oncol, 2018. **29**(12): p. 2363-2370.
8. Ennishi, D., et al., *Toward a New Molecular Taxonomy of Diffuse Large B-cell Lymphoma.* Cancer Discov, 2020. **10**(9): p. 1267-1281.
9. Miyawaki, K., et al., *A germinal center-associated microenvironmental signature reflects malignant phenotype and outcome of DLBCL.* Blood Adv, 2022. **6**(7): p.

2388-2402.

10. Schmitz, R., et al., *Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma.* N Engl J Med, 2018. **378**(15): p. 1396-1407.

11. Lou, X., et al., *CCL8 as a promising prognostic factor in diffuse large B-cell lymphoma via M2 macrophage interactions: A bioinformatic analysis of the tumor microenvironment.* Front Immunol, 2022. **13**: p. 950213.

12. Gong, Z., J. Zhang, and W. Guo, *Tumor purity as a prognosis and immunotherapy relevant feature in gastric cancer.* Cancer Med, 2020. **9**(23): p. 9052-9063.

13. Lou, S., et al., *Comprehensive Characterization of Tumor Purity and Its Clinical Implications in Gastric Cancer.* Front Cell Dev Biol, 2021. **9**: p. 782529.

14. Zhao, Y., et al., *Tumor purity-associated genes influence hepatocellular carcinoma prognosis and tumor microenvironment.* Front Oncol, 2023. **13**: p. 1197898.

15. Zhang, C., et al., *Tumor Purity as an Underlying Key Factor in Glioma.* Clin Cancer Res, 2017. **23**(20): p. 6279-6291.

16. Yoshihara, K., et al., *Inferring tumour purity and stromal and immune cell admixture from expression data.* Nat Commun, 2013. **4**: p. 2612.

17. Wang, M., D. Windgassen, and E.T. Papoutsakis, *Comparative analysis of transcriptional profiling of CD3+, CD4+ and CD8+ T cells identifies novel immune response players in T-cell activation.* BMC Genomics, 2008. **9**: p. 225.

18. Baghy, K., et al., *Proteoglycans in liver cancer.* World J Gastroenterol, 2016. **22**(1): p. 379-93.

19. Wight, T.N., *Provisional matrix: A role for versican and hyaluronan.* Matrix Biol, 2017.

60-61: p. 38-56.

20. Fujii, K., et al., *Versican upregulation in Sézary cells alters growth, motility and resistance to chemotherapy.* Leukemia, 2015. **29**(10): p. 2024-32.

21. Wight, T.N., et al., *Versican-A Critical Extracellular Matrix Regulator of Immunity and Inflammation.* Front Immunol, 2020. **11**: p. 512.

22. Song, J., et al., *Versican enrichment predicts poor prognosis and response to adjuvant therapy and immunotherapy in gastric cancer.* Front Immunol, 2022. **13**: p. 960570.

23. Wang, M.Q., et al., *VCAN, expressed highly in hepatitis B virus-induced hepatocellular carcinoma, is a potential biomarker for immune checkpoint inhibitors.* World J Gastrointest Oncol, 2022. **14**(10): p. 1933-1948.

24. Yang, H., et al., *Prognostic and immune-related value of complement C1Q (C1QA, C1QB, and C1QC) in skin cutaneous melanoma.* Front Genet, 2022. **13**: p. 940306.

25. Li, Z., et al., *Differentiation-related genes in tumor-associated macrophages as potential prognostic biomarkers in non-small cell lung cancer.* Front Immunol, 2023. **14**: p. 1123840.

26. Mangogna, A., et al., *Prognostic Implications of the Complement Protein C1q in Gliomas.* Front Immunol, 2019. **10**: p. 2366.

27. Jiang, J., et al., *Identification of TYROBP and C1QB as Two Novel Key Genes With Prognostic Value in Gastric Cancer by Network Analysis.* Front Oncol, 2020. **10**: p. 1765.

28. Chen, L.H., et al., *Complement C1q (C1qA, C1qB, and C1qC) May Be a Potential*

Prognostic Factor and an Index of Tumor Microenvironment Remodeling in Osteosarcoma. Front Oncol, 2021. **11**: p. 642144.

29. Zheng, Y., et al., *IRF4-activated TEX41 promotes the malignant behaviors of melanoma cells by targeting miR-103a-3p/C1QB axis.* BMC Cancer, 2021. **21**(1): p. 1339.

30. Chen, Z., et al., *A Machine Learning Model to Predict the Triple Negative Breast Cancer Immune Subtype.* Front Immunol, 2021. **12**: p. 749459.

31. Wang, Q., P. Li, and W. Wu, *A systematic analysis of immune genes and overall survival in cancer patients.* BMC Cancer, 2019. **19**(1): p. 1225.

32. Wang, J., et al., *Establishment and validation of immune microenvironmental gene signatures for predicting prognosis in patients with head and neck squamous cell carcinoma.* Int Immunopharmacol, 2021. **97**: p. 107817.

33. Zhang, Y., et al., *GM-CSF enhanced the effect of CHOP and R-CHOP on inhibiting diffuse large B-cell lymphoma progression via influencing the macrophage polarization.* Cancer Cell Int, 2021. **21**(1): p. 141.

34. Croci, G.A., et al., *SPARC-positive macrophages are the superior prognostic factor in the microenvironment of diffuse large B-cell lymphoma and independent of MYC rearrangement and double-/triple-hit status.* Ann Oncol, 2021. **32**(11): p. 1400-1409.

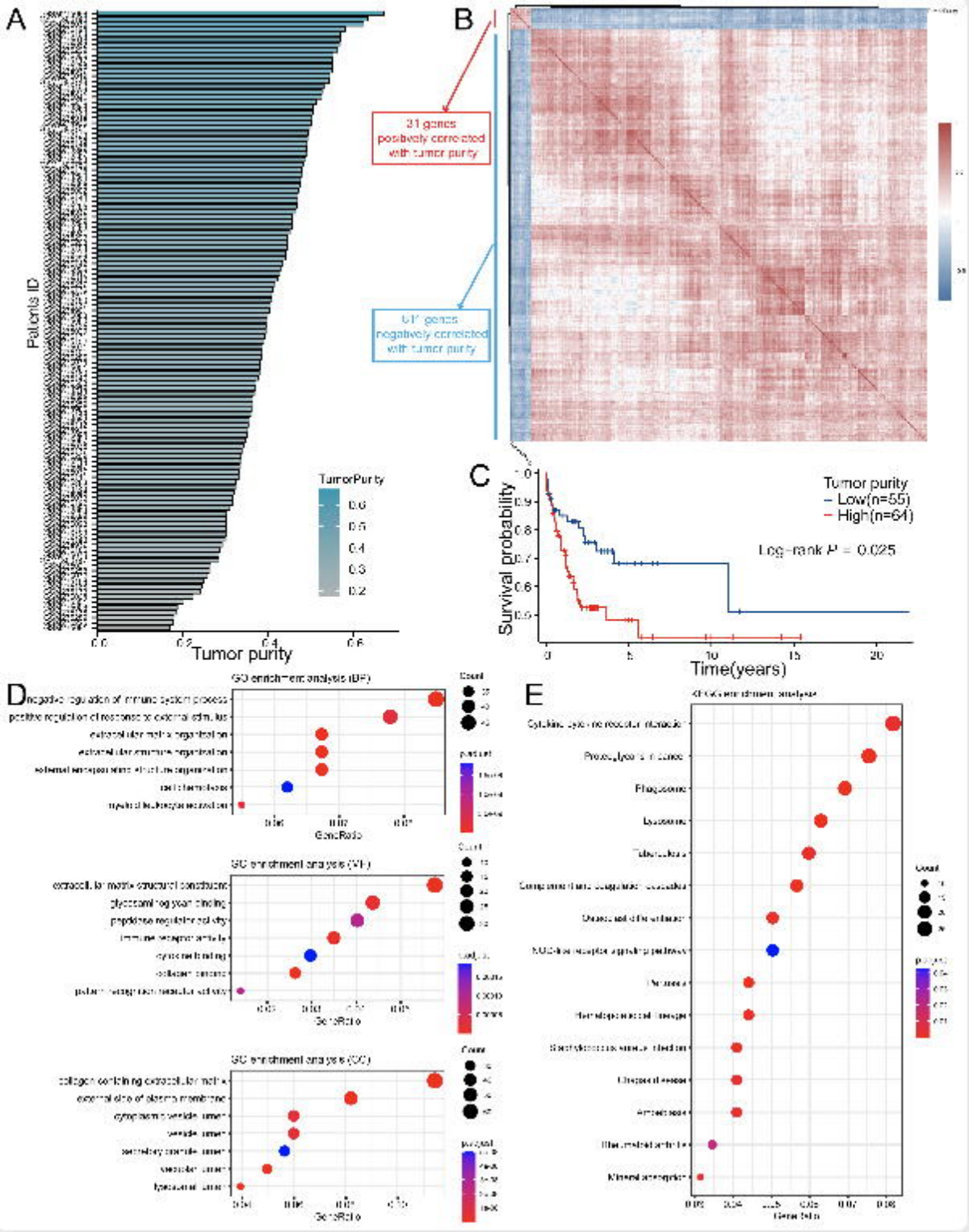
35. Farhood, B., M. Najafi, and K. Mortezaee, *CD8(+) cytotoxic T lymphocytes in cancer immunotherapy: A review.* J Cell Physiol, 2019. **234**(6): p. 8509-8521.

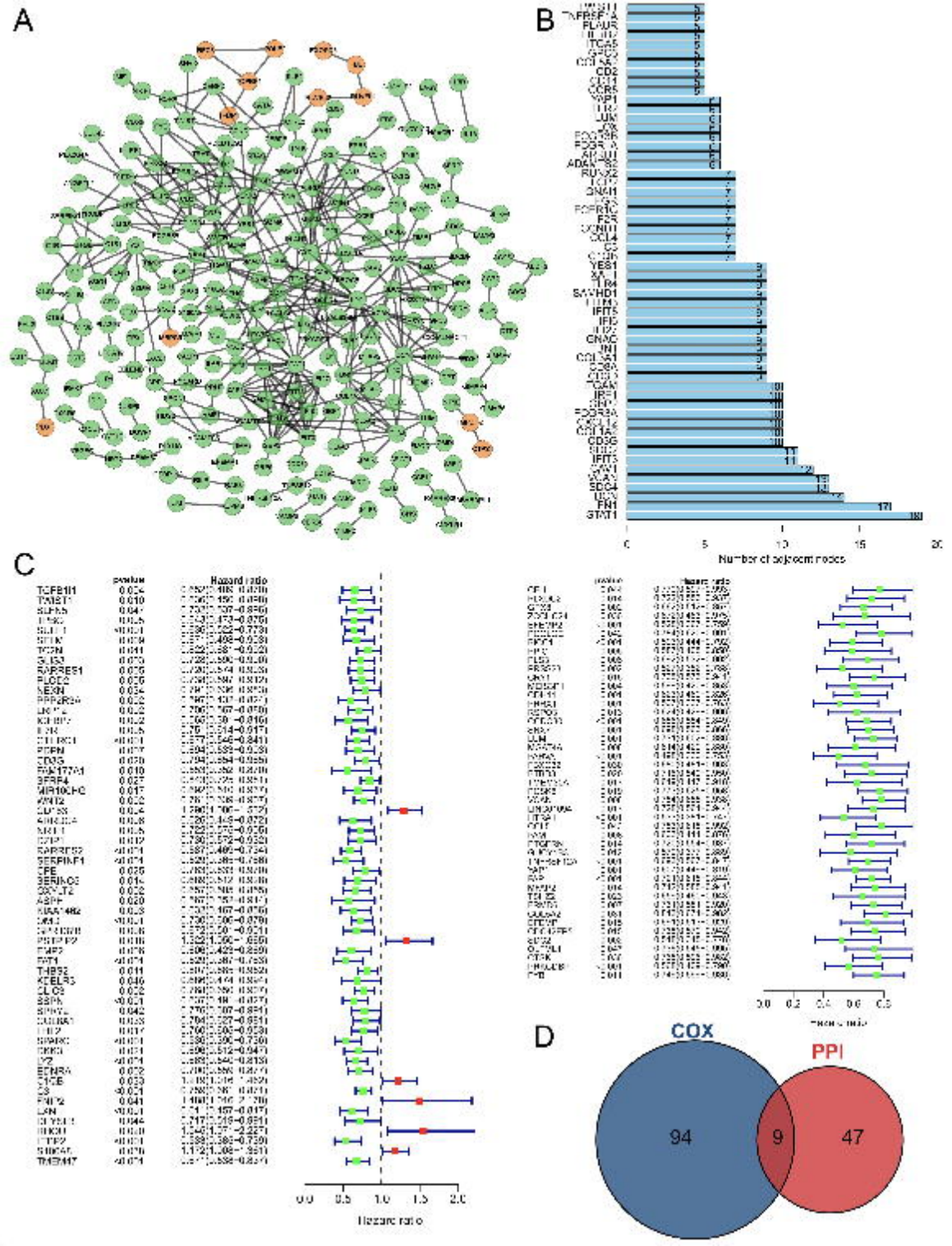
36. Roussel, M., et al., *Functional characterization of PD1+TIM3+ tumor-infiltrating T cells in DLBCL and effects of PD1 or TIM3 blockade.* Blood Adv, 2021. **5**(7): p. 1816-1829.

37. Xu-Monette, Z.Y., et al., *Immune Profiling and Quantitative Analysis Decipher the Clinical Role of Immune-Checkpoint Expression in the Tumor Immune Microenvironment of DLBCL.* Cancer Immunol Res, 2019. **7**(4): p. 644-657.

38. Hirani, P., et al., *Targeting Versican as a Potential Immunotherapeutic Strategy in the Treatment of Cancer.* Front Oncol, 2021. **11**: p. 712807.

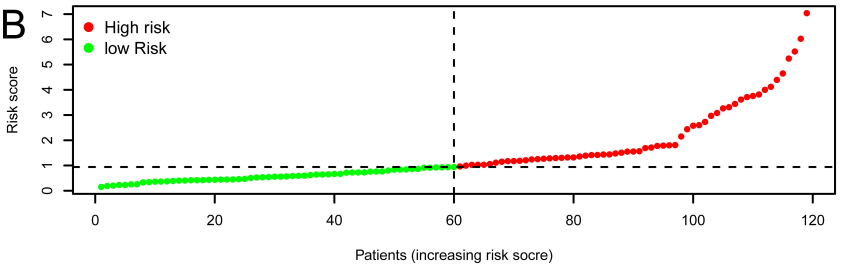
39. Huang, X.Y., et al., *Bioinformatics analysis of the prognosis and biological significance of VCAN in gastric cancer.* Immun Inflamm Dis, 2021. **9**(2): p. 547-559.

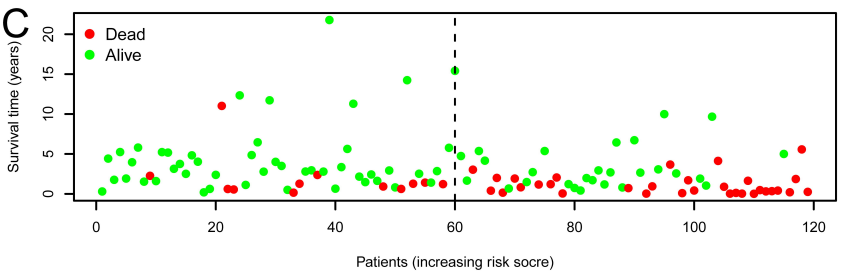

A
Patients ID
Tumor purity
TumorPurity
0.6
0.5
0.4
0.3
0.2
B
31 genes positively correlated with tumor purity
617 genes negatively correlated with tumor purity
C
Survival probability
Time(years)
Tumor purity
Low(n=55)
High(n=64)
Log-rank P = 0.025
D
GO enrichment analysis (BP)
negative regulation of immune system process
positive regulation of response to external stimulus
extracellular matrix organization
external encapsulating structure organization
cell chemotaxis
myeloid leukocyte migration
GeneRatio
Count
GO enrichment analysis (MF)
extracellular matrix structural constituent
glycosaminoglycan binding
peptidase regulator activity
immune receptor activity
cytokine binding
collagen binding
GeneRatio
GO enrichment analysis (CC)
collagen containing extracellular matrix
external side of plasma membrane
cytoplasmic vesicle lumen
vesicle lumen
secretory vesicle lumen
vesicular lumen
specific granule lumen
GeneRatio
E
Cytokine-cytokine receptor interaction
Phagosome
Lysosome
Tuberculosis
Complement and coagulation cascades
Osteoclast differentiation
NOD-like receptor signaling pathway
Influenza A
Herpes simplex virus 1 infection
Staphylococcus aureus infection
Measles
Amoebiasis
Rheumatoid arthritis
Malaria
Count
GeneRatio

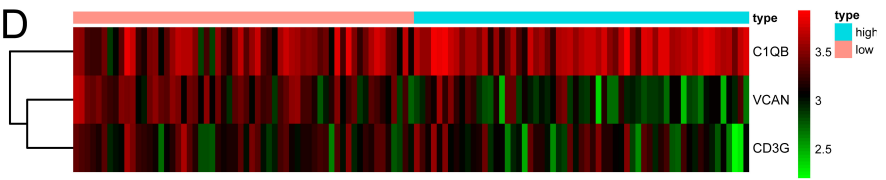

A
B
Number of adjacent nodes
C
Hazard ratio
D
COX
PPI
94
9
47

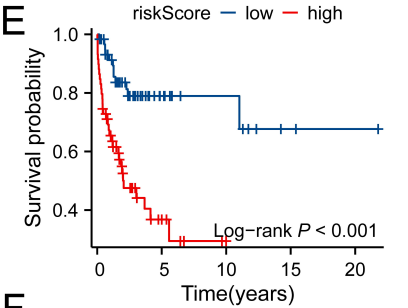
A

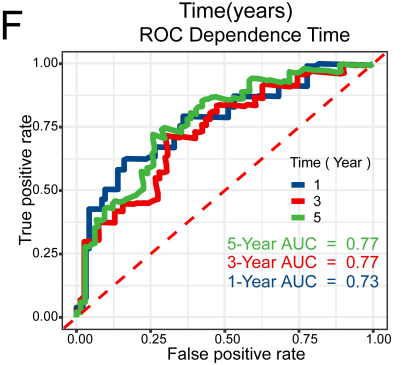
Gene	coef	HR(95%CI)	P
VCAN	-0.317	0.73(0.61–0.87)	<0.001
CD3G	-0.404	0.67(0.55–0.80)	<0.001
C1QB	0.383	1.47(1.20–1.79)	<0.001

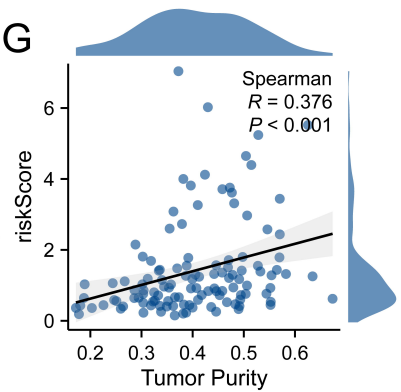


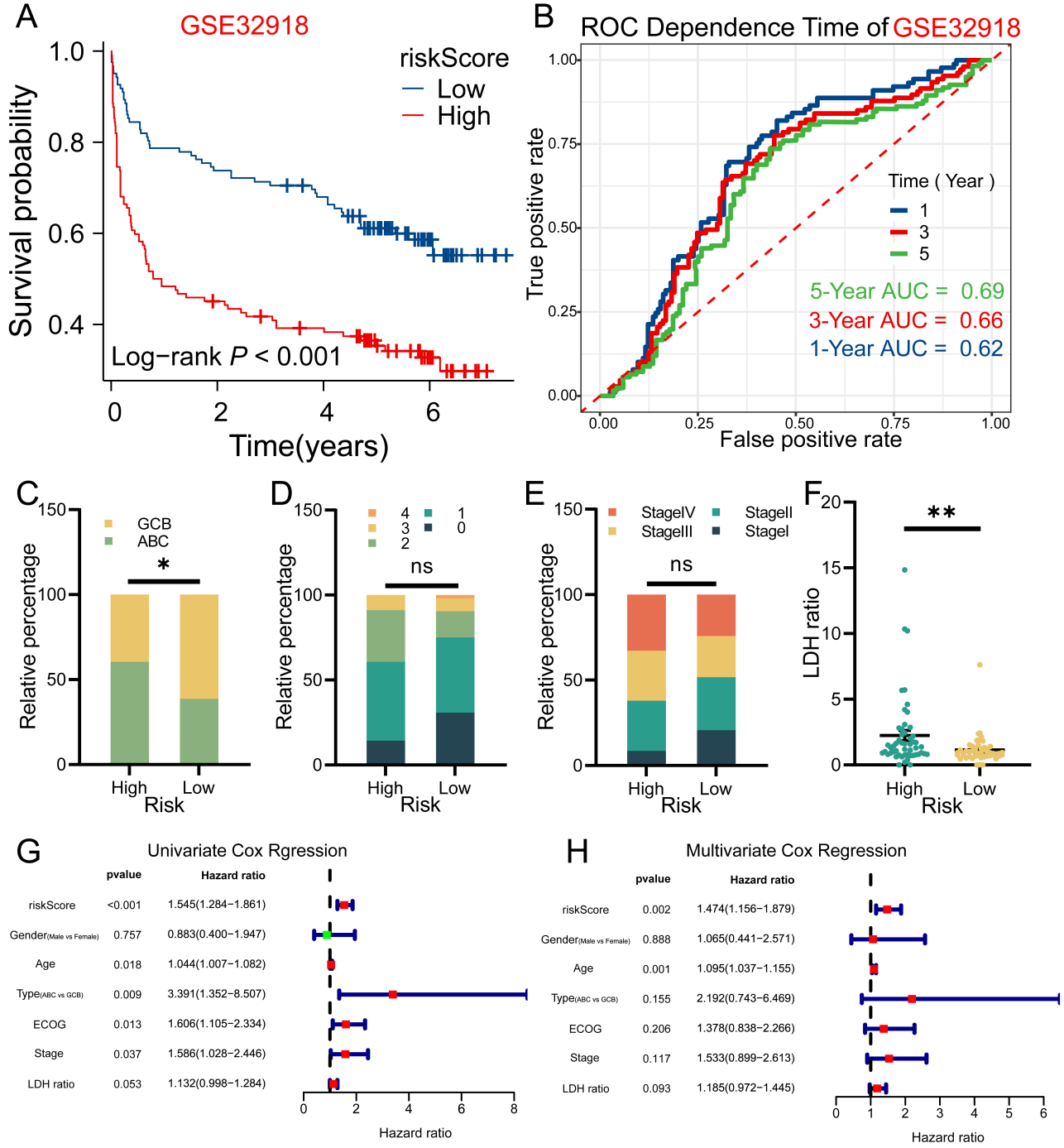


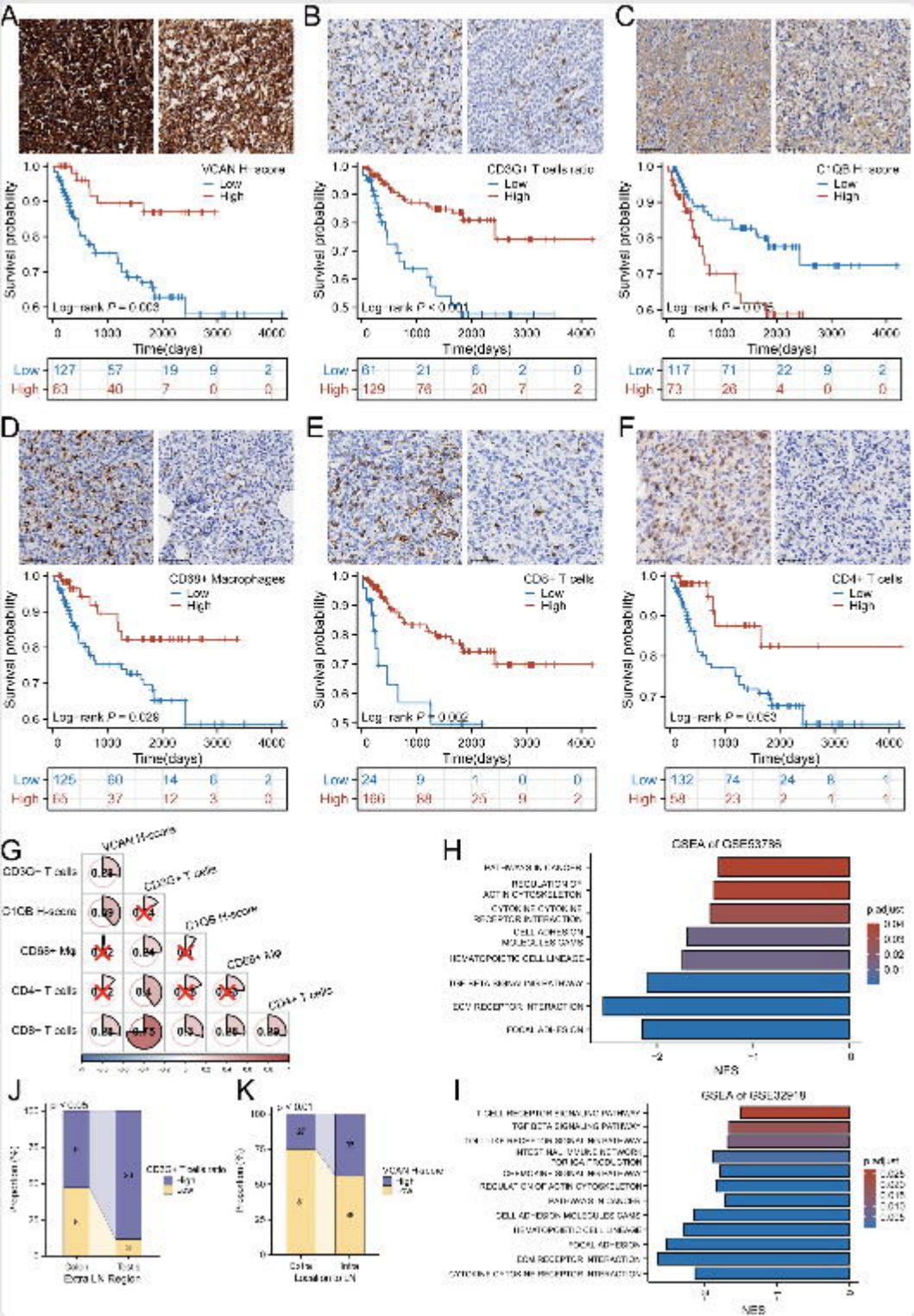


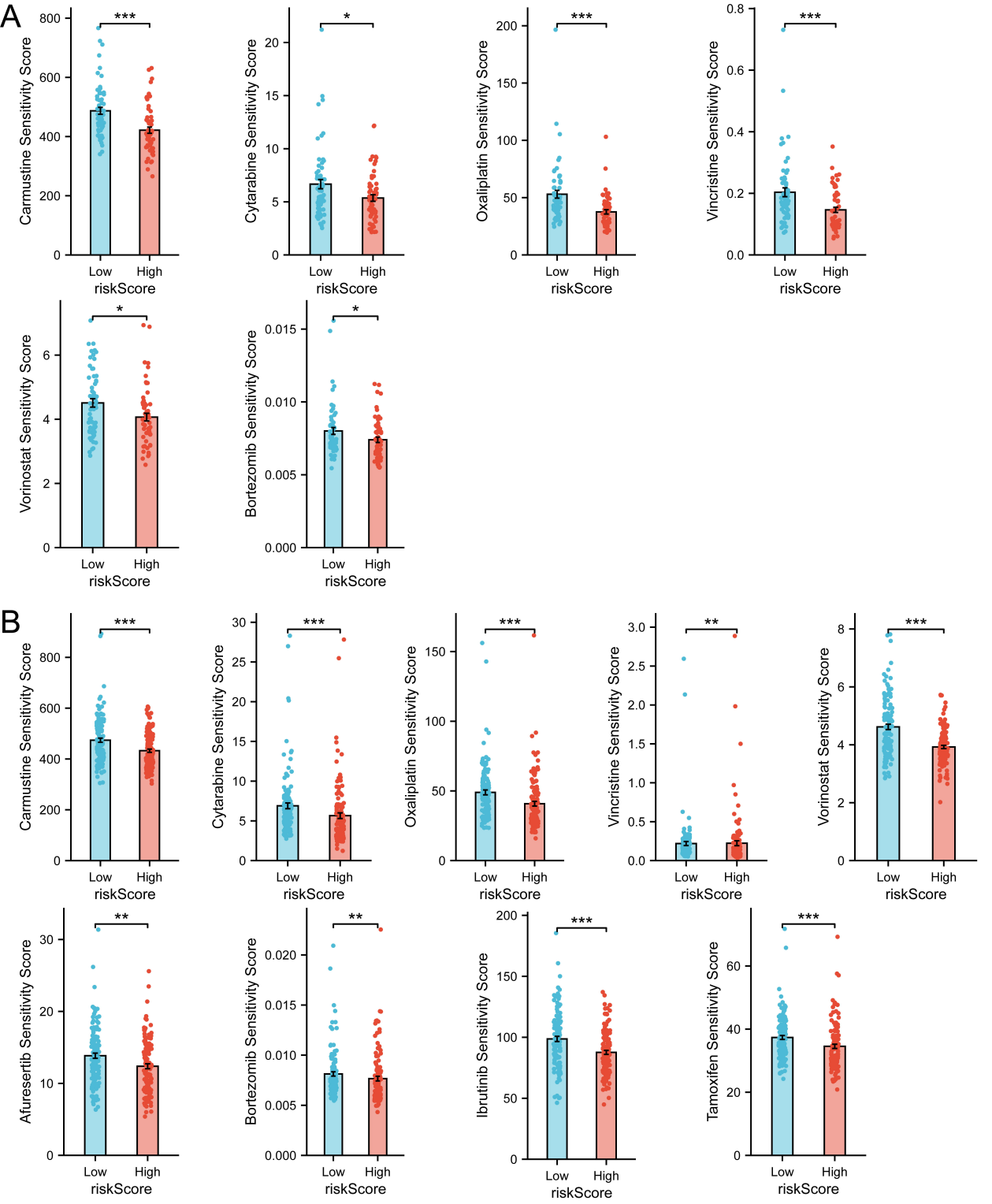







A
GSE32918
riskScore
Low
High
Survival probability
Log−rank P < 0.001
Time(years)
B
ROC Dependence Time of GSE32918
True positive rate
False positive rate
Time (Year)
1
3
5
5-Year AUC = 0.69
3-Year AUC = 0.66
1-Year AUC = 0.62
C
GCB
ABC
*
Relative percentage
High
Low
Risk
D
4
3
2
1
0
ns
Relative percentage
High
Low
Risk
E
StageIV
StageIII
StageII
StageI
ns
Relative percentage
High
Low
Risk
F
**
LDH ratio
High
Low
Risk
G
Univariate Cox Rgression
pvalue
Hazard ratio
riskScore <0.001 1.545(1.284−1.861)
Gender(Male vs Female) 0.757 0.883(0.400−1.947)
Age 0.018 1.044(1.007−1.082)
Type(ABC vs GCB) 0.009 3.391(1.352−8.507)
ECOG 0.013 1.606(1.105−2.334)
Stage 0.037 1.586(1.028−2.446)
LDH ratio 0.053 1.132(0.998−1.284)
Hazard ratio
H
Multivariate Cox Regression
pvalue
Hazard ratio
riskScore 0.002 1.474(1.156−1.879)
Gender(Male vs Female) 0.888 1.065(0.441−2.571)
Age 0.001 1.095(1.037−1.155)
Type(ABC vs GCB) 0.155 2.192(0.743−6.469)
ECOG 0.206 1.378(0.838−2.266)
Stage 0.117 1.533(0.899−2.613)
LDH ratio 0.093 1.185(0.972−1.445)
Hazard ratio


A
B
C
VCAN H-score
Low
High
Survival probability
Log-rank P = 0.003
Time(days)
Low 127 57 19 9 2
High 63 40 7 0 0
CD8G+ T cells ratio
Log-rank P < 0.001
Low 61 21 6 2 0
High 129 76 20 7 2
C1QB H-score
Log-rank P = 0.015
Low 117 71 22 9 2
High 73 26 4 0 0
D
E
F
CD8+ Macrophages
Log-rank P = 0.028
Low 125 80 14 6 2
High 65 37 12 3 0
CD8+ T cells
Log-rank P = 0.002
Low 24 9 1 0 0
High 166 88 25 9 2
CD4+ T cells
Log-rank P = 0.063
Low 132 74 24 8 1
High 58 23 2 1 1
G
VCAN H-score
CD8G+ T cells
C1QB H-score
CD8+ Ma
CD4+ T cells
CD8+ T cells
H
GSEA of GSE63788
p.adjust
NES
J
CD8G+ T cells ratio
High
Low
Proportion (%)
Extra LN
Intra LN
Extra LN Region
K
VCAN H-score
Location of LN
I
GSEA of GSE32918
NES


A
Carmustine Sensitivity Score
Cytarabine Sensitivity Score
Oxaliplatin Sensitivity Score
Vincristine Sensitivity Score
Vorinostat Sensitivity Score
Bortezomib Sensitivity Score
Low
High
riskScore
B
Carmustine Sensitivity Score
Cytarabine Sensitivity Score
Oxaliplatin Sensitivity Score
Vincristine Sensitivity Score
Vorinostat Sensitivity Score
Afuresertib Sensitivity Score
Bortezomib Sensitivity Score
Ibrutinib Sensitivity Score
Tamoxifen Sensitivity Score
Low
High
riskScore