

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

3 Zhenbang Ye<sup>1#</sup>, Ning Huang<sup>2#</sup>, Yongliang Fu<sup>1</sup>, Rongle Tian<sup>1</sup>, Liming Wang<sup>2\*</sup>, Wenting Huang<sup>1,3\*</sup>

4 <sup>1</sup>Department of Pathology, National Cancer Center/National Clinical Research Center for  
5 Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical  
6 College, Beijing, China;

7 <sup>2</sup>Department of Hepatobiliary Surgery, National Cancer Center/National Clinical Research Center  
8 for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical  
9 College, Beijing 100021, China

10 <sup>3</sup>Department of Pathology, National Cancer Center/National Clinical Research Center for  
11 Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking  
12 Union Medical College, Shenzhen, 518116, China.

13 #These authors contributed equally to this work.

14 \*Corresponding author

15 Wenting Huang, Ph.D.

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

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

23 E-mail: [huangwt@cicams.ac.cn](mailto:huangwt@cicams.ac.cn)

24

25 Liming Wang, M.D.

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

29 Email: [stewen\\_wang@sina.com](mailto:stewen_wang@sina.com)

30

31 **Abstract**

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

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

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

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

50

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

52

53 **Introduction**

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

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

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

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

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

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

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

98

## 99 **Materials and Methods**

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

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

108

### 109 **TPGs function analysis**

110 Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were  
111 executed to analyze the biological processes, cellular components, molecular functions and  
112 pathways related to the TPGs. The statistical significance was considered as  $p_{\text{adjust}} < 0.05$ .

113 The protein-protein interactions (PPI) analysis was utilized to investigate the interaction among  
114 TPGs, and those with interactive confidence greater than 0.90 on the STRING platform (version  
115 11.5) were selected to establish an interaction network with Cytoscape software (version 3.8.2).

116

### 117 **Prognostic model**

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



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

122

123 **Clinical specimens and follow-up**

124 190 patients from Cancer hospital Chinese Academy of Medical Sciences, the CHCAMS cohort,  
125 were enrolled in this study (**Supplementary table 1**). All patients received surgery or biopsy  
126 during September, 2010 and September, 2020, and then standard follow-ups were carried out until  
127 March, 2023. The overall survival (OS) was defined as the interval between the operation and  
128 death or the last follow-up. The specimens from the CHCAMS cohort were used for  
129 immunohistochemistry assay. The study was designed according to the Declaration of Helsinki  
130 and approved by the institutional ethics committee of Cancer Hospital Chinese Academy of  
131 Medical Sciences. Informed consent was taken from all the patients.

132

133 **IHC**

134 Paraffin embedded DLBCL tissues of CHCAMS Cohort were used for immunohistochemistry  
135 (IHC). After de-paraffinization and hydration, heat-induced method was performed for antigen  
136 retrieval. Primary antibody of VCAN (AB177480, 1:100, Abcam, USA), CD3G (AB134096,  
137 1:1000, Abcam, USA), C1QB (AB92508, 1:50, Abcam, USA), CD68 (303565, 1:1000, Abcam,  
138 USA), CD4 (ZM-0418, ZSGB-BIO, China), CD8 (ZA-0508, ZSGB-BIO, China), CD206(24595S,  
139 1:400, CST, USA), and CD32(15625-1-AP, 1:1000, proteintech, China) was incubated at 4°C  
140 overnight. Sections were washed with TBS-T buffer, and then incubated with secondary antibody,  
141 and finally stained with DAB. The quantitative analysis of the slices was conducted by  
142 QuPath-0.4.3. VCAN and C1QB were assessed by H-score, and CD3G, CD68, CD4, CD8, CD206,  
143 CD32 were assessed as the ratio of the corresponding positive cells among all cells.

144

145 **Drug sensitivity prediction**

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

150

151 **Statistical analysis**

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

160

161 **Results**

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

164 Based on GSE53786 dataset, we first assessed the tumor purity of DLBCL, which ranged from  
165 17.2% to 67.4% (**Figure 1A**). In order to screen out the TPGs, we then analyzed the correlation  
166 between genes expression and tumor purity. According to the thresholds mentioned above, 642  
167 genes were identified as TPGs, among which 31 genes were positively correlated with tumor  
168 purity, while 611 genes were negatively correlated with it (**Figure 1B**). In addition, tumor purity  
169 did have influence on the prognosis of DLBCL patients, which showed that patients with high  
170 tumor purity had lower OS rate than those with low tumor purity (**Figure 1C**,  $p = 0.025$ ).

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

176

177 **A prognostic model was constructed with three TPGs**

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

181 TPGs that were associated with the prognosis of DLBCL patients, and 103 genes were identified  
182 (**Figure 2C**). Interestingly, most of the TPGs were correlated with good outcome (with  $HR < 1$ ),  
183 and only six genes were associated with poor outcome (with  $HR > 1$ ). Through conducting  
184 intersection analysis, we found nine genes (LUM, VCAN, YAP1, COL5A2, SDC2, TWIST1,  
185 CD3G, C1QB and C3) were intersection genes, indicating that they had an active effect in  
186 modulating the tumor purity, as well as influencing the prognosis of DLBCL patients.  
187 After ascertaining the key genes, we tried to construct a prognostic model with them. The model  
188 was constructed by Cox regression, and the three selected genes (VCAN, CD3G, C1QB) and their  
189 parameters like coefficient, HR and 95%CI of HR, were shown in **Figure 3A**. It showed that  
190 VCAN and CD3G were correlated with good prognosis and C1QB was correlated with poor  
191 prognosis. All patients were divided into high and low-risk group according to the median value of  
192 riskScore (**Figure 3B**). As expected, the high-risk group has worse prognosis than the low-risk  
193 group (**Figure 3C**). In addition, the three genes were differentially expressed between high and  
194 low-risk group, with VCAN and CD3G showing high expression level in low-risk group, and  
195 C1QB showing high expression level in high-risk group, which was consistent with the coefficient  
196 (**Figure 3D**). To appraise the efficacy of these prognostic model, we conducted survival analysis  
197 and ROC analysis. High-risk group had lower OS rate than low-risk group (**Figure 3E**,  $p < 0.001$ ),  
198 and the areas under curve (AUC) for 1-year, 3-year and 5-year ROC were 0.73, 0.77 and 0.77  
199 respectively (**Figure 3F**). Just similar to tumor purity, the high riskScore indicated bad outcome,  
200 which was consistent with the positive correlation between tumor purity and riskScore (**Figure**  
201 **3G**). This TPGs signature prognostic model manifested satisfying prognostic efficacy.  
202 When we applied this model to GSE32918 dataset, it still did excellently and the results were in  
203 accordance with that in GSE53786 dataset (**Figure 4A, 4B; Supplementary Figure 2**). Next, we  
204 analyzed the relationship between riskScore and some clinicopathological characteristics provided  
205 in GSE53786 dataset. The results showed that high-risk group had more ABC type DLBCL, while  
206 low-risk group had more the GCB type DLBCL (**Figure 4C**). Besides, high-risk group displayed  
207 higher lactic dehydrogenase (LDH) ratio (**Figure 4F**). However, the Eastern Cooperative  
208 Oncology Group (ECOG) performance and stage was not associated with the riskScore (**Figure**  
209 **4D, 4E**). Still, the high-risk group has more Stage III and Stage IV patients, but less Stage I  
210 patients than low-risk group. Finally, we employed the univariate and multivariate analysis to

211 explore whether riskScore was an independent prognostic factor for DLBCL patients. As expected,  
212 the riskScore was associated with the poor prognosis (**Figure 4G**,  $p < 0.001$ , HR = 1.545, 95%CI  
213 1.284–1.861) and was an independent prognostic factor (**Figure 4H**,  $p = 0.002$ , HR = 1.474,  
214 95%CI 1.156–1.879).

215

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

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

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

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

246 In addition to the above analyses, we also explored the relationship between these three genes and  
247 location of DLBCL. It turned out that CD3G+ T cells ratio was higher in DLBCL originated from  
248 groin and testis, and VCAN featured higher expression in lymph node originated DLBCL (**Figure**  
249 **5J–K**,  $p < 0.05$ ,  $p < 0.01$ , **Supplementary figure 3B–K**).

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

253

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

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

259 As is shown in **Figure 6A** (prediction of GSE53786), patients in high-risk group could be  
260 sensitive to Carmustine, Cytarabine, Oxaliplatin, Vincristine, Vorinostat, and Bortezomib, but no  
261 drug could work better in low-risk group. And in GSE32918 (**Figure 6B**), Carmustine, Cytarabine,  
262 Oxaliplatin, Vorinostat, Afuresertib, Bortezomib, Ibrutinib and Tamoxifen could work better in  
263 high-risk group, and Vincristine (sensitivity score: low-risk vs high-risk =  $0.219 \pm 0.026$  vs  
264  $0.223 \pm 0.031$ ) could work better in low-risk group. The discrepancy between the prediction in two  
265 datasets might be due to the samples and sequencing platforms. However, the intersection analysis  
266 of the drugs to which the high-risk patients in both datasets could be sensitive revealed that  
267 Carmustine, Cytarabine, Oxaliplatin, Vorinostat and Bortezomib could be reliable candidates for  
268 treating high-risk patients based on the three TPGs signature prognostic model (**Supplementary**  
269 **table 2**).

270

271 **Discussion**

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

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

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

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

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

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

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

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

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

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



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

369

#### 370 **Author contributions**

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

378

#### 379 **Funding**

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

384

#### 385 **Acknowledgement**

386 Not applicable.

387

#### 388 **Conflicts of interest**

389 The authors made no disclosures.

390

391 **Figure 1 TPGs were screened out with GSE53786 dataset.**

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

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

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

405 **Figure 3 TPGs signature prognostic model was constructed.**

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

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

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

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

428 **Figure 5 The analysis of CHCAMS cohort.**

429 (A) The representative image of VCAN staining of high and low expression groups and the  
430 survival analysis based on VCAN expression. (B) The representative image of CD3G staining  
431 of high and low CD3G+ T cells ratio groups and the survival analysis based on CD3G+ T  
432 cells ratio. (C) The representative image of C1QB staining of high and low expression group  
433 and the survival analysis based on C1QB expression. (D) The representative image of CD68  
434 staining of high and low CD68+ macrophages ratio groups and the survival analysis based on  
435 CD68+ macrophages ratio. (E) The representative image of CD8 staining of high and low  
436 CD8+ T cells ratio groups and the survival analysis based on CD8+ T cells ratio. (F) The  
437 representative image of CD4 staining of high and low CD4+ T cells ratio groups and the  
438 survival analysis based on CD4+ T cells ratio. (G) The correlation between VCAN, CD3G+ T  
439 cells ratio, C1QB and CD68+ macrophages, CD8+ T cells and CD4+ T cells ratio. “×”  
440 means no statistical significance. (H–I) GSEA analysis based on the differentially expressed  
441 genes between high-risk and low-risk group in GSE53786 and GSE32918. (J) The CD3G+ T  
442 cells infiltration varied from colon to testis originating DLBCL in male. (K) The VCAN  
443 expression level was different between intra- and extra-lymph node DLBCL.

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

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

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

450 **Supplementary Figure 2 The correlation between tumor purity and prognosis and**

451 **riskScore in GSE32918 dataset.**

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

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

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

468

469   References:

- 470   1.     Alaggio, R., et al., *The 5th edition of the World Health Organization Classification of*  
471         *Haematolymphoid Tumours: Lymphoid Neoplasms*. *Leukemia*, 2022. **36**(7): p.  
472         1720-1748.
- 473   2.     Sehn, L.H. and G. Salles, *Diffuse Large B-Cell Lymphoma*. *N Engl J Med*, 2021.  
474         **384**(9): p. 842-858.
- 475   3.     Hans, C.P., et al., *Confirmation of the molecular classification of diffuse large B-cell*  
476         *lymphoma by immunohistochemistry using a tissue microarray*. *Blood*, 2004. **103**(1): p.  
477         275-82.
- 478   4.     Autio, M., et al., *Immune cell constitution in the tumor microenvironment predicts the*  
479         *outcome in diffuse large B-cell lymphoma*. *Haematologica*, 2021. **106**(3): p. 718-729.
- 480   5.     Joyce, J.A. and J.W. Pollard, *Microenvironmental regulation of metastasis*. *Nat Rev*  
481         *Cancer*, 2009. **9**(4): p. 239-52.
- 482   6.     Steen, C.B., et al., *The landscape of tumor cell states and ecosystems in diffuse large*  
483         *B cell lymphoma*. *Cancer Cell*, 2021. **39**(10): p. 1422-1437.e10.
- 484   7.     Ciavarella, S., et al., *Dissection of DLBCL microenvironment provides a gene*  
485         *expression-based predictor of survival applicable to formalin-fixed paraffin-embedded*  
486         *tissue*. *Ann Oncol*, 2018. **29**(12): p. 2363-2370.
- 487   8.     Ennishi, D., et al., *Toward a New Molecular Taxonomy of Diffuse Large B-cell*  
488         *Lymphoma*. *Cancer Discov*, 2020. **10**(9): p. 1267-1281.
- 489   9.     Miyawaki, K., et al., *A germinal center-associated microenvironmental signature*  
490         *reflects malignant phenotype and outcome of DLBCL*. *Blood Adv*, 2022. **6**(7): p.

- 491 2388-2402.
- 492 10. Schmitz, R., et al., *Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma*. N  
493 Engl J Med, 2018. **378**(15): p. 1396-1407.
- 494 11. Lou, X., et al., *CCL8 as a promising prognostic factor in diffuse large B-cell lymphoma*  
495 *via M2 macrophage interactions: A bioinformatic analysis of the tumor*  
496 *microenvironment*. Front Immunol, 2022. **13**: p. 950213.
- 497 12. Gong, Z., J. Zhang, and W. Guo, *Tumor purity as a prognosis and immunotherapy*  
498 *relevant feature in gastric cancer*. Cancer Med, 2020. **9**(23): p. 9052-9063.
- 499 13. Lou, S., et al., *Comprehensive Characterization of Tumor Purity and Its Clinical*  
500 *Implications in Gastric Cancer*. Front Cell Dev Biol, 2021. **9**: p. 782529.
- 501 14. Zhao, Y., et al., *Tumor purity-associated genes influence hepatocellular carcinoma*  
502 *prognosis and tumor microenvironment*. Front Oncol, 2023. **13**: p. 1197898.
- 503 15. Zhang, C., et al., *Tumor Purity as an Underlying Key Factor in Glioma*. Clin Cancer  
504 Res, 2017. **23**(20): p. 6279-6291.
- 505 16. Yoshihara, K., et al., *Inferring tumour purity and stromal and immune cell admixture*  
506 *from expression data*. Nat Commun, 2013. **4**: p. 2612.
- 507 17. Wang, M., D. Windgassen, and E.T. Papoutsakis, *Comparative analysis of*  
508 *transcriptional profiling of CD3+, CD4+ and CD8+ T cells identifies novel immune*  
509 *response players in T-cell activation*. BMC Genomics, 2008. **9**: p. 225.
- 510 18. Baghy, K., et al., *Proteoglycans in liver cancer*. World J Gastroenterol, 2016. **22**(1): p.  
511 379-93.
- 512 19. Wight, T.N., *Provisional matrix: A role for versican and hyaluronan*. Matrix Biol, 2017.

513 **60-61:** p. 38-56.

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

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

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

520 960570.

521 23. Wang, M.Q., et al., *VCAN, expressed highly in hepatitis B virus-induced hepatocellular*  
522 *carcinoma, is a potential biomarker for immune checkpoint inhibitors*. World J

523 Gastrointest Oncol, 2022. **14**(10): p. 1933-1948.

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

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

528 **14**: p. 1123840.

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

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

533 1765.

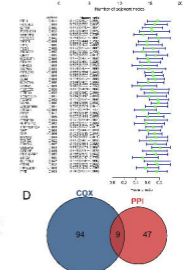
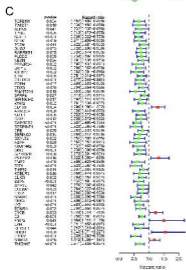
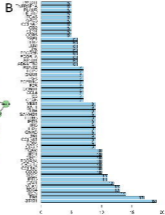
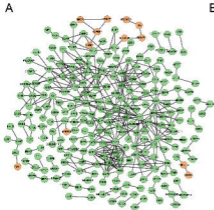
534 28. Chen, L.H., et al., *Complement C1q (C1qA, C1qB, and C1qC) May Be a Potential*

- 535            *Prognostic Factor and an Index of Tumor Microenvironment Remodeling in*  
536            *Osteosarcoma*. *Front Oncol*, 2021. 11: p. 642144.
- 537    29.    Zheng, Y., et al., *IRF4-activated TEX41 promotes the malignant behaviors of*  
538            *melanoma cells by targeting miR-103a-3p/C1QB axis*. *BMC Cancer*, 2021. 21(1): p.  
539            1339.
- 540    30.    Chen, Z., et al., *A Machine Learning Model to Predict the Triple Negative Breast*  
541            *Cancer Immune Subtype*. *Front Immunol*, 2021. 12: p. 749459.
- 542    31.    Wang, Q., P. Li, and W. Wu, *A systematic analysis of immune genes and overall*  
543            *survival in cancer patients*. *BMC Cancer*, 2019. 19(1): p. 1225.
- 544    32.    Wang, J., et al., *Establishment and validation of immune microenvironmental gene*  
545            *signatures for predicting prognosis in patients with head and neck squamous cell*  
546            *carcinoma*. *Int Immunopharmacol*, 2021. 97: p. 107817.
- 547    33.    Zhang, Y., et al., *GM-CSF enhanced the effect of CHOP and R-CHOP on inhibiting*  
548            *diffuse large B-cell lymphoma progression via influencing the macrophage polarization*.  
549            *Cancer Cell Int*, 2021. 21(1): p. 141.
- 550    34.    Croci, G.A., et al., *SPARC-positive macrophages are the superior prognostic factor in*  
551            *the microenvironment of diffuse large B-cell lymphoma and independent of MYC*  
552            *rearrangement and double-/triple-hit status*. *Ann Oncol*, 2021. 32(11): p. 1400-1409.
- 553    35.    Farhood, B., M. Najafi, and K. Mortezaee, *CD8(+) cytotoxic T lymphocytes in cancer*  
554            *immunotherapy: A review*. *J Cell Physiol*, 2019. 234(6): p. 8509-8521.
- 555    36.    Roussel, M., et al., *Functional characterization of PD1+TIM3+ tumor-infiltrating T cells*  
556            *in DLBCL and effects of PD1 or TIM3 blockade*. *Blood Adv*, 2021. 5(7): p. 1816-1829.



- 557 37. Xu-Monette, Z.Y., et al., *Immune Profiling and Quantitative Analysis Decipher the*  
558 *Clinical Role of Immune-Checkpoint Expression in the Tumor Immune*  
559 *Microenvironment of DLBCL*. *Cancer Immunol Res*, 2019. **7**(4): p. 644-657.
- 560 38. Hirani, P., et al., *Targeting Versican as a Potential Immunotherapeutic Strategy in the*  
561 *Treatment of Cancer*. *Front Oncol*, 2021. **11**: p. 712807.
- 562 39. Huang, X.Y., et al., *Bioinformatics analysis of the prognosis and biological significance*  
563 *of VCAN in gastric cancer*. *Immun Inflamm Dis*, 2021. **9**(2): p. 547-559.
- 564





**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

