Large-scale transcriptome data analysis reveals prognostic signature genes for overall survival prediction in diffuse large B cell lymphoma

Mengmeng Pan (panmengmeng520@126.com)
Tongji University https://orcid.org/0000-0001-5060-2753

Pingping Yang
Tongji University

Fangce Wang
Tongji University

Xiu Luo
Tongji University

Bing Li
Tongji University

Yi Ding
Tongji University

Huina Lu
Tongji University

Yan Dong
Tongji University

Wenjun Zhang
Tongji University

Bing Xiu
Tongji University

Aibin Liang
Tongji University

Primary research

Keywords: Diffuse large B cell lymphoma, overall survival, prognosis, biomarkers, risk score

Posted Date: July 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-43517/v1

Abstract

BACKGROUND With the improvement of clinical treatment outcomes in Diffuse large B cell lymphoma (DLBCL), the high rate of relapse in DLBCL patients is still an established barrier, due to the therapeutic strategy selection based on potential target remains unsatisfactory. Therefore, there is an urgent need in further exploration of prognostic biomarkers so as to improve the prognosis of DLBCL.

METHODS The univariable and multivariable Cox regression models were employed to screen out gene signatures for DLBCL overall survival prediction. The differential expression analysis was used to identify representative genes in high-risk and low-risk groups, respectively, by student t test and fold change. The functional difference between the high-risk and low-risk groups were identified by the gene set enrichment analysis.

RESULTS We conducted a systematic data analysis to screen the candidate genes significantly associated with overall survival of DLBCL in three NCBI Gene Expression Omnibus (GEO) datasets. To construct a prognostic model, five genes (*CEBPA*, *CYP27A1*, *LST1*, *MREG*, and *TARP*) were then screened and tested using the multivariable Cox model and the stepwise regression method. Kaplan-Meier curve confirmed the good predictive performance of the five-gene Cox model. Thereafter, the prognostic model and the expression levels of the five genes were validated by means of an independent dataset. All five genes were significantly favorable for the prognosis in DLBCL, both in training and validation datasets. Additionally, further analysis revealed the independence and superiority of the prognostic model in risk prediction. Functional enrichment analysis revealed some vital pathways resulting in unfavorable outcome and potential therapeutic targets in DLBCL.

CONCLUSION We developed a five-gene Cox model for the clinical outcome prediction of DLBCL patients. Meanwhile, potential drug selection using this model can help clinicians to improve the clinical practice for the patients.

Background

Diffuse large B cell lymphoma (DLBCL) is the most common type of aggressive non-Hodgkin lymphoma with an annual incidence of 1–5/10,000[1, 2]. DLBCL is an aggressive and potentially curable hematological malignancy, which makes an early diagnosis and effective treatments essential for patients. R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, prednisone) is the current standard first line treatment of DLBCL [3]. Despite the high rate of complete response (76%), approximately 40% of patients will relapse, and the molecular mechanism underlying recurrence remains largely unknown[4]. DLBCL displays tremendous clinical, genetical and molecular heterogeneity. The International Prognostic Index (IPI) has been used to predict the prognosis of patients with DLBCL for nearly 30 years, there still have minority of patients whose clinical process were not in accord with the IPI stratification [5]. Gene expression profiling has helped identify two major subtypes, known as germinal center B-cell-like (GCB) and activated B-cell-like (ABC), of which patients of the ABC group exhibit a

generally worse prognosis [6]. However, the high prices and strict requirements regarding tissue limit the routine use of this method. Therefore, efforts have been made to find novel biomarkers with prognostic values in order to improve therapeutic strategy selection based on potential targets[7].

Currently, various markers are defined through immunophenotyping, such as CD5, CD30, BCL2, MYC and TP53[8, 9]. CD5 promotes downstream B-cell receptor signaling, is associated with ABC subtype and more aggressive clinical traits. Patients with $CD30^{+}$ DLBCL, which leads to the downregulation of NF-κB and B-cell receptor signaling, tends to exhibit a better prognosis[10, 11]. Meanwhile, in patients with the GCB subtype, BCL2 and MYC rearrangements would lead to worse prognosis[12]. TP53 mutation also adversely affects patients' prognosis[13]. Based on the new integrated genetic map, B. Chapuy, *et al* identified distinct subsets, including a previously unrecognized group of low-risk ABC-DLBCLs, two GCB-DLBCLs subsets with different prognosis and an ABC/GCB-independent group[14]. In addition, R. Schmitz, *et al* uncovered some previously unknown subtypes of DLBCL by differences in gene-expression signatures and responses to immunochemotherapy [15]. The subset of high-risk patients requires more intensive therapy, the personalized therapy based on patient's histological and molecular-genetic characteristics will bring greater benefits to patients. Therefore, further exploration of prognostic indicators is still needed to distinguish DLBCL patients with varied prognosis.

Materials And Methods

Data collection

The gene expression data and corresponding clinical information were collected from NCBI Gene Expression Omnibus (GEO) database with accession numbers of GSE32918 (n = 172), GSE4475 (n = 166), GSE69051 (n = 172), and GSE10846 (n = 414). The expression values were normalized by the data submitters, and discretized by median, which were used for downstream analysis.

Cox proportional hazard model

The univariable Cox proportional hazard model was used to screen the prognostic genes in the first three datasets. Subsequently, the gene signatures jointly identified in the three datasets were then subjected for the multivariable Cox model, and the stepwise regression method was used to determine the best model. The expression threshold for splitting high-expression and low-expression was selected based on the minimal P-value of the log-rank tests in the training set. The threshold in validation set was selected by the quantile in the training set. The risk scores for the samples of training and validation sets were estimated by the multivariable Cox model. The high- and low-risk groups were stratified based on the median of the risk scores in the training set. The independence of the risk stratification on the cofactors was also assessed by multivariable Cox model.

Differential Gene Expression Analysis

The differential gene expression analysis was conducted to identify the genes that upregulated or downregulated in a specified group as compared the other one. The Wilcoxon rank-sum test and fold change methods were employed, and the thresholds of adjusted p-value and log2-fold change were determined at 0.05 and 0.5.

The Pathway Enrichment Analysis

The upregulated genes in each risk group were subjected to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis, respectively. Hypergeometric test was applied to test the statistical significance of the pathways. The threshold for adjusted P-value was determined at 0.05.

The Drug-target Identification

The therapeutic targets were selected from the upregulated genes in each risk group. The drugs and upregulated genes were mapped by the R package maftools with *drugInteractions*.

Results

Systematic identification of prognostic gene signatures for overall survival prediction

To identify the prognostic gene signatures, we collected three public DLBCL datasets with accession numbers of GSE32918 (n = 172), GSE4475 (n = 166) and GSE69051 (n = 172) from Gene Expression Omnibus (GEO) database as depicted in the flow chart in Figure S1. Subsequently, univariable Cox regression analysis was conducted, and identified 763, 685, and 589 genes associated with overall survival (OS) based on the three gene expression datasets (Fig. 1A, log-rank test, $P < 0.01$), respectively. Particularly, *CEBPA*, *CSF2RA*, *CYP27A1*, *LST1*, *MREG*, *SCPEP1*, and *TARP* were found to be significantly associated with OS in all the three datasets at the stringent threshold (Fig. 1A). Furthermore, the three datasets were merged as training set (n = 510), and multivariable Cox regression model was then built based on these genes using the merged dataset. A stepwise method was used to select a subset of those gene signatures and a multivariable Cox regression model with highest performance. Specifically, five genes including *CEBPA*, *CYP27A1*, *LST1*, *MREG*, and *TARP* were retained in the multivariable Cox model (Table 1), which was termed as five-gene Cox model, and all of them were favorable for the prognosis (Fig. 1B).

Performance Validation In An Independent Dataset

To evaluate the performance of the multivariable model in risk prediction, we first calculated the risk scores for the DLBCL samples in the training set, and stratified these samples into high and low risk

groups by the median of risk scores. The high-risk group exhibited worse prognosis than the low-risk group (Fig. 2A, P < 0.0001). Moreover, we also collected an independent gene expression dataset of 414 samples with long-term follow-up. Similarly, the risk scores were predicted and used to stratify the samples in the validation set. Consistently, the two groups also had significant prognostic difference (Fig. 2B, P < 0.0001). Furthermore, the five gene signatures were found to be expressed higher in low-risk group than high-risk group in both the training and validation sets (Figs. 2C-2D). These results indicated that the five gene signatures were robust and consistently associated with OS in both training and validation datasets.

The five-gene Cox model is superior to other gene expression-based Cox models

To demonstrate the superiority of the five-gene Cox model based on the five gene signatures, we compared its performance with three sets of gene signatures[16–18] in the validation set. Based on the trained models by the gene signatures, the samples in the validation set could also be stratified into high- and low-risk groups. The three sets of gene signatures had the capability of predicting the prognosis of DLBCL patients (Fig. 3A-C). The gene signatures proposed by Rosenwald *et al.* [17] had the worst performance (Fig. 3B). However, our proposed five gene signatures showed the highest statistical significance than the other sets (Fig. 3D), suggesting that the Cox model based on the five gene signatures was superior to other models.

The five-gene-based risk stratification is a prognostic factor independent of clinical factors

To further investigate the robustness of the five-gene Cox model, we tested the independence of the five-gene-based risk stratification in the validation set. As the IPI scoring system was a well-recognized factor for prognostic risk prediction and widely applied in clinical practice[19], the samples were first divided into two groups with high (> = 3) or low (< 3) IPI scores, considering age, serum lactate dehydrogenase (LDH), Eastern Cooperative Oncology Group (ECOG) Performance Status, Ann Arbor stage, and extranodal infiltration sites[5]. As shown in Fig. 4A, the risk scores estimated by the five-gene Cox model showed no significant difference between the two groups. Moreover, the differences were not also observed between the four stages. These results suggested that the risk scores were not associated with IPI scoring system and tumor stage.

Notably, the samples could be classified into four groups by combining the IPI scoring system and the five-gene-based risk stratification, and the four groups exhibited significantly prognostic difference (Fig. 4B, P < 0.0001). It should be noted that the differences of overall survival were not observed between the two groups with the worse prognosis, but the samples with IPI > = 3 in high-risk group still had shorter overall survival than samples with IPI > = 3 in the low-risk group based on the KM curve.

Moreover, we also tested whether the risk stratification was independent of the subtypes. Consistently, the three subtypes, including ABC, GCB and unclassified subtypes, could be further stratified into high- and low-risk groups, and the samples in high-risk group had significantly shorter survival than those in low-risk group (Fig. 4C-E). In addition, we also fitted the IPI scoring system, stage, subtype and risk

stratification into a multivariable Cox model, and found that the risk stratification was still statistically significant with these prognostic factors as cofactors (Table 2). These results further demonstrated that the five-gene-based risk stratification was an independent prognostic factor for DLBCL risk prediction.

The molecular characteristics and potential drugs for the two risk groups

To reveal the molecular characteristics of the two risk groups, we compared the gene expression profiles of high-risk with those of low-risk group. The differentially expressed genes (DEGs) were then selected by Wilcoxon rank-sum test and fold change (Adjusted *P*-value < 0.05 and log2-fold change > 0.5). Moreover, the overrepresentation enrichment analysis (ORA) was employed to identify the pathways potentially involved in the DLBCL progression (Fig. 5A). Specifically, cell cycle-related pathway and those associated with genomic stability maintenance, such as Fanconi anemia pathway and Mismatch repair, were highly upregulated in high-risk group (Adjusted *P*-value < 0.05). In contrast, immune-related pathways such as rheumatoid arthritis, antigen processing and presentation, allograft rejection, hematopoietic cell lineage, and Th1 and Th2 cell differentiation were upregulated in low-risk group (Adjusted *P*-value < 0.05).

Among the DEGs, 29 genes were found to be therapeutic targets specifically upregulated in the two risk groups (Fig. 5B). As we known, CD20 (also termed *MS4A1*) is expressed on the surface of normal B lymphocytes and almost all DLBCL cases. At present, RITUXIMAB, a chimeric monoclonal antibody directed against the CD20, combine with intensive chemotherapy (CHOP) is the standard therapy for DLBCL. Besides, two cell cycle kinases, AURKB and CDK1, were upregulated in high-risk group, and BARASERTIB and DINACICLIB might be the potential drugs for treating DLBCL with high-risk factors. For the low-risk group, some immune checkpoint proteins and inhibitors were identified, such as PDCD1 (PD-1), CD274 (PD-L1), CTLA4, and their corresponding drugs, suggesting that the low-risk samples might benefit from inhibiting the immune checkpoint pathway.

Discussion

DLBCL is a remarkably heterogeneous disease, both histologically and genetically. Despite significant advances in subtype classification of DLBCL, accurate prediction of prognosis remains a challenge. With the development of high throughput sequencing technology, some potential prognostic genomic markers for DLBCL patients have been identified[17, 20, 21]. However, the number of prognostic markers is still limited. There is an urgent need to screen out more biomarkers to improve the accuracy of prognostic prediction.

In the present study, we identified potential gene candidates by using the univariable Cox regression analysis to examine associations between gene expression and patient prognosis of three DLBCL cohorts in GEO. To further narrow down these gene signatures, multivariate Cox analysis was carried out on the merged datasets. A stepwise approach was used to select a subset of gene candidates with highest performance, and a risk model was established for predicting DLBCL prognosis based on the expression levels of five genes including *CEBPA*, *CYP27A1*, *LST1*, *MREG*, and *TARP*. We evaluated the model performance using an independent gene expression dataset and compared it with previously reported models. Our five-gene based risk model showed improved robustness, accuracy, and efficiency compared

to those models and was demonstrated to be an independent prognostic factor for overall survival in patients with DLBCL. Subsequently, we compared the gene expression profiles of high-risk with those of low-risk group and performed ORA to identify pathways potentially involved in the DLBCL progression. Thus, we believe that our five-gene-based risk scoring model can be used for refining DLBCL subtypes and potentially improving patient therapy.

According to the multivariable Cox model, high expression of the five genes were all associated with a favorable survival outcome. CEBPA is a transcription factor playing roles in regulating proliferation and differentiation of many cell types[22]. Within the hematopoietic system, inactivation mutation of CEBPA blocks the granulocytic differentiation in acute myeloid leukemia (AML) [23]. In addition, it has been reported that CEBPA-regulated PER2 activation is a potential tumor suppressor pathway in diffuse large B-cell lymphoma (DLBCL)[24]. CYP27A1, a cytochrome P450 oxidase family member, is closely related to the proliferation of multiple tumor cells, such as prostate, breast and colon cancer[25–27]. LST1 is encoded within the TNF region of the human MHC which regulates lymphocyte proliferation [28]. MREG is reported to suppress thyroid cancer cell invasion and proliferation through PI3K/Akt-mTOR signaling pathway [29]. The biological role of these genes in DLBCL need to be further investigated.

The overrepresentation enrichment analysis of DEGs suggests that the abnormal cell cycle progression and increased genomic instability contribute to the rapid progression of DLBCL. Inhibitors of cell cycle kinases, such as BARASERTIB and DINACICLIB, may be effective in high-risk patients. On the contrary, genes related to immune-related pathways, such as antigen processing and presentation, Th1 and Th2 cell differentiation, were enrichment in low-risk group, suggesting that activated host immune response may predict favorable prognosis and response to therapy. These findings provide novel clues into the explanation of the mechanism of DLBCL.

The prognostic model we proposed is helpful for further risk stratification in genetic level on the basis of the current traditional typing, but this study still has some limitations. Some potential prognostic factors may be not included in the model such as the racial factors and the roles that the five genes play in DLBCL required further experimental validation.

Conclusions

To sum up, our research indicates that the five-gene prognostic model is a reliable tool for predicting the OS of DLBCL patients and providing some hints on drug selection, which can assist clinicians in selecting personalized treatment, although specific drug selection requires further molecular biology research and clinical trials.

Abbreviations

DLBCL
Diffuse large B cell lymphoma

IPI
International Prognostic Index
GCB
germinal center B-cell-like
ABC
activated B-cell-like
GEO
Gene Expression Omnibus
LDH
serum lactate dehydrogenase
ECOG
Eastern Cooperative Oncology Group
CHOP
combine with intensive chemotherapy

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Participants gave their written informed consent for the materials to appear in publications without limit on the duration of publication.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors have declared that no competing interests exist.

Funding
This work was sponsored by Shanghai Sailing Program (No. 19YF1444300), Program of Outstanding Young Scientists of Tongji Hospital of Tongji University (No. HBRC1802), Youth project of scientific research project of Shanghai Health and Family Planning Commission (No.20174Y0110), and the Key Project of Natural Science Foundation of China (No. 81830004).

Author contributions

BING XIU and AIBIN LIANG conceived and designed the experiments. MENGMENG PAN, PINGPING YANG, and FANGCE WANG acquired data, related materials, and analysis tools. MENGMENG PAN, XIU LUO, and BING LI analyzed the data. MENGMENG PAN, PINGPING YANG, and FANGCE WANG wrote the paper. YI DING, HUINA LU, and YAN DONG revised the paper. All authors read and approved the final manuscript.

Acknowledgements
Not applicable.

References

1. Marangon AV, Colli CM, Cardozo DM, Visentainer JEL, Sell AM, Guimaraes F, Marques SBD, Lieber SR, Aranha FJP, Zulli R, et al. Impact of SNPs/Haplotypes of and on the Development of Diffuse Large B-Cell Lymphoma. J Immunol Res. 2019;2019:2137538.
2. Li S, Young KH, Medeiros LJ. Diffuse large B-cell lymphoma. Pathology. 2018;50(1):74–87.
3. Coiffier B, Lepage E, Briere J, Herbrecht R, Tilly H, Bouabdallah R, Morel P, Van Den Neste E, Salles G, Gaulard P, et al. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. N Engl J Med. 2002;346(4):235–42.
4. Coiffier B, Thieblemont C, Van Den Neste E, Lepeu G, Plantier I, Castaigne S, Lefort S, Marit G, Macro M, Sebban C, et al. Long-term outcome of patients in the LNH-98.5 trial, the first randomized study comparing rituximab-CHOP to standard CHOP chemotherapy in DLBCL patients: a study by the Groupe d'Etudes des Lymphomes de l'Adulte. Blood. 2010;116(12):2040–5.
5. **A predictive model for aggressive non-Hodgkin's lymphoma**. *N Engl J Med* 1993, 329(14):987–994.
6. Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, Xu W, Tan B, Goldschmidt N, Iqbal J, et al. Stromal gene signatures in large-B-cell lymphomas. N Engl J Med. 2008;359(22):2313–23.
7. Cabanillas F, Shah B. Advances in Diagnosis and Management of Diffuse Large B-cell Lymphoma. Clin Lymphoma Myeloma Leuk. 2017;17(12):783–96.
8. Pierce JMR, Mehta A. Diagnostic, prognostic and therapeutic role of CD30 in lymphoma. Expert Rev Hematol. 2017;10(1):29–37.
9. Zhao P, Li L, Zhou S, Qiu L, Qian Z, Liu X, Meng B, Zhang H. CD5 expression correlates with inferior survival and enhances the negative effect of p53 overexpression in diffuse large B-cell lymphoma. Hematol Oncol. 2019;37(4):360–7.
10. Bhatt G, Maddocks K, Christian B. CD30 and CD30-Targeted Therapies in Hodgkin Lymphoma and Other B cell Lymphomas. Curr Hematol Malig Rep. 2016;11(6):480–91.
11. Thakral B, Medeiros LJ, Desai P, Lin P, Yin CC, Tang G, Khoury JD, Hu S, Xu J, Loghavi S, et al. Prognostic impact of CD5 expression in diffuse large B-cell lymphoma in patients treated with rituximab-EPOCH. Eur J Haematol. 2017;98(4):415–21.
12. Visco C, Tzankov A, Xu-Monette ZY, Miranda RN, Tai YC, Li Y, Liu W-m, d'Amore ESG, Li Y, Montes-Moreno S, et al. Patients with diffuse large B-cell lymphoma of germinal center origin with BCL2

translocations have poor outcome, irrespective of MYC status: a report from an International DLBCL rituximab-CHOP Consortium Program Study. Haematologica. 2013;98(2):255–63.

13. Xu-Monette ZY, Wu L, Visco C, Tai YC, Tzankov A, Liu W-m, Montes-Moreno S, Dybkaer K, Chiu A, Orazi A, et al. Mutational profile and prognostic significance of TP53 in diffuse large B-cell lymphoma patients treated with R-CHOP: report from an International DLBCL Rituximab-CHOP Consortium Program Study. Blood. 2012;120(19):3986–96.
14. Chapuy B, Stewart C, Dunford AJ, Kim J, Kamburov A, Redd RA, Lawrence MS, Roemer MGM, Li AJ, Ziepert M, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nature medicine. 2018;24(5):679–90.
15. Schmitz R, Wright GW, Huang DW, Johnson CA, Phelan JD, Wang JQ, Roulland S, Kasbekar M, Young RM, Shaffer AL, et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. N Engl J Med. 2018;378(15):1396–407.
16. Lossos IS. **Diffuse Large B Cell Lymphoma: From Gene Expression Profiling to Prediction of Outcome**. 2008, 14(1):108–111.
17. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med. 2002;346(25):1937–47.
18. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM: **A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma**. *Proceedings of the National Academy of Sciences* 2003, **100**(17):9991–9996.
19. Martelli M, Ferreri AJM, Agostinelli C, Di Rocco A, Pfreundschuh M, Pileri SA. Diffuse large B-cell lymphoma. Crit Rev Oncol Hematol. 2013;87(2):146–71.
20. Lossos IS. Diffuse large B cell lymphoma: from gene expression profiling to prediction of outcome. Biol Blood Marrow Transplant. 2008;14(1 Suppl 1):108–11.
21. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proc Natl Acad Sci USA. 2003;100(17):9991–6.
22. Gery S, Gombart AF, Yi WS, Koeffler C, Hofmann W-K, Koeffler HP. Transcription profiling of C/EBP targets identifies Per2 as a gene implicated in myeloid leukemia. Blood. 2005;106(8):2827–36.
23. Wang X, Scott E, Sawyers CL, Friedman AD. C/EBPalpha bypasses granulocyte colony-stimulating factor signals to rapidly induce PU.1 gene expression, stimulate granulocytic differentiation, and limit proliferation in 32D cl3 myeloblasts. Blood. 1999;94(2):560–71.
24. Thoennissen NH, Thoennissen GB, Abbassi S, Nabavi-Nouis S, Sauer T, Doan NB, Gery S, Müller-Tidow C, Said JW, Koeffler HP. Transcription factor CCAAT/enhancer-binding protein alpha and critical circadian clock downstream target gene PER2 are highly deregulated in diffuse large B-cell lymphoma. Leuk Lymphoma. 2012;53(8):1577–85.
25. Ji Y-C, Liu C, Zhang X, Zhang C-S, Wang D, Zhang Y. Intestinal bacterium-derived cyp27a1 prevents colon cancer cell apoptosis. Am J Transl Res. 2016;8(10):4434–9.

26. Alfaqih MA, Nelson ER, Liu W, Safi R, Jasper JS, Macias E, Geradts J, Thompson JW, Dubois LG, Freeman MR, et al. CYP27A1 Loss Dysregulates Cholesterol Homeostasis in Prostate Cancer. Cancer research. 2017;77(7):1662–73.
27. Kimbung S, Chang C-Y, Bendahl P-O, Dubois L, Thompson JW, McDonnell DP, Borgquist S. Impact of 27-hydroxylase (CYP27A1) and 27-hydroxycholesterol in breast cancer. Endocr Relat Cancer. 2017;24(7):339–49.
28. Rollinger-Holzinger I, Eibl B, Pauly M, Griesser U, Hentges F, Auer B, Pall G, Schratzberger P, Niederwieser D, Weiss EH, et al. LST1: a gene with extensive alternative splicing and immunomodulatory function. J Immunol. 2000;164(6):3169–76.
29. Meng X, Dong Y, Yu X, Wang D, Wang S, Chen S, Pang S. MREG suppresses thyroid cancer cell invasion and proliferation by inhibiting Akt-mTOR signaling. Biochem Biophys Res Commun. 2017;491(1):72–8.

Tables

Table 1
The statistics for the gene signatures in the multivariable Cox model.

Gene	coef	exp(coef)	se(coef)	z	Pr(>|z|)
CEBPA	-0.384	0.681	0.180	-2.138	3.25E-02
CYP27A1	-0.390	0.677	0.187	-2.086	3.69E-02
LST1	-0.468	0.626	0.178	-2.631	8.50E-03
MREG	-0.420	0.657	0.170	-2.471	1.35E-02
TARP	-0.292	0.746	0.156	-1.873	6.11E-02

Table 2
The statistics for the risk stratification and prognostically clinical factors in the multivariable Cox model.

Variables	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Subtype					
ABC	-	-	-	-	-
GCB	-0.931	0.394	0.202	-4.604	**4.15E-06**
Unclassified	-0.726	0.484	0.267	-2.718	**6.56E-03**
Stage					
1	-	-	-	-	-
2	1.074	2.926	0.412	2.605	**9.18E-03**
3	0.641	1.899	0.441	1.455	1.46E-01
4	1.023	2.781	0.427	2.395	**1.66E-02**
Risk Stratification					
high-risk	-	-	-	-	-
low-risk	-0.583	0.558	0.178	-3.273	**1.06E-03**
IPI					
< 3	-	-	-	-	-
>= 3	0.939	2.558	0.214	4.398	**1.09E-05**

Table 3
The performance of the three stratifications in the validation data.

Studies	coef	exp(coef)	se(coef)	z	Pr(>|z|)	C-index
This study	-0.659	0.518	0.158	-4.182	**2.89E-05**	**0.601**
Lossos *et al.*	-0.646	0.524	0.158	-4.099	4.16E-05	0.574
Rosenwald *et al.*	-0.459	0.632	0.158	-2.898	3.75E-03	0.559
Wright *et al.*	-0.606	0.545	0.156	-3.874	1.07E-04	0.570

Figures

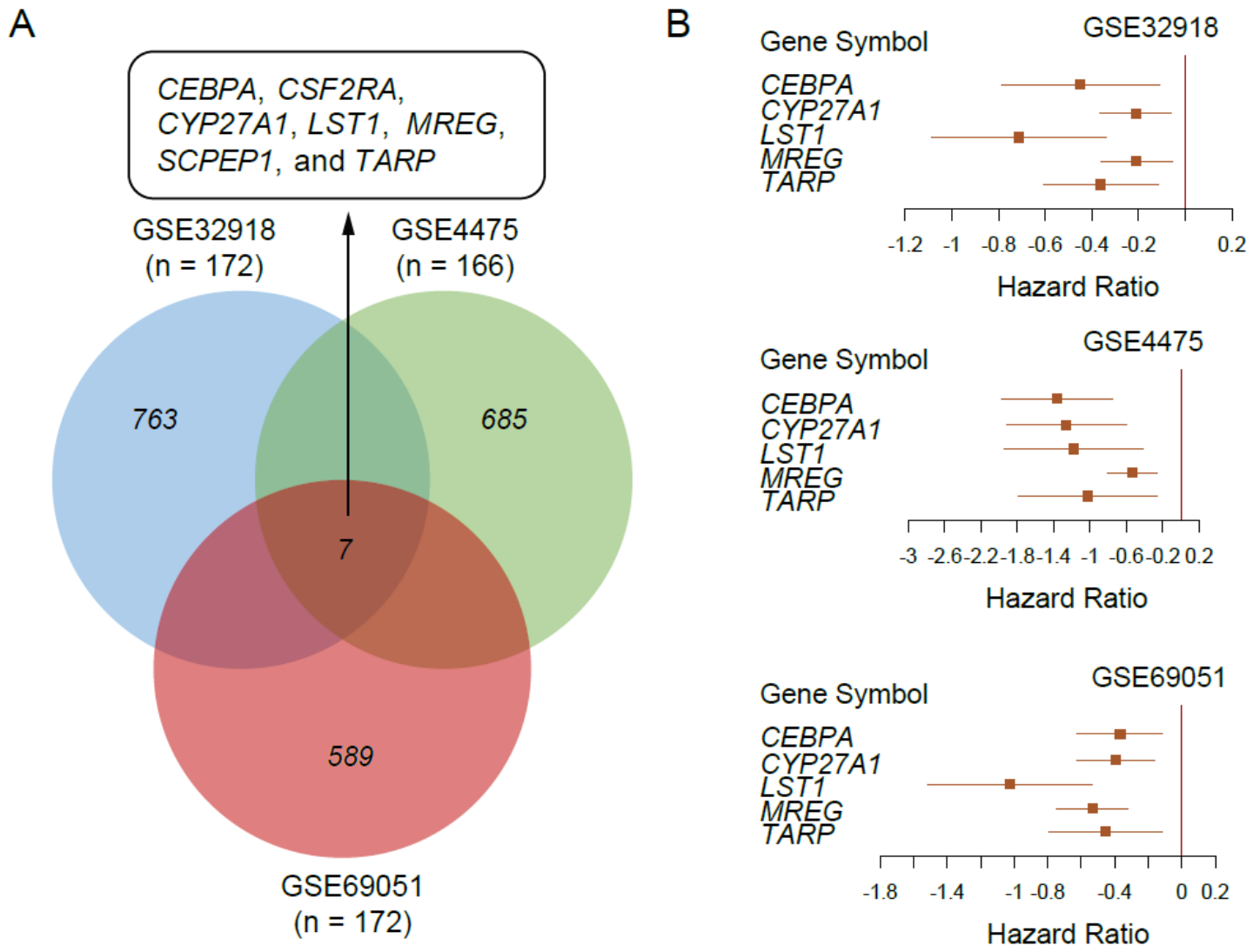

Figure 1

Screening a five-gene Cox model in the public DLBCL datasets from Gene Expression Omnibus (GEO). (A). Venn diagram summarizing the overlap between the prognostic genes identified by univariable Cox regression analysis in three public DLBCL datasets with accession numbers of GSE32918 (n = 172), GSE4475 (n = 166) and GSE69051 (n = 172). (B) The forest plots represent the association of the five gene signatures with overall survival in the three public DLBCL datasets.

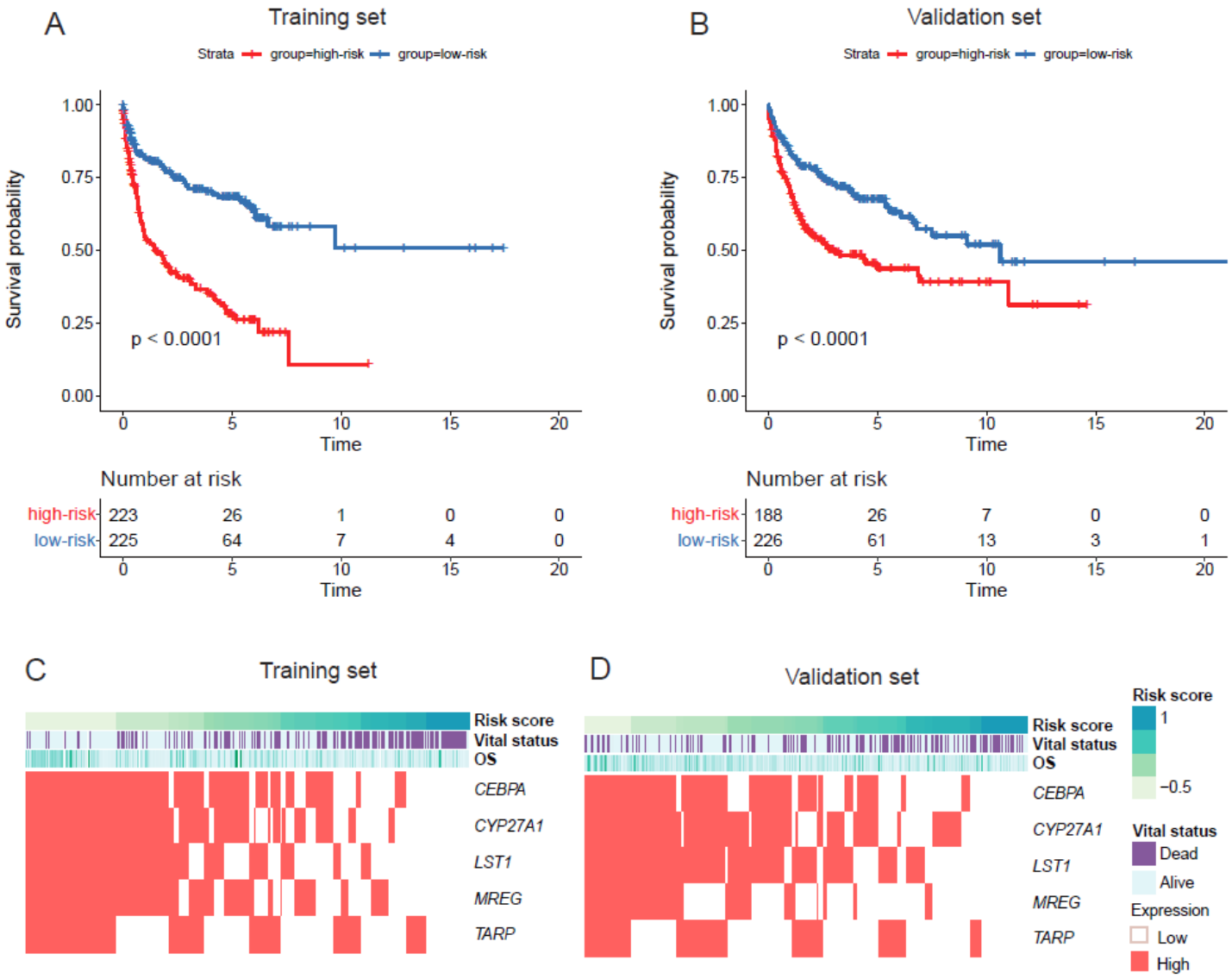

Figure 2

The performance of the five gene signatures in predicting the patients' risk. K-M curves for the prognostic model in the training datasets (A) and in the validation datasets (B). The expression patterns of the five prognostic genes in training (C) and validation (D) sets. The risk scores were estimated by the linear predictors of the Cox model.

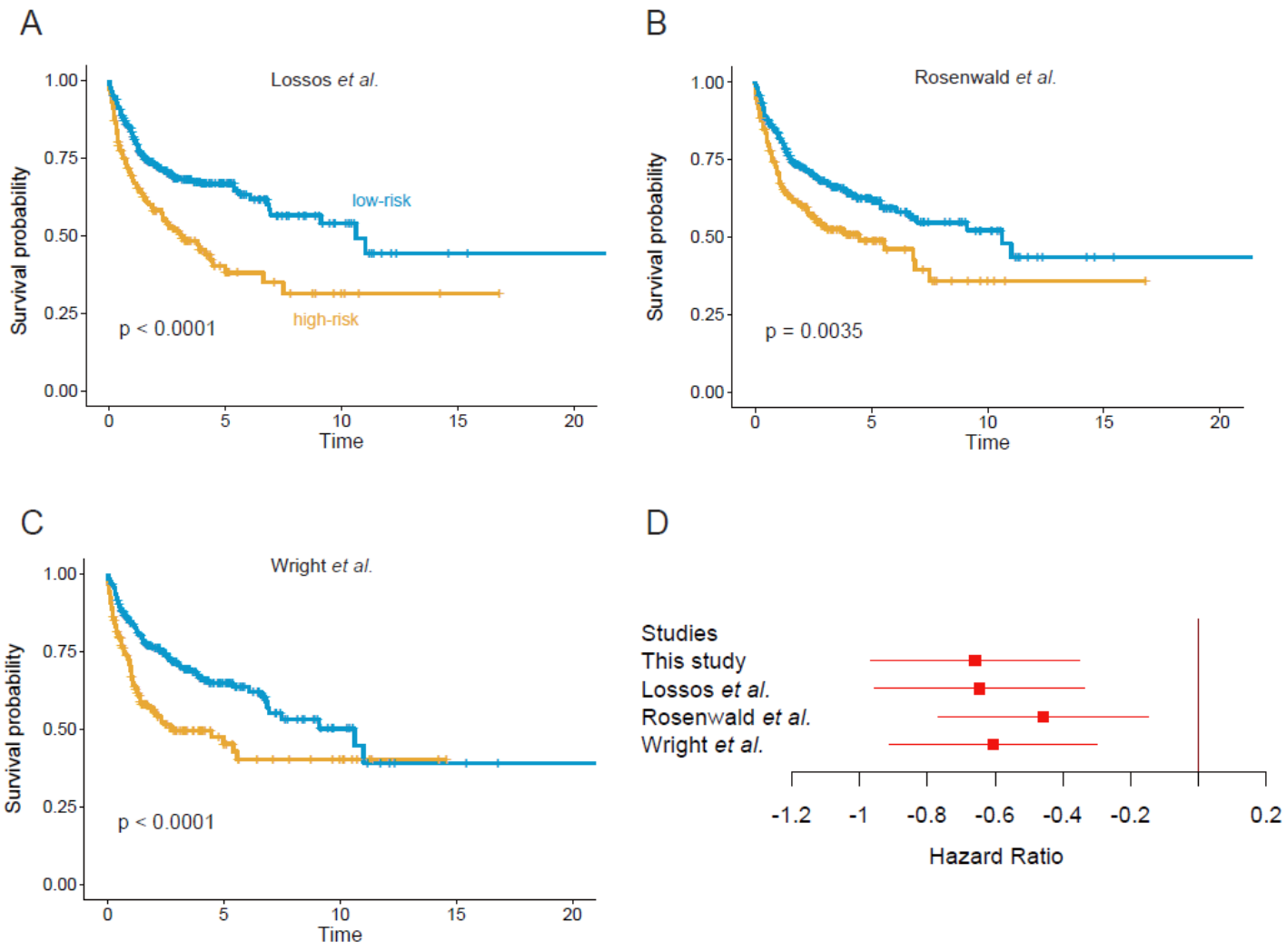

Figure 3

The Cox model based on the five gene signatures was superior to other models. (A-C) K-M curves for the three other prognostic models in the validation datasets. The K-M survival curves show the overall survival based on the relatively high- and low-risk patients divided by the optimal cut-off point from each model, respectively. P values were shown in the picture. (D) The performance of the four prognostic models in the validation datasets with accession numbers of GSE10846 (n = 414). The log2-hazard ratios and 95% confidence intervals were denoted by the red boxes and lines.

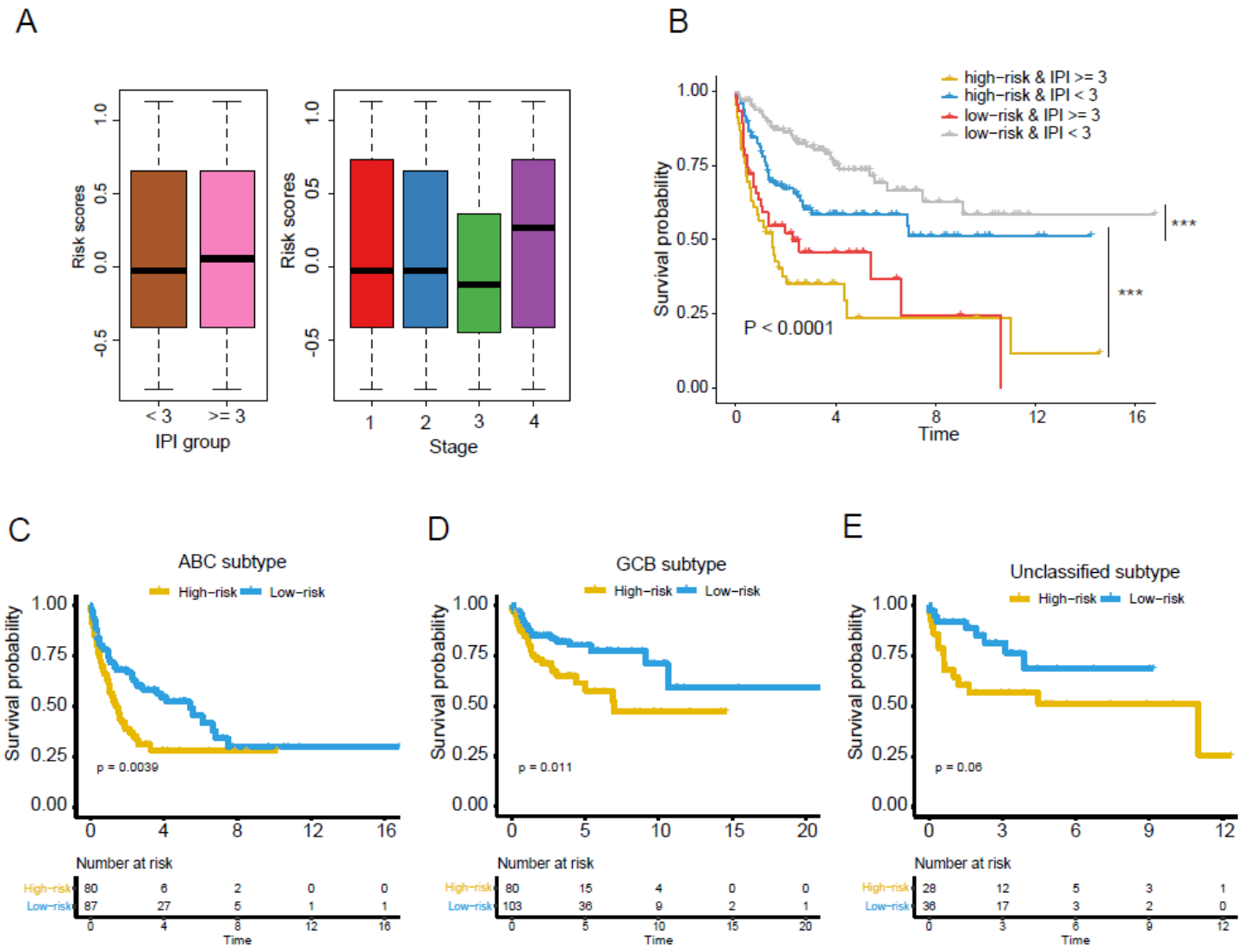

Figure 4

The risk stratification based on the five prognostic genes is independent of clinical factors. (A) The risk scores in different IPI groups (left panel) and clinical stages (right panel). The boxes show the median and the interquartile range (IQR) of the risk scores grouped by the IPI scoring system and clinical stage in the validation dataset. There are no significant differences between those groups (P > 0.05). (B) Kaplan-Meier survival curves show the overall survival of samples grouped by combining the IPI scoring system and the five-gene-based risk stratification. ***, P < 0.0001. The differences of overall survival between the high-risk and low-risk groups in specific subtype (C: ABC subtype; D: GCB subtype; E: unclassified subtype).

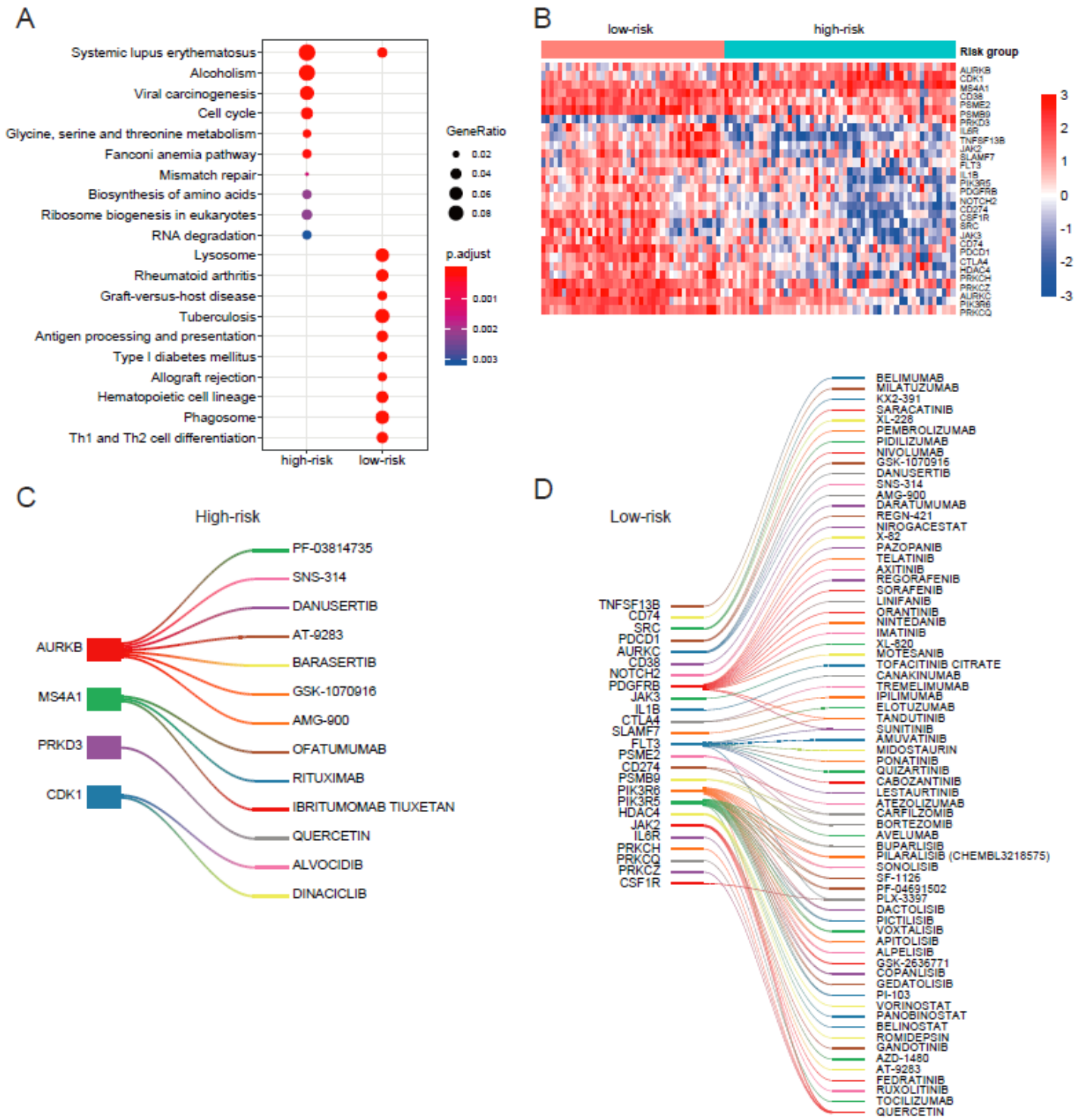

Figure 5

The molecular characteristics and potential drugs for the two risk groups (A) The top-ten GO terms enriched by the upregulated genes in high-risk and low-risk groups. The dots size and color represent the ratio of gene counts and statistical significance, respectively. (B)The heat map of the expression of 29 therapeutic targets genes among the DEGs, for the low-risk vs high-risk groups in the validation datasets. The DEGs were selected by Wilcoxon rank-sum test and fold change (Adjusted P-value < 0.05 and log2-

fold change > 0.5). (C)The upregulated cell cycle kinases and their potential drugs in high-risk group. (D) The upregulated immune checkpoint proteins and the corresponding drugs in the low-risk group.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DLBCLflowchart3.pdf