



Identification and validation of 5-methylcytosine-associated genes in diffuse large B-cell lymphoma

Cheng Xing^{a,b,c}, Shicong Zhu^d, Wenzhe Yan^{a,b,c}, Hongkai zhu^{a,b,c},
Zineng Huang^{a,b,c}, Yan Zhao^{a,b,c}, Wancheng Guo^{a,b,c}, Huifang Zhang^{a,b,c},
Le Yin^{a,b,c}, Xueqin Ruan^{a,b,c}, Zeyue Deng^{a,b,c}, Peilong Wang^{a,b,c}, Zhao Cheng^{a,b,c},
Zhihua Wang^{a,b,c,*}, Hongling Peng^{a,b,c,e,**}

^a Department of Hematology, The Second Xiangya Hospital, Central South University, Changsha, Hunan, China

^b Institute of Molecular Hematology, Central South University, Changsha, Hunan, China

^c Hunan Engineering Research Center of Cell Immunotherapy for Hematopoietic Malignancies, Changsha, Hunan, China

^d Department of Geriatrics, The Second Xiangya Hospital, Central South University, Changsha, Hunan, China

^e Hunan Key Laboratory of Tumor Models and Individualized Medicine, Changsha, Hunan, China

ARTICLE INFO

Keywords:

Diffuse large B-Cell lymphoma
5-Methylcytosine
Risk model
Prognosis
Biomarker

ABSTRACT

5-methylcytosine modifications play a significant role in carcinogenesis; however, studies exploring 5-methylcytosine-related genes in diffuse large B-cell lymphoma patients are lacking. In this study, we aimed to understand the potential role and clinical prognostic impact of 5-methylcytosine regulators in diffuse large B-cell lymphoma and identify a prognostic biomarker based on 5-methylcytosine-associated genes. Gene expression profiles and corresponding clinical information of diffuse large B-cell lymphoma patients and normal controls were obtained from The Cancer Genome Atlas, Gene Expression Omnibus, and Genotype-Tissue Expression databases. Diffuse large B-cell lymphoma was divided into three clusters according to the 5-methylcytosine regulators, and differentially expressed genes were screened among the three clusters. Univariate Cox and Lasso-Cox regression analyses were used to screen prognostic genes and construct a prognostic risk model. Kaplan-Meier curve analysis, univariate and multivariate Cox regression analyses, and time-dependent receiver operator characteristic curve analysis were used to evaluate prognostic factors. GSVA was used to enrich potential pathways associated with 5-methylcytosine modification patterns. SsGSEA and CIBERSORT were used to assess immune cell infiltration. Six 5-methylcytosine-related genes (*TUBB4A*, *CD3E*, *ZNF681*, *HAP1*, *IL22RA2*, and *POSTN*) were used to construct a prognostic risk model, which was proved to have a good predictive effect. In addition, univariate and multivariate Cox regression risk scores were independent prognostic factors for diffuse large B-cell lymphoma. Further analysis showed that the 5-methylcytosine risk score was significantly correlated with immune cell infiltration and immune checkpoint of diffuse large B-cell lymphoma. Our study reveals for the first time a potential role for 5-methylcytosine modifications in diffuse large B-cell lymphoma, provides novel insights for future studies on diffuse large B-cell lymphoma, and offers potential prognostic biomarkers and therapeutic targets for patients with diffuse large B-cell lymphoma.

* Corresponding author. Department of Hematology, The Second Xiangya Hospital, Central South University, Changsha, Hunan, China.

** Corresponding author. Department of Hematology, The Second Xiangya Hospital, Central South University, Changsha, Hunan, China.

E-mail addresses: wangzhihuazsc@csu.edu.cn (Z. Wang), penghongling@csu.edu.cn (H. Peng).

<https://doi.org/10.1016/j.heliyon.2023.e22209>

Received 13 March 2023; Received in revised form 1 November 2023; Accepted 6 November 2023

Available online 10 November 2023

2405-8440/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Diffuse large B-cell lymphoma (DLBCL), the most common type of non-Hodgkin's lymphoma, is a heterogeneous disease with different clinical manifestations, genetic characteristics, treatment responses, and prognoses [1]. Rituximab has been a breakthrough in the treatment of DLBCL, with over half of patients with DLBCL receiving the R-CHOP regimen, which includes cyclophosphamide, doxorubicin, vincristine, and prednisone plus rituximab; however, 30–40 % of patients still relapse or even fail to respond to R-CHOP therapy [2]. Over 50 % of chemotherapy-sensitive patients with relapsed/refractory DLBCL who are administered high-dose chemotherapy followed by autologous stem cell transplantation eventually relapse [3]. The advent of anti-CD19 chimeric antigen receptor (CAR) T-cell therapy has solved the treatment dilemma for some relapsed/refractory DLBCL patients. Although CAR T-cell therapies have achieved response rates in relapsed/refractory DLBCL patients, they are associated with significant toxicities in the form of cytokine-release syndrome and immune effector cell-associated neurotoxicity syndrome [4–6]. Due to these toxicities, the application of CAR T-cell therapies has been limited among elderly and unfit patients. Therefore, novel and more effective treatment strategies for patients with DLBCL are urgently required.

The advent of high-throughput sequencing technology has helped researchers gain a comprehensive understanding of the gene expression profile of DLBCL and identify many diagnostic, prognostic, and therapeutic biomarkers for DLBCL [7–9]. RNA modifications, including N6-methyladenosine (m6A), 5-methylcytosine (m5C), N1-methyladenosine (m1A), and pseudouridine (Ψ), play important roles in carcinogenesis [10,11]. Recently, the less studied m5C modification has received increasing attention, and growing evidence suggests that m5C regulates RNA stabilization, splicing, nuclear export, transcription, and translation, thereby mediating biological functions, such as cell proliferation, differentiation, apoptosis, and senescence [12,13]. Similar to m6A modifications, m5C RNA methylation is dynamically regulated by the corresponding m5C regulators and can be functionally classified into three isoforms -"writers", "erasers", and "readers". The RNA C5-cytosine methyltransferase NSUN2 is upregulated in several types of cancer including pancreatic cancer, nasopharyngeal carcinoma, uveal melanoma, and gastric cancer [14–16]. In gastric cancer, NSUN2 has been found to promote cancer cell proliferation, migration, and invasion [17]. Similarly, the enzymes NSUN4, NSUN5, NSUN6, and NSUN7 have been associated with the development of colorectal cancer, hepatocellular carcinoma, pancreatic cancer, and glioma [18–21]. Abnormal expression and mutations of the *TET* family and *ALKBH1* "erasers" are also associated with several malignancies [22–25]. In renal cell carcinoma, high expression of *ALKBH1* is correlated with malignant features of the tumor [26]. Furthermore, the m5C binding protein YBX1 has been found to maintain the stability of m5C-containing oncogenes and promote bladder cancer progression [27]. YBX1 has also been found to promote tumor progression in breast, pancreatic, and non-small cell lung cancers [28–30], and mediate resistance to the first-line chemotherapy drug sorafenib in hepatocellular carcinoma [31].

As far as we are aware, no studies have specifically investigated the potential of RNA m5C modification as a combined prognostic factor in DLBCL. The purpose of this research was to explore the clinical value of m5C regulators in DLBCL and to establish a prognostic biomarker for this condition based on m5C regulators, which has not been done before. Additionally, we examined the relationship between the risk of m5C modification and the distribution of immune cells, and found evidence indicating a potential correlation with immunotherapy. Our findings offer new insights for further research on m5C modifications and personalized treatment of DLBCL.

2. Materials and methods

2.1. Data collection

RNA sequencing data and clinical information (FPKM values) of patients with DLBCL and normal subjects were obtained from the Genotype-Tissue Expression (GTEx), The Cancer Genome Atlas (TCGA), and Gene Expression Omnibus (GEO) databases. We gathered 48 DLBCL sample datasets from TCGA database and 444 normal control datasets from the GTEx database, removed batch effects, merged the data, and normalized the data using $\log_2(\text{FPKM}+1)$ values. The GSE10846 and GSE181063 datasets were extracted from the GEO database, and \log_2 transformation was performed to normalize mRNA expression to eliminate batch effects for subsequent analyses. Finally, the training cohort ($n = 228$) and internal validation cohort ($n = 152$) were randomly (5:2) stratified from the 380 DLBCL samples in GSE10846. The external validation cohort consisted of 1310 DLBCL samples from the GSE181063 dataset to validate the prognostic value of m5C regulator signatures. The RCircos package in R was used to analyze and visualize the copy number variants (CNVs) derived from the TCGA database [32].

2.2. Collection and gene expression analysis of m5C regulators

Based on the relevant literature [10,33–37], we selected 22 m5C regulators for further analysis, excluding those that are not expressed in DLBCL, including eight "writers" (*DNMT3B*, *NSUN2*, *NSUN3*, *NSUN4*, *NSUN5*, *NSUN6*, *NSUN7*, and *NOP2*), four "erasers" (*TET1*, *TET2*, *TET3*, and *TDG*), and ten "readers" (*MBD1*, *MBD2*, *MBD3*, *MBD4*, *MECP2*, *NEIL1*, *NTHL1*, *SMUG1*, *UHRF1*, and *UHRF2*). We removed m5C regulators with very low expression (mean expression <0.5) in DLBCL samples and used the "limma" R package to compare the expression of the m5C regulators between different groups. Spearman correlation analysis of DLBCL samples from TCGA database was performed for differentially expressed m5C regulators using "corrplot" R package, and statistical significance was defined as p value < 0.05.

2.3. Consensus clustering analysis based on m5C regulators

According to the different expression patterns of 22 m5C regulators, DLBCL samples in the GSE10846 dataset were clustered into several subgroups by the “ConsensusClusterPlus” R package [38]. The “pca3d” and “rgl” R packages were used for principal component analysis (PCA) to evaluate sample clustering [39]. The association of different m5C subgroups with clinical information, such as age, gender, stage, molecular subtype, Eastern Cooperative Oncology Group (ECOG) performance status, and lactate dehydrogenase (LDH) ratio, was assessed using chi-square test. In addition, the Kaplan-Meier (K-M) overall survival curves for different subgroups were plotted by the R package “survival”. The “GSEABase”, “GSVA”, and “pheatmap” R packages were used for gene differential analysis to evaluate the enriched m5C-related pathways [40]. The gene set “c2_cp.kegg.v7.4_syms” was retrieved from the MSigDB database as the background, and statistical significance was defined as $|\log \text{ Fold change (FC)}| > 1$ and FDR value < 0.05 .

2.4. Construction and validation of a m5C prognostic risk score model in DLBCL

Differentially expressed genes (DEGs) between different clusters were screened with the empirical Bayesian approach using the “limma” R package ($|\log \text{ FC}| > 1$, FDR < 0.05) [41]. Moreover, we performed univariate Cox regression analysis to screen out candidate genes with independent prognostic value. Further, relevant genes were filtered again to construct an m5C risk prediction model with minimal risk of overfitting using the “glmnet” and “survival” R packages, together with Lasso-Cox regression analysis [42]. Finally, a risk score based on m5C modification features was constructed using the formula: Risk score = $\sum_{i=1}^n \text{Coefficient}_i * \text{Expression}_i$. The formula was applied to calculate risk scores for the training cohort, internal validation cohort, and external validation cohort and to divide DLBCL patients into low-risk and high-risk groups depending on their median m5C risk score. The survival difference between the two subgroups was compared using the “survivor” and “survminer” R packages. To confirm the predictive reliability of the model, receiver operating characteristic (ROC) curve analysis at 1-, 3-, and 5-years was performed using the “time-ROC” R package [43].

2.5. Establishment of the predictive nomogram

Data on clinical and pathological characteristics of patients, with respect to age, subtype, ECOG status, staging, LDH ratio, and risk score, were obtained from the GEO database. Univariate and multivariate Cox regression analyses were conducted to determine independent prognostic factors in patients with DLBCL. Using the “rms” R package, a nomogram was constructed based on risk scores and clinical variables significantly correlated with DLBCL prognosis, and the predictive prognostic ability for the nomogram was assessed using a calibration plot and ROC curve.

2.6. Immune analyses

Using the R packages “GSEABase” and “GSVA”, we performed single-sample gene set enrichment analysis (ssGSEA) [40] to evaluate the enrichment scores for 16 different classes of immune cells and 13 different immune-related pathway activities in the three m5C clusters of DLBCL. Further, the Cell Type Identification by Estimation of Relative Subtypes of RNA Transcripts (CIBERSORT) [44] algorithm was used to assess the relative proportions of 22 types of infiltrating immune cells to examine differences in immune cell subtypes among the m5C high-risk and low-risk groups. Subsequently, we calculated the correlation between risk scores and the enrichment fraction of immune cells, as well as the activity of immune-related pathways. Differences in the expression of potential immune checkpoint genes between the high-risk and low-risk groups were compared using the software package “ggpubr” R.

2.7. GSEA and functional enrichment analysis

Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) pathway analysis was applied to determine the potential molecular mechanisms associated with m5C risk [45]. Subsequently, GSEA was performed to screen for the most prominent enrichment pathways in the high-risk and low-risk groups ($|\log \text{ FC}| > 1$ and FDR < 0.05) [46].

2.8. RNA extraction and RT-PCR

Total RNA from DLBCL tissues and normal tissues was extracted by TRIzol reagent (TaKaRa, Shiga, Japan). Immediately afterwards, the *Evo M-MLV* RT Mixing Kit (Accurate Biotechnology (Hunan) Co., Ltd., China) was employed for cDNA synthesis, followed by the qRT-PCR analysis using the SYBR® Green Premix Pro Taq HS qPCR Kit (Accurate Biotechnology (Hunan) Co, Ltd., China). The mRNA levels between the experimental and control groups were normalized using β -actin. The sequences of the primers used are presented in Table S1.

2.9. Statistical analysis

All statistical analyses performed in our study were carried out using R software. Differences with $p < 0.05$ (*), 0.01 (**), and 0.001 (***) were considered to be statistically significant unless otherwise stated.

3. Results

3.1. Landscape and multi-omics analysis of m5C regulators in DLBCL

Apart from *DNMT3A*, *NSUN1*, *DNMT1*, *DNMT2*, *YBX1*, *ALYREF*, and *UNG*, which are genes with very low expression levels in DLBCL, a comparison of mRNA expression levels for 22 m5C regulators between DLBCL samples and normal tissues was performed using datasets obtained from TCGA and GTEx databases. Compared to normal controls, 19 of the 22 m5C regulators were differentially expressed in DLBCL samples. Among these genes, 13 m5C regulators (*DNMT3B*, *NSUN2*, *NSUN3*, *NSUN4*, *NSUN5*, *NSUN6*, *TET3*, *TDG*, *MBD2*, *MBD4*, *NTHL1*, *UHRF1*, and *NOP2*) were upregulated and six (*NSUN7*, *TET2*, *MBD1*, *MECP2*, *NEIL1*, and *UHRF2*) were downregulated (Fig. 1A). The 22 m5C regulators were significantly correlated with each other (Fig. 1B). Further CNV analysis revealed that alterations in CNV of genes were very frequent; among the m5C regulators, *MBD2*, *SMUG1*, *MECP2*, *NSUN5*, *UHRF2*, *NTHL1*, *MBD1*, *NSUN3*, *NSUN2*, *NSUN6*, and *DNMT3B* showed increased copy numbers, whereas the copy numbers of *TET3*, *TDG*, *NEIL1*, *NSUN4*, *MBD4*, *TET2*, *TET1*, *NOP2*, and *UHRF1* were reduced (Fig. 1C). Fig. 1D summarizes the categories, correlations, and prognosis of m5C regulators in TCGA-DLBCL cohort.

3.2. Depicting m5C clusters and clinical implications

Based on the expression levels of these m5C regulators, a consensus clustering analysis was conducted with the GSE10846 dataset. The clustering algorithm performed the best when $k = 3$, and the different clusters had significant prognostic values (Fig. 2A–C). PCA (Fig. 2D) indicated that the samples in clusters 1, 2, and 3 were significantly clustered, further demonstrating that the m5C-based DLBCL clusters were reliable. M5C regulators and their associated clinical and pathological features were correlated, as shown in

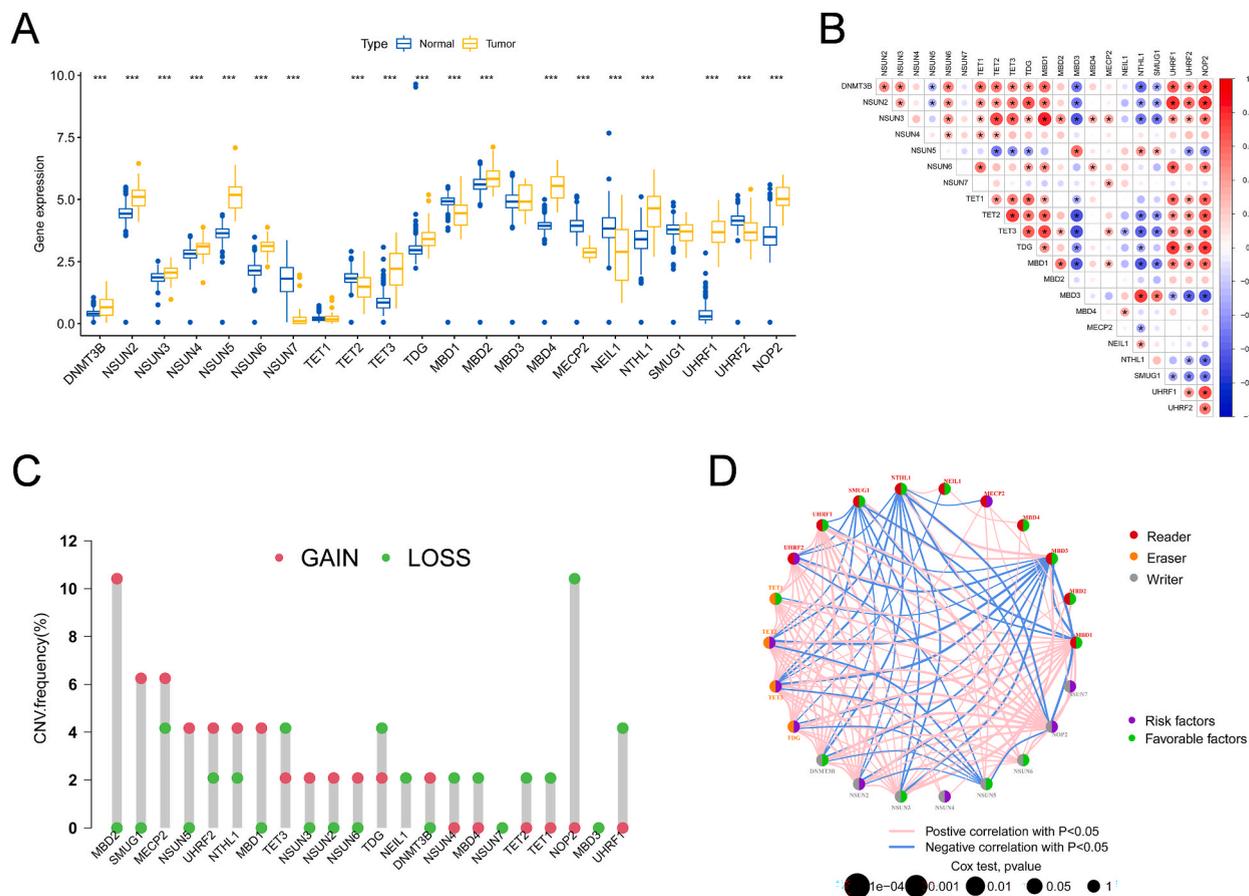


Fig. 1. Expression of 22 m5C regulators in DLBCL tissues and normal tissues. (A) Gene expression of m5C regulators in DLBCL tissues and normal tissues. $***p < 0.001$; $**p < 0.01$; $*p < 0.05$. (B) Correlation between the expression levels of m5C regulators in DLBCL. Darker shades of blue indicate a stronger negative correlation, and darker shades of red indicate a stronger positive correlation. * indicates $p < 0.05$. (C) CNV frequency of m5C regulators in TCGA-DLBCL cohort. Red dots: increased copy number; green dots: decreased copy number. (D) Gene network of m5C regulators. M5C, 5-methylcytosine; DLBCL, diffuse large B-cell lymphoma; CNV, copy number variation. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

the heat map, and the m5C regulators were clearly differentially expressed in the three clusters (Fig. 2F). Moreover, the difference in survival rates between DLBCL patients in clusters 1, 2, and 3 was found to be statistically significant using the K-M analysis, with patients in cluster 3 demonstrating a markedly better prognosis than those in clusters 1 and 2 ($p < 0.001$) (Fig. 2E).

Subsequently, we conducted a gene set variation enrichment analysis to determine the different biological processes and potential pathways of enrichment among the three types of m5C modification patterns (Fig. 3A–B, Fig. S1). Cluster 1 was significantly enriched in mismatch repair and DNA replication and homologous recombination; cluster 2 was enriched in glycosaminoglycan biosynthesis, glycosphingolipid biosynthesis, and glycerophospholipid metabolism; and cluster 3 was markedly enriched in limonene and pinene degradation, ascorbate and aldarate metabolism, and the interconversion of pentose and glucuronate. Furthermore, the immune status between the three m5C clusters in DLBCL was analyzed. SsGSEA revealed apparent differences in the enrichment fractions for most types of infiltrating immune cells (15 of the 16 groups) and 13 immune-related pathways between m5C clusters 1–3 (Fig. 3C–D). According to our research, it appears that the patterns associated with m5C could potentially have an impact on various biosynthetic and metabolic processes in DLBCL cells, as well as the immune microenvironment of DLBCL.

We analyzed DEGs among clusters 1, 2, and 3. Further studies found that a total of 116 DEGs were identified for the three clusters ($FDR < 0.05$ and $|\log FC| > 1$) (Fig. S2). Using GO analysis, we found that the DEGs were significantly enriched in interferon- γ -mediated signaling pathways, positive T cell selection, cell-cell junctions, adherens junctions, heart development, and dopamine-related pathways (Fig. 3E). In addition, the KEGG pathway enrichment analysis (Fig. 3F) revealed that these DEGs were involved in hematopoietic cell lineage, longevity regulating pathways, Th17 cell differentiation, systemic lupus erythematosus, pyrimidine metabolism, and acute myeloid leukemia. Overall, DEGs were associated with hematopoietic development, immune response, and tumor development and metastasis, and the associated dysfunction of m5C may affect the development of DLBCL and the immune microenvironment.

3.3. Identification of m5C regulator signatures in DLBCL

Using univariate Cox regression analysis, we identified 76 genes correlated with prognostic value from a total of 116 DEGs. Subsequently, as shown in Fig. 4A, LASSO regression analysis was performed on these genes to further define candidate genes with

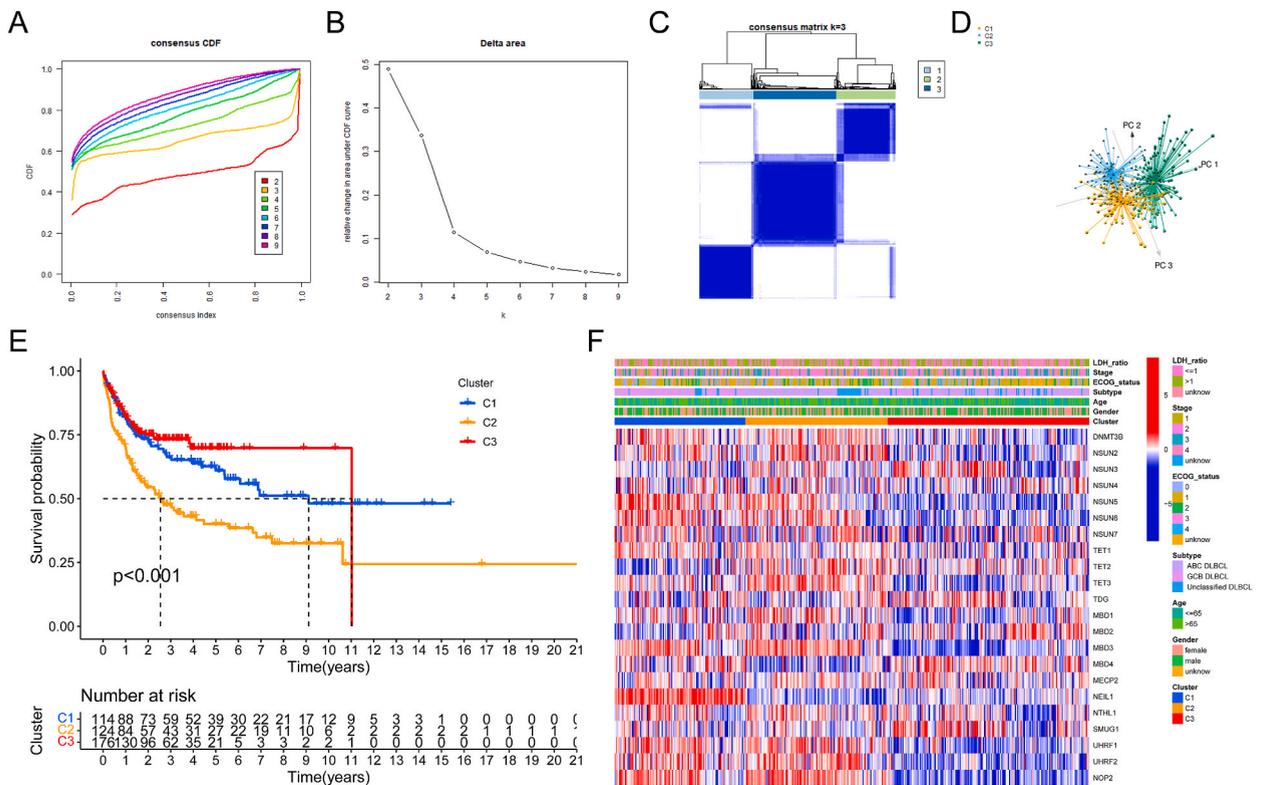


Fig. 2. Consensus clustering of m5C regulators. (A, B) Relative change in the area under the consensus clustering CDF and CDF curve when $k = 2-9$. (C) The consensus clustering matrix for $k = 3$. (D) Principal component analysis of three m5C clusters. (E) Heat map and clinical features of m5C regulator expression levels. Gender, age, LDH ratio, stage, ECOG status, and subtype were clinical parameters with statistically significant differences between the groups classified by m5C. (F) K-M survival analysis of the three m5C clusters. CDF, cumulative distribution function; m5C, 5-methylcytosine; DLBCL, diffuse large B-cell lymphoma; LDH, lactate dehydrogenase; ECOG, eastern cooperative oncology group; K-M, Kaplan-Meier.

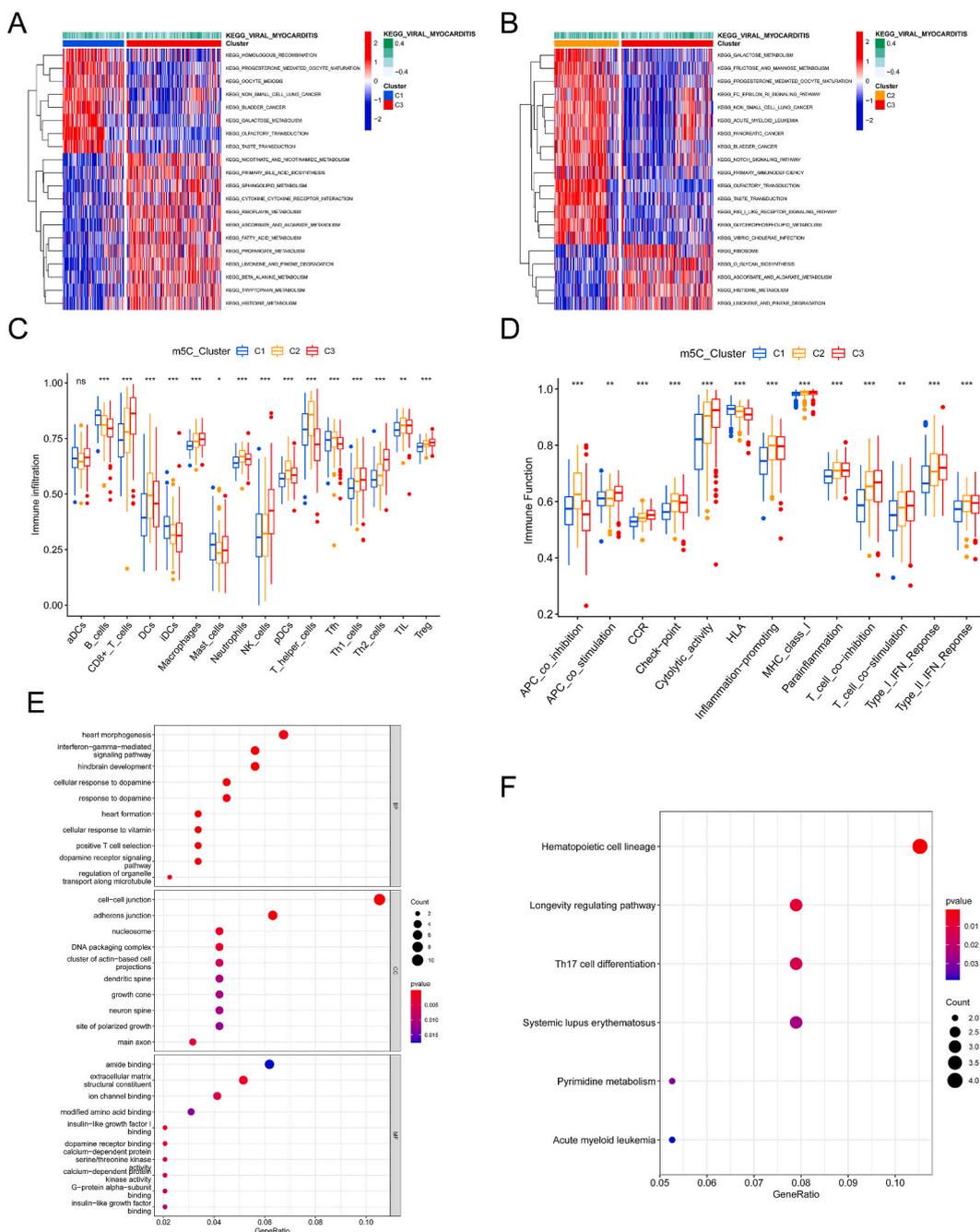


Fig. 3. Interactions and associations among clusters. (A, B) Gene set variation analysis of the different biological processes and potential pathways between C1 and C3 and between C2 and C3. Red: activated pathways; blue: inhibited pathways. (C, D) Sixteen infiltrating immune cell types (C) and 13 immune-related pathways (D) were assessed with ssGSEA in three m5C modification patterns. ****p* < 0.001; ***p* < 0.01; **p* < 0.05; ns, no significant difference. (E, F) Functional annotation of overlapping m5C phenotype-associated genes using Gene Ontology analysis (E) and KEGG enrichment methods (F). SsGSEA, gene set enrichment analysis; m5C, 5-methylcytosine; DLBCL, diffuse large B-cell lymphoma; KEGG, kyoto encyclopedia of genes and genomes. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

independent prognostic value. Finally, we carried out a multivariate Cox regression analysis to construct a model of prognostic risk using six m5C-related genes (*TUBB4A*, *CD3E*, *ZNF681*, *HAP1*, *IL22RA2*, and *POSTN*). Patients from the GEO database were then randomly separated into a training cohort and an internal validation cohort (3:2). Depending on gene expression levels and regression coefficients, a specific risk score was calculated for each individual with DLBCL in the training cohort: risk score = 0.357438413056158 × Expression of *TUBB4A* - 0.420001557858442 × Expression of *CD3E* - 0.177316474524244 × Expression of

$ZNF681 + 0.185246234028066 \times \text{Expression of } HAP1 - 0.371417012273356 \times \text{Expression of } IL22RA2 - 0.161281821509798 \times \text{Expression of } POSTN$. As illustrated in Fig. 4C, we divided the DLBCL patients from the training cohort equally into a low-risk group and a high-risk group based on their risk scores. Patients in the low-risk group had fewer deaths and longer survival than those in the high-risk group (Fig. 4D). The alluvial diagram revealed that the majority of DLBCL patients in cluster 3, who had the best prognosis,

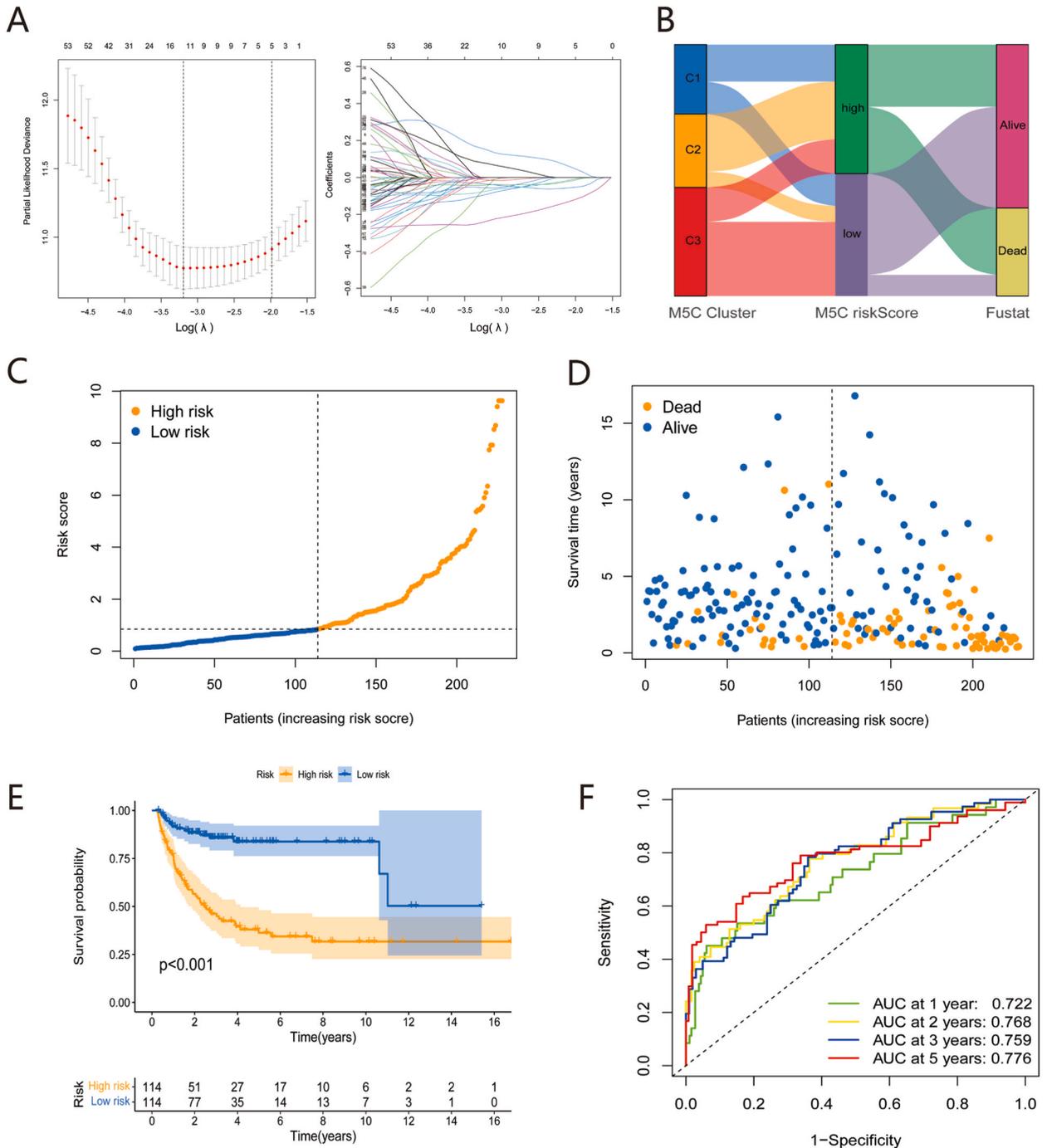


Fig. 4. Risk models for m5C-associated genes. (A) Lasso-Cox regression analysis. (B) Alluvial diagram illustrating the relationship between the three clusters, m5C risk, and patient survival status. (C) The model divides the patients in the training set into low-risk and high-risk groups. (D) Number of survivors and deaths in these high-risk and low-risk groups. (E) K-M curve of the high- and low-risk groups. The survival rate of patients in the low-risk group was higher than that in the high-risk group ($p < 0.001$). (F) ROC curves for 1-, 3-, and 5-year survival for the training set, with AUC values of 0.722, 0.759, and 0.776, respectively. M5C, 5-methylcytosine; K-M, Kaplan-Meier; ROC, receiver operator characteristic; AUC, area under the ROC curve.

were in the low-risk group, and the majority of patients in cluster 2, who had the worst prognosis, were in the high-risk group (Fig. 4B). Further, the K-M survival curve indicated that the survival probability of patients in the high-risk group was significantly lower than that of patients in the low-risk group ($p < 0.001$) (Fig. 4E). In the training cohort, the results showed that the areas under the ROC curves (AUC) were 0.722, 0.759, and 0.776 for the 1-year, 3-year, and 5-year survival, respectively (Fig. 4F). Prognostic analyses of subgroups with different clinical characteristics also indicated poor prognoses in the high-risk group (Fig. 5A–E). The above findings indicate that the m5C signatures in the training cohort exhibit a certain predictive power.

Next, the prognostic signatures of m5C were further validated in a training cohort using the internal validation cohort as well as another GEO dataset as an external validation cohort. Fig. 6A, D and 6B, E present the m5C risk scores and survival data for patients in the validation cohorts. Analysis of the K-M survival curve showed that the survival rate was higher in the low-risk patients than in the high-risk group, with a statistically significant difference ($p = 0.026, p < 0.001$) (Fig. 6C, F). In the internal validation cohort, the AUC values were 0.667 and 0.640 for the predicted 1-year and 3-year survival curves, respectively (Fig. 6G). Meanwhile, we found that the external validation cohort also exhibited superior reproducibility, with AUC values of 0.670, 0.662, and 0.648 for the predicted 1-, 3-, and 5-year survival rates, respectively (Fig. 6H); these findings indicated good prognosis prediction in both validation cohorts. Collectively, these findings confirm that the risk model can robustly and accurately predict the prognosis of DLBCL.

3.4. Pathological features and independent prognostic analysis of risk scores

To further assess whether the m5C signature is an independent prognostic factor for DLBCL, univariate and multivariate Cox regression analyses were performed on the GEO cohort. The univariate Cox regression analysis revealed that age (hazard ratio (HR) = 1.025, $p < 0.001$), subtype (HR = 0.558, $p < 0.001$), ECOG status (HR = 1.576, $p < 0.001$), stage (HR = 1.534, $p < 0.001$), LDH ratio (HR = 1.210, $p < 0.001$), and risk score (HR = 1.105, $p < 0.001$) were all strongly associated with the prognosis of DLBCL (Fig. 7A). Multivariate Cox regression analysis showed that age (HR = 1.030, $p < 0.001$), subtype (HR = 0.668, $p = 0.007$), ECOG status (HR = 1.310, $p < 0.013$), stage (HR = 1.408, $p < 0.001$), LDH ratio (HR = 1.117, $p = 0.016$), and risk score (HR = 1.100, $p < 0.001$) were all independent prognostic factors of DLBCL (Fig. 7B).

Subsequently, we established a prognostic nomogram by combining the risk score with all important clinical indicators. The predictive nomogram was constructed by selecting gender, ECOG status, LDH ratio, stage, subtype, age, and risk score from the GEO cohort (Fig. 7C). However, the calibration curve of the nomogram showed moderate agreement between the predicted and actual

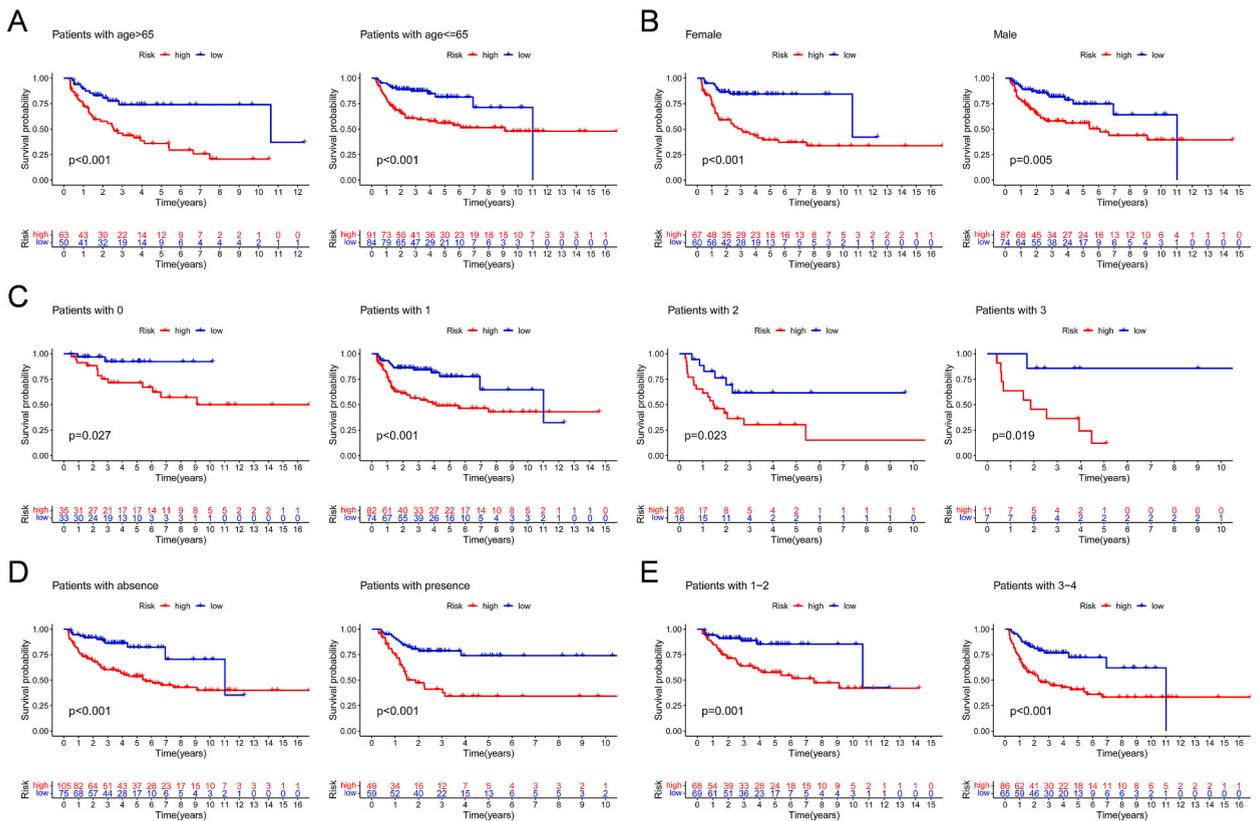


Fig. 5. K-M curves showing age (>65, ≤65) (A), gender (female, male) (B), ECOG status (0, 1, 2, 3) (C), extranodal involvement (absence, presence) (D), and stage (stage 1–2, 3–4) (E) for the high- and low-risk groups. K-M, Kaplan-Meier; ECOG, eastern cooperative oncology group.

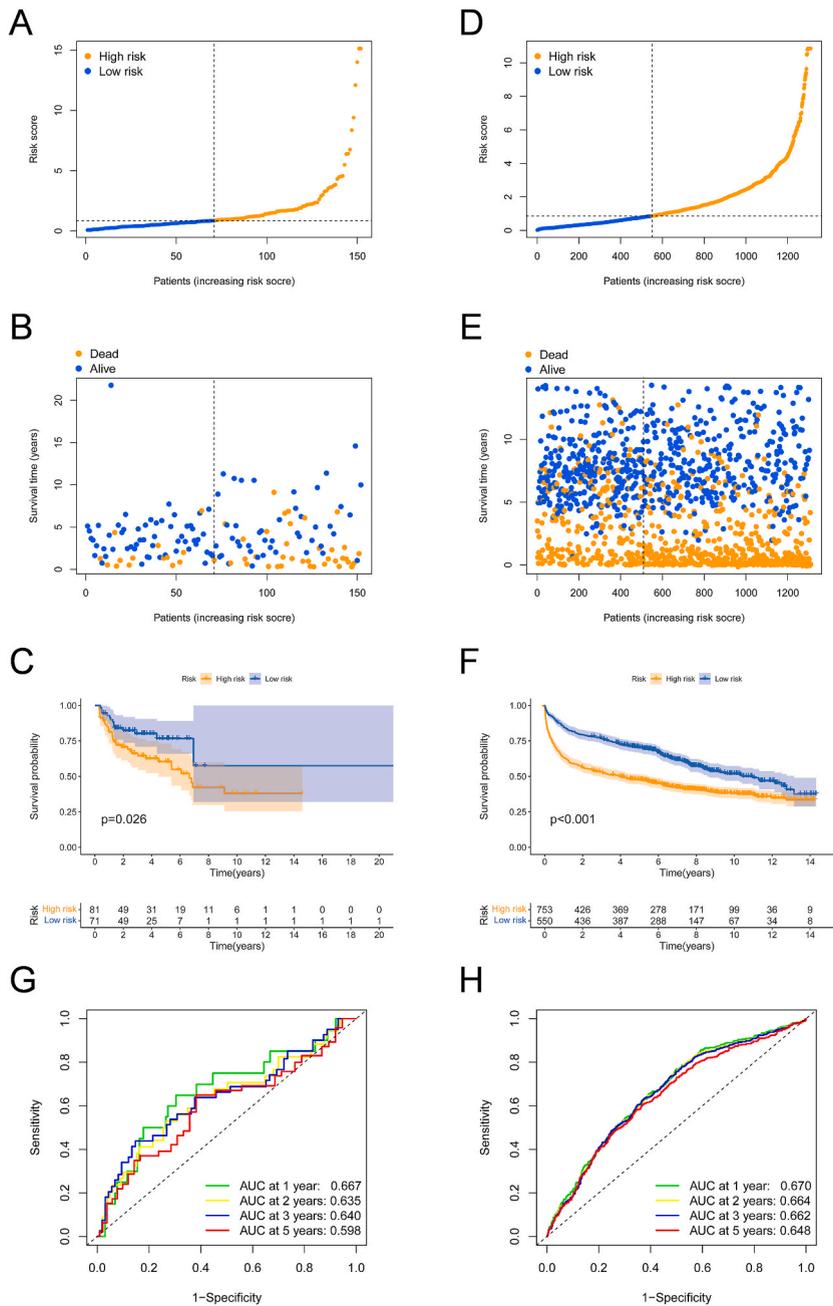


Fig. 6. Validation of the prognostic model. (A-C) Risk scores, survival, and K-M curves for the internal validation set. (D-F) Risk scores, survival, and K-M curves for the external validation set. (G, H) ROC curves of the internal and external validation sets. K-M, Kaplan-Meier; ROC, receiver operator characteristic.

survival (Fig. S3). Decision curve analysis and ROC curve with acceptable accuracy (Fig. 7D-E) indicated good predictive efficacy.

3.5. Estimation of m5C risk and tumor immune microenvironment characteristics

Immune cell infiltration was investigated in the two groups. CIBERSORT analysis indicated that the proportions of immunosuppressive cells, such as regulatory T cells, M2 macrophages, and resting natural killer cells, were increased in the high-risk group, whereas those of gamma delta T cells and resting mast cells were significantly decreased in the high-risk group (Fig. 8A). In addition, ssGSEA results indicated that the m5C risk score was negatively related to most immune cell types, including aDCs, CD8⁺ T cells, macrophages, natural killer cells, and tumor-infiltrating lymphocytes. Except APC co-inhibition, 12 immune-related pathways were

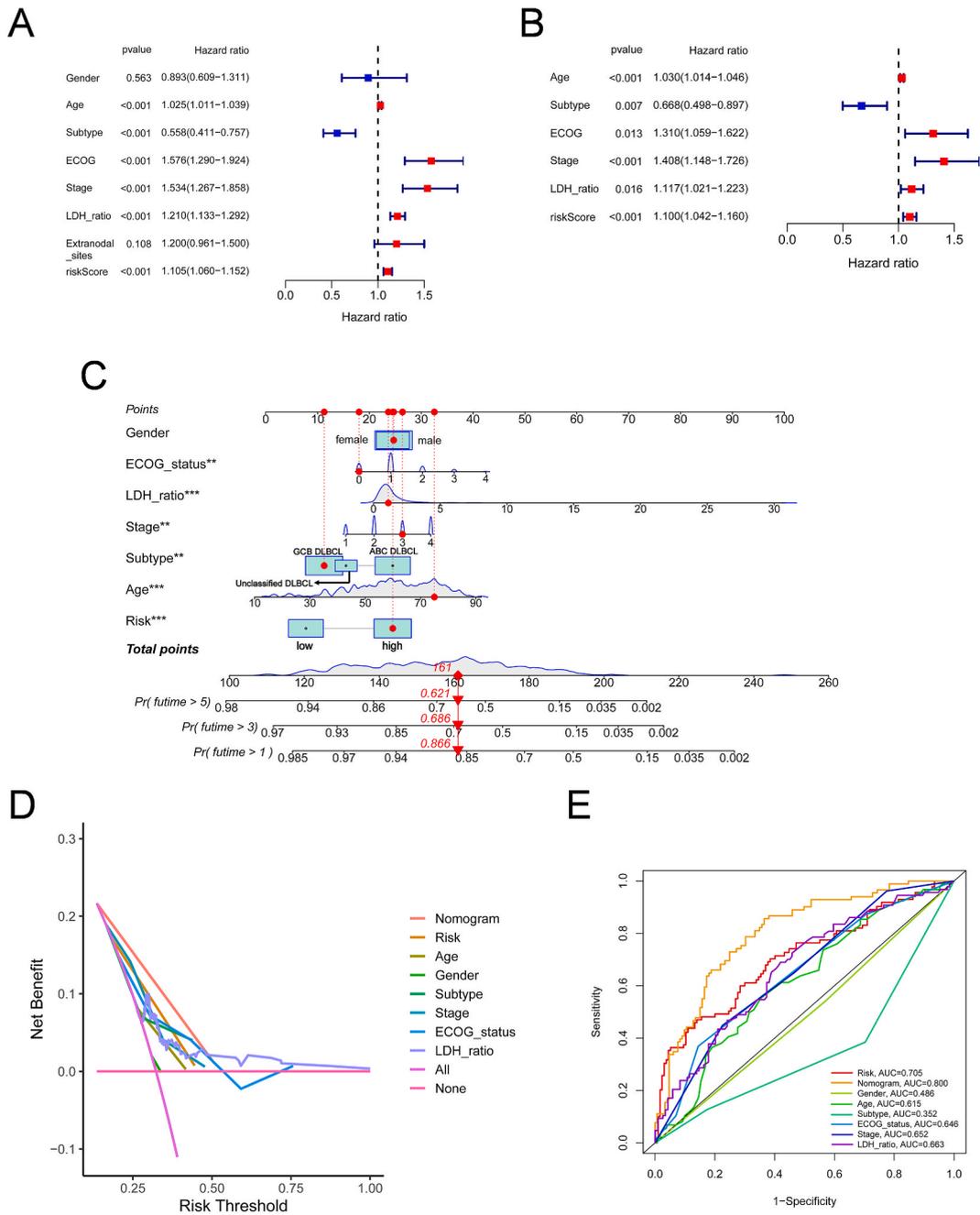


Fig. 7. (A) Univariate Cox regression analysis of the risk scores versus clinical factors. (B) Multivariate Cox regression analysis of risk scores and clinical factors. (C) Nomogram combining gender, ECOG status, LDH ratio, stage, subtype, age, and risk score to predict overall survival at 1, 3, and 5 years. (D, E) Decision curve analysis and ROC curves show good accuracy. LDH, lactate dehydrogenase; ECOG, eastern cooperative oncology group; ROC, receiver operator characteristic.

negatively correlated with the m5C risk score (Fig. 8B). These findings suggest that changes in immune cell infiltration induced by m5C risk may significantly affect the production of specific immune cell types, thereby affecting the response to immunotherapy.

Immune checkpoints are a major mechanism of tumor immune escape and a potential target for tumor immunotherapy. Consequently, we analyzed the expression levels of immune checkpoint genes in the two risk groups and identified 22 DEGs. The expression levels of *CD40*, *TNFRSF14*, *CD274*, *LGALS9*, *CD70*, *CD200R1*, *TNFSF9*, *TNFRSF18*, *PDCD1*, *TNFSF14*, *IDO2*, *TNFSF18*, and *HHLA2* in the high-risk group were obviously higher than those in the low-risk group, whereas the expression levels of *CD86*, *TIGIT*, *NRP1*, *CD28*, *ICOS*, *TNFRSF8*, *CD160*, *CD40LG*, and *PDCD1LG2* were the opposite (Fig. 8C). In addition, we observed that the risk score and expression levels of *CD274*, *LGALS9*, *CD70*, and *TNFSF9* were positively correlated ($R = 0.18, p = 0.00029$; $R = 0.29, p = 1.3e^{-8}$; $R =$

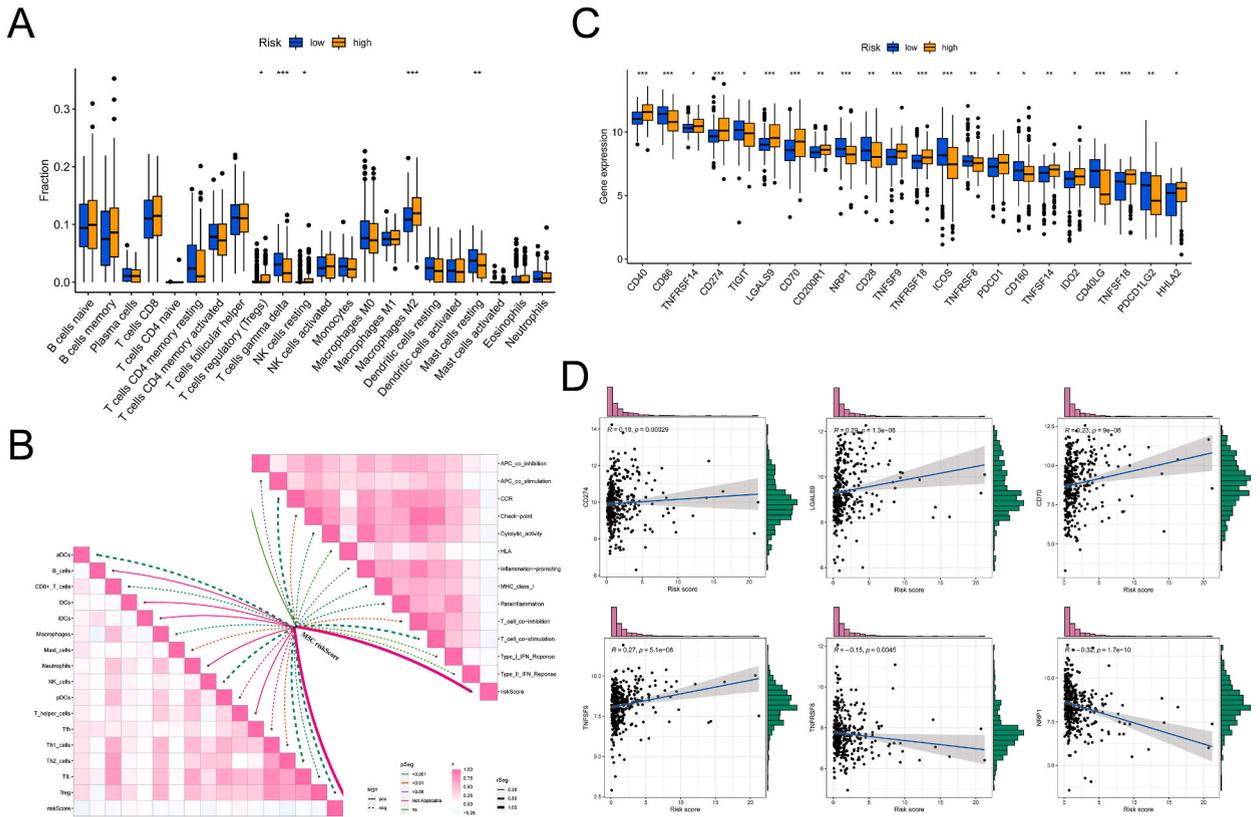


Fig. 8. Immune characteristics of different risk groups. (A) CIBERSORT analysis of the proportions of 22 immune-infiltrating cells in the high- and low-risk groups. (** $p < 0.001$; * $p < 0.01$; * $p < 0.05$). (B) Correlation of m5C risk with 16 types of immune cells and 13 immune-related pathways. Solid lines represent positive correlations, and dashed lines represent negative correlations. (C) Differential expression levels of immune checkpoints in different risk groups. (** $p < 0.001$; * $p < 0.01$; * $p < 0.05$). (D) Correlation between risk scores and expression levels of *CD274*, *LGALS9*, *CD70*, *TNFSF9*, *TNFRSF8*, and *NRP1*. CIBERSORT, cell-type identification by estimating relative subsets of RNA transcripts; m5C, 5-methylcytosine.

0.23, $p = 9e^{-6}$; $R = 0.27$, $p = 5.1e^{-8}$) and negatively correlated with *TNFRSF8* and *NRP1* ($R = -0.15$, $p = 0.0045$; $R = -0.32$, $p = 1.7e^{-10}$) (Fig. 8D).

3.6. Pathways enriched in the low- and high-risk groups

GSEA was performed to reveal potential differences in biological functions between the two risk groups of DLBCL. The top five

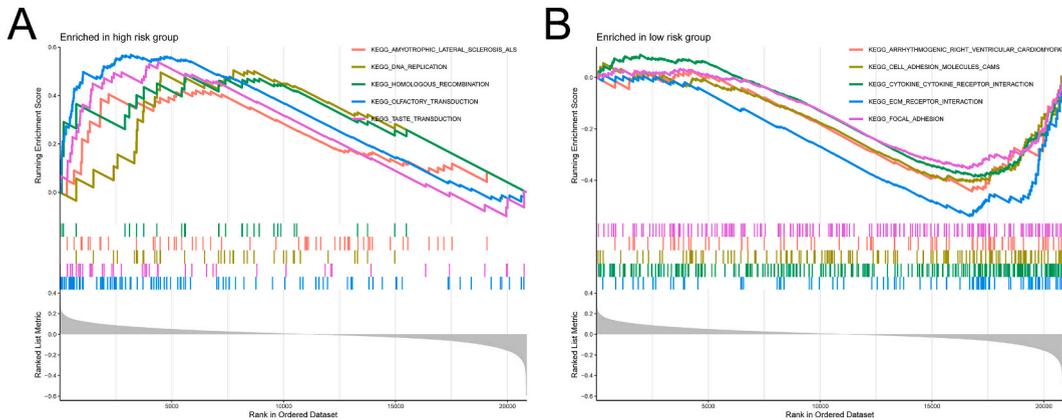


Fig. 9. Enrichment analysis of different risk subgroups. (A) Top five KEGG enriched pathways in the high-risk group. (B) Top five KEGG enriched pathways in the low-risk group. KEGG, kyoto encyclopedia of genes and genomes.

pathways significantly enriched in the low- and high-risk groups, respectively, were selected. The top five enriched pathways in the high-risk group included olfactory transduction, taste transduction, DNA replication, homologous recombination, and amyotrophic lateral sclerosis (Fig. 9A). The top five enriched pathways in the low-risk group included ECM receptor interaction, arrhythmogenic right ventricular cardiomyopathy, cell adhesion molecules, cytokine-cytokine receptor interaction, and focal adhesion (Fig. 9B).

3.7. Validation of signature-related genes in clinical samples

To further evaluate the expression of m5C-related genes, their mRNA expression levels were examined in DLBCL tissues and normal tissues. As shown in Fig. S4, *TUBB4A* expression in tumor tissue was obviously higher than that in normal tissue. In contrast, *CD3E*, *ZNF681*, and *POSTN* expression levels in tumor tissue were lower than those in normal tissue. Due to the large heterogeneity of DLBCL patients, *HAP1* and *IL22RA2* expression did not show significant statistical differences between tumor and normal tissues; however, the expression trends of *HAP1* and *IL22RA2* in tumor and normal tissues tended to be consistent with our predictions. Taken together, the expression of the six m5C-related genes in clinical tissue samples was basically consistent with that in public databases.

4. Discussion

Considering the complex phenotypes and genetic heterogeneity of DLBCL, subtype classification, prognosis prediction, and precision treatment are challenging tasks. With the advent of high-throughput sequencing, more than 170 RNA modifications have been discovered, particularly m6A, m5C, and m1A modifications [47]. Recently, more and more studies have found that m5C modifications are implicated in various biological processes, including DNA damage repair [48], nuclear-cytoplasmic shuttling, mRNA stability [49, 50], tumor development and progression, and metastasis [12,14,16]. Despite the rapid pace of research on the role of m5C in cancer, few studies have investigated the clinical significance of m5C regulators in DLBCL. In the present study, we assessed the prognostic value of 22 m5C regulators by dividing patients with DLBCL into three clusters and screening for candidate genes between the clusters. Then, a prognostic prediction model of m5C-related genes in patients with DLBCL was constructed and validated. The model could satisfactorily distinguish the prognosis of high- and low-risk patients with DLBCL. It is understood that this study is the first to use the m5C signature to predict the prognosis of DLBCL patients. Furthermore, we systematically explored the influence of m5C risk on the immune microenvironment.

We identified 22 m5C regulators expressed in patients with DLBCL. Most m5C regulators were differentially expressed in DLBCL and had complex regulatory networks. First, we performed a clustering consensus analysis based on m5C features and identified three molecular subtypes associated with m5C that exhibited significantly different prognoses, with cluster 3 presenting a significantly better prognosis than clusters 1 and 2. The DEGs among the three clusters were primarily enriched in hematopoietic development, immune response, and tumor development and metastatic pathways. This finding suggests that m5C may influence the development of DLBCL and the immune microenvironment. Interestingly, these DEGs are also highly enriched in cardiac development and dopamine-related pathways [51–54]. We thus hypothesized that m5C may have important effects on the development and progression of cardiovascular and neuropsychiatric diseases.

Our prognostic model involved six genes associated with m5C. According to the risk values of each gene, *TUBB4A* and *HAP1* were positively associated with the m5C risk in DLBCL patients, and *CD3E*, *ZNF681*, *IL22RA2*, and *POSTN* were negatively associated with the m5C risk. *TUBB4A* is a member of the β -microtubulin family; previous studies have shown it is oncogenic, with a high expression associated with aggressive prostate cancer development and metastasis [55]. In addition, *TUBB4A* mutations can lead to various diseases [56,57]. *CD3E* is associated with severe immunodeficiency [58], and the risk of head and neck squamous cell carcinoma recurrence is higher in patients with low *CD3E* expression levels [59]. *CD3E* is also associated with squamous cell carcinoma of the cervix, where high expression is associated with better prognosis [60]. Similarly, a bioinformatics study of patients with bladder cancer showed that patients with high expression levels of *CD3E* present a better prognosis [61]. Few studies have focused on the role of *ZNF681*; however, as a member of the zinc finger protein family, it has been speculated to be crucial in tumor development. Nevertheless, its exact role needs to be further elucidated. *IL22RA2* is an anti-tumor factor that induces the anti-tumor effects of lymphotoxin. It is expressed at reduced levels in colorectal tumors [62]. *HAP1* is primarily expressed in the nervous system, which is important to maintain neuronal survival [63]. Recently, *HAP1* was found to be associated with the regulation of gene transcription and vesicular transport [64]. In an *in vitro* study of breast cancer, *HAP1* was found to act as an oncogene, and overexpression of *HAP1* increased the radiosensitivity of breast cancer cells [65] and promoted apoptosis of tumor cells [66]. In addition, studies on the role of *HAP1* in acute lymphoblastic leukemia found that *HAP1* knockdown significantly reduced L-asparaginase-induced apoptosis [67]; this finding provides new avenues for developing more effective individualized therapies for patients with L-asparaginase-resistant acute lymphoblastic leukemia. However, the role of *HAP1* in DLBCL has not been reported. Furthermore, whether its role in DLBCL is a risk factor as we predicted needs further investigation. *POSTN* has been proven to stimulate tumor progression in various tumor cells [68,69]. Interestingly, *POSTN* was found to promote tumor invasiveness in prostate cancer, while it inhibited it in bladder cancer [70]; downregulation of *POSTN* was found to be significantly associated with high-grade bladder cancer [71]. Another research [72] demonstrated a biphasic effect of *POSTN* in pancreatic cancer development. *POSTN* role in tumor development and progression may be associated with tissue specificity. Gene expression profiling of the CD5 DLBCL subtype associated with poor prognosis revealed downregulation of *POSTN* expression [73]; these findings are similar to the prediction of the present study that *POSTN* expression may be associated with good prognosis in DLBCL. However, the mechanism underlying *POSTN* expression requires further investigation. In conclusion, the role of these six m5C-related genes in patients with DLBCL remains unclear, and more clinical data and experimental studies are required.

Surprisingly, our prognostic risk model based on m5C features could well predict the prognosis of DLBCL patients. The survival rate of DLBCL patients was significantly different between high-risk and low-risk groups. The model exhibited good prognostic prediction performance in the training as well as the internal and external validation sets. Notably, the risk score was identified as an independent prognostic factor. Importantly, we successfully constructed a model by combining m5C features and the prognostic nomogram of clinical parameters. ROC analysis further confirmed that the model had moderate accuracy in prognostic prediction, suggesting that the prognostic model based on m5C features has some clinical applicability.

Furthermore, the risk scores were negatively correlated with the abundances of most immune cells and immune-related pathways. In the high-risk group, there was an increase in immunosuppressive cells, such as regulatory T cells, M2 macrophages, and resting natural killer cells. The expression of most immune checkpoint-related genes (*CD40*, *TNFRSF14*, *CD274*, *LGALS9*, *CD70*, *CD200R1*, *TNFSF9*, *TNFRSF18*, *PDCD1*, *TNFSF14*, *IDO2*, *TNFSF18*, and *HHLA2*) was also upregulated in the high-risk group, suggesting that immunotherapy might be more effective in DLBCL patients at high risk of m5C features.

Although the present study provides new avenues for elucidating the pathogenesis of DLBCL and identifying prospective therapeutic targets for individualized therapy of patients, there are also some limitations. First, an m5C signature model was constructed to predict the prognosis of DLBCL, but more scientific studies are needed to evaluate the accuracy of the model. Second, our results are mainly derived from bioinformatics analysis, and more evidence is needed to demonstrate the exact molecular mechanisms underlying the role of the six m5C-related genes used to construct this model in the relationship between DLBCL development and the characteristics of the immune microenvironment.

In conclusion, we revealed the role of m5C modifications in DLBCL. A prognostic risk score model based on m5C-related genes was constructed to identify potential prognostic biomarkers for DLBCL. The m5C risk score has strong independent predictive power and can reliably forecast the prognosis of DLBCL. Our comprehensive analysis of m5C modifications provides new research directions for the study of DLBCL and contributes to the investigation of targets for the individualized treatment of DLBCL.

Data availability

The datasets utilized to claim the findings of our research are publicly available from the TCGA database, GTEx database and GEO database, GSE10846 and GSE181063.

Ethics statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Ethics Committee of the Second Xiangya Hospital of Central South University (protocol code: ethical review (2023) 059).

Funding statement

This work was supported by the Natural Science Foundation of Hunan Province, China [No. 2022JJ70058; No. 2022JJ30830; No. 2020JJ5840], the National Nature Science Foundation of China [No. 81670160; No. 82070175; No. 81600183; No. 81900170] and the Changsha Municipal Natural Science Foundation [No. Kq2202405].

CRediT authorship contribution statement

Cheng Xing: Writing – review & editing, Writing – original draft, Conceptualization. **Shicong Zhu:** Methodology. **Wenzhe Yan:** Methodology. **Hongkai zhu:** Data curation. **Zineng Huang:** Data curation. **Yan Zhao:** Data curation. **Wancheng Guo:** Data curation. **Huifang Zhang:** Data curation. **Le Yin:** Data curation. **Xueqin Ruan:** Data curation. **Zeyue Deng:** Formal analysis. **Peilong Wang:** Software. **Zhao Cheng:** Project administration. **Zhihua Wang:** Funding acquisition, Conceptualization. **Hongling Peng:** Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

m5C	5-methylcytosine
AUC	Area under the ROC curve
CIBERSORT	Cell-type identification by estimating relative subsets of RNA transcripts
CNV	Copy number variation
DEGs	Differentially expressed genes
DLBCL	Diffuse large B-cell lymphoma
ECOG	Eastern cooperative oncology group

GEO	Gene expression omnibus
GSEA	Gene set enrichment analysis
HR	Hazard ratio
ssGSEA	Gene set enrichment analysis
K-M	Kaplan-Meier
KEGG	Kyoto encyclopedia of genes and genomes
LDH	Lactate dehydrogenase
m1A	N1-methyladenosine
m6A	N6-methyladenosine
ROC	Receiver operator characteristic
TCGA	The cancer genome atlas
GSVA	Gene set variation analysis

Appendix B. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.heliyon.2023.e22209>.

References

- [1] S. Li, K.H. Young, L.J. Medeiros, Diffuse large B-cell lymphoma, *Pathology* 50 (1) (2018) 74–87.
- [2] A. Patriarca, G. Gaidano, Investigational drugs for the treatment of diffuse large B-cell lymphoma, *Expert Opin. Invest. Drugs* 30 (1) (2021) 25–38.
- [3] P. Skrabek, S. Assouline, A. Christofides, D. MacDonald, A. Prica, R. Sangha, B.A. Matthews, L.H. Sehn, Emerging therapies for the treatment of relapsed or refractory diffuse large B cell lymphoma 26 (4) (2019) 253–265.
- [4] S.S. Neelapu, F.L. Locke, N.L. Bartlett, L.J. Lekakis, D.B. Miklos, C.A. Jacobson, I. Braunschweig, O.O. Oluwole, T. Siddiqi, Y. Lin, J.M. Timmerman, P.J. Stiff, J. W. Friedberg, I.W. Flinn, A. Goy, B.T. Hill, M.R. Smith, A. Deol, U. Farooq, P. McSweeney, J. Munoz, I. Avivi, J.E. Castro, J.R. Westin, J.C. Chavez, A. Ghobadi, K.V. Komanduri, R. Levy, E.D. Jacobsen, T.E. Witzig, P. Reagan, A. Bot, J. Rossi, L. Navale, Y. Jiang, J. Aycok, M. Elias, D. Chang, J. Wieszorek, W.Y. Go, Axicabtagene ciloleucel CAR T-cell therapy in refractory large B-cell lymphoma, *N. Engl. J. Med.* 377 (26) (2017) 2531–2544.
- [5] S.J. Schuster, M.R. Bishop, C.S. Tam, E.K. Waller, P. Borchmann, J.P. McGuirk, U. Jäger, S. Jaglowski, C. Andreadis, J.R. Westin, I. Fleury, V. Bachanova, S. R. Foley, P.J. Ho, S. Mielke, J.M. Magenu, H. Holte, S. Pantano, L.B. Pacaud, R. Awasthi, J. Chu, Ö. Anak, G. Salles, R.T. Maziarz, Tisagenlecleucel in adult relapsed or refractory diffuse large B-cell lymphoma, *N. Engl. J. Med.* 380 (1) (2019) 45–56.
- [6] J.S. Abramson, M.L. Palomba, L.I. Gordon, M.A. Lunning, M. Wang, J. Arnason, A. Mehta, E. Purev, D.G. Maloney, C. Andreadis, A. Sehgal, S.R. Solomon, N. Ghosh, T.M. Albertson, J. Garcia, A. Kostic, M. Mallaney, K. Ogasawara, K. Newhall, Y. Kim, D. Li, T. Siddiqi, Lisocabtagene maraleucel for patients with relapsed or refractory large B-cell lymphomas (TRANSCEND NHL 001): a multicentre seamless design study, *Lancet* 396 (10254) (2020) 839–852.
- [7] C. Zhou, Y. Cui, H. Sun, F. Yang, H. Zhao, L. Huangfu, J. Zhang, Identification of key mutations in central nervous diffuse large B-cell lymphoma (DLBCL) by comprehensive analysis between sequencing and TCGA database, *Transl. Cancer Res.* 10 (6) (2021) 2632–2642.
- [8] A. Reddy, J. Zhang, N.S. Davis, A.B. Moffitt, C.L. Love, A. Waldrop, S. Leppa, A. Pasanen, L. Meriranta, M.L. Karjalainen-Lindsberg, P. Nørgaard, M. Pedersen, A. O. Gang, E. Høgdall, T.B. Heavican, W. Lone, J. Iqbal, Q. Qin, G. Li, S.Y. Kim, J. Healy, K.L. Richards, Y. Fedoriw, L. Bernal-Mizrachi, J.L. Koff, A.D. Staton, C. R. Flowers, O. Paltiel, N. Goldschmidt, M. Calaminici, A. Clear, J. Gribben, E. Nguyen, M.B. Czader, S.L. Ondrejka, A. Collie, E.D. Hsi, E. Tse, R.K.H. Au-Yeung, Y.L. Kwong, G. Srivastava, W.W.L. Choi, A.M. Evens, M. Pilichowska, M. Sengar, N. Reddy, S. Li, A. Chadburn, L.I. Gordon, E.S. Jaffe, S. Levy, R. Rempel, T. Tzeng, L.E. Happ, T. Dave, D. Rajagopalan, J. Datta, D.B. Dunson, S.S. Dave, Genetic and functional drivers of diffuse large B cell lymphoma, *Cell* 171 (2) (2017) 481–494.e15.
- [9] F. Marchesi, G. Regazzo, F. Palombi, I. Terrenato, A. Sacconi, M. Spagnuolo, S. Donzelli, M. Marino, C. Ercolani, A. Di Benedetto, G. Blandino, G. Ciliberto, A. Mengarelli, M.G. Rizzo, Serum miR-22 as potential non-invasive predictor of poor clinical outcome in newly diagnosed, uniformly treated patients with diffuse large B-cell lymphoma: an explorative pilot study, *J. Exp. Clin. Cancer Res.* : CR 37 (1) (2018) 95.
- [10] P. Nombela, B. Miguel-López, S. Blanco, The role of m(6)A, m(5)C and Ψ RNA modifications in cancer: novel therapeutic opportunities, *Mol. Cancer* 20 (1) (2021) 18.
- [11] D. Wiener, S. Schwartz, The epitranscriptome beyond m(6)A, *Nat. Rev. Genet.* 22 (2) (2021) 119–131.
- [12] M. Li, Z. Tao, Y. Zhao, L. Li, J. Zheng, Z. Li, X. Chen, 5-methylcytosine RNA methyltransferases and their potential roles in cancer, *J. Transl. Med.* 20 (1) (2022) 214.
- [13] H. Song, J. Zhang, B. Liu, J. Xu, B. Cai, H. Yang, J. Straube, X. Yu, T. Ma, Biological roles of RNA m(5)C modification and its implications in Cancer immunotherapy, *Biomark. Res.* 10 (1) (2022) 15.
- [14] S.Y. Chen, K.L. Chen, L.Y. Ding, C.H. Yu, H.Y. Wu, Y.Y. Chou, C.J. Chang, C.H. Chang, Y.N. Wu, S.R. Wu, Y.C. Hou, C.T. Lee, P.C. Chen, Y.S. Shan, P.H. Huang, RNA bisulfite sequencing reveals NSUN2-mediated suppression of epithelial differentiation in pancreatic cancer, *Oncogene* 41 (22) (2022) 3162–3176.
- [15] X. Tong, Y. Xiang, Y. Hu, Y. Hu, H. Li, H. Wang, K.N. Zhao, X. Xue, S. Zhu, NSUN2 promotes tumor progression and regulates immune infiltration in nasopharyngeal carcinoma, *Front. Oncol.* 12 (2022), 788801.
- [16] G. Luo, W. Xu, X. Chen, S. Wang, J. Wang, F. Dong, D.N. Hu, P.S. Reinach, D. Yan, NSUN2-mediated RNA m(5)C modification modulates uveal melanoma cell proliferation and migration, *Epigenetics* (2022) 1–12.
- [17] Y. Hu, C. Chen, X. Tong, S. Chen, X. Hu, B. Pan, X. Sun, Z. Chen, X. Shi, Y. Hu, X. Shen, X. Xue, M. Lu, NSUN2 modified by SUMO-2/3 promotes gastric cancer progression and regulates mRNA m5C methylation, *Cell Death Dis.* 12 (9) (2021) 842.
- [18] Z. Jiang, S. Li, M.J. Han, G.M. Hu, P. Cheng, High expression of NSUN5 promotes cell proliferation via cell cycle regulation in colorectal cancer, *Am. J. Transl. Res.* 12 (7) (2020) 3858–3870.
- [19] R. Yang, X. Liang, H. Wang, M. Guo, H. Shen, Y. Shi, Q. Liu, Y. Sun, L. Yang, M. Zhan, The RNA methyltransferase NSUN6 suppresses pancreatic cancer development by regulating cell proliferation, *EBioMedicine* 63 (2021), 103195.
- [20] K. Sato, K. Tahata, K. Akimoto, Five genes associated with survival in patients with lower-grade gliomas were identified by information-theoretical analysis, *Anticancer Res.* 40 (5) (2020) 2777–2785.
- [21] Y. He, X. Yu, J. Li, Q. Zhang, Q. Zheng, W. Guo, Role of m(5)C-related regulatory genes in the diagnosis and prognosis of hepatocellular carcinoma, *Am. J. Transl. Res.* 12 (3) (2020) 912–922.
- [22] N.P. Archer, V. Perez-Andreu, M.E. Scheurer, K.R. Rabin, E.C. Peckham-Gregory, S.E. Plon, R.C. Zabriskie, P.A. De Alarcon, K.S. Fernandez, C.R. Najera, J. J. Yang, F. Antillon-Kluschmann, P.J. Lupo, Family-based exome-wide assessment of maternal genetic effects on susceptibility to childhood B-cell acute lymphoblastic leukemia in hispanics, *Cancer* 122 (23) (2016) 3697–3704.

- [23] L. Wang, X. Feng, Z. Jiao, J. Gan, Q. Meng, Characterization of the prognostic and diagnostic values of ALKBH family members in non-small cell lung cancer, *Pathol. Res. Pract.* 231 (2022), 153809.
- [24] M.G. García, A. Carella, R.G. Urduñuño, G.F. Bayón, V. Lopez, J.R. Tejedor, M.I. Sierra, E. García-Torano, P. Santamarina, R.F. Perez, C. Mangas, A. Astudillo, M. D. Corte-Torres, I. Sáenz-de-Santa-María, M.D. Chiara, A.F. Fernández, M.F. Fraga, Epigenetic dysregulation of TET2 in human glioblastoma, *Oncotarget* 9 (40) (2018) 25922–25934.
- [25] H. Takai, K. Masuda, T. Sato, Y. Sakaguchi, T. Suzuki, T. Suzuki, R. Koyama-Nasu, Y. Nasu-Nishimura, Y. Katou, H. Ogawa, Y. Morishita, H. Kozuka-Hata, M. Oyama, T. Todo, Y. Ino, A. Mukasa, N. Saito, C. Toyoshima, K. Shirahige, T. Akiyama, 5-Hydroxymethylcytosine plays a critical role in glioblastomagenesis by recruiting the CHTOP-methylosome complex, *Cell Rep.* 9 (1) (2014) 48–60.
- [26] L. Li, C. Zhu, Q. Xu, S. Xu, J. Ye, D. Xu, L. Wang, S. Gan, B. Liu, C. Tang, ALKBH1 contributes to renal cell carcinoma progression by reducing N6-methyladenine of GPR137, *Eur. J. Clin. Invest.* 53 (7) (2023), e13986.
- [27] X. Chen, A. Li, B.F. Sun, Y. Yang, Y.N. Han, X. Yuan, R.X. Chen, W.S. Wei, Y. Liu, C.C. Gao, Y.S. Chen, M. Zhang, X.D. Ma, Z.W. Liu, J.H. Luo, C. Lyu, H.L. Wang, J. Ma, Y.L. Zhao, F.J. Zhou, Y. Huang, D. Xie, Y.G. Yang, 5-methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs, *Nat. Cell Biol.* 21 (8) (2019) 978–990.
- [28] Y. Zhang, Y.X. Huang, D.L. Wang, B. Yang, H.Y. Yan, L.H. Lin, Y. Li, J. Chen, L.M. Xie, Y.S. Huang, J.Y. Liao, K.S. Hu, J.H. He, P.E. Saw, X. Xu, D. Yin, LncRNA DSCAM-AS1 interacts with YBX1 to promote cancer progression by forming a positive feedback loop that activates FOXA1 transcription network, *Theranostics* 10 (23) (2020) 10823–10837.
- [29] H. Zhang, H. Yu, D. Ren, Y. Sun, F. Guo, H. Cai, C. Zhou, Y. Zhou, X. Jin, H. Wu, CBX3 regulated by YBX1 promotes smoking-induced pancreatic cancer progression via inhibiting SMURF2 expression, *Int. J. Biol. Sci.* 18 (8) (2022) 3484–3497.
- [30] H. Su, G. Fan, J. Huang, X. Qiu, LncRNA HOXC-AS3 promotes non-small-cell lung cancer growth and metastasis through upregulation of YBX1, *Cell Death Dis.* 13 (4) (2022) 307.
- [31] J. Xu, L. Ji, Y. Liang, Z. Wan, W. Zheng, X. Song, K. Gorshkov, Q. Sun, H. Lin, X. Zheng, J. Chen, R.A. Jin, X. Liang, X. Cai, CircRNA-SORE mediates sorafenib resistance in hepatocellular carcinoma by stabilizing YBX1, *Signal Transduct. Targeted Ther.* 5 (1) (2020) 298.
- [32] C.H. Mermel, S.E. Schumacher, B. Hill, M.L. Meyerson, R. Beroukhi, G. Getz, GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers, *Genome Biol.* 12 (4) (2011) R41.
- [33] Z. Xu, S. Chen, Y. Zhang, R. Liu, M. Chen, Roles of m5C RNA modification patterns in biochemical recurrence and tumor microenvironment characterization of prostate adenocarcinoma, *Front. Immunol.* 13 (2022), 869759.
- [34] Y.-T. Chen, J.-Y. Shen, D.-P. Chen, C.-F. Wu, R. Guo, P.-P. Zhang, J.-W. Lv, W.-F. Li, Z.-X. Wang, Y.-P. Chen, Identification of cross-talk between m6A and 5mC regulators associated with onco-immunogenic features and prognosis across 33 cancer types, *J. Hematol. Oncol.* 13 (1) (2020) 22.
- [35] Y.-S. Chen, W.-L. Yang, Y.-L. Zhao, Y.-G. Yang, Dynamic transcriptomic m5C and its regulatory role in RNA processing 12 (4) (2021) e1639.
- [36] R. García-Vilchez, A. Sevilla, S. Blanco, Post-transcriptional regulation by cytosine-5 methylation of RNA, *Biochim. Biophys. Acta. Gene Regulatory Mech.* 1862 (3) (2019) 240–252.
- [37] H. Shi, P. Chai, R. Jia, X. Fan, Novel insight into the regulatory roles of diverse RNA modifications: Re-defining the bridge between transcription and translation, *Mol. Cancer* 19 (1) (2020) 78.
- [38] M.D. Wilkerson, D.N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking, *Bioinformatics* 26 (12) (2010) 1572–1573.
- [39] A. Abid, M.J. Zhang, V.K. Bagaria, J. Zou, Exploring patterns enriched in a dataset with contrastive principal component analysis, *Nat. Commun.* 9 (1) (2018) 2134.
- [40] S. Hänzelmann, R. Castelo, J. Guinney, GSEA: gene set variation analysis for microarray and RNA-seq data, *BMC Bioinf.* 14 (2013) 7.
- [41] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Res.* 43 (7) (2015) e47.
- [42] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for cox's proportional hazards model via coordinate descent, *J. Stat. Software* 39 (5) (2011) 1–13.
- [43] P. Blanche, J.F. Dartigues, H. Jacqmin-Gadda, Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks, *Stat. Med.* 32 (30) (2013) 5381–5397.
- [44] A.M. Newman, C.L. Liu, M.R. Green, A.J. Gentles, W. Feng, Y. Xu, C.D. Hoang, M. Diehn, A.A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles, *Nat. Methods* 12 (5) (2015) 453–457.
- [45] G. Yu, L.G. Wang, Y. Han, Q.Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, *OMICS* 16 (5) (2012) 284–287.
- [46] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. U. S. A.* 102 (43) (2005) 15545–15550.
- [47] P. Boccaletto, F. Stefaniak, A. Ray, A. Cappannini, S. Mukherjee, E. Purta, M. Kurkowska, N. Shirvanizadeh, E. Destefanis, P. Groza, G. Avşar, A. Romitelli, P. Pir, E. Dassi, S.G. Conticello, F. Aguiló, J.M. Bujnicki, MODOMICS: a database of RNA modification pathways. 2021 update, *Nucleic Acids Res.* 50 (D1) (2022) D231–d235.
- [48] H. Chen, H. Yang, X. Zhu, T. Yadav, J. Ouyang, S.S. Truesdell, J. Tan, Y. Wang, M. Duan, L. Wei, L. Zou, A.S. Levine, S. Vasudevan, L. Lan, m(5)C modification of mRNA serves a DNA damage code to promote homologous recombination, *Nat. Commun.* 11 (1) (2020) 2834.
- [49] Y. Yang, L. Wang, X. Han, W.L. Yang, M. Zhang, H.L. Ma, B.F. Sun, A. Li, J. Xia, J. Chen, J. Heng, B. Wu, Y.S. Chen, J.W. Xu, X. Yang, H. Yao, J. Sun, C. Lyu, H. L. Wang, Y. Huang, Y.P. Sun, Y.L. Zhao, A. Meng, J. Ma, F. Liu, Y.G. Yang, RNA 5-methylcytosine facilitates the maternal-to-zygotic transition by preventing maternal mRNA decay, *Mol. Cell* 75 (6) (2019) 1188–1202.e11.
- [50] X. Yang, Y. Yang, B.F. Sun, Y.S. Chen, J.W. Xu, W.Y. Lai, A. Li, X. Wang, D.P. Bhattarai, W. Xiao, H.Y. Sun, Q. Zhu, H.L. Ma, S. Adhikari, M. Sun, Y.J. Hao, B. Zhang, C.M. Huang, N. Huang, G.B. Jiang, Y.L. Zhao, H.L. Wang, Y.P. Sun, Y.G. Yang, 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m(5)C reader, *Cell Res.* 27 (5) (2017) 606–625.
- [51] Y. Wang, T. Jiang, J. Xu, Y. Gu, Y. Zhou, Y. Lin, Y. Wu, W. Li, C. Wang, B. Shen, X. Mo, X. Wang, B. Zhou, C. Ding, Z. Hu, Mutations in RNA methyltransferase gene NSUN5 confer high risk of outflow tract malformation, *Front. Cell Dev. Biol.* 9 (2021), 623394.
- [52] H. Ghanbarian, N. Wagner, B. Polo, D. Baudouy, J. Kiani, J.F. Michiels, F. Cuzin, M. Rassoulzadegan, K.D. Wagner, Dnmt2/Trdmt1 as mediator of RNA polymerase II transcriptional activity in cardiac growth, *PLoS One* 11 (6) (2016), e0156953.
- [53] F. Wang, Y. Yang, X. Lin, J.Q. Wang, Y.S. Wu, W. Xie, D. Wang, S. Zhu, Y.Q. Liao, Q. Sun, Y.G. Yang, H.R. Luo, C. Guo, C. Han, T.S. Tang, Genome-wide loss of 5-hmC is a novel epigenetic feature of Huntington's disease, *Hum. Mol. Genet.* 22 (18) (2013) 3641–3653.
- [54] L. Xiang, G. Huang, W. Shu, C. Gong, N. Cao, R. Chen, J. Li, H. Lu, G. Jiang, Role of chromatin remodeling genes and TETs in the development of human midbrain dopaminergic neurons, *Stem Cell Rev. Rep.* 16 (4) (2020) 718–729.
- [55] S. Gao, S. Wang, Z. Zhao, C. Zhang, Z. Liu, P. Ye, Z. Xu, B. Yi, K. Jiao, G.A. Naik, S. Wei, S. Rais-Bahrami, S. Bae, W.H. Yang, G. Sonpavde, R. Liu, L. Wang, TUBB4A interacts with MYH9 to protect the nucleus during cell migration and promotes prostate cancer via GSK3 β / β -catenin signalling, *Nat. Commun.* 13 (1) (2022) 2792.
- [56] C. Simons, N.I. Wolf, N. McNeil, L. Caldovic, J.M. Devaney, A. Takanohashi, J. Crawford, K. Ru, S.M. Grimmond, D. Miller, D. Tonduti, J.L. Schmidt, R. S. Chudnow, R. van Coster, L. Lagae, J. Kisler, J. Sperner, M.S. van der Knaap, R. Schiffmann, R.J. Taft, A. Vanderver, A de novo mutation in the β -tubulin gene TUBB4A results in the leukoencephalopathy hypomyelination with atrophy of the basal ganglia and cerebellum, *Am. J. Hum. Genet.* 92 (5) (2013) 767–773.
- [57] D. Kancheva, T. Chamova, V. Guerguelcheva, V. Mitev, D.N. Azmanov, L. Kalaydjieva, I. Tournev, A. Jordanova, Mosaic dominant TUBB4A mutation in an inbred family with complicated hereditary spastic paraplegia 30 (6) (2015) 854–858.
- [58] B. Erman, S. Firtina, T. Fisgin, C. Bozkurt, F.E. Cipe, Biallelic form of a known CD3E mutation in a patient with severe combined immunodeficiency, *J. Clin. Immunol.* 40 (3) (2020) 539–542.

- [59] C. Lecerf, M. Kamal, S. Vacher, W. Chemlali, A. Schnitzler, C. Morel, C. Dubot, E. Jeannot, D. Meseure, J. Klijanienko, O. Mariani, E. Borcoman, V. Calugaru, N. Badois, A. Chilles, M. Lesnik, S. Krhili, O. Choussy, C. Hoffmann, E. Piaggio, I. Bieche, C. Le Tourneau, Immune gene expression in head and neck squamous cell carcinoma patients, *Eur. J. Cancer* 121 (2019) 210–223.
- [60] S. Punt, J.J. Houwing-Duistermaat, I.A. Schulkens, V.L. Thijssen, E.M. Osse, C.D. de Kroon, A.W. Griffioen, G.J. Fleuren, A. Gorter, E.S. Jordanova, Correlations between immune response and vascularization qRT-PCR gene expression clusters in squamous cervical cancer, *Mol. Cancer* 14 (2015) 71.
- [61] Y. Liu, Y. Wu, P. Zhang, C. Xu, Z. Liu, C. He, Y. Liu, Z. Kang, CXCL12 and CD3E as indicators for tumor microenvironment modulation in bladder cancer and their correlations with immune infiltration and molecular subtypes, *Front. Oncol.* 11 (2021), 636870.
- [62] J. Kempfski, A.D. Giannou, K. Riecken, L. Zhao, B. Steglich, J. Lücke, L. Garcia-Perez, K.F. Karstens, A. Wöstemeier, M. Nawrocki, P. Pelczar, M. Witkowski, S. Nilsson, L. Konczalla, A.M. Shiri, J. Kempfska, R. Wahib, L. Brockmann, P. Huber, A.C. Gnirck, J.E. Turner, D.E. Zazara, P.C. Arck, A. Stein, R. Simon, A. Daubmann, J. Meiners, D. Perez, T. Strowig, P. Koni, A.A. Kruglov, G. Sauter, J.R. Izbicke, A.H. Guse, T. Rösch, A.W. Lohse, R.A. Flavell, N. Gagliani, S. Huber, IL22BP mediates the antitumor effects of lymphotoxin against colorectal tumors in mice and humans, *Gastroenterology* 159 (4) (2020) 1417–1430.e3.
- [63] S.H. Li, S.H. Hosseini, C.A. Gutekunst, S.M. Hersch, R.J. Ferrante, X.J. Li, A human HAP1 homologue. Cloning, expression, and interaction with huntingtin, *J. Biol. Chem.* 273 (30) (1998) 19220–19227.
- [64] X. Zhao, A. Chen, Z. Wang, X.H. Xu, Y. Tao, Biological functions and potential therapeutic applications of huntingtin-associated protein 1: progress and prospects, *Clin. Transl. Oncol. : Off. Publ. Fed. Spanish Oncol. Soc. Natl. Cancer Instit. Mexico* 24 (2) (2022) 203–214.
- [65] J. Wu, J.-y. Zhang, L. Yin, J.-z. Wu, W.-j. Guo, J.-f. Wu, M. Chen, Y.-y. Xia, J.-h. Tang, Y.-c. Ma, X. He, HAP1 gene expression is associated with radiosensitivity in breast cancer cells, *Biochem. Biophys. Res. Commun.* 456 (1) (2015) 162–166.
- [66] L. Zhu, X. Song, J. Tang, J. Wu, R. Ma, H. Cao, M. Ji, C. Jing, Z. Wang, Huntingtin-associated protein 1: a potential biomarker of breast cancer, *Oncol. Rep.* 29 (5) (2013) 1881–1887.
- [67] J.K. Lee, S. Kang, X. Wang, J.L. Rosales, X. Gao, H.-G. Byun, Y. Jin, S. Fu, J. Wang, K.-Y. Lee, HAP1 loss confers l-asparaginase resistance in ALL by downregulating the calpain-1-Bid-caspase-3/12 pathway, *Blood* 133 (20) (2019) 2222–2232.
- [68] K. Ratajczak-Wielgomas, A. Kmiecik, P. Dziegiel, Role of periostin expression in non-small cell lung cancer: periostin silencing inhibits the migration and invasion of lung cancer cells via regulation of MMP-2 expression, *Int. J. Mol. Sci.* 23 (3) (2022).
- [69] L. Morra, H. Moch, Periostin expression and epithelial-mesenchymal transition in cancer: a review and an update, *Virchows Arch. : Int. J. Pathol.* 459 (5) (2011) 465–475.
- [70] C.J. Kim, K. Sakamoto, Y. Tambe, H. Inoue, Opposite regulation of epithelial-to-mesenchymal transition and cell invasiveness by periostin between prostate and bladder cancer cells, *Int. J. Oncol.* 38 (6) (2011) 1759–1766.
- [71] C.J. Kim, N. Yoshioka, Y. Tambe, R. Kushima, Y. Okada, H. Inoue, Periostin is down-regulated in high grade human bladder cancers and suppresses in vitro cell invasiveness and in vivo metastasis of cancer cells, *Int. J. Cancer* 117 (1) (2005) 51–58.
- [72] A. Kanno, K. Satoh, A. Masamune, M. Hirota, K. Kimura, J. Umino, S. Hamada, A. Satoh, S. Egawa, F. Motoi, M. Unno, T. Shimosegawa, Periostin, secreted from stromal cells, has biphasic effect on cell migration and correlates with the epithelial to mesenchymal transition of human pancreatic cancer cells 122 (12) (2008) 2707–2718.
- [73] M. Suguro, H. Tagawa, Y. Kagami, M. Okamoto, K. Ohshima, H. Shiku, Y. Morishima, S. Nakamura, M. Seto, Expression profiling analysis of the CD5+ diffuse large B-cell lymphoma subgroup: development of a CD5 signature, *Cancer Sci.* 97 (9) (2006) 868–874.