An NCCN-IPI based immune-related gene prognostic model for diffuse large B-cell lymphoma

Shidai Mu[1,#,*], Deyao Shi[2,#], Lisha Ai[1], Fengjuan Fan[1], Fei Peng[1], Chunyan Sun[1*], Yu Hu[1*]

1 Institution of Hematology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430022, China

2 Department of Orthopaedics, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430022, China

[#] equally contributed

[*] Correspondence to: Shidai Mu, Chunyan Sun and Yu Hu

Abstract

Background: An enhanced International Prognostic Index (NCCN-IPI) was built to better discriminate diffuse large B-cell lymphoma (DLBCL) patients in the rituximab era. However, there is an urgent need to identify novel valuable biomarkers in the context of targeted therapies, such as immune checkpoint blockade (ICB) therapy.

Methods: Gene expression data and clinical information were obtained from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) datasets. 73 immune-related hub genes in DLBCL patients with different IPI levels were identified by weighted gene co-expression network analysis (WGCNA), and 4 genes were selected to construct an IPI-based immune-related prognostic model (IPI-IPM). Afterward, the genetic, somatic mutational and molecular profiles of IPI-IPM subgroups were analyzed, as well as the potential clinical response of ICB in different IPI-IPM subgroups.

Results: The IPI-IPM was constructed base on the expression of LCN2, CD5L, NLRP11 and SERPINB2, where high-risk patients had shorter overall survival (OS) than low-risk patients, consistent with the results in the GEO cohorts. The comprehensive results showed that a high IPI-IPM risk score was correlated with immune-related signaling pathways, high KMT2D and CD79B mutation rates, high infiltration of CD8+ T cells and macrophages (M1, M2), as well as up-regulation of inhibitory immune checkpoints including PD-L1, LAG3 and BTLA, indicating more potential response to ICB therapy.

Conclusion: The IPI-IPM has independent prognostic significance for DLBCL patients, which provides an immunological perspective to elucidate the mechanisms on tumor progression and drug resistance, also sheds a light on developing immunotherapy for DLBCL.

Keywords: DLBCL, NCCN-IPI, Immune prognostic model, Nomogram, Immunotherapy

Introduction

Diffuse large B-cell lymphoma (DLBCL) accounts for about 40% of non-Hodgkin B-cell lymphoma (NHL), usually presents with advanced stage, both in nodal and in extra-nodal symptomatic disease, with a median age of 60[1, 2]. Although current frontline DLBCL therapy (the standard R-CHOP chemotherapy regimen) is associated with a high complete response rates of 70–80%, 10% to 15% of DLBCL patients are refractory, and almost 40% of cases experience relapse within 2–3 years after initial response[3, 4]. An enhanced International Prognostic Index (NCCN-IPI) was built to better discriminate low- and high-risk subgroups in the rituximab era, which still needs to be further investigated on the robust capacity for risk stratification in the context of targeted therapies[5, 6]. Therefore, there is an urgent need to explore potential molecular mechanism and identify more key biomarkers and therapeutic targets.

Accumulating evidence has shed light on the prognostic role of tumor microenvironments (TME) in immune checkpoint blockade therapy (ICB), which was mostly composed of a variety of immune cells (T-, NK-, and B-cells as well as macrophages) and stroma (blood vessels and extra-cellular matrix)[7-9]. The TME in DLBCL could be categorized as "inflamed" (with two main subtypes: immune suppressed and immune evasion) and "non-inflamed" ("immune excluded"), which are not equally represented among cases of DLBCL, with the majority of DLBCLs in a "non-inflamed" landscape[10, 11]. Here, we applied Estimation of STromal and Immune cells in Malignant Tumors using Expression data (ESTIMATE) to investigate the fraction of stromal and immune cells in tumor samples using gene expression signature, which helps in elucidating the facilitating roles of TME to tumor initiation and progression[12].

In this study, we identified immune-related hub genes in DLBCL patients with different IPI levels by weighted gene co-expression network analysis (WGCNA), and constructed an IPI-based immune-related prognostic model (IPI-IPM). We then characterized the genetic, somatic mutational and molecular profile of IPI-IPM subgroups, investigated the expression of several inhibitory immune checkpoints between low- and high-risk subgroups, and applied tumor immune dysfunction and exclusion (TIDE) and Immune Cell Abundance Identifier (ImmuCellAI) to roughly predict clinical response of ICB in different IPI-IPM subgroups. The results showed that IPI-IPM was a promising prognostic biomarker, which also had potential for use in patient management.

Results

Identification of immune-related genes in DLBCL patients with different IPI levels

A flowchart was diagramed to demonstrate the procedure and result of our study (**Figure 1**). Clinical information of 566 DLBCL patients were obtained from the TCGA database (TCGA-DLBC, CTSP-DLBCL1, NCICCR-DLBCL). Among DLBCL patients, 321 (56.71%) were male and 245 (43.29%) were female. Age of the patients at initial diagnosis ranged from 14 to 92 (median = 62). Other clinical characteristics including

follow-up period, Ann Arbor stages, LDH ratio, ECOG performance status, number of extranodal sites and therapy were all documented (**Supplementary material 1**). Because gene expression data was collected from 3 projects, the principal component analysis was performed to show there was no obvious batch effect among the samples (**Figure S1A**). After excluding samples without IPI score or with IPI score crossing low/high groups (such as 1-5 or 2-3), 458 DLBCL patients were divided into low-IPI groups (n = 231, 112 at low risk (IPI = 0-1) and 119 at low-intermediate risk (IPI = 2)), and high-IPI groups (n = 227, 82 at high-intermediate risk (IPI = 3), 145 at high risk (IPI = 4-5)).

A total 2334 genes (713 up-regulated and 1621 down-regulated) were detected significantly differentially expressed in the RNA-Seq data (**Figure 2A-B and Supplementary material 2**). Besides, all genes with the top 50% variance among samples were included in WGCNA. All clinical characteristics were enrolled as trait variables, and the best β value in the co-expression network was calculated to be 9 **(Figure S2B)**. Index for clustering of module eigengenes was modified to be 0.65, so as to construct reasonable number of merged modules **(Figure S2B)**. As shown in the Module-trait relationship, 10 modules were significantly correlated with IPI group (**Figure 2C**), 6 modeules (Tan, Lightcyan, Royalblue, Midnightblue, Skyblue3, Darkgreen) out of which were significantly associated with gene significance of IPI group (**Figures 2D**).

Moreover, we collected 4686 immune-related genes in the combination with 6 independent databases (**Supplementary material 3**), and identified 73 genes as the IPI-based immune related genes for further analysis (**Figure S2C**).

Construction and validation of an IPI-based immune-related prognostic model

In the training cohort (n = 563), 21 out of the 73 genes were significantly correlated with OS in the univariate Cox regression analysis (**Supplementary material 4**). Next, we applied the Lasso penalized Cox regression to identify the optimal number of genes (n = 5) for risk score model (**Figure 3A-B**). As a result of stepwise multivariate Cox regression and AIC analysis, 4 out of 5 genes were selected to construct the most optimal IPI-based immune-related prognostic model (IPI-IPM) (**Figure 3C**). Risk score = (expression level of LCN2 * 0.001866 + expression level of CD5L * 0.007183 + expression level of NLRP11 * (-0.046541) + expression level of SERPINB2 * 0.027145). According to the ROC curve of the median survival time, we identified 1.07 as the cut-off value (**Figure S3A-B**). Then we calculated the risk score of each patient and divided them into high and low risk groups. As shown in **Figure 3D**, Kaplan-Meier survival analysis showed shorter OS of patients in the high-risk score group (log-rank P = 5.223e-06). The distribution of the risk score, survival status and the 4-gene expression levels between low- and high-risk score groups was shown in **Figure 3E**. As shown in the time-dependent ROC curves, AUCs were 0.694, 0.733, 0.705 for the 1, 3, 5-year, respectively (**Figure 3F**). With 1000 cycles of bootstrapping, the C-index of the risk score model was 0.65 (95% CI: 0.60-0.69, P = 1.87e-09). The results showed that IPI-IPM endowed a good capacity in OS prediction.

Moreover, 335 patients with all available clinicopathologic parameters were enrolled for further analysis. The risk scores, along with age, LDH ratio and number of extranodal site were shown to be independent prognostic factors for OS (**Figure 4A**), which were integrated to construct a nomogram model showing IPI-

IPM with the highest risk points (ranging from 0 to 100, **Figure 4B**). As shown in **Figure 4C**, AUCs were 0.875, 0.799, 0.808 and 0.837 for the 1, 3, 5-year and the median survival time, respectively. The C-index for the nomogram was 0.74 (95% CI: 0.689-0.796, P = 1.19e-18). Moreover, the bias-corrected lines for the nomogram were shown close to the ideal line in 1,3,5-year and the median survival time periods (**Figure S3C**). As shown in the DCA analysis, the nomogram along with risk score from IPI-IPM showed relatively high net benefit (**Figure 4D**). Altogether, these results suggested that the nomogram had excellent capacity and consistency for OS prediction in the training cohort.

414 patients from GEO (GSE10846) with survival data and genome expression microarray data were enrolled for further validation of IPI-IPM. The risk score for each patient was calculated and all patients were divided into low and high risk score groups likewise. Kaplan-Meier survival analysis showed significantly shorter OS of patients in the high-risk score group (log-rank P = 0.003619, **Figure 4E**). And similar nomogram model was also constructed, presenting the risk scores as the major contributor. Likewise, relevant analyses were performed to evaluate the the capacity of IPI-IPM and nomogram model in OS prediction (**Figure S3D-I**). Taken the results of training and testing cohorts together, the nomogram combining IPI-IPM with clinical characteristics (age, LDH ratio and number of extranodal site) was an excellent model for predicting short-term or long-term OS in DLBCL patients, which might guide the therapeutic strategy decision and long-term prognosis follow-up.

Immune related molecular characteristics of IPI-IPM subgroups

As shown in the **Figure 5A**, Pearson correlation coefficient of every 2 genes from IPI-IPM was less than 0.1, suggesting that these 4 genes were independently expressed in DLBCL patients. Compared to the low-risk group, A total 3103 genes (1281 up-regulated and 1822 down-regulated) were detected significantly differentially expressed in the high-risk group (**Figure 5B-C and Supplementary material 5**). Additionally, t-SNE was applied to show an obvious genetic diversity between samples in the high- and low- risk groups (**Figure 5D**). The standard GSEA was performed to display that few gene sets were enriched in the high risk group. Details were all documented in **Figure S4** and **Supplementary material 6.** Furthermore, GSVA was applied as the extension of GSEA to show that several immune related signaling pathways were significantly highly enriched in the high risk group (**Figure 5E**). Based on the somatic mutational data of 37samples from TCGA，most mutations in low-risk group were missense mutation, while more nonsense mutations were identified in high-risk group (**Figure 6A**). Besides, the mutation frequency of top10 genes in high-risk group was much higher than that in low-risk group. Furthermore, we investigated specific mutation sites of key genes corresponding to their amino acids location, including KMT2D, CARD11, CD79B and KIRREL3 (**Figure 6B**).

To gain further molecular insight into the immune characteristics of IPI-IPM, we selected 412 genes by intersecting DEGs with immune related gene sets. 79 genes correlating with risk scores were furthermore identified to be IPI-IPM associated immune genes (absolute Pearson correlation coefficient $\geqslant$ 0.2). Over representation analysis was applied to identify the enriched biological functions and pathways, such as humor

or mucosal immune response, and leukocyte or granulocyte chemotaxis (**Figure 7A-B**). Detailed results were listed in **Supplementary material 7**. Besides, a PPI network via the STRING database was built to identify top10 degree genes via the cytoscape cytoHubba plugin, including CXCL1, LCN2, SLPI, S100A7, LTF, SAA1, S100A12, CXCL5, CRISP3, and MUC7 (**Figure 7C**).

Immune related TME characteristics of IPI-IPM subgroups

With the ESTIMATE algorithm the immune scores of high risk group was significantly higher than low risk group, and there was no significant difference of the ESTIMATE scores or stromal scores between these 2 IPI-IPM groups (**Figure 8A**). Additionally, CIBERSORT was applied to analyze the infiltrating abundances of various immune cell types in different IPI-IPM subgroups (**Figure 8B**). $CD8^+$ T cells, activated memory $CD4^+$ T cells, resting NK cells, and macrophages (M1, M2) were highly infiltrated in the high risk group, while regulatory T cells (Tregs), activated NK cells, non-activated macrophages (M0) and resting mast cells were more abundant in the low risk group (**Figure 8C**).

Moreover, by exploring the TIMER database for relationship between immune infiltration and gene expression/mutation, we found that expression of LCN2 was positively correlated with infiltrating of B cells, and the expression of NLRP11 was negatively correlated with infiltrating of Neutrophils (**Figure S5A**). Besides, there was significant correlation between the infiltration of B cells, Macrophages, and the copy number alteration (CNA) of CD5L, NLRP11 (**Figure S5B**).

Immune checkpoint blockage response prediction of IPI-IPM subgroups

The clinical development of cancer immunotherapies, along with advances in genomic analysis, has validated the important role of TME in predicting for sensitivity to immune checkpoint blockade therapy (ICB). We investigate the expression of several inhibitory immune checkpoints between high- and low-risk groups using TMM normalized counts data. As shown in **Figure 9A**, the expression of PD-L1, IDO-1, IDO-2 and LAG3 were significantly up-regulated in high-risk group, as well as BTLA, VISTA, KIR2DL1, KIR2DL3, KIR2DS4.

We then used TIDE to assess the potential clinical efficacy of immunotherapy in different IPI-IPM subgroups, where higher TIDE prediction score (based on T cell dysfunction and T cell exclusion) represented higher potential for immune evasion, thus less benefit from ICB. As shown in **Figure 9B**, the high-risk group was correlated with a lower T cell exclusion score, lower FAP^+ CAFs, and higher IFN-γ. Furthermore, we used another bioinformatic tool based on ssGSEA - ImmuCellAI to predict the abundance of critical TILs that related to the response to immunotherapy. There were more cytotoxic $CD8^+$ T cells (CTL) and mucosal-associated invariant T cells (MAIT), but less exhausted $CD8^+$ T cells, natural and induced Tregs in high-risk groups (**Figure 9C**).

Discussion

Recent groundbreaking insights into the pronounced genomic heterogeneity of DLBCL have identified potential biomarkers for patient diagnosis, prognostication, and therapy, paving the way for a standardized application of precision medicine[1, 3, 4, 6, 13, 14]. The enhanced NCCN-IPI was built to stratify prognostically relevant subgroups of DLBCL patients with R-CHOP therapy, the robustness of which however needs to be further investigated in the context of targeted therapies and novel biomarkers. In the current study, we used WGCNA to profile IPI-based immune-related gene set and constructed a 4-gene IPI-IPM (SERPINB2, CD5L, LCN2 and NLRP11), with shorter OS in high-risk patients and longer OS in low-risk patients in both TCGA and GEO cohorts.

SerpinB2 belongs to a superfamily of serpins, which has diverse roles in cancer development and metastasis via ECM remodeling[15]. SerpinB2 has been shown to be associated with the survival in various cancer types including lung, breast, bladder, colorectal, and ovarian cancers[15]. Besides, SerpinB2 is upregulated in activated macrophages, monocytes, and fibroblasts, also exerting a protective effect on the TCDD-mediated immunosuppression of the B cells[16]. CD5-like protein (CD5L) is a secreted glycoprotein mainly from macrophages, which involves in infection, atherosclerosis, and cancer by regulating inflammatory responses[17]. Sanjurjo et al. have demonstrated that CD5L drives macrophage polarization toward an anti-inflammatory and pro-resolving phenotype. Moreover, CD5L has been shown to enhance the proliferation of liver cancer cells and protect them from cisplatin-induced apoptosis, consistent with elevated CD5L expression favoring worse prognosis[18]. Lipocalin-2 (LCN2) is a secreted glycoprotein of the adipokine superfamily, dysregulation of which has been tied to obesity, metabolic syndrome, cardiovascular diseases, and cancer[19]. Several studies have revealed elevated LCN2 expression in multiple cancers, and correlated LCN2 overexpression with the progression and poor prognosis of breast, gastric, and pancreatic cancer. Additionally, Gomez-Chou et al. firstly showed that LCN2 modulated pro-inflammatory factors production in pancreatic stromal cells and reduced macrophages infiltration in the stroma of pancreatic ductal adenocarcinoma (PDAC)[20]. NLRP11, a primate-specific member of the NOD-like receptor (NLR), is highly expressed in the testes, ovaries, lungs and other various tissues, with the important role in both development and immune regulation[21]. Ellwanger et al. reported that NLRP11 is expressed in a panel of immune cells and functions as a negative regulator of inflammatory responses by inhibiting NF-κB and IFN pathways. Furthermore, the differential expression of NLRP11 in B cell lymphoma lines and cancer entities suggested the potential role of NLRP11 in the malignant transformation process.

Additionally, we demonstrated that the risk score remained an independent prognostic factor after the modification of clinical characteristics, suggesting the promising potential of local immune status in accurate prognosis. Therefore, we developed a nomogram model combining the risk score and other clinical features (age, LDH ratio and number of extranodal site), to predict OS probability of DLBCL patients in 1-, 3- and 5-year and the median survival time, respectively. And the calibration curve showed satisfactory agreement

between the observed values and the predicted values in 1-, 3-, 5-year and the median survival time periods. The result of DCA analysis also presented the nomogram with relatively high net benefit. Moreover, our nomogram provides a complementary perspective on individualizing tumors and develops an individual scoring system for patients, thus arising to be a promising tool for clinicians in the future.

The overall somatic mutational profile showed more nonsense mutations in high-risk group, and the largest difference in mutations was KMT2D (43.75% in the high risk samples vs. 19.05% in the low risk samples). KMT2D is a tumor suppressor gene in DLBCL and genetic ablation of KMT2D in a BCL2-overexpression driven model of B-cell lymphoma promotes a higher penetrance of DLBCL[22]. Additionally, higher mutation frequency of CD79B was found in the high risk group, indicating that high-risk DLBCLs promote proliferation through chronic active B-cell receptor (BCR) signaling and NF-κB constitutive activation[23]. Therefore, high-risk DLBCL patients with high KMT2D and CD79B mutations have a worse outcome, in agreement with our survival results.

In the GSEA analysis between the low and high risk group patients, few immune-related gene sets, such as immune response, cytokine/chemokine signaling, and NOD-like receptor signaling pathway, were enriched for the high-risk group, but not in the low-risk group. Therefore, we speculated that the local immune signature conferred an intense immune phenotype in the high-risk group and a weaker immune phenotype in the low risk group. Moreover, 79 immune-related DEGs correlating with risk scores were identified as IPI-IPM associated immune genes sets. Several immune-related pathways including immune response and chemotaxis were enriched, and top10 degree genes such as CXCL1, CXCL5, LCN2, S100A7, and MUC7, were identified through a PPI network via the STRING database.

Moreover, we evaluated the tumor purity and immune/stromal cells infiltration where the immune scores of high-risk group was significantly higher than that of low-risk group. Additionally, the composition of some immune cells was different between two IPI-IPM subgroups, where CD8+ T cells and macrophages (M1, M2) were highly infiltrated in high-risk group, while regulatory T cells (Tregs) and non-activated macrophages (M0) were more abundant in the low-risk group. Several studies have reported an inferior effect of tumor-infiltrating cytotoxic cells on the outcome in DLBCL, the reason for which is that malignant cells with high CD8+ T cells infiltration might be more resistant to cell-mediated killing, and these T cells might be in a dysfunctional state, thus inducing immune suppression[8, 9]. Also, an increase in the M2 component of TAM has been shown to correlate with a poor prognosis in DLBCL, where M2 macrophages demonstrates a high rate of PD-L1 expression, likely allowing tumor cells to escape immune control.

Next, we explored the expression of several inhibitory immune checkpoints between IPI-IPM subgroups so as to predict the response for immunotherapy, where the expression of PD-L1, LAG3 and BTLA were significantly up-regulated in high-risk group. Although PD-L1 expression was shown to correlate with

clinical response to PD-1 blockade and prognosis in solid tumors, the expression patterns of PD-1 and PD-L1 are complex and variable in lymphoid malignancies[24, 25]. Kim et al. has found that high tumoral PD-L1 expression is associated with poor prognosis in primary DLBCL of the central nervous system[26]. Combination with PD-L1 blockade, other immune checkpoint inhibitors, such as CTLA4, TIM3, LAG3, and BTLA, have emerged as the novel strategy for lymphoma immunotherapy[10, 27]. Keane et al. has reported the co-expression of LAG3 with PD-1 and TIM-3 on CD4+ Tregs and CD8+ TILs, and high LAG3 expression is associated with poor outcome of DLBCL patients[28]. Therefore, the risk score from our model was compatible with the ability of tumor-infiltrating immune cells to determine the expression of immune checkpoints, suggesting that the poor prognosis of high-risk group may be due to the stronger immunosuppressive TME, and high-risk patients will benefit more from immune checkpoint inhibitors than low-risk patients, thereby resulting in a better prognosis.

Furthermore, we used TIDE to assess the potential clinical efficacy of immunotherapy in different IPI-IPM subgroups, where higher TIDE prediction score (based on T cell dysfunction and T cell exclusion) represented higher potential for immune evasion, thus less benefit from ICB. In our study, there is no significant difference on TIDE score and T cells dysfunction between these 2 subgroups, while the high-risk group was correlated with a lower T cell exclusion score (using MDSC, M2-TAM and CAF gene signature). Additionally, ImmuCellAI was applied to show more CD8+ CTLs, but less exhausted CD8+ T cells and Tregs in high-risk groups. The complex immune-escape mechanisms in DLBCL may be the most important reason for the inconsistent results compared to that in most solid tumor data sets[2, 3, 10, 11]. Besides, Muris el al. has found that a high percentage of activated CTLs is a strong indicator for an unfavorable clinical outcome in patients with primary nodal DLBCL[29]. Also, more samples are warranted to further validation.

Our research provides new insights into the TME and immune-related therapies for DLBCL. However, it is noteworthy that some limitations came out because the conclusion was drawn from data from retrospective studies, and prospective studies are warranted to further confirm our results. In addition, functional and mechanistic studies of the genes in our risk model should be conducted to support their clinical application.

Materials and Methods

Data selection and acquisition

The study reported herein fully satisfies the TCGA publication requirements (http://cancergenome.nih.gov/publications/publicationguidelines). Gene expression data and the corresponding clinical data for DLBCL samples (Project: TCGA-DLBC, CTSP-DLBCL1, NCICCR-DLBCL) were acquired from the Cancer Genome Atlas (TCGA) website (https://portal.gdc.cancer.gov/repository) through the TCGAbiolinks[30] R package in R software (version 4.0.2, https://www.r-project.org) and Rstudio software (Version 1.3.1073, https://rstudio.com). Among them, available gene expression quantification data of 570 samples were downloaded using the Illumina HTSeq-Counts and HTSeq-FPKM workflow types. The

latest Homo_sapiens.GRCh38.101.chr.gtf file (https://www.ensembl.org) was used for gene symbol annotation. The trimmed mean of M values (TMM) method was applied for the normalization of downloaded gene expression data by using the edgeR R package[31]. Sequencing data of low abundance were eliminated.

The Principal components analysis

The Principal components analysis[32] for samples of different projects was performed and visualized by using the psych and factoextra packages of R.

Identification of differentially expressed genes

458 samples were divided to two groups according to IPI scores, and he edgeR R package[31] was applied to identify differentially expressed genes (DEGs) between high- and low- IPI groups. The differential expression was defined with a $|\log_2$ fold-change (FC)$| > 1$ and the false discovery rate (FDR) value < 0.01.

Gene functional enrichment analysis

The clusterProfiler R package[33] was used for both over representation analysis and gene set enrichment analysis. Analysis of Gene Ontology (GO)[34], Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway[35], and Reactome pathway[36] was contained in the present study. $P < 0.05$ was considered statistically significance. The standard gene set enrichment analysis was performed by using the normalized counts data via the GSEA software (http://software.broadinstitute.org/gsea). The threshold for GSEA was set at the nominal p-value < 0.05, FDR <0.25 and | normalized enrichment score (NES) | > 1.0. The non-parametric gene set variation analysis was further performed with the GSVA package of R. The annotated Hallmark gene sets, Canonical pathways gene sets (KEGG and Reactome) and Ontology gene sets (GO biological process) were selected as the reference gene sets.

Weighted gene co-expression network analysis

Weighted gene co-expression network analysis (WGCNA)[37, 38] is commonly used for analyzing high-throughput gene expression data with different characteristics, so as to mine gene co-expression networks and intramodular hub genes based on pairwise correlations in genomic applications. In the present study, we applied the WGCNA R package[38] to analyze key gene clusters that were most relevant to IPI scores in DLBCL samples.

Construction and validation of IPI-based immune-related prognostic model (IPI-IPM)

73 IPI-based immune related genes were selected to construct the prognostic risk model. The training cohort (563 samples from TCGA) was used for the construction of IPI-IPM, while the testing cohort (414 patients from Gene Expression Omnibus (GEO), GSE10846) was used for external validation. The Survival R package was utilized to analyze the correlation between the expression of objective gene sets and DLBCL patients' overall survival (OS). The univariate Cox regression analysis was used to screen genes with $P < 0.05$. Lasso (least absolute shrinkage and selection operator) regression analysis was applied for variable selection and regularization to enhance the prediction accuracy and interpretability[39]. Then the multivariate Cox regression analysis was carried out to select the optimal gene sets, according to the method of Akaike information criterion (AIC)[40]. For each sample, the risk score = SUM (the normalized expression level of

each gene * the corresponding regression coefficient).

563 DLBCL patients in the training cohort were divided to low- and high- risk score groups according to the cut-off value identified in the receiver operating characteristic (ROC) curve of the median survival time. Then, Kaplan-Meier survival analysis and time-dependent ROC curve analysiswere performed to evaluate the prognostic significance and accuracy of IPI-IPM. Besides, Harrell's concordance index (C-index) was calculated by using the survcomp R package[41].

The univariate and multivariate Cox regression analyses were performed on the risk score and all available clinicopathologic parameters, such as age, gender, Ann Arbor clinical stage, LDH ratio, ECOG performance status, number of extranodal site. Then, all independent prognostic factors were retained to construct a prognostic nomogram for OS probability assessment by using the rms and mstate R packages[42]. The discriminative efficacy, consistency and clinical judgment utility of the nomogram was evaluated by time-dependent ROC curve[43],C-index, time-dependent Calibration plots, and decision curve analysis (DCA) using the rmda R package[44]. As for the external validation, all above methods were used to evaluate the prognostic model and nomogram in the testing cohort.

Comprehensive analysis of molecular and immune characteristics in different IPI-IPM subgroups

DEGs between high and low risk score groups were analyzed following the criteria of $|\log2FC| > 1$ and FDR value < 0.01. The gene expression of samples between IPI-IPM subgroups were analyzed with the t-distributed stochastic neighbor embedding (t-SNE)[45] method by using the Rtsne package of R and then visualized on the 3D map with the scatterplot3d package of R. And the quantity and quality of gene mutations were analyzed in IPI-IPM subgroups by using the Maftools R package[46]. The intersection of DEGs and immune related gene set was used to construct a protein-protein interaction (PPI) network[47] based on the STRING database[48], and the Cytoscape plugin cytoHubba was utilized to identify top10 degree genes in the network[49].

The ESTIMATE algorithm (Estimation of Stromal and Immune cells in Malignant Tumor tissues using Expression data)[12], a bioinformatic tool for assessing tumor purity and stromal/immune cells infiltration in tumor tissues, was used to calculate the corresponding infiltrating scores of the 570 DLBCL samples in the present study. Besides, the gene expression data of 570 DLBCL samples were imported into CIBERSORT (HTTPS://cibersort.stanford.edu/)[50] and iterated 1000 times to estimate the relative proportion of 22 types of immune cells. The TIMER database[51] was explored to evaluate the correlation between immune infiltration and critical gene expression/mutation. TIDE[52] and immuneAI[53] were explored to assess the potential clinical efficacy of immunotherapy in different IPI-IPM subgroups.

Data Analysis

All statistical data was analyzed in the R software (version 4.0.2. An independent t-test was applied to compare continuous variables between two groups. Immune scores calculated via ESTIMATE between groups were compared by the Wilcoxon test). Statistical tests were two-tailed with statistical significance level set at $P < 0.05$.

Acknowledgements

Not applicable

Author contributions

SM, DS, CS and YH conceived and designed the study. SM and DS did the literature research, performed study selection, data extraction, statistical analysis and wrote the draft. LA, FF and FP participated in the extraction and analysis of data. All the authors read and approved the final version of the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest.

Funding

This work was supported by grants from the National Natural Science Foundation of China (No. 81974007 to Chunyan Sun) and the National Key R&D Program of China (Grant No 2019YFC1316204 to Yu Hu).

Data Availability Statement

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

References

1. Pasqualucci, L, and Dalla-Favera, R (2018). Genetics of diffuse large B-cell lymphoma. *Blood* **131**: 2307-2319.
2. Solimando, AG, Annese, T, Tamma, R, Ingravallo, G, Maiorano, E, Vacca, A, Specchia, G, and Ribatti, D (2020). New Insights into Diffuse Large B-Cell Lymphoma Pathobiology. *Cancers (Basel)* **12**.
3. El Hussein, S, Shaw, KRM, and Vega, F (2020). Evolving insights into the genomic complexity and immune landscape of diffuse large B-cell lymphoma: opportunities for novel biomarkers. *Mod Pathol* **33**: 2422-2436.
4. Wang, L, Li, LR, and Young, KH (2020). New agents and regimens for diffuse large B cell lymphoma. *J Hematol Oncol* **13**: 175.
5. Zhu, Z, Jin, Z, Zhang, H, Zhang, M, and Sun, D (2020). Integrative Clustering Reveals a Novel Subtype of Soft Tissue Sarcoma With Poor Prognosis. *Frontiers in genetics* **11**: 69.
6. Wight, JC, Chong, G, Grigg, AP, and Hawkes, EA (2018). Prognostication of diffuse large B-cell lymphoma in the molecular era: moving beyond the IPI. *Blood Rev* **32**: 400-415.
7. Ciavarella, S, Vegliante, MC, Fabbri, M, De Summa, S, Melle, F, Motta, G, De Iuliis, V, Opinto, G, Enjuanes, A, Rega, S*, et al.* (2018). Dissection of DLBCL microenvironment provides a gene expression-based predictor of survival applicable to formalin-fixed paraffin-embedded tissue. *Ann Oncol* **29**: 2363-2370.
8. Hopken, UE, and Rehm, A (2019). Targeting the Tumor Microenvironment of Leukemia and Lymphoma. *Trends Cancer* **5**: 351-364.
9. Autio, M, Leivonen, SK, Bruck, O, Mustjoki, S, Jorgensen, JM, Karjalainen-Lindsberg, ML, Beiske, K, Holte, H, Pellinen, T, and Leppa, S (2020). Immune cell constitution in the tumor microenvironment predicts the outcome in diffuse large B-cell lymphoma. *Haematologica*.
10. Xu-Monette, ZY, Xiao, M, Au, Q, Padmanabhan, R, Xu, B, Hoe, N, Rodriguez-Perales, S, Torres-Ruiz, R, Manyam, GC, Visco, C*, et al.* (2019). Immune Profiling and Quantitative Analysis Decipher the Clinical Role of Immune-Checkpoint Expression in the Tumor Immune Microenvironment of DLBCL. *Cancer Immunol Res* **7**: 644-657.
11. Kline, J, Godfrey, J, and Ansell, SM (2020). The immune landscape and response to immune checkpoint blockade therapy in lymphoma. *Blood* **135**: 523-533.
12. Yoshihara, K, Shahmoradgoli, M, Martínez, E, Vegesna, R, Kim, H, Torres-Garcia, W, Treviño, V, Shen, H, Laird, PW, Levine, DA*, et al.* (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications* **4**: 2612.
13. Zhou, M, Zhao, H, Xu, W, Bao, S, Cheng, L, and Sun, J (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol Cancer* **16**: 16.
14. Tian, L, He, Y, Zhang, H, Wu, Z, Li, D, and Zheng, C (2018). Comprehensive analysis of differentially expressed profiles of lncRNAs and mRNAs reveals ceRNA networks in the transformation of diffuse large B-cell lymphoma. *Oncol Lett* **16**: 882-890.
15. Ramnefjell, M, Aamelfot, C, Helgeland, L, and Akslen, LA (2017). Low expression of SerpinB2 is associated with reduced survival in lung adenocarcinomas. *Oncotarget* **8**: 90706-90718.
16. Dornbos, P, Warren, M, Crawford, RB, Kaminski, NE, Threadgill, DW, and LaPres, JJ (2018). Characterizing Serpinb2 as a Modulator of TCDD-Induced Suppression of the B Cell. *Chem Res Toxicol* **31**: 1248-1259.
17. Sanjurjo, L, Aran, G, Tellez, E, Amezaga, N, Armengol, C, Lopez, D, Prats, C, and Sarrias, MR (2018). CD5L Promotes M2 Macrophage Polarization through Autophagy-Mediated Upregulation of ID3. *Front Immunol* **9**: 480.
18. Aran, G, Sanjurjo, L, Barcena, C, Simon-Coma, M, Tellez, E, Vazquez-Vitali, M, Garrido, M, Guerra, L, Diaz, E, Ojanguren, I*, et al.* (2018). CD5L is upregulated in hepatocellular carcinoma and promotes liver cancer cell proliferation and antiapoptotic responses by binding to HSPA5 (GRP78). *FASEB J* **32**: 3878-3891.
19. Santiago-Sanchez, GS, Pita-Grisanti, V, Quinones-Diaz, B, Gumpper, K, Cruz-Monserrate, Z, and Vivas-Mejia, PE (2020). Biological Functions and Therapeutic Potential of Lipocalin 2 in Cancer. *Int*

J Mol Sci **21**.

20. Gomez-Chou, SB, Swidnicka-Siergiejko, AK, Badi, N, Chavez-Tomar, M, Lesinski, GB, Bekaii-Saab, T, Farren, MR, Mace, TA, Schmidt, C, Liu, Y*, et al.* (2017). Lipocalin-2 Promotes Pancreatic Ductal Adenocarcinoma by Regulating Inflammation in the Tumor Microenvironment. *Cancer Res* **77**: 2647-2660.
21. Ellwanger, K, Becker, E, Kienes, I, Sowa, A, Postma, Y, Cardona Gloria, Y, Weber, ANR, and Kufer, TA (2018). The NLR family pyrin domain-containing 11 protein contributes to the regulation of inflammatory signaling. *The Journal of biological chemistry* **293**: 2701-2710.
22. Saffie, R, Zhou, N, Rolland, D, Onder, O, Basrur, V, Campbell, S, Wellen, KE, Elenitoba-Johnson, KSJ, Capell, BC, and Busino, L (2020). FBXW7 Triggers Degradation of KMT2D to Favor Growth of Diffuse Large B-cell Lymphoma Cells. *Cancer Res* **80**: 2498-2511.
23. Takeuchi, T, Yamaguchi, M, Kobayashi, K, Miyazaki, K, Tawara, I, Imai, H, Ono, R, Nosaka, T, Tanaka, K, and Katayama, N (2017). MYD88, CD79B, and CARD11 gene mutations in CD5-positive diffuse large B-cell lymphoma. *Cancer* **123**: 1166-1173.
24. Xu-Monette, ZY, Zhou, J, and Young, KH (2018). PD-1 expression and clinical PD-1 blockade in B-cell lymphomas. *Blood* **131**: 68-83.
25. Xie, M, Huang, X, Ye, X, and Qian, W (2019). Prognostic and clinicopathological significance of PD-1/PD-L1 expression in the tumor microenvironment and neoplastic cells for lymphoma. *International immunopharmacology* **77**: 105999.
26. Kim, S, Nam, SJ, Park, C, Kwon, D, Yim, J, Song, SG, Ock, CY, Kim, YA, Park, SH, Kim, TM*, et al.* (2019). High tumoral PD-L1 expression and low PD-1(+) or CD8(+) tumor-infiltrating lymphocytes are predictive of a poor prognosis in primary diffuse large B-cell lymphoma of the central nervous system. *Oncoimmunology* **8**: e1626653.
27. Chen, BJ, Dashnamoorthy, R, Galera, P, Makarenko, V, Chang, H, Ghosh, S, and Evens, AM (2019). The immune checkpoint molecules PD-1, PD-L1, TIM-3 and LAG-3 in diffuse large B-cell lymphoma. *Oncotarget* **10**: 2030-2040.
28. Keane, C, Law, SC, Gould, C, Birch, S, Sabdia, MB, Merida de Long, L, Thillaiyampalam, G, Abro, E, Tobin, JW, Tan, X*, et al.* (2020). LAG3: a novel immune checkpoint expressed by multiple lymphocyte subsets in diffuse large B-cell lymphoma. *Blood Adv* **4**: 1367-1377.
29. Muris, JJ, Meijer, CJ, Cillessen, SA, Vos, W, Kummer, JA, Bladergroen, BA, Bogman, MJ, MacKenzie, MA, Jiwa, NM, Siegenbeek van Heukelom, LH*, et al.* (2004). Prognostic significance of activated cytotoxic T-lymphocytes in primary nodal diffuse large B-cell lymphomas. *Leukemia* **18**: 589-596.
30. Colaprico, A, Silva, TC, Olsen, C, Garofano, L, Cava, C, Garolini, D, Sabedot, TS, Malta, TM, Pagnotta, SM, Castiglioni, I*, et al.* (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research* **44**: e71.
31. Robinson, MD, McCarthy, DJ, and Smyth, GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26**: 139-140.
32. Groth, D, Hartmann, S, Klie, S, and Selbig, J (2013). Principal components analysis. *Methods in molecular biology (Clifton, NJ)* **930**: 527-547.
33. Yu, G, Wang, LG, Han, Y, and He, QY (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology* **16**: 284-287.
34. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic acids research* **47**: D330-d338.
35. Ogata, H, Goto, S, Sato, K, Fujibuchi, W, Bono, H, and Kanehisa, M (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* **27**: 29-34.
36. Fabregat, A, Jupe, S, Matthews, L, Sidiropoulos, K, Gillespie, M, Garapati, P, Haw, R, Jassal, B, Korninger, F, May, B*, et al.* (2018). The Reactome Pathway Knowledgebase. *Nucleic acids research* **46**: D649-d655.
37. Maertens, A, Tran, V, Kleensang, A, and Hartung, T (2018). Weighted Gene Correlation Network Analysis (WGCNA) Reveals Novel Transcription Factors Associated With Bisphenol A Dose-Response. *Frontiers in genetics* **9**: 508.
38. Langfelder, P, and Horvath, S (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**: 559.

39. Steyerberg, EW, Vickers, AJ, Cook, NR, Gerds, T, Gonen, M, Obuchowski, N, Pencina, MJ, and Kattan, MW (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology (Cambridge, Mass)* **21**: 128-138.
40. Yamaoka, K, Nakagawa, T, and Uno, T (1978). Application of Akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations. *Journal of pharmacokinetics and biopharmaceutics* **6**: 165-175.
41. Schröder, MS, Culhane, AC, Quackenbush, J, and Haibe-Kains, B (2011). survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics (Oxford, England)* **27**: 3206-3208.
42. de Wreede, LC, Fiocco, M, and Putter, H (2010). The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Computer methods and programs in biomedicine* **99**: 261-274.
43. Heagerty, PJ, Lumley, T, and Pepe, MS (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**: 337-344.
44. Tataranni, T, and Piccoli, C (2019). Dichloroacetate (DCA) and Cancer: An Overview towards Clinical Applications. *Oxidative medicine and cellular longevity* **2019**: 8201079.
45. Cieslak, MC, Castelfranco, AM, Roncalli, V, Lenz, PH, and Hartline, DK (2020). t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis. *Marine genomics* **51**: 100723.
46. Mayakonda, A, Lin, DC, Assenov, Y, Plass, C, and Koeffler, HP (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome research* **28**: 1747-1756.
47. Bader, GD, and Hogue, CW (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics* **4**: 2.
48. Szklarczyk, D, Morris, JH, Cook, H, Kuhn, M, Wyder, S, Simonovic, M, Santos, A, Doncheva, NT, Roth, A, Bork, P*, et al.* (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research* **45**: D362-d368.
49. Shannon, P, Markiel, A, Ozier, O, Baliga, NS, Wang, JT, Ramage, D, Amin, N, Schwikowski, B, and Ideker, T (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**: 2498-2504.
50. Chen, B, Khodadoust, MS, Liu, CL, Newman, AM, and Alizadeh, AA (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods in molecular biology (Clifton, NJ)* **1711**: 243-259.
51. Li, T, Fu, J, Zeng, Z, Cohen, D, Li, J, Chen, Q, Li, B, and Liu, XS (2020). TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic acids research* **48**: W509-w514.
52. Jiang, P, Gu, S, Pan, D, Fu, J, Sahu, A, Hu, X, Li, Z, Traugh, N, Bu, X, Li, B*, et al.* (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med* **24**: 1550-1558.
53. Miao, YR, Zhang, Q, Lei, Q, Luo, M, Xie, GY, Wang, H, and Guo, AY (2020). ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy. *Adv Sci (Weinh)* **7**: 1902880.

Figure legends

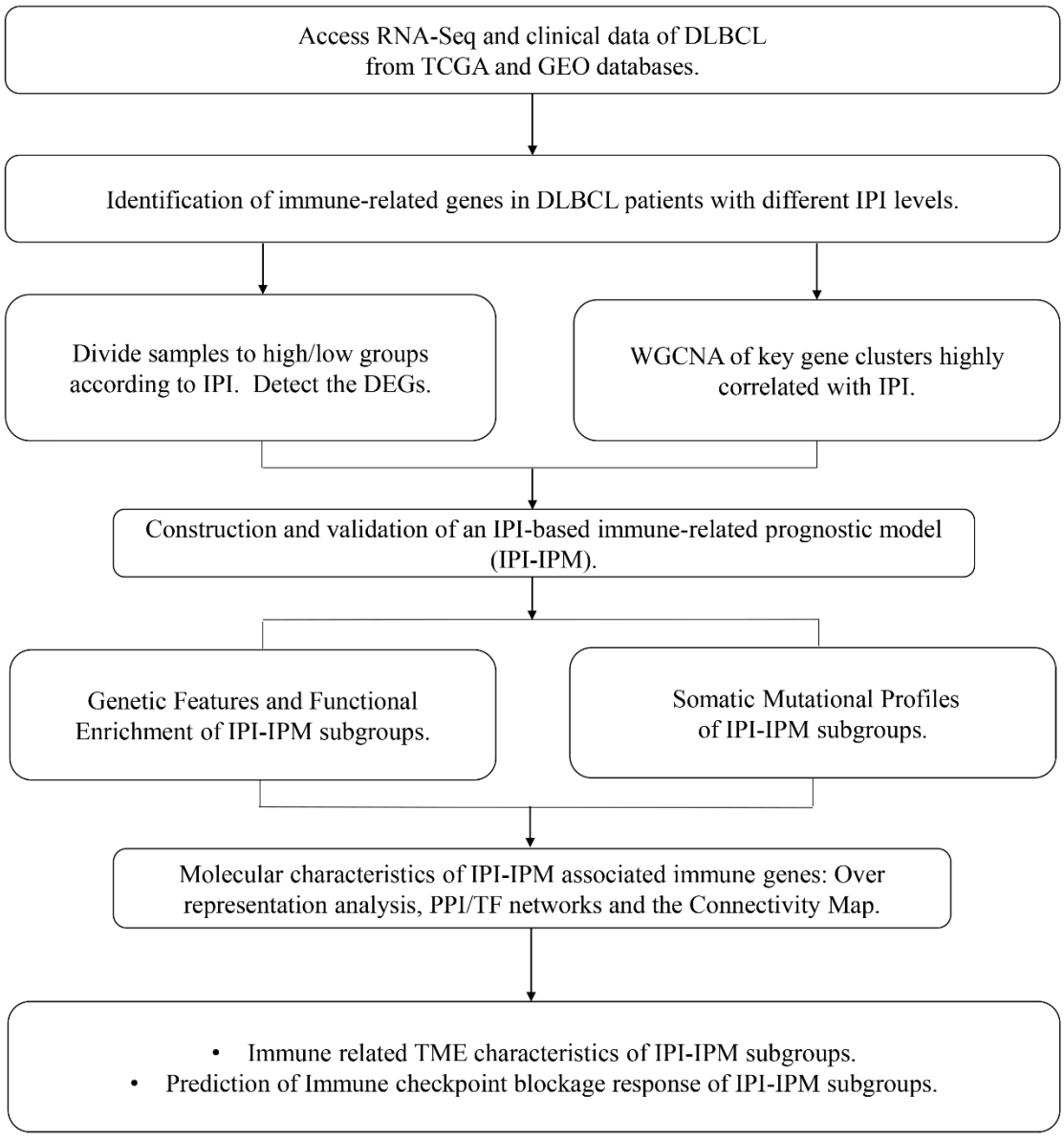

Figure 1. A flow chart for the process of the present study.

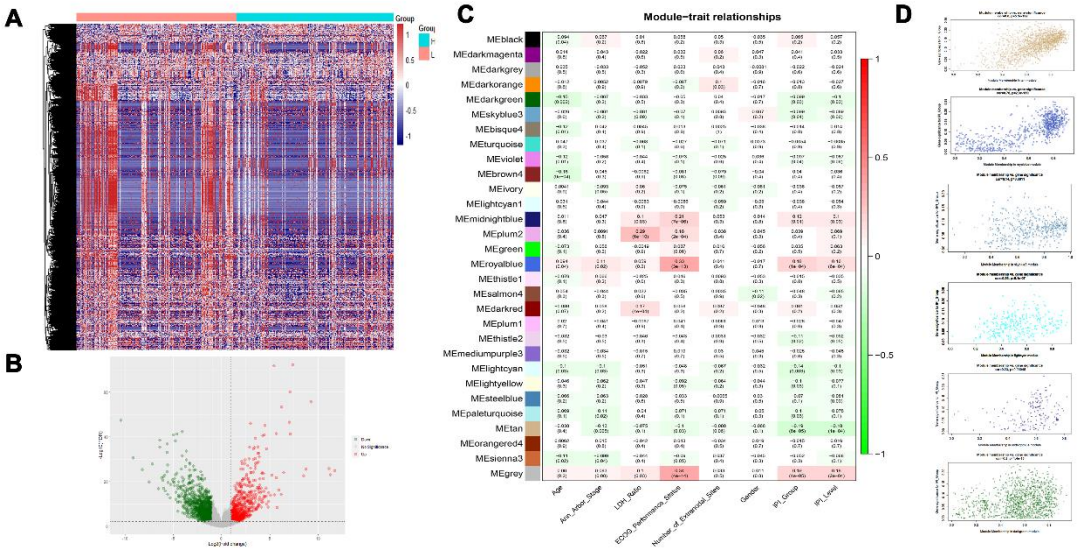

Figure 2. Identification of immune-related genes in DLBCL patients with different IPI levels. (A-B) Heatmap and volcano plot of differentially expressed genes between the high and low IPI groups. (C-D) Identification of IPI related gene clusters using WGCNA: Analysis and visualization of Module-trait relationship (C); and Correlation between gene module membership and gene significance for IPI in six modules (D).

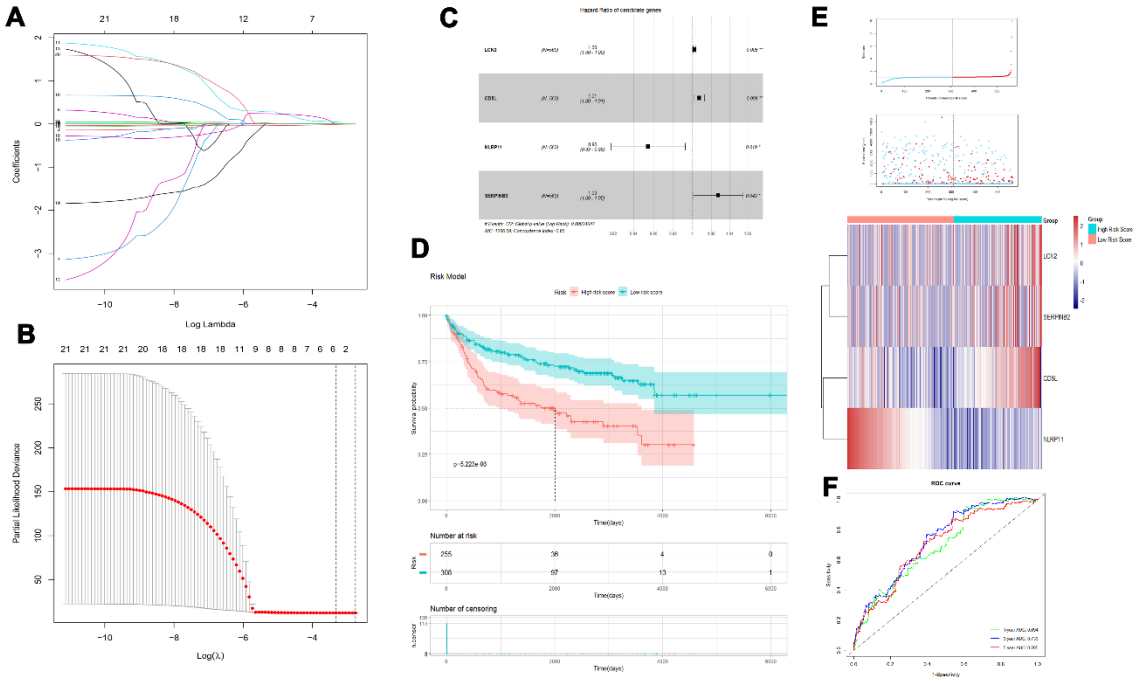

Figure 3. Construction and validation of an IPI-based immune-related prognostic model. (A and B) Plots of the Lasso penalized Cox regression. (C) A forest plot of the genes that compose the IPI-based immune-related prognostic model. (D) Survival analysis of overall survival between high and low IPI-IPM risk groups. (E) Risk score, survival status and the gene expression of each patient in high and low IPI-

IPM risk groups. (F) time-dependent ROC curves for the IPI-IPM risk score model.

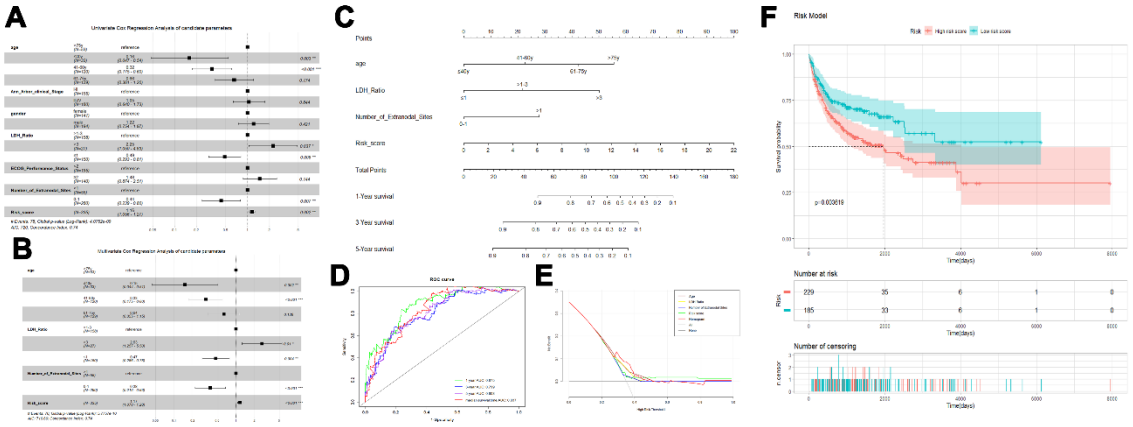

Figure 4. Construction and validation of an IPI-based immune-related nomogram model. (A-B) The univariate analysis and multivariate analysis of IPI-IPM risk score and clinicopathologic parameters. (C) Nomogram for the prediction of the survival probability of 1-, 3-, and 5-year overall survival. (D) time-dependent ROC curves for the Nomogram. (E) The DCA analysis of all parameters in the nomogram. (F) Survival analysis of overall survival between high and low risk groups in the testing (GSE10846) cohort.

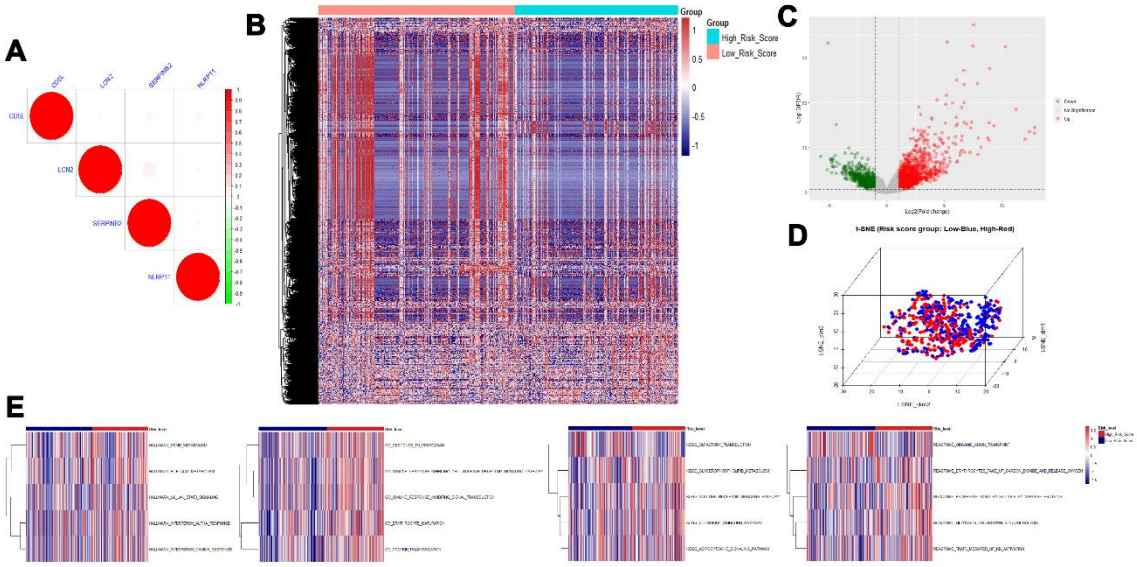

Figure 5. Molecular characteristics of IPI-IPM associated immune genes. (A) Pearson correlation analysis of the genes that compose the IPI-IPM. (B-C) Heatmap and volcano plot of differentially expressed genes between high and low IPI-IPM risk groups. (D) The t-SNE algorithm were applied to show the difference of DLBCL patients between the high and low IPI-IPM risk groups. (E) GSVA heatmaps of IPI-IPM risk groups regarding the Hallmark gene sets, GO biological processes terms, KEGG pathways and Reactome pathways.

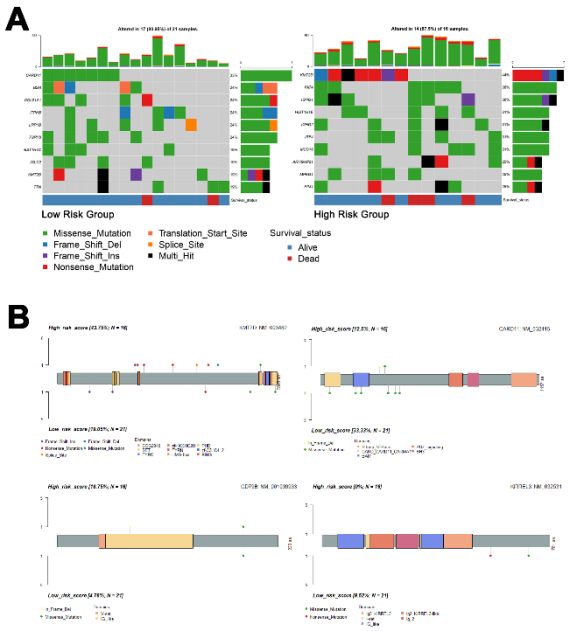

Figure 6. Somatic mutational profiles of IPI-IPM subgroups.

(A) Differentially mutated genes between high and low IPI-IPM risk groups. Top 10 Mutated genes (rows) are ordered by mutation rate. Samples (columns) are be classified into high and low IPI-IPM risk groups. The color-coding legends indicates the mutation types and survival status of patients. (B) Lollipop plots for amino acid changes of KMT2D, CARD11, CD79B, and KIRREL3.

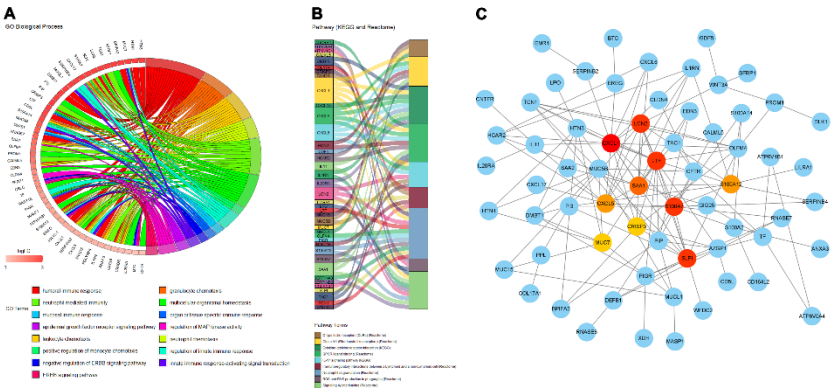

Figure 7. Immune characteristics of IPI-IPM subgroups.

(A and B) Over representative analysis: A Chord map of the enriched biological processes (A) and a Sankey plot of the enriched pathways (B); (C) a protein-protein interaction network based on the STRING database.

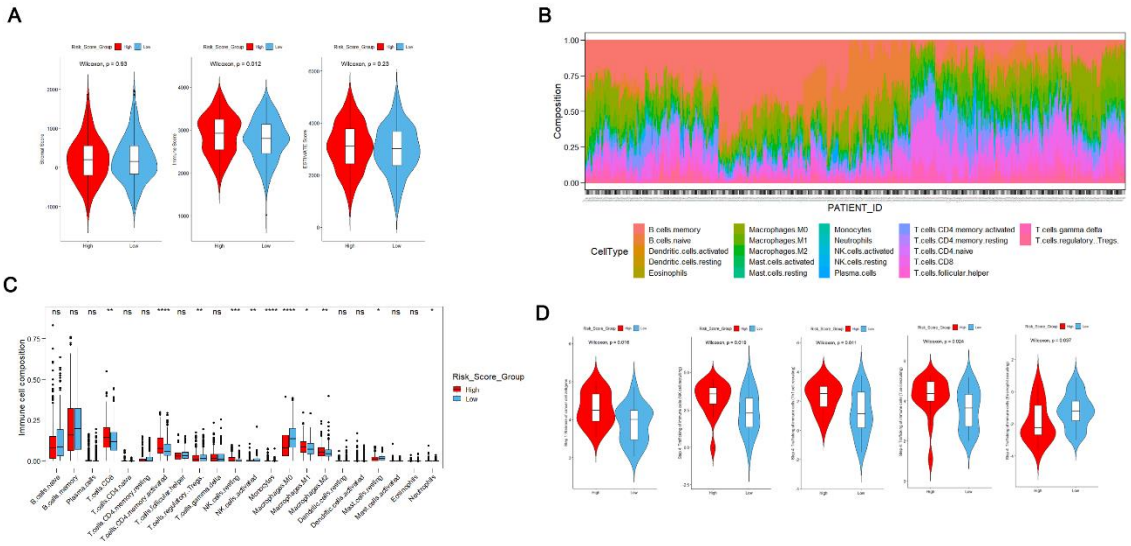

Figure 8. Tumor immune microenvironment (TME) characteristics of IPI-IPM subgroups. (A) Analysis of tumor purity and immune/stromal cells infiltration by using the ESTIMATE algorithm. (B and C) Analysis of immune cells infiltration by using the CIBERSORT algorithm: Relative proportion of each type of cells infiltration in DLBCL patients (B) and a bar plot visualizing significantly differentially infiltrated cells between high and low IPI-IPM risk groups (C). (D) Profile of anticancer immune microenvironment based on the TIP database between high and low IPI-IPM risk groups.

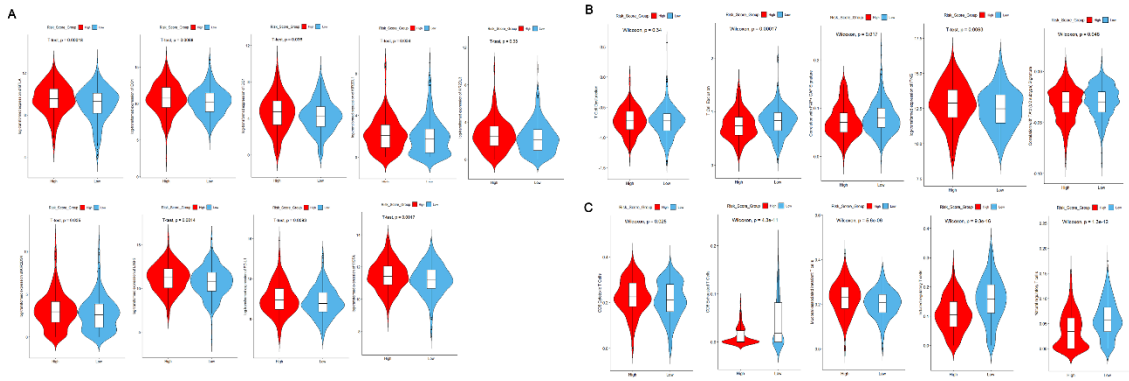

Figure 9. Immune checkpoint blockage (ICB) response prediction of IPI-IPM subgroups. (A) The expression of inhibitory immune checkpoints between high and low IPI-IPM risk groups. (B) Analysis of T cell dysfunction, T cell exclusion, FAP+ CAFs, and IFN-γ signature between high and low IPI-IPM risk groups based on TIDE algorithm. (C) Prediction of the infiltration of cytotoxic CD8+ T cells, mucosal-associated invariant T cells, exhausted CD8+ T cells, natural and induced Tregs between high and low IPI-IPM risk groups based on ImmuCellAI database.