Identification and Validation of a DNA Damage Repair-Related Signature for Diffuse Large B-Cell Lymphoma

Yang Li, Xiyang Liu, Yu Chang, Bingjie Fan, Chenxing Shangguan, Huan Chen, and Lei Zhang

Department of Oncology, The First Affiliated Hospital of Zhengzhou University, No. 1 Jianshe East Road, Zhengzhou 450000, China

Correspondence should be addressed to Lei Zhang; zhang_lei@zzu.edu.cn

Received 18 July 2022; Accepted 27 September 2022; Published 14 October 2022

Academic Editor: Federico Zambelli

Copyright © 2022 Yang Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background. Diffuse large B-cell lymphoma (DLBCL) is the most common subtype of non-Hodgkin’s lymphoma in adults, whose prognostic scoring system remains to be improved. Dysfunction of DNA repair genes is closely associated with the development and prognosis of diffuse large B-cell lymphoma. The aim of this study was to establish and validate a DNA repair-related gene signature associated with the prognosis of DLBCL and to investigate the clinical predictive value of this signature. Methods. DLBCL cases were obtained from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases. One hundred ninety-nine DNA repair-related gene sets were retrieved from the GeneCards database. The LASSO Cox regression was used to generate the DNA repair-related gene signature. Subsequently, the level of immune cell infiltration and the correlation between the gene signature and immune cells were analyzed using the CIBERSORT algorithm. Based on the Genomics of Drug Sensitivity in Cancer (GDSC) database, the relationship between the signature and drug sensitivity was analyzed, and together with the nomogram and gene set variation analysis (GSVA), the value of the signature for clinical application was evaluated. Results. A total of 14 DNA repair genes were screened out and included in the final risk model. Subgroup analysis of the training and validation cohorts showed that the risk model accurately predicted overall survival of DLBCL patients, with patients in the high-risk group having a worse prognosis than patients in the low-risk group. Subsequently, the risk score was confirmed as an independent prognostic factor by multivariate analysis. Furthermore, by CIBERSORT analysis, we discovered that immune cells, such as regulatory T cells (Tregs), activated memory CD4+ T cells, and gamma delta T cells showed significant differences between the high- and low-risk groups. In addition, we found some interesting associations of our signature with immune checkpoint genes (CD96, TGFBR1, and TIGIT). By analyzing drug sensitivity data in the GDSC database, we were able to identify potential therapeutics for DLBCL patients stratified according to our signature. Conclusions. Our study identified and validated a 14-DNA repair-related gene signature for stratification and prognostic prediction of DLBCL patients, which might guide clinical personalization of treatment.

1. Introduction

Diffuse large B-cell lymphoma (DLBCL), with an estimated 150,000 new cases per year worldwide, is the most common subtype of non-Hodgkin lymphoma among adults, accounting for 30%-40% of all cases [1]. Due to its heterogeneity in clinical manifestations, histomorphology, and molecular and genetic levels, approximately 40% of patients still cannot achieve long-term survival with standard treatment rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) [2, 3]. The prognosis of patients with recurrence or refractory after first-line treatment is poor; only a minority of these patients can benefit from “salvage treatments” consisting of high-dose chemotherapy and autologous stem cell transplantation (ASCT) [2, 4]. Therefore, to improve prognosis, it is essential to stratify patients, according to risk level at the time of diagnosis and individualized treatment.

The International Prognostic Index (IPI) is a clinically validated tool for estimating the risk of death in patients with DLBCL that takes into account a variety of factors, including patient’s age, performance status, number of...
extranodal sites of disease, hemoglobin level, lactate dehydrogenase level, and presence or absence of B symptoms [5]. However, since DLBCL is a genetic and molecular heterogeneity disease with a wide range of clinical presentations and outcomes, the IPI does not reflect the underlying molecular and biological mechanisms [6].

DNA damage can be a result of a variety of different factors, including exposure to ultraviolet light, radiation, and carcinogens. DNA repair or DNA damage response plays a crucial role in fixing DNA damage and maintaining genetic stability. In some types of cancer, the DNA repair machinery is defective, and cells are more likely to mutate, which allows the mutated gene products to exert their effects without restraint. And DLBCL is a kind of cancer that could arise when the DNA repair machinery malfunctions. In recent years, the role of DNA repair genes in tumorigenesis and prognosis of DLBCL has been extensively studied. For example, F53 is an important DNA damage response gene, which has been found mutated in about 20% of cases of DLBCL [7, 8]. Moreover, mutations of the TP53 tumor suppressor gene have been reported to be associated with poor survival in DLBCL [8, 9]. Another DNA damage response gene, ATM, which is one of the most frequently mutated genes in DLBCL, was correlated with a worse prognosis and proposed to be a potential new prognostic biomarker [10]. However, most studies have examined the role of DNA repair genes in DLBCL independently without considering the simultaneous changes of multiple genes. Therefore, to date, no comprehensive signature based on DNA repair genes has been constructed yet.

In the current study, we aimed to identify novel prognostic biomarkers for DLBCL that are associated with DNA repair genes. We established a prognosis-related signature based on 14 DNA repair genes through LASSO Cox regression analysis, which could predict the prognosis of DLBCL patients. In addition, we also found that the signature we constructed was highly correlated with immune cell infiltration and drug sensitivity. Finally, we created a predictive nomogram, as well as a ceRNA network with the DNA repair gene signature. In summary, we have constructed a novel DNA repair gene-related prognostic model, which might provide new insights into the prediction and treatment of DLBCL in clinics.

2. Materials and Methods

2.1. Data Collection. Series Matrix File data for GSE31312 were downloaded from the NCBI Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). The GSE31312 is based on the GPL570 Affymetrix Human Genome U133 Plus 2.0 Array, which consists of 470 DLBCL patients with complete expression profiles and a follow-up time of >0 days. A total of 47 original DLBCL mRNA expression data was obtained from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/) as a validation cohort to examine the signature we established.

2.2. Establishment and Validation of Prognostic Model. After DNA repair genes were selected, univariate Cox analysis was first performed to identify potential DNA repair genes significantly related to prognosis \((p < 0.001)\). Then, these selected prognosis-related genes from the univariate Cox analysis were further included in the LASSO Cox regression analysis to build an optimal prognostic model using the R package “glmnet” [11]. Finally, a risk score formula was created by combining the relative expression of each gene and the gene expression levels weighted by the regression coefficients. Based on the risk score formula, patients were stratified into low-risk and high-risk groups using the median risk score as the cut-off value. Kaplan-Meier (K-M) survival curves and a log-rank test were used to assess the survival differences between the two groups. The time-dependent receiver operating characteristic (ROC) curves of 1-year, 2-year, and 3-year survival were performed to evaluate the validity of the model predictions. Survival analysis and the ROC curves were based on the survival and survivalROC packages [12]. The R package “rms” was used to generate the nomogram and calibration map.

2.3. Functional Annotation via Gene Ontology (GO) Analyses. The functional annotation of the prognostic genes was performed using the R package “ClusterProfiler” (version 3.14.3) to comprehensively investigate the functional relevance of these prognostically relevant genes [13]. The adjusted \(p < 0.05\) and \(q < 0.05\) were considered significant for GO analysis.

2.4. The Correlation Analysis between Prognostic Model and Drug Sensitivity. Based on the largest pharmacogenomic database, Genomics of Drug Sensitivity in Cancer (GDSC; https://www.cancerrxgene.org/), we used the R package “pRProphetic” to predict the chemosensitivity of each tumor sample by evaluating the half maximum inhibitory concentration (IC50) of each specific chemotherapeutic drug. The IC50 of each drug was estimated by ridge regression analysis, and the prediction accuracy was evaluated by 10-fold cross-validation with the GDSC training set. Default values were selected for all parameters, including “combat” to remove the batch effect, and duplicate gene expression was summarized as the mean [14].

2.5. Immune Infiltration Analysis. Normalized GSE31312 expression data were analyzed using the CIBERSORT algorithm (CIBERSORT R script v1.03) to estimate the relative proportion of 22 infiltrating immune cells [15], and Spearman correlation analysis was performed to evaluate the potential correlation between the gene signature and immune cells. \(p < 0.05\) was considered statistically significant.

2.6. Gene Set Variation Analysis (GSVA). Gene set variation analysis (GSVA) is a nonparametric, unsupervised method for assessing transcriptomic gene set enrichment. GSVA converts a gene-by-sample gene expression matrix into a gene set-by-sample pathway enrichment matrix by synthetically assessing gene sets of interest to determine the biological function of samples [16]. In this study, gene sets were downloaded from the Molecular Signatures Database (version 7.0), and the GSVA R package (version 1.34.0) was used.
to calculate the score of both low- and high-risk groups for individual samples.

2.7. ceRNA Network. Competitive endogenous RNAs (ceRNAs) have attracted considerable academic interest in recent years and represent a new model for regulating gene expression. Compared with the miRNA regulatory network, the ceRNA regulatory network is more sophisticated and complex involving more RNA molecules, including mRNA, pseudogenes, miRNA, IncRNA, etc. [17]. The NPInter database (version 4.0) is commonly used to query the relationship between IncRNA and miRNA. In this study, the NPInter database was used to predict IncRNA-miRNA interactions. In addition, FunRich software (version 3.1.3) was used to reversely predict the interactions between mRNA and miRNA. Then, the IncRNA-miRNA-mRNA network was constructed by combining IncRNA-miRNA interactions and mRNA-miRNA interactions and visualized using Cytoscape (version 3.7.1).

2.8. Motif Enrichment Analysis. Gene expression is regulated by transcription factors (TFs). Different transcription factors regulate different genes, which in turn cause cells to exhibit different states or types. To identify binding motifs of transcription factors enriched on 14 model genes, RcisTarget (version 1.6.0) was used to perform TF motif enrichment analysis to identify overrepresented TF binding motifs and to identify candidate transcription factors, where we applied a subset of 4.6 k motifs (rcistarget.hg19.motifdb.cisbpont.500bp) to the gene-motif ranking database [18, 19].

2.9. Statistical Analysis. K-M survival curves and log-rank test were performed to compare the differences in overall survival (OS) between 2 different groups. The univariate and multivariate Cox proportional hazard models were used to calculate the hazard ratio (HR) of the variables and to determine independent prognostic factors of DLBCL. R software (version 3.6) was used for all statistical analyses. All statistical tests were two-sided, and \( p < 0.05 \) was considered statistically significant.

3. Results

3.1. Identification of Prognosis-Related DNA Repair Genes in DLBCL. We downloaded the original mRNA expression data (FPKM) of DLBCL from the GEO database \((n = 470)\) and obtained a total of 199 DNA repair-related gene sets (relevance score > 20) from the GeneCards database (https://www.genecards.org/) in the meantime, of which 186 were included in the data from GEO. To further identify key genes in DNA repair-related gene sets, we first screened out 29 prognosis-related DNA repair genes \((p < 0.001)\) for further investigation by collecting survival data from DLBCL patients using univariate Cox analysis (Figure 1(a), Table S1).

Twenty-nine prognosis-related DNA repair genes were enriched by GO analysis, and these genes were mainly associated with the binding of damaged DNA, response to radiation, telomere maintenance, catalytic activity, action on DNA, replication fork, DNA repair complex, and other pathways (Figure 1(b)). We also constructed a PPI network of these prognostic genes that was visualized using Cytoscape software (Figure 1(c)). The PPI network included 265 nodes and 1109 edges, of which RPA3 was identified as the most highly linked protein with the highest degrees (Table S2). The presence of a large number of links in this network indicates that these proteins interact frequently, suggesting that they may influence the prognosis of DLBCL patients.

3.2. Selection of the DNA Repair Genes and Construction of the Prognostic Gene Signature. The 29 selected prognosis-related DNA repair genes were further included in the LASSO Cox regression analysis to eliminate the overfitting genes and construct an optimal prognosis-related signature for DLBCL. Patients in the GEO database were randomly divided into a training cohort and a validation cohort in a 4:1 ratio. We found that the prognostic model achieved the best performance when the 14 DNA repair genes were included (Figure 2(a)). Then, the optimal regression coefficients of these 14 genes were determined via LASSO Cox regression analysis (Figures 2(b) and 2(c), Table S3) for subsequent modelling (risk score = XRCC2 \*(-0.27798) + TMYMP \*(-0.06018) + RAD1* (0.011479) + PCNA* (0.015372) + UBC* (0.027425) + CDT1* (0.03864) + sCO2* (0.041364) + POLK* (0.049191) + DDB2* (0.055706) + WRN* (0.06327) + RPA3* (0.063596) + PRIMPOL* (0.116672) + RAD23B* (0.12546) + PMS1* (0.160487)). Based on the risk score, patients were divided into a high-risk group and a low-risk group. K-M survival analysis showed that the OS of the high-risk group was significantly lower than that of the low-risk group in both training and validation cohorts \((p < 0.001,\) Figures 3(a) and 3(b)). Moreover, the areas under the ROC curve (AUC) for 1 year, 2 years, and 3 years in the training cohort were 0.7, 0.74, and 0.74, respectively. And the AUC of the internal validation cohort was 0.94 at 1 year, 0.69 at 2 years, and 0.69 at 3 years, indicating that the model had good verification efficiency (Figures 3(c) and 3(d)).

3.3. External Validation. DLBCL patients from the TCGA database were downloaded to evaluate the robustness of the prognostic model we created \((n = 47)\). Based on the risk score, patients were divided into a high-risk group \((n = 23)\) and a low-risk group \((n = 24)\). The difference in survival between the two groups was assessed by K-M analysis. We found that the results of the TCGA validation cohort were consistent with those of the entire GEO cohort, suggesting that high-risk patients were associated with an inferior prognosis (Figure 4(a)). To further determine the predictive accuracy of the external validation dataset, we constructed ROC curves, which showed strong efficacy of the model in predicting prognosis (AUC of 1, 2, and 3 years were 0.73, 0.82, and 0.82, respectively) (Figure 4(b)).

3.4. The Relationship between the Prognostic Signature and the Immune Microenvironment. The tumor microenvironment is mainly composed of tumor-associated fibroblasts, immune cells, extracellular matrix, multiple growth factors, inflammatory factors, and specific physicochemical features (e.g., hypoxia and low pH), as well as the cancer cells
<table>
<thead>
<tr>
<th>Gene</th>
<th>( p ) value</th>
<th>Hazard ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPA3</td>
<td>&lt; 0.001</td>
<td>1.476 (1.302–1.673)</td>
</tr>
<tr>
<td>PRIMPOL</td>
<td>&lt; 0.001</td>
<td>1.382 (1.226–1.559)</td>
</tr>
<tr>
<td>HUS1</td>
<td>&lt; 0.001</td>
<td>1.368 (1.212–1.544)</td>
</tr>
<tr>
<td>RAD1</td>
<td>&lt; 0.001</td>
<td>1.328 (1.187–1.487)</td>
</tr>
<tr>
<td>RAD23B</td>
<td>&lt; 0.001</td>
<td>1.407 (1.224–1.619)</td>
</tr>
<tr>
<td>ERCC1</td>
<td>&lt; 0.001</td>
<td>1.344 (1.189–1.518)</td>
</tr>
<tr>
<td>WRN</td>
<td>&lt; 0.001</td>
<td>1.323 (1.171–1.496)</td>
</tr>
<tr>
<td>XRCC4</td>
<td>&lt; 0.001</td>
<td>1.307 (1.161–1.473)</td>
</tr>
<tr>
<td>PRIM2</td>
<td>&lt; 0.001</td>
<td>1.359 (1.185–1.559)</td>
</tr>
<tr>
<td>PCNA</td>
<td>&lt; 0.001</td>
<td>1.421 (1.214–1.662)</td>
</tr>
<tr>
<td>PRIM1</td>
<td>&lt; 0.001</td>
<td>1.305 (1.150–1.481)</td>
</tr>
<tr>
<td>DDB2</td>
<td>&lt; 0.001</td>
<td>1.290 (1.143–1.456)</td>
</tr>
<tr>
<td>SUCLA2</td>
<td>&lt; 0.001</td>
<td>1.345 (1.166–1.552)</td>
</tr>
<tr>
<td>DCLRE1C</td>
<td>&lt; 0.001</td>
<td>1.318 (1.153–1.506)</td>
</tr>
<tr>
<td>XRCC2</td>
<td>&lt; 0.001</td>
<td>0.836 (0.766–0.912)</td>
</tr>
<tr>
<td>ALKBH1</td>
<td>&lt; 0.001</td>
<td>1.254 (1.121–1.402)</td>
</tr>
<tr>
<td>RNF168</td>
<td>&lt; 0.001</td>
<td>1.280 (1.130–1.449)</td>
</tr>
<tr>
<td>ATR</td>
<td>&lt; 0.001</td>
<td>1.245 (1.112–1.394)</td>
</tr>
<tr>
<td>SCO2</td>
<td>&lt; 0.001</td>
<td>1.260 (1.118–1.420)</td>
</tr>
<tr>
<td>MLH1</td>
<td>&lt; 0.001</td>
<td>1.233 (1.100–1.383)</td>
</tr>
<tr>
<td>POLE2</td>
<td>&lt; 0.001</td>
<td>1.244 (1.102–1.404)</td>
</tr>
<tr>
<td>PMS1</td>
<td>&lt; 0.001</td>
<td>1.226 (1.095–1.372)</td>
</tr>
<tr>
<td>MBD4</td>
<td>&lt; 0.001</td>
<td>1.220 (1.089–1.366)</td>
</tr>
<tr>
<td>UBC</td>
<td>&lt; 0.001</td>
<td>1.429 (1.166–1.732)</td>
</tr>
<tr>
<td>RAD51C</td>
<td>&lt; 0.001</td>
<td>1.246 (1.097–1.415)</td>
</tr>
<tr>
<td>TYMP</td>
<td>&lt; 0.001</td>
<td>0.818 (0.728–0.919)</td>
</tr>
<tr>
<td>CDT1</td>
<td>&lt; 0.001</td>
<td>1.231 (1.090–1.391)</td>
</tr>
<tr>
<td>XPA</td>
<td>&lt; 0.001</td>
<td>1.233 (1.090–1.394)</td>
</tr>
<tr>
<td>POLK</td>
<td>&lt; 0.001</td>
<td>1.212 (1.082–1.358)</td>
</tr>
</tbody>
</table>

**Figure 1:** Continued.
Figure 1: Identification of DNA repair genes associated with prognosis. (a) Forest plot shows the univariate Cox results of the 29 prognosis-related DNA repair genes ($p < 0.001$). (b) The GO enrichment analysis. BP: biological process; CC: cellular component; MF: molecular function. (c) PPI network indicating the interactions of the prognosis-related DNA repair genes.
themselves, which significantly affect the biological behavior, functions, and multidrug resistance of cancer cells. We further explored the potential molecular mechanism affecting the progression of diffuse large B-cell lymphoma by analyzing the relationship between the prognostic signature and tumor immune infiltration. The immune cell infiltration level of each sample was quantified using the CIBERSORT algorithm, and the correlation between 22 immune cells was further analyzed (Figures 5(a) and 5(b)), which showed that naive B cells were negatively correlated with memory B cells ($r = -0.44$) and gamma delta T cells were negatively related with resting NK cells ($r = -0.43$). Furthermore, we compared the differences in immune cell content between the low-risk group and the high-risk group. As shown, patients in the high-risk group had significantly fewer memory B cells ($p = 0.044$), naive CD4+ T cells ($p = 0.001$), and regulatory T cells (Tregs) ($p < 0.001$), while naive B cells ($p = 0.006$), activated memory CD4+ T cells ($p < 0.001$), and gamma delta T cells ($p < 0.001$) were significantly higher.

The relationship between the signature and immunoregulatory genes was further analyzed. The differences in the expression of genes such as immune-related chemokines, immune inhibitors, immune stimulators, major histocompatibility complex (MHC), and immune receptors between high- and low-risk groups are shown in Figures 6(a)–6(e). We also examined the correlation between risk score and immunoinhibitory genes. The correlation heat map showed that several genes such as CD96, TGFBR1, and TIGIT were strongly positively correlated with the risk score (Figure 6(f)).

3.5. Activation of Signaling Pathways Related to the Signature. GSVA was used to identify the specific signaling pathways involved in the high- and low-risk groups to explore the potential molecular mechanisms by which signature influences tumor progression. The GSVA results showed that the
The high-risk group was enriched in the signaling pathways MYC_TARGETS, DNA_REPAIR, P53_PATHWAY, and OXIDATIVE_PHOSPHORYLATION, which are associated with DNA damage and DNA repair, and the abnormality of these pathways played a crucial role in tumorigenesis. Meanwhile, signaling pathways such as WNT_BETA_CATENIN_SIGNALING, KRAS_SIGNALING_DN, and MYOGENESIS were enriched in the low-risk group (Figure 7(a)). All these showed that the alterations of these signaling pathways affected the prognosis of patients in the high- and low-risk groups, indicating that the prognostic model we constructed is representative.

3.6. The Relationship between Signature and Drug Sensitivity.
Immunochemotherapy, as one of the common treatments for DLBCL, is typically used in combination with common targeted therapies such as CD20 antibodies and tyrosine kinase inhibitors. Sensitivity to immunochemotherapy is also an important prognostic factor for DLBCL. Chemotherapy in combination with immunotherapy (e.g., rituximab) has been shown to be effective in the treatment of DLBCL. Based on the drug sensitivity data from the GDSC database, we further investigated the correlation between the signature and sensitivity to common chemotherapeutic agents and targeted drugs. As shown in Figure 7(b), the high-risk group
was more sensitive to mitomycin C, while the IC50 values of gefitinib, dasatinib, and imatinib were significantly higher in the high-risk group than in the low-risk group, suggesting that DLBCL patients in the high-risk group were more resistant to these molecularly targeted drugs; i.e., the high-risk group was less sensitive to gefitinib, dasatinib, and imatinib. Our signature correlated with sensitivity to multiple drugs, which may indicate that the signature has potential predictive properties for sensitivity to chemotherapy. Overall, patients in the high-risk group may benefit more from chemotherapeutic agents than from these tyrosine kinase inhibitors (TKIs).

3.7. Incidence Risk and Independent Prognostic Analysis. To demonstrate the role of the signature in assessing the incidence risk and prognosis of DLBCL, we constructed a predictive nomogram based on age, gender, Ann Arbor staging (AAS), N extranodal, ECOG, IPI, and the risk score derived from our signature. By converting the regression coefficient of each variable from multivariate logistic regression analysis into a 0- to 100-point scale, we created the alignment diagram to predict the 5- and 7-year survival probability of DLBCL patients. Among all variables, the risk score contributed significantly to the nomogram prediction model, and the risk score was assigned 100 points in our nomogram (Figure 8(a)). The predictive accuracy of our nomogram was measured using calibration curves with 1000 bootstrap samples to reduce overfit bias by comparing the concordance between the OS predicted by the nomogram and the OS observed, underpinning its suitability for predicting survival in DLBCL (Figure 8(b)). In addition, we performed multivariate Cox regression analysis and confirmed that the risk score was an independent prognostic factor for DLBCL patients ($p < 0.001$) (Figure 9).

3.8. ceRNA Network. First, FunRich was used for reverse prediction of 14 model genes, and 50 miRNAs were predicted that formed 69 mRNA-miRNA pairs. Then, reverse prediction of lncRNAs was performed using NPInter based on the 50 miRNAs, resulting in a total of 10,248 mRNA-miRNA-lncRNA pairs. To further screen the most important lncRNAs, we used maximum information coefficients (MIC) to calculate the correlation between lncRNAs and model genes and screened out 1785 lncRNAs ($\text{cor} > 0.2$). These 1785 lncRNAs were crossed with 10,248 mRNA-miRNA-lncRNA pairs to finally obtain 948 mRNA-miRNA-lncRNA pairs and construct a comprehensive ceRNA network associated with model genes (Figure 10).

3.9. Motif Enrichment Analysis. We used 14 model genes for motif enrichment analysis and found that they were regulated by multiple transcription factors. After a series of analyses to calculate enrichment using a recovery-based method, selecting significant motifs and motif TF annotation, and identifying the genes with the best enrichment for each motif, the transcription factor NRF1 emerged as the most important regulator in our gene set, MOT annotated as cisbp__M5688, in which a total of five model genes were enriched in this motif with a normalized enrichment score (NES) of 9.75. Here, we show all enriched motifs and the corresponding transcription factors of the model genes (Figures 11(a) and 11(b), Figure S1).

4. Discussion

DLBCL is the most common subtype of adult non-Hodgkin’s lymphoma, with a 40% recurrence rate in the era of rituximab [20, 21]. To improve the prognosis of patients with DLBCL, it is critical to identify novel
Figure 5: Immune cell distribution profiles. (a) Landscape of the composition of the 22 immune cells in each sample. (b) Correlation heat map of the 22 immune cells. The blue boxes indicate negative correlations, and the red boxes indicate positive correlations. (c) Violin plot comparing the proportion of immune cells between the high- and low-risk groups based on the risk score. The yellow bars represent the low-risk groups, and the blue bars represent the high-risk groups. (d) Correlation between risk score and infiltrating immune cells.
Figure 6: Continued.
Figure 6: Continued.
prognostic factors that can predict disease progression. Traditional prognostic stratification schemes such as the IPI and cell origin classification (COO) can be used to classify patients into clinical risk groups, which in turn helps in clinical decision making. However, these classification schemes have several limitations, and a consensus is emerging that DLBCL with the same clinical features, but different molecular and genetic characteristics, may represent a distinct entity and respond differently to specific therapy. Therefore, a more refined and accurate classification system is needed. In this study, a molecular classification for DLBCL was established to improve the prognostic stratification of patients with DLBCL. In addition, the important role of DNA repair genes in the development and progression of DLBCL was identified, which can be used as a prognostic factor.

Recent advances in molecular and cytogenetic techniques have allowed refinement of the classification of DLBCL. Several distinct genetic abnormalities, including somatic mutations in genes such as MYD88, CD79B, and MYC, have been identified in DLBCL, and single or combined mutations have been shown to have prognostic value in DLBCL subtypes [22–24]. Bioinformatics has been at the forefront in advancing the field of cancer research and has been applied to many different areas, including the prognosis of diffuse large B-cell lymphoma. For example, TME immune cell signatures [21] and IncRNA signatures [25] have been identified for subtype classification and prognosis prediction of DLBCL patients. In addition, some integrated models such as the model incorporating pharmacogenomic gene signatures and clinical information as well as the IPI-based immunoprediction model have been validated to have better predictive power for DLBCL patients [26]. However, there is no study yet to declare whether there is a DNA repair-related prognostic signature capable of predicting the outcome of DLBCL.

DNA repair or DNA damage response (DDR) is a cellular program that detects DNA damage and triggers DNA repair to protect genetic material from damage. In a subset of DLBCL, mutations of genes involved in DNA damage response and DNA repair genes are identified, and defects of some DNA repair factors are associated with patient prognosis [27–29]. The aim of this study was to evaluate the prognostic value of DNA repair genes that play important roles in DLBCL and to develop a prognostic signature that can stratify DLBCL patients. First, we identified twenty-nine DNA repair genes that significantly correlate with overall survival in DLBCL patients. We then subjected them to LASSO Cox regression analysis to generate a signature of 14 DNA repair-related genes. Patients were divided into low-risk and high-risk groups according to the prognostic risk score formula. K-M analysis estimated survival probabilities and showed that the prognostic model based on these 14 DNA repair genes, including XRCC2, TYMP, RAD1, PCNA, UBC, CDT1, sCO2, POLK, DDB2, WRN, RPA3, PRIMPOL, RAD23B, and PMS1, was useful and robust in predicting survival in both our training and validation.
FIGURE 7: Continued.
GEO-DLBCL cohorts as well as the external validation TCGA-DLBCL cohort. In addition, based on the risk score and other independent factors, we constructed a predictive nomogram to predict the prognosis of DLBCL patients to estimate their survival probability and identify high-risk patients. Among all variables, risk score significantly contributed to survival prediction. Moreover, the risk score based on the prognostic model proved to be a useful independent prognostic factor by multivariate analysis. Furthermore, we constructed the DNA repair-associated ceRNA network based on the hypothesis that RNAs interact by competing for a limited pool of microRNAs to reveal a novel mechanism of interaction between RNAs and probably provide insights for further studies on potential lncRNA biomarkers in DLBCL [17]. Finally, motif enrichment analysis was performed to identify the mechanisms controlling the transcription of these DNA repair genes. Interestingly, the transcription factor NRF1 was found to be the major regulator in our gene set. NRF1, a transcription factor, has been shown to be an important component of the DNA repair network in response to UVB-induced DNA damage, and NRF1 can also regulate circRNAs to promote lymphoma progression [30, 31]. All these suggest that the DNA repair gene signature is a robust prognostic stratification system that may serve as a novel prognostic indicator for DLBCL.

X-ray Repair Cross-Complementing 2 (XRCC2) is one of many proteins that regulate the repair of DNA double-strand breaks (DSBs) by homologous recombination [32]. Studies on genetic variations in XRCC2 (rs3218536, Arg188His) have revealed some evidence of association with breast cancer risk and poor survival [33–35]. Thymidine phosphorlyase (TYMP) is a catabolic enzyme in thymidine metabolism that has been attributed a dual role in promoting angiogenesis and inhibiting apoptosis in many tumors [36, 37]. In DLBCL, TYMP has been shown to correlate with the development of a nongerminal center origin and a worse outcome [38]. It has been previously demonstrated that RAD1 can form a heterotrimeric complex (9-1-1 complex) that is activated to arrest cell cycle progression in response to DNA damage or incomplete DNA replication [39]. In ovarian cancer, RAD1 acts as a tumor suppressor in a BRCA-like manner [40]. Proliferating Cell Nuclear Antigen (PCNA) is a protein known to play a key role in DNA replication and repair in non-Hodgkin’s lymphoma. It interacts with many different proteins involved in replication, repair, and tumorigenesis and is an important prognostic factor in non-Hodgkin’s lymphoma [41–43]. Ubiquitin C (UBC) attaches to abnormal proteins to mark them for destruction by the degradation process known as ubiquitination. UBC has a variety of functions, a major one of which is the regulation of DNA repair [44].

Figure 7: GSVA analysis and drug sensitivity. (a) Differences in signaling pathways enriched between the high-risk and low-risk groups. Blue bars indicate activated signaling pathways in the high-risk group, and green bars indicate activated signaling pathways in the low-risk group. (b) Comparison of IC50 of chemotherapeutic agents and targeted drugs between the high-risk and low-risk groups.
Figure 8: Predictive nomogram. (a) Nomogram predicting 5-year and 7-year survival probabilities in DLBCL patients. (b) Calibration curves of observed and predicted OS, to verify the accuracy of the nomogram.
UBC has also been found in many human cancer specimens [45, 46]; however, the relationship between UBC and DLBCL needs further research. The checkpoint protein chromatin licensing and DNA replication factor 1 (CDT1) is a key regulator of DNA replication licensing and DNA damage repair [47]. Deregulation of CDT1 has been shown to trigger a DNA damage response that contributes to genomic instability, and CDT1 has been reported to function as an oncogene in various human cancers [48–50]. However, the relationship between CDT1 and DLBCL is still unknown.

sCO2 is essential for the proper assembly of cytochrome c oxidase (COX), the terminal component of the mitochondrial respiratory chain [51]. As a transcriptional target of p53, sCO2 plays an important role in oxidative phosphorylation [52]. To date, the effect of sCO2 in different cancer types remains controversial. Low sCO2 expression has been associated with significantly worse prognosis in patients with breast cancer, while high sCO2 expression predicted worse survival in human lung adenocarcinoma [53, 54]. DNA polymerase kappa (POLK) is a member of the POLK family of DNA polymerases specifically involved in DNA repair. Dysregulation of POLK can lead to genetic instability in human cells, which is associated with some cancers. However, the role of POLK in DLBCL has not been adequately evaluated [55–57]. DNA damage is recognized by specific proteins that can initiate a signal transduction cascade. One of these proteins is damage-specific DNA-binding protein 2 (DDB2), which plays an important role in the initial steps of nuclear excision repair (NER) [58]. DDB2 appears to have a dual function in cancer, sometimes with tumor suppressor properties and sometimes as an oncogene [59–61]. It has been reported that the expression of WRN is related to the expression of Myc oncoprotein, which has been implicated in the pathogenesis of DLBCL [62, 63]. Replication protein A3 (RPA3) is part of the heterotrimeric replication protein A complex, which plays an essential role in DNA replication and DNA repair by binding to single-stranded DNA and damaged DNA on chromatin. RPA3 expression has been shown to correlate with poor prognosis and radioresistance in various cancers [64–66]. PRIMPOL is thought to contribute to cellular tolerance to DNA damage by facilitating damage bypass [67]. Further studies have shown that loss of PRIMPOL leads to increased cellular sensitivity to DNA cross-linking agents such as

![Forest plot of multivariable Cox analyses of the risk score and clinical characteristics.](image-url)
mitomycin C, suggesting that repriming of PRIMPOL may represent a novel target for improving the sensitivity of cancer cells to DNA-damaging chemotherapies [68, 69]. Rad23 homolog B plays a critical role in nucleotide excision repair (NER) in humans [70]. The study demonstrated that depletion of RAD23B abrogated the accumulation of ubiquitinated p53 and suppressed apoptosis after mitomycin C (MMC) exposure [71]. PMS1 protein homolog 1 (PMS1) is required...
Figure 11: Motif enrichment analysis. (a) AUC histogram estimating the overrepresentation of each motif on our model genes. (b) Selection of the best enriched genes for each motif. The red line is the mean of the recovery curve for each motif, the green line is the mean + standard deviation, and the blue line is the recovery curve for the current motif. The maximum distance point (mean + sd) between the current motif and the green curve is the selected maximum enrichment level.
for DNA mismatch repair, which is also involved in the repair of DNA damage caused by cisplatin, alkylating agents, and oxidation [72, 73]. Loss of function of PMS1 leads to microsatellite instability and colorectal cancer [74]. The role of this protein in DLBCL has not yet been established.

Additionally, to further explore the immune infiltration-related mechanism affecting DLBCL progression, the estimated proportions of 22 tumor-infiltrating immune cell types in DLBCL were evaluated by CIBERSORT analysis. It was found that memory B cells, naive CD4+ T cells and regulatory T cells (Tregs), naive B cells, activated memory CD4+ T cells, and gamma delta T cells showed significant differences between high- and low-risk groups. Recent studies suggest that the composition of the TME is essential for the pathogenesis of lymphoma [21, 75]. Regulatory T cells (Tregs) are important modulators of the interaction between lymphoma cells and the host microenvironment [76]. In DLBCL, intratumoral FOXP3+ Tregs are reported to be associated with superior prognosis, especially in the GCB type, but with poor prognosis in the non-GCB type [77, 78]. According to our results, patients in the high-risk group had significantly fewer regulatory T cells (Tregs), implying that Tregs may serve as suppressors. Gamma delta T cells (γδ T cells), as a bridge between innate and adaptive responses, can exhibit dual antitumor and protumor activity in cancer [79]. Although the γδ T cells make up only a minor proportion in TME, they have been demonstrated to be an important component of tumor-infiltrated lymphocytes in many types of cancer [80–82]. A previous study has shown a significant abundance of γδ T cells in DLBCL patients [83], but the role of γδ T cells in DLBCL is still unknown. In our study, we found that γδ T cells were positively correlated with the high-risk score, which may define the protumor role in DLBCL, but we need more evidence to further validate the function of γδ T cells in DLBCL. Activated memory CD4+ T cells are an important infiltrate in the microenvironment of DLBCL [84]. Tumors from DLBCL patients were enriched in CD4+ T memory cells that displayed high coexpression of TIGIT and PD-1, which may be potential targets for novel therapeutic intervention in DLBCL [85]. Immune checkpoint inhibition or blockade is a cancer treatment strategy that harnesses the power of the immune system to attack cancer cells. It has had a profound impact on the prognosis of many cancers and has recently been approved as an important treatment option for patients with DLBCL. Therefore, in our model, we investigated the relationship between risk score and immunoregulatory genes, especially genes related to immune checkpoints. We found that the expression of CD96, TGFBR1, and TIGIT was significantly increased in high-risk patients. Taken together, our analysis of immune infiltration suggests that the strong expression of immune inhibitors and immunosuppressive TME surrounding the tumor contribute to the poor prognosis of the high-risk group. Furthermore, we hypothesize that high-risk patients may respond more effectively to therapy with TIGIT, CD96, and TGFBR1 checkpoint inhibitors which provides us with additional clues for immunotherapy candidates.

Our GSVA analysis uncovered distinct molecular mechanisms between the low- and high-risk groups, with several DNA repair-related gene sets or signaling pathways, including the unfolded protein response pathway, the P53 pathway, and MYC targets, being most enriched in the high-risk group, whereas the Wnt/β-catenin and KRAS pathways were enriched in the low-risk group. MYC is thought to be associated with the development of multiple cancers, including DLBCL [86, 87]. And dysregulation of MYC can directly impair DNA repair and induce genomic instability [88, 89]. The P53 protein is an important tumor suppressor that plays a role in numerous cellular processes, including cell cycle arrest, apoptosis, and DNA repair. Conversely, loss of p53 function leads to loss of these processes, resulting in mutations and cancer development [90]. Unfolded protein response (UPR) is a conserved signaling pathway that responds to protein misfolding and the accumulation of unfolded proteins in cells [91]. When cells are under stress, both the UPR and the DNA damage response are activated, allowing cells to repair damage [92, 93]. Therefore, we hypothesize that the high-risk group may have a higher degree of malignancy, and thus, more damage occurs, leading to excessive activation of DNA repair signaling pathways. While the low-risk group showed a positive correlation with Wnt/β-catenin [94] and KRAS signaling [95], both are drivers or participants in various cancers, and their crosstalk between the DNA damage response has been validated, which may provide combination strategies for the treatment of cancer to overcome the dysregulation and deficits in DNA repair [96, 97]. In summary, these GSVA results describe the different molecules and mechanisms involved in the progress of low-risk and high-risk DLBCL, which provide a basis for future targeted therapeutic studies.

Subsequently, we performed a multimics analysis to investigate the clinical utility of the prognostic model. First, analysis of drug sensitivity data from the GDSC database showed that the high-risk group was more sensitive to mitomycin C, whereas gefitinib, dasatinib, and imatinib were more effective in the low-risk group. Dasatinib is a tyrosine kinase inhibitor that can inhibit Src family kinases and BTK. In vitro studies show that dasatinib is toxic to ABC DLBCL cell lines that rely on chronic active BCR signaling, suggesting that this drug is a candidate for the treatment of DLBCL [98, 99]. Imatinib, a PDGFRA-selective tyrosine kinase inhibitor, has been validated to be effective in blocking lymphoma growth in both human xenograft and murine allograft models [100]. Although mitomycin C is not a commonly used chemotherapeutic agent for the treatment of DLBCL, the role of these TKIs as a therapeutic strategy in DLBCL has been demonstrated. Of note, the sensitivity to these chemotherapeutic agents and targeted therapies could be evaluated by our risk model. This suggests that our model may guide the personalized treatment of DLBCL patients. Subsequently, the risk score of our model was included together with other widely used prognostic indicators to construct the predictive nomogram. The risk score contributed more to survival prediction compared with the other indicators and was also confirmed as an independent prognostic factor. Therefore, our risk model could be a promising tool for clinical applications in the future.
However, there are still some limitations in our current study. Our risk model derived from bioinformatic analysis needs to be further confirmed by functional and mechanistic experiments, as only three of the genes included in the signature, TYMP, PCNA, and WRN, have been previously studied as prognostic markers for DLBCL. In addition, our current study is retrospective, so prospective studies are needed to further confirm our findings.

5. Conclusions

In summary, our study established a novel robust 14-gene signature based on DNA repair genes that could be used for the stratification and prognostic prediction of DLBCL patients. Our study also provided new insights into the relationship between this signature and the immune microenvironment as well as immunotherapy to guide the clinical treatment of DLBCL patients.

Abbreviations

DLBCL: Diffuse large B-cell lymphoma
TCGA: The Cancer Genome Atlas
GEO: Gene Expression Omnibus
GDSC: Genomics of Drug Sensitivity in Cancer
GSVA: Gene set variation analysis
ASCT: Autologous stem cell transplantation
IPI: International Prognostic Index
ROC: Receiver operating characteristic
IC50: Half-maximal inhibitory concentration
ceRNA: Competing endogenous RNA
OS: Overall survival
HR: Hazard ratio
AUC: Area under the curve
MHC: Major histocompatibility complex
AAS: Ann Arbor staging
COO: Cell of origin.

Data Availability

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

Conflicts of Interest

The authors report no conflicts of interest in this work.

Authors’ Contributions

Yang Li and Xiyang Liu contributed equally to this work.

Acknowledgments

The present study was supported by the Medical Science and Technology Project of Henan Province in 2019 (grant no. LHGJ20190029).

Supplementary Materials

Supplementary 1. Table S1: 29 prognosis-related DNA repair genes were significantly predictive of prognosis in DLBCL patients by univariate Cox regression analysis (p < 0.001).
Supplementary 2. Table S2: top 28 genes in the PPI network.
Supplementary 3. Table S3: coefficients of 14 selected DNA repair genes.
Supplementary 4. Figure S1: the motif enrichment information and its annotation.

References


