Identification of Predicting Diagnostic Gene Biomarkers through Machine Learning combined with Pan-cancer in Patients with Systemic Lupus Erythematosus

Jinfeng Zhan
Health Management Medicine Department, The Second Affiliated Hospital of Nanchang University, Nanchang, China

Ruoying Cheng
The Second Clinical Medical College of Nanchang University, Nanchang, Jiangxi, China

Qi Liu
The Second Clinical Medical College of Nanchang University, Nanchang, Jiangxi, China

Yuxin Zu
The Second Clinical Medical College of Nanchang University, Nanchang, Jiangxi, China

Kaibo Hu
The Second Clinical Medical College of Nanchang University, Nanchang, Jiangxi, China

Zhongbin Xia (✉ 408979872@qq.com)
Health Management Medicine Department, The Second Affiliated Hospital of Nanchang University, Nanchang, China

Article

Keywords: SLE, machine learning, diagnostic gene biomarkers, pan-cancer, immune cell infiltration

Posted Date: April 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2801641/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Background

Early diagnosis of systemic lupus erythematosus (SLE) is crucial for treatment and reducing mortality. In this research, we set out to explore several important biomarkers for the diagnosis of SLE through machine learning combined with pan-cancer, and to investigate the significance of immune cell infiltration.

Methods

Gene expression profiles for three human SLE and control samples were downloaded from the GEO database. The LASSO regression model and support vector machine recursive feature elimination (SVM-RFE) analysis were used to identify candidate biomarkers. The area under the receiver operating characteristic curve (AUC) value was obtained for the evaluation of the discriminations. The GSE20864 dataset was then further validated for the expression and diagnostic value of SLE biomarkers. The structure of 22 immune cell fractions in SLE was explored through the pooled cohort of CIBERSORT. The Cancer Genome Atlas (TCGA), Treatment-Applicable Research to Generate Effective Treatments (TARGET) and Genotype-Tissue Expression (GTEx) datasets are sources of pan-cancer and normal tissue data. And associated pan-cancer analysis was carried out.

Results

We identified a total of 52 differently expressed genes, of which 23 were significantly upregulated and 29 were significantly downregulated. After further screening, we found four genes as candidate biomarkers (FOS, IFI27, ANKH, and IGF2BP2). Three signature genes (ANKH, IFI27, IGF2BP2) are excellent diagnostic values in treat and control groups. In immune cell analysis, 11 immune cells showed significant differences between SLE patients and normal humans. We also analysed the effect of IGF2BP2 in the tumour process. The IGF2BP2 gene was low expressed in 9 tumors and high expressed in 22 tumors. In addition, we also found a strong correlation between IGF2BP2 expression in tumour cells and patient prognosis. Our study also indicates that IGF2BP2 may have the ability to predict the efficacy of ICI in the corresponding cancer.

Conclusion

We explored several important biomarkers for the diagnosis of SLE through machine learning and found that IGF2BP2 has a sufficient role in the immune microenvironment of most of the tumors.

Key Points

We explored several significant biomarkers for SLE diagnosis through machine learning and fined that IGF2BP2 plays a sufficient role in the immune microenvironment of most tumors.
Impact Statement

The manuscript aimed to explore several significant biomarkers for SLE diagnosis through machine learning combined with Pan-cancer. And research the significance of immune cell infiltration. We explored several significant biomarkers for SLE diagnosis through machine learning and fined that IGF2BP2 plays a sufficient role in the immune microenvironment of most tumors. We have subjected the screened genes to pan-cancer analysis to assess their relevance to clinical prognosis and associated signaling pathways by conducting pan-cancer analysis, which has the potential to identify new immunotherapeutic targets. This could provide new ideas for the diagnosis of SLE.

Introduction

Systemic lupus erythematosus (SLE) is a classic auto-immune connective tissue disease in which the immune system attacks healthy tissue, usually involving the skin, kidneys, joints, lungs, blood and central nervous system [1, 2]. It is distinguished by the overproduction of autoantibodies and a systemic or organ-specific inflammatory response, leading to damage to a number of tissues and organs, including the skin, heart, lungs, intestines and kidneys [3–6]. The pathological damage is a direct result of the excessive production of immune complexes and their impaired clearance and consequent deposition in the tissue, which makes SLE a harmful disease with high morbidity and mortality [7]. According to statistics, SLE is most common in people aged 15 to 45 years, especially women of childbearing age, with a prevalence ratio of 1:7 to 1:9 for both sexes [8]. The global prevalence of SLE is about (30–150) per 100,000 people. The highest estimation of the incidence and a prevalence of SLE were in North America [23.2/100 000 person-years (95% CI: 23.4, 24.0) and 241/100 000 people (95% CI: 130, 352), respectively] [9]. In Asia, the annual incidence of SLE ranges from 2.8 to 8.6 per 100,000 person-years, with a prevalence of 26.5 to 103 per 100,000 person-years [10]. To improve the clinical management and survival of patients with SLE, a rapid and accurate diagnosis of SLE should be the first thing on our minds. Further laboratory tests including high erythrocyte sedimentation rate, antiphospholipid antibodies, anti-dsDNA antibodies, and anti-nuclear antibodies (ANA) can help dragonize SLE [11]. However, this does not allow for an early diagnosis of SLE. The reasons for the disease remain uncertain, with many studies suggesting that genetic, endocrine, infectious, immune abnormalities, and some environmental factors are involved in the development of SLE [12]. In particular, genetic factors play a crucial role in the pathogenesis of AMI. It is therefore important to explore the molecular features and mechanisms of SLE pathogenesis in order to provide new strategies for effective prevention, diagnosis and treatment of SLE.

In recent studies on SLE, microarrays based on high-throughput platforms have been widely used to explore and identify promising biomarkers in order to diagnose and prognosticate the disease at the genomic level [13]. Using bioinformatic analysis and protein blotting, a study found that the expression level of hsa_circ_0000479 could serve as a new diagnostic biomarker for SLE [14]. In a clinical study, Li, H., et al. found that serum levels of PGLYRP2 were greatly elevated in patients with SLE [15]. Both the increase in miR-183-5p and the decrease in miR-374b-3p were validated by qPCR and both showed good
diagnostic performance for SLE. Furthermore, the expression level of miR-183-5p was positively correlated with the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) and the number of anti-dsDNA antibodies [16]. Therefore, it is essential to explore specific SLE biomarkers to enhance the appropriate clinical management of SLE patients and to speeding up the development of new therapeutical agents.

In this study, we explored several significant biomarkers for SLE diagnosis through machine learning. Moreover, in recent years, cancer immunotherapy has gradually come into the limelight [17, 18]. We have subjected the screened genes to pan-cancer analysis to assess their relevance to clinical prognosis and associated signaling pathways by conducting pan-cancer analysis, which has the potential to identify new immunotherapeutic targets[19].

Materials And Methods

Microarray Data Information

The series of matrix files of the GSE20864 and GSE61635 datasets were downloaded from http://www.ncbi.nlm.nih.gov/geo/ and these datasets were based on the GPL570 platform. The GSE20864 dataset included 21 SLE and 45 controls collected from peripheral blood cells, whereas the GSE61635 dataset included 79 SLE and 39 controls collected from the blood. And the GSE37573 dataset included 2 SLE and 2 controls which was based on GPL 210 platforms. Probes for multiple genes were removed and changed to gene symbols based on the files annotated by their probes. The mean of the probes was then calculated as the final expression value of the gene for subsequent studies. And then statistical analyses were performed using the R statistical environment (version 3.6.3).

Data Handling and Analysis of Variance

The limma package of R (http://www.bioconductor.org/limma/) was used for the process of data preprocessing including background correction, normalization, and analysis of differential gene expression between 102 SLE and 86 control samples. After the adjustment the false discovery rate P-value <0.05 and |log FC|>1.0 were defined as the threshold points for DEGs.

Functional Enrichment Analysis

Using R’s “clusterProfiler” and “DOSE” packages, we ran enrichment analysis for disease ontology (DO) terms on DEGs[20]. To reveal biological functions, the "clusterProfiler" R package including Kyoto Gene and Genome Encyclopedia (KEGG) pathway enrichment, Gene Set Enrichment Analysis (GSEA), Disease Ontology (DO) and Gene Ontology (GO) classification was used for annotation analysis to detect the best genetic biomarkers for the diagnosis of SLE. The “c2.cp.kegg.v7.0.symbols.gmt” from the Molecular
Signatures Database (MSigDB) was used as the reference gene set. A gene set was defined as the significant enrichment if a P-value <0.05 and a false discovery rate <0.025 were met.

### Screening for the Best Diagnostic Biomarkers for SLE

In machine learning, the support vector machine (SVM), a supervised machine-learning algorithm, was applied for regression and classification[21]. Gene screening using the least absolute shrinkage and selection operator (LASSO) and support vector machine recursive feature elimination (SVM-RFE) to identify the best diagnostic genetic biomarkers for SLE. Then the random forest package (“rpart” package) in R was used to build the random forest model and the e1071 package in R was used to build the SVM model. And the SVM-RFE method could remove the least important features and select the most important features based on the weights of classifiers[22]. Therefore, support vector machine recursive feature elimination (SVM-RFE) can be applied to select admissible signals to identify the set of genes with the best differentiation potential. Finally, the expression levels of the candidate genes were further analysed in the GSE20864 dataset.

### Diagnostic Value of Feature Biomarkers in SLE

For each gene, receiver operating characteristic (ROC) curve was constructed and the area under the ROC curve (AUC) value of four genes was assessed by numeric integration of the receiver operating characteristic curve, in which the corresponding cutoff points of the ROC curve were ascertained by the Youden's Index. The binary regression model was used to calculate the sensitivity, specificity, and 95% confidence intervals (CIs).

### Discovery of Immune Cell Subtypes

Based on the gene expression profiles in four genes (IFI127, ANKH, FOS, IGF2BP2), to assess the relative proportions of 22 types of infiltrating immune cells in SLE patients. We used a bioinformatics algorithm called CIBERSORT (https://cibersortx.stanford.edu/) to calculate immune cell infiltrations. The immune cell abundance prediction was estimated using a reference set with 22 types of immune cell. To analyze how these four genes function through immune cells, we used the “barplot” package in R. Then the R package "corHeatmap" was used for correlation analysis and visualisation of 22 infiltrating immune cells. Violin plots were additionally drawn using the 'vioplot' package in the R language for the purpose of visualising variations in immune cell infiltration between the four genes and the control samples.

### Analysis of the Association between Genes and Immune Infiltrating Cells
Spearman's rank correlation analysis in the R software explored the association of genetic biomarkers with the level of infiltrating immune cells. This was visualised using the graphical techniques of the 'Lollipop' package.

Data Acquisition

First, we used the Genotype-Tissue Expression (GTEx, https://commonfund.nih.gov/GTEx/) data set. The database is sequenced by collecting data from normal human tissue, so it can be used to study the gene expression between different tissues and between normal and diseased tissues. In addition, the Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov) had characterized more than 20,000 primary cancer samples and corresponding non-carcinoma samples from 33 types of cancers. And Therapeutically Applicable Research to Generate Effective Treatments (TARGET, https://ocg.cancer.gov/programs/target) is a database for pediatric tumors. So we decided that data were downloaded from TCGA database and TARGET database. Then We analysed the differential expression of the IGF2BP2 gene in 34 tumours and adjacent normal tissues by comparing normal tissue data with tumour data. The threshold for differential expression of the genes was $^*p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Prognostic Analysis of IGF2BP2

First, we download the relevant information from the UCSC Xena database, including overall survival (OS) time, progression-free survival (PFS) time, disease-free survival (DFS) time, and disease-specific survival (DSS) time. Then Kaplan–Meier model and univariate Cox regression were used to assess the prognostic role of IGF2BP2 for a specific prognosis type in each cancer. And the cut-off point is selected by the "surv-cutoffpoint" function of the "survminer" R package (0.4.9). Finally, the result of the log-rank P-value of the K-M method and hazard ratio (HR) with a 95% confidence interval (95% CI) were presented in the form of a heatmap.

Gene Set Enrichment Analyses

The “gmt” file of the hallmark gene set (h.all.v7.4.symbols.gmt) containing 50 hallmark gene sets was downloaded from the website of Molecular Signatures Database (MSigDB, https://www.gseamsigdb.org/gsea/index.jsp). And then the Normalized Enrichment Score (NES) and False Discovery Rate (FDR) were then calculated for each cancer type. The Gene Set Enrichment Analysis was conducted using the R packages “clusterProfiler,” and “GSVA” [20]. Lastly the results were summarized in the bubble plot depicted by the R package “ggplot2.”

Immunotherapy Prediction Analysis
The Spearman correlation analysis was performed to show the relationship between IGF2BP2 and tumor mutational burden (TMB) and microsatellite instability (MSI). Numerous studies have explored the significance of using TMB as a biomarker for predicting which patients would be most responsive to checkpoint inhibitors[23]. So we decided that TMB measures the mutation number in a specific cancer genome. The TMB and MSI scores were obtained from TCGA. The horizontal axis of the graph represents the correlation coefficient between IGF2BP2 and TMB or MSI, and the vertical coordinates represent the different types of cancer. The size of the dots in the graph represents the magnitude of the correlation coefficient, and the different colors represent the importance of the p-value. SPSS version 19.0 (IBM Corporation) was used for all of the data and all graphs were made by GraphPad Prism 8 software (GraphPad Software, CA, USA) and R language. All statistical analyses were carried out using R (version 3.6.3). LASSO regression algorithm using the 'glmnet' package of R to identify genes significantly associated with differentiation in SLE and controls. Receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) can be used to determine the diagnostic effectiveness of the diagnostic biomarkers. Then we analysed the relationship between gene biomarker expression and infiltrating immune cells using Spearman's correlation.

**Statistical Analyses**

SPSS version 19.0 (IBM Corporation) was used for all of the data and all graphs were made by GraphPad Prism 8 software (GraphPad Software, CA, USA) and R language. All statistical analyses were carried out using R (version 3.6.3). LASSO regression algorithm using the 'glmnet' package of R to identify genes significantly associated with differentiation in SLE and controls. Receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) can be used to determine the diagnostic effectiveness of the diagnostic biomarkers. Then we analysed the relationship between gene biomarker expression and infiltrating immune cells using Spearman's correlation.

In addition to the analysis of pan-cancer we have done the following studies. First, we selected clinical indicators, including OS, disease-specific survival (DSS), progression-free interval (PFI) and disease-free interval (DFI). Then the Wilcoxon rank-sum test was used to compare the IGF2BP2 expression levels between tumor and normal tissues. Paired t-tests can assess IGF2BP2-associated cancer characteristics. Univariate Cox regression analysis and the Kaplan–Meier method (log-rank test) were employed to assess the prognostic role of IGF2BP2 expression in each cancer. Finally, Spearman correlation analysis was performed to evaluate the statistical relationships between IGF2BP2 and other factors, such as TMB, MSI and other cancer tissues.

**Results**

**Identification of DEGs in SLE**
After preprocessing the downloaded data, we used the R package "limma" to analyse the data for differentially expressed genes. Finally, a total of 52 DEGs were identified in the SLE samples, with heatmaps showing the expression of these genes in the control and SLE groups (Figure 1A). The volcano plot showed that 29 of these genes were significantly upregulated and 23 were significantly downregulated (Figure 1B).

**Functional Enrichment Analysis**

GO enrichment analysis manifested that Using the criteria of P-value < 0.05, GO enrichment results revealed that 10 BP terms, 10 CC terms, and 10 MF term were enriched in our study. These 52 DEGs were mainly enriched in biological process (BP) of viral life cycle, response to the virus, type I interferon signaling pathway, cellular response to type I interferon and viral life cycle; in the same way what were enriched in cellular component (CC) were cell cortex, haptoglobin-hemoglobin complex, hemoglobin complex, and the cortical cytoskeleton; the most significant molecular function (MF) included transmembrane transporter, haptoglobin binding, oxygen carrier activity, and oxygen binding(Figure 2A). The results of the KEGG enrichment analysis indicated significant enrichment in the SLE group mainly involved malaria, Coronavirus disease-COVID-19, and nicotine addiction pathways (Figure 2B).

The DO enrichment analysis was performed to determine the enrichment of these DEGs in terms of diseases. The results showed that the DEGs were enriched for diseases mainly associated with brain diseases, epileptic syndromes, malaria, parasitic protozoa, infectious diseases and parasitic infectious diseases (Figure 2C).

The GSEA analysis showed that treat group functional enrichment pathways were mainly reflected in the NOD-like receptor signaling pathway, RIG-like receptor signaling pathway, and Toll-like receptor signaling pathway (Figure 2D). Control group mainly involved hematopoietic cell lineage pathway and ribosome pathway (Figure 2E).

**Identification and Validation of Diagnostic Gene Biomarkers**

In order to select and validate the most significantly characteristic DEGs of SLE, we take two different algorithms. By using the LASSO regression algorithm, we obtained 12 DEGs between SLE and the control groups (Figure 3A). A subset of the SVM-RFE algorithm was used to determine the 13 DEGs (Figure 3B). At least four features (FOS, IFI27, ANKH, and IGF2BP2) were selected simultaneously by the two algorithms and finally identified as the diagnostic gene biomarkers for SLE (Figure 3C). The difference in expression of the four-character genes in the control and treat groups is shown in the box plots (Figure 3D-G). The results showed that the expression of IGF2BP2 was significantly higher in the con group than in the treat group, and the expression of three characteristic genes, ANKH, IFI27 and FOS, were all higher in the treat group than in the con group.
Validating the Expression and Diagnostic Value of four genes in an Independent Cohort

In order to predict the diagnostic value of the four diagnostic genes, we analyzed the receiver operating characteristic (ROC) curves for these four genes. The results indicated that the three signature genes (ANGH, IFI27, IGF2BP2) showed excellent diagnostic value in SLE and control groups. First, we found a significant difference in the expression of ANKH between Con and Treat patients. The corresponding AUC for the treat group was 0.851 (95%CI:0.782-0.915), and the corresponding AUC for the control group was 0.568 (95%CI:0.417-0.712) (Figure 4A & Figure 4E). Then, we observed the expression of FOS between Con patients and Treat, which was no significant difference in FOS expression. The corresponding AUC for the treat group was 0.782 (95%CI:0.712-0.845), and the corresponding AUC for the control group was 0.658 (95%CI:0.516-0.777) (Figure 4B & Figure 4F). IFI27 expression was significantly different between Con and Treat patients. The corresponding AUC for the treat group was 0.728 (95%CI:0.648-0.799), and the corresponding AUC for the control group was 0.964 (95%CI:0.898-1.000) (Figure 4C & Figure 4G). IGF2BP2 expression was down-regulated in the treat group. The corresponding AUC for the treat group was 0.890 (95%CI:0.835-0.943), and the corresponding AUC for the control group was 0.818 (95%CI:0.686-0.930) (Figure 4D & Figure 4H).

Immune Cell Infiltration

The ratio of 22 immune cells in SLE patients to healthy people is shown directly in the bar chart (Figure 5A). In addition, we use violin plots to test whether genetic differences are significant. The proportions of neutrophils (P<0.001), monocytes (P=0.010), macrophages M1 (P=0.001), macrophages M0 (P=0.003), plasma cells (P=0.011) and dendritic cells activated (P<0.001) were significantly lower in the control group than in the treat group. However, the proportion of CD8 T cells (P=0.002), NK cells resting (P<0.001), CD4 memory activated T cells (P<0.001), CD4 memory resting T cells (P<0.001), Macrophages M2 (P<0.001) were significantly higher in the control group than in the treat group (Figure 5C). The correlation between the 22 immune cells is shown in the diagram and we visualised their relationship using a heat map (Figure 5B).

Correlation Analysis Between the Four Genes and Infiltrating Immune Cells

We also conducted the immune-correlation analyses between the expression of individual genes and immune cells. IGF2BP2 was negatively correlated with Dendrite cells activated (r= R=-0.3, P=4.8e-05), Neutrophils (r=-0.32, P=1.1e-05), Plasma cells (r=-0.21, P=0.0052), and positively correlated with Dendrite cells resting (r=0.19, P=0.013), Macrophages M2 (r=0.2, P=0.007), NK cells resting (r=0.16, P=0.037), T cells CD4 naive (r=0.23, P=0.0022) (Figure 6A-G). IFI127 was negatively correlated with T cells CD8 (r=-0.24, P=0.0011), B cells memory (r=-0.18, P=0.017), and positively correlated with T cells follicular
helper \( (r=0.16, P=0.036) \), Macrophages M1 \( (r=0.48, P=1.3e-11) \), Monocytes \( (r=0.22, P=0.0035) \), NK cells activated \( (r=0.19, P=0.011) \) (Figure 6H-L). FOS was positively correlated with Macrophages M0 \( (r=0.27, P=3e-04) \), and was negatively correlated with T cells regulatory (Tregs) \( (r=-0.17, P=0.02) \), Mast cells resting \( (r=-0.19, P=0.0094) \), T cells CD4 memory activated \( (r=-0.35, P=2.5e-06) \) (Figure 6M-P). ANKH was positively correlated with T cells gamma delta \( (r=0.16, P=0.011) \) (Figure 6T). IGF2BP2 was positively correlated with CD4 naive T cells \( (P=0.002) \), Macrophages M2 \( (P=0.007) \), resting dendritic cells \( (P=0.013) \), resting NK cells\( (P=0.037) \), and negatively correlated with Plasma cells \( (P=0.005) \), activated dendritic cells \( (P<0.001) \) neutrophils \( (P<0.001) \).

**Analysis of IGF2BP2 expression in tumors and prognostic potential**

To reveal differences in IGF2BP2 expression in different tumours, we compared the data from three databases. Data for tumours were retrieved from the Target and TCGA databases, and data for healthy human were retrieved from the GTEx database. (Figure 7A). Normal tissue data \( (n = 10456) \), tumour data \( (n = 11537) \). The results showed that IGF2BP2 gene was lowly expressed in 9 tumors (GBMLGG, LGG, BRCA, LUAD, KIPAN, PRAD, KIRC, PCPG, ACC) compared to normal tissues. In 22 tumors, GBM, CESC, ESCA, STES, KIRP, COAD, COADREAD, STAD, HNSC, LUSC, LIHC, WT, SKCM, THCA, READ, OV, PAAD, TGCT, UCS, ALL, LAML, CHOL, the IGF2BP2 gene showed high expression. This suggests that IGF2BP2 is not equally expressed in cancers.

To investigate the relationship between IGF2BP2 and prognostic potential in tumours, the Kaplan-Meier method (log-rank test) and univariate Cox regression were used to analyse four prognostic indicators (OS, DSS, DFI, PFI) in 37 cancers. The heat map shows the prognostic potential of IGF2BP2 expression in 37 tumours. The results revealed that the prognosis of most cancers presented a high correlation with IGF2BP2, except for BRCA, CHOL, GBM, KICH, OV, READ, SKCM, STAD, STES and TGCT (Figure 7B). Specifically, IGF2BP2 was a risk factor for poor prognosis in ACC, BLCA, CESC, COAD, COADREAD, DLBC, GBMLGG, HNSC, KIPAN, KIRC, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, SARC, UCEC risk factors. In addition, IGF2BP2 was a good prognostic factor in ESCA, THCA, THYM, UCS and UVM. IGF2BP2 was a risk factor for all four prognostic survival indicators in BLCA, GBMLGG, LGG, LUAD and MESO, and was significantly associated with poor prognosis. We determined the relationship between IGF2BP2 gene expression and patient prognosis in 39 tumours \( (p<0.05) \) by drawing forest plots based on univariate cox regression analysis. The results of the forest plot (Figure 7C) showed that downregulation of IGF2BP2 expression was associated with a delayed time to OS: GBMLGG\( (P = 6.8e-55, HR = 1.70[95\% CI,1.58–1.82])\), LGG\( (P = 9.4e-33, HR = 1.87[95\% CI,1.66–2.10])\), KIRC\( (P = 9.2e-9, HR = 1.40[95\% CI,1.24–1.57])\), PAAD\( (P = 4.4e-5, HR = 1.54[95\% CI,1.25–1.90])\), MESO\( (P = 1.9e-4, HR = 1.36[95\% CI,1.15–1.61])\), HNSC\( (P = 4.4e-4, HR = 1.21[95\% CI,1.09–1.35])\), ACC\( (P = 6.9e-4, HR = 1.47[95\% CI,1.16–1.85])\), LIHC\( (P = 2.0e-3, HR = 1.21[95\% CI,1.07–1.37])\), LUAD\( (P = 5.2e-3, HR = 1.18[95\% CI,1.05–1.32])\), BLCA\( (P = 9.8e-3, HR = 1.30[95\% CI,1.15–1.46])\).
Many studies have shown that IGF2BP2 is strongly associated with the prognosis of LUAD, LGG and PAAD. Therefore, we analyzed Kaplan-Meier curves for LUAD, LGG and PAAD and found that higher IGF2BP2 expression was associated with better survival outcomes (Figure 7H, Figure 7I, Figure 7K), suggesting that IGF2BP2 is a prognostic for OS in LUAD, LGG and PAAD biomarkers of OS in LUAD, LGG and PAAD. Finally, we found that high expression of IGF2BP2 in BLCA, ACC, GBMLGG, KIRC and MESO had a more favourable prognosis (Figure 7D, Figure 7E, Figure 7F, Figure 7G, Figure 7J). IGF2BP2 has a prognostic role in predicting cancer prognosis.

Set enrichment analysis of IGF2BP2 reveals its relevance in cancer immune processes

GSEA analysis of IGF2BP2 reveals its importance in tumour immune processes. We analysed the expression of IGF2BP2 in different tumour types, selecting the top 30% and bottom 30% expression difference groups, and based on these differences we performed GSEA enrichment analysis for 33 tumours (Figure 8A). We identified several pathways involved in immunity in 33 cancers. TNFA-signaling-via-NFKB, IFN-α response, IFN-γ response, allograft rejection pathways and inflammatory response were significantly enriched in various cancers, especially in ACC, BLCA, BRCA, DLBC, LGG, PCPG, and PRAD. These results suggest that IGF2BP2 may be closely involved in the binding of ligand receptors to cancer cells and immune cells. Furthermore, we found that differential expression of IGF2BP2 was associated with transformation of the epithelial-mesenchymal pathway in different cancers, with significant positive correlations for BLCA, BRCA, CHOL, DLBC, GBM, HNSC, LGG, LUAD, MESO, PCPG, PRAD, SARC, TGCT and THCA, suggesting that IGF2BP2 may have a role in tumour invasion of normal tissue. In addition, MYC targets, MITOTIC_SPINOLE, E2F targets, G2M_CHECKPOIMT were also closely associated with IGF2BP2 expression in cancer. These results indicate that IGF2BP2 may be involved in multiple pathways of cancer immunity.

The role of IGF2BP2 in the response to ICI treatment

To assess the impact of IGF2BP2 expression on the efficacy of immunotherapy. We calculated the correlation between IGF2BP2 expression and two biomarkers of immune checkpoint inhibitor (ICI) efficacy prediction, TMB and MIS. The results showed that IGF2BP2 expression was positively correlated with TMB scores in BRCA, GBMLGG, HNSC, LGG, LUAD, PAAD, STAD and THYM, and negatively correlated with TMB scores in UCEC and KIRC (Figure 8B). In addition, IGF2BP2 expression was found to be positively correlated with MSI in BLCA, BRCA, KIPAN, LUSC, STAD, STES and TGCT; and negatively in DLBC and GBMLGG. (Figure 8C). Our findings suggest that IGF2BP2 may be useful in predicting the efficacy of ICI.
Correlation between IGF2BP2 gene expression and cancer cell stemness

To further analyze the relationship between IGF2BP2 gene expression and cancer cell stemness, we calculated DNAss and RNAss based on DNA methylation and gene expression in pan-cancer, and performed Spearman correlation analysis. The results showed that DNAss based on DNA methylation was shown to be strongly associated with IGF2BP2 expression in most cancer types except DLBC, UVM, CESC, PCPG, KIRP, LAML, KIRC, CHOL and GBM. The correlation was particularly obvious in ACC, YHYM, LGG and GBMLGG (Figure 8D). In STES, STAD and KIPAN, A substantial relationship between CNV (copy number variations) based RNAss and IGF2BP2 expression was proved. In contrast, this correlation was not evident in MESO, CESC, HNSC, UCEC, KIRC, PAAD, ACC and KICH (Figure 8E).

Discussion

SLE is a common autoimmune disease, often associated with vital organ involvement. Although symptoms and signs are the basis for the diagnostic and disease assessment of SLE, due to the complexity and diversity of its clinical manifestations, early definitive diagnosis, disease assessment and treatment monitoring still require reference to laboratory indicators\[13, 24\]. Therefore, the search for genetic markers is important for the diagnosis of SLE.

In the present study, by functional enrichment analysis, 52 DEGs were strongly correlated with the viral life cycle, response to the virus, type I interferon signaling pathway, malaria, and brain disease in SLE. It has been shown that infection by various pathogens, such as cytomegalovirus, EBV, and human microvirus B19, may be involved in the development and progression of systemic lupus erythematosus [25–27]. As early as 2003, significant expression of interferon-regulated genes in SLE, IFN acts in SLE in multiple ways, including expressing ISG through the IFN signaling pathway, prompting other cells to function, and forming interactions with other cytokines [28–30]. Important observations suggest that the African genome includes genes selected for the presence of malaria that protect from systemic lupus erythematosus in the individuals chronically exposed to malaria, but are dangerous factors for SLE in the absence of malaria [31]. The presence of central nervous system damage is a clinical sign of critical illness and one of the main causes of death in SLE[32–34]. Those further justify the results of our analysis.

We used two different algorithms, LASSO and SVM-RFE, for simultaneous selection and finally obtained four diagnostic biomarkers, IGF2BP2, ANGH, IFI27 and FOS. Among the four diagnostic biomarkers screened, IGF2BP2 was significantly downregulated in gene expression in SLE patients. Through data analysis, we validated that these four genes have better expression and diagnostic genes. Insulin-like growth factor 2 (IGF2) mRNA-binding protein 2 (IGF2BP2) is an RNA-binding protein that is a retrotranscriptional regulator of mRNA targeting, stability and translation control. IGF2BP2 is located in the cytoplasm and can also enter the nucleus, where it has a processing effect on RNA. IGF2BP2 not only binds and promotes IGF2 mRNA translation to increase IGF2 protein expression, but also binds and
stabilizes other genes through specific functional domains to affect ATP production and regulate cellular Energy metabolism. Dysregulation of IGF2BP2 is often associated with diseases such as insulin resistance, diabetes or cancer in humans. Some experiments with IGF2BP2 knockout mice have shown that IGF2BP2 is a tumor promoter, driving cancer development. IGF2BP2 may affect SLE by participating in multiple pathways that influence cell proliferation and differentiation, and the exact mechanisms need to be further investigated [35–37].

Mutations in the progressive ankylosis protein (ANKH) cause craniometaphyseal dysplasia (CMD) [38, 39]. It has been shown that the normal ANKH protein has a strong anti-calcification effect and that mice with this gene defect show progressive postnatal ankylosis in the joints of the limbs and spine, narrowing of the joint space, ossification of the spinal ligaments and intervertebral discs, and changes similar to those seen in human ankylosing spondylitis[40]. Therefore, ANKH may play a role in SLE by being involved in the inflammatory pathway.

It has been shown in many studies that IFI27 expression is increased in SLE patients and that it plays an important role in the development of SLE [41, 42]. This is consistent with our findings. The Fos gene family comprises four members: FOS, FOSB, FOSL1, and FOSL2. In some situations, the expression of FOS genes is correlated with apoptosis. Sustained expression of FOS proteins is a hallmark of terminal cell differentiation and a precursor to the onset of cell death, as well as suppressing the expressed of a numbers of genes that maintain cell survival. Therefore, we hypothesize that the FOS gene may affect SLE by regulating apoptotic cell death [43–45].

Through pan-cancer analysis, we found that IGF2BP2 had a high mutation rate and showed significantly high expression in more than half of the cancers and that the gene was sufficiently correlated with immune system-related pathways and most immune cells, suggesting that IGF2BP2 plays a sufficient role in the immune microenvironment of most tumors. In addition, we also calculated the expression of IGF2BP2 in 33 cancers in relation to TMB and MSI [46]. It has been shown that the gene encoding IGF2BP2 is expanded and is overexpressed in many forms of human cancer [47]. Over the past few years, microsatellite instability (MSI) has become a primary predictor of the effectiveness of immune checksite blocks and tumor mutation burden(TMB) is emerging as a potential biomarker[48]. Interestingly, IGF2BP2 was found to be remarkably correlated with TMB and MSI in most cancers, providing an important rationale for IGF2BP2 as a new future tumour immunotherapy target [49].

Conclusions

This research shows that the SLE-associated gene IGF2BP2 is of outstanding value in the diagnosis of SLE at an early stage. We subjected the screened genes to pan-cancer analysis and the results showed that IGF2BP2 plays an adequate part in the immune microenvironment of most tumours. However, future diagnostic studies of SLE need to be further explored.

Declarations
Authors’ Contributions

JFZ and RYC were responsible for the entire project. XYZ, QL and ZBX conduct study selection, data abstraction, and statistical analysis and interpretation of data. JFZ, ZBX, RYC, QL, YXZ, KBH was responsible for drafting and revising the final review paper.

Declaration of Conflicting Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of Data and Materials

The data that support the findings of this study are available in Gene Expression Omnibus, and they were downloaded from http://www.ncbi.nlm.nih.gov/geo/.

References


**Figures**

**Figure 1**

Outline of genes differentially expression in SLE versus controls.

(A) Heat map of 52 SLE-related DEGs.

(B) Volcano map of genes that are differentially expressed among SLE patients and controls.
Figure 2

Analysis of functional enrichment of SLE-associated DEGs.

(A) Enrichment of GO (Gene Ontology) biological process pathway analysis.

(B) Enrichment of KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis.
Identification and validation of diagnostic signature biomarkers. Two algorithms are applied for feature selection. the LASSO (A) on the one hand and the SVM-RFE (B) algorithm on the other.

(A) LASSO regression for feature gene screening.

(B) SVM-RFE screening of feature genes.

(C) Four features (FOS, IFI27, ANKH, and IGF2BP2) selected by the two algorithms.
(D-G) Differential expression of the four features among the treatment and control groups.

Figure 4

Validating the Expression and Diagnostic Value of four genes in an Independent Cohort

(A, E) The corresponding AUC for the treat group of ANKH, and the corresponding AUC for the control group of ANKH.

(B, F) The corresponding AUC for the treat group of FOS, and the corresponding AUC for the control group of FOS.

(C, G) The corresponding AUC for the treat group of IFI27, and the corresponding AUC for the control group of IFI27.

(D, H) The corresponding AUC for the treat group of IGF2BP2, and the corresponding AUC for the control group of IGF2BP2.
Figure 5

Distribution and visualization regarding immune cell infiltration.

(A) Proportion of 22 immune cells in SLE patients and normal subjects.

(B) A correlation matrix composed of all 22 immune cell subtypes.

(C) The 22 immune cell subtypes compared between SLE patients and normal subjects.
Figure 6

Analysis of correlations among four genes and infiltrating immune cells.

(A-S) The immune-correlation analyses between the expression of individual genes. Correlation between IGF2BP2 (A-G), IFI27 (H-M), FOS (N-Q), ANKH (R, S) and infiltrating immune cells in SLE patients.

(T) Correlation between IGF2BP2.
Figure 7

Analysis of IGF2BP2 expression in tumors and prognostic potential.

(A) IGF2BP2 expression level in tumour and normal groups.

(B) The prognosis role of in pan-cancer.

(C) The forest plot of OS association in 39 tumour types.

(D–K) Kaplan-Meier analysis of the relationship among TREM2 expression and OS.
Figure 8

Enrichment analysis of IGF2BP2 revealing its relevance in cancer immune processes.

(A) IGF2BP2-related cancer characteristics of GSEA analysis among 33 cancer types.

(B) Radar map of among IGF2BP2 expression and TMB.
(C) Radar map of among IGF2BP2 expression and MIS.

(D, E) Spearman correlation analysis of DNAss(D) and RNAss(E) based on DNA methylation and gene expression in pan-cancer.