ORIGINAL RESEARCH

Integrated Approach for Biomarker Discovery and Mechanistic Insights into the Co-Pathogenesis of Type 2 Diabetes Mellitus and Non-Hodgkin Lymphoma

Yidong Zhu [1],*, Jun Liu[1],*, Bo Wang[2]

[1]Department of Traditional Chinese Medicine, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, 200072, People's Republic of China; [2]Department of Endocrinology, Yangpu Hospital, Tongji University School of Medicine, Shanghai, 200090, People's Republic of China

*These authors contributed equally to this work

Correspondence: Bo Wang, Email bowang13241@163.com

Background: Type 2 diabetes mellitus (T2DM) is associated with an increased risk of non-Hodgkin lymphoma (NHL), but the underlying mechanisms remain unclear. This study aimed to identify potential biomarkers and elucidate the molecular mechanisms underlying the co-pathogenesis of T2DM and NHL.
Methods: Microarray datasets of T2DM and NHL were downloaded from the Gene Expression Omnibus database. Subsequently, a protein-protein interaction network was constructed based on the common differentially expressed genes (DEGs) between T2DM and NHL to explore regulatory interactions. Functional analyses were performed to explore underlying mechanisms. Topological analysis and machine learning algorithms were applied to refine hub gene selection. Finally, quantitative real-time polymerase chain reaction was performed to validate hub genes in clinical samples.
Results: Intersection analysis of DEGs from the T2DM and NHL datasets identified 81 shared genes. Functional analyses suggested that immune-related pathways played a significant role in the co-pathogenesis of T2DM and NHL. Topological analysis and machine learning identified three hub genes: *GZMM*, *HSPG2*, and *SERPING1*. Correlation analysis revealed significant correlations between these hub genes and immune cells, underscoring the importance of immune dysregulation in shared pathogenesis. The expression of these genes was successfully validated in clinical samples.
Conclusion: This study suggested the pivotal role of immune dysregulation in the co-pathogenesis of T2DM and NHL and identified and validated three hub genes as key contributors. These findings provide insight into the complex interplay between T2DM and NHL.
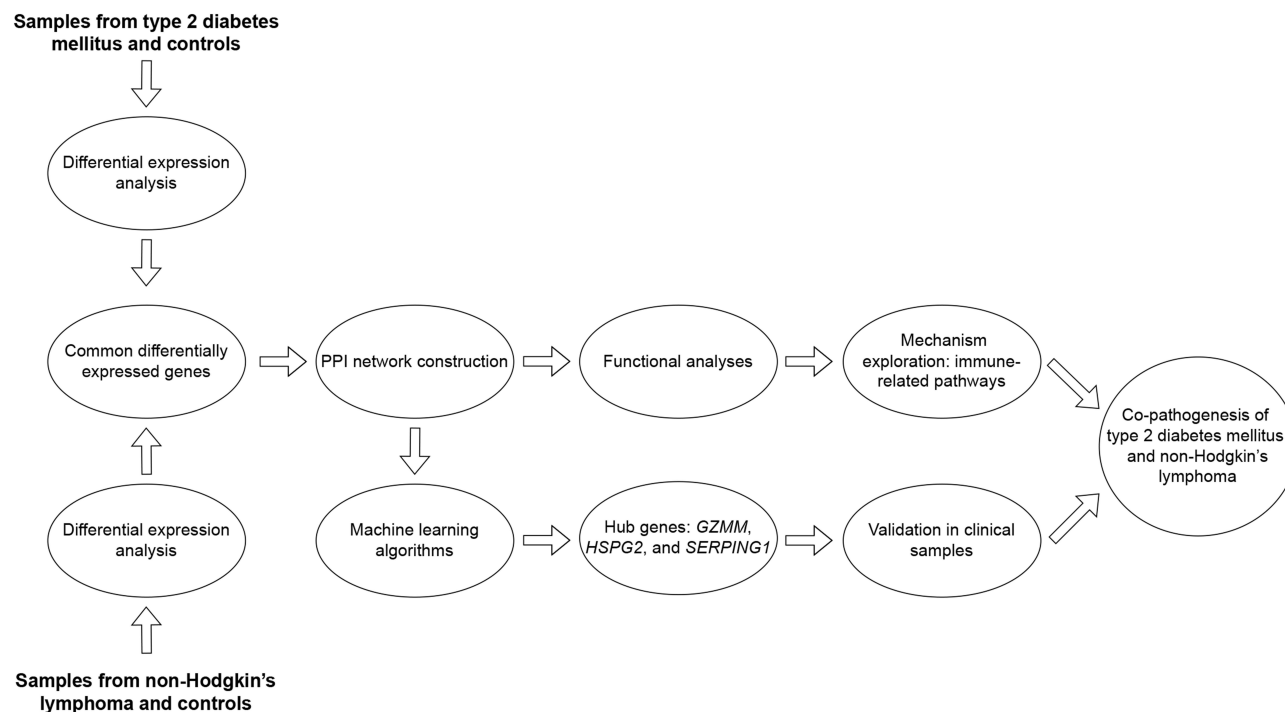**Keywords:** type 2 diabetes mellitus, non-Hodgkin lymphoma, immunity, microarray analysis, machine learning

Introduction

T2DM is a chronic disorder characterized by insulin resistance and progressive pancreatic β-cell dysfunction, leading to metabolic disturbances and hyperglycemia.[1] It is the most prevalent form of diabetes, accounting for 90–95% of all diabetes cases.[2] Over the past few decades, the incidence of T2DM has risen at an alarming rate, posing a significant global public health challenge.[3] Environmental factors including obesity, sedentary lifestyles, and unhealthy diets, as well as genetic predispositions, contribute to the various pathophysiological disruptions responsible for glucose homeostasis dysfunction in T2DM.[4] This condition is associated with numerous complications, including cardiovascular disease, nephropathy, neuropathy, and retinopathy, which contribute significantly to global morbidity and mortality.[5]

Emerging evidence highlights a connection between T2DM and an increased risk of various cancers such as liver,[6] gallbladder,[7] prostate,[8] gastric,[9] lung,[10] and oral cancers.[11] Moreover, among hematological malignancies, previous research suggests that T2DM is linked to a heightened risk of NHL, leukemia, and myeloma.[12–14] It is worth mentioning that a growing

267

Graphical Abstract

Samples from type 2 diabetes mellitus and controls

Differential expression analysis

Common differentially expressed genes → PPI network construction → Functional analyses → Mechanism exploration: immune-related pathways

Differential expression analysis

Machine learning algorithms → Hub genes: *GZMM*, *HSPG2*, and *SERPING1* → Validation in clinical samples

Co-pathogenesis of type 2 diabetes mellitus and non-Hodgkin's lymphoma

Samples from non-Hodgkin's lymphoma and controls

body of research has identified a potential association between T2DM and NHL, a heterogeneous group of malignant lymphoid tumors.[15–18] NHL is the most prevalent hematological malignancy worldwide, accounting for approximately 3% of all cancer diagnoses and deaths, and its incidence has steadily increased in recent decades.[19] The development of NHL is influenced by complex factors, including genetic susceptibility, immune dysfunction, and viral infections, such as those caused by the Epstein-Barr virus and human immunodeficiency virus.[19–21] However, the precise mechanisms underlying the association between T2DM and NHL remain poorly understood. Further research is crucial to elucidate the underlying mechanisms and identify potential biomarkers that could enable targeted therapeutic strategies for coexisting conditions.

In recent years, advancements in high-throughput technologies, such as microarray analysis, have enabled researchers to profile gene expression patterns comprehensively, thereby offering valuable insights into the molecular mechanisms underlying complex diseases.[22,23] Coupled with the power of machine learning algorithms, these tools allow the identification of critical biomarkers and key genes associated with disease progression, as well as the discovery of potential therapeutic targets.[24,25] Machine learning facilitates the integration and analysis of large, multidimensional datasets and provides a robust framework for distinguishing meaningful biological signals from noise.[26–28] In this study, we aimed to identify the hub genes and elucidate the mechanisms underlying the co-pathogenesis of T2DM and NHL by integrating microarray analysis, machine learning, and experimental validation.

Materials and Methods
Data Collection
This study was conducted in accordance with the Declaration of Helsinki and received approval from the Ethics Committee of Shanghai Yangpu Hospital (LL-012). Microarray datasets were obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). The GSE25724 dataset, which included samples from T2DM patients and non-diabetic controls, was used for T2DM-related analysis. The GSE25638 dataset, which included samples from patients with various types of NHL and normal controls, was used for NHL analysis. Detailed information on these datasets is

provided in [Supplementary Table S1](#). To ensure consistency and reliability across datasets, raw data were normalized to address potential batch effects. Peripheral blood samples were obtained from three groups at Shanghai Yangpu Hospital: five patients with T2DM, five patients diagnosed with NHL, and five healthy donors. T2DM samples were selected based on the American Diabetes Association diagnostic criteria. Inclusion criteria required patients to meet at least one of the following: fasting plasma glucose ≥126 mg/dL (7.0 mmol/L), 2-hour plasma glucose ≥200 mg/dL (11.1 mmol/L) during an OGTT, or HbA1c ≥6.5%. NHL samples were included after diagnostic confirmation via morphological examination, immunohistochemistry, flow cytometry, and molecular genetic testing. Healthy donors were required to be free from significant illnesses (eg, chronic diseases, autoimmune disorders, infections, or malignancies). The general exclusion criteria included patients with significant comorbidities (eg, autoimmune diseases, active infections, or other malignancies), individuals with other types of diabetes (eg, type 1, gestational, or secondary diabetes), patients taking medications affecting glucose metabolism or immune function (eg, corticosteroids), individuals under 18 or over 75 years of age, pregnant or lactating women, and cases with poor-quality blood samples (eg, hemolysis or clotting).

Identification of DEGs

Differential expression analysis was conducted on the above-mentioned datasets to identify DEGs using the "LIMMA" (version 3.56.2) package. DEGs were selected based on the criteria of adjusted p-value < 0.05 and |log fold change| > 0.585. The "VennDiagram" (version 1.7.3) package was used to identify genes common to both T2DM and NHL.

Construction of a PPI Network

To analyze the regulatory interactions among the identified DEGs, a PPI network was constructed using the STRING database (http://string-db.org). The resulting network was then imported into Cytoscape software (version 3.8.2) for further analysis. Within Cytoscape, the MCODE algorithm was used to identify key functional modules. The selection criteria for the MCODE analysis were set as follows: degree cutoff = 2, node score cutoff = 0.2, k-core value = 2, and maximum depth = 100. A topological analysis was performed using the CytoHubba plugin to rank the most significant genes in the PPI network. Based on a previous study,[29] we identified ten genes with the highest degree values as key genes. To further investigate the relationships between these key genes, a co-expression network was constructed using GeneMANIA (http://www.genemania.org/), a tool that provides insights into the functional associations between gene sets.

GO and KEGG Analyses

GO and KEGG analyses were conducted to explore the underlying mechanisms. The enrichment analysis helps to identify the potential mechanisms through which these genes contribute to disease development, offering theoretical support for further investigation. A significance threshold of $p < 0.05$ was applied in both analyses. The results were visualized using the "clusterProfiler" (version 4.8.1) and "enrichplot" (version 1.20.0) packages.

Selection of Hub Genes Using Machine Learning Algorithms

Advanced machine learning techniques were used to identify hub genes from the key genes in the PPI network, ensuring a rigorous and reliable selection process. The LASSO regression, recognized for variable selection and regularization to prevent overfitting, uses one standard error criterion to balance model complexity and performance.[30] The SVM algorithm is an effective supervised learning method known for its ability to handle high-dimensional data by maximizing the separation between classes.[31] To enhance predictive accuracy, we applied SVM-RFE, which iteratively refines the feature set by removing less important features. In addition, the RF algorithm, an ensemble learning method, effectively manages unbalanced data and estimates feature importance.[32] The final set of hub genes was derived from an intersection analysis of the results of the LASSO logistic regression, SVM-RFE, and RF methods. The implementation and visualization of these machine learning algorithms were carried out using the "glmnet" (version 4.1.7), "e1071" (version 1.7.13), and "randomForest" (version 4.7.1.1) packages, respectively.

GSEA and Correlation Analysis

GSEA was performed to identify the pathways significantly associated with the hub genes. Moreover, ssGSEA was performed to quantify the enrichment levels of 28 immune cells in the T2DM and control groups, as well as in the NHL and control groups, offering a comprehensive overview of immune cell involvement. Spearman correlation analysis was used to investigate the relationship between the hub genes and infiltrating immune cells. The analysis and visualization were conducted using several packages, including "vioplot" (version 0.4.0), "reshape2" (version 1.4.4), "ggplot2" (version 3.4.2), and "org.Hs.eg.db" (version 3.17.0).

qRT-PCR

Total RNA was extracted from the samples using TRIzol reagent (Takara, Japan), according to the manufacturer's instructions. The isolated RNA was reverse-transcribed into cDNA using the RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific, USA). Subsequently, qRT-PCR was conducted in 96-well plates using the SYBR Green PCR Master Mix (KAPA, Japan) and the Applied Biosystems 7900HT Fast Real-Time PCR System (Thermo Fisher Scientific, USA). Gene expression levels were calculated using the $2^{-\Delta\Delta Ct}$ algorithm, with GAPDH as the internal control. Primer sequences used for qRT-PCR are listed in Supplementary Table S2.

Statistical Analysis

Data analysis and visualization were performed using R software (version 4.3.0) and GraphPad Prism (version 8.0.1). For group comparisons involving normally distributed quantitative variables, the Student's t-test was applied, while the Wilcoxon test was used for non-normally distributed data. Statistical significance was determined using the following thresholds: $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$.

Results

Identification of DEGs

The flowchart in Figure 1 illustrates the analyses performed in this study. In the GSE25724 dataset, we identified 1807 DEGs between patients with T2DM and controls, including 754 upregulated and 1053 downregulated genes (Figure 2A and B). Similarly, we identified 2487 DEGs between patients with NHL and controls in the GSE25638 dataset, consisting of 1494 upregulated and 993 downregulated genes (Figure 2C and D). Intersection analysis of DEGs from the T2DM and NHL datasets identified 81 common genes, of which 52 were upregulated and 29 were downregulated (Figure 2E and F). GO analysis revealed that these common DEGs were significantly involved in immune-related biological processes, including response to type II interferon, lymphocyte differentiation, complement activation, mononuclear cell differentiation, and immune response activation (Figure 2G). KEGG analysis further revealed significant enrichment in immune-related pathways, including primary immunodeficiency, complement and coagulation cascades, pertussis, hematopoietic cell lineage, and *Staphylococcus aureus* infection (Figure 2H).

Construction of the PPI Network

The PPI network was constructed to analyze the shared DEGs between T2DM and NHL (Figure 3A). Using the MCODE plug-in in Cytoscape, we identified two interconnected gene modules comprising 14 common DEGs (Figure 3B and C). GO analysis indicated that these genes were significantly enriched in immune-related biological processes, including complement activation, immune response activation, and type I interferon-mediated signaling pathway (Figure 3D). KEGG analysis further showed significant involvement of immune-related pathways, such as complement and coagulation cascades, *Staphylococcus aureus* infection, and primary immunodeficiency (Figure 3E).

Subsequently, we identified the top ten genes with the highest degree values as key genes using topological analysis: *ZAP70, IL7R, GZMM, CD8A, GBP2, HSPG2, LTBP2, IFITM1, SERPING1*, and *SAMHD1* (Figure 4A). To gain deeper functional insights, we used the GeneMANIA database to construct a co-expression network and analyze the interactions among these genes (Figure 4B). Functional analysis showed that these genes were significantly involved in immune-related processes, including lymphocyte differentiation, T cell differentiation, positive regulation of leukocyte activation,
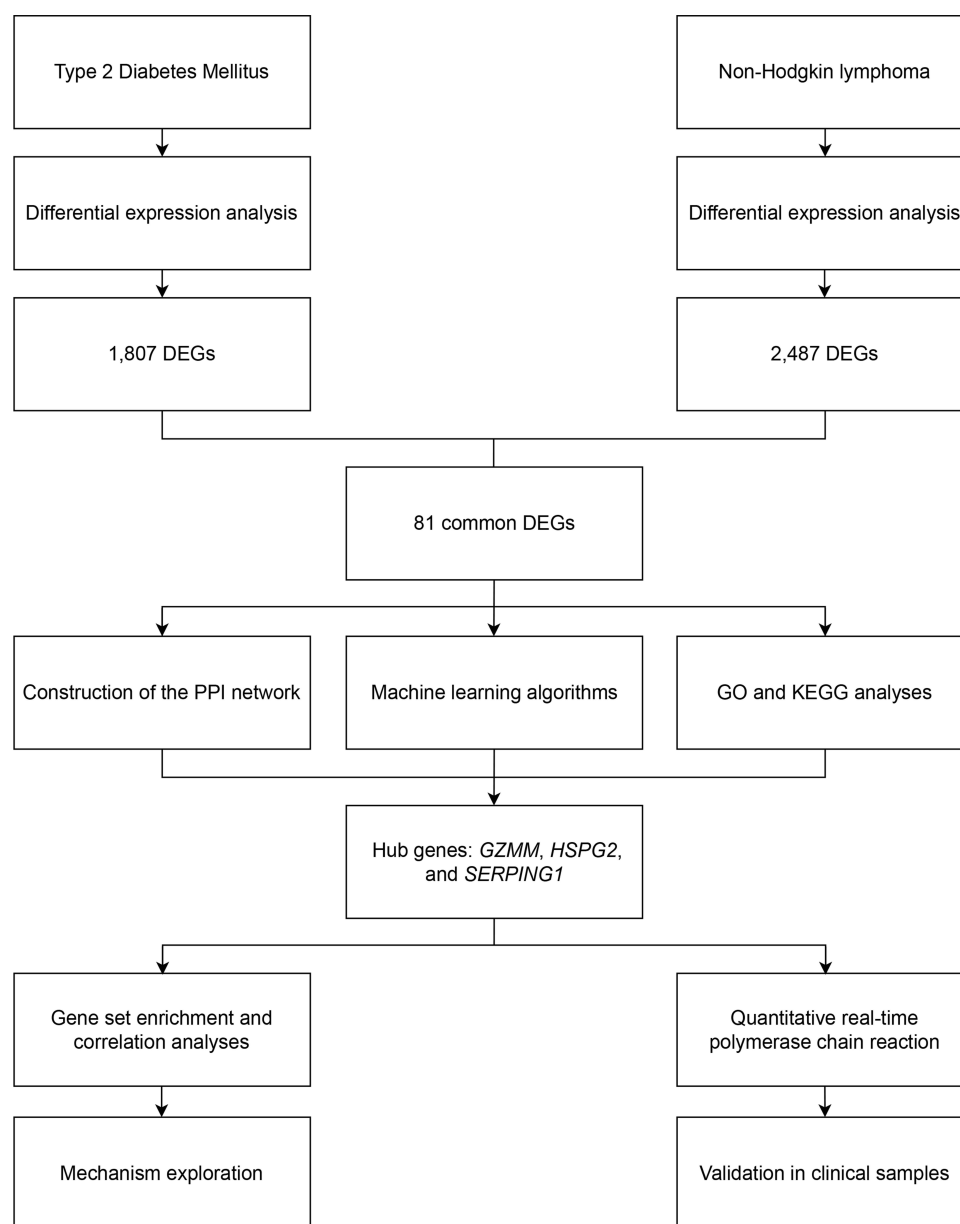
Figure 1 Flowchart of this study.

positive regulation of cell activation, response to type I interferon, cellular response to type I interferon, and receptor signaling pathway via STAT.

Selection of Hub Genes Using Machine Learning Algorithms

Multiple machine learning algorithms were used to identify hub genes from the set of ten key genes. Using LASSO regression, four genes were identified based on the variables (Figure 5A and B; Supplementary Table S3). The SVM-RFE method identified six genes with an accuracy of 0.98 and an error of 0.02 (Figure 5C and D; Supplementary Table S4). Moreover, the RF algorithm was used to analyze the relationship between the error rate and the number of classification trees, ultimately identifying seven genes of relative importance (Figure 5E and F; Supplementary Table S5). By cross-referencing the results of all three algorithms, we identified three overlapping hub genes: *GZMM, HSPG2*, and *SERPING1* (Figure 5G).
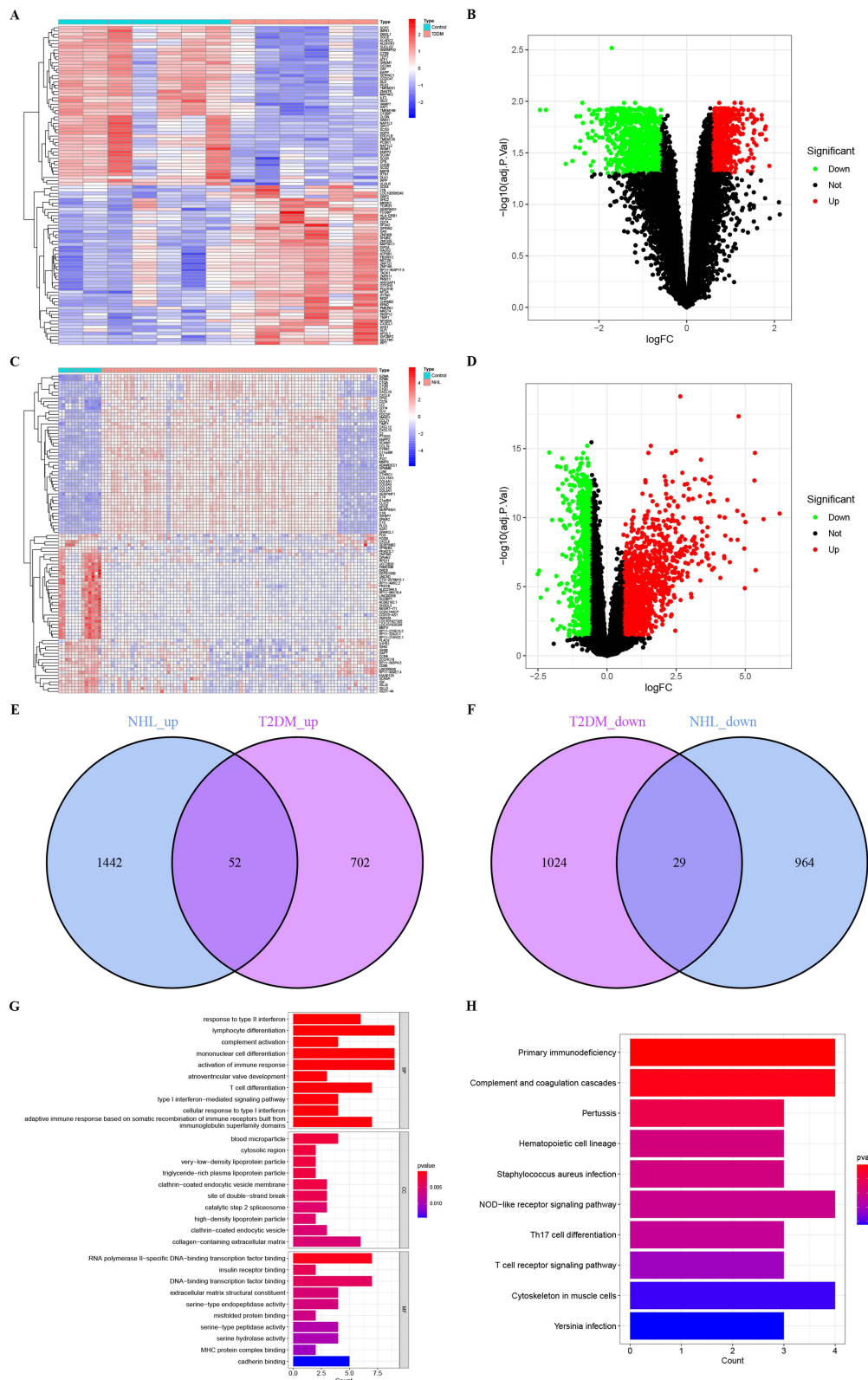
Figure 2 Identification of differentially expressed genes (DEGs). (**A** and **B**) Heatmap (**A**) and volcano plot (**B**) of the DEGs of T2DM. (**C** and **D**) Heatmap (**C**) and volcano plot (**D**) of the DEGs of NHL. (**E**) Venn diagram of the common upregulated DEGs. (**F**) Venn diagram of the common downregulated DEGs. (**G**) The barplot of the Gene Ontology analysis. (**H**) The barplot of the Kyoto Encyclopedia of Genes and Genomes analysis.
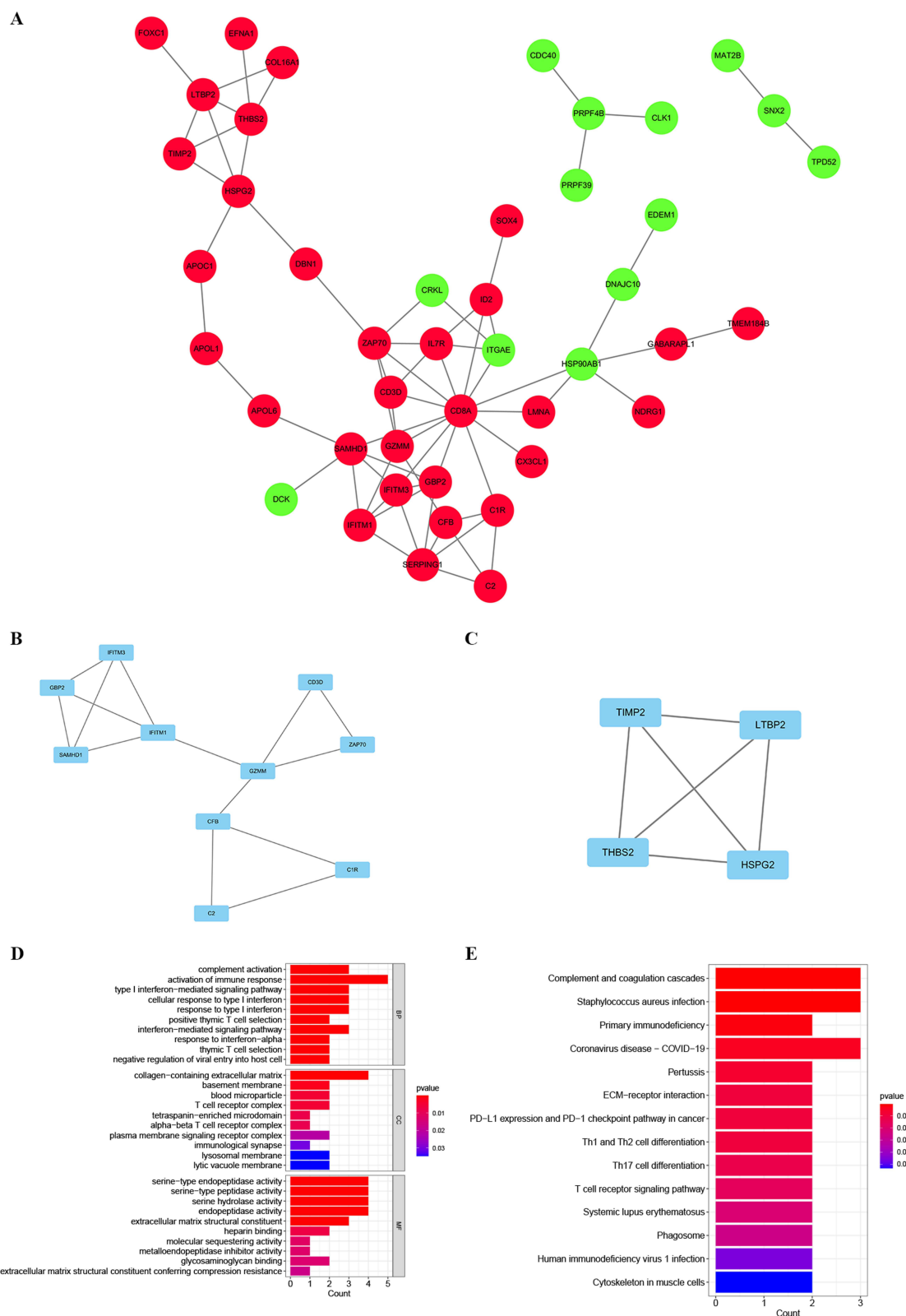
Figure 3 Construction of the protein-protein interaction (PPI) network and module analysis. (**A**) The PPI network of the common differentially expressed genes between type 2 diabetes mellitus and non-Hodgkin lymphoma, with downregulated genes denoted by green dots and upregulated genes indicated by red dots. (**B** and **C**) Two interconnected gene modules. (**D**) The barplot of the Gene Ontology analysis. (**E**) The barplot of the Kyoto Encyclopedia of Genes and Genomes analysis.
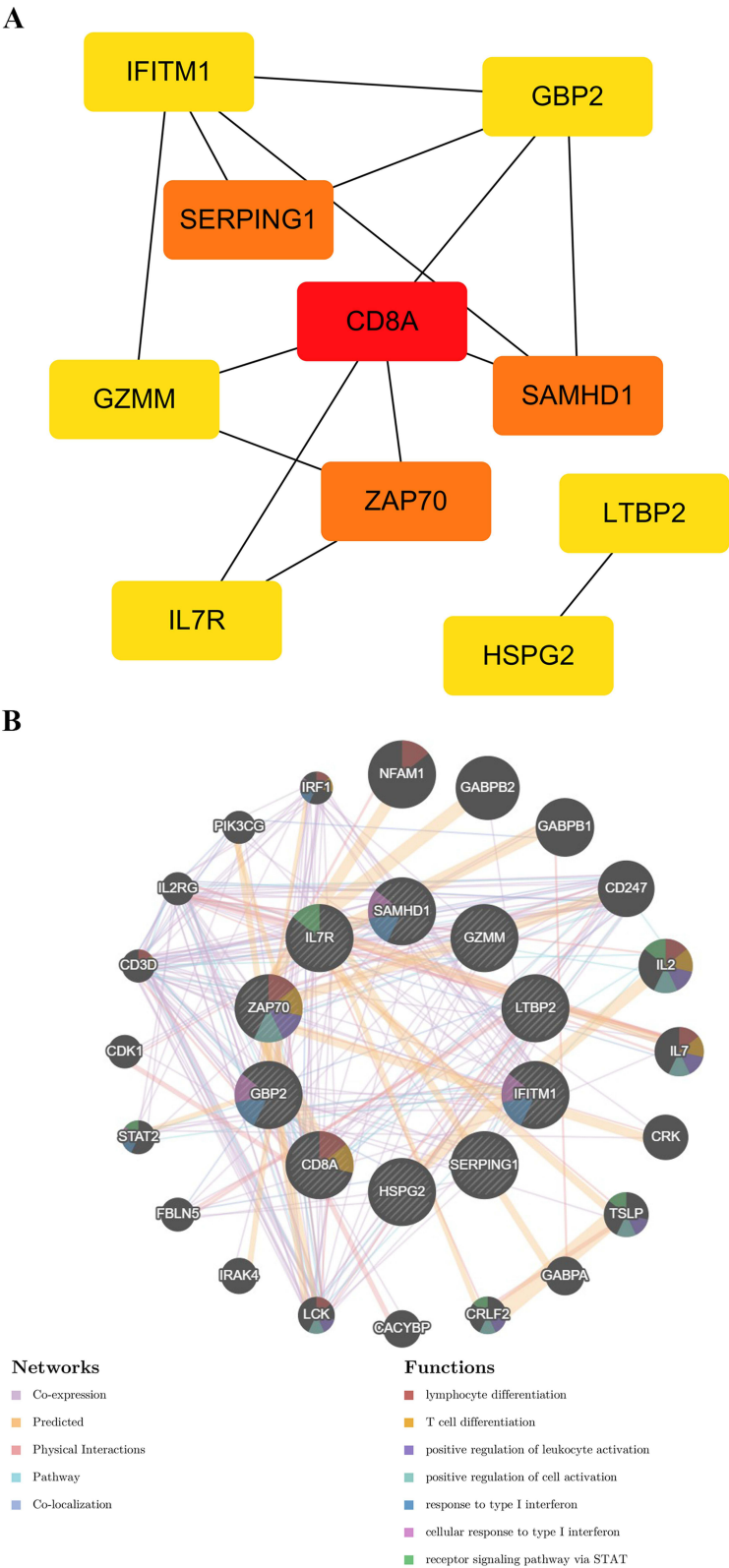
Figure 4 The top 10 genes in the protein-protein interaction network. (**A**) The ten genes with the highest degree values. Genes are represented using a color gradient, with darker hues indicating higher degree values. Bright red corresponds to the highest degree, while yellow represents the lowest. (**B**) The co-expression network of key genes derived from the GeneMANIA database.
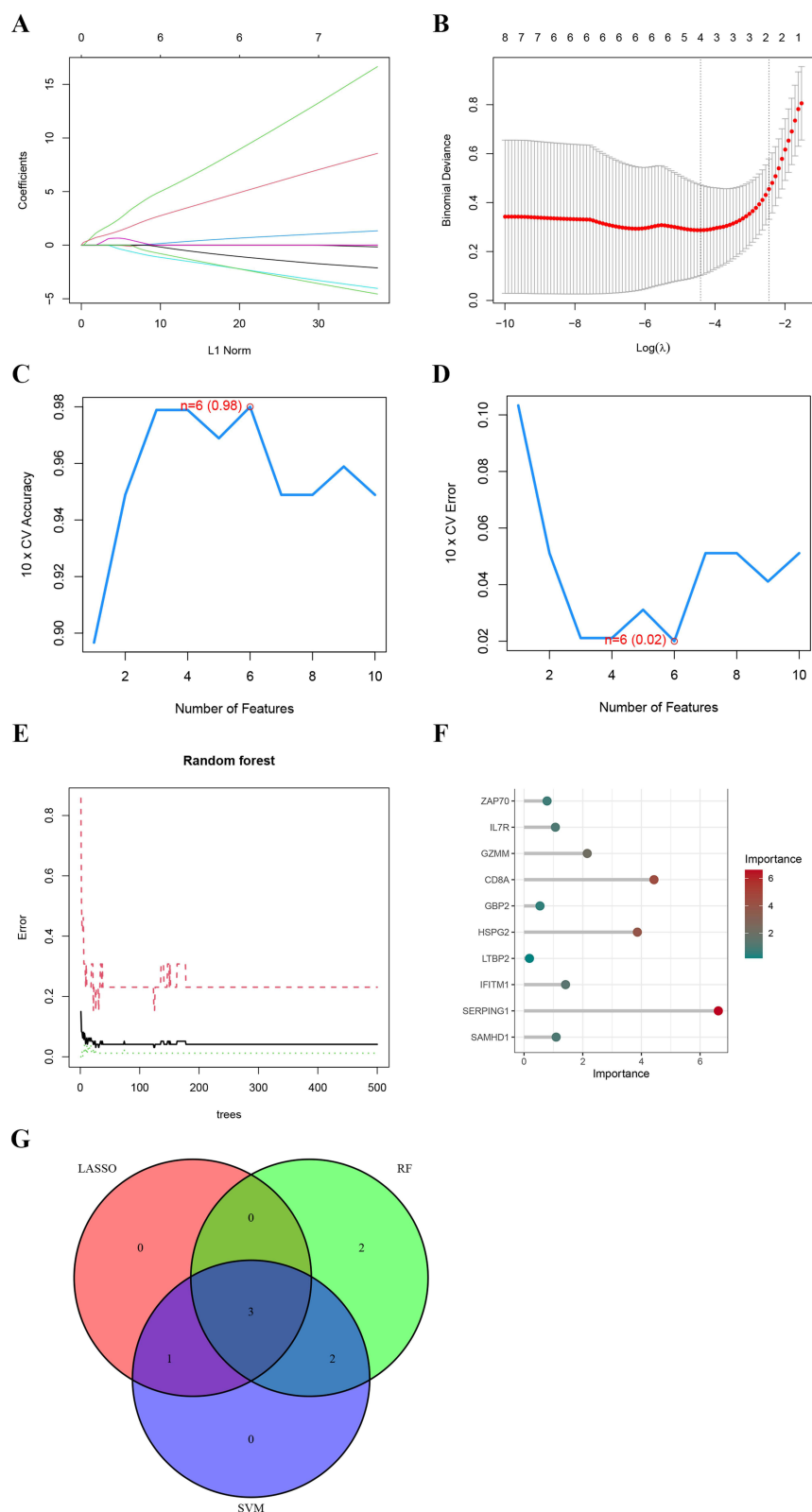
Figure 5 Selection of hub genes using machine learning algorithms. (**A** and **B**) Regression coefficients (**A**) and cross-validation curve (**B**) derived from the least absolute shrinkage and selection operator regression algorithm. (**C** and **D**) Cross-validation accuracy (**C**) and error (**D**) in the support vector machine-recursive feature elimination algorithm. (**E**) Relationship between error rate and quantity of trees in the random forest algorithm. (**F**) Importance of feature genes. (**G**) Venn diagram of the hub genes from the intersection of results from the above-mentioned algorithms.

GSEA and Correlation Analysis

To investigate the underlying mechanisms, GSEA was conducted to identify pathways significantly associated with hub genes. *GZMM* was predominantly enriched in complement and coagulation cascades, natural killer cell-mediated cytotoxicity, and T-cell receptor signaling pathway (Figure 6A). *HSPG2* was enriched in complement and coagulation cascades, ECM-receptor interaction, and cancer pathways (Figure 6B). *SERPING* was primarily enriched in complement and coagulation cascades, cytokine-cytokine receptor interaction, and ECM-receptor interaction (Figure 6C). Collectively, these findings indicated that all hub genes were enriched in immune-related pathways. Further analysis using ssGSEA revealed significant differences in the enrichment levels of 19 of the 28 immune cells between T2DM patients and controls (Figure 6D). In patients with NHL, all 28 immune cells showed significant differences compared to the controls (Figure 6E). These findings suggested a strong association between immune cell infiltration and both diseases. Furthermore, correlation analysis demonstrated significant associations between the hub genes and immune cells (Figure 6F).

qRT-PCR

To validate these findings, qRT-PCR was performed on clinical samples from patients with T2DM, patients with NHL, and controls. As anticipated, T2DM patients showed significantly higher expression levels of *GZMM, HSPG2*, and *SERPING1* than the controls (Figure 7A–C). Similarly, NHL patients exhibited a consistent pattern, with all three hub genes showing significant upregulation compared to controls (Figure 7D–F).

Discussion

Previous studies have indicated that T2DM may increase the risk of developing NHL, although the mechanisms underlying this association remain unclear. In this study, functional analyses suggested an important role of immune-related pathways in the co-pathogenesis of T2DM and NHL. Identifying biomarkers is crucial to elucidate the molecular connections between these diseases. By integrating microarray analysis and machine learning, we identified three hub genes involved in the co-pathogenesis of T2DM and NHL. These genes demonstrated significant correlations with immune cells, further supporting the potential role of immune dysregulation in the shared pathogenesis. Experimental validation using clinical samples was performed to ensure the reliability of the hub genes in clinical practice. These findings provide valuable insights into the shared pathogenic mechanisms of T2DM and NHL, emphasizing the genetic changes that inform molecular pathways and could guide future research and therapeutic strategies.

Functional analyses highlighted the role of immune-related pathways in the co-pathogenesis of T2DM and NHL. Immune cell analysis further demonstrated a significant association between both diseases and immune cell infiltration, highlighting immune dysfunction as a common feature. Patients with T2DM demonstrate increased susceptibility to infections, resulting in higher morbidity and mortality rates than non-diabetic individuals.[33] This increased susceptibility is largely due to abnormalities in both the innate and adaptive immune responses that interact with each other during the progression of T2DM.[34–36] Specific changes, such as altered T cell and macrophage proliferation as well as impaired NK cell and B cell function, reflect an overall dysfunction in immune regulation in patients with T2DM.[37] These immune alterations suggest that patients with T2DM may be immunodeficient, increasing their vulnerability to various diseases, including cancer. On the other hand, NHL is associated with dysregulation of the immune system. Individuals with compromised immune systems, such as those with HIV/AIDS or those who have undergone organ transplantation, and patients with autoimmune disorders involving chronic immune activation and inflammation, are known to have an elevated risk of NHL.[38–40] Disruptions in immune function promote the neoplastic transformation of blood cells into lymphoid malignancies through mechanisms, such as chronic immune stimulation, persistent inflammation, defective immune surveillance, and impaired anticancer immunity.[40,41] Therefore, it is plausible that chronic immune dysfunction in T2DM may create an environment of sustained inflammation and impaired immune surveillance, facilitating the development of lymphoid malignancies such as NHL.

Ten key genes were identified in the PPI network of the DEGs shared between T2DM and NHL: *ZAP70, IL7R, GZMM, CD8A, GBP2, HSPG2, LTBP2, IFITM1, SERPING1*, and *SAMHD1. ZAP70* is a cytoplasmic tyrosine kinase that plays an essential role in T-cell receptor signaling.[42,43] Previous research has identified *ZAP70* as a promising biomarker for follicular lymphoma and implicated it in the development and progression of this disease through
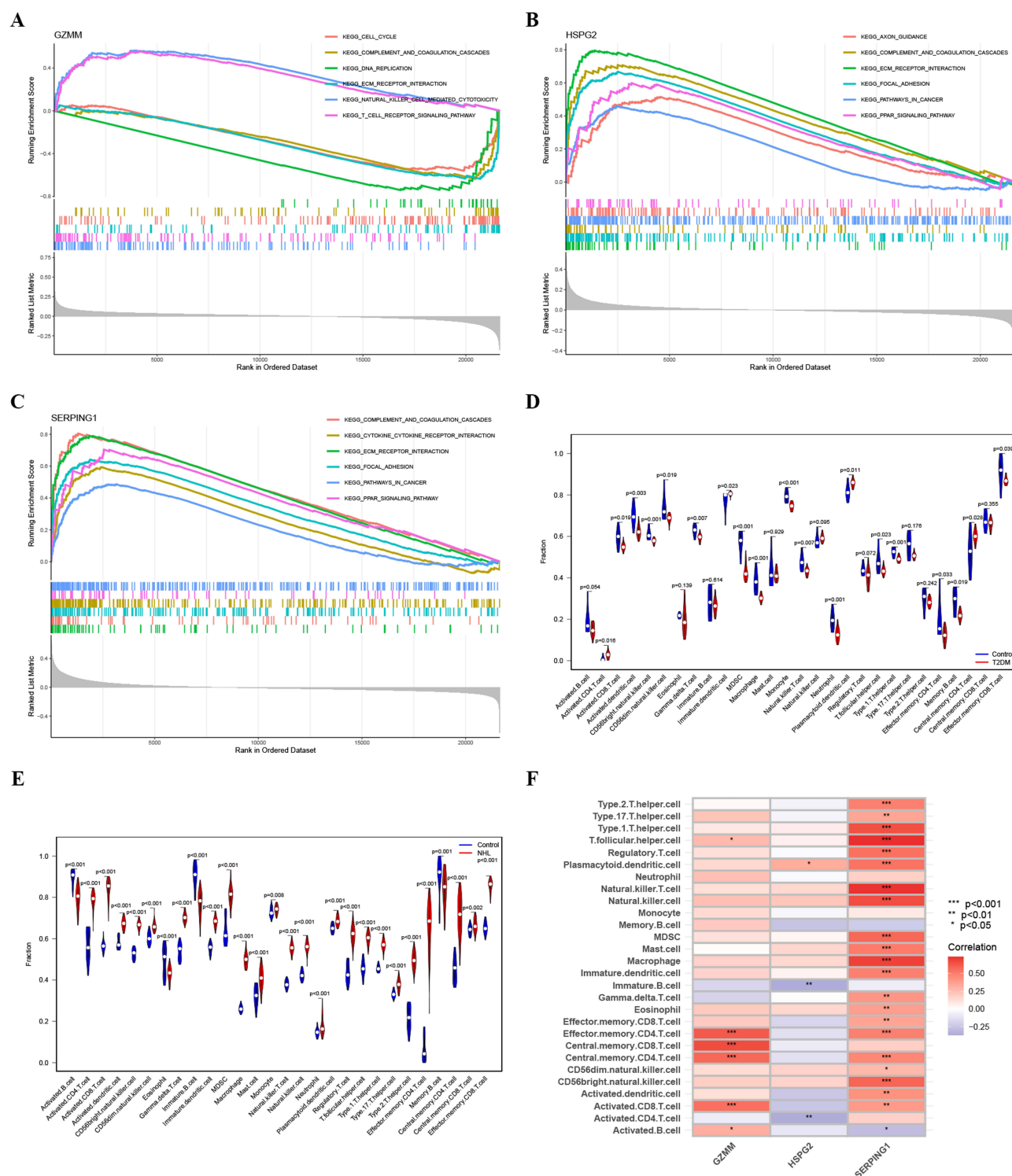
Figure 6 Gene set enrichment analysis (GSEA) and correlation analysis. (**A-C**) GSEA of the significantly enriched pathways of *GZMM* (**A**), *HSPG2* (**B**), and *SERPING1* (**C**). (**D**) Violin plot of immune cells between T2DM and control samples. (**E**) Violin plot of immune cells between NHL and control samples. (**F**) Heatmap of correlations between hub genes and immune cells. Statistical significance was determined using the following thresholds: *p < 0.05, **p < 0.01, and ***p < 0.001.

immune-related pathways.[44] *IL7R*, the receptor for interleukin-7, is critical for B cell development and T cell maturation.[45,46] *GZMM*, a serine protease expressed in cytotoxic lymphocytes, contributes to tumor cell destruction, inhibits cytomegalovirus replication, and plays a role in inflammation.[47] *CD8A* encodes the alpha chain of the CD8 glycoprotein, which is predominantly expressed in cytotoxic T cells and is vital for antigen recognition and immune
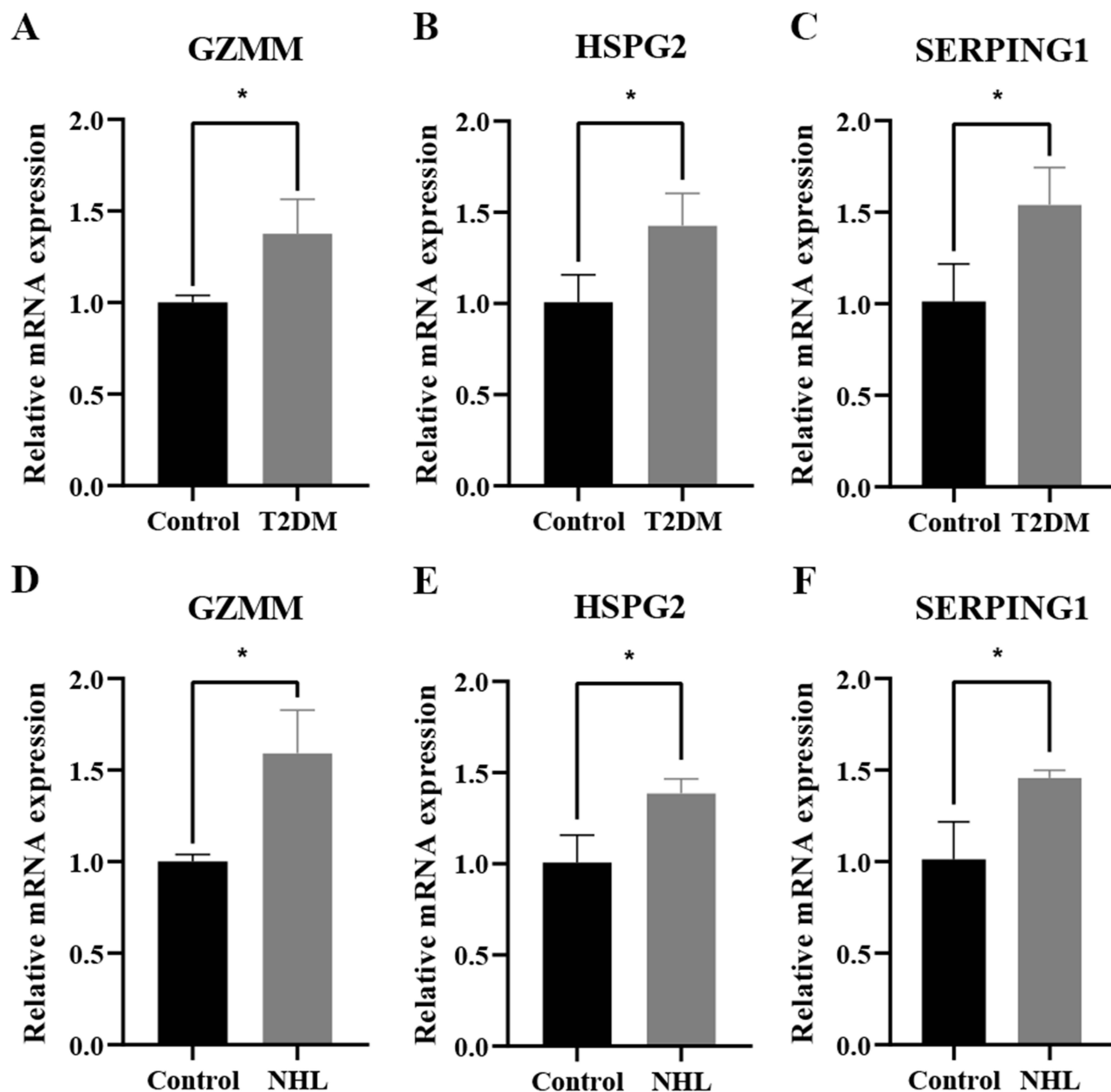
Figure 7 qRT-PCR. (**A-C**) Expression levels of *GZMM* (**A**), *HSPG2* (**B**), and *SERPING1* (**C**) between T2DM and control samples. (**D-F**) Expression of *GZMM* (**D**), *HSPG2* (**E**), and *SERPING1* (**F**) between NHL and control samples. Statistical significance was determined using the following threshold: *$p < 0.05$.

responses.[48,49] *GBP2*, part of the GTPase family, is essential in innate immunity against bacterial, viral, and protozoan pathogens and is significantly upregulated by interferon-γ.[50,51] *HSPG2* encodes perlecan, a protein that plays a key role in ECM and immune signaling.[52,53] *LTBP2*, a member of the ECM glycoprotein family, regulates the activity and function of transforming growth factor-beta, which has immune-modulating properties.[54,55] *IFITM1* is an interferon-induced transmembrane protein that is involved in antiviral immunity, immune response regulation, and cellular membrane function.[56,57] *SERPING1*, a plasma protein, regulates immune responses and blood clotting by inhibiting key enzymes in the complement, coagulation, and fibrinolytic systems.[58] *SAMHD1*, a triphosphohydrolase, restricts viral replication by degrading intracellular deoxynucleoside triphosphate in non-dividing cells.[59-61] All these genes share a connection with immune system regulation, highlighting the immune dysregulation underlying the co-pathogenesis of T2DM and NHL.

To further explore the biomarker genes involved in the co-pathogenesis of T2DM and NHL, we applied machine learning algorithms that were known for their ability to identify hidden patterns and enhance predictive model accuracy.[26,62,63] This approach led to the identification of three hub genes as key contributors to shared pathogenesis: *GZMM, HSPG2*, and *SERPING1*. Validation with clinical samples confirmed the reliability of these genes in clinical practice. GSEA revealed that all three hub genes were enriched in the immune-related pathways. Correlation analysis also revealed significant associations between these hub genes and immune cells, further supporting the hypothesis that immune dysregulation played a central role in the co-pathogenesis of T2DM and NHL. By identifying biomarker genes and exploring the involved mechanisms connecting these two diseases, this study provides a foundation for future research and potential therapeutic development. However, further research is required to elucidate the specific mechanisms by which these genes mediate the complex interplay between T2DM and NHL.

In previous studies, hub genes associated with T2DM and NHL were explored separately.[64–68] In contrast, our study focused on identifying shared hub genes and the mechanisms underlying these two diseases. Functional analyses highlighted the critical role of immune-related pathways in the co-pathogenesis of T2DM and NHL. Additionally, using a combination of microarray analysis and machine learning, we identified three hub genes that showed strong correlations with immune cells, further emphasizing the role of immune dysregulation in shared pathogenesis. Furthermore, hub genes were successfully validated in clinical samples. Our study introduced a novel integrative approach combining microarray analysis, machine learning, and experimental validation to provide a comprehensive understanding of the mechanisms underlying the association between T2DM and NHL. However, it is important to acknowledge the limitations of this study. First, the analyzed datasets from the GEO database primarily focused on B-cell NHL, because data on T-cell NHL were less comprehensive. Although the immunological similarities between the two subtypes suggest potential parallels, future studies should include T-cell NHL to validate the generalizability of our findings. Second, the relatively small sample size of the clinical specimens used for experimental validation may limit the robustness of our conclusions. Larger sample sizes are necessary in future studies to strengthen our results. Finally, although potential hub genes were identified, the exact mechanisms in mediating the interaction between T2DM and NHL remained unclear. Further experimental studies are required to fully elucidate these mechanisms. Addressing these limitations will be the focus of future studies to achieve a more comprehensive understanding of the shared pathophysiology between T2DM and NHL.

Conclusion

Our study suggested that immune-related pathways played a significant role in the shared pathogenesis of T2DM and NHL. By integrating microarray analysis, machine learning, and experimental validation, we identified and validated three key hub genes that were critical contributors to co-pathogenesis. These findings enhance our understanding of the complex interplay between T2DM and NHL and may pave the way for future therapeutic advancements.

Abbreviations

T2DM, Type 2 diabetes mellitus; NHL, Non-Hodgkin lymphoma; GEO, Gene Expression Omnibus; OGTT, Oral glucose tolerance test; HbA1c, Hemoglobin A1c; DEGs, Differentially expressed genes; PPI, Protein-protein interaction; MCODE, Molecular complex detection; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; LASSO, Least absolute shrinkage and selection operator; SVM, Support vector machine; RFE, Recursive feature elimination; RF, Random forest; GSEA, Gene set enrichment analysis; ssGSEA, Single-sample gene set enrichment analysis; qRT-PCR, Quantitative real-time PCR; *ZAP70*, Zeta chain of T cell receptor associated protein kinase 70; *IL7R*, Interleukin-7 receptor; *GZMM*, Granzyme M; *CD8A*, CD8 subunit alpha; *GBP2*, Guanylate binding protein 2; *HSPG2*, Heparan sulfate proteoglycan 2; *LTBP2*, Latent transforming growth factor beta binding protein 2; *IFITM1*, Interferon-induced transmembrane protein 1; *SERPING1*, Serpin family G member 1; *SAMHD1*, SAM and HD domain containing deoxynucleoside triphosphate triphosphohydrolase 1; ECM, Extracellular matrix.

Data Sharing Statement

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

Ethics Approval and Consent to Participate

This study was approved by the Ethics Committee of Shanghai Yangpu Hospital (LL-012). All procedures were performed in accordance with the principles of the Declaration of Helsinki.

Consent for Publication

Informed consent was obtained from all participants.

Acknowledgments

We thank the contributions of the GEO project.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

Disclosure

The authors declare no conflicts of interest regarding the publication of this paper.

References

1. Ahmad E, Lim S, Lamptey R, Webb DR, Davies MJ. Type 2 diabetes. *Lancet*. 2022;400(10365):1803–1820. doi:10.1016/S0140-6736(22)01655-5
2. ElSayed NA, Aleppo G, Aroda VR, et al. 2. classification and diagnosis of diabetes: standards of care in diabetes-2023. *Diabetes Care*. 2023;46 (Suppl 1):S19–s40. doi:10.2337/dc23-S002
3. Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. *Lancet*. 2017;389(10085):2239–2251. doi:10.1016/S0140-6736(17)30058-2
4. DeFronzo RA, Ferrannini E, Groop L, et al. Type 2 diabetes mellitus. *Nat Rev Dis Primers*. 2015;1:15019. doi:10.1038/nrdp.2015.19
5. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat Rev Endocrinol*. 2018;14 (2):88–98. doi:10.1038/nrendo.2017.151
6. Wang Y, Wang B, Yan S, et al. Type 2 diabetes and gender differences in liver cancer by considering different confounding factors: a meta-analysis of cohort studies. *Ann Epidemiol*. 2016;26(11):764–772. doi:10.1016/j.annepidem.2016.09.006
7. Gu J, Yan S, Wang B, et al. Type 2 diabetes mellitus and risk of gallbladder cancer: a systematic review and meta-analysis of observational studies. *Diabetes/Metab Res Rev*. 2016;32(1):63–72. doi:10.1002/dmrr.2671
8. Bansal D, Bhansali A, Kapil G, Undela K, Tiwari P. Type 2 diabetes and risk of prostate cancer: a meta-analysis of observational studies. *Prostate Cancer Prostatic Dis*. 2013;16(2):151–8,s1. doi:10.1038/pcan.2012.40
9. Miao ZF, Xu H, Xu YY, et al. Diabetes mellitus and the risk of gastric cancer: a meta-analysis of cohort studies. *Oncotarget*. 2017;8 (27):44881–44892. doi:10.18632/oncotarget.16487
10. Lee JY, Jeon I, Lee JM, Yoon JM, Park SM. Diabetes mellitus as an independent risk factor for lung cancer: a meta-analysis of observational studies. *European J Cancer*. 2013;49(10):2411–2423. doi:10.1016/j.ejca.2013.02.025
11. Gong Y, Wei B, Yu L, Pan W. Type 2 diabetes mellitus and risk of oral cancer and precancerous lesions: a meta-analysis of observational studies. *Oral Oncol*. 2015;51(4):332–340. doi:10.1016/j.oraloncology.2015.01.003
12. Castillo JJ, Mull N, Reagan JL, Nemr S, Mitri J. Increased incidence of non-Hodgkin lymphoma, leukemia, and myeloma in patients with diabetes mellitus type 2: a meta-analysis of observational studies. *Blood*. 2012;119(21):4845–4850. doi:10.1182/blood-2011-06-362830
13. Yan P, Wang Y, Fu T, Liu Y, Zhang ZJ. The association between type 1 and 2 diabetes mellitus and the risk of leukemia: a systematic review and meta-analysis of 18 cohort studies. *Endocr J*. 2021;68(3):281–289. doi:10.1507/endocrj.EJ20-0138
14. Atchison EA, Gridley G, Carreon JD, Leitzmann MF, McGlynn KA. Risk of cancer in a large cohort of U.S. veterans with diabetes. *Int J Cancer*. 2011;128(3):635–643. doi:10.1002/ijc.25362
15. Maskarinec G, Brown SM, Lee J, et al. Association of obesity and type 2 diabetes with non-Hodgkin lymphoma: the multiethnic cohort. *Cancer Epidemiol Biomarkers Prevention*. 2023;32(10):1348–1355. doi:10.1158/1055-9965.EPI-23-0565
16. Wang Z, Phillips LS, Rohan TE, et al. Diabetes, metformin use and risk of non-Hodgkin's lymphoma in postmenopausal women: a prospective cohort analysis in the women's health initiative. *Int J Cancer*. 2023;152(8):1556–1569. doi:10.1002/ijc.34376
17. Wang Y, Liu X, Yan P, Bi Y, Liu Y, Zhang ZJ. Association between type 1 and type 2 diabetes and risk of non-Hodgkin's lymphoma: a meta-analysis of cohort studies. *Diabetes Metabolism*. 2020;46(1):8–19. doi:10.1016/j.diabet.2019.04.006
18. Tseng CH. Diabetes and non-Hodgkin's lymphoma: analyses of prevalence and annual incidence in 2005 using the national health insurance database in Taiwan. *Ann Oncol*. 2012;23(1):153–158. doi:10.1093/annonc/mdr334

19. Thandra KC, Barsouk A, Saginala K, Padala SA, Barsouk A, Rawla P. Epidemiology of Non-Hodgkin's Lymphoma. *Med Sci*. 2021;9(1). doi:10.3390/medsci9010005

20. Berndt SI, Vijai J, Benavente Y, Camp NJ. Distinct germline genetic susceptibility profiles identified for common non-Hodgkin lymphoma subtypes. *Leukemia*. 2022;36(12):2835–2844. doi:10.1038/s41375-022-01711-0

21. Falchi L. Immune dysfunction in non-Hodgkin lymphoma: avenues for new immunotherapy-based strategies. *Curr Hematol Malignancy Rep*. 2017;12(5):484–494. doi:10.1007/s11899-017-0410-1

22. Agapito G, Arbitrio M. Microarray data analysis protocol. *Methods mol Biol*. 2022;2401:263–271.

23. Kotlyar M, Wong SWH, Pastrello C, Jurisica I. Improving analysis and annotation of microarray data with protein interactions. *Methods mol Biol*. 2022;2401:51–68.

24. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat Rev mol Cell Biol*. 2022;23(1):40–55. doi:10.1038/s41580-021-00407-0

25. Ledesma D, Symes S, Richards S. Advancements within modern machine learning methodology: impacts and prospects in biomarker discovery. *Curr Med Chem*. 2021;28(32):6512–6531. doi:10.2174/0929867328666210208111821

26. Binson VA, Thomas S, Subramoniam M, Arun J, Naveen S, Madhu S. A review of machine learning algorithms for biomedical applications. *Annals of Biomedical Engineering*. 2024;52(5):1159–1183. doi:10.1007/s10439-024-03459-3

27. Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: a brief primer. *Behavior Ther*. 2020;51(5):675–687. doi:10.1016/j.beth.2020.05.002

28. Lee YW, Choi JW, Shin EH. Machine learning model for predicting malaria using clinical information. *Comput. Biol. Med*. 2021;129:104151. doi:10.1016/j.compbiomed.2020.104151

29. Zeng J, Lai C, Luo J, Li L. Functional investigation and two-sample Mendelian randomization study of neuropathic pain hub genes obtained by WGCNA analysis. *Front Neurosci*. 2023;17:1134330. doi:10.3389/fnins.2023.1134330

30. McNeish DM. Using lasso for predictor selection and to assuage overfitting: a method long overlooked in behavioral sciences. *Multivariate Behav Res*. 2015;50(5):471–484. doi:10.1080/00273171.2015.1036965

31. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*. 2018;15(1):41–51.

32. Rigatti SJ. Random Forest. *J Insurance Med*. 2017;47(1):31–39.

33. de Lourdes Ochoa-González F, González-Curiel IE, Cervantes-Villagrana AR, Fernández-Ruiz JC, Castañeda-Delgado JE. Innate immunity alterations in type 2 diabetes mellitus: understanding infection susceptibility. *Curr Mol Med*. 2021;21(4):318–331. doi:10.2174/1566524020999200831124534

34. Richardson VR, Smith KA, Carter AM. Adipose tissue inflammation: feeding the development of type 2 diabetes mellitus. *Immunobiology*. 2013;218(12):1497–1504. doi:10.1016/j.imbio.2013.05.002

35. Lee MS. Role of innate immunity in the pathogenesis of type 1 and type 2 diabetes. *J Korean Med Sci*. 2014;29(8):1038–1041. doi:10.3346/jkms.2014.29.8.1038

36. SantaCruz-Calvo S, Bharath L, Pugh G, et al. Adaptive immune cells shape obesity-associated type 2 diabetes mellitus and less prominent comorbidities. *Nat Rev Endocrinol*. 2022;18(1):23–42. doi:10.1038/s41574-021-00575-1

37. Zhou T, Hu Z, Yang S, Sun L, Yu Z, Wang G. Role of adaptive and innate immunity in type 2 diabetes mellitus. *Journal of Diabetes Research*. 2018;2018:7457269. doi:10.1155/2018/7457269

38. Ballow M, Sánchez-Ramón S, Walter JE. Secondary immune deficiency and primary immune deficiency crossovers: hematological malignancies and autoimmune diseases. *Front Immunol*. 2022;13:928062. doi:10.3389/fimmu.2022.928062

39. Baecklund E, Smedby KE, Sutton LA, Askling J, Rosenquist R. Lymphoma development in patients with autoimmune and inflammatory disorders--what are the driving forces? *Semi Cancer Biol*. 2014;24:61–70. doi:10.1016/j.semcancer.2013.12.001

40. Ponce RA, Gelzleichter T, Haggerty HG, et al. Immunomodulation and lymphoma in humans. *J Immunotoxicol*. 2014;11(1):1–12. doi:10.3109/1547691X.2013.798388

41. Stevens WB, Netea MG, Kater AP, van der Velden WJ. 'Trained immunity': consequences for lymphoid malignancies. *Haematologica*. 2016;101(12):1460–1468. doi:10.3324/haematol.2016.149252

42. Ashouri JF, Lo WL, Nguyen TTT, Shen L, Weiss A. ZAP70, too little, too much can lead to autoimmunity. *Immunol Rev*. 2022;307(1):145–160. doi:10.1111/imr.13058

43. Au-Yeung BB, Shah NH, Shen L, Weiss A. ZAP-70 in signaling, biology, and disease. *Ann Rev Immunol*. 2018;36(1):127–156. doi:10.1146/annurev-immunol-042617-053335

44. Zhu Y, Jin X, Liu J, Yang W. Identification and functional investigation of hub genes associated with follicular lymphoma. *Biochem. Genet*. 2024. doi:10.1007/s10528-024-10831-4

45. Barata JT, Durum SK, Seddon B. Flip the coin: IL-7 and IL-7R in health and disease. *Nat Immunol*. 2019;20(12):1584–1593. doi:10.1038/s41590-019-0479-x

46. Wang C, Kong L, Kim S, et al. The Role of IL-7 and IL-7R in cancer pathophysiology and immunotherapy. *Int J mol Sci*. 2022;23(18):10412.

47. de Poot SA, Bovenschen N. Granzyme M: behind enemy lines. *Cell Death Differ*. 2014;21(3):359–368. doi:10.1038/cdd.2013.189

48. Dumontet E, Osman J, Guillemont-Lambert N, Cros G, Moshous D, Picard C. Recurrent respiratory infections revealing CD8α deficiency. *J Clin Immunol*. 2015;35(8):692–695. doi:10.1007/s10875-015-0213-x

49. Bernardo I, Mancebo E, Aguiló I, et al. Phenotypic and functional evaluation of CD3+CD4-CD8- T cells in human CD8 immunodeficiency. *Haematologica*. 2011;96(8):1195–1203. doi:10.3324/haematol.2011.041301

50. Braun E, Hotter D, Koepke L, et al. Guanylate-binding proteins 2 and 5 exert broad antiviral activity by inhibiting furin-mediated processing of viral envelope proteins. *Cell Rep*. 2019;27(7):2092–104.e10. doi:10.1016/j.celrep.2019.04.063

51. Tretina K, Park ES, Maminska A, MacMicking JD. Interferon-induced guanylate-binding proteins: guardians of host defense in health and disease. *J Exp Med*. 2019;216(3):482–500. doi:10.1084/jem.20182031

52. Hayes AJ, Farrugia BL, Biose IJ, Bix GJ, Melrose J. Perlecan, A multi-functional, cell-instructive, matrix-stabilizing proteoglycan with roles in tissue development has relevance to connective tissue repair and regeneration. *Front Cell Develop Biol*. 2022;10:856261. doi:10.3389/fcell.2022.856261

53. Melrose J. Perlecan, a modular instructive proteoglycan with diverse functional properties. *Int J Biochem Cell Biol*. 2020;128:105849. doi:10.1016/j.biocel.2020.105849

54. Bodmer NK, Knutsen RH, Roth RA, et al. Multi-organ phenotypes in mice lacking latent TGFβ binding protein 2 (LTBP2). *Developmental Dynamics*. 2024;253(2):233–254. doi:10.1002/dvdy.651

55. Robertson IB, Horiguchi M, Zilberberg L, Dabovic B, Hadjiolova K, Rifkin DB. Latent TGF-β-binding proteins. *Matrix Biol*. 2015;47:44–53. doi:10.1016/j.matbio.2015.05.005

56. Yánez DC, Ross S, Crompton T. The IFITM protein family in adaptive immunity. *Immunology*. 2020;159(4):365–372. doi:10.1111/imm.13163

57. Gómez-Herranz M, Taylor J, Sloan RD. IFITM proteins: understanding their diverse roles in viral infection, cancer, and immunity. *J Biol Chem*. 2023;299(1):102741. doi:10.1016/j.jbc.2022.102741

58. Drouet C, López-Lera A, Ghannam A, et al. SERPING1 variants and C1-INH biological function: a close relationship with C1-INH-HAE. *Front Aller*. 2022;3:835503. doi:10.3389/falgy.2022.835503

59. Coggins SA, Mahboubi B, Schinazi RF, Kim B. SAMHD1 Functions and Human Diseases. *Viruses*. 2020;12(4):382. doi:10.3390/v12040382

60. Chen S, Bonifati S, Qin Z, St Gelais C, Wu L. SAMHD1 suppression of antiviral immune responses. *Trend Microbiol*. 2019;27(3):254–267. doi:10.1016/j.tim.2018.09.009

61. Deutschmann J, Gramberg T. SAMHD1 … and Viral Ways around It. *Viruses*. 2021;13(3):395. doi:10.3390/v13030395

62. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers*. 2021;25(3):1315–1360.

63. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *New Engl J Med*. 2023;388(13):1201–1208. doi:10.1056/NEJMra2302038

64. Lin Y, Li J, Wu D, Wang F, Fang Z, Shen G. Identification of hub genes in type 2 diabetes mellitus using bioinformatics analysis. *Diab Metab Syndrome Obesity*. 2020;13:1793–1801. doi:10.2147/DMSO.S245165

65. Li J, Yan N, Li X, He S, Yu X. Identification and analysis of hub genes of hypoxia-immunity in type 2 diabetes mellitus. *Front Genetics*. 2023;14:1154839. doi:10.3389/fgene.2023.1154839

66. Li Q, Meng Y, Hu L, Charwudzi A, Zhu W, Zhai Z. Integrative analysis of hub genes and key pathway in two subtypes of diffuse large B-cell lymphoma by bioinformatics and basic experiments. *J Clin Lab Analysis*. 2021;35(11):e23978. doi:10.1002/jcla.23978

67. Zhang Q, Wang M. Identification of hub genes and key pathways associated with follicular lymphoma. *Contrast Media mol Imag*. 2022;2022 (1):5369104. doi:10.1155/2022/5369104

68. Doughan A, Salifu SP. Genes associated with diagnosis and prognosis of Burkitt lymphoma. *Iet Syst Biol*. 2022;16(6):220–229. doi:10.1049/syb2.12054