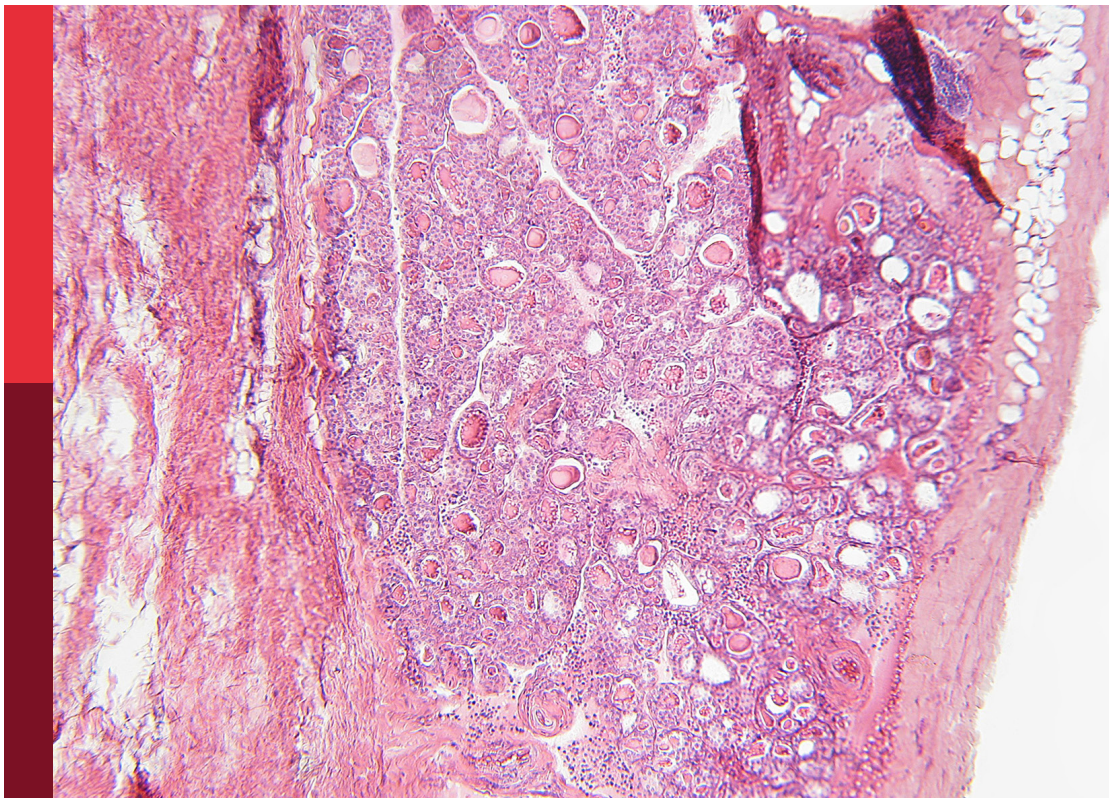
frontiers | Research Topics

# Machine learning-assisted diagnosis and treatment of endocrine-related diseases

**Edited by**
Qiuming Yao, Prem Prakash Kushwaha and Wenjie Shi

**Coordinated by**
Yutao Wang

**Published in**
Frontiers in Endocrinology



ISSN 1664-8714
ISBN 978-2-8325-4235-4
DOI 10.3389/978-2-8325-4235-4

## About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

# Machine learning-assisted diagnosis and treatment of endocrine-related diseases

**Topic editors**
Qiuming Yao — Fudan University, China
Prem Prakash Kushwaha — Case Western Reserve University, United States
Wenjie Shi — Otto von Guericke University Magdeburg, Germany

**Topic Coordinator**
Yutao Wang — Chinese Academy of Medical Sciences and Peking Union Medical College, China

**Citation**
Yao, Q., Kushwaha, P. P., Shi, W., Wang, Y., eds. (2023). *Machine learning-assisted diagnosis and treatment of endocrine-related diseases*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4235-4

# Table of contents

frontiers | Frontiers in Endocrinology

TYPE Editorial
PUBLISHED 18 December 2023
DOI 10.3389/fendo.2023.1305897



OPEN ACCESS

EDITED AND REVIEWED BY
Tom Michoel,
University of Bergen, Norway

*CORRESPONDENCE
Ulf D. Kahlert
ulf.kahlert@med.ovgu.de
Wenjie Shi
wenjie.shi@ovgu.de

RECEIVED 02 October 2023
ACCEPTED 06 December 2023
PUBLISHED 18 December 2023

CITATION
Zhang H, Kahlert UD and Shi W (2023) Editorial: Machine learning-assisted diagnosis and treatment of endocrine-related diseases. *Front. Endocrinol.* 14:1305897. doi: 10.3389/fendo.2023.1305897



# Editorial: Machine learning-assisted diagnosis and treatment of endocrine-related diseases

Heng Zhang[1], Ulf D. Kahlert[2*] and Wenjie Shi[2*]

[1]Department of Laboratory, Shandong Daizhuang Hospital, Jining, China, [2]Molecular and Experimental Surgery, University Clinic for General-, Visceral-, Vascular- and Trans-Plantation Surgery, Medical Faculty University Hospital Magdeburg, Otto-von Guericke University, Magdeburg, Germany

KEYWORDS

**machine learning, endocrine disorders, diagnosis, treatment, artificial intelligence**

Editorial on the Research Topic
Machine learning-assisted diagnosis and treatment of endocrine-related diseases

The endocrine system is an important regulatory system in the human body. It regulates physiological processes such as growth, development, metabolism, and reproduction through the secretion, transfer, and feedback of hormones. Various factors, including environmental, genetic, and lifestyle factors, can lead to the development of endocrine diseases associated with dysfunction of the pituitary, adrenal, and thyroid glands, as well as diabetes. Changes in living environments and lifestyles have led to steady increases in the incidence of endocrine diseases. If not detected and treated in time, these can lead to complications and cause irreversible damage to health.

The application of machine learning in medical fields has great prospects (1). This Research Topic is the use of machine learning to establish models for the prediction, diagnosis, and management of endocrine-related diseases for the benefit of both patients and clinical practice. The following is a brief overview of some of the included articles:

The model built by Wang et al. using the Gaussian Naive Bayes (GNB) algorithm allows improved prediction of prolonged hospitalization of patients following acute ischemic stroke (AIS). This will assist in policy adjustments for improved resource utilization, thereby alleviating the increasingly heavy economic burden caused by AIS.

Chen et al. used a two-sample Mendelian randomization (MR) study to explore the relationship between thyroid function and cholelithiasis. Their findings showed that low-density lipoprotein cholesterol (LDL-C) and apolipoprotein B mediated the effects of FT4 on cholelithiasis risk, with patients with high FT4 levels showing delayed or reduced risk of the long-term effects of cholelithiasis.

Zeng et al. analyzed data from the Genomics of Drug Sensitivity in Cancer (GDSC) and The Cancer Genome Atlas (TCGA) databases, using Spearman correlation analysis to identify the key genes that influence Gemcitabine (GEM) efficacy. It was

found that *CALB2* and *GPX3* could be used as biomarkers for prognosis prediction in some forms of colorectal cancer (CRC) as well as potential target genes of GEM, providing new ideas for the development of new combined targeted drugs for colorectal cancer.

Liu et al. established a machine learning-based choice for some type 2 diabetic kidney disease(T2DKD) risk prediction model based on clinical data from a multi-center retrospective database and verified its effectiveness. While the model was found to be helpful for the diagnosis of T2DKD, further investigation using additional data is required.

Huang et al. used the ESTIMATE algorithm to conduct a series of bioinformatics analyses on breast cancer (BC) samples from the TCGA database to identify genes associated with the tumor microenvironment (TME). The authors found a significant correlation between KLRB1 and the BC TME, suggesting its use as a prognostic marker and therapeutic target, providing a new direction for the treatment of BC.

Li et al. identified stratified prognostic biomarkers for serous ovarian cancer (SOC) after investigation of immune infiltration, drug sensitivity, and genes associated with the epithelial-mesenchymal transition (EMT). This lays a foundation for in-depth investigation of the role of the EMT in SOC immune regulation and changes in related pathways. It also suggests effective solutions for the early diagnosis and clinical treatment of ovarian cancer.

Xu et al. identified the C11 cluster that specifically expresses HSPB6 in fibroblasts as key to the development of pancreatic adenocarcinoma (PAAD). The authors used comprehensive bioinformatics analyses and constructed a nine-gene prognostic model using tumor-related PAAD prognostic genes in the C11 subgroup. The RiskScore may have reliable clinical potential for the prognostic prediction of PAAD.

With the increased informatization of society and further accumulation of data, the use of machine learning will become more common in medical applications requiring the processing of large amounts of data. Machine learning can process large amounts of data that humans cannot handle and is a powerful and versatile tool for future medicine (2). The application of machine learning in the diagnosis and treatment of endocrine diseases will become increasingly widespread and mature, as is currently seen in the prediction, diagnosis, and management of diseases such as diabetes (3), thyroid disease (4), and neuroendocrine tumors (5).

## Author contributions

HZ: Formal analysis, Investigation, Methodology, Resources, Writing – original draft. UK: Conceptualization, Data curation, Project administration, Resources, Writing – review & editing. WS: Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

## Conflict of interest

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.

## Publisher's note



## References

1. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. *J Intern Med* (2018) 284(6):603–19. doi: 10.1111/joim.12822

2. Komuro J, Kusumoto D, Hashimoto H, Yuasa S. Machine learning in cardiology: Clinical application and basic research. *J Cardiol* (2023) 82(2):128–33. doi: 10.1016/j.jjcc.2023.04.020

3. Afsaneh E, Sharifdini A, Ghazzaghi H, Ghobadi MZ. Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. *Diabetol Metab Syndr* (2022) 14 (1):196. doi: 10.1186/s13098-022-00969-9

4. Cao CL, Li QL, Tong J, Shi LN, Li WX, Xu Y, et al. Artificial intelligence in thyroid ultrasound. *Front Oncol* (2023) 13:1060702. doi: 10.3389/fonc.2023.1060702

5. Ramesh S, Dolezal JM, Pearson AT. Applications of deep learning in endocrine neoplasms. *Surg Pathol Clin* (2023) 16(1):167–76. doi: 10.1016/j.path.2022.09.014

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 29 March 2023
DOI 10.3389/fendo.2023.1166740



OPEN ACCESS

EDITED BY
Wenjie Shi,
Otto von Guericke University Magdeburg, Germany

REVIEWED BY
Yufei Liu,
Fudan University, China
Weixing Wang,
Renmin Hospital of Wuhan University, China

*CORRESPONDENCE
Kai Liu
liuk@jlu.edu.cn

SPECIALTY SECTION
This article was submitted to Systems Endocrinology, a section of the journal Frontiers in Endocrinology

RECEIVED 15 February 2023
ACCEPTED 20 March 2023
PUBLISHED 29 March 2023

CITATION
Chen J, Zhou H, Jin H and Liu K (2023) The causal effects of thyroid function and lipids on cholelithiasis: A Mendelian randomization analysis. *Front. Endocrinol.* 14:1166740. doi: 10.3389/fendo.2023.1166740



# The causal effects of thyroid function and lipids on cholelithiasis: A Mendelian randomization analysis

Junhong Chen, Hao Zhou, Hengwei Jin and Kai Liu*

Department of Hepatobiliary and Pancreatic Surgery II, General Surgery Center, The First Hospital of Jilin University, Changchun, China

**Objective:** To investigate the relationship between function of thyroid, lipids, and cholelithiasis and to identify whether lipids mediate the causal relationship between function of thyroid and cholelithiasis.

**Methods:** A Mendelian randomization (MR) study of two samples was performed to determine the association of thyroid function with cholelithiasis. A two-step MR was also performed to identify whether lipid metabolism traits mediate the effects of thyroid function on cholelithiasis. A method of inverse variance weighted (IVW), weighted median method, maximum likelihood, MR-Egger, MR-robust adjusted profile score (MR-RAPS) method, and MR pleiotropy residual sum and outlier test (MR-PRESSO) methods were utilized to obtain MR estimates.

**Results:** The IVW method revealed that FT4 levels were correlated with an elevated risk of cholelithiasis (OR: 1.149, 95% CI: 1.082–1.283, $P$ = 0.014). Apolipoprotein B (OR: 1.255, 95% CI: 1.027–1.535, $P$ = 0.027) and low-density lipoprotein cholesterol (LDL-C) (OR: 1.354, 95% CI: 1.060–1.731, $P$ = 0.016) were also correlated with an elevated risk of cholelithiasis. The IVW method demonstrated that FT4 levels were correlated with the elevated risk of apolipoprotein B (OR: 1.087, 95% CI: 1.019–1.159, $P$ = 0.015) and LDL-C (OR: 1.084, 95% CI: 1.018–1.153, $P$ = 0.012). Thyroid function and the risk of cholelithiasis are mediated by LDL-C and apolipoprotein B. LDL-C and apolipoprotein B had 17.4% and 13.5% of the mediatory effects, respectively.

**Conclusions:** We demonstrated that FT4, LDL-C, and apolipoprotein B had significant causal effects on cholelithiasis, with evidence that LDL-C and apolipoprotein B mediated the effects of FT4 on cholelithiasis risk. Patients with high FT4 levels should be given special attention because they may delay or limit the long-term impact on cholelithiasis risk.

KEYWORDS

**thyroid function, lipid metabolism traits, cholelithiasis, Mendelian randomization, mediation effects**

# Introduction

Cholelithiasis, a prevalent condition affecting approximately 10–20% of the global adult population, has experienced a recent upsurge in incidence (1). The association between cholelithiasis and the onset of the gallbladder, pancreatic, and colorectal cancers is well established (2). While the majority of affected adults remain asymptomatic, the economic and societal burdens of cholelithiasis can be substantial in the event of symptomatology or complications (3, 4).

Cholelithiasis remains a prevalent gastrointestinal disorder for which the pathophysiology is still unknown. Recent clinical observational studies have shed light on a potential association between thyroid function and cholelithiasis. Notably, a study encompassing a cohort of 3,749 subjects aged 20 to 79 demonstrated an independent correlation between cholelithiasis and elevated serum thyroid stimulating hormone (TSH) levels (5). Furthermore, patients with cholelithiasis exhibited a significantly higher prevalence of both subclinical and clinical hypothyroidism (6). However, conventional observational studies are limited in their ability to determine causal effects and account for potential confounding factors.

Several convincing studies state a positive correlation between high cholesterol levels and the development of cholelithiasis (7). Furthermore, serum FT4 levels are positively correlated with total triglycerides (TG) and LDL-C (8). It was also identified that lipid metabolism traits might mediate the causal effects of thyroid function on cholelithiasis.

Utilizing genetic variants as instrument variables (IVs), MR analysis has emerged as a powerful tool for determining the causal relationship between risk factors and diseases (9). Large-scale genome-wide summary association studies (GWAS) also allow for the systematic investigation of the causal effects of exposures on outcomes using MR (10). In the current investigation, MR analysis was used to evaluate the relationship between thyroid function, lipids, and cholelithiasis and to determine whether lipids mediate the causal effects of thyroid function on cholelithiasis.

# Materials and methods

## Study design and GWAS statistics source

The total effects were determined using a two-sample MR to evaluate the association between thyroid function and cholelithiasis. Another two-step MR was performed to assess whether lipid metabolism traits mediate the effects of thyroid function on cholelithiasis. First, we explore the relationship between function of thyroid and lipid metabolism traits. Second, we investigated the effects of lipid metabolism traits on cholelithiasis risk.

We retrieved the GWAS thyroid function summary data from the ThyroidOmics Consortium, which was formed to study the determinants and effects of thyroid disorders and thyroid function (11). In a meta-analysis, analyses of TSH comprised information from 22 different cohorts comprising 54,288 individuals, FT4 analyses had data from 19 cohorts with 49,269 individuals, hypothyroidism data from 53,423 individuals, and hyperthyroidism data from 51,823 subjects (11). The UK Biobank (UKB) provided summary statistical data for lipids (LDL-C, apolipoprotein B, and TG) (12). The sample size for LDL-C, apolipoprotein B, and TG was 440,546, 439,214, and 441,015, respectively. The UKB data came from a prospective cohort study that enrolled over 500,000 males and females (40–69 years old at baseline) between 2006 and 2010 (13). FinnGen Biobank of European ancestry provided the GWAS associated with cholelithiasis (19,023 cases and 195,144 controls). FinnGen is a large public-private partnership that aims to collect and analyze genomic and health data from 500,000 participants in Finnish biobanks. Table 1 contains detailed information.

## Selection of genetic instrumental variables

We identified single-nucleotide polymorphisms (SNPs) with genome-wide significance ($P< 5 \times 10^{-8}$), linkage disequilibrium (LD), and an $r^{2<}$ 0.001 threshold within a 10,000 kb window (14). We used PhenoScanner, a genotype-to-phenotype cross-reference (www.phenoscanner.medschl.cam.ac.uk), to look for secondary phenotypes associated with the selected instruments. In the present study, when cholelithiasis was identified as the outcome, blood glucose, BMI, and cholecystitis were identified as confounding factors. The palindromic variants were removed for incompatible alleles. When the SNPs were unavailable in the outcomes GWAS datasets, proxy SNPs were used. The final IVs for the subsequent MR study consisted of the strictly chosen SNPs. F-statistic was calculated to assess the strength of the selected SNPs

TABLE 1 Details of GWAS included in MR analyses.

| Traits | Consortia | Ethnicity | Sample size |
|---|---|---|---|
| Cholelithiasis (n, %) | FinnGen Biobank | European | 214,167 |
| FT4 (mmol/L) | The ThyroidOmics Consortium | European | 49,269 |
| TSH (mmol/L) | The ThyroidOmics Consortium | European | 54,288 |
| Hyperthyroidism (n, %) | The ThyroidOmics Consortium | European | 51,823 |
| Hypothyroidism (n, %) | The ThyroidOmics Consortium | European | 53,423 |
| LDL-C (mmol/L) | UK Biobank | European | 440,546 |
| Apolipoprotein B (mmol/L) | UK Biobank | European | 439,214 |
| Triglyceride (mmol/L) | UK Biobank | European | 441,016 |

according to the following equation:

$$F = \frac{R^2(N-1-K)}{(1-R^2)K}$$

Where $R^2$ is the portion of exposure variance explained by the IVs, N is the sample size, and K is the number of IVs. F-statistic ≧ 10 indicates no strong evidence of weak instrument bias.

## Replicative analysis

The Global Lipids Genetics Consortium (GLGC) was the source for the summary statistics of LDL-C and apolipoprotein B, while cholelithiasis was procured from the UK Biobank to serve as a replicative analysis.

## The proportion of mediation effects

The total effects of exposure on an outcome can be divided into indirect and direct effects (15). After adjusting for LDL-C, apolipoprotein B, and TG, MR revealed direct effects of thyroid function on cholelithiasis risk. The product method was used to calculate the indirect effects of lipid traits by multiplying the effects of thyroid function on lipid traits and the effects of lipid traits on cholelithiasis (16). The following equation was used to calculate the proportion of the mediation effects (17):

$$E(\%) = \frac{\sum_{k=1}^{K} \beta_1 * \beta_{2k}}{\sum_{k=1}^{K} \beta_3 + \beta_1 * \beta_{2k}}$$

Where β1 represents the MR effects of thyroid function on mediator k by two-step MR, β2 represents the MR effects of mediator k on cholelithiasis risk by two-step MR, and β3 represents the MR effects of thyroid function on cholelithiasis risk by two-sample MR.

## Statistical analysis

For MR analysis, five different methods [inverse-variance weighted (IVW), weighted median, MR Egger, maximum likelihood, and MR-robust adjusted profile score (MR-RAPS)] were used. To estimate the causal effects, the IVW method combines the wald ratios of the causal effects of each SNP. Each IV in this method must fulfill the three MR assumptions, or the derived estimates may be biased in the case of horizontal pleiotropy (18, 19). Compared with other MR methods, the maximum likelihood method provides an estimator with the lowest standard error under almost all conditions (20). MR-RAPS has been performed to model the random-effects distribution of pleiotropic genetic variation effects (21).

The MR-Egger analysis was used to evaluate the potential pleiotropic effects of SNP. The intercept *P*-value indicated if horizontal pleiotropy interfered with the MR estimates in MR Egger analysis. There was significant pleiotropy if the intercept *P*-value< 0.05. Cochrane's Q value was utilized to evaluate the heterogeneity among SNPs in IVW estimates. By excluding each SNP from the analysis, leave-one-out sensitivity analyses can be used to decide whether a single SNP has a significant effect on the overall result. The outlier variants were determined *via* the MR pleiotropy residual sum and outlier test (MR-PRESSO). The funnel plot was used to demonstrate the symmetrical distribution of the selected SNPs. A *P*-value< 0.05 was regarded as significant.

The statistical analyses were conducted utilizing the TwoSampleMR R package (version 0.5.5), MR-RAPS (version 1.0), and MR-PRESSO (version 1.0) with R software 4.1.2.

# Results

## MR analysis between thyroid function and cholelithiasis

Finally, 12 SNPs were used as IVs for FT4, 36 SNPs for TSH, 6 SNPs for hyperthyroidism, and 6 SNPs for hypothyroidism. Supplementary Tables 1–4 provided detailed information. In our study, F-statistic for each instrument-exposure association ranged from 19.867 to 37.321, demonstrating the less possibility of weak instrumental variable bias in the final results (Supplementary Table 11). The IVW method revealed that FT4 levels were correlated with an elevated risk of cholelithiasis (OR: 1.149, 95% CI: 1.082–1.283, $P = 0.014$, Figure 1). Weighted median (OR: 1.152, 95% CI: 1.002–1.326, $P = 0.046$), maximum likelihood (OR: 1.152, 95% CI: 1.042–1.274, $P = 0.006$), and MR-RAPS (OR: 1.090, 95% CI: 1.020–1.161, $P = 0.018$) indicated consistent results, while MR Egger illustrated negative results. In addition, the IVW method revealed that TSH, hyperthyroidism, and hypothyroidism were not related to the risk of cholelithiasis (Figure 1). Similar results are obtained through the implementation of alternative methodologies and the use of replicative analyses, as demonstrated by the findings presented in Supplementary Table 12.

The MR-PRESSO method also revealed that high FT4 levels were correlated with an elevated risk of cholelithiasis, and no outlier SNPs were recognized (Table 2). The present MR analysis revealed no horizontal pleiotropy and heterogeneity (Supplementary Table 10). Moreover, the stability of the results was indicated by the symmetrical shape of the funnel plot, as presented in Supplementary Figure 1. The findings of additional analyses related to thyroid function using MR are summarized in Table 2; Supplementary Table 10. Furthermore, the use of the leave-one-out sensitivity analysis demonstrated the robustness of the results, as depicted in Supplementary Figure 6.

## MR analysis between lipids and cholelithiasis

Finally, 155 SNPs were used as IVs for LDL-C, 277 SNPs for TG, and 179 SNPs for apolipoprotein B. Supplementary Tables 5–7 contain detailed information. The IVW method demonstrated a significant correlation between LDL-C and an elevated risk of

10.3389/fendo.2023.1166740

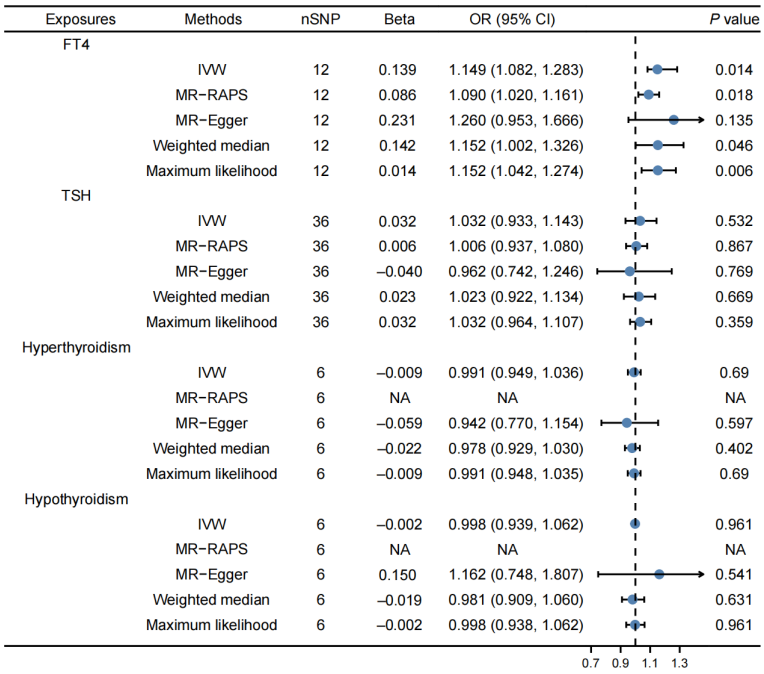



FIGURE 1
Effects of thyroid function on cholelithiasis. FT4 levels were correlated with an elevated risk of cholelithiasis, while TSH, hyperthyroidism, and hypothyroidism were not related to the risk of cholelithiasis.

cholelithiasis (OR: 1.354, 95% CI: 1.060–1.731, $P = 0.016$, Figure 2). The MR-PRESSO method also revealed that LDL-C was linked to an elevated risk of cholelithiasis (Table 2). The result was consistent after the outliners were removed. In this MR study, no horizontal pleiotropy was observed, but there was heterogeneity (Supplementary Table 10). The IVW method revealed that apolipoprotein B was correlated with an elevated risk of cholelithiasis (OR: 1.255, 95% CI: 1.027–1.535, $P = 0.027$, Figure 2). However, there was no correlation between TG and cholelithiasis risk. Moreover, the results of the study were observed to be stable based on the symmetrical funnel plots (Supplementary Figures 2, 3) and the leave-one-out method (Supplementary Figures 7, 8). Additionally, the replicative analysis further confirmed the consistency of the findings (Supplementary Table 12).

## MR analysis between thyroid function and lipids

Supplementary Tables 8, 9 contain detailed information on IVs. The IVW method revealed that FT4 levels were correlated with an elevated risk of LDL-C (OR: 1.084, 95% CI: 1.018–1.153, $P = 0.012$, Figure 3) and apolipoprotein B (OR: 1.087, 95% CI: 1.019–1.159, $P = 0.015$, Figure 4). The MR-PRESSO method produced consistent

TABLE 2 MR-PRESSO estimates between exposures and outcomes.

| Exposures | Outcomes | Raw estimates | | | Outlines corrected estimates | | |
|---|---|---|---|---|---|---|---|
| | | N | Beta | P value | N | Beta | P value |
| FT4 | Cholelithiasis | 12 | 0.138 | 0.017 | 12 | NA | NA |
| TSH | Cholelithiasis | 36 | 0.032 | 0.542 | 35 | 0.010 | 0.772 |
| Hyperthyroidism | Cholelithiasis | 6 | 0.002 | 0.903 | 6 | NA | NA |
| Hypothyroidism | Cholelithiasis | 6 | −0.002 | 0.950 | 6 | NA | NA |
| LDL-C | Cholelithiasis | 155 | 0.167 | 0.021 | 149 | 0.185 | 0.011 |
| Triglyceride | Cholelithiasis | 277 | 0.013 | 0.795 | 260 | 0.035 | 0.358 |
| Apolipoprotein B | Cholelithiasis | 179 | 0.299 | 0.025 | 163 | 0.143 | <0.001 |
| FT4 | LDL-C | 12 | 0.721 | 0.014 | 12 | NA | NA |
| FT4 | Apolipoprotein B | 12 | 0.071 | 0.003 | 12 | NA | NA |

10.3389/fendo.2023.1166740

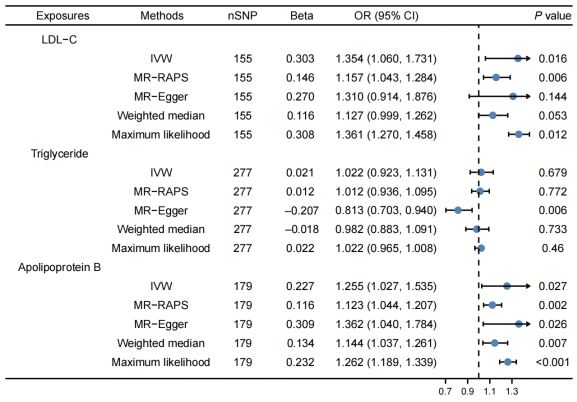



FIGURE 2
Effects of lipids on cholelithiasis. LDL-C and apolipoprotein B levels were correlated with an elevated risk of cholelithiasis, while triglyceride was not related to the risk of cholelithiasis.

results, with no outlier SNPs identified (Table 2). In the present MR analysis, there was no horizontal pleiotropy, but there was heterogeneity (Supplementary Table 10).

In addition, the symmetrical nature of the funnel plots provides evidence for the stability of the findings as demonstrated in Supplementary Figures 4, 5. The leave-one-out approach further reinforces the reliability of the results, as evidenced by the consistency of the outcomes shown in Supplementary Figures 9, 10. Additionally, the replicative analysis, as illustrated in Supplementary Table 12, yields comparable findings, further consolidating the robustness of the results.

## The proportion of the mediatory effects of apolipoprotein B and LDL-C

The present analysis revealed that apolipoprotein B and LDL-C mediated the effects of thyroid function on cholelithiasis risk. LDL-C and apolipoprotein B had 17.4% and 13.6% of the mediatory effects, respectively.

# Discussion

In the current study, Mendelian randomization was employed to investigate the causal associations between thyroid function and lipids

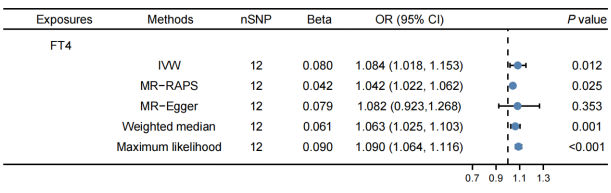



FIGURE 3
Effects of FT4 on LDL-C. FT4 levels were correlated with an elevated risk of LDL-C.

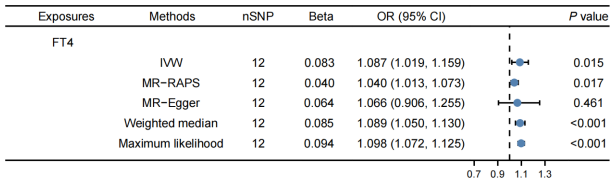



FIGURE 4
Effects of FT4 on apolipoprotein B. FT4 levels were correlated with an elevated risk of apolipoprotein B.

in relation to cholelithiasis. The results indicate a positive association between FT4 levels and cholelithiasis risk. Additionally, elevated levels of LDL-C and apolipoprotein B were also significantly associated with an increased risk of cholelithiasis. MR analysis revealed that LDL-C and apolipoprotein B accounted for 17.4% and 13.5% of the mediatory effects, respectively. These findings provide important insights into the role of thyroid function and lipid metabolism traits in the pathogenesis of cholelithiasis, which may have implications for developing preventive and therapeutic strategies for this disease.

In prior studies, observational analyses have predominantly established a correlation between thyroid function and the presence of cholelithiasis. As reported by J. Inkinen in previous research, a significant association was detected between the occurrence of common bile duct stones and pre-existing hypothyroidism (22). A study was conducted on a sample of 3,749 individuals aged between 20 and 79 years, which revealed a statistically significant and independent association between increased serum thyrotropin (TSH) levels and cholelithiasis in males (5). However, no such relation was identified in the female population. In contrast, some studies have suggested a greater susceptibility of women to both cholelithiasis and thyroid disorders (23). Animal models have also indicated that hyperthyroidism may be a predisposing factor for cholelithiasis. Furthermore, a Chinese researcher has proposed that dysfunction of the thyroid, including both hyperthyroidism and hypothyroidism, can promote the formation of gallstones through various pathways (24). However, the existence of residual confounding, reverse causation, or both, has been raised as potential explanations for the observed associations. In the present study, we found that neither TSH levels, hyperthyroidism, nor hypothyroidism were significantly related to the risk of cholelithiasis. Notably, this is the first MR analysis to demonstrate that elevated FT4 levels confer an increased risk of cholelithiasis.

Although various investigations have been performed to clarify the association between lipids and cholelithiasis, the findings remain controversial. Several convincing studies revealed a positive association between high cholesterol levels and cholelithiasis (2). In a case-control study, Fu et al. found that increased serum LDL-C and apolipoprotein B were an index of cholesterol stones (25). However, Tang et al. identified that apolipoprotein A, B, high serum HDL, and lower LDL are risk factors for cholelithiasis in a study with 109 sample sizes. The studies mentioned above had limited sample sizes. LDL-C and apolipoprotein B were also correlated with an elevated risk of cholelithiasis in our large-scale MR study.

The underlying mechanisms of thyroid function and cholelithiasis remain unknown. Cholelithiasis can be caused by various factors, considering how thyroid hormones affect the balance of cholesterol, the amount of bile produced, biliary secretion, and motility of the gallbladder (26). Thyroid hormones have been shown to influence enterohepatic circulation and detoxification (27–29), as well as nuclear receptor-mediated LITH gene expression (2, 30–33). Thyroid dysfunction and lipid homeostasis were two other underlying mechanisms. We identified that LDL-C and apolipoprotein B mediate the effects of thyroid function on the cholelithiasis risk. LDL-C and apolipoprotein B had 17.4% and 13.5% of the mediatory effects, respectively. Reduced bile acid production by the conventional (CYP7A1 and CYP8B1) and alternative (CYP27A1) pathways were found in an *in vitro* research on primary human hepatocytes (34). It revealed that hyperthyroidism could cause a disturbance in the composition of bile through dysregulation of lipid homeostasis.

The present study has several strengths. First, it was the first MR to examine how lipids and thyroid function affect cholelithiasis using large-scale GWAS from UKB, Finngen Biobank, and the ThyroidOmics Consortium. Second, because the IVs we selected were located on a different chromosome, any possible gene-gene interaction may have few effects on the predicted value (35). Third, we used several stable methods to obtain the MR effects, such as MR-PRESSO and MR-RAPS. Furthermore, we assessed horizontal pleiotropy. Finally, using the two-step MR analysis, we identified that LDL-C and apolipoprotein B acted as mediators of the causal pathway from FT4 levels to cholelithiasis risk.

The present study has some limitations. First, there was potential heterogeneity due to differences in health status, age, or gender. Second, all participants were of European ancestry, which may restrict the applicability of the results to other races and ethnicities. Third, any potential nonlinear relationships or stratification effects result from the GWAS data. Fourth, TSH and FT4 levels were obtained from different cohorts which may have an impact on the reliability of the results. Finally, because confounding and mediation cannot be statistically differentiated, mediation analysis was critically dependent on the accurate characterization of the causal relationships (36).

## Conclusion

In conclusion, we demonstrated that FT4, LDL-C, and apolipoprotein B had significant causal effects on cholelithiasis, with evidence that the LDL-C and apolipoprotein B mediated the effects of FT4 on cholelithiasis risk. Patients with high FT4 levels should be given special attention because they may delay or limit the long-term impact on cholelithiasis risk.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

## Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

## Funding

This research was funded by the Jilin Science and Technology Development Program (CN) (No. 20200201426JC).

## Conflict of interest

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.

## Publisher's note



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1166740/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Funnel plot of the association between FT4 and cholelithiasis.

SUPPLEMENTARY FIGURE 2
Funnel plot of the association between LDL-C and cholelithiasis.

SUPPLEMENTARY FIGURE 3
Funnel plot of the association between apolipoprotein B and cholelithiasis.

SUPPLEMENTARY FIGURE 4
Funnel plot of the association between FT4 and LDL-C.

SUPPLEMENTARY FIGURE 5
Funnel plot of the association between FT4 and apolipoprotein B.

SUPPLEMENTARY FIGURE 6
Leave-one-out sensitivity analysis of the association between FT4 and cholelithiasis.

SUPPLEMENTARY FIGURE 7
Leave-one-out sensitivity analysis of the association between LDL-C and cholelithiasis.

SUPPLEMENTARY FIGURE 8
Leave-one-out sensitivity analysis of the association between apolipoprotein B and cholelithiasis.

SUPPLEMENTARY FIGURE 9
Leave-one-out sensitivity analysis of the association between FT4 and LDL-C.

SUPPLEMENTARY FIGURE 10
Leave-one-out sensitivity analysis of the association between FT4 and apolipoprotein B.

# References

1. Zdanowicz K, Daniluk J, Lebensztejn DM, Daniluk U. The etiology of cholelithiasis in children and adolescents-a literature review. *Int J Mol Sci* (2022) 23 (21). doi: 10.3390/ijms232113376

2. Pak M, Lindseth G. Risk factors for cholelithiasis. *Gastroenterol Nurs* (2016) 39 (4):297–309. doi: 10.1097/SGA.0000000000000235

3. Di Ciaula A, Wang DQ, Portincasa P. An update on the pathogenesis of cholesterol gallstone disease. *Curr Opin Gastroenterol* (2018) 34(2):71–80. doi: 10.1097/MOG.0000000000000423

4. Everhart JE, Ruhl CE. Burden of digestive diseases in the united states part III: Liver, biliary tract, and pancreas. *Gastroenterology* (2009) 136(4):1134–44. doi: 10.1053/j.gastro.2009.02.038

5. Völzke H, Robinson DM, John U. Association between thyroid function and gallstone disease. *World J Gastroenterol* (2005) 11(35):5530–4. doi: 10.3748/wjg.v11.i35.5530

6. Kulkarni V, Ramteke H, Lamture Y, Gharde P. A review of synchronous findings of hypothyroidism and cholelithiasis. *Cureus* (2022) 14(10):e30316. doi: 10.7759/cureus.30316

7. Andreotti G, Chen J, Gao YT, Rashid A, Chang SC, Shen MC, et al. Serum lipid levels and the risk of biliary tract cancers and biliary stones: A population-based study in China. *Int J Cancer* (2008) 122(10):2322–9. doi: 10.1002/ijc.23307

8. Du FM, Kuang HY, Duan BH, Liu DN, Yu XY. Effects of thyroid hormone and depression on common components of central obesity. *J Int Med Res* (2019) 47 (7):3040–9. doi: 10.1177/0300060519851624

9. Emdin CA, Khera AV, Kathiresan S. Mendelian randomization. *Jama* (2017) 318 (19):1925–6. doi: 10.1001/jama.2017.17219

10. Sekula P, Del Greco MF, Pattaro C, Köttgen A. Mendelian randomization as an approach to assess causality using observational data. *J Am Soc Nephrol* (2016) 27 (11):3253–65. doi: 10.1681/ASN.2016010098

11. Teumer A, Chaker L, Groeneweg S, Li Y, Di Munno C, Barbieri C, et al. Genome-wide analyses identify a role for SLC17A4 and AADAT in thyroid hormone regulation. *Nat Commun* (2018) 9(1):4455. doi: 10.1038/s41467-018-06356-1

12. Richardson TG, Sanderson E, Palmer TM, Ala-Korpela M, Ference BA, Davey Smith G, et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable mendelian randomisation analysis. *PloS Med* (2020) 17(3):e1003062. doi: 10.1371/journal.pmed.1003062

13. Collins R. What makes UK biobank special? *Lancet* (2012) 379(9822):1173–4. doi: 10.1016/S0140-6736(12)60404-8

14. Park S, Lee S, Kim Y, Lee Y, Kang MW, Kim K, et al. Atrial fibrillation and kidney function: a bidirectional mendelian randomization study. *Eur Heart J* (2021) 42 (29):2816–23. doi: 10.1101/2020.07.31.20166207

15. Carter AR, Sanderson E, Hammerton G, Richmond RC, Davey Smith G, Heron J, et al. Mendelian randomisation for mediation analysis: Current methods and challenges for implementation. *Eur J Epidemiol* (2021) 36(5):465–78. doi: 10.1007/s10654-021-00757-1

16. Zhao SS, Holmes MV, Zheng J, Sanderson E, Carter AR. The impact of education inequality on rheumatoid arthritis risk is mediated by smoking and body mass index: Mendelian randomization study. *Rheumatology* (2022) 61(5):2167–75. doi: 10.1093/rheumatology/keab654

17. Xu L, Borges MC, Hemani G, Lawlor DA. The role of glycaemic and lipid risk factors in mediating the effect of BMI on coronary heart disease: A two-step, two-sample mendelian randomisation study. *Diabetologia* (2017) 60(11):2210–20. doi: 10.1007/s00125-017-4396-y

18. Pierce BL, Burgess S. Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *Am J Epidemiol* (2013) 178 (7):1177–84. doi: 10.1093/aje/kwt084

19. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol* (2016) 40(4):304–14. doi: 10.1002/gepi.21965

20. Milligan BG. Maximum-likelihood estimation of relatedness. *Genetics* (2003) 163(3):1153–67. doi: 10.1093/genetics/163.3.1153

21. Fu Y, Xu F, Jiang L, Miao Z, Liang X, Yang J, et al. Circulating vitamin c concentration and risk of cancers: A mendelian randomization study. *BMC Med* (2021) 19(1):171. doi: 10.1186/s12916-021-02041-1

22. Inkinen J, Sand J, Nordback I. Association between common bile duct stones and treated hypothyroidism. *Hepatogastroenterology* (2000) 47(34):919–21.

23. Bauer M, Glenn T, Pilhatsch M, Pfennig A, Whybrow PC. Gender differences in thyroid system function: Relevance to bipolar disorder and its treatment. *Bipolar Disord* (2014) 16(1):58–71. doi: 10.1111/bdi.12150

24. Wang Y, Yu X, Zhao QZ, Zheng S, Qing WJ, Miao CD, et al. Thyroid dysfunction, either hyper or hypothyroidism, promotes gallstone formation by different mechanisms. *J Zhejiang Univ Sci B* (2016) 17(7):515–25. doi: 10.1631/jzus.B1500210

25. Fu X, Gong K, Shen T, Shao X, Li G, Wang L, et al. Gallstones and their chemical types in relation to serum lipids and apolipoprotein levels. *Chin Med J* (1997) 110 (5):384–7.

26. Laukkarinen J, Sand J, Nordback I. The underlying mechanisms: How hypothyroidism affects the formation of common bile duct stones-a review. *HPB Surg* (2012) 2012:102825. doi: 10.1155/2012/102825

27. Chiang JYL, Ferrell JM. Bile acid metabolism in liver pathobiology. *Gene Expr* (2018) 18(2):71–87. doi: 10.3727/105221618X15156018385515

28. Xiao L, Pan G. An important intestinal transporter that regulates the enterohepatic circulation of bile acids and cholesterol homeostasis: The apical sodium-dependent bile acid transporter (SLC10A2/ASBT). *Clin Res Hepatol Gastroenterol* (2017) 41(5):509–15. doi: 10.1016/j.clinre.2017.02.001

29. Cai JS, Chen JH. The mechanism of enterohepatic circulation in the formation of gallstone disease. *J Membr Biol* (2014) 247(11):1067–82. doi: 10.1007/s00232-014-9715-3

30. Joshi AD, Andersson C, Buch S, Stender S, Noordam R, Weng LC, et al. Four susceptibility loci for gallstone disease identified in a meta-analysis of genome-wide association studies. *Gastroenterology* (2016) 151(2):351–363.e28. doi: 10.1053/j.gastro.2016.04.007

31. Ferkingstad E, Oddsson A, Gretarsdottir S, Benonisdottir S, Thorleifsson G, Deaton AM, et al. Genome-wide association meta-analysis yields 20 loci associated with gallstone disease. *Nat Commun* (2018) 9(1):5101. doi: 10.1038/s41467-018-07460-y

32. Weber SN, Bopp C, Krawczyk M, Lammert F. Genetics of gallstone disease revisited: updated inventory of human lithogenic genes. *Curr Opin Gastroenterol* (2019) 35(2):82–7. doi: 10.1097/MOG.0000000000000511

33. Jiang ZY, Parini P, Eggertsen G, Davis MA, Hu H, Suo GJ, et al. Increased expression of LXR alpha, ABCG5, ABCG8, and SR-BI in the liver from normolipidemic, nonobese Chinese gallstone patients. *J Lipid Res* (2008) 49(2):464–72. doi: 10.1194/jlr.M700295-JLR200

34. Ellis EC. Suppression of bile acid synthesis by thyroid hormone in primary human hepatocytes. *World J Gastroenterol* (2006) 12(29):4640–5. doi: 10.3748/wjg.v12.i29.4640

35. Wang T, Xu L. Circulating vitamin e levels and risk of coronary artery disease and myocardial infarction: A mendelian randomization study. *Nutrients* (2019) 11(9). doi: 10.3390/nu11092153

36. MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the mediation, confounding and suppression effect. *Prev Sci* (2000) 1(4):173–81. doi: 10.1023/A:1026595011371

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 09 May 2023
DOI 10.3389/fendo.2023.1180732



OPEN ACCESS

EDITED BY
Wenjie Shi,
Otto von Guericke University Magdeburg, Germany

REVIEWED BY
Guo Liu,
Shenzhen University, China
Yunwei Han,
The Affiliated Hospital of Southwest Medical University, China

*CORRESPONDENCE
Baoqiang Song
songbq2012@163.com
Jungang Ma
31403699@qq.com

†These authors have contributed equally to this work

RECEIVED 06 March 2023
ACCEPTED 03 April 2023
PUBLISHED 09 May 2023

CITATION
Cui Z, Liang Z, Song B, Zhu Y, Chen G, Gu Y, Liang B, Ma J and Song B (2023) Machine learning-based signature of necrosis-associated lncRNAs for prognostic and immunotherapy response prediction in cutaneous melanoma and tumor immune landscape characterization. *Front. Endocrinol.* 14:1180732. doi: 10.3389/fendo.2023.1180732



# Machine learning-based signature of necrosis-associated lncRNAs for prognostic and immunotherapy response prediction in cutaneous melanoma and tumor immune landscape characterization

Zhiwei Cui[1†], Zhen Liang[1†], Binyu Song[1†], Yuhan Zhu[1], Guo Chen[1], Yanan Gu[1], Baoyan Liang[1], Jungang Ma[2*] and Baoqiang Song[1*]

[1]Department of Plastic and Reconstructive Surgery, Xijing Hospital, Fourth Military Medical University, Xi'an, China, [2]Department of Cancer Center, Daping Hospital, Army Medical University, Chongqing, China

**Background:** Cutaneous melanoma (CM) is one of the malignant tumors with a relative high lethality. Necroptosis is a novel programmed cell death that participates in anti-tumor immunity and tumor prognosis. Necroptosis has been found to play an important role in tumors like CM. However, the necroptosis-associated lncRNAs' potential prognostic value in CM has not been identified.

**Methods:** The RNA sequencing data collected from The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression Project (GTEx) was utilized to identify differentially expressed genes in CM. By using the univariate Cox regression analysis and machine learning LASSO algorithm, a prognostic risk model had been built depending on 5 necroptosis-associated lncRNAs and was verified by internal validation. The performance of this prognostic model was assessed by the receiver operating characteristic curves. A nomogram was constructed and verified by calibration. Furthermore, we also performed sub-group K-M analysis to explore the 5 lncRNAs' expression in different clinical stages. Function enrichment had been analyzed by GSEA and ssGSEA. In addition, qRT-PCR was performed to verify the five lncRNAs' expression level in CM cell line (A2058 and A375) and normal keratinocyte cell line (HaCaT).

**Results:** We constructed a prognostic model based on five necroptosis-associated lncRNAs (AC245041.1, LINC00665, AC018553.1, LINC01871, and AC107464.3) and divided patients into high-risk group and low-risk group depending on risk scores. A predictive nomogram had been built to be a prognostic indicator to clinical factors. Functional enrichment analysis showed that immune functions had more relationship and immune checkpoints were

more activated in low-risk group than that in high-risk group. Thus, the low-risk group would have a more sensitive response to immunotherapy.

**Conclusion:** This risk score signature could be used to divide CM patients into low- and high-risk groups, and facilitate treatment strategy decision making that immunotherapy is more suitable for those in low-risk group, providing a new sight for CM prognostic evaluation.

KEYWORDS

**cutaneous melanoma (CM), necroptosis, long non-coding RNAs (lncRNAs), prognostic signature, tumor immune function**

# 1 Introduction

Cutaneous melanoma (CM) is considered to be a malignant tumor that develops from melanocytes. CM is the result of a genetic mutation caused by ultraviolet ray radiation (1). Though it is not a high-incidence disease, it has a relatively high lethality rate. Multiple studies have revealed that CM occurs for 4% of skin cancers but 75% of skin cancer-related mortality (2). CM is one of the most immunogenic carcinomas and hence has a substantial potential for a positive response to immunotherapy (3). However, due to the early metastases, selecting the proper treatment strategy is critical for CM. Therefore, early detection of CM and stratified risk assessment are essential for CM treatment (4).

Necroptosis is a type of controlled cell death that mimics both necrosis and apoptosis. Plasma membrane permeabilization occurs rapidly during necroptosis. In this process, the cell content is released and then subsequently exposed to a variety of cytokines, chemokines, and damage-associated molecular patterns. The immunogenic nature of necroptotic cancer cells and their capacity to effectively trigger anti-tumor immunity are both attributed to necroptosis, which is becoming increasingly recognized as being crucial in cancer (5, 6). For example, the decrease in the expression of necroptotic factors such as receptor-interacting protein kinase-3 leads to a worse prognosis in breast cancer (7, 8). Similarly, the decrease of necroptotic factors' expression, such as receptor-interacting protein kinase-3 and mixed lineage kinase domain-like protein, leads to reduced overall survival (OS) (9).

Long non-coding RNAs (lncRNAs) are those having transcripts that are 200 nucleotides or longer. Multiple studies have shown that lncRNAs are involved in cancer. lncRNA-Gm31932 uses the miR-344d-3-5p/Prc1 axis, for example, to induce cell cycle arrest and differentiation in melanoma (10). LncRNA TINCR suppresses melanoma cell proliferation and invasion by regulating the miR-424-5p/LATS1 axis, and it also upregulates apoptosis (11). Based on the important role of lncRNA in melanoma, several prognostic signatures have been established, like lncRNAs which are associated with autophagy (12), ferroptosis (2), pyroptosis (13), etc.

To create a unique predictive signature, this work analyzed the relationship between necroptosis-associated lncRNAs and clinicopathological features, and immune infiltration of individuals with CM. Internal verification and gene enrichment analysis (GSEA) were used to assess the robustness of the model as well as the potential mechanisms, respectively.

# 2 Methods

## 2.1 Data collection

The Cancer Genome Atlas (TCGA, n=471) was used to acquire RNA transcriptome datasets and associated clinical information of CM, and a synthetic data matrix concerning healthy skin data was obtained from Genotype-Tissue Expression Project (GTEx, n=234). The data were merged and processed using the R "Limma" tool. Furthermore, the lncRNA expression values and survival rates for 471 CM patients were determined.

## 2.2 Analysis of necroptosis-associated genes

The 201 necroptosis-associated genes were obtained from GeneCards (https://www.genecards.org/) and other published studies (Table S1). The "Limma" R-package was used to distinguish differentially expressed genes (DEGs) related to necroptosis in normal and CM tissues using a False Discovery Rate (FDR) of < 0.05 and a |log2 fold change (FC) >1| threshold (14). To determine which genes belong to both DEGs and necroptosis-associated, a Venn diagram was constructed. To visually represent the expression level of overlapping genes, a volcano image and a heatmap were created. The "ggplot2" package was utilized to conduct the analysis of the Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Ontology (GO).

10.3389/fendo.2023.1180732

## 2.3 Necroptosis-associated lncRNAs signature associated with prognostic significance

The screening criteria for identifying the necroptosis-associated lncRNAs in CM samples with expression values were the correlation coefficients $|R| > 0.3$ and $p < 0.001$. Furthermore, to determine the necroptosis-associated lncRNA predictive signature and to assess the relationships between overall survival (OS) and necroptosis-associated lncRNAs in CM, the "survival" R package was used to perform a univariate Cox regression (uni-Cox) analysis at a significance level of $p < 0.001$. Then, the R package "caret" was employed to randomly classify the CM samples into the training and testing cohorts.

To identify the most important necroptosis-associated lncRNA with CM patients, we performed LASSO-penalized Cox regression analysis by the "glmnet" R package. The selection of variables in the Cox model was performed using the lasso method, and a risk signature was generated using the "survminer" package in R. The following calculations were utilized to determine the risk score: Risk score = sum (each lncRNA's expression × corresponding coefficient). Low-risk (LR) and high-risk (HR) groups of CM patients were defined using the median risk score in both the training and testing datasets. The clinical information for the entire set was shown in Table 1. Potential lncRNA expression in normal and CM tissues was plotted using a heat map generated using the "pheatmap" R package.

The relationship between candidate lncRNA and mRNA was visualized by the Cytoscape diagram. Furthermore, "gg alluvial" in the R package was employed to visualize the distribution of the 5 candidate genes in LR and HR groups. The Kaplan-Meier (K-M) survival analysis and the correlation between risk score and survival time were performed using the "survival" and "survminer" packages in R. 1-year, 3-year, and 5-year ROC analyses were conducted using the "timeROC" R-package.

The "survival" R-package was used to create a prediction model, which included univariate and multivariate independent prognostic studies to determine the correlation between clinical features, risk score, and patient OS. A heatmap was made to illustrate the distribution of clinical features and potential lncRNAs in LR and HR groups. ROC analysis on risk scores and clinical features was done with the "survival ROC" package.

## 2.4 Nomogram and calibration

Using the risk score, age, and T, N stage, the R-package "rms" was used to create a nomogram for 1-year, 3-year, and 5-year OS. Using a calibration chart and ROC curves, we analyzed the nomogram's prognostic accuracy.

TABLE 1 Different clinicopathological features of the necrosis-associated risk subgroups in TCGA-SKCM.

| Clinical variables | Total (N=360) | Risk-group | | *p* value |
|---|---|---|---|---|
| | | high (n=178) | low (n=182) | |
| **Gender** | | | | 0.177 |
| Female | 138 | 62 | 76 | |
| Male | 222 | 116 | 106 | |
| **Age** | | | | 0.962 |
| <65 | 231 | 114 | 117 | |
| ≥65 | 129 | 64 | 65 | |
| **Stage** | | | | 0.200 |
| High stage | 168 | 77 | 91 | |
| Low stage | 192 | 101 | 91 | |
| **T** | | | | 0.025 |
| T1-T2 | 134 | 56 | 78 | |
| T3-T4 | 226 | 122 | 104 | |
| **N** | | | | 0.185 |
| N1-N2 | 313 | 159 | 154 | |
| N3-N4 | 47 | 19 | 28 | |
| **M** | | | | 0.840 |
| M0 | 343 | 170 | 173 | |
| M1 | 17 | 8 | 9 | |

## 2.5 Function enrichment analyses

The predominant route genes were analyzed using GSEA. GSEA 4.1.0 was employed to conduct the analysis. The cutoffs for statistical significance were minimal (FDR<0.25 and $p<0.05$). Here, the GSVA software and ssGSEA were used to determine the infiltration scores of 16 immunological cells and the 13 immune-related pathway activities.

## 2.6 Cell culture

Human normal keratinocyte cell line (HaCaT) and human melanoma cell lines (A2058 and A375) were purchased from American Type Culture Collection (ATCC). HaCaT and A375 cell lines were cultured in DMEM (Dulbecco's Modified Eagle Medium) (Gibco, Grand Island, NY) and A2058 cell line was cultured in DMEM/F-12 (Gibco, Grand Island, NY) and both were added 10% fetal bovine serum (Yeasen, Shanghai, China) at 37°C in an incubator with 5% CO2.

## 2.7 RNA extraction and quantitative real-time polymerase chain reaction

Total RNA of the three cell lines was extracted with TRIzol Reagent (Takara, Kusatsu, Japan). cDNA was synthesized with Hifair® III 1st Strand cDNA Synthesis SuperMix (11141ES60, Yeasen) according to standard protocol. The cDNA was used as a template and lncRNA expression was quantified using SYBR Green Master Mix (11184ES08, Yeasen). The primer pairs were synthesized by Tsingke Biotechnology (Beijing, China), and the primer pairs are listed in Table S2. All samples were normalized to GAPDH, and the $2^{-\Delta\Delta Ct}$ method was used to evaluate relative expression levels.

## 2.8 Statistical analysis

R 4.0.2 version and Prism 5 were used to conduct all statistical analyses. Analysis of continuous data was conducted using the Wilcoxon test, while analysis of categorical data was conducted using the Chi-square or Fisher tests. The log-rank test and the K-M technique were used to compare the OS rates of patients in the HR and LR groups. The collected findings were also considered significant at the $p < 0.05$ level.

# 3 Results

## 3.1 Necroptosis-associated lncRNAs in CM patients and functional analyses

Figure 1 shows how the study progressed. We got 234 normal samples and 471 CM samples from TCGA and GTEx. Finally, 79 predictive necroptosis-associated DEGs (correlation coefficients>0.3 and $p$<0.001) were identified by comparing the expression of 201 necroptosis-associated genes to 16,977 DEGs (|Log 2 FC|>1 and $p$<0.05) between normal and CM samples (Figure 2A). 52 of them had higher regulation, while the remaining 14 had lower regulation (Figure 2B). The heatmap was used to show the amount of expression of the 79 overlapped genes in the normal and CM tissue (Figure 2C). The 79 overlapping genes were enriched in processes related to necroptosis, systemic lupus erythematosus, and NOD-like receptor signaling pathway, according to KEGG enrichment analysis (Figure 2D). The 79 overlapping genes were shown to be enriched in biological processes related to necroptotic and protein secretion, including the necroptotic process, programmed necrotic cell death, and control of protein and peptide secretion, according to GO enrichment analysis (Figure 2E).

## 3.2 Development of the necroptosis-associated lncRNA predictive signature

We identified 880 necroptosis-associated lncRNAs (Table S3). For the subsequent study, the expression levels of 880 lncRNAs associated with necroptosis and the clinical details of 454 melanoma samples were used. Using uni-Cox regression analysis, we identified 34 lncRNAs associated with necroptosis that were statistically significant predictors of OS ($p$<0.001) (Figure 3A). After running the Lasso regression on these lncRNAs, we found that 5 of them were associated with necroptosis in melanoma when the first-rank value of Log(λ) was the minimum likelihood of deviance, hence preventing the prognostic signature from overfitting (Figures 3B, C). The model contains the lncRNAs: AC245041.1, LINC00665, AC018553.1, LINC01871, and AC107464.3. The formula for the risk score is as follows: risk score = (-0.365383*expression of AC107464.3) + (0.26265*expression of LINC00665) + (0.56689*expression of AC245411.1) + (-0.45893*expression of LINC01871) + (0.48615*expression of AC018553.1). A heatmap depicts the differential expression of the five lncRNAs between normal and CM tissue (Figure 3D).

Figure 4A depicts a network of the prognostic lncRNAs and the mRNAs they are linked to. Furthermore, this study identified AC107464.3 and LINC01871 as protective genes, whereas others were identified as risk factors (Figure 4B). It was also discovered

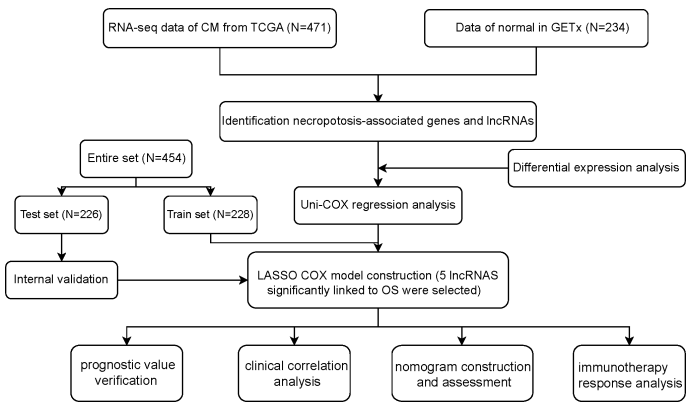



FIGURE 1
Flowchart of the study.

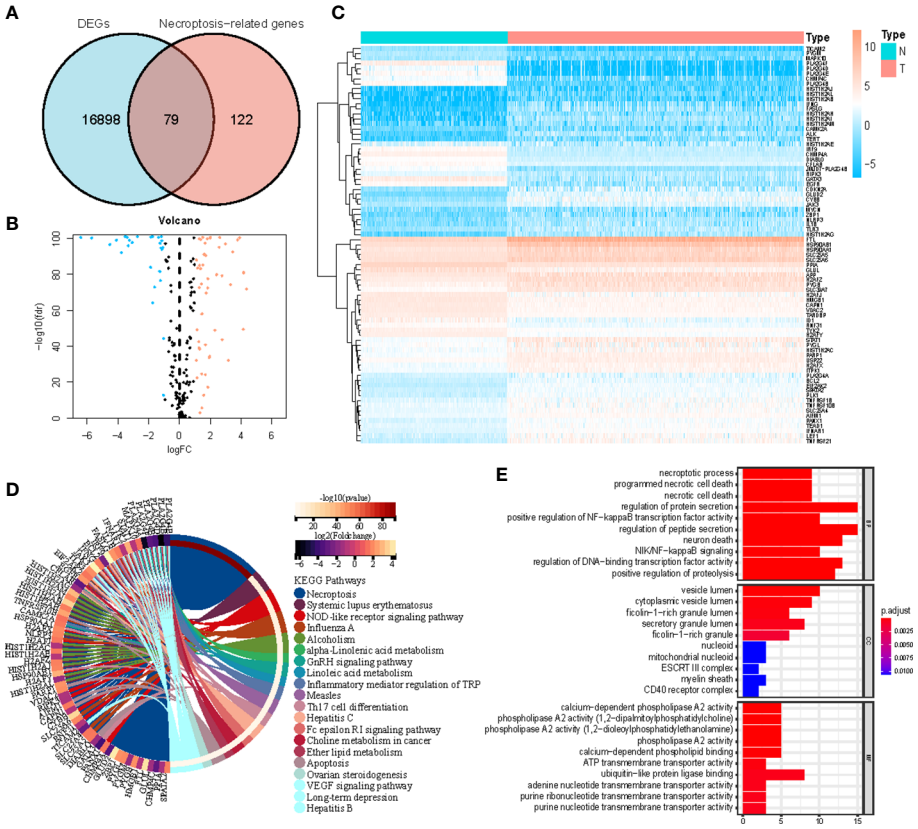



FIGURE 2
**(A)** Veen diagram of candidate necroptosis-associated differentially expressed genes (DEGs); **(B)** Volcano plot of 79 necroptosis-associated DEGs; **(C)** Heatmap visualizing the expression of necroptosis-associated DEGs; **(D, E)** KEGG and GO functional enrichment analysis of necroptosis-associated genes.

that the LR and HR groups express the five lncRNAs in distinct ways (Figure 4C). LR and HR groups of CM patients were created using median risk scores (Figures 4E, F). When comparing the HR group to the LR group, Figure 4D shows that the HR group's OS is significantly shorter ($p$<0.001). ROC study indicated that the risk signature had reasonable predictive accuracy at 1-year (ROC = 0.758), 2-year (ROC = 0.701), and 3-year (ROC = 0.727) (Figure 4G).

## 3.3 Independent prognostic factors and clinical correlation analysis of NALncSig

Univariate and multivariate Cox regression (multi-COX) analyses show that the newly identified risk signature is an independent prognostic factor for CM patients. The hazard ratio (HR) of the risk score and 95% confidence interval (CI) were 1.354 and 1.240-1.478 ($p$<0.001) in uni-Cox regression, respectively, 1.385 and 1.260-1.523 ($p$<0.001) in multi-Cox regression (Figures 5A, B). In addition, we found the other three independent prognostic parameters, age (1.020 and 1.009-1.030; $p$<0.001), T stage (1.309 and 1.115-1.536; $p$<0.001), and N stage (1.319 and 1.052-1.655; $p$=0.016) (Figure 5B). Our NALncSig was substantially correlated with age, T stage, and the N stage, according to the heatmap of clinical characteristics and risk groupings (Figure 5C). Additionally, various melanoma prognostic indicators were chosen for comparison to determine whether the NALncSig had the capacity for consistent and reliable performance. The ROC curve of the risk score and the clinicopathological criteria was performed. The area under the ROC curve (AUC) for our NALncSig curve was 0.707, which was significantly higher than the that for age (AUC = 0.618), gender (AUC = 0.486), stage (AUC = 0.592), T stage (AUC = 0.683), M stage (AUC = 0.508), and N stage (AUC = 0.571) (Figure 5D).

## 3.4 Construction of nomogram

We constructed a nomogram using four independent prognostic factors risk score, age, T, and N ($p$<0.05 in multi-Cox) to predict the 1-, 3-, and 5-year OS incidences of CM patients

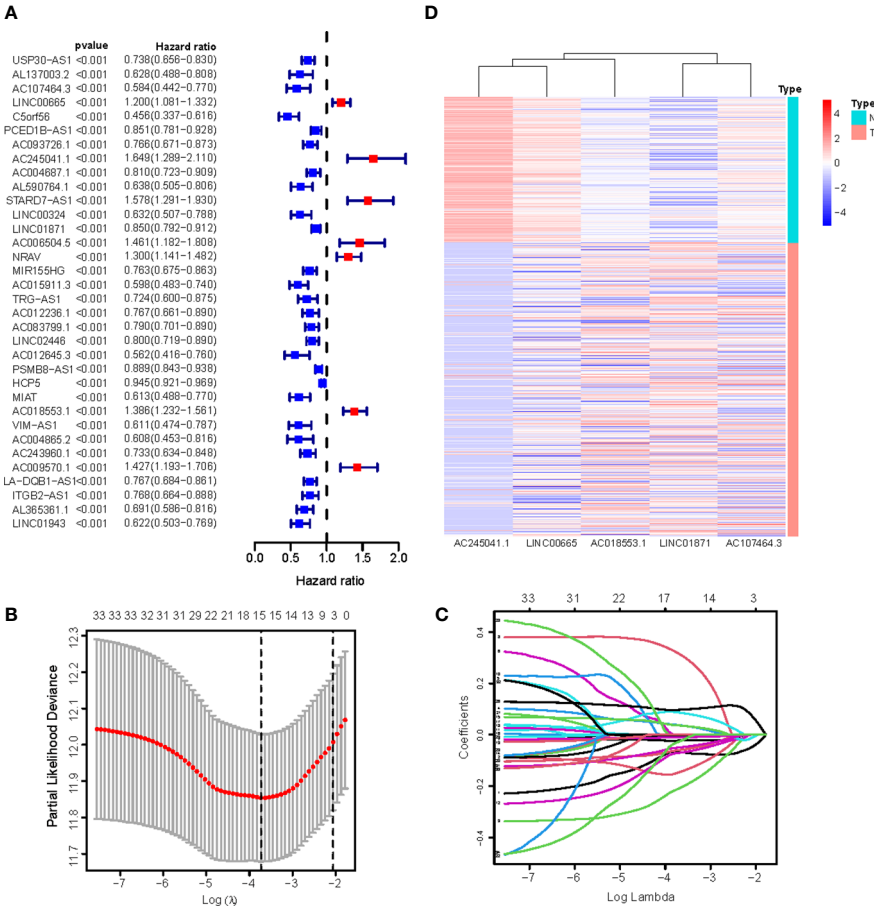



FIGURE 3
**(A)** 34 prognostic necroptosis-associated lncRNAs extracted by univariate Cox regression analysis; **(B)** The 10-fold cross-validation for variable selection in the LASSO model; **(C)** The LASSO coefficient profile of 16 necroptosis-associated lncRNAs; **(D)** Heatmap visualizing the expression of differentially expressed necroptosis-associated lncRNAs.

(Figure 6A). Further confirmation of the nomogram's accuracy in predicting these outcomes was obtained using 1-, 3-, and 5-year calibration plots (Figures 6B–D).

## 3.5 Relationship between the NALncSig and CM patient's prognosis in different clinicopathological variables

CM patients were divided into groups according to conventional clinicopathologic parameters such as age, gender, grade, and TNM stage. Except for patients with metastases (M1), the OS of patients in the HR groups was much lower than that of patients in the LR groups, demonstrating that the predictive signature can reliably predict the prognosis of CM patients (Figures 7A-M). Additionally, clinical data revealed that the expression of AC245041.1 had a difference between stage I and II, stage II and III (Figure 8B), and the expression of LINC01871 had a difference between stage I and II (Figure 8E), while the relationship between the expression of LNC00665, AC107464.3 and AC018553.1 and pathologic stage was not statistically significant (Figures 8A, C, D). These results indicate that various clinical variables have a certain effect on the expression and risk score of the necroptosis-associated lncRNA. Because of this, the pathological status of patients ought to be taken into consideration while developing the risk model that is used to evaluate the prognosis of patients.

## 3.6 Internal validation of NALncSig

To evaluate the feasibility of utilizing the NALncSig to predict OS in the entire TCGA dataset, we randomly divided the entire cohort into the training (N=228) and testing (N=226) cohorts using the same algorithm and regression coefficient (β). As expected, the K-M survival analysis confirmed what was observed in the entire dataset: CM patients in the LR group had longer OS than those in the HR group in both the training and testing cohorts ($p$<0.0001)

10.3389/fendo.2023.1180732

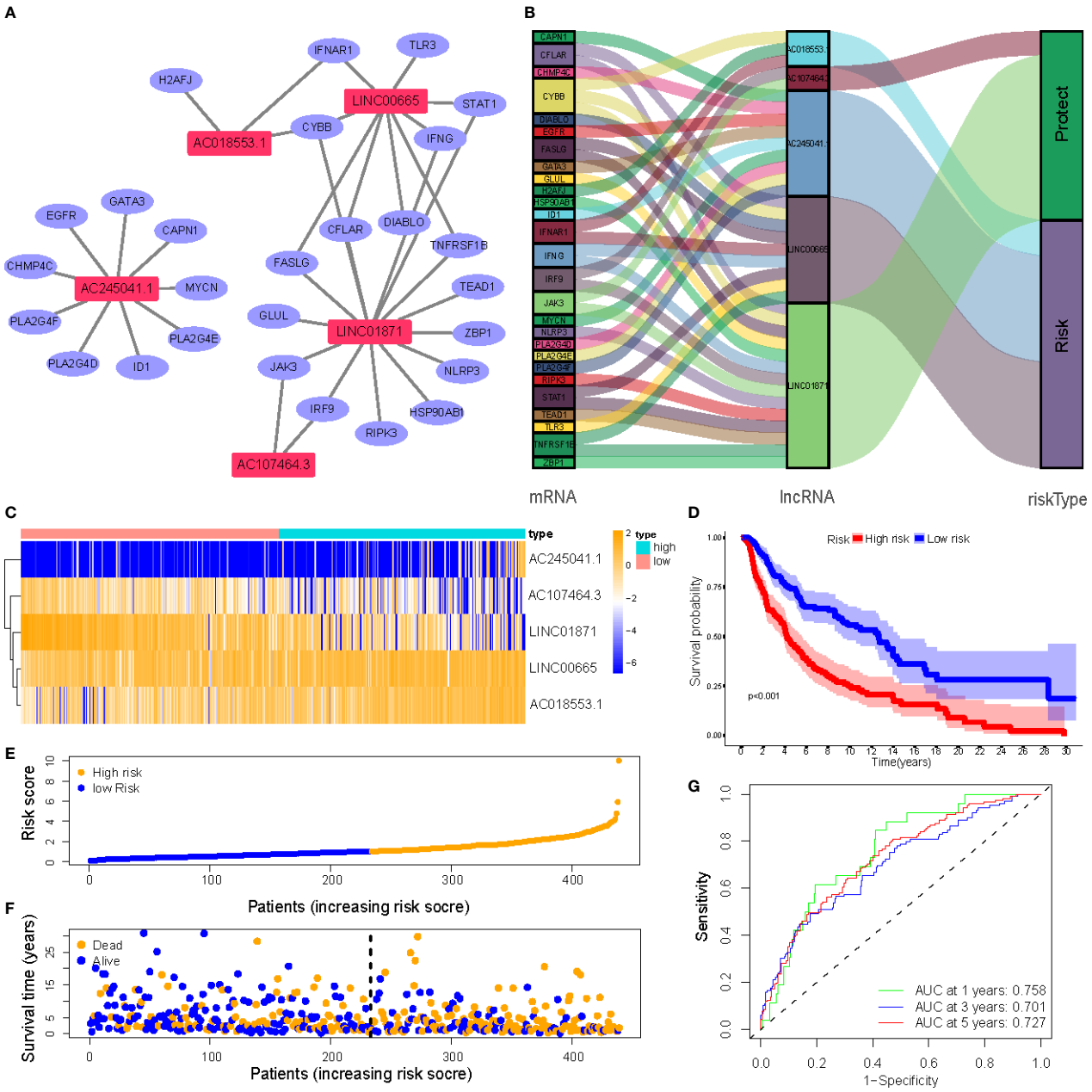



FIGURE 4
**(A)** Correlation network of prognostic lncRNAs and their associated genes; **(B)** The Sankey diagram of the prognostic lncRNAs and the related mRNAs; **(C)** Heatmap of 5 lncRNAs expression in the entire sets; **(D)** Kaplan–Meier (K-M) survival curves of OS of patients between low- and high-risk groups in the entire sets; **(E)** Exhibition of necroptosis-associated lncRNAs model based on risk score of the entire sets; **(F)** Survival time and survival status between low- and high-risk groups in the entire sets; **(G)** The 1-, 3-, and 5-year ROC curves of the entire sets.

(Figures 9A, C). The predictive efficacy of the risk score was further evaluated using a ROC curve and the AUC value of 1-, 3-, and 5-year OS was 0.746, 0.667, and 0.721, respectively in the training cohort (Figure 9B). At the same time, findings from the testing cohort demonstrated AUC value of 1-, 3-, and 5-year OS was 0.775, 0.731, and 0.736, respectively (Figure 9D). In predicting the OS of CM patients, the prognostic NALncSig demonstrated improved overall sensitivity and specificity. HR melanoma patients have shorter survival times, according to the training cohort (Figure 9E). The testing cohort revealed the same outcome concurrently (Figure 9F).

## 3.7 Immune infiltration characteristics and pathways involved

Overall, 15 immune cells, including $CD8^+$ T cells, dendritic cells (DCs), natural killer cells (NK cells), macrophages, and almost all

10.3389/fendo.2023.1180732

immunological activity were more strongly activated in the LR group (Figures 10A, B). Almost all the immune checkpoints expressed more activity in the LR group, such as CTLA4, and CD274 (programmed cell death-1, ligand 1, PD-L1) (Figure 10C). We observed that the LR group had increased PDL1 expression (Figure 10D). Principal component analysis (PCA) maps were also used to display the distribution of patients according to necroptosis-associated genes, necroptosis-associated lncRNAs, and five-lncRNA prognostic signature. The results showed that the five-lncRNA prognostic signature was more suitable to distinguish the risk categories of CM patients (Figures 11A-C).

## 3.8 Identification of GSEA-derived NALncSig

Using GSEA, we compared the two risk groups to determine which biological processes were highly enriched in one over the other. The samples in the HR group have higher levels of glycosylphosphatidylinositol (GPI)-anchor biosynthesis, cell TCA cycle, and aminoacyl-tRNA biosynthesis. While the samples from the group in the LR group are richer in antigen processing and presentation, cell adhesion, and chemokine signaling (Figures 11D, E).

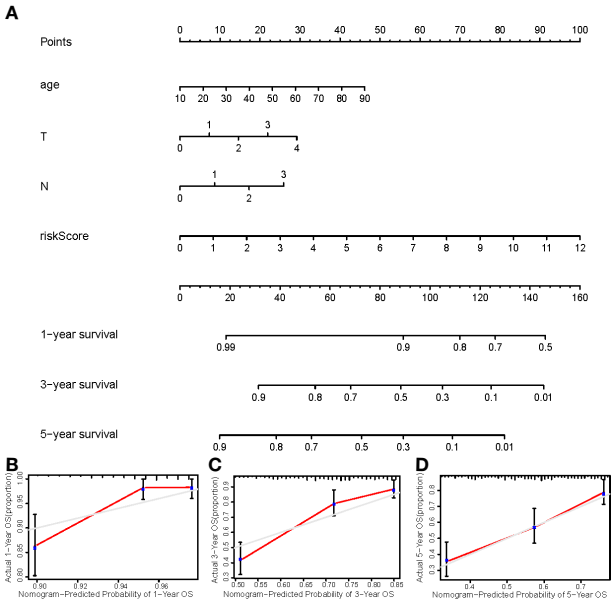



FIGURE 6
**(A)** The nomogram that integrated the risk score, age, and tumor stage predicted the probability of the 1-, 2-, and 3-year OS; **(B–D)** The calibration curves for 1-, 3-, and 5-year OS.

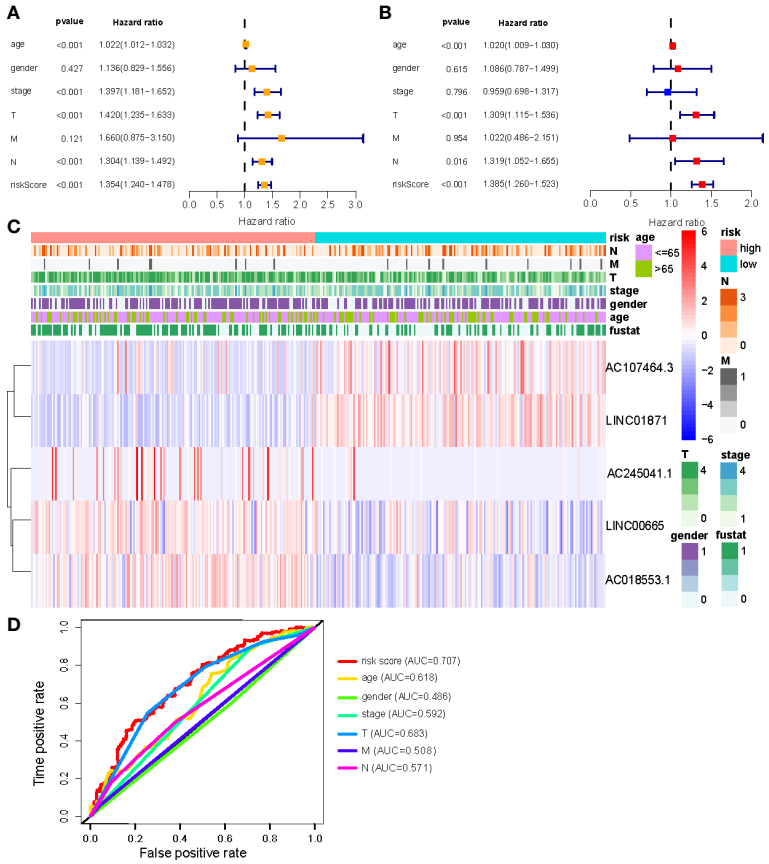



FIGURE 5
**(A, B)** Uni-Cox and Multi-Cox analyses of clinical factors and risk score with OS; **(C)** Heatmap of clinicopathological features and hub lncRNAs expression in two risk subgroups; **(D)** ClinicalROC curves to forecast overall survival of patients.

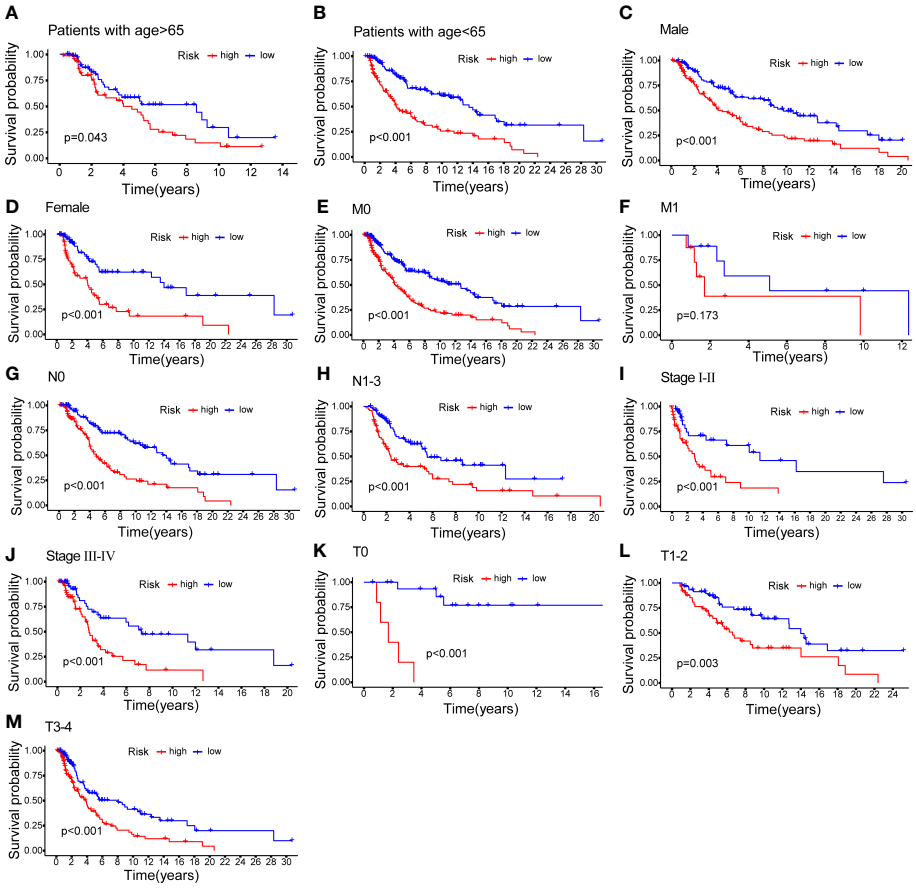



FIGURE 7
K–M methods for the two risk groups (high vs. low) categorized by clinical variables, comprising age **(A, B)**; gender **(C, D)**; M **(E, F)**; N **(G, H)**; stage **(I, J)** and T **(K–M)**.

## 3.9 Validation of the expression of necroptosis-associated lncRNA in cell lines

To get a further assessment for the expression of necroptosis-associated lncRNA in CM, we selected two melanoma cell lines (A2058 and A375) and a normal keratinocyte cell line (HaCaT) to compare the lncRNAs' expression. The expression of LINC00665 and AC018553.1 in both melanoma cell lines are higher than that in HaCaT (Figures 12A, C). While the expression of LINC01871 and AC107464.3 have both low-expression in A2058 and A375 than HaCaT (Figures 12D, E). However, the expression of AC245041.1 showed no statistical significance (Figure 12B).

# 4 Discussion

Melanoma mortality has steadily increased over the last few decades, becoming one of the most dangerous types of human cancer (15, 16). As a result, advancements in the molecular characterization and stratification of CM are critical for achieving advances in the disease's treatment. Finding potential prognostic biomarkers is essential. Necroptosis has a significant role in melanoma invasion, migration, and metastasis since it is implicated in several key ways including the upregulation of death receptors and activation of caspase-8, and mitochondrial complex I inhibition (17, 18) Most studies have focused on the effects of necroptosis on tumor development, treatment resistance, and metastasis; instead, few studies have investigated the potential predictive usefulness of lncRNAs associated with necroptosis in cancer, especially melanoma (19–23).

In this study, a total of 454 patients with CM were randomly divided into a training group, a testing group, as well as a combined group. We summarized 201 necroptosis-associated genes from GeneCards and previous literature. In the combined cohort, we analyzed genes that were differently expressed in GTEx and TCGA. To construct a novel prognostic model, we used uni-Cox analysis and Lasso regression to narrow down the pool of candidate lncRNAs to five (AC245041.1, LINC00665, AC018553.1, LINC01871, and AC107464.3). And these lncRNAs are experimentally verified by qRT-PCR. In these groups, the patients in the HR group had a shorter OS than those in the LR group. Subgroups based on pathological types, grades, M, and N were all shown to be of comparable importance in the prognostic analysis. The multi-Cox regression model identified the necroptosis-associated lncRNA signature as a significant independent risk factor for CM prognosis. A nomogram map was constructed to

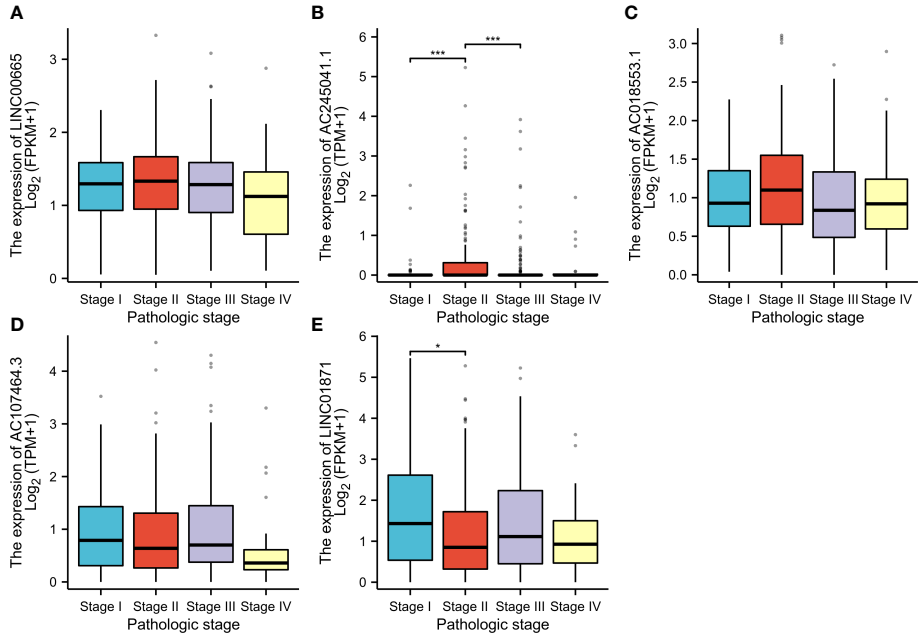



FIGURE 8
According to the clinical data, the relationship between pathologic stage and LINC00665 **(A)**; AC245041.1 **(B)**; AC245041.1 **(C)**; AC107464. 3 **(D)**; LINC01871 **(E)**. *p<0.05, ***p<0.001.

predict the survival of CM patients at 1, 3, and 5 years. The risk scores are the representation of the clinical outcomes. There was a poor prognosis for patients who scored highly on the necroptosis-associated lncRNAs signature. Also, compared with previously established signature, our signature has a better performance in AUC value (24, 25). ssGSEA showed that the LR group had more immune cell infiltration and more immune functions performing than the HR group. When looking at patients' immune systems, the principal component analysis revealed that the five-lncRNA prognostic signature varied by patient. PCA showed that the five-lncRNA prognostic signature could differentiate patients according to their immune status. These findings support the use of necroptosis-associated lncRNA as a prognostic signature for CM, particularly when compared to the predictive ability to exist signatures.

Among these five lncRNAs, LINC01871 and AC107464.3 are protective factors, while AC245041.1, AC018553.1, and LINC01871 are risk factors. And these findings were experimentally verified by qRT-PCR. These lncRNAs are widely studied in different types of cancer. For example, AC245041.1 has been certificated to participate in angiogenesis, cell adhesion, wound healing, and extracellular matrix organization processes in stomach adenocarcinoma (26). By regulating the miR-224-5p/VMA21 axis, LINC00665 promotes melanoma cell growth and migration (27). Similarly, silencing LINC00665 inhibits cutaneous melanoma progression *via* the miR-339-3p/TUBB axis (28). AC018553.1 has been identified as a biomarker in melanoma (29, 30). Meanwhile, AC107464.3 has been associated with a lower risk of developing breast cancer (31). LINC01871 has been proven to be a prognostic factor in breast cancer associated with necroptosis (32), autophagy (33), and ferroptosis (34). As was mentioned before, these lncRNAs play significant roles in different cancers to induce either anti- or pro-tumor effects which is consistent with our findings. Consequently, lncRNA-targeted therapy holds a great deal of potential.

The efficiency of immunological functions is greatly influenced by immune cells(35, 36). Our ssGSEA results show the proportion of the infiltrating immune cells in groups. We found that most immune cells are remarkably higher in the LR risk group, such as $CD8^+$ T cells, DCs, NK cells, and macrophages. Based on single-cell studies, Sukumar et al. shows melanoma reactivity is linked to a high level of dysfunctionality in $CD8^+$ T cells. A rise in $CD8^+$ T cell memory, on the other hand, is associated with anti-melanoma effects (37). $CD8^+$ T cells release perforin and granules to induce melanoma cell apoptosis (38). In previous studies, $CD103^+$ DCs are proven to be critical tumor-draining antigen-presenting cells driving $CD8^+$ T cells to elicit strong T cell response in melanoma (39, 40). NK cells exert cytotoxic activity, facilitating the eradication of melanoma (41, 42). Moreover, macrophage activation leads to the phagocytosis of apoptotic or dead melanoma cells or debris (43). Different immune functions collaborate to maintain the immune balance of the tumor microenvironment. Our result shows that HLA, MHC class I, check-point, APC co-inhibition and co-stimulation activities are enhanced. There is evidence that in melanoma, a lack of HLA expression and downregulation of MHC molecules causes T cells to evade immune recognition and subsequently leads to reduced infiltration, which suggests a poor prognosis for the patients (44). It has also been demonstrated that co-stimulatory molecules enhance $CD8^+$ tumor-infiltrating lymphocyte expansion and also effector-memory (45, 46). The precise balance between costimulatory and coinhibitory signals determines how effectively the immune system responds. These

10.3389/fendo.2023.1180732

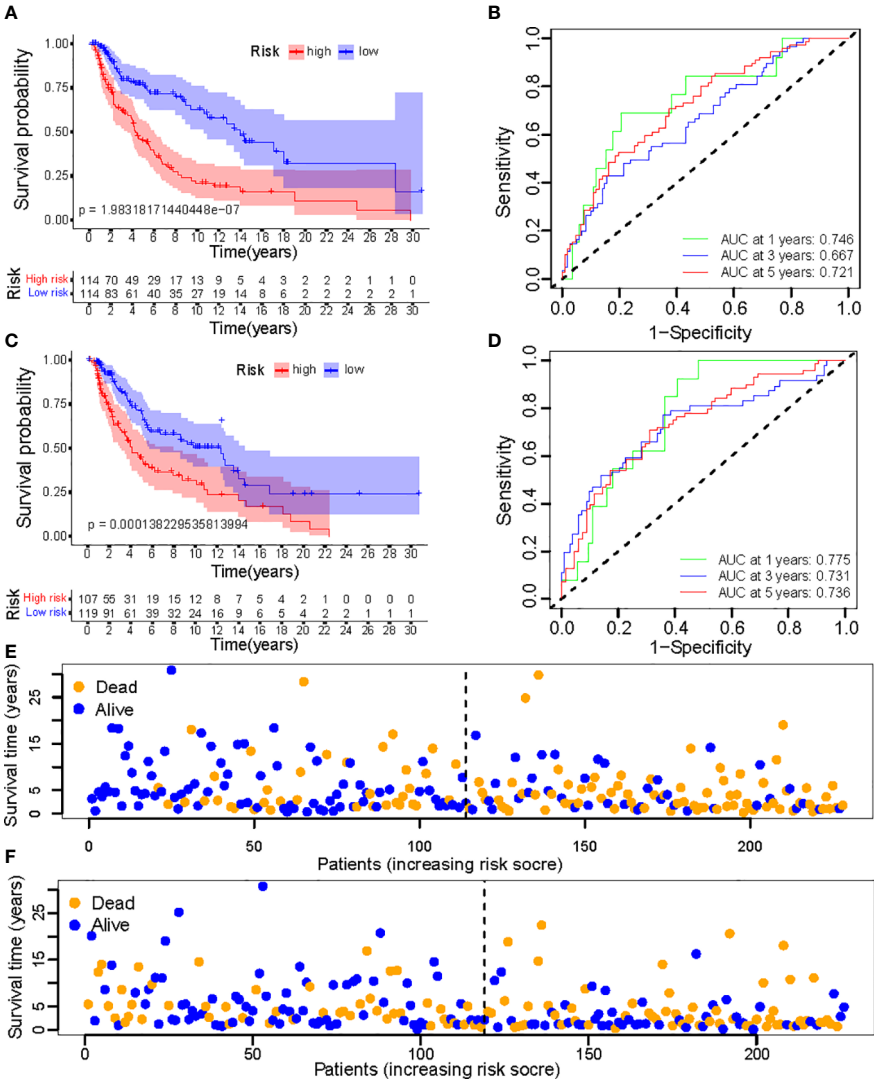



FIGURE 9
**(A, C)** Kaplan–Meier survival curves of OS of patients between low- and high-risk groups in the learning and training set; **(B, D)** The calibration curves for 1-, 3-, and 5-year OS; **(E, F)** Distribution of survival status and risk score.

findings suggest that our conclusion aligns with earlier research suggesting different levels of immune cell infiltration, which leads to melanoma of varying malignancy.

The immunological checkpoint typically has a detrimental effect on immune system control, which is essential for preserving self-tolerance (47). However, tumor cells frequently alter the immunological checkpoint in the tumor, which prevents the immune system from acting as effectively as it could against the tumor (48). Our results showed that checkpoints' expression is relatively high in the LR group, such as B- and T-lymphocyte attenuator (BTLA), Cytotoxic T lymphocyte associate protein-4 (CTLA4), and programmed cell death protein 1(PD-1). The application of anti-CTLA-4 antibodies like Ipilimumab, and anti-programmed cell death 1 (PD-1) antibody-like Nivolumab have led to a long-term disease control in melanoma patients (49). Programmed cell death-ligand 1 (PDL-1), which is the ligand of PD-1, delivers the inhibitory signals together with PD-1 (50). PDL-1 signals present fresh drug development targets and have the potential to be accurate treatment response indicators in melanoma. Our findings show that the LR group expresses more immune checkpoint blockade-related genes than the HR group, which may cause self-destruction and apoptosis of tumor cells. This result is consistent with previous studies (3, 51). The distinct risk score groupings in this prognostic signature could result in a variable potential for immune treatment, which is crucial in practical application that the LR group receive a better outcome by immunotherapy. In particular, the treatment targeting PDL-1 can achieve a good result and is a priority for future research.

According to earlier research, the immune system changed during melanomagenesis and reduced anti-tumor immunity (3). This study's GSEA enrichment analysis found that the HR cohort was enriched for aminoacyl tRNA biosynthesis and citrate cycle TCA cycle, while the LR cohort was enriched for antigen processing and presentation and cell adhesion, which may impede immune escape and metastasis. A

10.3389/fendo.2023.1180732

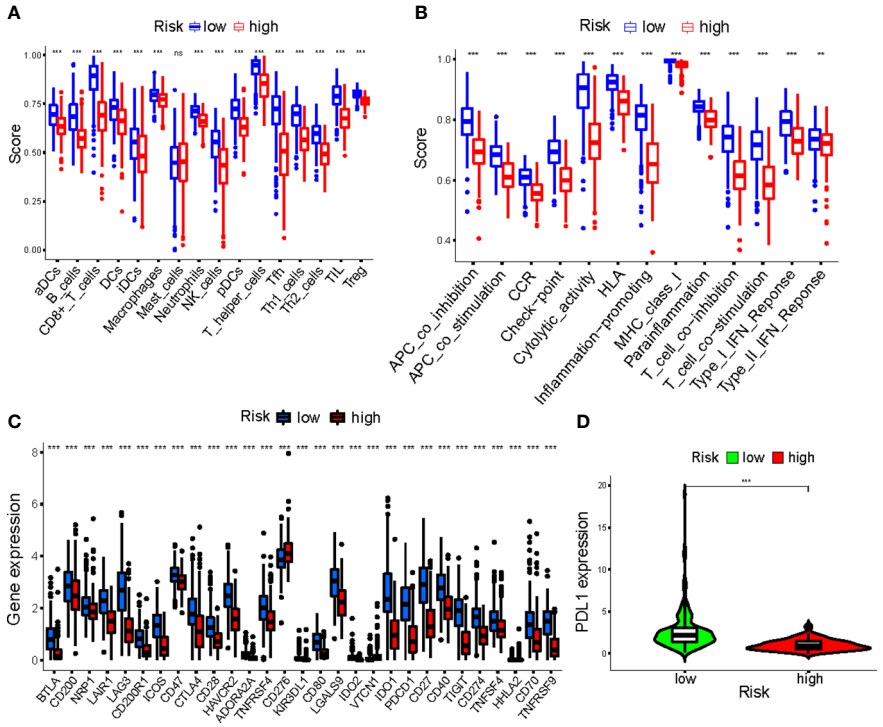



FIGURE 10
**(A, B)** The ssGSEA scores of immune cells and immune functions in clusters; **(C)** The difference of checkpoints expression in clusters; **(D)** The difference of PDL1 expression in clusters. ***p<0.001. ns, no significance.

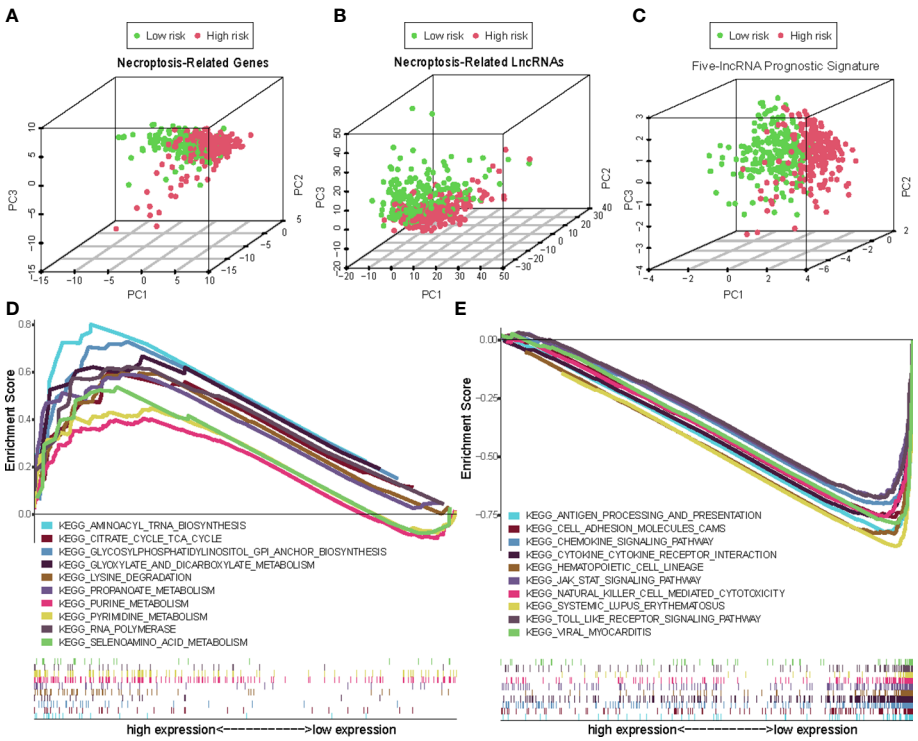



FIGURE 11
The principal component analysis (PCA) maps show the distribution of patients based on the **(A)** Necroptosis-associated gene sets; **(B)** Necroptosis-associated lncRNAs sets; **(C)** The five-lncRNA prognostic signature; **(D, E)** GSEA of high- and low- risk groups.

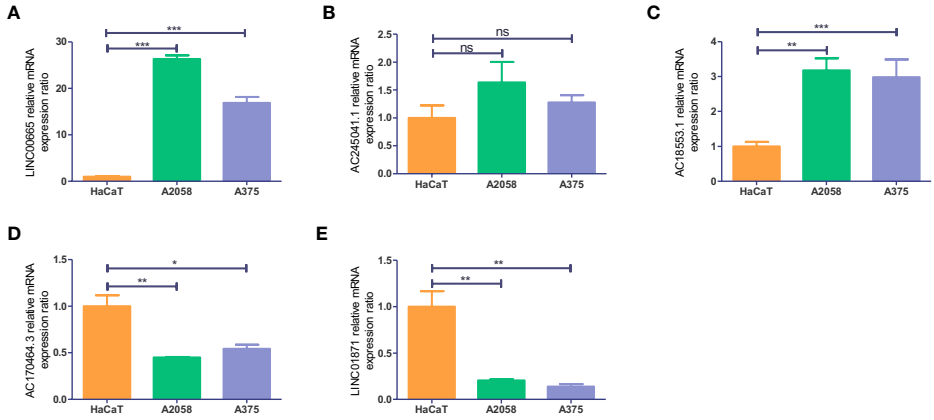



FIGURE 12
Validation of the expression level of the five necroptosis-associated lncRNAs in cell lines. Expression analysis of **(A)** LINC00665; **(B)** AC245041.1; **(C)** AC018553.1; **(D)** AC107464.3 **(E)** LINC01871. *p<0.05, **p<0.005,***p<0.001. ns, no significance.

previous study demonstrated that defects in antigen presentation can predict outcomes to immune checkpoint blockade in melanoma (52). Also, an intact MHC class II antigen presentation pathway improves survival in melanoma(53). The above may be the reason for the better prognosis for the LR group. Meanwhile, it has been discovered that melanoma growth and immune response are influenced by metabolic regulation and metabolic interactions between cancer cells and the microenvironment (54, 55), which was consistent with our result that the HR group is enriched in biosynthesis and metabolism, which may help the proliferation and metastasis of CM.

However, the model we developed still had certain weaknesses despite our attempts to address them using several different approaches. There were problems with the research that was inevitable given that it was conducted in hindsight. Additionally, further research is needed that integrates biochemical tests with clinical prognostic information to properly identify how these lncRNAs impact the prognosis of CM patients through necroptosis.

In conclusion, a signature of five necroptosis-associated lncRNAs was created in this investigation to predict the prognosis of CM. Our results also indicated that NALncSig is a separate risk factor for CM. We hope to provide a new reference for the current prognostic assessment of CM and to shed new light on treatment strategies.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

## Author contributions

Conception and design: ZWC and ZL; acquisition, analysis, or interpretation of data: GC, BYL, and YNG; drafting the work: ZWC, BYS, and YHZ; revising the work: JGM and BQS. All authors contributed to the article and approved the submitted version.

## Funding

This study was funded by the National Natural Science Foundation of China (82072182) and Shaanxi Science and Technology Coordination and Innovation Project Grants (2020SF-179).

## Conflict of interest

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.

## Publisher's note



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1180732/full#supplementary-material

# References


1. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, et al. Whole-genome landscapes of major melanoma subtypes. *Nature* (2017) 545:175–80. doi: 10.1038/nature22071

2. Sun S, Zhang G, Zhang L. A novel ferroptosis-related lncRNA prognostic model and immune infiltration features in skin cutaneous melanoma. *Front Cell Dev Biol* (2021) 9:790047. doi: 10.3389/fcell.2021.790047

3. Marzagalli M, Ebelt ND, Manuel ER. Unraveling the crosstalk between melanoma and immune cells in the tumor microenvironment. *Semin Cancer Biol* (2019) 59:236–50. doi: 10.1016/j.semcancer.2019.08.002

4. Ma Y, Wang N, Yang S. Skin cutaneous melanoma properties of immune-related lncRNAs identifying potential prognostic biomarkers. *Aging (Albany NY)* (2022) 14:3030–48. doi: 10.18632/aging.203982

5. Krysko O, Aaes TL, Kagan VE, D'herde K, Bachert C, Leybaert L, et al. Necroptotic cell death in anti-cancer therapy. *Immunol Rev* (2017) 280:207–19. doi: 10.1111/imr.12583

6. Gong Y, Fan Z, Luo G, Yang C, Huang Q, Fan K, et al. The role of necroptosis in cancer biology and therapy. *Mol Cancer* (2019) 18:100. doi: 10.1186/s12943-019-1029-8

7. Koo GB, Morgan MJ, Lee DG, Kim WJ, Yoon JH, Koo JS, et al. Methylation-dependent loss of RIP3 expression in cancer represses programmed necrosis in response to chemotherapeutics. *Cell Res* (2015) 25:707–25. doi: 10.1038/cr.2015.56

8. Stoll G, Ma Y, Yang H, Kepp O, Zitvogel L, Kroemer G. Pro-necrotic molecules impact local immunosurveillance in human breast cancer. *Oncoimmunology* (2017) 6: e1299302. doi: 10.1080/2162402X.2017.1299302

9. Moriwaki K, Bertin J, Gough PJ, Orlowski GM, Chan FK. Differential roles of RIPK1 and RIPK3 in TNF-induced necroptosis and chemotherapeutic agent-induced cell death. *Cell Death Dis* (2015) 6:e1636. doi: 10.1038/cddis.2015.16

10. Wang D, Chen J, Li B, Jiang Q, Liu L, Xia Z, et al. A noncoding regulatory RNA Gm31932 induces cell cycle arrest and differentiation in melanoma *via* the miR-344d-3-5p/Prc1 (and Nuf2) axis. *Cell Death Dis* (2022) 13:314. doi: 10.1038/s41419-022-04736-6

11. Han X, Jia Y, Chen X, Sun C, Sun J. lncRNA TINCR attenuates the proliferation and invasion, and enhances the apoptosis of cutaneous malignant melanoma cells by regulating the miR−424−5p/LATS1 axis. *Oncol Rep* (2021) 46:238. doi: 10.3892/or.2021.8189

12. Shu Q, Zhou Y, Zhu Z, Chen X, Fang Q, Zhong L, et al. A novel risk model based on autophagy-related LncRNAs predicts prognosis and indicates immune infiltration landscape of patients with cutaneous melanoma. *Front Genet* (2022) 13:885391. doi: 10.3389/fgene.2022.885391

13. Zhong J, Wang Z, Houssou Hounye A, Liu J, Zhang J, Qi M. A novel pyroptosis-related LncRNA signature predicts prognosis and indicates tumor immune microenvironment in skin cutaneous melanoma. *Life Sci* (2022) 307:120832. doi: 10.1016/j.lfs.2022.120832

14. Peng G, Chi H, Gao X, Zhang J, Song G, Xie X, et al. Identification and validation of neurotrophic factor-related genes signature in HNSCC to predict survival and immune landscapes. *Front Genet* (2022) 13:1010044. doi: 10.3389/fgene.2022.1010044

15. Mohammadpour A, Derakhshan M, Darabi H, Hedayat P, Momeni M. Melanoma: where we are and where we go. *J Cell Physiol* (2019) 234:3307–20. doi: 10.1002/jcp.27286

16. Teixido C, Castillo P, Martinez-Vila C, Arance A, Alos L. Molecular markers and targets in melanoma. *Cells* (2021) 10:2320. doi: 10.3390/cells10092320

17. Ashrafizadeh M, Mohammadinejad R, Tavakol S, Ahmadi Z, Roomiani S, Katebi M. Autophagy, anoikis, ferroptosis, necroptosis, and endoplasmic reticulum stress: potential applications in melanoma therapy. *J Cell Physiol* (2019) 234:19471–9. doi: 10.1002/jcp.28740

18. Liu N, Li Y, Chen G, Ge K. Evodiamine induces reactive oxygen species-dependent apoptosis and necroptosis in human melanoma a-375 cells. *Oncol Lett* (2020) 20:121. doi: 10.3892/ol.2020.11983

19. Ando Y, Ohuchida K, Otsubo Y, Kibe S, Takesue S, Abe T, et al. Necroptosis in pancreatic cancer promotes cancer cell migration and invasion by release of CXCL5. *PloS One* (2020) 15:e0228015. doi: 10.1371/journal.pone.0228015

20. Tang R, Xu J, Zhang B, Liu J, Liang C, Hua J, et al. Ferroptosis, necroptosis, and pyroptosis in anticancer immunity. *J Hematol Oncol* (2020) 13:110. doi: 10.1186/s13045-020-00946-7

21. Bolik J, Krause F, Stevanovic M, Gandraß M, Thomsen I, Schacht SS, et al. Inhibition of ADAM17 impairs endothelial cell necroptosis and blocks metastasis. *J Exp Med* (2022) 219:e20201039. doi: 10.1084/jem.20201039

22. Yan J, Wan P, Choksi S, Liu ZG. Necroptosis and tumor progression. *Trends Cancer* (2022) 8:21–7. doi: 10.1016/j.trecan.2021.09.003

23. Zhang T, Wang Y, Inuzuka H, Wei W. Necroptosis pathways in tumorigenesis. *Semin Cancer Biol* (2022) 86:32–40. doi: 10.1016/j.semcancer.2022.07.007

24. Guo S, Chen J, Yi X, Lu Z, Guo W. Identification and validation of ferroptosis-related lncRNA signature as a prognostic model for skin cutaneous melanoma. *Front Immunol* (2022) 13:985051. doi: 10.3389/fimmu.2022.985051

25. Rong J, Wang H, Yao Y, Wu Z, Chen L, Jin C, et al. Identification of m7G-associated lncRNA prognostic signature for predicting the immune status in cutaneous melanoma. *Aging (Albany NY)* (2022) 14:5233–49. doi: 10.18632/aging.204151

26. Han C, Zhang C, Wang H, Li K, Zhao L. Angiogenesis-related lncRNAs predict the prognosis signature of stomach adenocarcinoma. *BMC Cancer* (2021) 21:1312. doi: 10.1186/s12885-021-08987-y

27. Wang X, Wang Y, Lin F, Xu M, Zhao X. Long non-coding RNA LINC00665 promotes melanoma cell growth and migration *via* regulating the miR-224-5p/VMA21 axis. *Exp Dermatol* (2022) 31:64–73. doi: 10.1111/exd.14246

28. Liu Y, Ma S, Ma Q, Zhu H. Silencing LINC00665 inhibits cutaneous melanoma *in vitro* progression and induces apoptosis *via* the miR-339-3p/TUBB. *J Clin Lab Anal* (2022) 36:e24630. doi: 10.1002/jcla.24630

29. Li Z, Wei J, Zheng H, Zhang Y, Song M, Cao H, et al. The new horizon of biomarker in melanoma patients: a study based on autophagy-related long non-coding RNA. *Med (Baltimore)* (2022) 101:e28553. doi: 10.1097/MD.0000000000028553

30. Qiu Y, Wang HT, Zheng XF, Huang X, Meng JZ, Huang JP, et al. Autophagy-related long non-coding RNA prognostic model predicts prognosis and survival of melanoma patients. *World J Clin cases* (2022) 10:3334–51. doi: 10.12998/wjcc.v10.i11.3334

31. Shi GJ, Zhou Q, Zhu Q, Wang L, Jiang GQ. A novel prognostic model associated with the overall survival in patients with breast cancer based on lipid metabolism-related long noncoding RNAs. *J Clin Lab Anal* (2022) 36:e24384. doi: 10.1002/jcla.24384

32. Tao S, Tao K, Cai X. Necroptosis-associated lncRNA prognostic model and clustering analysis: prognosis prediction and tumor-infiltrating lymphocytes in breast cancer. *J Oncol* (2022) 2022:7099930. doi: 10.1155/2022/7099930

33. Wu Q, Li Q, Zhu W, Zhang X, Li H. Identification of autophagy-related long non-coding RNA prognostic signature for breast cancer. *J Cell Mol Med* (2021) 25:4088–98. doi: 10.1111/jcmm.16378

34. Xu Z, Jiang S, Ma J, Tang D, Yan C, Fang K. Comprehensive analysis of ferroptosis-related LncRNAs in breast cancer patients reveals prognostic value and relationship with tumor immune microenvironment. *Front Surg* (2021) 8:742360. doi: 10.3389/fsurg.2021.742360

35. Chi H, Peng G, Wang R, Yang F, Xie X, Zhang J, et al. Cuprotosis programmed-Cell-Death-Related lncRNA signature predicts prognosis and immune landscape in PAAD patients. *Cells* (2022) 11:3436. doi: 10.3390/cells11213436

36. Chi H, Xie X, Yan Y, Peng G, Strohmer DF, Lai G, et al. Natural killer cell-related prognosis signature characterizes immune landscape and predicts prognosis of HNSCC. *Front Immunol* (2022) 13:1018685. doi: 10.3389/fimmu.2022.1018685

37. Sukumar M, Liu J, Ji Y, Subramanian M, Crompton JG, Yu Z, et al. Inhibiting glycolytic metabolism enhances CD8+ T cell memory and antitumor function. *J Clin Invest* (2013) 123:4479–88. doi: 10.1172/JCI69589

38. Thor Straten P, Guldberg P, Grønbaek K, Hansen MR, Kirkin AF, Seremet T, et al. *In situ* T cell responses against melanoma comprise high numbers of locally expanded T cell clonotypes. *J Immunol* (1999) 163:443–7. doi: 10.4049/jimmunol.163.1.443

39. Van Lint S, Wilgenhof S, Heirman C, Corthals J, Breckpot K, Bonehill A, et al. Optimized dendritic cell-based immunotherapy for melanoma: the TriMix-formula. *Cancer Immunol Immunother* (2014) 63:959–67. doi: 10.1007/s00262-014-1558-3

40. Roberts EW, Broz ML, Binnewies M, Headley MB, Nelson AE, Wolf DM, et al. Critical role for CD103(+)/CD141(+) dendritic cells bearing CCR7 for tumor antigen trafficking and priming of T cell immunity in melanoma. *Cancer Cell* (2016) 30:324–36. doi: 10.1016/j.ccell.2016.06.003

41. Mgrditchian T, Arakelian T, Paggetti J, Noman MZ, Viry E, Moussay E, et al. Targeting autophagy inhibits melanoma growth by enhancing NK cells infiltration in a CCL5-dependent manner. *Proc Natl Acad Sci USA* (2017) 114:E9271–e9279. doi: 10.1073/pnas.1703921114

42. Bhat H, Zaun G, Hamdan TA, Lang J, Adomati T, Schmitz R, et al. Arenavirus induced CCL5 expression causes NK cell-mediated melanoma regression. *Front Immunol* (2020) 11:1849. doi: 10.3389/fimmu.2020.01849

43. Huang Z, Gan J, Long Z, Guo G, Shi X, Wang C, et al. Targeted delivery of let-7b to reprogramme tumor-associated macrophages and tumor infiltrating dendritic cells for tumor rejection. *Biomaterials* (2016) 90:72–84. doi: 10.1016/j.biomaterials.2016.03.009

44. Sucker A, Zhao F, Real B, Heeke C, Bielefeld N, Maßen S, et al. Genetic evolution of T-cell resistance in the course of melanoma progression. *Clin Cancer Res* (2014) 20:6593–604. doi: 10.1158/1078-0432.CCR-14-0567

45. Chacon JA, Wu RC, Sukhumalchandra P, Molldrem JJ, Sarnaik A, Pilon-Thomas S, et al. Co-Stimulation through 4-1BB/CD137 improves the expansion and function of CD8(+) melanoma tumor-infiltrating lymphocytes for adoptive T-cell therapy. *PloS One* (2013) 8:e60031. doi: 10.1371/journal.pone.0060031

46. Chacon JA, Sarnaik AA, Pilon-Thomas S, Radvanyi L. Triggering co-stimulation directly in melanoma tumor fragments drives CD8(+) tumor-infiltrating lymphocyte expansion with improved effector-memory properties. *Oncoimmunology* (2015) 4: e1040219. doi: 10.1080/2162402X.2015.1040219

47. Kalaora S, Nagler A, Wargo JA, Samuels Y. Mechanisms of immune activation and regulation: lessons from melanoma. *Nat Rev Cancer* (2022) 22:195–207. doi: 10.1038/s41568-022-00442-9

48. Chi H, Peng G, Yang J, Zhang J, Song G, Xie X, et al. Machine learning to construct sphingolipid metabolism genes signature to characterize the immune landscape and prognosis of patients with uveal melanoma. *Front Endocrinol (Lausanne)* (2022) 13:1056310. doi: 10.3389/fendo.2022.1056310

49. Larkin J, Chiarion-Sileni V, Gonzalez R, Grob JJ, Rutkowski P, Lao CD, et al. Five-year survival with combined nivolumab and ipilimumab in advanced melanoma. *N Engl J Med* (2019) 381:1535–46. doi: 10.1056/NEJMoa1910836

50. Watanabe N, Gavrieli M, Sedy JR, Yang J, Fallarino F, Loftin SK, et al. BTLA is a lymphocyte inhibitory receptor with similarities to CTLA-4 and PD-1. *Nat Immunol* (2003) 4:670–9. doi: 10.1038/ni944

51. Wei J, Kishton RJ, Angel M, Conn CS, Dalla-Venezia N, Marcel V, et al. Ribosomal proteins regulate MHC class I peptide generation for immunosurveillance. *Mol Cell* (2019) 73:1162–73.e5. doi: 10.1016/j.molcel.2018.12.020

52. Thompson JC, Davis C, Deshpande C, Hwang WT, Jeffries S, Huang A, et al. Gene signature of antigen processing and presentation machinery predicts response to checkpoint blockade in non-small cell lung cancer (NSCLC) and melanoma. *J Immunother Cancer* (2020) 8:e000974. doi: 10.1136/jitc-2020-000974

53. Buetow KH, Meador LR, Menon H, Lu YK, Brill J, Cui H, et al. High GILT expression and an active and intact MHC class II antigen presentation pathway are associated with improved survival in melanoma. *J Immunol* (2019) 203:2577–87. doi: 10.4049/jimmunol.1900476

54. Ruocco MR, Avagliano A, Granato G, Vigliar E, Masone S, Montagnani S, et al. Metabolic flexibility in melanoma: a potential therapeutic target. *Semin Cancer Biol* (2019) 59:187–207. doi: 10.1016/j.semcancer.2019.07.016

55. Tasdogan A, Faubert B, Ramesh V, Ubellacker JM, Shen B, Solmonson A, et al. Metabolic heterogeneity confers differences in melanoma metastatic potential. *Nature* (2020) 577:115–20. doi: 10.1038/s41586-019-1847-2

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 24 May 2023
DOI 10.3389/fendo.2023.1196293



OPEN ACCESS

EDITED BY
Prem P. Kushwaha,
Case Western Reserve University,
United States

REVIEWED BY
Alpna Tyagi,
University of Colorado Anschutz Medical Campus, United States
Raushan Kumar,
Allahabad University, India

*CORRESPONDENCE
Lu Xie
xielu@sibpt.com
Lei Liu
liulei@fudan.edu.cn
Yin Wang
chinawangyin@foxmail.com

RECEIVED 29 March 2023
ACCEPTED 10 May 2023
PUBLISHED 24 May 2023

CITATION
Chen Y, Yao L, Zhao S, Xu M, Ren S, Xie L, Liu L and Wang Y (2023) The oxidative aging model integrated various risk factors in type 2 diabetes mellitus at system level. *Front. Endocrinol.* 14:1196293. doi: 10.3389/fendo.2023.1196293



# The oxidative aging model integrated various risk factors in type 2 diabetes mellitus at system level

Yao Chen[1], Lilin Yao[1], Shuheng Zhao[1], Mengchu Xu[1], Siwei Ren[1], Lu Xie[2]*, Lei Liu[3]* and Yin Wang[1,4]*

[1]Department of Biomedical Engineering, School of Intelligent Medicine, China Medical University, Shenyang, Liaoning, China, [2]Shanghai-MOST Key Laboratory of Health and Disease Genomics & Institute for Genome and Bioinformatics, Shanghai Institute for Biomedical and Pharmaceutical Technologies, Shanghai, China, [3]Intelligent Medicine Institute, Fudan University, Shanghai, China, [4]Key Laboratory of GI Cancer Etiology and Prevention in Liaoning Province, The First Hospital of China Medical University, Shenyang, China

**Background:** Type 2 diabetes mellitus (T2DM) is a chronic endocrine metabolic disease caused by insulin dysregulation. Studies have shown that aging-related oxidative stress (as "oxidative aging") play a critical role in the onset and progression of T2DM, by leading to an energy metabolism imbalance. However, the precise mechanisms through which oxidative aging lead to T2DM are yet to be fully comprehended. Thus, it is urgent to integrate the underlying mechanisms between oxidative aging and T2DM, where meaningful prediction models based on relative profiles are needed.

**Methods:** First, machine learning was used to build the aging model and disease model. Next, an integrated oxidative aging model was employed to identify crucial oxidative aging risk factors. Finally, a series of bioinformatic analyses (including network, enrichment, sensitivity, and pan-cancer analyses) were used to explore potential mechanisms underlying oxidative aging and T2DM.

**Results:** The study revealed a close relationship between oxidative aging and T2DM. Our results indicate that nutritional metabolism, inflammation response, mitochondrial function, and protein homeostasis are key factors involved in the interplay between oxidative aging and T2DM, even indicating key indices across different cancer types. Therefore, various risk factors in T2DM were integrated, and the theories of oxi-inflamm-aging and cellular senescence were also confirmed.

**Conclusion:** In sum, our study successfully integrated the underlying mechanisms linking oxidative aging and T2DM through a series of computational methodologies.

KEYWORDS

**oxidative stress, type 2 diabetes mellitus, energy metabolism, aging, pan-cancer analysis**

# 1 Introduction

Type 2 diabetes mellitus (T2DM) is a chronic endocrine metabolic disease caused mostly by insulin dysfunction. The increasing prevalence of diabetes has resulted in a great economic burden in many countries (1). According to statistics, there are approximately 536.6 million people with diabetes worldwide, and this number is expected to rise to approximately 783.2 million in 2045, with T2DM accounting for approximately 90% (1, 2). Therefore, it is imperative to study the etiology of T2DM in depth.

Various reports have shown that T2DM is closely related to aging, with aging being one of the most vital risk factors for T2DM (3, 4). Adipose tissue (AT) is redistributed during aging, which affects the sensitivity of insulin (5). Furthermore, the normal function of pancreatic beta cells also declines (3), and aging causes inflammation and low nutritional status, affecting the endocrine system (6). Additionally, a series of risk factors for T2DM are vital to other age-related diseases, such as Alzheimer's disease (AD), cardiovascular disease (CVD), and cancer (7–10).

During the aging process, oxidative stress accumulates, leading to an energy imbalance that is key to T2DM (11, 12). For example, oxidative intermediates can damage pancreatic beta cells and exacerbate insulin resistance (13). Moreover, accumulated reactive oxygen species also accelerate aging-related DNA damage and induce cellular senescence (14, 15). With increasing age, the free radical dynamic balance in cells is gradually broken, causing an increase in free radical concentration and inducing the oxidation reaction, leading to T2DM (16). In addition, oxidative stress is closely interrelated with inflammation (17) by activating multiple transcription factors in the inflammatory response (18). Furthermore, abnormal oxidative stress dysregulates the balance of energy metabolism during T2DM development (19–23). In summary, the potential mechanism by which aging-related oxidative stress (often described as "oxidative aging" (24)) triggers T2DM needs to be further studied at the system level (Figure 1A).

With the development of artificial intelligence, many research results on diabetes have utilized machine learning (ML), which can gain useful information from original profiles. ML can be widely used in the risk prediction, prognosis, and treatment of clinical diseases such as cardiovascular disease and cancer (25, 26). Recently, it was reported that ML can predict the occurrence of T2DM and its complications, as well as identify key markers in T2DM (27–29). Additionally, Mendelian randomization (MR) is conducive to integrating biological information (30, 31). Although numerous studies have revealed some risk factors/mechanisms associated with T2DM, the underlying mechanism between

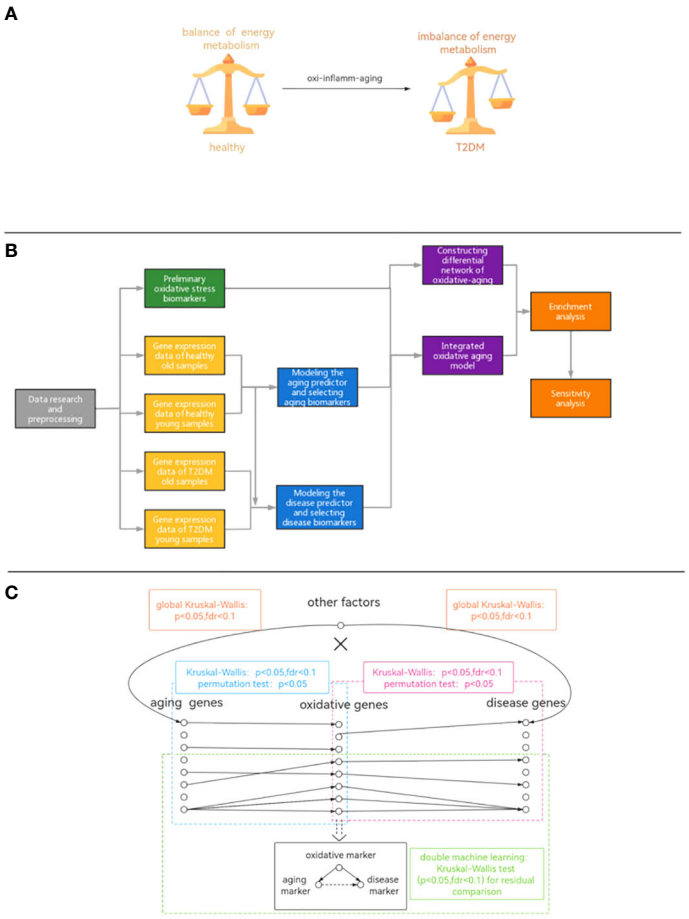



FIGURE 1
**(A)** Diagram of the hypothetical mechanism. **(B)** The workflow of our study. **(C)** The pipeline of integrated oxidative aging model.

oxidative aging and T2DM is still unclear and requires further exploration.

To further explore the potential mechanisms between oxidative aging and T2DM, a series of computational studies was performed in this paper (Figures 1B, C): (1) Machine learning was used to identify aging and disease (T2DM) markers. (2) An integrative model was built to further explore essential relationships between oxidative aging (aging-related oxidative stress) and T2DM (Figure 1C). (3) Network analysis, enrichment analysis and sensitivity analysis were used to investigate the underlying mechanisms between oxidative aging and T2DM markers. (4) Relative biological functions of identified oxidative aging markers were further validated across different cancer types. As a result, the underlying mechanisms of T2DM (i.e., nutritional metabolism, inflammatory response, mitochondrial function and protein homeostasis) were integrated, which can also provide key indices in cancers.

# Results

## 2.1 Modeling prediction models and identifying relative biomarkers

The gene expression profiles were obtained from the GEO database, including 489 samples and 12,958 genes (Tables S1–S3). These genes were ranked by the ReliefF algorithm, and then the aging predictor and disease predictor were built using the k-nearest neighbors (kNN; k=3 with the correlation distance) algorithm, optimized by 10-fold cross-validation. The accuracy of the aging predictor in the test set was 0.70455 and 0.7279 in the aging and disease predictors (Figure 1; Table 1), respectively. Furthermore, the ROC area under the curve (AUC) for the aging and disease predictor models were 0.7712 and 0.72788 (Figure 2), respectively. As a result, our predictors were sufficiently accurate in both aging and disease models.

Both aging and disease markers have meaningful biological functions. For example, OSBPL1A (oxysterol binding protein-like 1A, ReliefF weight=0.058) was the top aging marker. OSBPL1A is one of a set of intracellular lipid receptors and is closely related to lipid metabolism and cholesterol metabolism (32, 33). TIGD4 (tigger transposable element derived 4, ReliefF weight=0.0253), as the top disease marker, was related to glycogen metabolism. In sum, the abnormal metabolism of lipids, cholesterol and glycogen can lead to T2DM (34). These results indicated the crucial role of energy metabolism in T2DM.

## 2.2 Identifying the oxidative-aging risk factors by the integrated prediction model

The integrated oxidative aging model was built to explore essential relationships among aging, oxidative and T2DM markers (details are shown in Materials and Methods 5.3, with a total of 11829 "aging-oxidative-disease" triples). The top 10 aging, oxidative and disease markers are shown in Table 2, including relative experimental details (35–43). For example, ADP-ribosylarginine hydrolase (ADPRH) is the top aging marker, participating in the regulation of various cellular processes, including both immunity and aging (44). ADPRH adversely influences the immune system via CD8+ T cells, hence promoting an imbalance in energy metabolism (45). TPST1 (tyrosyl protein sulfotransferase 1) is the top disease marker, catalyzing the posttranslational sulfation of tyrosine residues within acidic motifs of many polypeptides in all multicellular organisms (46). TPST1 promoted the secretion of some cytokines and then induced the inflammatory response (47). COX5A (cytochrome C oxidase subunit 5A) is the top oxidative marker related to mitochondrial function (48), which induces an imbalance in energy metabolism and insulin resistance (35). In addition, the predictor accuracy calculated by the selected disease markers was 0.7662 (Table 1). In sum, these results indicated that the integrated oxidative aging model could identify essential relationships in T2DM, even with enough prediction ability.

## 2.3 Sensitivity analysis further highlighted the imbalance of energy metabolism in T2DM

The Markov chain Monte Carlo (MCMC) method was used to evaluate the sensitive relationship between oxidative aging and T2DM. As a result, a series of triples were identified as key components (2501 out of 11829) in the integrated oxidative aging model.

The top 10 sensitive relationships (by calculating the absolute differential frequency) are shown in Table 3, where the top relationship was "OSBPL7-COX7C-TM6SF1" (difference=-0.03935). Additionally, Table 3 also displayed experimental details of relative oxidative markers (49–55). OSBPL7 (oxysterol binding protein like 7) is an oxysterol-binding protein-like (OSBPL) family member involved in lipid binding and transport and induces cholesterol efflux (56, 57). COX7C (cytochrome C oxidase subunit 7C) is an enzyme in the electron transport chain related to cellular respiration and is also a potential biomarker of diabetes mellitus (58, 59). Transmembrane 6 superfamily member 1

TABLE 1 The accuracy of aging predictor and disease predictor.

| | The accuracy of training datasets | The accuracy of test datasets | Markers used for classification |
|---|---|---|---|
| The aging model | 0.7552 | 0.70455 | 304 |
| The disease predictor | 0.8328 | 0.7279 | 299 |
| The integrated oxidative aging model | 0.8485 | 0.7662 | 282 |

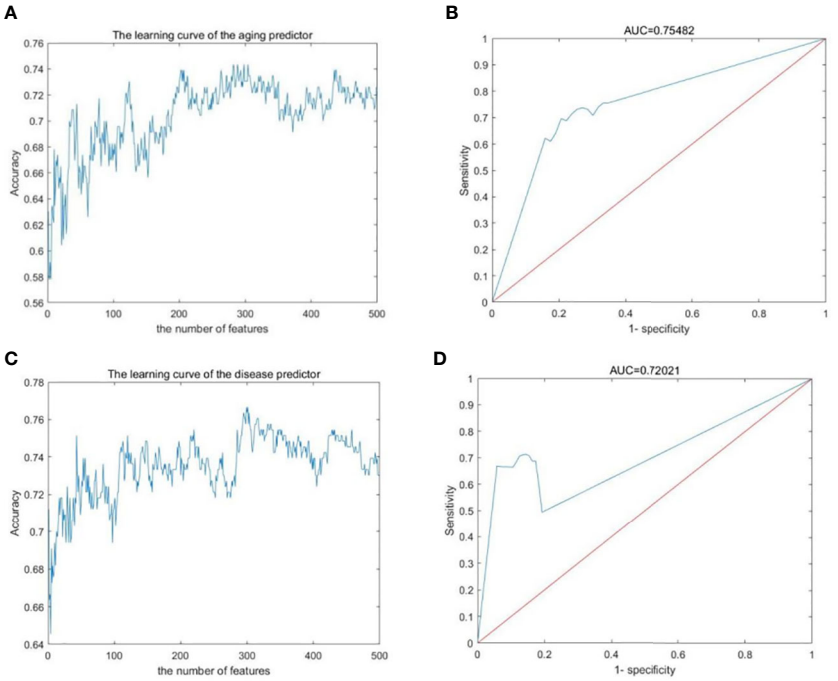



FIGURE 2
Machine learning results. **(A, B)** Aging predictor from our previous study, selecting the number of aging markers. **(C, D)** The improved inflamm-aging predictor, selecting the number of disease markers. **(A, C)** Learning curve for the training dataset. **(B, D)** The ROC curve for the test dataset.

TABLE 2 The top 10 aging markers, disease markers and oxidative markers from the integrated oxidative model.

| Aging marker | Times | Disease marker | Times | Oxidative marker | Times | Experimental results of the oxidative marker | Reference | Experimental method |
|---|---|---|---|---|---|---|---|---|
| ADPRH | 21 | TPST1 | 26 | COX5A | 86 | COX5A is related to mitochondrial dysfunction in insulin resistance. | (35) | Western blotting |
| OAS3 | 14 | PGK1 | 25 | CYB5B | 83 | CYB5B is related to diabetic retinopathy. | (36) | Quantitative PCR |
| RNF10 | 13 | ADM | 25 | ERCC8 | 77 | Loss of ERCC8 will have insulin-dependent diabetes with Cockayne syndrome. | (37) | DNA hybridization |
| LMO7 | 11 | PLAC8 | 24 | ANXA1 | 62 | ANXA1 is related to weight gain and diet-induced insulin resistance. | (38) | Flow cytometry |
| KATNB1 | 10 | ITGB5 | 22 | ATRN | 62 | | | |
| PLD1 | 10 | STEAP4 | 21 | BAK1 | 59 | BAK1 is related to mitochondria-dependent programmed cell death. | (39) | Cell culture of hepathocellular carcinoma and renal epithelial |
| PTPLB | 9 | TMEM163 | 21 | CD36 | 58 | CD36 is a key molecule to limit β-cell function in T2DM associated with obesity. | (40) | Western blot analysis |
| ATP1B3 | 9 | KDELR3 | 20 | CYCS | 55 | CYCS affects the expression level of β cells through regulating the production of mitochondrial ROS. | (41) | Western blot analysis |
| PABPC3 | 7 | SCD | 19 | ALOX5 | 53 | ALOX5 can lead to inflammation in patients with T2DM. | (42) | Normal fasting glucose and normal glucose tolerance |
| AQR | 7 | PELO | 19 | CAT | 53 | CAT belongs to peroxidase, which can affect the oxidative metabolism of fatty acid. | (43) | Cell culture of human fibroblasts |

TABLE 3 The top 10 pairs with the greatest absolute difference frequency.

| Aging marker | Oxidative marker | Disease marker | Difference | Experimental results of the oxidative marker | Reference | Experimental method |
|---|---|---|---|---|---|---|
| OSBPL7 | COX7C | TM6SF1 | -0.039347869 | COX7C activity is associated with pancreatic β-cells. | (49) | OGTT testing |
| DNAJA3 | MYC | GSTZ1 | -0.037033525 | MYC is a key factor for proliferation of pancreatic β-cells. | (50) | Western blot analysis and real-time PCR |
| OSBPL7 | COX7C | SLC25A37 | -0.036323552 | COX7C activity is associated with pancreatic β-cells. | (49) | OGTT testing |
| OSBPL7 | MGAT3 | SF3A2 | -0.0357659 | MGAT3 plays role in lipid homeostasis. | (51) | Mouse model:oral administration of isoindoline-5-sulfonamide |
| TTC25 | COX7A1 | CMTM8 | -0.03019756 | COX7A1 activity is associated with pancreatic β-cells. | (49) | OGTT testing |
| OSBPL7 | MGAT3 | RECK | -0.027526704 | MGAT3 plays role in lipid homeostasis. | (51) | Mouse model:oral administration of isoindoline-5-sulfonamide |
| MTUS1 | ISCU | ATP5J | -0.019164871 | ISCU can cause Friedreich ataxia (FRDA), which is related to diabetes. | (52) | Cell culture of endocardium |
| SLC23A2 | GCH1 | SPI1 | 0.01780268 | GCH1 is related to endothelial dysfunction in T2DM. | (53) | Venous occlusion plethysmography |
| EPN1 | IL18BP | MRPL11 | 0.017333862 | IL18BP is related to inflammatory response, which plays important roles in diabetic nephropathy. | (54) | Cell culture of human proximal tubular epithelial and western blot analysis |
| EPN1 | PARK7 | NFKBIA | 0.014743576 | PARK7 participates in glucose homeostasis and then induces insulin resistance. | (55) | Quantitative PCR analysis and western blotting analyses |

(TM6SF1) participates in regulating transmembrane transport in macrophages (60). Overall, these results indicated that oxidative stress played an important role in the development of T2DM.

The top sensitive aging, disease, oxidative markers (evaluated by the occurrence times, also along with relative experimental details (61–69)) and are also shown in Table 4. For example, the top aging marker was HPS1 (Hermansky-Pudlak Syndrome 1 gene), inducing the biogenesis of lysosome-associated cellular organelles (70), which regulates the aging process through sphingolipids (71). The top disease marker was PPP1R15A (protein phosphatase 1 regulatory subunit 15A). PPP1R15A plays an important role in insulin resistance via energy metabolism (72, 73). The top oxidative marker was ATOX1 (antioxidant 1 copper chaperone). It has been reported that ATOX1 can regulate the copper level in the cell and maintain the redox balance as a defense antioxidant (74, 75). In short, the sensitivity analysis emphasized the crucial relationship among aging, oxidative stress and T2DM.

## 2.4 Underlying oxidative-aging mechanisms based on enrichment analysis

To further explore the underlying mechanisms between oxidative aging and T2DM, the shortest path between each pair of oxidative aging and disease markers was identified, and then enrichment analysis was performed based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and biological process (BP) terms in Gene Ontology (GO). As a result, relative enrichment results were summarized in Figures 3, S1, as well as Tables 5 (75–88), 6 (89–102) and S4 (76–81, 83–85, 103, 104, 111, 113), S5 (89–91, 94, 96, 103, 105–110).

The top 10 KEGG pathways are shown in Tables 5, S3. The most enriched KEGG pathway was “Parkinson’s Disease” (enriched in 1213 shortest paths). It has been reported that Parkinson’s disease (PD) and T2DM have common pathological mechanisms (76–78, 111). For example, oxidative stress and mitochondrial dysfunction are involved in both T2DM and PD pathogenesis (77). Strikingly, there are also a series of common biological pathways in T2DM, PD and cancer, such as mitochondrial dysfunction and protein homeostasis (112). Furthermore, the most significant KEGG pathway with the minimum FDR was “Leishmania Infection” (FDR=0.0000629) (Figure 3B), indicating the inflammatory response in the immune system (82, 113, 114). Notably, the inflammatory response is also often closely related to cancer (113). The classical aging pathway, the “mTOR signaling pathway” was also enriched in shortest pathway (Figure S2), indicating the interrelationship between oxidative aging and T2DM.

The top 10 BP terms are shown in Tables 6, S4. For example, the top enriched BP term was “Regulation of aerobic respiration” (enriched in 29 shortest paths), which was related to energy and

TABLE 4 The top 10 aging markers with the most paired with oxidative markers after sensitive analysis.

| Aging marker | Times | Disease marker | Times | Oxidative marker | Times | Experimental results of the oxidative marker | Reference | Experimental method |
|---|---|---|---|---|---|---|---|---|
| HPS1 | 13 | PPP1R15A | 18 | ATOX1 | 39 | ATOX1 can protect pancreatic β-cells and induce diabetes mellitus. | (61) | Western blot analysis |
| SCARB1 | 10 | ALDH4A1 | 16 | APEX1 | 31 | APEX1 is associated with diabetic retinopathy. | (62) | Western blot analysis |
| TMPO | 7 | G0S2 | 15 | APP | 29 | APP is related to protein accumulation, and then leads to T2DM. | (63) | *In vitro* aggregation assay |
| MRPL10 | 7 | CALML4 | 15 | ALDH3B1 | 27 | ALDH3B1 is related to lipid peroxidation. | (64) | Western blot analysis |
| TTC25 | 6 | STXBP2 | 15 | AXL | 25 | AXL is involved in diabetic vascular disease. | (65) | OGTT testing |
| MYLK | 6 | ZCCHC14 | 15 | AKT1 | 24 | AKT1 is related to insulin resistance. | (66) | Western blotting analysis and real-time PCR |
| RPS4Y1 | 5 | MCEE | 15 | ARNTL | 21 | ARNTL regulates lipid metabolism and diet-induced insulin resistance. | (67) | Plasma metabolites analysis |
| ESCIT | 5 | HIST1H2AC | 15 | ADAM9 | 20 | ADAM9 is a potential novel target for regulating the function of diabetic EPCs. | (68) | Western blotting |
| FKBP1B | 5 | PDLIM1 | 15 | ATRN | 20 | | | |
| PTPLB | 5 | MRPL18 | 14 | CAMKK2 | 20 | CAMKK2 plays role in diet-induced obesity, glucose intolerance and insulin resistance. | (69) | Immunoblotting |

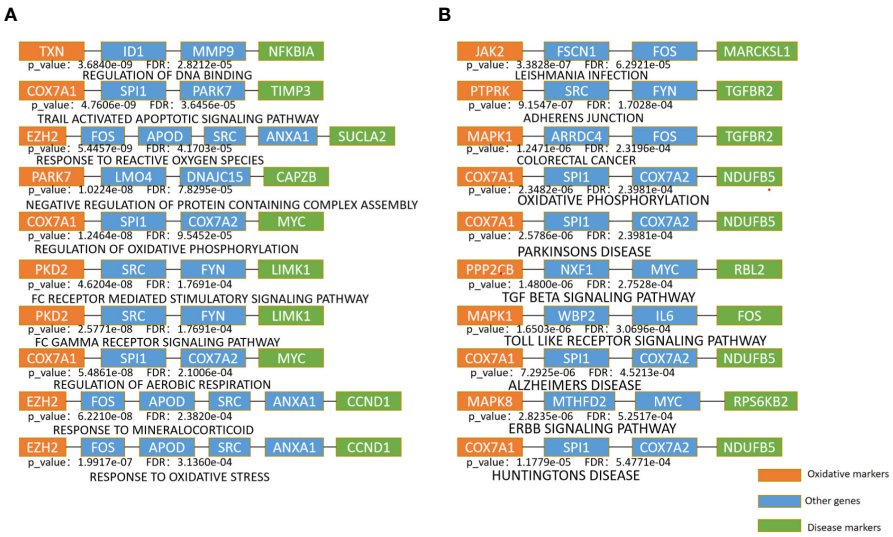



FIGURE 3
Enrichment analysis of the shortest path of KEGG and BP **(A)**. The top 10 pathway enrichment with the minimum FDR in BP terms **(B)**. The top 10 pathway enrichment with the minimum FDR in KEGG pathways.

mitochondrial function (96). In addition, reactive oxygen species (ROS) are byproducts of aerobic respiration that control various cellular functions (97). The BP term with the minimum FDR was "Regulation of DNA binding" (FDR=0.0000282) (Figure 3A), which is vital to T2DM by dysregulating mitochondria and energy metabolism (100). Obviously, the accumulation of DNA damage is also a hallmark of cancer (115). Overall, these results identified various aspects of risk factors for T2DM, such as oxidative stress, aging, energy metabolism and immune systems.

## 2.5 Network markers revealed key mechanisms between aging and T2DM

Network markers were identified by calculating the betweenness in the shortest path of each "oxidative-disease" pair, where the top markers are shown in Table 7. For example, the top network marker was SCD (stearyl-coenzyme A desaturase), which is mainly expressed in adipose tissue and can catalyze the synthesis of monounsaturated fatty acids (116). In addition, SCD can affect lipid metabolism and

TABLE 5 The top 10 enriched KEGG pathways.

| KEGG | Enriched shortest paths | Functions | Reference |
|---|---|---|---|
| PARKINSON DISEASE | 1213 | (1) T2DM and Parkinson Disease have shared pathological mechanism. (2) T2DM is a determinant of Parkinson Disease risk and progression. | (76–78) |
| OXIDATIVE PHOSPHORYLATION | 1175 | Causing metabolic alterations at the organism level through producing energy-rich molecules like ATP. | (87) |
| ALZHEIMERS DISEASE | 1128 | (1) T2DM is modifiable risk factor for Alzheimer's Disease. (2) Insulin resistance is a common mechanism between Alzheimer's Disease and T2DM. | (79, 80) |
| HUNTINGTONS DISEASE | 1115 | T2DM and Huntington's Disease have shared treatment method. | (81) |
| LEISHMANIA INFECTION | 702 | Related to the immune system. | (82) |
| CARDIAC MUSCLE CONTRACTION | 357 | Related to insulin sensitivity and mitochondrial function. | (88) |
| TOLL LIKE RECEPTOR SIGNALING PATHWAY | 239 | Producing and releasing various inflammatory mediators and triggering immune response. | (83) |
| COLORECTAL CANCER | 163 | T2DM is the risk factor for colorectal cancer. | (84) |
| ADHERENS JUNCTION | 145 | Regulating insulin vesicle trafficking. | (85) |
| T CELL RECEPTOR SIGNALING PATHWAY | 84 | Related to immune system. | (86) |

TABLE 6 The top 10 enriched BP terms.

| BP | Enriched shortest paths | Functions | Reference |
|---|---|---|---|
| RESPONSE TO REACTIVE OXYGEN SPECIES | 418 | (1) Modifying cell signaling proteins and then mediating T2DM. (2) As a central mechanism for the development of T2DM. | (89, 101) |
| RESPONSE TO OXIDATIVE STRESS | 308 | (1) Causing the function of pancreatic beta cells damaged. (2) Related to insulin resistance. | (90, 91) |
| CELLULAR RESPONSE TO REACTIVE OXYGEN SPECIES | 260 | (1) Maintaining the cellular redox homeostasis. (2) Related to mitochondrial oxidative stress and cell senescence. | (92, 93) |
| CELLULAR RESPONSE TO CHEMICAL STRESS | 179 | Regulating the cellular redox state. | (94) |
| RESPONSE TO OXYGEN CONTAINING COMPOUND | 142 | Controlling the intracellular metabolism and energy metabolism. | (95) |
| REGULATION OF AEROBIC RESPIRATION | 131 | (1) Regulating the level of glucose metabolism. (2) Reactive oxygen species (ROS) are a byproduct of aerobic respiration and signaling molecules, which controls various cellular functions. | (96, 97) |
| CELLULAR RESPONSE TO OXYGEN CONTAINING COMPOUND | 104 | Disorder of glucose and lipid metabolism is an important cause for the development of T2DM. | (102) |
| AEROBIC RESPIRATION | 101 | Regulating energy metabolism,and then affecting T2DM. | (98) |
| REGULATION OF GLYCOLYTIC PROCESS | 98 | Producing energy and inducing mitochondrial dysfunction and oxidative stress. | (99) |
| REGULATION OF DNA BINDING | 91 | Regulating the function of mitochondrial. | (100) |

mediate steroidogenesis, playing an important role in insulin resistance (117, 118). Furthermore, SCD participates in mediating the inflammatory reaction, which promotes the progression of cancer (119). Moreover, there were also a series of shortest paths through SITR1 (Figure S3, where permutation p-value=0.002 and 0, before and after sensitive analysis), which was as a clssical aging marker. Thus, network markers indicate the crucial role of oxidative stress dysfunction, along with energy metabolism, in T2DM.

TABLE 7 The top 10 genes with the highest number before and after sensitive analysis.

| Before sensitive analysis | | | After sensitive analysis | | |
|---|---|---|---|---|---|
| Gene Symbol | Betweenness | P-value | Gene Symbol | Betweenness | P-value |
| SCD | 3403 | 0 | SCD | 355 | 0 |
| MARCKSL1 | 2049 | 0 | MRPL11 | 325 | 0 |
| APOD | 1910 | 0 | ATOX1 | 300 | 0 |
| FOS | 1827 | 0 | COX7A2 | 223 | 0 |
| ATOX1 | 1462 | 0 | FOS | 212 | 0 |
| PCGF2 | 1288 | 0 | NENF | 164 | 0 |
| COX7A2 | 1266 | 0 | ISCU | 163 | 0 |
| MRPL11 | 1135 | 0 | COX4I1 | 151 | 0 |
| OGT | 1098 | 0 | HYAL2 | 146 | 0 |
| NDUFA8 | 10003 | 0 | MMP9 | 101 | 0 |

## 2.6 Pan-cancer analysis further verified the mechanism of oxidative aging in T2DM

Pan-cancer analysis was used to further verify the relative functions of T2DM oxidative aging markers in cancer. For example, oxidative aging markers in the integrated model were used to evaluate the survival index across different cancer types. There were 9 out of 15 cancer types with significant results (including BLCA, COAD, KIRC, KIRP, LUSC, PRAD, READ, THCA and UCEC, shown in Figure 4). These results suggest that oxidative aging markers can also be used as relative risk factors in cancer.

Additionally, both the commonality and specificity across 15 cancer types were investigated based on enrichment analysis. The top 10 common KEGG pathways are shown in Figures 5, S4, where "Alzheimer's Disease" was the top KEGG pathway. Alzheimer's disease (AD) and cancer share common risk factors. For example, aging is one of the greatest risk factors for the development of Alzheimer's disease, and the risk of cancer also increases with increasing age (120). In addition, some cancer patients may have a higher risk of Alzheimer's disease (121). Figures 6, S2 showed the top 10 common BP terms in 15 cancers. "Regulation of cellular respiration" was the top BP term, indicating the key role of energy metabolism in cancer (122). Cellular respiration participates in energy metabolism and is also a hallmark of many cancers (123). The specific enrichment results within each cancer are also summarized in Tables 8, 9, S6, S7 (112, 120–163), indicating a series of oxidative aging-related risk factors in cancer, such as the

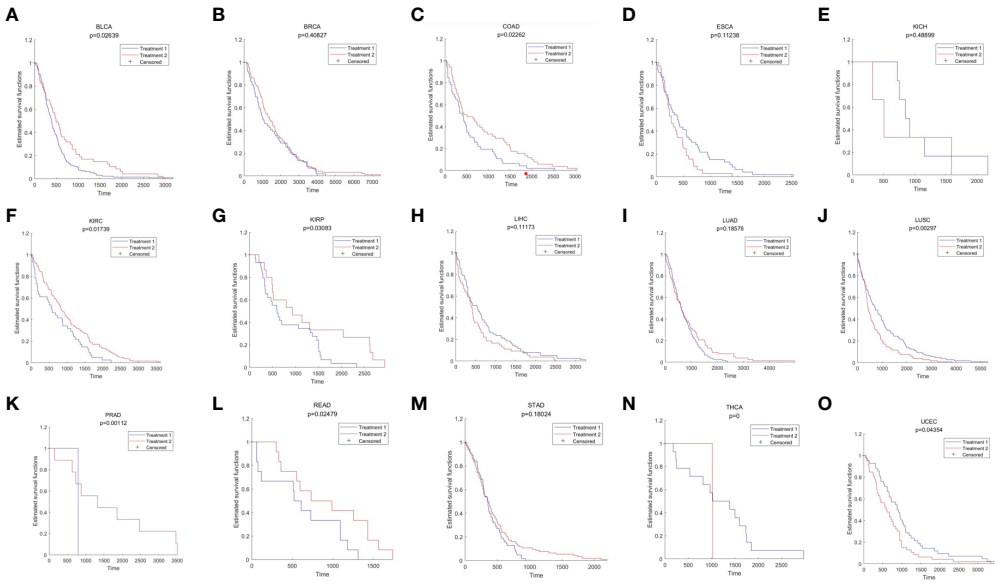



FIGURE 4
The results of survival analysis across different cancer types. **(A)** BLCA; **(B)** BRCA; **(C)** COAD; **(D)** ESCA; **(E)** KICH; **(F)** KIRC; **(G)** KIRP; **(H)** LIHC; **(I)** LUAD; **(J)** LUSC; **(K)** PRAD; **(L)** READ; **(M)** STAD; **(N)** THCA; **(O)** LIHC.

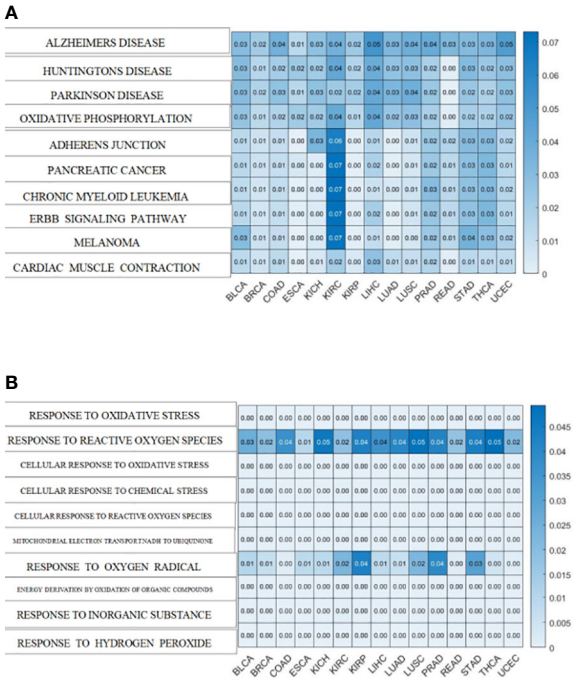



FIGURE 5
The enrichment analysis shared by cancers. **(A)**KEGG pathways enriched in 15 cancers **(B)**. BP terms enriched in 15 cancers.

inflammatory response, energy metabolism and mitochondrial function. Overall, our results highlighted a series of crucial functions related to oxidative aging, which can also be used to study potential mechanisms in cancer.

# 3 Discussion

It is well known that aging-related oxidative stress plays a crucial role in T2DM (3). However, the essential relationship among aging, oxidative stress and T2DM still needs to be explored in more depth. In this paper, a series of computational methods were performed to explore these relationships in T2DM as well as the relative mechanisms. First, both the aging model and disease model were optimized, and relative aging markers and disease markers were identified. Next, the integrated oxidative aging model was built to identify essential "aging-oxidative-disease" relationships. Finally, network analysis, enrichment analysis, sensitivity analysis and pan-cancer analysis were used to further explore the potential mechanisms between oxidative aging and T2DM. As a result, various risk factors in T2DM were integrated.

Our results highlighted that energy metabolism was vital to the development of T2DM. For example, the integrated oxidative aging model identified a series of key markers in T2DM that were closely related to energy metabolism. OSBPL1A and T1GD4 participate in nutritional metabolism; the former is mainly involved in lipid metabolism and cholesterol metabolism, and the latter is mainly related to glycogen metabolism (32–34). ADPRH and PPP1R15A can lead to energy metabolism imbalance (35, 63). COX5A can affect mitochondrial function, and ATOX1 is the redox catalyst, both of which can affect energy metabolism through mitochondrial dysfunction (39, 65). Furthermore, as the top network marker, SCD is mainly expressed in adipose tissue and can catalyze the synthesis of monounsaturated fatty acids (116). It can affect lipid metabolism and mediate steroidogenesis, which plays an important role in insulin resistance (117, 118). SIRT1 was also identified by calculating the betweenness. In MCMC, the greatest difference in the absolute value pair was "OSBPL7-COX7C-TM6SF1", where OSBPL7 participates in lipid binding and transport (49, 50) and COX7C is related to cellular respiration as a potential biomarker of diabetes (51, 52). The classical energy metabolism pathway, "mTOR signaling pathway", was also identified using the enrichment analysis, indicating the key interaction between oxidative aging and T2DM.

Protein homeostasis is also involved in the progression of T2DM. For instance, amyloid precursor protein (APP) is an

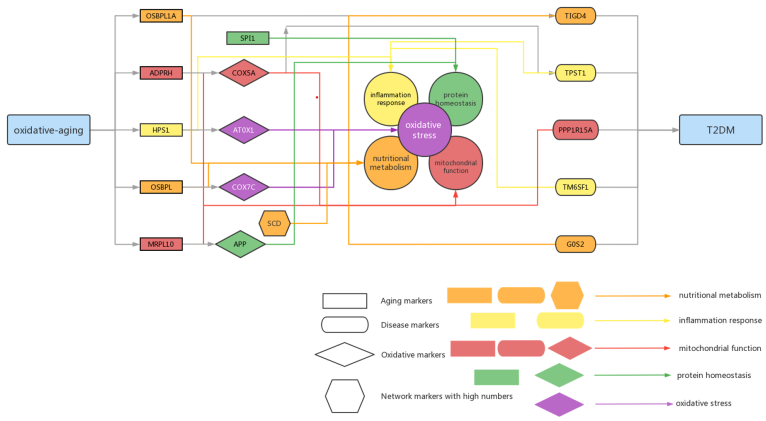



FIGURE 6
Summarized mechanisms of oxidative-aging in T2DM Rectangle genes represent aging markers, oval genes represent disease markers, rhombus genes represent oxidative markers, hexagon genes represent network markers with high numbers. Orange arrows indicate the gene involved in nutritional metabolism, yellow arrows indicate the gene involved in inflammation response, red arrows indicate the gene associated with mitochondrial function, green arrows indicate the gene associated with protein homeostasis.

10.3389/fendo.2023.1196293

TABLE 8 KEGG pathways in each cancer with the minimum FDR.

| Type of cancer | FDR | KEGG | Functions | Reference |
|---|---|---|---|---|
| BLCA | 8.72e-05 | ERBB SIGNALING PATHWAY | Related to human cancer pathogenesis. | (124) |
| BLCA | 8.72e-05 | PROGESTERONE MEDIATED OOCYTE MATURATION | The source of immune cells and macrophages. | (125) |
| BLCA | 8.72e-05 | PANCREATIC CANCER | A fatal malignancy with an aggressive disease course. | (126) |
| BRCA | 7.82e-06 | INTESTINAL IMMUNE NETWORK FOR IGA PRODUCTION | Related to immune system. | (127) |
| COAD | 9.79e-06 | HUNTINGTONS DISEASE | Cancer and Huntington's Disease have common pathogenesis. | (128) |
| ESCA | 1.08e-07 | PARKINSONS DISEASE | Parkinson Disease and cancer share some common biological pathways, such as mitochondrial dysfunction and protein homeostasis. | (112) |
| KICH | 7.64e-07 | OXIDATIVE PHOSPHORYLATION | Cancer cells utilize certain pathways to enhance oxidative phosphorylation. | (129) |
| KICH | 7.64e-07 | PARKINSONS DISEASE | Parkinson Disease and cancer share some common biological pathways, such as mitochondrial dysfunction and protein homeostasis. | (112) |
| KIRC | 1.47e-05 | RENAL CELL CARCINOMA | Main factor contributed to kidney cancer. | (130) |
| KIRC | 1.47e-05 | MELANOMA | The most lethal form of skin cancer. | (131) |
| KIRP | 4.32e-07 | ALZHEIMERS DISEASE | (1) age is the risk factor for the development of Alzheimer's Disease and cancer. (2) some cancer patients may have a higher risk of Alzheimer's Disease. | (120, 121) |
| LIHC | 1.31e-04 | ERBB SIGNALING PATHWAY | Related to human cancer pathogenesis. | (124) |
| LIHC | 1.31e-04 | PANCREATIC CANCER | A fatal malignancy with an aggressive disease course. | (126) |
| LUAD | 5.99e-08 | OXIDATIVE PHOSPHORYLATION | Cancer cells utilize certain pathways to enhance oxidative phosphorylation. | (129) |
| LUAD | 5.99e-08 | PARKINSONS DISEASE | Parkinson Disease and cancer share some common biological pathways, such as mitochondrial dysfunction and protein homeostasis. | (112) |
| LUSC | 2.54e-05 | BLADDER CANCER | The ninth most common malignancy worldwide. | (132) |
| PRAD | 2.13e-04 | OXIDATIVE PHOSPHORYLATION | Cancer cells utilize certain pathways to enhance oxidative phosphorylation. | (129) |
| PRAD | 2.13e-04 | PARKINSONS DISEASE | Parkinson Disease and cancer share some common biological pathways, such as mitochondrial dysfunction and protein homeostasis. | (112) |
| READ | 4.88e-05 | ADHERENS JUNCTION | Downregulation of E-cadherin, the two major components of adherens junctions, and p120, is a frequently recurrent hallmark of carcinomas. | (133) |
| READ | 4.88e-05 | GLIOMA | The most malignant and aggressive form of brain tumors, accounting for the majority of brain cancer-related deaths. | (134) |
| READ | 4.88e-05 | MELANOMA | The most lethal form of skin cancer. | (131) |
| STAD | 2.94e-05 | MELANOMA | The most lethal form of skin cancer. | (131) |
| THCA | 5.08e-05 | GAP JUNCTION | Genetic or acquired alterations of connexin proteins have been implicated in cancer. | (135) |
| UCEC | 6.41e-06 | ALZHEIMERS DISEASE | (1) age is the risk factor for the development of Alzheimer's Disease and cancer. (2) some cancer patients may have a higher risk of Alzheimer's Disease. | (120, 121) |

TABLE 9 BP terms in each cancer with the minimum FDR.

| Type of cancer | FDR | BP | Functions | Reference |
|---|---|---|---|---|
| BLCA | 7.53e-07 | NEGATIVE REGULATION OF INSULIN SECRETION INVOLVED IN CELLULAR RESPONSE TO GLUCOSE STIMULUS | Creating conditions that force cancer cells to rely more on metabolites and limited factors. | (136) |
| BRCA | 2.09e-05 | REGULATION OF OXIDATIVE PHOSPHORYLATION | Playing a crucial role in cancer progression. | (137) |
| BRCA | 2.09e-05 | RESPONSE TO HEPATOCYTE GROWTH FACTOR | The Cancer cell growth, survival, and migration of cancer cell are relied on an HGF-dependent manner. | (138) |
| COAD | 9.24e-07 | CELLULAR RESPONSE TO CADMIUM ION | Cadmium is an established carcinogen in both humans and animals. | (139) |
| ESCA | 9.12e-07 | NEGATIVE REGULATION OF PROTEIN CATABOLIC PROCESS | Playing dual roles in tumorigenesis and cancer progression. | (140) |
| KICH | 2.36e-06 | RESPONSE TO HYDROGEN PEROXIDE | The progression of cancer is related to effect of hydrogen peroxide. | (141) |
| KIRC | 2.86e-05 | RESPONSE TO IMMOBILIZATION STRESS | Enhancing the ability of some cancer cells to enter a dormant state. | (142) |
| KIRP | 1.76e-06 | MITOCHONDRIAL ELECTRON TRANSPORT NADH TO UBIQUINONE | Cancer cell propagation is closely related to the regulation of the electron transport chain. | (143) |
| LIHC | 3.87e-07 | CELLULAR RESPONSE TO HYDROGEN PEROXIDE | Regulating catalase expression to target the redox state of cancer cells. | (144) |
| LUAD | 2.37e-07 | ELECTRON TRANSPORT CHAIN | Electrons originating from different metabolic processes are guided into the mitochondrial electron transport chain (ETC) to drive the oxidative phosphorylation process. | (145) |
| LUSC | 2.74e-05 | CELLULAR RESPIRATION | Tumors gain energy mainly from glucose to lactate and only partially through cellular respiration involving oxygen. | (146) |
| PRAD | 3.42e-04 | RESPONSE TO OXIDATIVE STRESS | Related to cancer, which can regulate the progression of cancer. | (147) |
| READ | 2.02e-05 | CELLULAR RESPONSE TO REACTIVE OXYGEN SPECIES | ROS dynamically affect the tumor microenvironment, and are known to initiate cancer angiogenesis, metastasis, and survival at various concentrations. | (148) |
| STAD | 6.00e-05 | POSITIVE REGULATION OF CYTOSOLIC CALCIUM ION CONCENTRATION | Cancer cell proliferation and apoptosis depend on the intracellular Ca (2+) concentration. | (149) |
| THCA | 7.46e-05 | REGULATION OF NUCLEOCYTOPLASMIC TRANSPORT | The nucleocytoplasmic transport of macromolecules is critical for both cellular physiology and pathology, playing an important role in the treatment of cancer. | (150) |
| UCEC | 2.10e-07 | AEROBIC RESPIRATION | Alterations in cancer glucose metabolism include leading to a shift in metabolism from aerobic respiration to glycolysis. | (151) |

oxidative marker identified by MCMC that promotes the secretion of amyloid proteins (164). SPI1 (Spi-1 Proto-Oncogene) was involved in the negative regulation of protein, which caused restraint of aerobic glycolysis (165) (Figure 3). In summary, both APP and SPI1 are related to protein homeostasis and even accelerate the development of both T2DM and neurodegenerative diseases (NDs). That is, protein homeostasis is a common mechanism in both T2DM and ND (166, 167).

The inflammatory response also plays an important role in the development of T2DM. For example, the aging marker HPS1 affects the biogenesis of lysosome-associated cellular organelles and even participates in regulating cellular inflammation (61, 62). The disease marker TPST1 induces the secretion of some cytokines, along with the inflammatory response (37, 38). TM6SF1, as one of the key markers identified by MCMC, was involved in transmembrane transport in macrophages, thus highlighting the key role of the immune system in T2DM (53).

Furthermore, there are a series of experiments and relative clinical stastic results also revealed significant relationships between the identified oxidative aging markers and T2DM. For example, it has been reported that *in vitro* oxidative stress in mammalian skeletal muscle leads to substantial insulin resistance to distal insulin signaling and glucose transport activity (p=9.2e-05) (168). Chronic oxidative stress can also leads to decreased responsiveness to insulin, ultimately leading to diabetes reported by Alina Berdichevsky et al (p=0.01) (169). Besides, NFKBIA affects the

wound healing in diabetic foot ulceration (DFU) (p=0.006) (170), MYC and SCD are related to pyroptosis and immune infiltration in T2DM (p=0.001) (171). The experiment of Parker C. Wilson et al using single-nucleus RNA sequencing has been revealed that GCH1 is associated with early-stage diabetic nephropathy (p=4.88e-09) (172) In short, our results also presented key clinical indices with the help of the integrated oxidative model.

T2DM is associated with an increased risk of developing cancers, such as COAD, PRAD, and THCA (30). It is well known that T2DM and cancer have common risk factors, such as oxidative stress, energy metabolism, inflammation and protein homeostasis (22, 23, 173). Our results also proved that inflammation and energy metabolism were common risk factors in cancers, and even survival analysis further verified the key role of oxidative aging markers across different cancer types. Oxidative stress may lead to chronic inflammation, which in turn can induce most chronic diseases, including both cancer and T2DM. In addition, oxidative stress can damage the normal function of mitochondria as well as energy metabolism, which plays an important role in the development of T2DM and cancer. In short, various risk factors related to oxidative aging were also confirmed in cancer.

According to the oxi-inflamm-aging theory, the aging process is regulated by chronic oxidative stress, as well as the inflammatory response (174). It is well known that dysregulated oxidative stress triggers a series of signaling pathways, thus leading to pancreatic beta cell damage (175). In addition, the cellular senescence theory also highlights cellular inflammation and the oxidative stress response during the aging process (176, 177). That is, cellular senescence may also play an important role in the pathogenesis of T2DM (i.e., through the mTOR signaling pathway) (177, 178). Furthermore, these risk factors even interact with each other and then promote T2DM. For example, the imbalance of energy metabolism could interact with a series of pathways, such as lipid accumulation, chronic inflammation and insulin resistance, triggering T2DM progression (179). It has been reported that normal homeostasis in the insulin-driven immunometabolic network is vital to the preservation of insulin sensitivity in healthy aging (180). Here, our work also highlighted the interaction between the immune system and energy metabolism in the development of T2DM (Figure 3; Tables 5, 6), which is also crucial in cancer (Figures 4, 5). With the help of the integrated oxidative aging model, our study revealed that oxidative stress was interrelated with various aging-related risk factors in T2DM (Tables 2–6), such as the inflammatory response, mitochondrial function and protein homeostasis. These results further confirmed both the oxi-inflamm-aging and cellular senescence theories. Overall, potential aging-related mechanisms in T2DM were integrated in the context of oxidative stress (Figure 6).

# 4 Conclusion

In this study, machine learning was performed to predict aging and T2DM, and then relative biomarkers were identified. An integrated oxidative aging model was built to explore the essential relationship between oxidative aging and T2DM. The key roles of nutritional metabolism, the inflammatory response, mitochondrial function and protein homeostasis in T2DM were highlighted in our work with the help of sensitivity analysis, enrichment analysis, network analysis and pan-cancer analysis. In conclusion, various risk factors were integrated in the development of T2DM as well as cancer based on oxidative aging.

# 5 Materials and methods

## 5.1 Data and preprocessing

All gene expression data were downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), including GSE362, GSE15790, GSE18732, GSE29221, GSE29226, GSE29231, GSE37171, GSE38642, GSE76894, and GSE182120. These datasets were from eight different platforms: GPL96, GPL97, GPL8450, GPL9486, GPL6947, GPL570, GPL6244, and GPL17586.

The gene expression profiles were processed as follows:

(1) Only the samples with both the age and phenotype index (i.e., type 2 diabetes versus control) were retained; otherwise, they were deleted.

(2) The gene expression matrix for each dataset was integrated by summarizing the probe number within the gene symbol.

(3) The total data matrix was integrated, and the missing gene expression values were filled with values of 0.

(4) Genes with missing values ≥ 30% were deleted.

(5) The gene expression matrix was transformed by logarithmic transformation if it contained outliers.

(6) Based on the mean and the standard deviation of gene expression for control individuals, the z-score normalization was performed for both T2DM and control samples.

(7) The singular value decomposition (SVD) method was performed to eliminate the intersample variation based on the top three principal components of the control samples.

(8) The z score was then utilized to normalize all samples based on the mean and the standard deviation of the control samples.

(9) The training set and the test set were randomly divided according to a ratio of approximately 2:1.

As a result, a total of 489 samples were obtained, including 208 samples of healthy aged people (age > 50 years old, 145 training datasets + 63 test datasets), 131 samples of healthy young people (age ≤ 50, 90 + 41), 110 samples of T2DM aged people (age > 50, 75 + 35) and 40 samples of T2DM young people (age ≤ 50, 25 + 15), containing 12958 gene symbols (Tables S1–S3).

We also obtained paired gene expression (RNAseq) profiles ("Batch effects normalized mRNA data") and clinical data from the TCGA database through the xena platform (https://xenabrowser.net/hub/). Cancer types with ≥10 adjacent normal samples were retained. As a result, there were 15 cancer types used in this work: BLCA (408 cancer samples and 19 adjacent

normal samples), BRCA(1102 + 113), COAD(451 + 41), ESCA (185 + 11), KICH(66 + 25), KIRC(534 + 72), KIRP(291 + 32), LIHC (376 + 50), LUAD(517 + 59), LUSC(504 + 51), PRAD(498 + 52), READ(160 + 10), STAD(414 + 35), THCA(513 + 59) and UCEC (533 + 22). The tumor expression profiles from the same patient were averaged. Genes with missing values ≥30% were deleted.

## 5.2 Modeling the aging model and disease model

After randomization as well as a random disorder, the healthy population samples were divided into a training dataset and a test dataset. The ratio of training dataset samples to test dataset samples was close to 2:1. The ReliefF algorithm was used to select key features, and then the first 500 models were studied to train predictors. The optimal model was selected by 10-fold cross-validation. To verify the accuracy of the aging predictor, the selected model was verified in the test dataset.

(1) In the aging model, the normal aged group (age>50) was labeled 1, and the young healthy group (age ≤ 50) was labeled 0; in the disease model, the T2DM group was labeled 1, and the control group (age ≤ 50) was labeled 0.

(2) The 12958 genes were sorted by the ReliefF algorithm;

(3) The predictor was generated using the k-nearest neighbor (kNN, k=3, correlation distance) algorithm. The optimal model was selected by 10-fold cross-validation, where the model with the highest accuracy rate was chosen.

(4) The identified features were considered aging and disease markers. As a result, 304 aging markers and 299 disease markers were identified.

## 5.2 Identifying essential relationships in T2DM by an integrated oxidative aging model

The integrated oxidative aging model was built to identify the essential relationship among aging, oxidative stress and T2DM. The computational pipeline was referred to by Mendelian randomization (MR), although it was not as strict as MR (Figure 1C).

In this model, the aging-related oxidative stress markers were considered oxidative aging markers, where the relative aging/disease markers were identified in "Methods 5.2". As a result, the essential relationships among aging, oxidative stress and disease (T2DM) markers were identified as key "aging-oxidative-disease" triples in T2DM.

MR is a statistical method for assessing the causal relationship between risk factors and outcomes based on observational data (181, 182). The causal relationships between the instrumental variables, risk factors, and outcome variables were assessed as follows.

(1) There was a correlation between the instrumental variable and the risk factor.

(2) There was no correlation between the instrumental variable and the confounding factor.

(3) There was no correlation between the instrumental variable and the outcome variable after deleting the effect from the risk factor.

Here, the aging marker was used as the auxiliary variable (similar to the instrumental variable in MR), and the oxidative stress markers were used as the candidate risk factor. Then, aging-related oxidative ("oxidative aging") markers were identified as the risk factor, and disease markers were used as the outcome variable. That is, the integrated oxidative aging model aimed to explore essential relationships among aging, oxidative stress and disease markers in T2DM. This model was performed as follows:

(1) Oxidative markers were obtained as candidate risk factors based on Biological Processes (BP) of Gene Ontology (GO) through the Gene Set Enrichment Analysis (GSEA) platform (http://www.gsea-msigdb.org/gsea/downloads.jsp, "OXIDATIVE" was taken as the keyword). As a result, 310 candidate oxidative markers were selected.

(2) The correlation (differential coexpression) pattern was used to select aging markers that strongly correlated with candidate oxidative stress markers with the help of the Kruskal–Wallis test. Here, the differential coexpression was calculated as follows:

$$p = \text{Kruskal} - \text{Wallis test}(\text{aging\_marker} * \text{oxidative\_marker}, \text{phenotype}) \quad (1)$$

where the phenotype could be defined as 1 (T2DM) and 0 (control).

Furthermore, both a p-value<0.05 and Benjamini–Hochberg false discovery rate (FDR)<0.1 were used to select strongly correlated aging markers.

(3) To reduce the correlation between the auxiliary variable (aging marker) and confounding factors, as well as further select a strong correlation between the aging marker and the candidate oxidative marker, a permutation test was performed by generating the simulated aging markers from the same number of randomly selected markers to each candidate oxidative marker; this process was repeated 1000 times, and then the p-value was calculated as the proportion of occurrence times (larger than the real mean difference) of the absolute difference between T2DM and control in 1000 permutations. The relationship between each aging marker and the candidate oxidative marker was retained if the permutation P<0.05.

(4) Correlation (differential coexpression) was used to select oxidative markers that strongly correlated with disease markers with the help of the Kruskal–Wallis test. Here, the differential coexpression was calculated as follows:

$$p = \text{Kruskal} - \text{Wallis test}(\text{oxidative\_marker.} * \text{disease\_marker}, \text{phenotype}) \quad (2)$$

where the phenotype could be defined as 1 (T2DM) and 0 (control).

Furthermore, both a p-value<0.05 and Benjamini–Hochberg false discovery rate (FDR)<0.1 were used to select strongly correlated oxidative markers.

(5) To reduce the correlation between the risk factor (oxidative marker) and confounding factors, as well as further select a strong correlation between the oxidative marker and the disease marker, a permutation test was performed by generating the simulated oxidative markers from the same number of randomly selected markers to each disease marker; this process was repeated 1000 times, and then the p-value was calculated as the proportion of occurrence times (larger than the real mean difference) of the absolute difference between T2DM and control in 1000 permutations. The relationship between each aging marker and the candidate oxidative marker was retained if the permutation P<0.05.

(6) The direct relationships for any other factors (genes) were found to reduce the correlation between the auxiliary variable (aging marker) and confounding factors. If there was another factor (gene) that was directly correlated (differentially coexpressed) to both the aging marker and the disease marker, then the relationship from aging to disease was deleted.

$$p = \text{Kruskal} - \text{Wallis test}(\text{aging\_marker.} * \text{other\_gene}, \text{phenotype}) \quad (3)$$

$$p = \text{Kruskal} - \text{Wallis test}(\text{disease\_marker.} * \text{other\_gene}, \text{phenotype}) \quad (4)$$

where the phenotype could be defined as 1 (T2DM) and 0 (control).

Furthermore, both a p-value<0.05 and Benjamini–Hochberg false discovery rate (FDR)<0.1 were used to filter out any direct relationships.

(7) To filter out the effect of horizontal pleiotropy, the aging–disease relationship was further examined by comparing the correlation between each aging and disease marker, through the oxidative marker or otherwise. Herein, steps ①–③ were used to calculate the correlations between auxiliary variables and outcome variables without the background of the risk factor, and step ④ was used to calculate the correlations between auxiliary variables and outcome variables with the context of the risk factor.

① The residual of each disease marker ("residual A") was calculated based on the oxidative marker:

$$\text{residual\_}A = disease\_mar\ker - b_1 * oxidative\_mar\ker \quad (5)$$

where $b_1$ is the regression coefficient.

② The residual of each aging marker ("residual B") was calculated based on the oxidative marker:

$$\text{residual\_}B = aging\_mar\ker - b_2 * oxidative\_mar\ker \quad (6)$$

where $b_2$ is the regression coefficient.

③ The abovementioned two residuals were further compared, and the residual of the disease marker was calculated (as "residual C"):

$$\text{residual\_}C = residual\_A - b_3 * residual - \_B \quad (7)$$

where $b_3$ is the regression coefficient.

④ The residual of the disease marker ("residual D") was calculated based on the aging marker.

⑤ The difference (between "residual C" and "residual D") was tested between the T2DM and control subgroups using the Kruskal–Wallis test (P<0.05 and FDR<0.1).

Finally, the essential relationship among the aging marker, oxidative marker and disease marker was retained. Thus, 11829 "aging-oxidative-disease" triples were identified, including 105 aging markers, 83 oxidative markers and 282 disease markers. Thus, these 83 oxidative markers were used as oxidative aging markers (risk factors), and 282 disease markers were also used to discriminate the T2DM phenotype.

## 5.4 Sensitivity analysis using the MCMC method

To further explore the relationship among aging, oxidative stress and T2DM, sensitivity analysis was performed based on the Markov chain Monte Carlo (MCMC) method, where "aging-oxidative-disease" triples identified by MR were further evaluated as a candidate relationship. The MCMC method is used to sample certain posterior distributions in a high-dimensional space based on a given probabilistic background. The key step of MCMC is to construct a Markov chain whose equilibrium distribution is equal to the target probability distribution. The steps were as follows:

(1) Constructing the transfer cores of the ergodic Markov chain. The prior distribution of each parameter was normally distributed based on all identified markers in each group (i.e., T2DM and control), respectively.

(2) Simulate the chains until equilibrium is reached. The Metropolis–Hastings sampling method was used to determine whether the new sample ($\theta^*$) was acceptable based on the $\alpha$ value.

$$\alpha = \frac{P(\theta^* | X) * q(\theta^n \rightarrow \theta^*)}{P(\theta^n | X) * q(\theta^n \rightarrow \theta^*)} \quad (8)$$

where $P(\theta^n | X)$ and $P(\theta^* | X)$ are the posterior probability of the nth accepted sample, the new sample $q(\theta^n \rightarrow \theta^*)$ is the transition probability from the nth accepted sample to the new sample, and $q(\theta^* \rightarrow \theta^n)$ is the transition probability from the new sample to the *n-th* accepted sample.

In this work, the disease score was used to evaluate the simulated samples, with 1000 random samples used as candidate samples for each group (i.e., T2DM or control). The disease score was calculated by comparing the distance between normal and T2DM training samples based on the 282 disease markers identified by the integrated oxidative aging model:

$$\begin{aligned} &\text{disease\_score} \\ &= \sum_{k=1}^{7} \text{distance\_of\_neareast\_neighbour\_in\_control} \\ &- \sum_{k=1}^{7} \text{distance\_of\_neareast\_neighbour\_in\_T2DM} \end{aligned} \quad (9)$$

(3) Performing the global sensitivity analysis

The correlation index was used to evaluate each "aging-oxidative-disease" triple in the accepted samples (including both T2DM and control):

$$\text{correlation\_index} = \frac{\text{disease\_marker} - \text{aging\_marker}}{\text{oxidative\_marker} - \text{aging\_marker}} \quad (10)$$

As a result, the correlation index was calculated in each "aging-oxidative-disease" triple for all accepted samples. Then, the Kruskal–Wallis test was used to evaluate each correlation index in each "aging-oxidative-disease" triple, where p-value<0.05 and FDR<0.1 were set as the threshold. Finally, 2501 "aging-oxidative-disease" triples were identified as sensitive relationships, including 41 aging markers, 37 oxidative markers and 61 disease markers.

## 5.5 Constructing the differential coexpression network

To further reveal the relationship between "oxidative aging" and T2DM, a differential coexpression network was constructed by the following steps:

(1) The Pearson correlation coefficient for each pair of genes was calculated based on the T2DM and control groups.
(2) The Benjamini–Hochberg FDR method was used to adjust the p-values of the correlation coefficient.
(3) The relationship between each gene pair was retained if the coefficient value in T2DM had the opposite sign (i.e., + or -) to that in control, as well as p< 0.05 and FDR< 0.1.
(4) The shortest path between each pair of oxidative aging and disease markers was selected based on the differential coexpression network using the Dijkstra algorithm.

## 5.6 Enrichment analysis

The gene functions were further explored by enrichment analysis of the shortest pathway. Gene Ontology (GO) terms and KEGG pathways for the GSEA platform were obtained from gene set enrichment analysis (http://software.broadinstitute.org/gsea/downloads.jsp, version 7.5). The hypergeometric distribution was used to test the degree of enrichment of the GO BP and KEGG pathways. Hypergeometric test formula:

$$P(X \geq x) = 1 - \sum_{k=0}^{x-1} \frac{C_M^k \times C_{N-M}^{n-k}}{C_M^k} \quad (11)$$

where $N$ is the total number of genes in the gene set, $M$ is the number of known genes (such as KEGG pathway or BP terms), which is the number of genes identified in each shortest pathway, and $k$ is the number of common genes between known genes and candidate genes identified in each "oxidative-disease" shortest pathway. The p-value of each path was controlled using the Benjamin-Hochberg method. Finally, pathways with p<0.05 and FDR<0.1 were retained.

## 5.7 Identifying network markers

The subnetwork with the shortest pathways among the selected "oxidative-disease" pairs was constructed, and genes in the subnetwork were sorted by their betweennesses in descending order. To test whether the top betweenness genes were hubs in the background network, we ran a permutation to count the occurrence time of the top genes in the shortest paths between randomly selected genes (containing the same numbers of "oxidative-disease" pairs, based on the identified "aging-oxidative-disease" triples) when they had greater betweennesses than those in our study. We repeated this process 1000 times, and the p-value was calculated as the proportion of occurrence times of the top betweenness genes in 1000 permutations.

## 5.8 Pan-cancer analysis

The survival analysis was performed based on the oxidative aging markers (identified by the integrated oxidative aging model in 5.3) for each cancer using the Kaplan–Meier method. The tumor samples of each cancer were divided into two groups based on the mean value of the oxidative aging markers. Then, the Kaplan–Meier method was used to evaluate the survival difference between these two groups, and the significance was estimated by the log-rank test. A p-value<0.05 was considered statistically significant.

Genes were considered differentially expressed if they satisfied the following criteria:

(1) Fold change>2;
(2) p-value<0.05 in the Kruskal–Wallis test;
(3) Benjamin-Hochberg false discovery rate (FDR)<0.1.

Then, the differential expression networks were constructed for each cancer, where the details were also the same as 5.5. As a result, each shoreat pathway was selected from each pair of oxidative aging markers and differentially expressed genes (as disease markers in cancer) using the Dijkstra algorithm. Furthermore, enrichment analysis was performed by the "oxidative-disease" shortest

pathway for each cancer type, where both $p<0.05$ and FDR<0.1 were used.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

## Author contributions

LX, LL, and YW designed the study. YC, LY, SZ and YW analyzed the data. YC, LY and YW interpreted the results. YC, MX, SR and YW visualized the results. All authors wrote and revised the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by National Natural Science Foundation of China (32000478 to YW) and the Shanghai Municipal Health Commission and Collaborative Innovation Cluster Project (No. 2019CXJQ02), the National Key R&D Program of China (No. 2018YFA0107800), the National Natural Science Foundation of China (No. 81974010), the Provincial Natural Science Foundation of Hunan Province (No. 2021JJ40963). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Conflict of interest

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.

## Publisher's note



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1196293/full#supplementary-material

SUPPLEMENTARY TABLE 1
The detailed datasets used in this work.

SUPPLEMENTARY TABLE 2
The gene symbols used in this work.

## References

1. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* (2022) 183:109119. doi: 10.1016/j.diabres.2021.109119

2. Artasensi A, Pedretti A, Vistoli G, Fumagalli L. Type 2 diabetes mellitus: a review of multi-target drugs. *Molecules* (2020) 25(8):1987. doi: 10.3390/molecules25081987

3. Gunasekaran U, Gannon M. Type 2 diabetes and the aging pancreatic beta cell. *Aging (Albany NY).* (2011) 3(6):565–75. doi: 10.18632/aging.100350

4. Pengpid S, Peltzer K. Prevalence and correlates of undiagnosed, diagnosed, and total type 2 diabetes among adults in Morocco, 2017. *Sci Rep* (2022) 12(1):16092. doi: 10.1038/s41598-022-20368-4

5. Jura M, Kozak LP. Obesity and related consequences to ageing. *Age (Dordr).* (2016) 38(1):23. doi: 10.1007/s11357-016-9884-3

6. van den Beld AW, Kaufman JM, Zillikens MC, Lamberts SWJ, Egan JM, van der Lely AJ. The physiology of endocrine systems with ageing. *Lancet Diabetes Endocrinol* (2018) 6(8):647–58. doi: 10.1016/S2213-8587(18)30026-3

7. Diniz Pereira J, Gomes Fraga V, Morais Santos AL, Carvalho MDG, Caramelli P, Braga Gomes K. Alzheimer's disease and type 2 diabetes mellitus: a systematic review of proteomic studies. *J Neurochem* (2021) 156(6):753–76. doi: 10.1111/jnc.15166

8. Dal Canto E, Ceriello A, Rydén L, Ferrini M, Hansen TB, Schnell O, et al. Diabetes as a cardiovascular risk factor: an overview of global trends of macro and micro vascular complications. *Eur J Prev Cardiol* (2019) 26(2_suppl):25–32. doi: 10.1177/2047487319878371

9. van Eersel ME, Joosten H, Gansevoort RT, Dullaart RP, Slaets JP, Izaks GJ. The interaction of age and type 2 diabetes on executive function and memory in persons aged 35 years or older. *PloS One* (2013) 8(12):e82991. doi: 10.1371/journal.pone.0082991

10. Gallagher EJ, LeRoith D. Obesity and diabetes: the increased risk of cancer and cancer-related mortality. *Physiol Rev* (2015) 95(3):727–48. doi: 10.1152/physrev.00030.2014

11. Frijhoff J, Winyard PG, Zarkovic N, Davies SS, Stocker R, Cheng D, et al. Clinical relevance of biomarkers of oxidative stress. *Antioxid Redox Signal* (2015) 23(14):1144–70. doi: 10.1089/ars.2015.6317

12. Sies H. Oxidative stress: a concept in redox biology and medicine. *Redox Biol* (2015) 4:180–3. doi: 10.1016/j.redox.2015.01.002

13. Giacco F, Brownlee M. Oxidative stress and diabetic complications. *Circ Res* (2010) 107(9):1058–70. doi: 10.1161/CIRCRESAHA.110.223545

14. Beckman KB, Ames BN. The free radical theory of aging matures. *Physiol Rev* (1998) 78(2):547–81. doi: 10.1152/physrev.1998.78.2.547

15. Golden TR, Hinerfeld DA, Melov S. Oxidative stress and aging: beyond correlation. *Aging Cell* (2002) 1(2):117–23. doi: 10.1046/j.1474-9728.2002.00015.x

16. Oliveira BF, Nogueira-Machado JA, Chaves MM. The role of oxidative stress in the aging process. *Sci World J* (2010) 10:1121–8. doi: 10.1100/tsw.2010.94

17. Mancini A, Di Segni C, Raimondo S, Olivieri G, Silvestrini A, Meucci E, et al. Thyroid hormones, oxidative stress, and inflammation. *Mediators Inflamm* (2016) 2016:6757154. doi: 10.1155/2016/6757154

18. Hussain T, Tan B, Yin Y, Blachier F, Tossou MC, Rahu N. Oxidative stress and inflammation: what polyphenols can do for us? *Oxid Med Cell Longev* (2016) 2016:7432797. doi: 10.1155/2016/7432797

19. Grevendonk L, Connell NJ, McCrum C, Fealy CE, Bilet L, Bruls YMH, et al. Impact of aging and exercise on skeletal muscle mitochondrial capacity, energy metabolism, and physical function. *Nat Commun* (2021) 12(1):4773. doi: 10.1038/s41467-021-24956-2

20. Lennicke C, Cochemé HM. Redox metabolism: ROS as specific molecular regulators of cell signaling and function. *Mol Cell* (2021) 81(18):3691–707. doi: 10.1016/j.molcel.2021.08.018

21. Tramunt B, Smati S, Grandgeorge N, Lenfant F, Arnal JF, Montagner A, et al. Sex differences in metabolic regulation and diabetes susceptibility. *Diabetologia* (2020) 63(3):453–61. doi: 10.1007/s00125-019-05040-3

22. Reuter S, Gupta SC, Chaturvedi MM, Aggarwal BB. Oxidative stress, inflammation, and cancer: how are they linked? *Free Radic Biol Med* (2010) 49 (11):1603–16. doi: 10.1016/j.freeradbiomed.2010.09.006

23. Tan YT, Lin JF, Li T, Li JJ, Xu RH, Ju HQ. LncRNA-mediated posttranslational modifications and reprogramming of energy metabolism in cancer. *Cancer Commun (Lond)*. (2021) 41(2):109–20. doi: 10.1002/cac2.12108

24. Zhou X, Du HH, Jiang M, Zhou C, Deng Y, Long X, et al. Antioxidant effect of lactobacillus fermentum CQPC04-fermented soy milk on d-Galactose-Induced oxidative aging mice. *Front Nutr* (2021) 8:727467. doi: 10.3389/fnut.2021.727467

25. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK biobank participants. *PloS One* (2019) 14(5): e0213653. doi: 10.1371/journal.pone.0213653

26. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* (2021) 13(1):152. doi: 10.1186/s13073-021-00968-x

27. Deberneh HM, Kim I. Prediction of type 2 diabetes based on machine learning algorithm. *Int J Environ Res Public Health* (2021) 18(6):3317. doi: 10.3390/ijerph18063317

28. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, et al. Machine learning methods to predict diabetes complications. *J Diabetes Sci Technol* (2018) 12 (2):295–302. doi: 10.1177/1932296817706375

29. Li Z, Pan X, Cai YD. Identification of type 2 diabetes biomarkers from mixed single-cell sequencing data with feature selection methods. *Front Bioeng Biotechnol* (2022) 10:890901. doi: 10.3389/fbioe.2022.890901

30. Pearson-Stuttard J, Papadimitriou N, Markozannes G, Cividini S, Kakourou A, Gill D, et al. Type 2 diabetes and cancer: an umbrella review of observational and mendelian randomization studies. *Cancer Epidemiol Biomarkers Prev* (2021) 30 (6):1218–28. doi: 10.1158/1055-9965.EPI-20-1245

31. Swerdlow DI. Mendelian randomization and type 2 diabetes. *Cardiovasc Drugs Ther* (2016) 30(1):51–7. doi: 10.1007/s10557-016-6638-5

32. Wang Z, Wang F. Identification of ten-gene related to lipid metabolism for predicting overall survival of breast invasive carcinoma. *Contrast Media Mol Imaging* (2022) 2022:8348780. doi: 10.1155/2022/8348780

33. Tao JH, Wang XT, Yuan W, Chen JN, Wang ZJ, Ma YB, et al. Reduced serum high-density lipoprotein cholesterol levels and aberrantly expressed cholesterol metabolism genes in colorectal cancer. *World J Clin Cases* (2022) 10(14):4446–59. doi: 10.12998/wjcc.v10.i14.4446

34. Fiuza-Luces C, Santos-Lozano A, Llavero F, Campo R, Nogales-Gadea G, Díez-Bermejo J, et al. Muscle molecular adaptations to endurance exercise training are conditioned by glycogen availability: a proteomics-based analysis in the McArdle mouse model. *J Physiol* (2018) 596(6):1035–61. doi: 10.1113/JP275292

35. Gong YY, Liu YY, Li J, Su L, Yu S, Zhu XN, et al. Hypermethylation of Cox5a promoter is associated with mitochondrial dysfunction in skeletal muscle of high fat diet-induced insulin resistant rats. *PloS One* (2014) 9(12):e113784. doi: 10.1371/journal.pone.0113784

36. Peng L, Ma W, Xie Q, Chen B. Identification and validation of hub genes for diabetic retinopathy. *PeerJ* (2021) 9:e12126. doi: 10.7717/peerj.12126

37. Ting TW, Brett MS, Tan ES, Shen Y, Lee SP, Lim EC, et al. Cockayne syndrome due to a maternally-inherited whole gene deletion of ERCC8 and a paternally-inherited ERCC8 exon 4 deletion. *Gene* (2015) 572(2):274–8. doi: 10.1016/j.gene.2015.07.065

38. Akasheh RT, Pini M, Pang J, Fantuzzi G. Increased adiposity in annexin A1-deficient mice. *PloS One* (2013) 8(12):e82608. doi: 10.1371/journal.pone.0082608

39. Nechushtan A, Smith CL, Lamensdorf I, Yoon SH, Youle RJ. Bax and bak coalesce into novel mitochondria-associated clusters during apoptosis. *J Cell Biol* (2001) 153(6):1265–76. doi: 10.1083/jcb.153.6.1265

40. Nagao M, Esguerra JLS, Asai A, Ofori JK, Edlund A, Wendt A, et al. Potential protection against type 2 diabetes in obesity through lower CD36 expression and improved exocytosis in β-cells. *Diabetes* (2020) 69(6):1193–205. doi: 10.2337/db19-0944

41. Zhao Z, Zhang X, Zhao C, Choi J, Shi J, Song K, et al. Protection of pancreatic beta-cells by group VIA phospholipase A(2)-mediated repair of mitochondrial membrane peroxidation. *Endocrinology* (2010) 151(7):3038–48. doi: 10.1210/en.2010-0016

42. Heemskerk MM, Giera M, Bouazzaoui FE, Lips MA, Pijl H, van Dijk KW, et al. Increased PUFA content and 5-lipoxygenase pathway expression are associated with subcutaneous adipose tissue inflammation in obese women with type 2 diabetes. *Nutrients* (2015) 7(9):7676–90. doi: 10.3390/nu7095362

43. Ivashchenko O, Van Veldhoven PP, Brees C, Ho YS, Terlecky SR, Fransen M. Intraperoxisomal redox balance in mammalian cells: oxidative stress and interorganellar cross-talk. *Mol Biol Cell* (2011) 22(9):1440–51. doi: 10.1091/mbc.E10-11-0919

44. Rack JGM, Ariza A, Drown BS, Henfrey C, Bartlett E, Shirai T, et al. (ADP-ribosyl)hydrolases: structural basis for differential substrate recognition and inhibition. *Cell Chem Biol* (2018) 25(12):1533–1546.e12. doi: 10.1016/j.chembiol.2018.11.001

45. Zhang C, Wang L, Liu H, Deng G, Xu P, Tan Y, et al. ADPRH is a prognosis-related biomarker and correlates with immune infiltrates in low grade glioma. *J Cancer* (2021) 12(10):2912–20. doi: 10.7150/jca.51643

46. Ouyang Yb, Lane WS, Moore KL. Tyrosylprotein sulfotransferase: purification and molecular cloning of an enzyme that catalyzes tyrosine O-sulfation, a common posttranslational modification of eukaryotic proteins. *Proc Natl Acad Sci USA* (1998) 95 (6):2896–901. doi: 10.1073/pnas.95.6.2896

47. Westmuckett AD, Moore KL. Lack of tyrosylprotein sulfotransferase activity in hematopoietic cells drastically attenuates atherosclerosis in ldlr-/- mice. *Arterioscler Thromb Vasc Biol* (2009) 29(11):1730–6. doi: 10.1161/ATVBAHA.109.192963

48. Xiyang YB, Liu R, Wang XY, Li S, Zhao Y, Lu BT, et al. COX5A plays a vital role in memory impairment associated with brain aging via the BDNF/ERK1/2 signaling pathway. *Front Aging Neurosci* (2020) 12:215. doi: 10.3389/fnagi.2020.00215

49. Aharon-Hananel G, Romero-Afrima L, Saada A, Mantzur C, Raz I, Weksler-Zangen S. Cytochrome c oxidase activity as a metabolic regulator in pancreatic beta-cells. *Cells* (2022) 11(6):929. doi: 10.3390/cells11060929

50. Karslioglu E, Kleinberger JW, Salim FG, Cox AE, Takane KK, Scott DK, et al. cMyc is a principal upstream driver of beta-cell proliferation in rat insulinoma cell lines and is an effective mediator of human beta-cell replication. *Mol Endocrinol* (2011) 25 (10):1760–72. doi: 10.1210/me.2011-1074

51. Huard K, Londregan AT, Tesz G, Bahnck KB, Magee TV, Hepworth D, et al. Discovery of selective small molecule inhibitors of monoacylglycerol acyltransferase 3. *J Med Chem* (2015) 58(18):7164–72. doi: 10.1021/acs.jmedchem.5b01008

52. Rouault TA, Tong WH. Iron-sulfur cluster biogenesis and human disease. *Trends Genet* (2008) 24(8):398–407. doi: 10.1016/j.tig.2008.05.008

53. Heitzer T, Krohn K, Albers S, Meinertz T. Tetrahydrobiopterin improves endothelium-dependent vasodilation by increasing nitric oxide activity in patients with type II diabetes mellitus. *Diabetologia* (2000) 43(11):1435–8. doi: 10.1007/s001250051551

54. Gu C, Liu S, Wang H, Dou H. Role of the thioredoxin interacting protein in diabetic nephropathy and the mechanism of regulating NOD-like receptor protein 3 inflammatory corpuscle. *Int J Mol Med* (2019) 43(6):2440–50. doi: 10.3892/ijmm.2019.4163

55. Pinto-Junior DC, Silva KS, Michalani ML, Yonamine CY, Esteves JV, Fabre NT, et al. Advanced glycation end products-induced insulin resistance involves repression of skeletal muscle GLUT4 expression. *Sci Rep* (2018) 8(1):8109. doi: 10.1038/s41598-018-26482-6

56. Chou CW, Hsieh YH, Ku SC, Shen WJ, Anuraga G, Khoa Ta HD, et al. Potential prognostic biomarkers of OSBPL family genes in patients with pancreatic ductal adenocarcinoma. *Biomedicines* (2021) 9(11):1601. doi: 10.3390/biomedicines9111601

57. Wright MB, Varona Santos J, Kemmer C, Maugeais C, Carralot JP, Roever S, et al. Compounds targeting OSBPL7 increase ABCA1-dependent cholesterol efflux preserving kidney function in two models of kidney disease. *Nat Commun* (2021) 12 (1):4662. doi: 10.1038/s41467-021-24890-3

58. Krishna S, Arrojo E Drigo R, Capitanio JS, Ramachandra R, Ellisman M, Hetzer MW. Identification of long-lived proteins in the mitochondria reveals increased stability of the electron transport chain. *Dev Cell* (2021) 56(21):2952–65. doi: 10.1016/j.devcel.2021.10.008

59. Wang X, Wang LT, Yu B. UBE2D1 and COX7C as potential biomarkers of diabetes-related sepsis. *BioMed Res Int* (2022) 2022:9463717. doi: 10.1155/2022/9463717

60. Zeng Z, Yu J, Yang Z, Du K, Chen Y, Zhou L. Investigation of M2 macrophage-related gene affecting patients prognosis and drug sensitivity in non-small cell lung cancer: evidence from bioinformatic and experiments. *Front Oncol* (2022) 12:1096449. doi: 10.3389/fonc.2022.1096449

61. Ahn EH, Kim DW, Shin MJ, Ryu EJ, Yong JI, Chung SY, et al. Tat-ATOX1 inhibits streptozotocin-induced cell death in pancreatic RINm5F cells and attenuates diabetes in a mouse model. *Int J Mol Med* (2016) 38(1):217–24. doi: 10.3892/ijmm.2016.2599

62. Jiang A, Gao H, Kelley MR, Qiao X. Inhibition of APE1/Ref-1 redox activity with APX3330 blocks retinal angiogenesis *in vitro* and *in vivo*. *Vision Res* (2011) 51(1):93–100. doi: 10.1016/j.visres.2010.10.008

63. Moreno-Gonzalez I, Edwards Iii G, Salvadores N, Shahnawaz M, Diaz-Espinoza R, Soto C. Molecular interaction between type 2 diabetes and alzheimer's disease through cross-seeding of protein misfolding. *Mol Psychiatry* (2017) 22(9):1327–34. doi: 10.1038/mp.2016.230

64. Wu Z, Chen A, Zhang G, Liu C, Yin S, Song R, et al. ALDH3B1 protects interfollicular epidermal cells against lipid peroxidation via the NRF2 pathway. *Cell Stress Chaperones* (2022) 27(6):703–15. doi: 10.1007/s12192-022-01306-9

65. Pei-Yuan Z, Yu-Wei L, Xiang-Nan Z, Song T, Rong Z, Xiao-Xiao H, et al. Overexpression of axl reverses endothelial cells dysfunction in high glucose and hypoxia. *J Cell Biochem* (2019) 120(7):11831–41. doi: 10.1002/jcb.28462

66. Yu N, Fang X, Zhao D, Mu Q, Zuo J, Ma Y, et al. Anti-diabetic effects of jiang tang xiao ke granule via PI3K/Akt signalling pathway in type 2 diabetes KKAy mice. *PloS One* (2017) 12(1):e0168980. doi: 10.1371/journal.pone.0168980

67. Pan X, Mota S, Zhang B. Circadian clock regulation on lipid metabolism and metabolic diseases. *Adv Exp Med Biol* (2020) 1276:53–66. doi: 10.1007/978-981-15-6082-8_5

68. Tian D, Xiang Y, Tang Y, Ge Z, Li Q, Zhang Y. Circ-ADAM9 targeting PTEN and ATG7 promotes autophagy and apoptosis of diabetic endothelial progenitor cells by sponging mir-20a-5p. *Cell Death Dis* (2020) 11(7):526. doi: 10.1038/s41419-020-02745-x

69. Anderson KA, Lin F, Ribar TJ, Stevens RD, Muehlbauer MJ, Newgard CB, et al. Deletion of CaMKK2 from the liver lowers blood glucose and improves whole-body glucose tolerance in the mouse. *Mol Endocrinol* (2012) 26(2):281–91. doi: 10.1210/me.2011-1299

70. Carmona-Rivera C, Simeonov DR, Cardillo ND, Gahl WA, Cadilla CL. A divalent interaction between HPS1 and HPS4 is required for the formation of the biogenesis of lysosome-related organelle complex-3 (BLOC-3). *Biochim Biophys Acta* (2013) 1833(3):468–78. doi: 10.1016/j.bbamcr.2012.10.019

71. Tang H, Huang X, Pang S. Regulation of the lysosome by sphingolipids: potential role in aging. *J Biol Chem* (2022) 298(7):102118. doi: 10.1016/j.jbc.2022.102118

72. Patel V, Bidault G, Chambers JE, Carobbio S, Everden AJT, Garcés C, et al. Inactivation of Ppp1r15a minimises weight gain and insulin resistance during caloric excess in female mice. *Sci Rep* (2019) 9(1):2903. doi: 10.1038/s41598-019-39562-y

73. Zhao Y, Li W, Zhang K, Xu M, Zou Y, Qiu X, et al. Revealing oxidative stress-related genes in osteoporosis and advanced structural biological study for novel natural material discovery regarding MAPKAPK2. *Front Endocrinol (Lausanne)* (2022) 13:1052721. doi: 10.3389/fendo.2022.1052721

74. Hatori Y, Lutsenko S. The role of copper chaperone Atox1 in coupling redox homeostasis to intracellular copper distribution. *Antioxidants (Basel)* (2016) 5(3):25. doi: 10.3390/antiox5030025

75. Hatori Y, Lutsenko S. An expanding range of functions for the copper chaperone/antioxidant protein Atox1. *Antioxid Redox Signal* (2013) 19(9):945–57. doi: 10.1089/ars.2012.5086

76. Hassan A, Sharma Kandel R, Mishra R, Gautam J, Alaref A, Jahan N. Diabetes mellitus and parkinson's disease: shared pathophysiological links and possible therapeutic implications. *Cureus* (2020) 12(8):e9853. doi: 10.7759/cureus.9853

77. Cheong JLY, de Pablo-Fernandez E, Foltynie T, Noyce AJ. The association between type 2 diabetes mellitus and parkinson's disease. *J Parkinsons Dis* (2020) 10(3):775–89. doi: 10.3233/JPD-191900

78. Chohan H, Senkevich K, Patel RK, Bestwick JP, Jacobs BM, Bandres Ciga S, et al. Type 2 diabetes as a determinant of parkinson's disease risk and progression. *Mov Disord* (2021) 36(6):1420–9. doi: 10.1002/mds.28551

79. Pugazhenthi S, Qin L, Reddy PH. Common neurodegenerative pathways in obesity, diabetes, and alzheimer's disease. *Biochim Biophys Acta Mol Basis Dis* (2017) 1863(5):1037–45. doi: 10.1016/j.bbadis.2016.04.017

80. Burillo J, Marqués P, Jiménez B, González-Blanco C, Benito M, Guillén C. Insulin resistance and diabetes mellitus in alzheimer's disease. *Cells* (2021) 10(5):1236. doi: 10.3390/cells10051236

81. Kreider RB, Kalman DS, Antonio J, Ziegenfuss TN, Wildman R, Collins R, et al. International society of sports nutrition position stand: safety and efficacy of creatine supplementation in exercise, sport, and medicine. *J Int Soc Sports Nutr* (2017) 14:18. doi: 10.1186/s12970-017-0173-z

82. Akhoundi M, Downing T, Votýpka J, Kuhls K, Lukeš J, Cannet A, et al. Leishmania infections: molecular targets and diagnosis. *Mol Aspects Med* (2017) 57:1–29. doi: 10.1016/j.mam.2016.11.012

83. Ashayeri Ahmadabad R, Mirzaasgari Z, Gorji A, Khaleghi Ghadiri M. Toll-like receptor signaling pathways: novel therapeutic targets for cerebrovascular disorders. *Int J Mol Sci* (2021) 22(11):6153. doi: 10.3390/ijms22116153

84. Yu GH, Li SF, Wei R, Jiang Z. Diabetes and colorectal cancer risk: clinical and therapeutic implications. *J Diabetes Res* (2022) 2022:1747326. doi: 10.1155/2022/1747326

85. Dissanayake WC, Sorrenson B, Shepherd PR. The role of adherens junction proteins in the regulation of insulin secretion. *Biosci Rep* (2018) 38(2):BSR20170989. doi: 10.1042/BSR20170989

86. Zikherman J, Au-Yeung B. The role of T cell receptor signaling thresholds in guiding T cell fate decisions. *Curr Opin Immunol* (2015) 33:43–8. doi: 10.1016/j.coi.2015.01.012

87. Nolfi-Donegan D, Braganza A, Shiva S. Mitochondrial electron transport chain: Oxidative phosphorylation, oxidant production, and methods of measurement. *Redox Biol* (2020) 37:101674. doi: 10.1016/j.redox.2020.101674

88. Kolwicz SC Jr, Purohit S, Tian R. Cardiac metabolism and its interactions with contraction, growth, and survival of cardiomyocytes. *Circ Res* (2013) 113(5):603–16. doi: 10.1161/CIRCRESAHA.113.302095

89. Zheng X, Narayanan S, Xu C, Eliasson Angelstig S, Grünler J, Zhao A, et al. Repression of hypoxia-inducible factor-1 contributes to increased mitochondrial reactive oxygen species production in diabetes. *Elife* (2022) 11:e70714. doi: 10.7554/eLife.70714

90. Leenders F, Groen N, de Graaf N, Engelse MA, Rabelink TJ, de Koning EJP, et al. Oxidative stress leads to β-cell dysfunction through loss of β-cell identity. *Front Immunol* (2021) 12:690379. doi: 10.3389/fimmu.2021.690379

91. Walton EL. Oxidative stress and diabetes: glucose response in the cROSsfire. *BioMed J* (2017) 40(5):241–4. doi: 10.1016/j.bj.2017.10.001

92. He L, He T, Farrar S, Ji L, Liu T, Ma X. Antioxidants maintain cellular redox homeostasis by elimination of reactive oxygen species. *Cell Physiol Biochem* (2017) 44(2):532–53. doi: 10.1159/000485089

93. Ray PD, Huang BW, Tsuji Y. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. *Cell Signal* (2012) 24(5):981–90. doi: 10.1016/j.cellsig.2012.01.008

94. Simaan H, Lev S, Horwitz BA. Oxidant-sensing pathways in the responses of fungal pathogens to chemical stress signals. *Front Microbiol* (2019) 10:567. doi: 10.3389/fmicb.2019.00567

95. Levonen AL, Hill BG, Kansanen E, Zhang J, Darley-Usmar VM. Redox regulation of antioxidants, autophagy, and the response to stress: implications for electrophile therapeutics. *Free Radic Biol Med* (2014) 71:196–207. doi: 10.1016/j.freeradbiomed.2014.03.025

96. Orang AV, Petersen J, McKinnon RA, Michael MZ. Micromanaging aerobic respiration and glycolysis in cancer cells. *Mol Metab* (2019) 23:98–126. doi: 10.1016/j.molmet.2019.01.014

97. Kasai S, Shimizu S, Tatara Y, Mimura J, Itoh K. Regulation of Nrf2 by mitochondrial reactive oxygen species in physiology and pathology. *Biomolecules* (2020) 10(2):320. doi: 10.3390/biom10020320

98. Toyoda Y, Saitoh S. Adaptive regulation of glucose transport, glycolysis and respiration for cell proliferation. *Biomol Concepts* (2015) 6(5-6):423–30. doi: 10.1515/bmc-2015-0018

99. Rabbani N, Thornalley PJ. Hexokinase-2 glycolytic overload in diabetes and ischemia-reperfusion injury. *Trends Endocrinol Metab* (2019) 30(7):419–31. doi: 10.1016/j.tem.2019.04.011

100. Klinge CM. Estrogenic control of mitochondrial function. *Redox Biol* (2020) 31:101435. doi: 10.1016/j.redox.2020.101435

101. Zhang J, Wang X, Vikash V, Ye Q, Wu D, Liu Y, et al. ROS and ROS-mediated cellular signaling. *Oxid Med Cell Longev* (2016) 2016:4350965. doi: 10.1155/2016/4350965

102. Xu L, Li Y, Yin L, Qi Y, Sun H, Sun P, et al. miR-125a-5p ameliorates hepatic glycolipid metabolism disorder in type 2 diabetes mellitus through targeting of STAT3. *Theranostics* (2018) 8(20):5593–609. doi: 10.7150/thno.27425

103. Coulthard LR, White DE, Jones DL, McDermott MF, Burchill SA. p38(MAPK): stress responses from molecular mechanisms to therapeutics. *Trends Mol Med* (2009) 15(8):369–79. doi: 10.1016/j.molmed.2009.06.005

104. Popa A, Georgescu M, Popa SG, Nica AE, Georgescu EF. New insights in the molecular pathways linking obesity, type 2 diabetes and cancer. *Rom J Morphol Embryol* (2019) 60(4):1115–25.

105. Hüttemann M, Lee I, Samavati L, Yu H, Doan JW. Regulation of mitochondrial oxidative phosphorylation through cell signaling. *Biochim Biophys Acta* (2007) 1773(12):1701–20. doi: 10.1016/j.bbamcr.2007.10.001

106. Andrae J, Gallini R, Betsholtz C. Role of platelet-derived growth factors in physiology and medicine. *Genes Dev* (2008) 22(10):1276–312. doi: 10.1101/gad.1653708

107. Drouin M, Saenz J, Chiffoleau E. C-type lectin-like receptors: head or tail in cell death immunity. *Front Immunol* (2020) 11:251. doi: 10.3389/fimmu.2020.00251

108. Ivanova EA, Orekhov AN. Monocyte activation in immunopathology: cellular test for development of diagnostics and therapy. *J Immunol Res* (2016) 2016:4789279. doi: 10.1155/2016/4789279

109. Coulthard MG, Morgan M, Woodruff TM, Arumugam TV, Taylor SM, Carpenter TC, et al. Eph/Ephrin signaling in injury and inflammation. *Am J Pathol* (2012) 181(5):1493–503. doi: 10.1016/j.ajpath.2012.06.043

110. Boekema EJ, Braun HP. Supramolecular structure of the mitochondrial oxidative phosphorylation system. *J Biol Chem* (2007) 282(1):1–4. doi: 10.1074/jbc.R600031200

111. Picca A, Guerra F, Calvani R, Romano R, Coelho-Júnior HJ, Bucci C, et al. Mitochondrial dysfunction, protein misfolding and neuroinflammation in parkinson's disease: roads to biomarker discovery. *Biomolecules* (2021) 11(10):1508. doi: 10.3390/biom11101508

112. Ejma M, Madetko N, Brzecka A, Guranski K, Alster P, Misiuk-Hojło M, et al. The links between parkinson's disease and cancer. *Biomedicines* (2020) 8(10):416. doi: 10.3390/biomedicines8100416

113. Singh N, Baby D, Rajguru JP, Patil PB, Thakkannavar SS, Pujari VB. Inflammation and cancer. *Ann Afr Med* (2019) 18(3):121–6. doi: 10.4103/aam.aam_56_18

114. Daryabor G, Atashzar MR, Kabelitz D, Meri S, Kalantar K. The effects of type 2 diabetes mellitus on organ metabolism and the immune system. *Front Immunol* (2020) 11:1582. doi: 10.3389/fimmu.2020.01582

115. Huang R, Zhou PK. DNA Damage repair: historical perspectives, mechanistic pathways and clinical translation for targeted cancer therapy. *Signal Transduct Target Ther* (2021) 6(1):254. doi: 10.1038/s41392-021-00648-7

116. Ntambi JM, Miyazaki M, Stoehr JP, Lan H, Kendziorski CM, Yandell BS, et al. Loss of stearoyl-CoA desaturase-1 function protects mice against adiposity. *Proc Natl Acad Sci USA* (2002) 99(17):11482–6. doi: 10.1073/pnas.132384699

117. Yuan X, Hu S, Li L, Liu H, He H, Wang J. Metabolomic analysis of SCD during goose follicular development: implications for lipid metabolism. *Genes (Basel)* (2020) 11(9):1001. doi: 10.3390/genes11091001

118. Yuan X, Abdul-Rahman I, Hu S, Li L, He H, Xia L, et al. Mechanism of SCD participation in lipid droplet-mediated steroidogenesis in goose granulosa cells. *Genes (Basel)* (2022) 13(9):1516. doi: 10.3390/genes13091516

119. Tang B, Qiu J, Hu S, Li L, Wang J. Role of stearyl-coenzyme a desaturase 1 in mediating the effects of palmitic acid on endoplasmic reticulum stress, inflammation, and apoptosis in goose primary hepatocytes. *Anim Biosci* (2021) 34(7):1210–20. doi: 10.5713/ajas.20.0444

120. Valentine D, Teerlink CC, Farnham JM, Rowe K, Kaddas H, Tschanz J, et al. Comorbidity and cancer disease rates among those at high-risk for alzheimer's disease: a population database analysis. *Int J Environ Res Public Health* (2022) 19(24):16419. doi: 10.3390/ijerph192416419

121. Kesler SR, Rao V, Ray WJ, Rao A. Alzheimer's disease neuroimaging initiative. probability of alzheimer's disease in breast cancer survivors based on gray-matter structural network efficiency. *Alzheimers Dement (Amst)* (2017) 9:67–75. doi: 10.1016/j.dadm.2017.10.002

122. Kroemer G, Pouyssegur J. Tumor cell metabolism: cancer's achilles' heel. *Cancer Cell* (2008) 13(6):472–82. doi: 10.1016/j.ccr.2008.05.005

123. Ojha R, Tantray I, Rimal S, Mitra S, Cheshier S, Lu B. Regulation of reverse electron transfer at mitochondrial complex I by unconventional notch action in cancer stem cells. *Dev Cell* (2022) 57(2):260–276.e9. doi: 10.1016/j.devcel.2021.12.020

124. Arteaga CL, Engelman JA. ERBB receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics. *Cancer Cell* (2014) 25(3):282–303. doi: 10.1016/j.ccr.2014.02.025

125. Dressing GE, Goldberg JE, Charles NJ, Schwertfeger KL, Lange CA. Membrane progesterone receptor expression in mammalian tissues: a review of regulation and physiological implications. *Steroids* (2011) 76(1-2):11–7. doi: 10.1016/j.steroids.2010.09.006

126. Goral V. Pancreatic cancer: pathogenesis and diagnosis. *Asian Pac J Cancer Prev* (2015) 16(14):5619–24. doi: 10.7314/apjcp.2015.16.14.5619

127. Ko HJ, Chang SY. Regulation of intestinal immune system by dendritic cells. *Immune Netw* (2015) 15(1):1–8. doi: 10.4110/in.2015.15.1.1

128. Zsindely N, Siági F, Bodai L. DNA Methylation in huntington's disease. *Int J Mol Sci* (2021) 22(23):12736. doi: 10.3390/ijms222312736

129. LeBleu VS, O'Connell JT, Gonzalez Herrera KN, Wikman H, Pantel K, Haigis MC, et al. PGC-1α mediates mitochondrial biogenesis and oxidative phosphorylation in cancer cells to promote metastasis. *Nat Cell Biol* (2014) 16(10):992–1003. doi: 10.1038/ncb3039

130. Gray RE, Harris GT. Renal cell carcinoma: diagnosis and management. *Am Fam Physician.* (2019) 99(3):179–84.

131. Davis LE, Shalin SC, Tackett AJ. Current state of melanoma diagnosis and treatment. *Cancer Biol Ther* (2019) 20(11):1366–79. doi: 10.1080/15384047.2019.1640032

132. Kader AK. Bladder cancer. *Sci World J* (2011) 11:2565–6. doi: 10.1100/2011/251920

133. Venhuizen JH, Jacobs FJC, Span PN, Zegers MM. P120 and e-cadherin: double-edged swords in tumor metastasis. *Semin Cancer Biol* (2020) 60:107–20. doi: 10.1016/j.semcancer.2019.07.020

134. Ludwig K, Kornblum HI. Molecular markers in glioma. *J Neurooncol* (2017) 134(3):505–12. doi: 10.1007/s11060-017-2379-y

135. Totland MZ, Rasmussen NL, Knudsen LM, Leithe E. Regulation of gap junction intercellular communication by connexin ubiquitination: physiological and pathophysiological implications. *Cell Mol Life Sci* (2020) 77(4):573–91. doi: 10.1007/s00018-019-03285-0

136. Buono R, Longo VD. Starvation, stress resistance, and cancer. *Trends Endocrinol Metab* (2018) 29(4):271–80. doi: 10.1016/j.tem.2018.01.008

137. Yu L, Lu M, Jia D, Ma J, Ben-Jacob E, Levine H, et al. Modeling the genetic regulation of cancer metabolism: interplay between glycolysis and oxidative phosphorylation. *Cancer Res* (2017) 77(7):1564–74. doi: 10.1158/0008-5472.CAN-16-2074

138. Owusu BY, Galemmo R, Janetka J, Klampfer L. Hepatocyte growth factor, a key tumor-promoting factor in the tumor microenvironment. *Cancers (Basel)* (2017) 9(4):35. doi: 10.3390/cancers9040035

139. Hartwig A. Cadmium and cancer. *Met Ions Life Sci* (2013) 11:491–507. doi: 10.1007/978-94-007-5179-8_15

140. Song Y, Xu Y, Pan C, Yan L, Wang ZW, Zhu X. The emerging role of SPOP protein in tumorigenesis and cancer therapy. *Mol Cancer* (2020) 19(1):2. doi: 10.1186/s12943-019-1124-x

141. Vilema-Enríquez G, Arroyo A, Grijalva M, Amador-Zafra RI, Camacho J. Molecular and cellular effects of hydrogen peroxide on human lung cancer cells: potential therapeutic implications. *Oxid Med Cell Longev* (2016) 2016:1908164. doi: 10.1155/2016/1908164

142. Lam T, Aguirre-Ghiso JA, Geller MA, Aksan A, Azarin SM. Immobilization rapidly selects for chemoresistant ovarian cancer cells with enhanced ability to enter dormancy. *Biotechnol Bioeng* (2020) 117(10):3066–80. doi: 10.1002/bit.27479

143. Gao X, Yang Y, Wang J, Zhang L, Sun C, Wang Y, et al. Inhibition of mitochondria NADH-ubiquinone oxidoreductase (complex I) sensitizes the radioresistant glioma U87MG cells to radiation. *BioMed Pharmacother* (2020), 129:110460. doi: 10.1016/j.biopha.2020.110460

144. Glorieux C, Calderon PB. Catalase, a remarkable enzyme: targeting the oldest antioxidant enzyme to find a new cancer treatment approach. *Biol Chem* (2017) 398(10):1095–108. doi: 10.1515/hsz-2017-0131

145. Raimondi V, Ciccarese F, Ciminale V. Oncogenic pathways and the electron transport chain: a dangeROS liaison. *Br J Cancer* (2020) 122(2):168–81. doi: 10.1038/s41416-019-0651-y

146. Thorne JL, Campbell MJ. Nuclear receptors and the warburg effect in cancer. *Int J Cancer* (2015) 137(7):1519–27. doi: 10.1002/ijc.29012

147. Ginckels P, Holvoet P. Oxidative stress and inflammation in cardiovascular diseases and cancer: role of non-coding RNAs. *Yale J Biol Med* (2022) 95(1):129–52.

148. Aggarwal V, Tuli HS, Varol A, Thakral F, Yerer MB, Sak K, et al. Role of reactive oxygen species in cancer progression: molecular mechanisms and recent advancements. *Biomolecules* (2019) 9(11):735. doi: 10.3390/biom9110735

149. Schwarz EC, Qu B, Hoth M. Calcium, cancer and killing: the role of calcium in killing cancer cells by cytotoxic T lymphocytes and natural killer cells. *Biochim Biophys Acta* (2013) 1833(7):1603–11. doi: 10.1016/j.bbamcr.2012.11.016

150. Cautain B, Hill R, de Pedro N, Link W. Components and regulation of nuclear transport processes. *FEBS J* (2015) 282(3):445–62. doi: 10.1111/febs.13163

151. Fogg VC, Lanning NJ, Mackeigan JP. Mitochondria in cancer: at the crossroads of life and death. *Chin J Cancer* (2011) 30(8):526–39. doi: 10.5732/cjc.011.10018

152. Law ML, Metzger JM. Cardiac myocyte intrinsic contractility and calcium handling deficits underlie heart organ dysfunction in murine cancer cachexia. *Sci Rep* (2021) 11(1):23627. doi: 10.1038/s41598-021-02688-z

153. Amelio I, Cutruzzolá F, Antonov A, Agostini M, Melino G. Serine and glycine metabolism in cancer. *Trends Biochem Sci* (2014) 39(4):191–8. doi: 10.1016/j.tibs.2014.02.004

154. Liu J, Wu Z, Han D, Wei C, Liang Y, Jiang T, et al. Mesencephalic astrocyte-derived neurotrophic factor inhibits liver cancer through small ubiquitin-related modifier (SUMO)ylation-related suppression of NF-κB/Snail signaling pathway and epithelial-mesenchymal transition. *Hepatology* (2020) 71(4):1262–78. doi: 10.1002/hep.30917

155. Schuldt M, Pei J, Harakalova M, Dorsch LM, Schlossarek S, Mokry M, et al. Proteomic and functional studies reveal detyrosinated tubulin as treatment target in sarcomere mutation-induced hypertrophic cardiomyopathy. *Circ Heart Fail* (2021) 14(1):e007022. doi: 10.1161/CIRCHEARTFAILURE.120.007022

156. Faas M, Ipseiz N, Ackermann J, Culemann S, Grüneboom A, Schröder F, et al. IL-33-induced metabolic reprogramming controls the differentiation of alternatively activated macrophages and the resolution of inflammation. *Immunity* (2021) 54(11):2531–2546.e5. doi: 10.1016/j.immuni.2021.09.010

157. Ward MH, Jones RR, Brender JD, de Kok TM, Weyer PJ, Nolan BT, et al. Drinking water nitrate and human health: an updated review. *Int J Environ Res Public Health* (2018) 15(7):1557. doi: 10.3390/ijerph15071557

158. Vultaggio-Poma V, Sarti AC, Di Virgilio F. Extracellular ATP: a feasible target for cancer therapy. *Cells* (2020) 9(11):2496. doi: 10.3390/cells9112496

159. Martínez-Reyes I, Cardona LR, Kong H, Vasan K, McElroy GS, Werner M, et al. Mitochondrial ubiquinol oxidation is necessary for tumour growth. *Nature* (2020) 585(7824):288–92. doi: 10.1038/s41586-020-2475-6

160. Radin DP, Tsirka SE. Interactions between tumor cells, neurons, and microglia in the glioma microenvironment. *Int J Mol Sci* (2020) 21(22):8476. doi: 10.3390/ijms21228476

161. Koike H, Iwasawa K, Ouchi R, Maezawa M, Giesbrecht K, Saiki N, et al. Modelling human hepato-biliary-pancreatic organogenesis from the foregut-midgut boundary. *Nature* (2019) 574(7776):112–6. doi: 10.1038/s41586-019-1598-0

162. Porporato PE, Filigheddu N, Pedro JMB, Kroemer G, Galluzzi L. Mitochondrial metabolism and cancer. *Cell Res* (2018) 28(3):265–80. doi: 10.1038/cr.2017.155

163. Zong WX, Rabinowitz JD, White E. Mitochondria and cancer. *Mol Cell* (2016) 61(5):667–76. doi: 10.1016/j.molcel.2016.02.011

164. Grimm MO, Mett J, Grimm HS, Hartmann T, Function APP. And lipids: a bidirectional link. *Front Mol Neurosci* (2017) 10:63. doi: 10.3389/fnmol.2017.00063

165. Wang J, Wang X, Guo Y, Ye L, Li D, Hu A, et al. Therapeutic targeting of SPIB/SPI1-facilitated interplay of cancer cells and neutrophils inhibits aerobic glycolysis and cancer progression. *Clin Transl Med* (2021) 11(11):e588. doi: 10.1002/ctm2.588

166. Sin O, Nollen EA. Regulation of protein homeostasis in neurodegenerative diseases: the role of coding and non-coding genes. *Cell Mol Life Sci* (2015) 72(21):4027–47. doi: 10.1007/s00018-015-1985-0

167. Press M, Jung T, König J, Grune T, Höhn A. Protein aggregates and proteostasis in aging: amylin and β-cell function. *Mech Ageing Dev* (2019) 177:46–54. doi: 10.1016/j.mad.2018.03.010

168. Archuleta TL, Lemieux AM, Saengsirisuwan V, Teachey MK, Lindborg KA, Kim JS, et al. Oxidant stress-induced loss of IRS-1 and IRS-2 proteins in rat skeletal muscle: role of p38 MAPK. *Free Radic Biol Med* (2009) 47(10):1486–93. doi: 10.1016/j.freeradbiomed.2009.08.014

169. Berdichevsky A, Guarente L, Bose A. Acute oxidative stress can reverse insulin resistance by inactivation of cytoplasmic JNK. *J Biol Chem* (2010) 285(28):21581–9. doi: 10.1074/jbc.M109.093633

170. Theocharidis G, Thomas BE, Sarkar D, Mumme HL, Pilcher WJR, Dwivedi B, et al. Single cell transcriptomic landscape of diabetic foot ulcers. *Nat Commun* (2022) 13(1):181. doi: 10.1038/s41467-021-27801-8

171. Song Y, He C, Jiang Y, Yang M, Xu Z, Yuan L, et al. Bulk and single-cell transcriptome analyses of islet tissue unravel gene signatures associated with pyroptosis and immune infiltration in type 2 diabetes. *Front Endocrinol (Lausanne)* (2023) 14:1132194. doi: 10.3389/fendo.2023.1132194

172. Wilson PC, Wu H, Kirita Y, Uchimura K, Ledru N, Rennke HG, et al. The single-cell transcriptomic landscape of early human diabetic nephropathy. *Proc Natl Acad Sci USA* (2019) 116(39):19619–25. doi: 10.1073/pnas.1908706116

173. Van Drie JH. Protein folding, protein homeostasis, and cancer. *Chin J Cancer* (2011) 30(2):124–37. doi: 10.5732/cjc.010.10162

174. Martínez de Toda I, Ceprián N, Díaz-Del Cerro E, de la Fuente M. The role of immune cells in oxi-Inflamm-Aging. *Cells* (2021) 10(11):2974. doi: 10.3390/cells10112974

175. Luc K, Schramm-Luc A, Guzik TJ, Mikolajczyk TP. Oxidative stress and inflammatory markers in prediabetes and diabetes. *J Physiol Pharmacol* (2019) 70 (6):809–24. doi: 10.26402/jpp.2019.6.01

176. Wu D, Bi X, Li P, Xu D, Qiu J, Li K, et al. Enhanced insulin-regulated phagocytic activities support extreme health span and longevity in multiple populations. *Aging Cell* (2023) 8:e13810. doi: 10.1111/acel.13810

177. Lananna BV, Musiek ES. The wrinkling of time: aging, inflammation, oxidative stress, and the circadian clock in neurodegeneration. *Neurobiol Dis* (2020) 139:104832. doi: 10.1016/j.nbd.2020.104832

178. Saoudaoui S, Bernard M, Cardin GB, Malaquin N, Christopoulos A, Rodier F. mTOR as a senescence manipulation target: a forked road. *Adv Cancer Res* (2021) 150:335–63. doi: 10.1016/bs.acr.2021.02.002

179. Palmer AK, Gustafson B, Kirkland JL, Smith U. Cellular senescence: at the nexus between ageing and diabetes. *Diabetologia* (2019) 62(10):1835–41. doi: 10.1007/s00125-019-4934-x

180. Gao W, Liu JL, Lu X, Yang Q. Epigenetic regulation of energy metabolism in obesity. *J Mol Cell Biol* (2021) 13(7):480–99. doi: 10.1093/jmcb/mjab043

181. Burgess S, Smith GD, Davie NM, Dudbridge F, Gill D, Glymour MM, et al. Guidelines for performing mendelian randomization investigations. version 2. *Wellcome Open Res* (2019) 4:186. doi: 10.12688/wellcomeopenres.15555.1

182. Burgess S, Foley CN, Allara E, Staley JR, Howson JMM, et al. A robust and efficient method for mendelian randomization with hundreds of genetic variants. *Nat Commun* (2020) 11:376. doi: 10.1038/s41467-019-14156-4

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 07 June 2023
DOI 10.3389/fendo.2023.1188547



OPEN ACCESS

EDITED BY
Prem P. Kushwaha,
Case Western Reserve University,
United States

REVIEWED BY
Wencai Liu,
The First Affiliated Hospital of Nanchang University, China
Li Ding,
The Affiliated Hospital of Xuzhou Medical University, China
Wei Huang,
Dongguan Tungwah Hospital, China

*CORRESPONDENCE
Rongjun Cui
cuirongjun@mdjmu.edu.cn

RECEIVED 17 March 2023
ACCEPTED 23 May 2023
PUBLISHED 07 June 2023

CITATION
Zhao G, Wang Z, Ji J and Cui R (2023) Effect of coffee consumption on thyroid function: NHANES 2007-2012 and Mendelian randomization.
*Front. Endocrinol.* 14:1188547.
doi: 10.3389/fendo.2023.1188547



# Effect of coffee consumption on thyroid function: NHANES 2007-2012 and Mendelian randomization

Guoxu Zhao[1], Zhao Wang[2], Jinli Ji[1] and Rongjun Cui[1]*

[1]Mudanjiang Medical University, Mudanjiang, China, [2]Chungnam National University School of Medicine, Daejeon Gwangyeoksi, Republic of Korea

**Background:** Coffee is one of the most consumed beverages worldwide, but the effects on the thyroid are unknown. This study aims to examine the association between coffee and thyroid function.

**Methods:** Participant data (≥ 20 years, n = 6578) for the observational study were obtained from NHANES 2007-2012. Analysis was performed using weighted linear regression models and multiple logistic regression models. Genetic datasets for Hyperthyroidism and Hypothyroidism were obtained from the IEU database and contained 462,933 European samples. Mendelian randomization (MR) was used for the analysis, inverse variance weighting (IVW) was the main method of analysis.

**Results:** In the model adjusted for other covariates, participants who drank 2-4 cups of coffee per day had significantly lower TSH concentrations compared to non-coffee drinkers (b=-0.23, 95% CI: -0.30, -0.16), but no statistically significant changes in TT4, FT4, TT3 and FT3. In addition, participants who drank <2 cups of coffee per day showed a low risk of developing subclinical hypothyroidism. (OR=0.60, 95% CI: 0.41, 0.88) Observational studies and MR studies have demonstrated both that coffee consumption has no effect on the risk of hyperthyroidism and hypothyroidism.

**Conclusions:** Our study showed that drinking <2 cups of coffee per day reduced the risk of subclinical hypothyroidism and drinking 2-4 cups of coffee reduced serum TSH concentrations. In addition, coffee consumption was not associated with the risk of hyperthyroidism and hypothyroidism.

KEYWORDS

**coffee, thyroid function, NHANES, Mendelian randomization, machine learning**

# 1 Introduction

Coffee is among the most consumed beverages worldwide, with approximately 2.25 billion cups consumed daily, amounting to around 500 billion cups per year, as reported by the National Coffee Association (1). It is estimated that over 1,000 compounds can be found in coffee, with the most common being caffeine, chlorogenic acid, and melanoidins. The health effects of coffee consumption have piqued academic interest for many years. Numerous studies have demonstrated that coffee consumption is linked to a reduced risk of various chronic diseases (including type 2 diabetes and cardiovascular disease), cancer, and neurodegenerative diseases such as Parkinson's disease (2). However, the impact on thyroid function remains unclear. The thyroid is the largest endocrine gland in the body (3). Thyroid hormone (TH) is synthesized and secreted by follicular epithelial cells, regulated by thyroid-stimulating hormone (TSH), and stored in the follicular compartment in colloidal form. TH broadly regulates growth, development, metabolism, and other bodily functions (4, 5). Thyroid hormones are iodides of tyrosine and primarily consist of triiodothyronine (T3) and thyroxine (T4). The active free thyroid hormones in the body include free triiodothyronine (FT3) and free thyroxine (FT4) (6). When FT3 and FT4 are depleted in the body, total T3 (TT3) and total T4 (TT4) are converted to FT3 and FT4, which continue to function as thyroid hormones (7, 8). Various lifestyle habits have been shown to cause changes in thyroid hormone levels, such as smoking, alcohol consumption, diet, and exercise (9). In previous animal experiments, caffeine, a substance found in coffee, inhibited TSH secretion by releasing hypothalamic growth inhibitory hormone after intraperitoneal injection into rats (10). However, no study on the effects on the thyroid gland after long-term coffee intake has been conducted. The National Health and Nutrition Examination Survey (NHANES) is a biennial survey of the U.S. population that employs a multi-stage probability sampling design, combining interviews, questionnaires, physical examinations, and laboratory data to assess the health and nutritional status of the population (11). Mendelian randomization (MR) analysis is widely used for causal inference in epidemiology, with the core concept being the use of genetic variation as an instrument variable (IV) to model and test the causal relationship between exposure factors and disease (12). The methodology of the MR study is akin to that of a randomized controlled trial (RCT) because parental alleles are randomly assigned to offspring according to Mendel's law during gamete formation, making the MR study equivalent to a naturally occurring RCT in a population (13). Simultaneously, genetic variants are formed before birth and persist throughout life, allowing MR studies to effectively avoid the influence of reverse causality (14). Therefore, in our study, we will employ a combination of NHANES and MR analysis in an observational study to explore whether coffee consumption causes changes in serum TT4, TT3, FT4, FT3, and TSH concentrations and the causal effect of coffee consumption on thyroid disorders, including hyperthyroidism and hypothyroidism.

# 2 Materials and methods

## 2.1 Study samples in NHANES

The full process of this study is shown in Figure 1. Our study selected data on sub-jects from 2007-2012, a time interval chosen because it was the only time interval during which data on thyroid function were collected by NHANES. A total of 30,442 participants from NHANES 2007-2012, subjects aged 20 years and older were enrolled in our study. In addition, among all subjects, we excluded the following individuals: (1) Individuals with missing thyroid function test indicators. (2) Individuals lacking coffee consumption data. (3) Individuals with incomplete data on

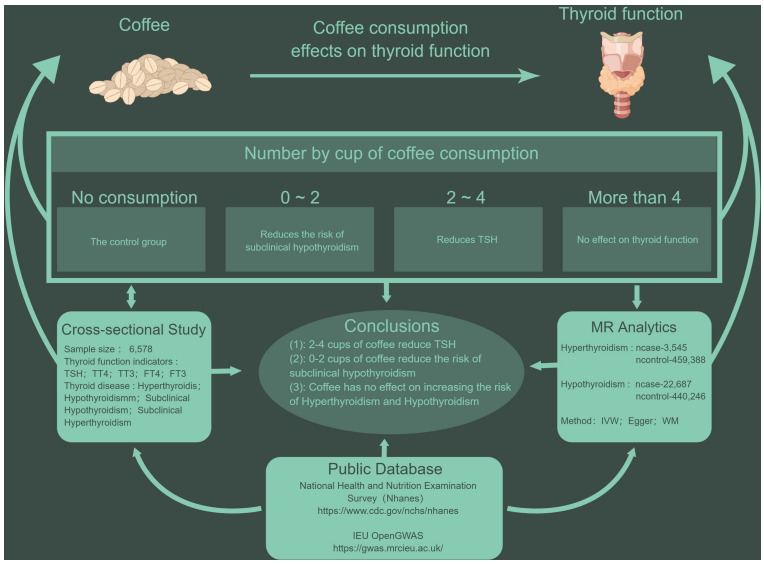



FIGURE 1
The line of research for this study.

education status, diabetes, hypertension, BMI, hyperlipidemia, alcohol consumption, and smoking status. Final study sample size n=6,578 (weighted n = 63,453,885).

## 2.2 Coffee consumption data collection

Food frequency questionnaires and two 24-hour dietary recalls collected during NHANES 2007-2012 were used to obtain coffee consumption data. As presented in previous studies, coffee consumption in NHANES was correlated with specific 8-digit food codes (codes beginning with 921) in the Food and Nutrition Database for Dietary Studies (FNDDS) (15). We selected participants who had two 24-hour dietary recalls and used the average coffee consumption for the two 24-hour recalls. In this study the size of a cup of coffee was defined as 283.5 grams (16). In addition, the coffee intake was further divided into four groups: 0, <2, 2-4, and >4 cups/day.

## 2.3 Thyroid function assessment

FT3, TT3, FT4, TT4 and TSH data in the study were obtained from NHANES laboratory data. A competitive binding immunoenzymatically assay was employed for T3, FT3, and T4 measurements, while a two-step enzyme immunoassay was utilized for FT4 assessment. The evaluation of TSH was carried out using the Access HYPER sensitive human thyroid-stimulating hormone (hTSH) assay, which is a 3rd generation two-site immunoenzymatically assay (17, 18). Diagnosis of hyperthyroidism based on TSH<0.45mIU/and FT4>1.6ng/dL, and hypothyroidism based on TSH>4.5mIU/L and FT4<0.6ng/dL. Sub-clinical hyperthyroidism was diagnosed according to TSH<0.45mIU/L and 0.6<FT4<1.6ng/dL, and subclinical hypothyroidism was diagnosed according to TSH>4.5mIU/L and 0.6<FT4<1.6ng/dL (19).

## 2.4 Acquisition of covariates

Covariates in the study included age, gender (Male and Female), ethnicity (White, Black, Mexican and Other), education (High School Grad/GED or Equivalent, less than 9th Grade, 9th-11th Grade, College graduate or above), body mass index (BMI), smoking (Never, Former, Now), and alcohol use (Never, Former, Mild, Moderate, Heavy). In addition, by screening related studies, we additionally included urine iodine, diabetes (No and Yes), hypertension (No and Yes) and hyperlipidemia (No and Yes) data for analysis (20–24).

## 2.5 Acquisition of IV of MR

The IV used in this study was obtained from a study by Li et al. (23). We used two groups of IV for MR analysis. The first group IV (IV1) uses the number of cups of coffee consumed per day (cups/day) by coffee consumers as an instrument, and the second group IV (IV2) compares regular versus infrequent coffee consumers. After screening for $P<5\times10^{-8}$ and excluding those SNPs in a state of linkage disequilibrium (r2 = 0.01, 10,000 kb), 6 and 3 SNPs were finally included in IV1 and IV2, respectively.

## 2.6 Genetic data of hyperthyroidism and hypothyroidism

Genetic data for hyperthyroidism were obtained from a study compiled and published by Ben Elsworth et al. in 2018, contains 3545 European cases and 459,388 European controls, including a total of 9,851,867 SNPs. Genetic data for hypothyroidism were also obtained from the Ben Elsworth study, which included 22,687 European cases and 440,246 European controls. Genetic data for hyperthyroidism and hypothyroid-ism can be obtained from https://gwas.mrcieu.ac.uk/datasets/ukb-b-20289/ and https://gwas.mrcieu.ac.uk/datasets/ukb-b-19732/.

## 2.7 Statistical analysis

All statistical analyses were performed in R software (4.2.1), and survey data in Table 1 were summarized by descriptive statistics, one-way ANOVA for continuous variables, and chi-square tests for categorical variables to assess the association between coffee consumption and other factors in different groups. We used the sampling weights provided by NHANES for weighting (21). Weighted generalized linear regression models were used to assess the relationship between adjusted coffee consumption and serum thyroid function indicators. In addition, we used weighted multivariate adjusted logistic regression to calculate the odds ratio (OR) and 95% confidence interval (CI) for thyroid disease in the different coffee consumption groups. All analyses considered the NHANES complex multi-stage sampling design and $p < 0.05$ was considered statistically significant. The MR analyses were all performed using the TwoSampleMR package (13). In this study, multiple SNPs were used as IVs for the MR study. The association of individual SNPs was first performed, and the Wald ratio was calculated for each SNP, and then the Wald ratios were combined using the inverse variance weighted (IVW) method to assess the association between coffee consumption and thyroid disease. To check the robustness of the results, we additionally used MR-Egger regression and weighted median estimator (WME) for additional analysis, these three calculation methods have different validity assumptions for IV. In addition, the mean pleiotropic effect of genetic variants could be assessed using the intercept of the MR-Egger regression (tested using $P < 0.05$) and the Cochran's Q test was used to determine the heterogeneity between the causal estimates of different genetic variants. Eventually a sensitivity analysis was also performed using the leave-one-out method.

10.3389/fendo.2023.1188547

TABLE 1 Baseline table of participants categorized by number of cups of coffee consumed.

| Variable | Total | 0 | <2 | 2-4 | >4 | P |
|---|---|---|---|---|---|---|
| N(Unweighted) | 6578(100.00) | 2741(41.66) | 1969(29.93) | 1204(18.30) | 664(10.09) | |
| Age | 51.05(50.28,51.82) | 46.36(45.15,47.57) | 52.84(51.74,53.94) | 56.19(54.96,57.43) | 55.24(53.73,56.75) | 0.01 |
| Sex | | | | | | 0.01 |
| Male | 3248(49.38) | 1358(49.40) | 876(43.86) | 618(52.73) | 396(59.71) | |
| Female | 3330(50.62) | 1383(50.60) | 1093(56.14) | 586(47.27) | 268(40.29) | |
| Eth | | | | | | 0.01 |
| white | 3201(48.66) | 1217(42.71) | 736(35.73) | 721(60.01) | 527(80.31) | |
| Black | 1298(19.73) | 779(29.83) | 345(16.44) | 142(10.32) | 32(4.99) | |
| Mexican | 1010(15.35) | 391(14.07) | 417(23.49) | 159(14.79) | 43(5.34) | |
| Other | 1069(16.25) | 354(13.39) | 471(24.34) | 182(14.88) | 62(9.36) | |
| Edu | | | | | | 0.01 |
| Less than 9th Grade | 723(10.99) | 230(8.62) | 285(15.77) | 151(14.40) | 57(8.63) | |
| 9-11th Grade | 1061(16.13) | 440(15.63) | 347(17.74) | 163(13.87) | 111(16.71) | |
| High School Grad/GED or Equivalent | 1562(23.75) | 671(24.50) | 434(22.85) | 287(24.56) | 170(26.78) | |
| College graduate or above | 3232(49.13) | 1400(51.25) | 903(43.64) | 603(47.17) | 326(47.87) | |
| Smoke | | | | | | 0.01 |
| Never | 3492(53.09) | 1734(64.81) | 1112(56.56) | 471(40.84) | 175(26.82) | |
| Former | 1745(26.53) | 533(18.60) | 528(25.94) | 456(36.98) | 228(36.72) | |
| Now | 1341(20.39) | 474(16.59) | 329(17.50) | 277(22.18) | 261(36.46) | |
| Alcohol User | | | | | | 0.01 |
| Never | 931(14.15) | 475(17.72) | 329(18.55) | 91(8.66) | 36(4.77) | |
| Former | 1299(19.75) | 504(17.55) | 375(18.33) | 262(23.84) | 158(23.84) | |
| Mild | 2081(31.64) | 761(27.37) | 628(29.29) | 465(36.93) | 227(37.90) | |
| Moderate | 972(14.78) | 406(14.87) | 275(14.21) | 177(13.29) | 114(15.61) | |
| Heavy | 1295(19.69) | 595(22.48) | 362(19.63) | 209(17.28) | 129(17.88) | |
| DM | | | | | | 0.01 |
| No | 5686(86.44) | 2433(88.57) | 1669(84.28) | 1020(84.30) | 564(86.14) | |
| Yes | 892(13.56) | 308(11.43) | 300(15.72) | 184(15.70) | 100(13.86) | |
| Hyperlipidemia | | | | | | 0.01 |
| No | 1617(24.58) | 797(28.38) | 428(20.86) | 255(22.45) | 137(22.32) | |
| Yes | 4961(75.42) | 1944(71.62) | 1541(79.14) | 949(77.55) | 527(77.68) | |
| Hypertension | | | | | | 0.09 |
| No | 3727(56.66) | 1642(58.61) | 1088(53.22) | 637(54.87) | 360(55.61) | |
| Yes | 2851(43.34) | 1099(41.39) | 881(46.78) | 567(45.13) | 304(44.39) | |
| Hyperthyroidism | | | | | | 0.33 |
| No | 6562(99.76) | 2734(99.72) | 1965(99.88) | 1202(99.86) | 661(99.42) | |
| Yes | 16(0.24) | 7(0.28) | 4(0.12) | 2(0.14) | 3(0.58) | |
| Hypothyroidism | | | | | | 0.17 |

*(Continued)*

10.3389/fendo.2023.1188547

TABLE 1 Continued

| Variable | Total | 0 | <2 | 2-4 | >4 | P |
|---|---|---|---|---|---|---|
| No | 6544(99.48) | 2728(99.68) | 1961(99.71) | 1197(99.78) | 658(99.11) | |
| Yes | 34(0.52) | 13(0.32) | 8(0.29) | 7(0.22) | 6(0.89) | |
| Subclinical Hyperthyroidism | | | | | | 0.96 |
| No | 6373(96.88) | 2661(96.80) | 1905(96.81) | 1162(96.41) | 645(96.63) | |
| Yes | 205(3.12) | 80(3.20) | 64(3.19) | 42(3.59) | 19(3.37) | |
| Subclinical Hypothyroidism | | | | | | 0.24 |
| No | 6339(96.37) | 2637(96.19) | 1905(97.27) | 1157(95.31) | 640(96.45) | |
| Yes | 239(3.63) | 104(3.81) | 64(2.73) | 47(4.69) | 24(3.55) | |
| BMI (kg/m $^2$) | 29.08(28.80,29.35) | 29.36(28.96,29.76) | 28.70(28.35,29.04) | 29.23(28.71,29.76) | 28.74(28.17,29.32) | 0.02 |
| TT3 (ng/dL) | 112.20(110.97,113.44) | 113.90(112.17,115.62) | 111.55(110.08,113.02) | 110.59(108.15,113.03) | 110.22(108.22,112.22) | 0.01 |
| TT4 (ug/dL) | 7.93(7.85,8.00) | 7.93(7.84,8.03) | 7.99(7.89,8.10) | 7.89(7.76,8.02) | 7.78(7.63,7.94) | 0.03 |
| FT3 (pg/mL) | 3.15(3.13,3.17) | 3.18(3.15,3.21) | 3.13(3.10,3.16) | 3.13(3.09,3.17) | 3.10(3.06,3.15) | 0.02 |
| FT4 (ng/dL) | 0.80(0.79,0.81) | 0.80(0.79,0.81) | 0.80(0.79,0.81) | 0.80(0.79,0.82) | 0.80(0.78,0.82) | 0.98 |
| TSH (mIU/L) | 2.01(1.90,2.12) | 1.97(1.86,2.08) | 2.02(1.86,2.17) | 1.95(1.83,2.07) | 2.30(1.79,2.81) | 0.52 |
| Urinary Iodine | 285.45(242.56,328.34) | 302.00(229.73,374.28) | 297.34(205.32,389.35) | 277.98(205.63,350.33) | 191.66(163.06,220.25) | 0.01 |

Edu, education, Eth, ethnicity Categorical variables show percentages, continuous variables show means and confidence intervals. A cup of coffee cup is defined as 283.5 grams.

# 3 Results

## 3.1 Baseline characteristics of participants

After screening as required, a total of 6578 subjects were included in this study, and among the general demographic data, 3330 (50.62%) were female, 3201 (48.66%) were white, 3232 (49.13%) were College graduate or above, 3492 (53.9%) were non-smokers, and 5647 (85.8%) were alcohol users above their corresponding groups. In addition, a total of 4961 cases were diagnosed with Hyperlipidemia, 2851 with Hyper-tension, and 892 with DM. 494 of these subjects were diagnosed with thyroid abnormalities, including 16 patients with hyperthyroidism, 34 with hypothyroidism, 205 with subclinical hyperthyroidism, and 239 with subclinical hypothyroidism (Table 1).

## 3.2 Relationship between coffee consumption and serum thyroid function indicators

In our study, 3837 (58.3%) were coffee drinkers, with <2 cups accounting for 29.93% of participants. After adjusting for gender, education, race, age, smoking, alcohol consumption, diabetes, hyperlipidemia, hypertension, BMI and urinary iodine, we found that those who consumed 2-4 cups of coffee per day had significantly lower levels of TSH compared to those who did not drink coffee. (b=-0.23, 95% CI: -0.30, -0.16), but there was no statistically significant association for TT4, FT4, TT3 and FT3 (p>0.05) (Table 2).

## 3.3 Relationship between coffee consumption and thyroid disorders

Weighted logistic regression results showed that consumption of 0-2 cups of coffee per day was negatively associated with the risk of developing subclinical hypothyroid-ism. (OR=0.60, 95% CI: 0.41, 0.88), but there was no significant association between coffee intake and thyroid disease in other coffee consumption categories (Table 3).

## 3.4 MR analysis

The results of the study showed that coffee consumption calculated by IVW, WME and MR-Egger regression methods for IV1 and IV2 were not statistically significant with hyperthyroidism or hypothyroidism (Table 4), implying that there was no causal relationship between coffee consumption on hyperthyroidism and hypothyroidism in the population. We then validated the reliability of our results. The results showed that there was no heterogeneity or pleiotropy in our study. (S1 and S2) (p>0.05). Sensitivity analysis of the IVW results using the leave-one-out method showed that the elimination of SNPs one by one did not reveal that a particular SNP caused a significant change in the results, and no SNPs with a strong

10.3389/fendo.2023.1188547

TABLE 2 Effect of coffee consumption on serum thyroid function indicators.

| Variable | Coffee | Beta | SE | t value | P |
|---|---|---|---|---|---|
| TSH | ref | ref | ref | ref | ref |
| | <2 | -0.06 | 0.09 | -0.62 | 0.54 |
| | 2-4 | -0.23 | 0.07 | -3.27 | 0.003 ** |
| | >4 | 0.11 | 0.26 | 0.44 | 0.66 |
| TT4 | ref | ref | ref | ref | ref |
| | <2 | -0.06 | 0.07 | -0.77 | 0.45 |
| | 2-4 | -0.07 | 0.08 | -0.81 | 0.43 |
| | >4 | -0.06 | 0.10 | -0.59 | 0.56 |
| TT3 | ref | ref | ref | ref | ref |
| | <2 | -0.43 | 0.91 | -0.47 | 0.64 |
| | 2-4 | 0.02 | 1.28 | 0.02 | 0.99 |
| | >4 | -0.67 | 1.25 | -0.53 | 0.60 |
| FT4 | ref | ref | ref | ref | ref |
| | <2 | -0.01 | 0.01 | -1.44 | 0.16 |
| | 2-4 | -0.01 | 0.01 | -1.00 | 0.32 |
| | >4 | -0.01 | 0.01 | -0.69 | 0.49 |
| FT3 | ref | ref | ref | ref | ref |
| | <2 | -0.01 | 0.02 | -0.60 | 0.56 |
| | 2-4 | 0.02 | 0.03 | 0.68 | 0.50 |
| | >4 | -0.02 | 0.02 | -1.03 | 0.31 |

**Tips less than 0.05, **Tips less than 0.01, Models adjusted for age, sex, education, ethnicity, smoking, alcohol, DM, hypertension, hyperlipidemia, BMI, and urinary iodine concentration.

TABLE 3 Effect of coffee consumption on serum thyroid function indicators.

| Variable | Coffee | OR | 95%CI | | P |
|---|---|---|---|---|---|
| Hyperthyroidism | ref | ref | ref | ref | ref |
| | <2 | 0.33 | 0.05 | 2.20 | 0.26 |
| | 2-4 | 0.41 | 0.04 | 4.03 | 0.45 |
| | >4 | 1.69 | 0.16 | 17.59 | 0.66 |
| Hypothyroidism | ref | ref | ref | ref | ref |
| | <2 | 0.94 | 0.26 | 3.36 | 0.92 |
| | 2-4 | 0.54 | 0.13 | 2.24 | 0.40 |
| | >4 | 1.98 | 0.52 | 7.45 | 0.32 |
| Subclinical. Hyperthyroidism | ref | ref | ref | ref | ref |
| | <2 | 0.93 | 0.61 | 1.40 | 0.72 |
| | 2-4 | 1.03 | 0.58 | 1.82 | 0.93 |
| | >4 | 0.98 | 0.46 | 2.07 | 0.96 |
| Subclinical. Hypothyroidism | ref | ref | ref | ref | ref |

*(Continued)*

TABLE 3 Continued

| Variable | Coffee | OR | 95%CI | | P |
|---|---|---|---|---|---|
| | <2 | 0.60 | 0.41 | 0.89 | 0.015* |
| | 2-4 | 0.89 | 0.59 | 1.36 | 0.61 |
| | >4 | 0.65 | 0.25 | 1.72 | 0.40 |

**Tips less than 0.05, **Tips less than 0.01, Models adjusted for age, sex, education, ethnicity, smoking, alcohol, DM, hypertension, hyperlipidemia, BMI, and urinary iodine concentration.

effect on the results were found in IV, indicating that the effect ORs derived from the previous IVW method were more robust.

# 4 Discussion

Coffee consumption ranks second only to water, yet the debate over its benefits and risks continues (25). Epidemiological studies investigating the relationship between coffee consumption and thyroid function are scarce. Hyperthyroidism is a condition resulting from excessive production of thyroid hormones. Common symptoms include weight loss, rapid heart rate, anxiety, and irritability. Serious complications may involve thyroid crisis, atrial fibrillation, and bone fractures. On the other hand, hypothyroidism is characterized by a deficiency in thyroid hormones, potentially leading to symptoms such as weight gain, fatigue, and depression. Severe complications of hypothyroidism include myxedema coma, cardiovascular disease, and pregnancy-related issues. Both hyperthyroidism and hypothyroidism can negatively impact a patient's quality of life and psychological well-being; therefore, timely diagnosis and treatment are crucial (26, 27). Animal studies have found that caffeine may be associated with a decrease in TSH concentration (10), suggesting that coffee consumption could have endocrine-disrupting effects on thyroid function. L-thyroxine is a synthetic form of thyroid hormone (28). In various studies, coffee consumption has been shown to interfere with the absorption of L-thyroxine (29–33). In a study conducted by Marija Andjelkovic (34), which included 150 patients on thyroid replacement therapy with cardiovascular disease, they found that cigarette smoking was a risk factor that decreased TSH levels in patients on thyroid replacement therapy but did not find an effect of coffee consumption on patients' TSH. An additional RCT included 11 healthy subjects, whose thyroid function levels were measured after receiving different types of coffee oil, and showed virtually no change in T4, T3, and TSH concentrations (35). However, these studies were influenced by the number of patients and the patients' own underlying diseases, so large sample studies are needed to explore the effects of coffee consumption on thyroid function. Subclinical hypothyroidism affects up to 10% of adults, and patients with subclinical hypothyroidism are at significant risk of progressing to hypothyroidism (36). Additionally, subclinical hypothyroidism is associated with a variety of cardiovascular diseases and all-cause mortality (37, 38). Our study revealed that coffee consumption of <2 cups per day effectively reduced the risk of developing subclinical hypothyroidism (OR=0.60, 95% CI 0.41-0.88) (Table 3), a finding with some clinical significance. We speculate that the mechanism may be due to coffee consumption helping to control TSH concentration within the normal range of 0.45 mIU/-4.5 mIU/L. Several studies exploring the effects of coffee consumption on humans using MR methods have been reported. Nordestgaard et al. (39) found that coffee consumption prevented symptomatic gallstone disease, while Kim et al. (40) showed that higher coffee consumption was associated with a lower risk of arrhythmias. These studies were analyzed based on MR methods. The results of this study, which analyzed data from subjects in NHANES, showed no effect of coffee consumption on the risk of hyperthyroidism and hypothyroidism. We then analyzed the causal relationship between coffee consumption and hyperthyroidism and hypothyroidism separately using the MR method, and the results remained non-significant, which validates our retrospective cross-sectional study and again supports the conclusion that coffee consumption is not associated with the risk of hyperthyroidism and hypothyroidism (Table 4). We also performed additional

TABLE 4 OR estimates and 95% CI for IVW, WME and MR-Egger regression.

| IV | Exposure | Outcome | Method | nSNP | OR | 95%CI | | P |
|---|---|---|---|---|---|---|---|---|
| IV_1 | Coffee | Hyperthyroidism | MR Egger | 6 | 1.00 | 1.00 | 1.01 | 0.16 |
| | Coffee | Hyperthyroidism | Weighted median | 6 | 1.00 | 1.00 | 1.00 | 0.24 |
| | Coffee | Hyperthyroidism | IVW | 6 | 1.00 | 1.00 | 1.00 | 0.17 |
| IV_2 | Coffee | Hyperthyroidism | MR Egger | 3 | 1.00 | 1.00 | 1.01 | 0.50 |
| | Coffee | Hyperthyroidism | Weighted median | 3 | 1.00 | 1.00 | 1.00 | 0.24 |
| | Coffee | Hyperthyroidism | IVW | 3 | 1.00 | 1.00 | 1.00 | 0.21 |

*(Continued)*

TABLE 4 Continued

| IV | Exposure | Outcome | Method | nSNP | OR | 95%CI | | P |
|---|---|---|---|---|---|---|---|---|
| IV_1 | Coffee | Hypothyroidism | MR Egger | 6 | 1.01 | 0.99 | 1.02 | 0.28 |
| | Coffee | Hypothyroidism | Weighted median | 6 | 1.01 | 1.00 | 1.01 | 0.06 |
| | Coffee | Hypothyroidism | IVW | 6 | 1.00 | 1.00 | 1.01 | 0.21 |
| IV_2 | Coffee | Hypothyroidism | MR Egger | 3 | 1.01 | 1.00 | 1.02 | 0.24 |
| | Coffee | Hypothyroidism | Weighted median | 3 | 1.00 | 1.00 | 1.01 | 0.06 |
| | Coffee | Hypothyroidism | IVW | 3 | 1.00 | 1.00 | 1.01 | 0.29 |

IVW, Inverse variance weighting.

validation of pleiotropy and heterogeneity, as well as sensitivity analysis of the results using the leave-one-out method, to make the results more reliable. Nevertheless, this study has several limitations, such as the fact that compounds in coffee are affected by various factors, including the type of coffee, brewing method, and degree of coffee roasting. Additionally, we were unable to distinguish between caffeinated and decaffeinated coffee consumption. In this study, the source of data on coffee consumption was the same as the two 24-hour recall interviews in NHANES. Due to the retrospective nature of data collection, it does not accurately reflect individuals' regular intake. To minimize bias, we excluded subjects who had only one 24-hour recall interview and averaged the data obtained from the two 24-hour recalls as coffee consumption. It has been suggested that two 24-hour recalls may be sufficient to assess daily dietary consumption (41). The data from the genome-wide association analysis used in our study were obtained from European populations, which limits the generalizability of our findings. Further research is needed among different ethnic groups to yield more comprehensive results.

In conclusion, our study has shed light on the relationship between coffee consumption and thyroid function, yet there are still aspects that warrant further exploration. To address the limitations of the current study and provide more robust evidence, large-scale, multi-ethnic, and prospective studies are required. As our understanding of the potential effects of coffee consumption on thyroid function grows, it will be crucial to consider various factors, such as the type of coffee, brewing method, and degree of coffee roasting, in future research.

## 5 Conclusions

Our study demonstrated that drinking <2 cups of coffee per day reduced the risk of subclinical hypothyroidism (OR=0.60, 95% CI: 0.41, 0.88) and 2-4 cups of coffee reduced serum TSH concentrations (b=-0.23, 95% CI: -0.30, -0.16) compared to no coffee consumption. In addition, coffee consumption was not associated with the risk of hyperthyroidism and hypothyroidism.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

## Ethics statement

The studies involving human participants were reviewed and approved by National Center for Health Statistics. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

Conceptualization: GZ and RC. Methodology: GZ. Software: JJ. Validation: GZ, RC, and ZW. Formal analysis: GZ. Investigation: JJ. Resources: ZW. Data curation: RC. Writing—original draft preparation: GZ. Writing—review and editing: RC. Visualization: JJ. Supervision: RC. Project administration: RC. Funding acquisition: RC. All authors have read and agreed to the published version of the manuscript. All authors contributed to the article.

## Funding

This research was funded by Natural Science Foundation of Heilongjiang Province, grant number SS2022H003. The APC was funded by SS2022H003.

## Acknowledgments

We would like to thank Figdraw (https://www.figdraw.com/) for the drawing help on Figure 1. Thanks to Zhang Jing (Shanghai Tongren Hospital) for his work on the NHANES database. His outstanding work, nhanesR package and webpage, makes it easier for us to explore NHANES database.

## Conflict of interest

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.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1188547/full#supplementary-material

## References


1. Surma S, Oparil S. Coffee and arterial hypertension. *Curr Hypertens Rep* (2021) 23:38. doi: 10.1007/s11906-021-01156-3

2. Pietzner M, Kohrle J, Lehmphul I, Budde K, Kastenmuller G, Brabant G, et al. A thyroid hormone-independent molecular fingerprint of 3,5-diiodothyronine suggests a strong relationship with coffee metabolism in humans. *Thyroid* (2019) 29:1743–54. doi: 10.1089/thy.2018.0549

3. Yen PM. Physiological and molecular basis of thyroid hormone action. *Physiol Rev* (2001) 81:1097–142. doi: 10.1152/physrev.2001.81.3.1097

4. Cheng SY, Leonard JL, Davis PJ. Molecular aspects of thyroid hormone actions. *Endocr Rev* (2010) 31:139–70. doi: 10.1210/er.2009-0007

5. Williams GR. Neurodevelopmental and neurophysiological actions of thyroid hormone. *J Neuroendocrinol* (2008) 20:784–94. doi: 10.1111/j.1365-2826.2008.01733.x

6. Brent GA. Mechanisms of thyroid hormone action. *J Clin Invest* (2012) 122:3035–43. doi: 10.1172/JCI60047

7. Brdar D, Gunjaca I, Pleic N, Torlak V, Knezevic P, Punda A, et al. The effect of food groups and nutrients on thyroid hormone levels in healthy individuals. *Nutrition* (2021) 91-92:111394. doi: 10.1016/j.nut.2021.111394

8. Aoun EG, Lee MR, Haass-Koffler CL, Swift RM, Addolorato G, Kenna GA, et al. Relationship between the thyroid axis and alcohol craving. *Alcohol Alcohol* (2015) 50:24–9. doi: 10.1093/alcalc/agu085

9. Babic Leko M, Gunjaca I, Pleic N, Zemunik T. Environmental factors affecting thyroid-stimulating hormone and thyroid hormone levels. *Int J Mol Sci* (2021) 22:6521. doi: 10.3390/ijms22126521

10. Spindel E, Arnold M, Cusack B, Wurtman RJ. Effects of caffeine on anterior pituitary and thyroid function in the rat. *J Pharmacol Exp Ther* (1980) 214:58–62.

11. Ahluwalia N, Dwyer J, Terry A, Moshfegh A, Johnson C. Update on NHANES dietary data: focus on collection, release, analytical considerations, and uses to inform public policy. *Adv Nutr* (2016) 7:121–34. doi: 10.3945/an.115.009258

12. Sanderson E. Multivariable mendelian randomization and mediation. *Cold Spring Harb Perspect Med* (2021) 11:a038984. doi: 10.1101/cshperspect.a038984

13. Emdin CA, Khera AV, Kathiresan S. Mendelian randomization. *JAMA* (2017) 318:1925–6. doi: 10.1001/jama.2017.17219

14. Ference BA, Holmes MV, Smith GD. Using mendelian randomization to improve the design of randomized trials. *Cold Spring Harb Perspect Med* (2021) 11: a040980. doi: 10.1101/cshperspect.a040980

15. Forbes GB. A note on the mathematics of "catch-up" growth. *Pediatr Res* (1974) 8:929–31. doi: 10.1203/00006450-197412000-00002

16. Wang M, Jian Z, Yuan C, Jin X, Li H, Wang K. Coffee consumption and prostate cancer risk: results from national health and nutrition examination survey 1999-2010 and mendelian randomization analyses. *Nutrients* (2021) 13:2317. doi: 10.3390/nu13072317

17. Kim K, Argos M, Persky VW, Freels S, Sargis RM, Turyk ME. Associations of exposure to metal and metal mixtures with thyroid hormones: results from the nhanes 2007-2012. *Environ Res* (2022) 212(Pt C):113413. doi: 10.1016/j.envres.2022.113413

18. Hollowell JG, Staehling NW, Flanders WD, Hannon WH, Gunter EW, Spencer CA, et al. Serum tsh, T(4), and thyroid antibodies in the united states population (1988 to 1994): national health and nutrition examination survey (Nhanes iii). *J Clin Endocrinol Metab* (2002) 87(2):489–99. doi: 10.1210/jcem.87.2.8182

19. Jain RB. Associations between the levels of thyroid hormones and lipid/lipoprotein levels: data from national health and nutrition examination survey 2007-2012. *Environ Toxicol Pharmacol* (2017) 53:133–44. doi: 10.1016/j.etap.2017.05.002

20. Gonzalez-Nunez A, Garcia-Solis P, Ramirez-Garcia SG, Flores-Ramirez G, Vela-Amieva M, Lara-Diaz VJ, et al. High iodine urinary concentration is associated with high TSH levels but not with nutrition status in schoolchildren of northeastern Mexico. *Nutrients* (2021) 13:3975. doi: 10.3390/nu13113975

21. He W, Li S, Wang B, Mu K, Shao X, Yao Q, et al. Dose-response relationship between thyroid stimulating hormone and hypertension risk in euthyroid individuals. *J Hypertens* (2019) 37:144–53. doi: 10.1097/HJH.0000000000001826

22. Su X, Peng H, Chen X, Wu X, Wang B. Hyperlipidemia and hypothyroidism. *Clin Chim Acta* (2022) 527:61–70. doi: 10.1016/j.cca.2022.01.006

23. Nerurkar PV, Gandhi K, Chen JJ. Correlations between coffee consumption and metabolic phenotypes, plasma folate, and vitamin B12: NHANES 2003 to 2006. *Nutrients* (2021) 13:1348. doi: 10.3390/nu13041348

24. Rong F, Dai H, Wu Y, Li J, Liu G, Chen H, et al. Association between thyroid dysfunction and type 2 diabetes: a meta-analysis of prospective observational studies. *BMC Med* (2021) 19:257. doi: 10.1186/s12916-021-02121-2

25. Li X, Cheng S, Cheng J, Wang M, Zhong Y, Yu AY. Habitual coffee consumption increases risk of primary open-angle glaucoma: a mendelian randomization study. *Ophthalmology* (2022) 129:1014–21. doi: 10.1016/j.ophtha.2022.04.027

26. Wiersinga WM, Poppe KG, Effraimidis G. Hyperthyroidism: aetiology, pathogenesis, diagnosis, management, complications, and prognosis. *Lancet Diabetes Endocrinol* (2023) 11(4):282–98. doi: 10.1016/S2213-8587(23)00005-0

27. Andreoli M, Centanni M. [Hypothyroidism: current clinical, physiopathological and therapeutic aspects]. *Recenti Prog Med* (1991) 82(6):344–51.

28. Coffee and Caffeine Genetics, C, Cornelis MC, Byrne EM, Esko T, Nalls MA, Ganna A, et al. Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Mol Psychiatry* (2015) 20:647–56. doi: 10.1038/mp.2014.107

29. Butt MS, Sultan MT. Coffee and its consumption: benefits and risks. *Crit Rev Food Sci Nutr* (2011) 51:363–73. doi: 10.1080/10408390903586412

30. Sue LY, Leung AM. Levothyroxine for the treatment of subclinical hypothyroidism and cardiovascular disease. *Front Endocrinol (Lausanne)* (2020) 11:591588. doi: 10.3389/fendo.2020.591588

31. Liwanpo L, Hershman JM. Conditions and drugs interfering with thyroxine absorption. *Best Pract Res Clin Endocrinol Metab* (2009) 23:781–92. doi: 10.1016/j.beem.2009.06.006

32. Benvenga S, Bartolone L, Pappalardo MA, Russo A, Lapa D, Giorgianni G, et al. Altered intestinal absorption of l-thyroxine caused by coffee. *Thyroid* (2008) 18:293–301. doi: 10.1089/thy.2007.0222

33. Wiesner A, Gajewska D, Pasko P. Levothyroxine interactions with food and dietary supplements-a systematic review. *Pharm (Basel)* (2021) 14:206. doi: 10.3390/ph14030206

34. Andjelkovic M, Jankovic S, Mitrovic M, Mladenovic V, Nikolic I, Zelen I, et al. Effects of cardiovascular drugs on TSH serum levels in patients on replacement therapy after thyroidectomy. *Int J Clin Pharmacol Ther* (2016) 54:628–33. doi: 10.5414/CP202606

35. Mensink RP, Lebbink WJ, Lobbezoo IE, Weusten-Van der Wouw MP, Zock PL, Katan MB. Diterpene composition of oils from arabica and robusta coffee beans and their effects on serum lipids in man. *J Intern Med* (1995) 237:543–50. doi: 10.1111/j.1365-2796.1995.tb00883.x

36. Biondi B, Cappola AR, Cooper DS. Subclinical hypothyroidism: a review. *JAMA* (2019) 322:153–60. doi: 10.1001/jama.2019.9052

37. Inoue K, Ritz B, Brent GA, Ebrahimi R, Rhee CM, Leung AM. Association of subclinical hypothyroidism and cardiovascular disease with mortality. *JAMA Netw Open* (2020) 3:e1920745. doi: 10.1001/jamanetworkopen.2019.20745

38. Manolis AA, Manolis TA, Melita H, Manolis AS. Subclinical thyroid dysfunction and cardiovascular consequences: an alarming wake-up call? *Trends Cardiovasc Med* (2020) 30:57–69. doi: 10.1016/j.tcm.2019.02.011

39. Nordestgaard AT, Stender S, Nordestgaard BG, Tybjaerg-Hansen A. Coffee intake protects against symptomatic gallstone disease in the general population: a mendelian randomization study. *J Intern Med* (2020) 287:42–53. doi: 10.1111/joim.12970

40. Kim EJ, Hoffmann TJ, Nah G, Vittinghoff E, Delling F, Marcus GM. Coffee consumption and incident tachyarrhythmias: reported behavior, mendelian randomization, and their interactions. *JAMA Intern Med* (2021) 181:1185–93. doi: 10.1001/jamainternmed.2021.3616

41. Knuppel S, Norman K, Boeing H. Is a single 24-hour dietary recall per person sufficient to estimate the population distribution of usual dietary intake? *J Nutr* (2019) 149:1491–2. doi: 10.1093/jn/nxz118

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 07 June 2023
DOI 10.3389/fendo.2023.1185799



OPEN ACCESS

EDITED BY
Wenjie Shi,
Otto von Guericke University Magdeburg, Germany

REVIEWED BY
Zhaomin Yao,
Northeastern University, China
Dan Zhang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Ke Xu
nsmcxuke@163.com
Juan Chen
chenjuan@usc.edu.cn
Bin Jiang
JiangBin@fsyy.usc.edu.cn

†These authors have contributed equally to this work

RECEIVED 14 March 2023
ACCEPTED 12 April 2023
PUBLISHED 07 June 2023

CITATION
Huang G, Xiao S, Jiang Z, Zhou X, Chen L, Long L, Zhang S, Xu K, Chen J and Jiang B (2023) Machine learning immune-related gene based on KLRB1 model for predicting the prognosis and immune cell infiltration of breast cancer. *Front. Endocrinol.* 14:1185799. doi: 10.3389/fendo.2023.1185799



# Machine learning immune-related gene based on KLRB1 model for predicting the prognosis and immune cell infiltration of breast cancer

Guo Huang[1,2†], Shuhui Xiao[3†], Zhan Jiang[3†], Xue Zhou[4], Li Chen[5], Lin Long[1], Sheng Zhang[6], Ke Xu[3*], Juan Chen[7*] and Bin Jiang[8*]

[1]Hengyang Medical School, University of South China, Hengyang, Hunan, China, [2]The Second Affiliated Hospital, Department of Breast and Thyroid Surgery, Hengyang Medical School, University of South China, Hengyang, Hunan, China, [3]Department of Oncology, Chongqing General Hospital, Chongqing, China, [4]Department of Oncology, The Affiliated Hospital of Southwest Medical University, Luzhou, China, [5]Department of Ultrasonography, Chengdu First People's Hospital, Chengdu, China, [6]Department of Radiology, Nanchong Central Hospital, The Second Clinical Medical College, North Sichuan Medical College, Nanchong, China, [7]The Second Affiliated Hospital, Department of Radiotherapy, Hengyang Medical School, University of South China, Hengyang, Hunan, China, [8]The Second Affiliated Hospital, Department of Burn and Plastic Surgery, Hengyang Medical School, University of South China, Hengyang, Hunan, China

**Objective:** Breast cancer is a prevalent malignancy that predominantly affects women. The development and progression of this disease are strongly influenced by the tumor microenvironment and immune infiltration. Therefore, investigating immune-related genes associated with breast cancer prognosis is a crucial approach to enhance the diagnosis and treatment of breast cancer.

**Methods:** We analyzed data from the TCGA database to determine the proportion of invasive immune cells, immune components, and matrix components in breast cancer patients. Using this data, we constructed a risk prediction model to predict breast cancer prognosis and evaluated the correlation between KLRB1 expression and clinicopathological features and immune invasion. Additionally, we investigated the role of KLRB1 in breast cancer using various experimental techniques including real-time quantitative PCR, MTT assays, Transwell assays, Wound healing assays, EdU assays, and flow cytometry.

**Results:** The functional enrichment analysis of immune and stromal components in breast cancer revealed that T cell activation, differentiation, and regulation, as well as lymphocyte differentiation and regulation, play critical roles in determining the status of the tumor microenvironment. These DEGs are therefore considered key factors affecting TME status. Additionally, immune-related gene risk models were constructed and found to be effective predictors of breast cancer prognosis. Further analysis through KM survival analysis and univariate and multivariate Cox regression analysis demonstrated that KLRB1 is an independent prognostic factor for breast cancer. KLRB1 is closely associated with immunoinfiltrating cells. Finally, *in vitro* experiments confirmed that overexpression of KLRB1 inhibits breast cancer cell proliferation, migration,

invasion, and DNA replication ability. KLRB1 was also found to inhibit the proliferation of breast cancer cells by blocking cell division in the G1/M phase.

**Conclusion:** KLRB1 may be a potential prognostic marker and therapeutic target associated with the microenzymic environment of breast cancer tumors, providing a new direction for breast cancer treatment.

KEYWORDS

breast cancer, KLRB1, prognostic model, immune-related gene, immune infiltration

# 1 Introduction

Breast cancer (BC) is the most predominant type of carcinoma in women, comprising 30% of all cancers affecting females and a mortality rate of around 15% (1). However, primary breast tumors alone are not the leading cause of death in BC patients, while drug resistance, recurrence, and metastasis are the leading causes of increased mortality. Following the onset of metastasis, the survival rate after five years is merely 25% (2). Around 2.09 million new cases of BC were reported globally in 2018, of which roughly 620,000 individuals died due to the disease. The incidence of BC is highest among Chinese women. The projected instances of BC in China are expected to rise to 2.5 million by 2021 (3).

The colonization of tumor cells in normal tissues, the stromal cells, and immune cells that coexist with these tumor cells and the factors they secrete, the vascular endothelial cells, and the extracellular matrix collectively form the tumor microenvironment (TME) (4). The tumor cells prompt the recruitment and activation of immune cells and stromal components, creating a tumor-suppressive inflammatory microenvironment during the initial tumor colonization or growth stages. As a result, this microenvironment impedes tumorigenesis and the advancement of the tumor. However, following sustained stimulation by tumor antigens and immune activation responses, the pertinent effector cells within the TME become exhausted or remodeled, rendering them incapable of fulfilling their usual functions or even facilitating the malignant manifestations of tumors. This, in turn, results in the formation of an immunosuppressive microenvironment.

These components are crucial in tumorigenesis, development, and immune escape (5). Tumor-associated macrophages can be polarized in TME as M1 type (classical activation) or M2 type (alternative activation). M1-type macrophages have a pro-inflammatory phenotype and inhibit tumor growth and metastasis by secreting associated inflammatory cytokines such as tumor necrosis factor-α (TNF-α) and interleukin-1β (IL-1β), which can be induced *in vitro* by LPS or IFN-γ (6, 7). M2-type macrophages, on the other hand, are immunosuppressive phenotypes that induce tumor cell invasion and migration by secreting immunosuppressive factors such as IL-10, which can be induced *in vitro* by IL-4 (8). Moreover, research has verified that the extent of immune cell infiltration is associated with the prognosis of individuals with cancer. Hence, evaluating the heterogeneity of TME and remodeling the immune microenvironment of tumors may represent a new avenue for treating cancer (9).

Killer cell lectin-like receptor B1 (KLRB1), encodes CD161, which belongs to the C-type lectin family and was initially defined as a receptor for natural killer (NK) cells. It was later discovered to also exist in subpopulations of CD4+ and CD8+ T cells (10). The attachment of CD161 on T cells provides a co-stimulatory signal for T cell receptor (TCR)-mediated activation (11). KLRB1 transcription is repressed in approximately 68% of people with non-small cell lung carcinoma, suggesting that KLRB1 could serve as a predictive tumor marker (12). In this study, identifying differentially expressed genes (DEGs) for stromal and immune components in cases of BC revealed that KLRB1 could be a potential marker affecting the tumor microenvironment of BC.

# 2 Materials and methods

## 2.1 Collection and preprocessing of BC data

The TCGA (https://portal.gdc.cancer.gov/repository) database was utilized to acquire BC mRNA data and their corresponding clinical data from 1109 patients, which were then presented in a standardized FPKM format.

## 2.2 ESTIMATE

The "Estimate" R package was utilized to calculate Stromal, Immune, and ESTIMATE scores (13).

## 2.3 DEGs identification according to stromal and immune scores

The 1109 BC cases were categorized into high- and low-score subgroups per the median comparison of stromal and immune scores and the DEGs between the two were determined through the Limma R package. The $|\log_2 FC|>1$ and $p<0.05$ served as the

identification criteria. Heat maps were drawn using the R package"pheatmap" (14).

## 2.4 Functional enrichment analysis

The DEGs co-expressed genes of Stromal score and Immune score were obtained using Venn diagrams for GO and KEGG (15) enrichment analysis.

## 2.5 Identification of potential prognostic DEGs with univariate Cox models

The LASSO Cox regression narrowed the range of prognostic DEGs to reduce the risk of overfitting (16). Multivariate Cox regression selected the DEGs most closely associated with survival which were utilized to construct risk models to predict patient survival. By using the standardized expression levels of all genes and their regression coefficient, the risk score for each patient was computed.

$$\text{Risk score} = \text{FOLR2} \times 0.016 + \text{PEX5L} \times 0.077 + \text{KLRB1} \times -0.171 + \text{EPYC} \times 0.041 + \text{BHLHE22} \times 0.064$$

The data were visualized in two dimensions utilizing principal component analysis (PCA) and t-distribution random neighborhood embedding (t-SNE) analysis through the "Rtsne" and "ggplot2" software packages. Furthermore, univariate and multivariate Cox regression analyses were conducted, and independent prognostic factors were identified using the "survival" package.

## 2.6 Building a prediction atlas

By selecting five independent predictive genes, a prediction atlas was constructed, and the robustness of the prediction model was assessed at 1-, 2-, and 3- years (17). The prediction atlas was corrected using calibration charts through a guided method of 1000 resamplings.

## 2.7 Gene set enrichment analysis

The R package "gsva" was utilized for single sample gene set concentration analysis (ssGSEA) to determine the signaling pathways that may be linked to both KLRB1 expression groups (18).

## 2.8 Estimation of TICs

The proportion of 22 TICs in BC samples was calculated utilizing the CIBERSORT algorithm (19), and the results were presented as bar graphs. The proportion of immune cells in tumor tissues with enhanced and reduced expression of KLRB1 was compared utilizing the Wilcoxon rank sum test, and the correlation between the proportion and KLRB1 expression was assessed.

## 2.9 KLRB1 differential expression and survival analysis

To compare the expression of KLRB1 mRNA between BC tissues and normal tissues, the Wilcoxon rank sum test was utilized, and the outcomes were visualized with the "ggpubr" R package. Survival of high and low KLRB1 expression was analyzed using the "survival" R package.

## 2.10 Drug sensitivity analysis and immunotherapy

In order to observe the differences in efficacy of chemotherapy drugs based on KLRB1 expression, we utilized the "pRophetic" package to calculate the half-maximal inhibitory concentration (IC50) of commonly used drugs for treating breast cancer (20). Moreover, we conducted an analysis of the correlation between KLRB1 expression and immunotherapy for breast cancer, utilizing the TCIA database.

## 2.11 Cell culture

Breast normal epithelial cell (MCF-10A) and breast cancer cell lines (MCF7, Hs 578T, HCC1937, MDA-MB-231) were retrieved from ATCC (Manassas, USA). MCF-10A, MCF7, and MDA-MB-231 cells were grown in DMEM high sugar medium (Gibico, China). HCC1937 cells were grown in 1640 medium (Gibico, China). The medium was supplemented with 10% FBS (Pricells, China) and 1% penicillin-streptomycin solution (Solarbio, China). Afterward, the cells were put in an incubator containing 5% CO2 at 37°C.

## 2.12 Cell infection with lentivirus

Once the cells attached and reached a density of 30%, MCF7, and MDA-MB-231 cells were seeded into 6-well plates and subjected to lentivirus infection (viral solution: medium = 1:1). In the infection process cells were incubated in polybrene (2 μg/Ml) in the incubator for a duration of 12 hours. Then the cultivation medium was replaced with a fresh one. After 72 hours of infection, the fusion rate of cells infected by virus infection was up to 80-90% under observation by fluorescence microscope. The cells were then passaged into multi-well plates for further culture. MCF7 and MDA-MB-231 cells were extracted for Western Blot to detect the infection efficiency.

## 2.13 RNA extraction and real-time PCR

TRIzol™ (TermoFisher, USA) was utilized to extract total RNA from the cells, and the front-strand cDNA synthesis kit (TaKaRa, Japan) was employed for reverse transcription, followed by real-time polymerase chain reaction (RT-PCR). The specific primers for KLRB1 and β-Actin were: KLRB1 forward primer 5'-GTTCCACCAAAGAATCCAGCCTG-3' and reverse primer 5'-AAGAGCCGTTTATCCACTTCCAG-3', β-Actin forward primer 5'- CACCATTGGCAATGAGGGTTCTC -3' and the reverse primer 5'- AGGTCTTTGCGTGTCCACGT-3'.

## 2.14 MTT assay

The mixture of MCF7 and MDA-MB-231 cells were grown in 96-well plates at a density of 5000 cells in each well. Furthermore, the edge wells were filled with 200 μl sterile PBS. 96-well plates were kept in an incubator with different time gradients according to the experimental needs. Afterward, 20 μl of MTT solution (5 mg/ml, Beijing Zhongguang Ltd.) was introduced into all wells, and the cells were subsequently incubated for an additional 4 hours. The MTT solution was gently aspirated from each well. Subsequently, 150 μl of DMSO was introduced into all wells, and the plate was shaken slowly for 10 minutes to completely dissolve the methane. The 96-well plate was placed in an enzyme calibrator to read the OD value of each well at 570 nm.

## 2.15 Wound healing assay

MCF7 and MDA-MB-231 cells were inoculated in 6-well plates and allowed to culture overnight. Once the cell density reached 90%, the cell surface was scratched with a 1000 μl pipette tip. Images of the scratches were captured using an inverted microscope (CKX31, Olympus, Japan). Then, cells were treated with serum-free medium after lysis and incubated for 24 hours, and photographs of the scratches were taken at the same location using a microscope.

## 2.16 Transwell assay

Pre-chilled DMEM medium and matrix gel were drawn through the pipette tip to configure the matrix gel working solution (matrix gel: DMEM medium = 1:5). To begin, 50 μl of the matrix gel working solution was aspirated from the bottom of the well. The Transwell and 24-well plates were then incubated in an incubator for 2 hours, and the matrix gel was observed to confirm its solidification. Next, the cell concentration was regulated to 2.5 X $10^6$ cells/mL based on cell counting. Following this, 200 μL of cell suspension was aspirated into the Transwell, and 500 μL of 20% complete medium was introduced to the well plate to ensure that the liquid level in the wells was in contact with the liquid level in the well plate. The migration assay was carried through for 24 hours, and the invasion assay for 36 hours. Transwell plates were removed and washed thrice for 5 minutes with PBS. Following this, cells were fixed using 4% paraformaldehyde for 20 minutes, and then the washing step was repeated again. The cells were left to air dry and then stained with crystal violet for 30 minutes. Following staining, the washing step was repeated again, and the plates were inverted to air dry.

## 2.17 Cell cycle

The cells were cultured at a density of $10\times10^4$ cells per well. After digestion with EDTA-free trypsin, the cells were collected by centrifugation, and the resulting supernatant was discarded. The cell suspension was gently agitated by adding pre-cooled PBS, followed by a second centrifugation step at 1800 rpm for 5 minutes. Afterward, the PBS was removed, and the cells were fixed overnight in 70% ice ethanol. The sample was centrifuged at 1800 rpm for 5 minutes, and the clear supernatant was discarded. Pre-cooled PBS was added to aspirate the suspended cells and the centrifugation step was executed again with subsequent discarding of the PBS. 100 μL of Rnase solution was utilized for resuspension of the cells, after which they were incubated at 37°C for 30 minutes. PI dye was introduced, and the cells were incubated for 30 minutes at 4°C before detection was performed.

## 2.18 Western blot

To prepare whole cell or tissue lysates, pre-cooled NP-40 buffer mixed with protease and phosphatase inhibitors (Roche, USA) was used, and the mixture was centrifuged at 12,000 × rpm at 4°C for 10 minutes. The resulting protein lysates were separated by SDS-PAGE and shifted to polyvinylidene fluoride (PVDF) membranes (Merck, Germany) which were incubated in 1× TBST buffer containing 5% non-skimmed milk powder for 2 hours at room temperature. PVDF membranes were incubated with the indicated primary antibodies and then washed with 1× TBST. Signal detection utilizing a Western ECL substrate kit (Bio-Rad, CA) was then performed using HRP-coupled secondary antibodies. The following antibodies were used for immunoblotting: KLRB1 (67537-1-Ig) and β-actin (81115-1-RR) were purchased from proteintech (China).

## 2.19 EdU Assay

Logarithmic growth phase cells were digested and centrifuged, each group of cells was adjusted to 3×104/mL with basal medium, 100ul was added to the plate with 5 side wells per group, and the cell adhesion was observed overnight. 100ul of EdU working solution per well and incubate in a cell culture incubator for 2 h. PBS was cleaned once, and 4% paraformaldehyde was added to fix at room temperature for 30min. Perforation of PBS solution of 0.3% TritonX and leave at room temperature for 15 min. Prepare the Click reaction solution according to the kit and add it to the well plate and incubate for 30 min at room temperature protected from light. Configure 1X Hoechst solution and incubate for 10 min at room temperature in the dark for nuclear staining. PBS is washed 3 times and post-photographed statistics are performed under a fluorescence microscope.

## 2.20 Statistical analysis

All experiments that required statistical analysis were conducted in triplicates. The experimental data were presented as mean ± standard deviation (mean ± SD). GraphPad Prism 9.0. was employed to execute the statistical analyses. Comparative assessment of two and multiple samples was executed through an unpaired Student's t-test and a one-way analysis of variance (ANOVA), respectively, for variability. A *p*-value< 0.05 was considered a statistically significant value.

# 3 Results

## 3.1 Identifying DEGs and enrichment analysis of BC TME in accordance with stromal and immune scores

Through the analysis of high- and low-scoring samples, it was found that DEGs of immune and stromal components have important roles in BC TME. The analysis identified a total of 832 DEGs through immune scoring, wherein 729 genes were overexpressed, and 103 genes had decreased expression (Figure 1A). In terms of stromal scoring, 773 DEGs were identified in total, comprising 634 upregulated genes and 139 downregulated genes (Figure 1B). A Venn diagram was used to identify DEGs associated with interstitial and prevalence scores, with 193 upregulated and 30 downregulated genes (Figures 1C, D). The identified DEGs could potentially be significant factors influencing the status of the TME. GO enrichment analysis highlighted that these DEGs are primarily involved in physiological processes such as T cell activation, differentiation, and regulation, as well as lymphocyte differentiation and regulation (Figure 1E). KEGG enrichment analysis also revealed that these DEGs help in cell adhesion molecule interactions, cytokine-receptor interactions, hematopoietic cell lineage, and viral proteins with cytokine and cytokine receptor interactions (Figure 1F).

## 3.2 Constructing and validating the stability of the prognostic model

The expression data of the 223 DEGs obtained were grouped according to the training set vs. validation set as 7:3. One-way Cox analysis was utilized to find nine prognosis-related DEGs. Subsequently, Lasso regression analysis was conducted on these nine DEGs to obtain the coefficients of five prognostic genes. (Figures 2A, B). Five prognostic genes (FOLR2, PEX5L, KLRB1, EPYC, and BHLHE22) that were significantly different in the multifactorial Cox regression model were identified (Figure 2C). For each BC patient, a risk score was quantified, and the individuals were classified per the median value into high-risk and low-risk groups. More individuals died in the high-risk group than those in the low-risk group. Conversely, the expression of KLRB1 was found to be elevated in the individuals in the low-risk group in comparison with the people in the high-risk group. (Figures 2D, E). According to Kaplan-Meier analysis, both the training and validation sets had a considerably increased overall survival rate for individuals belonging to the low-risk group in comparison with the individuals in the high-risk group ($P$<0.05) (Figures 2F, G). The respective three-year ROC curve areas for the training and validation sets were 0.759 and 0.741 (Figures 2H, I). The riskscore is an independent prognostic factor predicting breast cancer prognosis (Figures 3A, B). The results of both PCA analysis and t-SNE analysis indicated that individuals belonging to both risk groups were sorted into two distinct directions (Figures 3C–F). Multifactorial analyses were extracted from the training set and validation set data to create diagnostic curves to validate the scores of the five independent prognostic factors, the corresponding probabilities were found, and the survival probability of individuals at 1-, 2-, and 3- years was estimated (Figures 3G, H). The calibration plots demonstrated favorable performance in predicting the probability of survival beyond 3 years (Figures 3I, J).

## 3.3 Low expression of KLRB1 is correlated with poor prognosis in BC

To investigate the overall survival (OS) of the five genes related to prognosis, a Kaplan-Meier analysis was conducted, which revealed that individuals having reduced expression of KLRB1 had a remarkably shorter OS. However, the expression of the remaining four genes did not show any association with OS (Figures 4A–E). Then we found that KLRB1 was strongly associated with DSS and PFI in breast cancer (Figures 4F, G). The expression of KLRB1 was considerably lower in BC tissues (Figures 4H, I). The area under the ROC curve (AUC) for KLRB1 as a predictor of OS in BC was 0.71 (Figure 4J). The expression of KLRB1 depicted a positive association with patient age, gender, tumor size, and tumor stage (Figures 4K–P). We also assessed baseline data on high and low expression of KLRB1 in breast cancer. The KLRB1 is closely related to age, clinical stage, pathological type, estrogen receptor and molecular typing of breast cancer (Table 1). Considering that KLRB1 expression correlates with the molecular typing of breast cancer, we further analyzed and found that KLRB1 was also associated with the prognosis of Luminal A and Luminal B subtypes, but not statistically significant with HER-2 positivity and TNBC survival (Figures S1A–D). Finally, we also analyzed the correlation analysis of KLRB1 with ESR1, PGR, and ERBB2, and found that there was a negative correlation with ESR1, PGR and ERBB2 (Figures S1E–G).

## 3.4 Validation and enrichment analysis of the KLRB1 model in BC

Both univariate and multivariate Cox regression analyses highlighted that KLRB1 served as an independent prognostic factor in BC (Figures 5A, B). Based on KLRB1 and clinicopathological features, a prognostic scale was developed for the prediction of the prognosis of BC (Figure 5C). The calibration curve was approximately diagonal, indicating that the prognostic scoring scale had strong predictive power for survival at 1-, 3-, and 5- years (Figure 5D). The genome showing high expression of KLRB1 exhibited notable enrichment in immune-related activities, such as cell adhesion and signaling pathways including chemokine, T cell receptor, B cell receptor, NK cell regulatory, and JAK/STAT signaling pathway (Figure 5E). The KLRB1 low expression genome was mainly enriched in metabolic and biosynthetic pathways, including

10.3389/fendo.2023.1185799

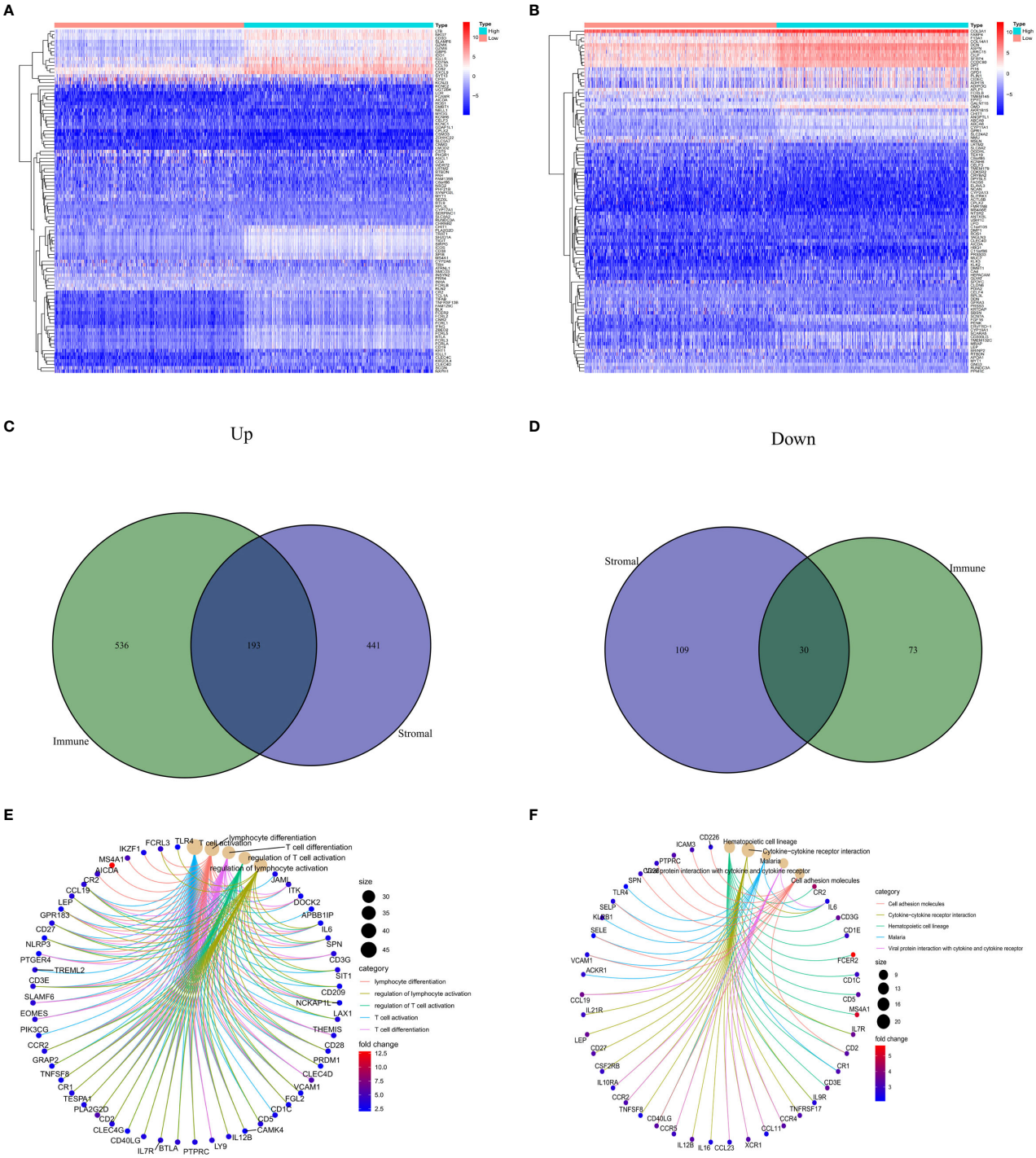



FIGURE 1
Comparison of stromal and immune scores in BC patients. **(A, B)** Heatmaps show the top 50 increases and downwards in immune scores and stromal scores. DEGs are tested using Wilcoxon rank and testing (P<0.05 |log2FC|<1). **(C, D)** Venn chart analysis of DEGs based on immune scores and stromalscores. **(E, F)** Analysis of biological functions and pathways associated with degs in BC, pValue <0.05 was considered significantly. BC: Breast cancer; DEGs: Different express genes.

unsaturated fatty acid biosynthesis, phosphatidyl inositol acyl-anchored biosynthesis, N-glycosylated biosynthesis, fructose and mannitol metabolism, and selenium amino acid metabolism (Figure 5F). In conclusion, these outcomes suggested that KLRB1 might be a potential indicator in TME.

## 3.5 KLRB1 affects the expression of immune cells in BC TME

To investigate the potential role of KLRB1 in the tumor microenvironment, a cell sorting algorithm was employed to

10.3389/fendo.2023.1185799

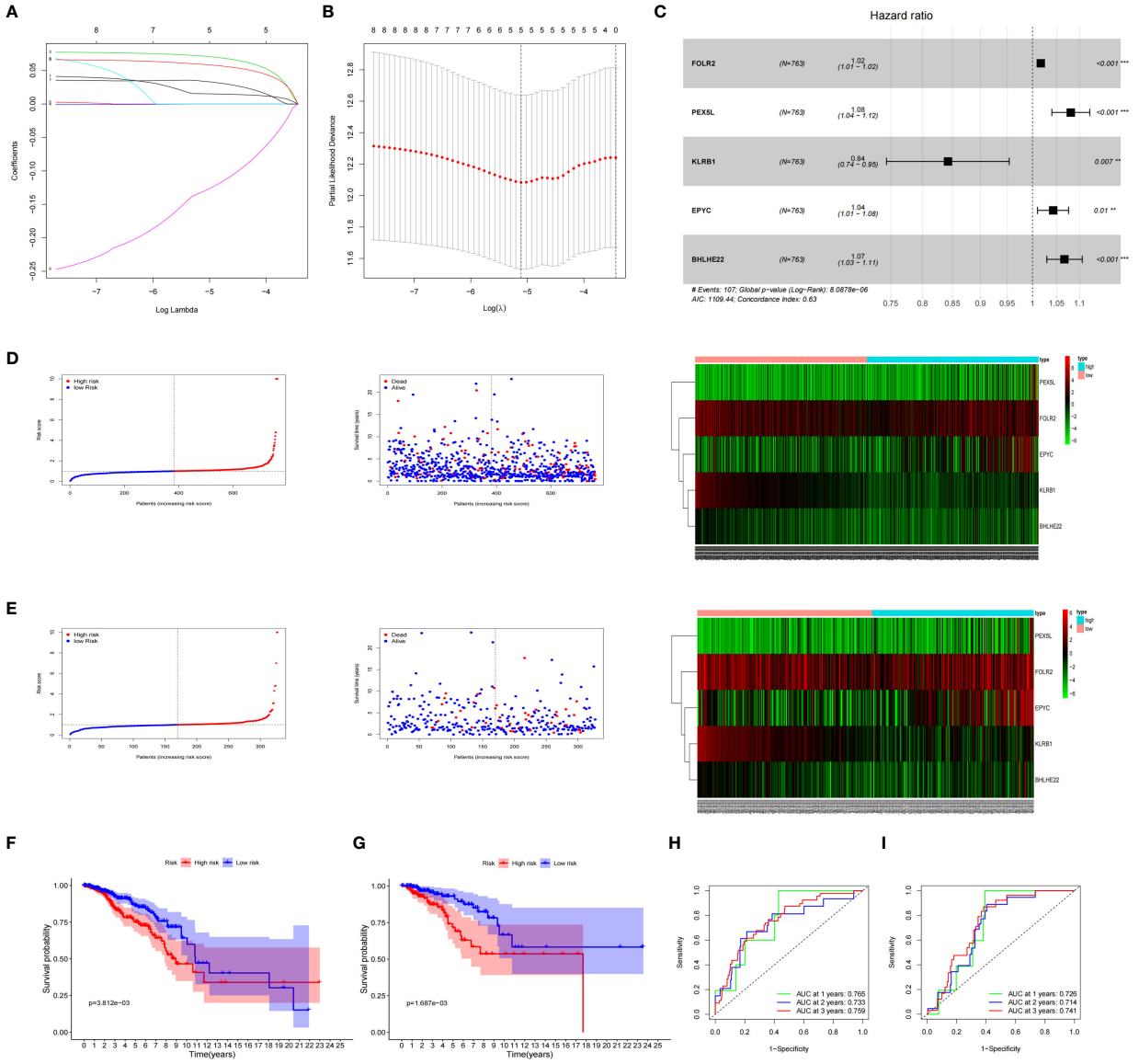



FIGURE 2
Prognosis of 5 genetic characteristic models in training and testing cohort. **(A)** LASSO regression of 5 os-related genes. **(B)** Cross-validated the method of adjusting parameter selection in LASSO regression. **(C)** Five gene forests map. Training cohort **(D)** and testing cohort **(E)** include median of the risk score, the status of OS and the expression spectrum of five immune genes. **(F, G)** Kaplan-Meier analysis survival rates for patients in high-risk and low-risk groups. **(H, I)** AUC time-dependent ROC curve evaluates the prognosis model for OS.

detect the proportion of 22 immune cells present in the BC microenvironment (Figure 6C). In addition, the group with high KLRB1 expression was scored for immunity using the ESTIMATE procedure, having a considerably higher immunity score, stromal score, and ESTIMATE score (Figure 6D). In addition, the Wilcoxon-Mann Whitney test showed that T cells, B cells, CD8 T cells, Th1 cells, DC cells, and cytotoxic cells were relatively high in the high KLRB1 expression group and relatively low in the low KLRB1 expression group (Figure 6A). Expression of KLRB1 was linked positively with the abundance of innate immune cells (Figure 6B), including T cells (r = 0.843), cytotoxic cells (r = 0.801), B cells (r = 0.719), Th1 cells (r = 0.597) and DC (r = 0.695), and CD8 T cells (r = 0627) (all $P< 0.001$).

## 3.6 KLRB1 expression associated with chemotherapy sensitivity and immunotherapy response

Specifically, patients with low expression of KLRB1 were found to be more sensitive to doxorubicin, paclitaxel, docetaxel, and 5-fluorouracil (Figures S2A–D). Currently, immunotherapy for breast cancer mainly focuses on PD-L1 and CTLA4. Interestingly, our study revealed that when both CTLA4 and PD-L1 were positive or when either CTLA4 or PD-L1 was positive, patients with low KLRB1 expression exhibited lower expression levels of both CTLA4 and PD-L1 than those who were negative for both CTLA4 and PD-L1 (Figures S2E–H).

10.3389/fendo.2023.1185799

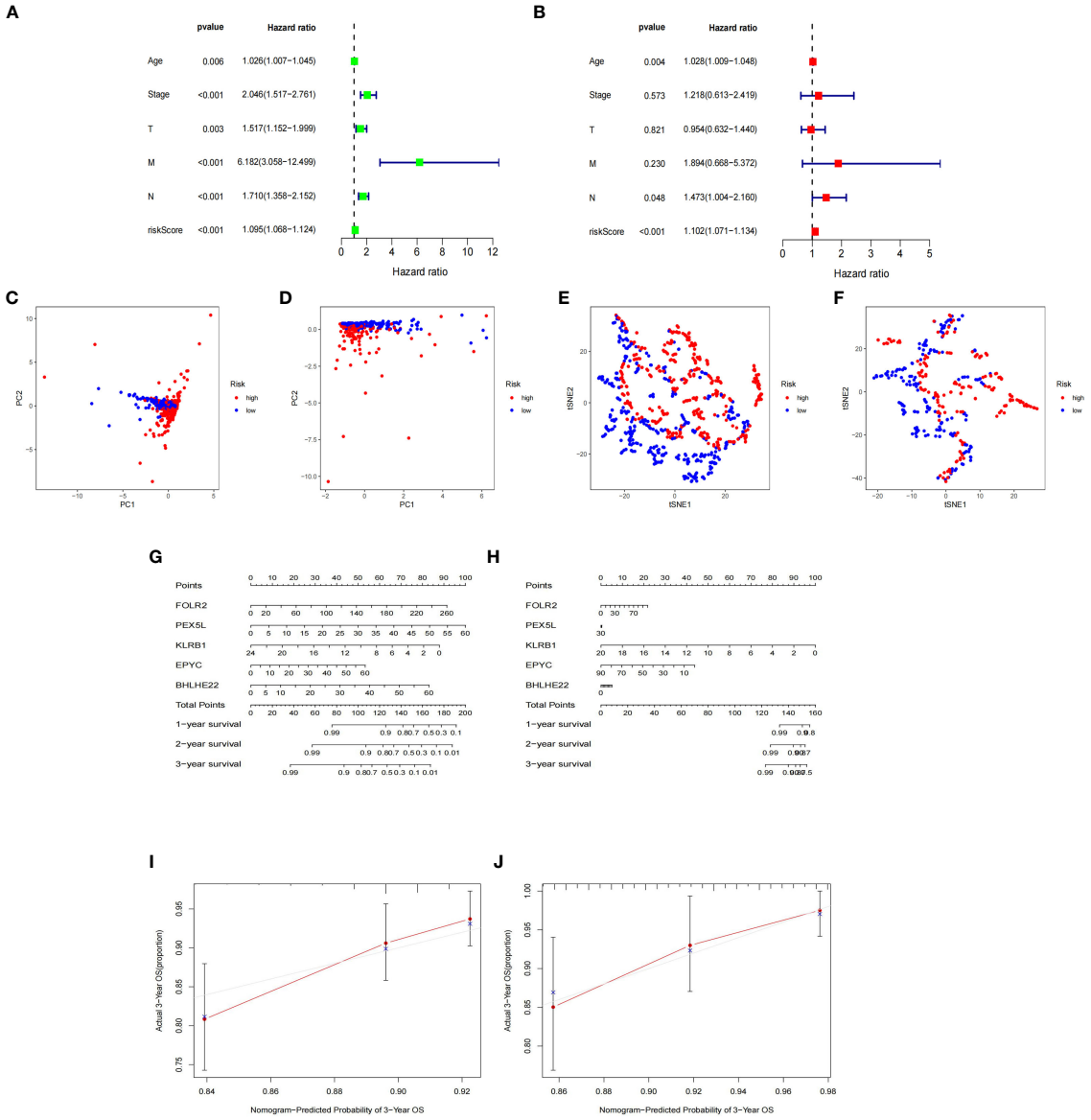



FIGURE 3
Nomogram based on the prognosis characteristics of the five genes is in the TCGA cohort. Univariate Cox **(A)** and multivariate Cox **(B)** regression analysis identified riskscore as a risk factor for breast cancer prognosis. PCA diagram **(C, D)** and t-SNE analysis **(E, F)** in Training cohort and testing cohort. Build a nomogram model of five genes and high and low risk to predict one, two, and three years of survival with Training cohort **(G)** and testing cohort **(H)**. The calibration chart shows that the predicted survival rate is consistent with the actual survival rates for 3 years with Training cohort **(I)** and testing cohort **(J)**.

## 3.7 Validation of KLRB1 *in vitro* assays

The qPCR experiments confirmed that KLRB1 expression was low in MCF7 and MDA-MB-231 cells (Figure 7A). Then we demonstrated that the KLRB1 protein is significantly lower than normal breast epithelial cells in breast cancer MCF7 and MDA-MB-231 cells (Figure 7B). After infection with KLRB1 lentivirus, we verified KLRB1 overexpression in MCF7 and MDA-MB-231 cells, confirming that BC cell lines stably expressing KLRB1 were constructed (Figure 7C). The MTT assay results demonstrated that the overexpressed KLRB1 significantly suppressed the proliferation and viability of both MCF7 and MDA-MB-231 cells (Figures 7D, E). The scratch assay results showed that KLRB1 could considerably inhibit the migrative capacity of MCF7 and MDA-MB-231 cells (Figures 7F, G). While the inhibition of their invasive and migrative abilities by KLRB1 was determined through Transwell assay (Figures 7H–K). Meanwhile, the EdU assay showed that KLRB1 inhibits the DNA replication capacity of cells (Figures 8A–C). Flow cytometry assays showed that KLRB1 can block MCF7 cells (Figures 8D, E) and MDA-MB-231 cells in the G1 phase (Figures 8F, G). The data suggested that KLRB1 might inhibit the proliferation of BC cells by preventing cells from dividing in the G1 cycle.

10.3389/fendo.2023.1185799

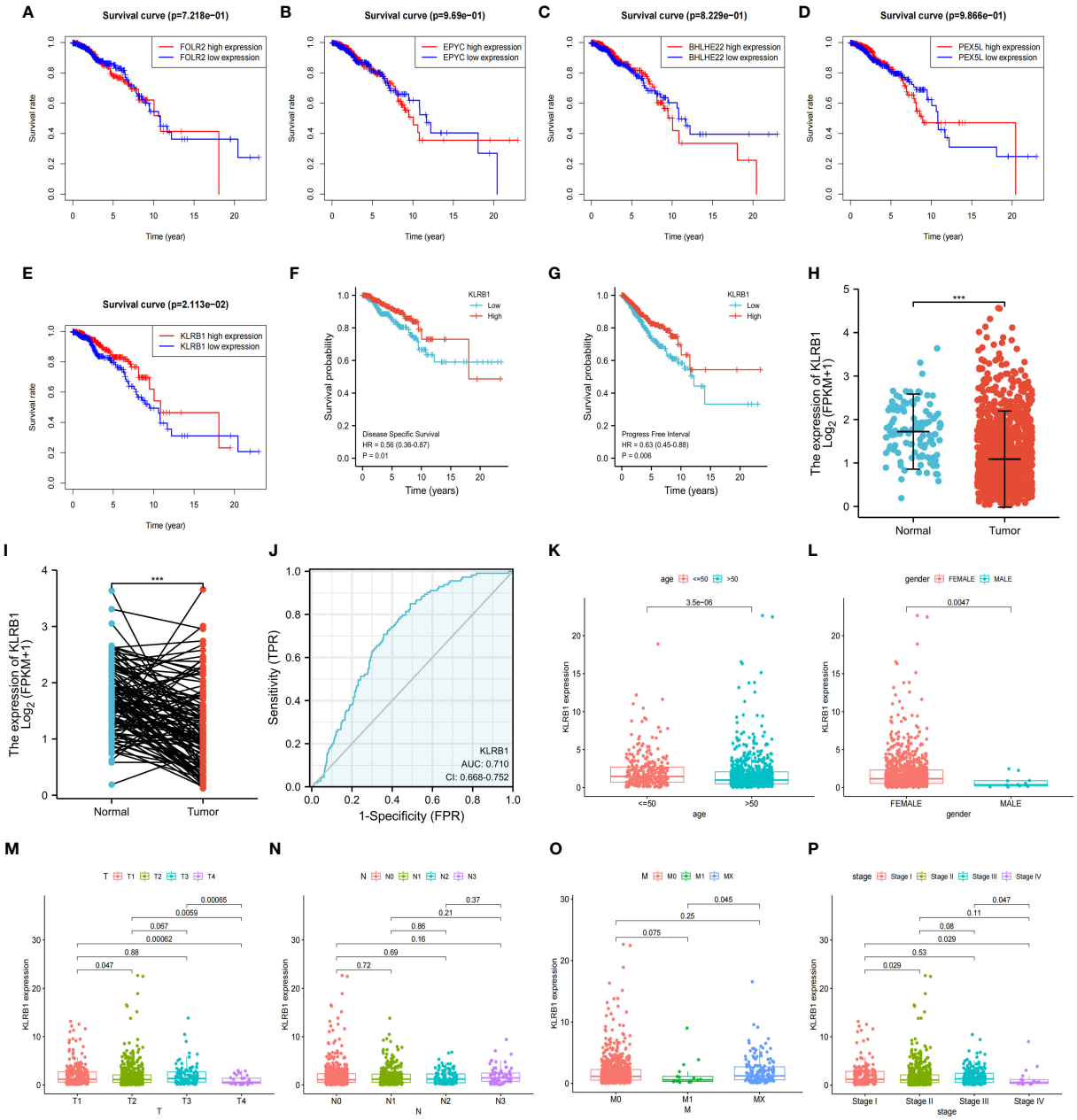



FIGURE 4
The relationship between the expression and total survival of 5 genes in BC. **(A–E)** The expression of five genes is associated with the overall survival prognosis of breast cancer. **(F, G)** The K-M analysis of KLRB1 with DSS and PFI in breast cancer patients. **(H)** KLRB1 mRNA is expressed at a low level in BC tissues. **(I)** Evaluate the level of KLRB1 mRNA in paired BC tissues. **(J)** Construct a ROC curve to predict the impact of KLRB1 on the overall survival of breast cancer patients. **(K–P)** Investigate the expression of KLRB1 in breast cancer patients with respect to age, gender, T size, N stage, M stage, and Stage. DSS, Disease Special Survival. PFI,Progression Free Interval.

# 4 Discussion

The objective of this research was to identify TME genes from the TCGA-BC database that could be used for the diagnosis and staging of TNM and prediction of OS in BC patients. The close association between KLRB1 and immune activity in TME was verified, highlighting its potential as a therapeutic target for BC patients. TMEs formed by tumors are dynamically unstable environments regulated by tumors. Their composition and function change at different stages of tumor development and are closely associated with tumor prognosis and overall disease process (21). The TME provides a favorable growth environment for tumors and promotes their malignant proliferation. However, they also interfere with the function of immune and stromal cells in the microenvironment, leading to tumor evasion of immune surveillance, promotion of metastasis, and drug tolerance (22). These findings highlight the significance of studying the interactions between tumor cells and immune cells and provide

10.3389/fendo.2023.1185799

TABLE 1 Baseline data sheet about KLRB1 express.

| Characteristic | Low expression of KLRB1 | High expression of KLRB1 | p |
|---|---|---|---|
| n | 541 | 542 | |
| Age, n (%) | | | < 0.001 |
| <=60 | 268 (24.7%) | 333 (30.7%) | |
| >60 | 273 (25.2%) | 209 (19.3%) | |
| T stage, n (%) | | | 0.142 |
| T1 | 132 (12.2%) | 145 (13.4%) | |
| T2 | 323 (29.9%) | 306 (28.3%) | |
| T3 | 61 (5.6%) | 78 (7.2%) | |
| T4 | 22 (2%) | 13 (1.2%) | |
| N stage, n (%) | | | 0.182 |
| N0 | 270 (25.4%) | 244 (22.9%) | |
| N1 | 172 (16.2%) | 186 (17.5%) | |
| N2 | 52 (4.9%) | 64 (6%) | |
| N3 | 32 (3%) | 44 (4.1%) | |
| M stage, n (%) | | | 0.044 |
| M0 | 448 (48.6%) | 454 (49.2%) | |
| M1 | 15 (1.6%) | 5 (0.5%) | |
| Pathologic stage, n (%) | | | 0.025 |
| Stage I | 86 (8.1%) | 95 (9%) | |
| Stage II | 323 (30.5%) | 296 (27.9%) | |
| Stage III | 105 (9.9%) | 137 (12.9%) | |
| Stage IV | 13 (1.2%) | 5 (0.5%) | |
| Histological type, n (%) | | | < 0.001 |
| Infiltrating Ductal Carcinoma | 416 (42.6%) | 356 (36.4%) | |
| Infiltrating Lobular Carcinoma | 66 (6.8%) | 139 (14.2%) | |
| ER status, n (%) | | | 0.021 |
| Negative | 102 (9.9%) | 138 (13.3%) | |
| Indeterminate | 1 (0.1%) | 1 (0.1%) | |
| Positive | 410 (39.6%) | 383 (37%) | |
| PR status, n (%) | | | 0.244 |
| Negative | 159 (15.4%) | 183 (17.7%) | |
| Indeterminate | 1 (0.1%) | 3 (0.3%) | |
| Positive | 352 (34%) | 336 (32.5%) | |
| HER2 status, n (%) | | | 0.236 |
| Negative | 256 (35.2%) | 302 (41.5%) | |
| Indeterminate | 6 (0.8%) | 6 (0.8%) | |
| Positive | 84 (11.6%) | 73 (10%) | |
| PAM50, n (%) | | | < 0.001 |
| Normal | 7 (0.6%) | 33 (3%) | |

*(Continued)*

TABLE 1 Continued

| Characteristic | Low expression of KLRB1 | High expression of KLRB1 | p |
|---|---|---|---|
| LumA | 277 (25.6%) | 285 (26.3%) | |
| LumB | 134 (12.4%) | 70 (6.5%) | |
| Her2 | 36 (3.3%) | 46 (4.2%) | |
| Basal | 87 (8%) | 108 (10%) | |

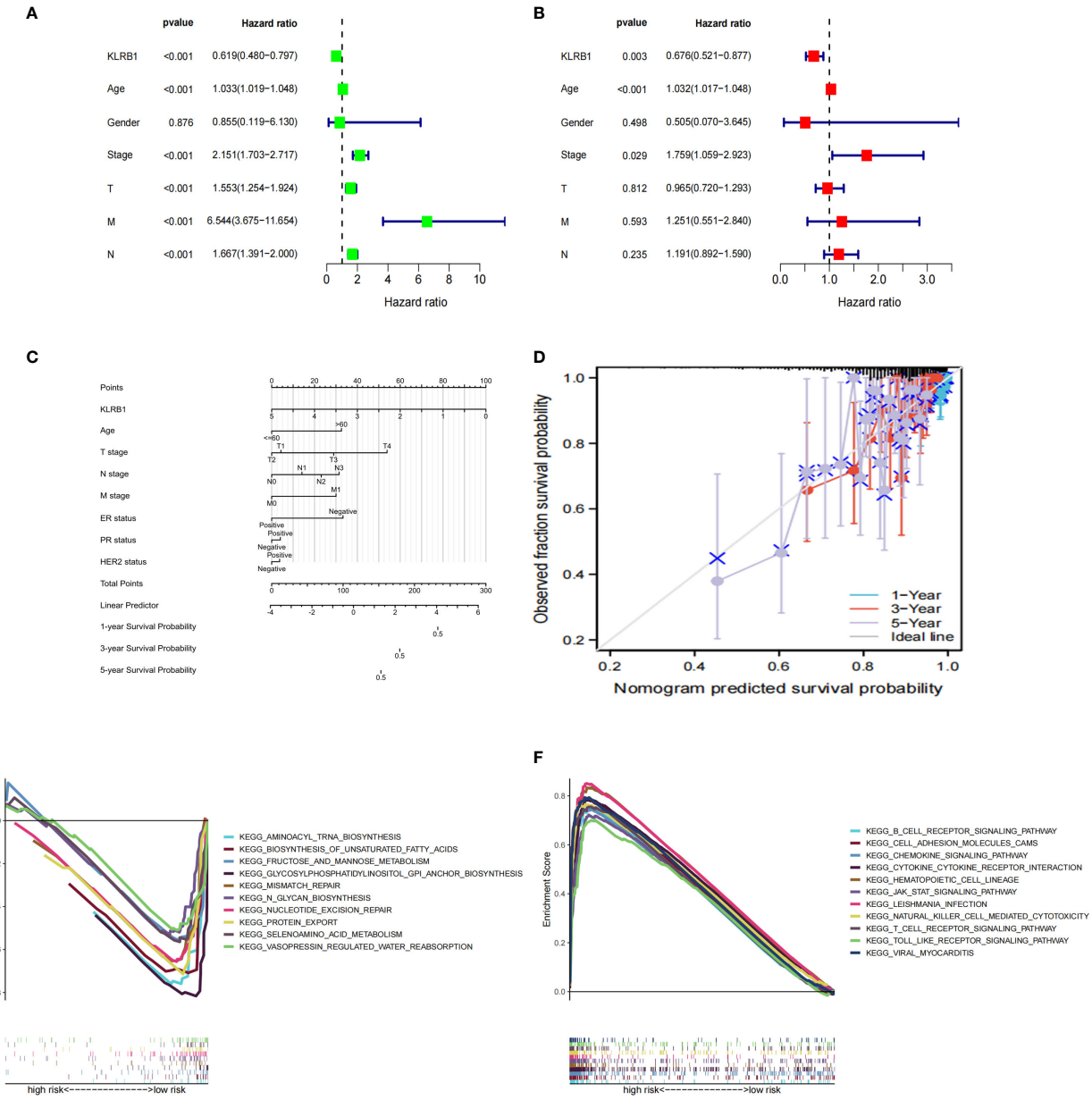



FIGURE 5
The prognostic value of KLRB1 in breast cancer. **(A, B)** Univariate and multivariate Cox regression analysis. **(C)** Nomogram predicting the probability of 1-year, 3-year, and 5-year OS for risk score and clinical characteristics. **(D)** Calibration curves of the nomogram for predicting of 1-, 3-, and 5-year OS in all BC patients. GO and KEGG enrichment analysis of KLRB1 associated genes. OS, overall survival. **(E)** The genome showing high expression of KLRB1 exhibited notable enrichment in immune-related activities. **(F)** The KLRB1 low expression genome was mainly enriched in metabolic and biosynthetic pathways.

10.3389/fendo.2023.1185799

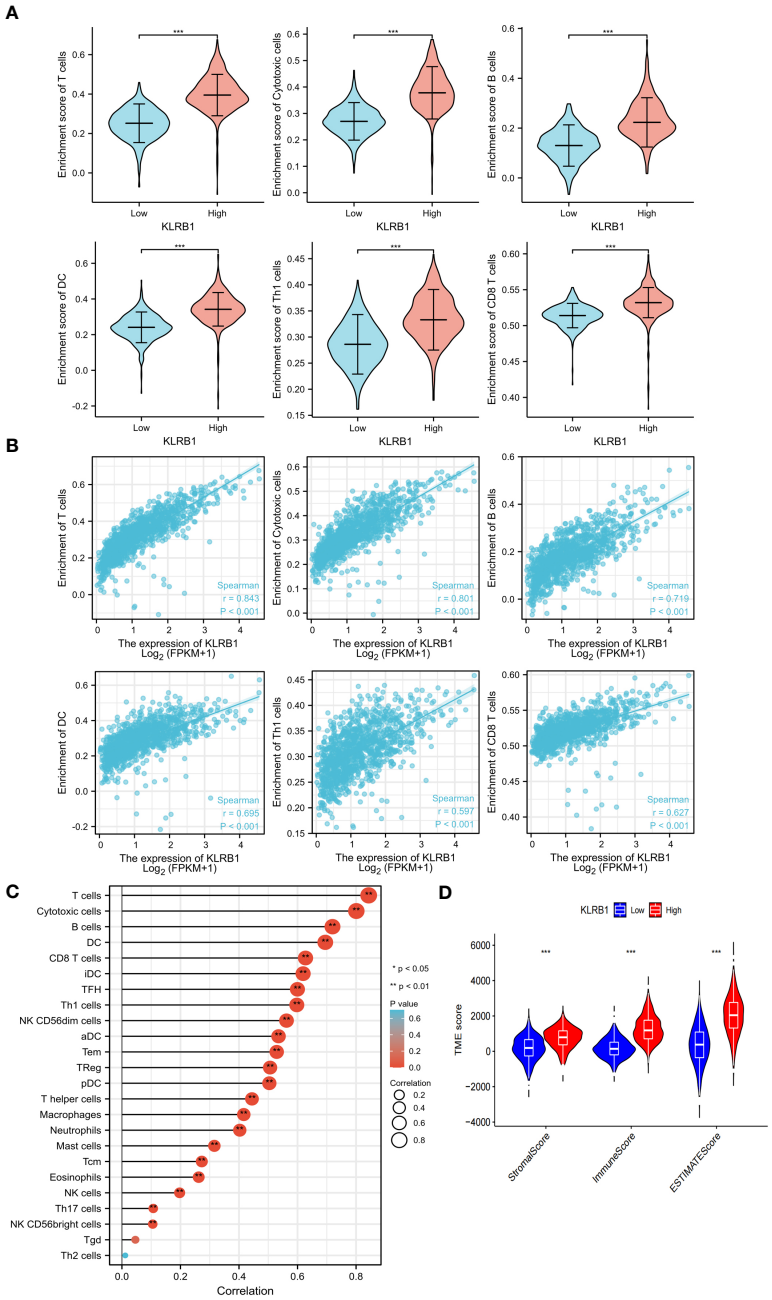



FIGURE 6
KLRB1 affects immune cell expression in the TME in breast cancer. The relationship between KLRB1 expression and the enrichment scores of different immune-infiltrating cells **(A)**. The abundance of immune-infiltrating cells **(B)**. The correlation of KLRB1 with 22 immune cells in the breast cancer tumor microenvironment (TME) are investigated **(C)**. The association between KLRB1 expression and immune scores, stromal scores, and ESTIMATE scores is examined **(D)**. *P<0.05, **P<0.01, ***P<0.001.

new perspectives for the development of more efficient therapeutic options. This analysis identified five immune-related genes, namely FOLR2, PEX5L, KLRB1, EPYC, and BHLHE22, that have prognostic significance.

Belonging to the soluble folate receptor family, FOLR2 is primarily expressed in the placenta, hematopoietic cells, and macrophages, where it is anchored to the extracellular surface by GPI (23). FOLR2 exhibits high expression levels in tumor-associated macrophages (TAMs) of ovarian cancer and can be selectively depleted by G5-MTXNps. In a mouse model of ovarian cancer, TAM depletion inhibited tumor growth. Furthermore, TAM depletion is linked with angiogenesis, which can overcome resistance to VEGF-A therapy when G5-MTXNps is combined with anti-VEGF-A therapy. Therefore, targeting FOLR2 in TAM could be a potential treatment for cancer patients (24). PEX5L is correlated with LINC00924 and serves as an independent predictor of peritoneal metastasis in gastric cancer. This finding indicates that targeting LINC00924/PEX5L could be a potential

10.3389/fendo.2023.1185799

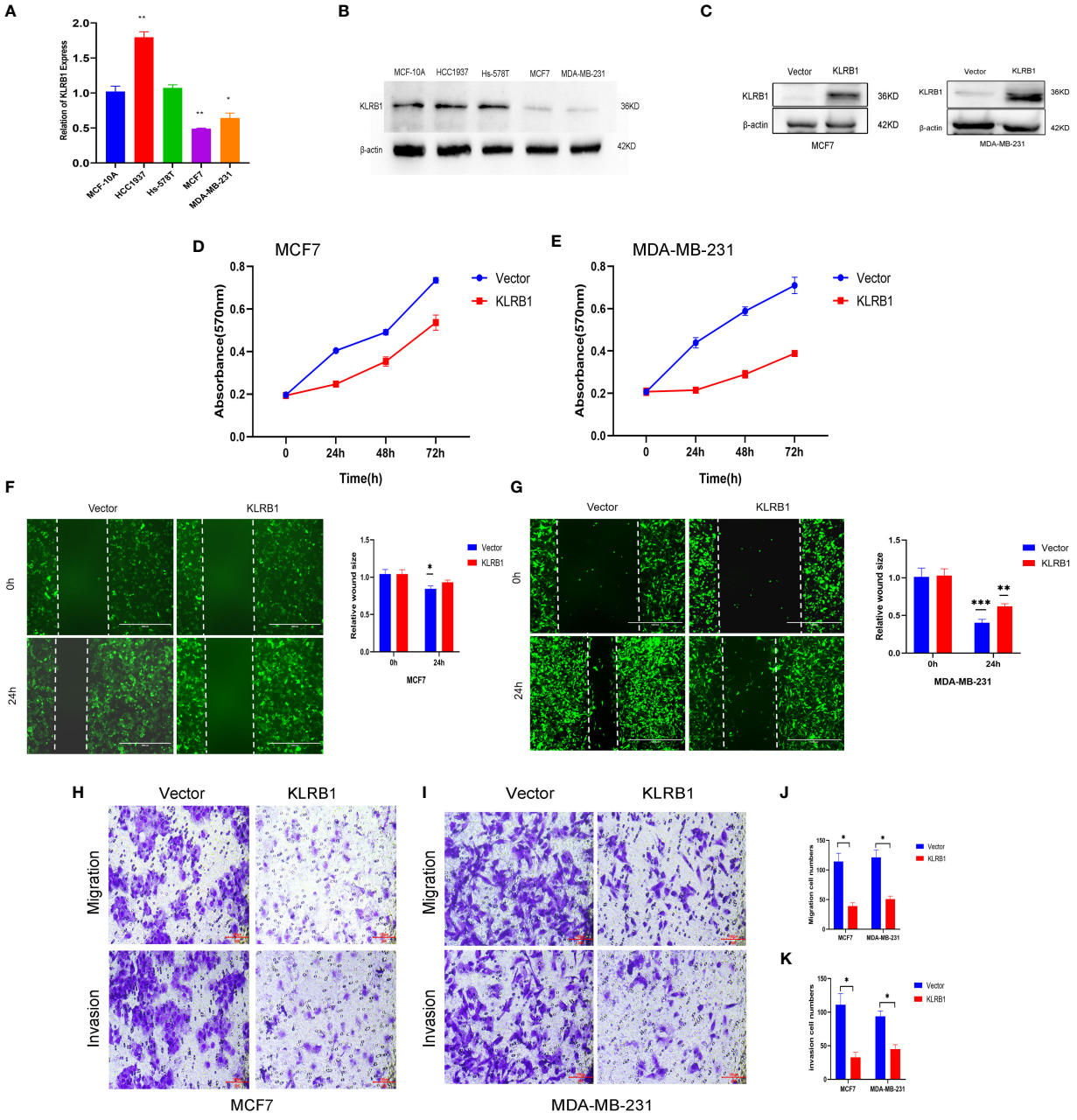



FIGURE 7
KLRB1 is involved in cell proliferation and migration in breast cancer cells. **(A)** qPCR confirmed KLRB1 was low expression in MCF7 and MDA-MB-231 cells. **(B)** KLRB1 were low express in MCF7 and MDA-MB-231 cells. **(C)** Western Blot verifies KLRB1 overexpression. **(D, E)** Overexpression of KLRB1 inhibits the proliferative activity of MCF7 and MDA-MB-231 cells. **(F, G)** KLRB1 inhibits the migration ability of MCF7 and MDA-MB-231 cells. **(H-K)** KLRB1 inhibits the migration and invasion ability of both MCF7 and MDA-MB-231 cells. *P<0.05, **P<0.01, ***P<0.001.

strategy for molecular targeted therapy (25). EPYC exhibits high expression in ovarian cancer and is significantly associated with both OS and disease-free survival (DFS) in patients with ovarian cancer (26). BHLHE2 methylation is increased considerably in healthy endometrium, endometrial hyperplasia, and type I and type II endometrial cancer, and might be a potential molecular target for predicting cervical cancer (27). In a comprehensive analysis of the entire cancer genome, the gene KLRB1 encoding CD161 showed a good clinical prognosis in breast, colorectal, prostate, melanoma, and neuroblastoma carcinomas (28–32).

The development of BC involves genetic and epigenetic changes in multiple genes. The analysis of multi-genomics (transcriptome, microbiome, epigenome, metabolome, and proteome) at different cellular levels provides new perspectives on the formation, diagnosis, and prognosis of BC. With the development of high-throughput technologies, the various mutations, methylation, copy

10.3389/fendo.2023.1185799

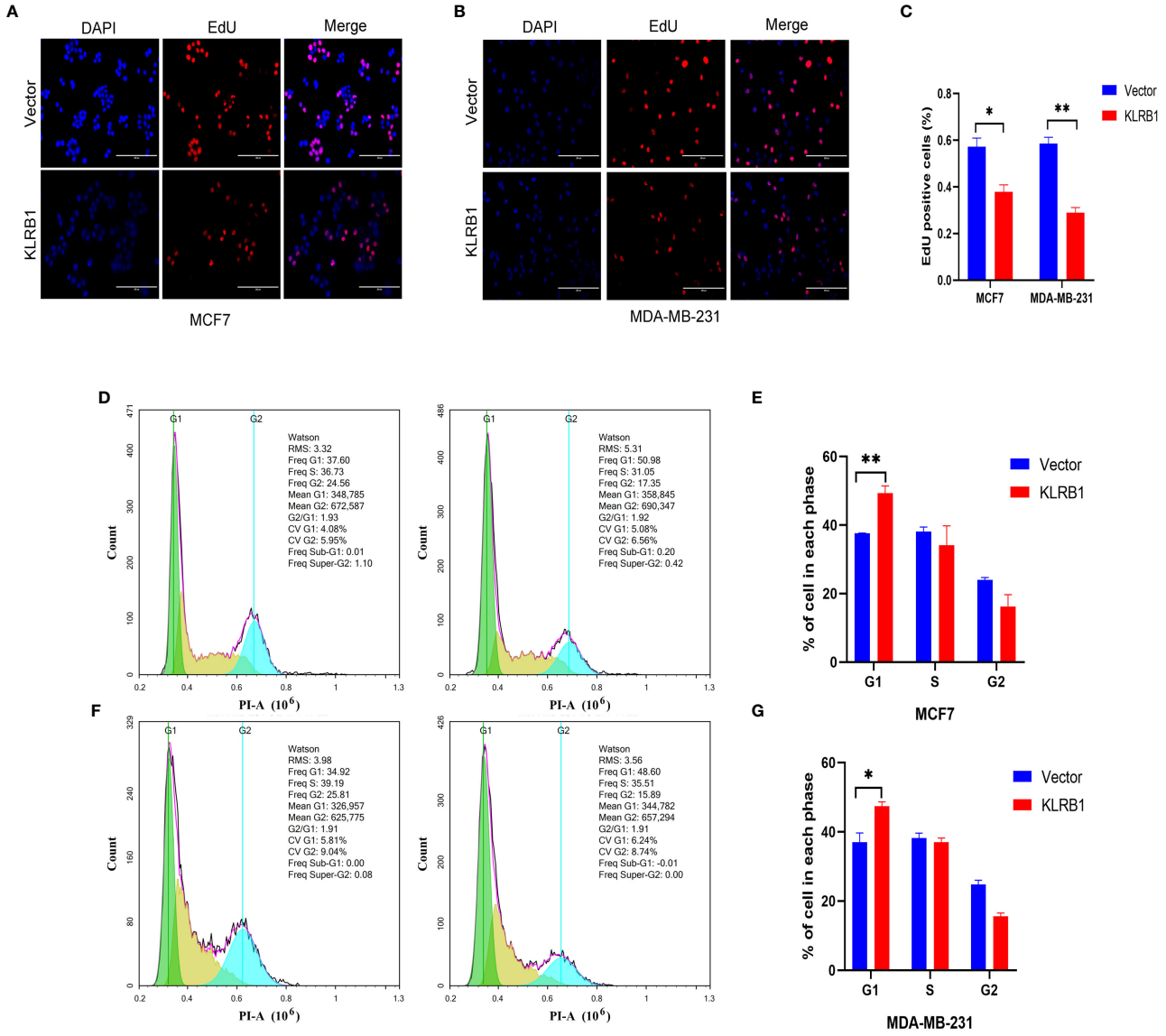



FIGURE 8
KLRB1 affects DNA replication and changes cell cycle in breast cancer cells. **(A–C)** KLRB1 inhibits the DNA replication ability of MCF7 and MDA-MB-231 cells. **(D, E)** KLRB1 blocks MCF7 cells in the G1 phase. **(F, G)** KLRB1 blocks MDA-MB-231 cells in the G1 phase. *P<0.05, **P<0.01.

number, and gene expression patterns have been identified for various cancer types. Copy number variation (CNV) is frequently regarded as a form of genetic variation and is involved in the pathogenesis of BC (33). BRCA1 and BRCA2 are the major BC-related genes, and women carrying BRCA1/2 mutations have a significantly increased risk of BC (34). DNA methylation is a crucial epigenetic modification that regulates gene transcription and maintains genomic stability. Altered methylation, commonly characterized by hypermethylation of proto-oncogenes and methylation of tumor suppressor genes, is critically involved in regulating gene expression in BC (35). The findings of this study suggested that aberrant KLRB1 expression might be due to a combination of copy number variants and methylation variants. Moreover, multi-omics analysis of these genes can help us better understand the molecular mechanisms linked with the development and progression of BC.

The study revealed that only KLRB1 was linked to prognosis in BC patients and had the potential to serve as a biomarker for BC. Individuals with elevated KLRB1 expression had longer survival rates compared to those with low KLRB1 expression. In stage T4 tumors, KLRB1 expression was significantly decreased, suggesting that decreased KLRB1 expression leads to the possibility of poor prognosis in patients, which is consistent with the findings of survival analysis. The above results suggest that KLRB1 expression is closely linked with clinicopathological parameters and poor prognosis. It implied the possible function of KLRB1 as a prognostic marker and therapeutic target for TME in BC. Hence, further analysis was executed to assess the link between KLRB1

expression and TME. The outcomes of GSEA highlighted that the group with over-expression of KLRB1 was mainly concentrated in the cell adhesion (36), and signaling pathways such as B cell receptor, T cell receptor (37), chemokine (38), JAK-STAT, VEGF signaling pathways, and other tumor development-associated pathways.

In this study, CIBERSORT analysis of TIC ratios in BC patients showed a positive correlation between T cells, cytotoxic cells, B cells, Th1 cells, DC cells, and CD8 T cells. The above immune effector cells are mainly responsible for cancer immunosurveillance. CD8 T cells are the primary effectors of the antitumor immune response with potent antitumor activity, and BC patients with high CD8 T cell expression generally have a more favorable prognosis (39), which is in agreement with the outcomes of the previous survival analysis. The analysis of tumor-associated macrophages, a primary component of the tumor stroma, and M2-type TAM concentrations within hypoxic tumor regions were conducted as they exhibit pro-angiogenic activity, and their levels increase with tumor progression (40). These findings align with the GSEA enrichment outcomes and provide further evidence of the validity of the conclusions of this study.

Based on these results, a link between the number of tumor immune infiltrating cells and BC survival can be demonstrated. This means maybe we can improve the prognosis of BC patients by targeting KLRB1 to eliminate the suppression of the immune microenvironment and enhance the immune response.

## 5 Conclusion

In conclusion, a series of bioinformatic analyses of BC samples from the TCGA database was performed, utilizing the ESTIMATE algorithm to determine genes associated with the TME. The association of KLRB1 with the tumor microenvironment of BC highlights that it can be utilized as a promising prognostic marker and therapeutic target. These findings offer a new direction for BC treatment.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.jianguoyun.com/#/sandbox/16cf098/3fb9b64503eda5cf/%2/.

## Author contributions

JC, BJ and KX designed the project. GH, SX, SZ, ZJ, XZ and LC wrote the paper. GH, SX, ZJ, LL and KX perform bioinformatics analysis and MTT Assay, Wound healing Assay, Transwell Assay, Western Blot, qPCR, Cell Cycle Assay. JC, BJ and KX have rigorously revised the final manuscript. All authors also read and agree to release versions of the manuscript.

## Funding

This study was funded by the Guiding Project of Clinical Medical Technology Innovation in Hunan Province (2021SK51706 and 2020SK51705).

## Acknowledgments

I would like to thank all the authors for their contributions to this article. We also acknowledge the TCGA data for providing data.

## Conflict of interest

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.

## Publisher's note



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1185799/full#supplementary-material

## References

1. Siegel R, Miller K, Jemal A. Cancer statistics, 2020. *CA: Cancer J Clin* (2020) 70 (1):7–30. doi: 10.3322/caac.21590

2. Miller K, Nogueira L, Mariotto A, Rowland J, Yabroff K, Alfano C, et al. Cancer treatment and survivorship statistics, 2019. *CA: Cancer J Clin* (2019) 69(5):363–85. doi: 10.3322/caac.21565

3. Lei S, Zheng R, Zhang S, Chen R, Wang S, Sun K, et al. Breast cancer incidence and mortality in women in China: temporal trends and projections to 2030. *Cancer Biol Med* (2021) 18(3):900–9. doi: 10.20892/j.issn.2095-3941.2020.0523

4. Kaymak I, Williams K, Cantor J, Jones R. Immunometabolic interplay in the tumor microenvironment. *Cancer Cell* (2021) 39(1):28–37. doi: 10.1016/j.ccell.2020.09.004

5. Toyama T, Iwase H, Yamashita H, Hara Y, Omoto Y, Sugiura H, et al. Reduced expression of the syk gene is correlated with poor prognosis in human breast cancer. *Cancer Lett* (2003) 189(1):97–102. doi: 10.1016/s0304-3835(02)00463-9

6. Ashida S, Yamawaki-Ogata A, Tokoro M, Mutsuga M, Usui A, Narita Y. Administration of anti-inflammatory M2 macrophages suppresses progression of

angiotensin ii-induced aortic aneurysm in mice. *Sci Rep* (2023) 13(1):1380. doi: 10.1038/s41598-023-27412-x

7. Kratochvill F, Neale G, Haverkamp J, Van de Velde L, Smith A, Kawauchi D, et al. Tnf counterbalances the emergence of M2 tumor macrophages. *Cell Rep* (2015) 12 (11):1902–14. doi: 10.1016/j.celrep.2015.08.033

8. Janes P, Vail M, Ernst M, Scott A. Eph receptors in the immunosuppressive tumor microenvironment. *Cancer Res* (2021) 81(4):801–5. doi: 10.1158/0008-5472.Can-20-3047

9. Akhand S, Liu Z, Purdy S, Abdullah A, Lin H, Cresswell G, et al. Pharmacologic inhibition of fgfr modulates the metastatic immune microenvironment and promotes response to immune checkpoint blockade. *Cancer Immunol Res* (2020) 8(12):1542–53. doi: 10.1158/2326-6066.Cir-20-0235

10. Maggi L, Santarlasci V, Capone M, Peired A, Frosali F, Crome S, et al. Cd161 is a marker of all human il-17-Producing T-cell subsets and is induced by rorc. *Eur J Immunol* (2010) 40(8):2174–81. doi: 10.1002/eji.200940257

11. Gentles A, Newman A, Liu C, Bratman S, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med* (2015) 21(8):938–45. doi: 10.1038/nm.3909

12. Kesselring R, Thiel A, Pries R, Wollenberg B. The number of Cd161 positive Th17 cells are decreased in head and neck cancer patients. *Cell Immunol* (2011) 269 (2):74–7. doi: 10.1016/j.cellimm.2011.03.026

13. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* (2013) 4:2612. doi: 10.1038/ncomms3612

14. Li H, Zhao X, Wang J, Zong M, Yang H. Bioinformatics analysis of gene expression profile data to screen key genes involved in pulmonary sarcoidosis. *Gene* (2017) 596:98–104. doi: 10.1016/j.gene.2016.09.037

15. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. Kegg: integrating viruses and cellular organisms. *Nucleic Acids Res* (2021) 49:D545–D51. doi: 10.1093/nar/gkaa970

16. Liu C, Wang X, Genchev G, Lu H. Multi-omics facilitated variable selection in cox-regression model for cancer prognosis prediction. *Methods (San Diego Calif)* (2017) 124:100–7. doi: 10.1016/j.ymeth.2017.06.010

17. Balachandran V, Gonen M, Smith J, DeMatteo R. Nomograms in oncology: more than meets the eye. *Lancet Oncol* (2015) 16(4):e173–80. doi: 10.1016/s1470-2045 (14)71116-7

18. Powers R, Goodspeed A, Pielke-Lombardo H, Tan A, Costello J. Gsea-incontext: identifying novel and common patterns in expression experiments. *Bioinf (Oxf Engl)* (2018) 34(13):i555–i64. doi: 10.1093/bioinformatics/bty271

19. Newman A, Liu C, Green M, Gentles A, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* (2015) 12 (5):453–7. doi: 10.1038/nmeth.3337

20. Chi H, Xie X, Yan Y, Peng G, Strohmer D, Lai G, et al. Natural killer cell-related prognosis signature characterizes immune landscape and predicts prognosis of hnscc. *Front Immunol* (2022) 13:1018685. doi: 10.3389/fimmu.2022.1018685

21. Risom T, Glass D, Averbukh I, Liu C, Baranski A, Kagel A, et al. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell* (2022) 185(2):299–310.e18. doi: 10.1016/j.cell.2021.12.023

22. Cao D, Naiyila X, Li J, Huang Y, Chen Z, Chen B, et al. Potential strategies to improve the effectiveness of drug therapy by changing factors related to tumor microenvironment. *Front Cell Dev Biol* (2021) 9:705280. doi: 10.3389/fcell.2021.705280

23. Holm J, Hansen S. Characterization of soluble folate receptors (Folate binding proteins) in humans. biological roles and clinical potentials in infection and malignancy. *Biochim Biophys Acta Proteins Proteomics* (2020) 1868(10):140466. doi: 10.1016/j.bbapap.2020.140466

24. Penn C, Yang K, Zong H, Lim J, Cole A, Yang D, et al. Therapeutic impact of nanoparticle therapy targeting tumor-associated macrophages. *Mol Cancer Ther* (2018) 17(1):96–106. doi: 10.1158/1535-7163.Mct-17-0688

25. Fang Y, Huang S, Han L, Wang S, Xiong B. Comprehensive analysis of peritoneal metastasis sequencing data to identify Linc00924 as a prognostic biomarker in gastric cancer. *Cancer Manage Res* (2021) 13:5599–611. doi: 10.2147/cmar.S318704

26. Lisowska K, Olbryt M, Student S, Kujawa K, Cortez A, Simek K, et al. Unsupervised analysis reveals two molecular subgroups of serous ovarian cancer with distinct gene expression profiles and survival. *J Cancer Res Clin Oncol* (2016) 142(6):1239–52. doi: 10.1007/s00432-016-2147-y

27. Liew P, Huang R, Wu T, Liao C, Chen C, Su P, et al. Combined genetic mutations and DNA-methylated genes as biomarkers for endometrial cancer detection from cervical scrapings. *Clin Epigenet* (2019) 11(1):170. doi: 10.1186/s13148-019-0765-3

28. Furuya H, Chan O, Pagano I, Zhu C, Kim N, Peres R, et al. Effectiveness of two different dose administration regimens of an il-15 superagonist complex (Alt-803) in an orthotopic bladder cancer mouse model. *J Trans Med* (2019) 17(1):29. doi: 10.1186/s12967-019-1778-6

29. Fergusson J, Hühn M, Swadling L, Walker L, Kurioka A, Llibre A, et al. Cd161 (Int)Cd8+ T cells: a novel population of highly functional, memory Cd8+ T cells enriched within the gut. *Mucosal Immunol* (2016) 9(2):401–13. doi: 10.1038/mi.2015.69

30. Konjević G, Mirjacić Martinović K, Vuletić A, Jović V, Jurisić V, Babović N, et al. Low expression of Cd161 and Nkg2d activating nk receptor is associated with impaired nk cell cytotoxicity in metastatic melanoma patients. *Clin Exp metastasis* (2007) 24 (1):1–11. doi: 10.1007/s10585-006-9043-9

31. Di W, Fan W, Wu F, Shi Z, Wang Z, Yu M, et al. Clinical characterization and immunosuppressive regulation of Cd161 (Klrb1) in glioma through 916 samples. *Cancer Sci* (2022) 113(2):756–69. doi: 10.1111/cas.15236

32. Ognibene M, De Marco P, Parodi S, Meli M, Di Cataldo A, Zara F, et al. Genomic analysis made it possible to identify gene-driver alterations covering the time window between diagnosis of neuroblastoma 4s and the progression to stage 4. *Int J Mol Sci* (2022) 23(12):6513. doi: 10.3390/ijms23126513

33. Aganezov S, Goodwin S, Sherman R, Sedlazeck F, Arun G, Bhatia S, et al. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res* (2020) 30(9):1258–73. doi: 10.1101/gr.260497.119

34. Hodgson D, Lai Z, Dearden S, Barrett J, Harrington E, Timms K, et al. Analysis of mutation status and homologous recombination deficiency in tumors of patients with germline Brca1 or Brca2 mutations and metastatic breast cancer: Olympiad. *Ann Oncol* (2021) 32(12):1582–9. doi: 10.1016/j.annonc.2021.08.2154

35. Batra R, Lifshitz A, Vidakovic A, Chin S, Sati-Batra A, Sammut S, et al. DNA Methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. *Nat Commun* (2021) 12(1):5406. doi: 10.1038/s41467-021-25661-w

36. Mori M, Hashimoto M, Matsuo T, Fujii T, Furu M, Ito H, et al. Cell-Contact-Dependent activation of Cd4 T cells by adhesion molecules on synovial fibroblasts. *Modern Rheumatol* (2017) 27(3):448–56. doi: 10.1080/14397595.2016.1220353

37. Nicol B, Salou M, Vogel I, Garcia A, Dugast E, Morille J, et al. An intermediate level of Cd161 expression defines a novel activated, inflammatory, and pathogenic subset of Cd8 T cells involved in multiple sclerosis. *J Autoimmun* (2018) 88:61–74. doi: 10.1016/j.jaut.2017.10.005

38. Fenoglio D, Poggi A, Catellani S, Battaglia F, Ferrera A, Setti M, et al. Vdelta1 T lymphocytes producing ifn-gamma and il-17 are expanded in hiv-1-Infected patients and respond to candida albicans. *Blood* (2009) 113(26):6611–8. doi: 10.1182/blood-2009-01-198028

39. Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner B, Weber K, et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol* (2018) 19(1):40–50. doi: 10.1016/s1470-2045(17)30904-x

40. Dallavalasa S, Beeraka N, Basavaraju C, Tulimilli S, Sadhu S, Rajesh K, et al. The role of tumor associated macrophages (Tams) in cancer progression, chemoresistance, angiogenesis and metastasis - current status. *Curr medicinal Chem* (2021) 28(39):8203–36. doi: 10.2174/0929867328666210720143721

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 19 June 2023
DOI 10.3389/fendo.2023.1196094



OPEN ACCESS

EDITED BY
Qiuming Yao,
Fudan University, China

REVIEWED BY
Quan Cheng,
Central South University, China
Liu Yunfei,
Ludwig Maximilian University of Munich, Germany

*CORRESPONDENCE
Qian Li
liqian841015@163.com

RECEIVED 29 March 2023
ACCEPTED 24 May 2023
PUBLISHED 19 June 2023

CITATION
Li Q, Xiao X, Feng J, Yan R and Xi J (2023) Machine learning-assisted analysis of epithelial mesenchymal transition pathway for prognostic stratification and immune infiltration assessment in ovarian cancer. *Front. Endocrinol.* 14:1196094. doi: 10.3389/fendo.2023.1196094



# Machine learning-assisted analysis of epithelial mesenchymal transition pathway for prognostic stratification and immune infiltration assessment in ovarian cancer

Qian Li*, Xiyun Xiao, Jing Feng, Ruixue Yan and Jie Xi

Department of Gynecology, Cangzhou Central Hospital, Cangzhou, Hebei, China

**Background:** Ovarian cancer is the most lethal gynaecological malignancy, and serous ovarian cancer (SOC) is one of the more important pathological subtypes. Previous studies have reported a significant association of epithelial tomesenchymal transition (EMT) with invasive metastasis and immune modulation of SOC, however, there is a lack of prognostic and immune infiltration biomarkers reported for SOC based on EMT.

**Methods:** Gene expression data for ovarian cancer and corresponding patient clinical data were collected from the TCGA database and the GEO database, and cell type annotation and spatial expression analysis were performed on single cell sequencing data from the GEO database. To understand the cell type distribution of EMT-related genes in SOC single-cell data and the enrichment relationships of biological pathways and tumour functions. In addition, GO functional annotation analysis and KEGG pathway enrichment analysis were performed on mRNAs predominantly expressed with EMT to predict the biological function of EMT in ovarian cancer. The major differential genes of EMT were screened to construct a prognostic risk prediction model for SOC patients. Data from 173 SOC patient samples obtained from the GSE53963 database were used to validate the prognostic risk prediction model for ovarian cancer. Here we also analysed the direct association between SOC immune infiltration and immune cell modulation and EMT risk score. and calculate drug sensitivity scores in the GDSC database.In addition, we assessed the specific relationship between GAS1 gene and SOC cell lines.

**Results:** Single cell transcriptome analysis in the GEO database annotated the major cell types of SOC samples, including: T cell, Myeloid, Epithelial cell, Fibroblast, Endothelial cell, and Bcell. cellchat revealed several cell type interactions that were shown to be associated with EMT-mediated SOC invasion and metastasis. A prognostic stratification model for SOC was constructed based on EMT-related differential genes, and the Kapan-Meier test showed that this biomarker had significant prognostic stratification value for several independent SOC databases. The EMT risk score has good stratification and identification properties for drug sensitivity in the GDSC database.

**Conclusions:** This study constructed a prognostic stratification biomarker based on EMT-related risk genes for immune infiltration mechanisms and drug sensitivity analysis studies in SOC. This lays the foundation for in-depth clinical studies on the role of EMT in immune regulation and related pathway alterations in SOC. It is also hoped to provide effective potential solutions for early diagnosis and clinical treatment of ovarian cancer.

KEYWORDS

serous ovarian cancer, epithelial mesenchymal transition, transcriptomics, single-cell sequencing, machine learning

# 1 Introduction

Ovarian cancer is a prevalent gynaecological malignancy that has a deleterious effect on women's health, accounting for approximately 3% of gynaecological malignancies worldwide. The incidence of ovarian cancer is positively correlated with the Human Development Index, and developing countries, mainly China, are at risk of increasing ovarian cancer incidence, according to GLOBOCAN 2020 (1). Global cancer statistics show that approximately 310,000 new cases of ovarian cancer occur each year, and approximately 150,000 people die from ovarian cancer (2, 3). The pathological types of ovarian cancer are also complex, with serous ovarian cancer (SOC) being one of the more important pathological subtypes. The mortality rate of SOC is also decreasing year by year with the improvement of medical treatment and the rapid progress in the development of therapeutic drugs (4–6). Despite the current improvements in the diagnosis and treatment of SOC, the 5-year survival rate of ovarian cancer patients has not improved significantly (7, 8). This shows that accelerating the research on SOC can not only improve the survival rate of ovarian cancer patients, but also reduce the disease burden caused by female malignancies and advance the development of healthcare.

Tumour invasion and metastasis is a complex process involving multiple genes and steps, with weakened adhesion and enhanced movement between tumour cells as the basis for invasion and metastasis. The potential of tumour invasion and metastasis depends on the interaction of internal environmental factors, of which epithelial-mesenchymal transition (EMT) is one of the main factors (9). EMT has also been shown to be a major cause of invasion and metastasis in epithelial ovarian cancer (10, 11). Epithelial mesenchymal transition (EMT) is a process by which epithelial cells lose their polarity and adhesion and acquire a mesenchymal phenotype. This process results in reduced adhesion and increased motility between tumour cells, involves multiple signalling pathways (12), is a necessary initial step for tumour cell invasion and metastasis, and is associated with malignant transformation and the development of metastasis, recurrence and drug resistance in a variety of malignancies, including ovarian cancer (13). In ovarian cancer, there is a small population of 'tumour initiating cells', characterised as mesenchymal and stem cells, which play a role in driving tumour initiation (14). The presence of TGFβ, a stimulating factor that induces EMT in ovarian follicular fluid, has been shown to inhibit PAX2, which maintains the differentiation of tubal epithelial cells, leading to the progression of intraepithelial tubal carcinoma to high-grade plasma SOC (15). It has also been shown that BRCA1 mutations induce EMT and tumourigenesis, often developing into highly invasive, poorly differentiated plasma ovarian cancer (16). Therefore, EMT plays an important role in both the pathogenesis of SOC and its development and invasion throughout the metastatic process. However, there is a lack of specific mechanisms of EMT-mediated SOC invasion and metastasis as well as single-cell descriptions of EMT-related genes in SOC. In addition, the construction of EMT-based prognostic stratification biomarkers for SOC is also important for the assessment of the efficacy of immunotherapy and the selection of therapies. Unfortunately, there is a lack of reported studies related to the above.

To address the existing and potential mechanistic evidence of EMT in the development of SOC, this paper will analyse the relationship between EMT and infiltrative metastasis of ovarian cancer in the hope of providing new ideas for the treatment of SOC. In this study, cell type annotation and spatial expression analysis of single cell sequencing data from the GEO database were performed using ovarian cancer gene expression data collected from the TCGA database and the GEO database and corresponding patient clinical data. To understand the cell type distribution of EMT-related genes in SOC single-cell data and the enrichment relationship between biological pathways and tumour function. Major differential genes for EMT were screened and a prognostic risk prediction model for SOC patients was constructed. We then used data from 173 SOC patient samples obtained from the GSE53963 database for the validation of the prognostic risk prediction model for ovarian cancer. In addition, GO functional annotation analysis and KEGG pathway enrichment analysis were performed on mRNAs predominantly expressed with EMT to predict the biological function of EMT in ovarian cancer. We also analysed the direct association of SOC immune infiltration and immune cell regulation with EMT risk scores, laying the foundation for in-depth clinical studies of EMT in SOC immune regulation and related pathway alterations. This study further explores the molecular mechanism of ovarian cancer metastasis and to provide potential markers and therapeutic targets for the diagnosis of early cancer metastasis in ovarian cancer.

# 2 Materials and methods

## 2.1 Data acquisition

Biological data refers to the measurement and collection of different kinds of genomic, transcriptomic, epigenomic, proteomic and metabolomic data of organisms by modern sequencing techniques as well as histological techniques. The main biological databases used in this paper include, gene expression databases (TCGA database, GEO database), single cell sequencing databases (GEO database) and gene enrichment analysis databases (GO database, KEGG database). The data for this study were obtained from The Cancer Genome Atlas (TCGA) database and individual clinical information data. The ovarian cancer mRNA gene data used in this paper were obtained from ovarian cancer gene expression data obtained from the TCGA database. Samples of the original RNA-Seq format ovarian cancer sequencing data were downloaded from this database with complete gene sequencing results and complete clinical information on the patients. In addition, bulk RNA-seq of 174 ovarian cancer patients from GSE53963, and single-cell scRNA-seq data of 5 high-grade SOC patients from GSE154600 were included for further analysis.

## 2.2 Single-cell sequencing annotation of SOC and spatial distribution profile of EMT-related markers

Single-cell transcriptome sequencing refers to the high-throughput sequencing of mRNAs after reverse transcriptional amplification at the individual cell level. By sequencing at the single cell level, single cell sequencing solves the problem of not being able to obtain information about the heterogeneity of different cells with tissue samples or having too small a sample size for routine sequencing, and provides a new direction for scientists to study the behaviour and mechanisms of individual cells. Due to cellular heterogeneity, the genetic information of cells of the same phenotype may differ significantly, and much of the low abundance information will be lost in the overall characterization. In order to compensate for the limitations of traditional high-throughput sequencing, single-cell sequencing has been developed. The basic steps of single cell sequencing: isolation of selected single cells → amplification → high throughput sequencing → data analysis using bioinformatics techniques. In this study, we accessed the NCBI GEO website: https://www.ncbi.nlm.nih.gov/geo/ to download the data. The SOC-related single cell data were searched and downloaded from the search box. First, single-cell analysis was performed on five high-grade SOC samples from the scRNA-seq data of GSE154600, downscaled to six cell populations. t-SNE's main use is to visualise and explore high-dimensional data. It was developed and published by Laurens van der Maatens and Geoffrey Hinton in JMLR Volume 9 (2008). The main goal of t-SNE is to transform multidimensional datasets into low-dimensional datasets. Compared to other dimensionality reduction algorithms, t-SNE works best for data visualisation. If we apply t-SNE to n-dimensional data, it will intelligently map n-dimensional data to 3d or even 2d data and the relative similarity of the original data is very good. Like PCA, t-SNE is not a linear dimensionality reduction technique, it follows non-linearity, which is the main reason why it can capture the complex flow structure of high-dimensional data. The main regions of the spatial distribution of single cells from five LIHC patients are depicted, thus providing insight into the expression of the main SOC tumour cell types and their relative proportions. In addition, we present a map of the spatial distribution of single cell data from five SOC patients. to understand the individual differences in cell annotation across patients. To understand the cell types and specific functional modalities of the major EMT roles in SOC, the spatial distribution of expression in the annotated major cell types was annotated using a variety of specific markers of EMT that have been widely formalised. The main EMT markers included were: EPCAM, LUM, RAMP2, COL3A1, CD79A, CD3D, CD8A, LYZ, CD68. We also compared and analysed the percentage and distribution levels of the different cell types in single cell samples from each SOC patient.

## 2.3 Signal communication and biological function enrichment analysis of SOC major annotated cell types

After initial pre-processing and downscaling analysis by scRNA-seq of GSE154600, we sought to understand the interactions between different cell types and signalling guidance. This is to better understand the microscopic role of EMT in the development of SOC. The intercellular communication network (ICN) is a weighted directed graph consisting of significant ligand-receptor pairs between interacting cell groups, showing the number of detected ligand-receptor interactions between different cell groups. cellChat uses the out-degree, in degree to infer the strength of different cell groups as senders, and receivers of signals during cellular communication. To further analyse intercellular communication in a more biologically meaningful way, ligand-receptor pairs are grouped into functionally relevant signalling pathways and CellChat is able to quantify the similarity between all significant signalling pathways, grouping them according to the similarity of their cellular communication networks. Pattern recognition was used to predict coordinated responses between cells. Therefore, we performed interaction resolution of single cell data from SOC by cellchat analysis. In addition, we calculated GSVA enrichment pathway scores for six cell populations using 50 Hallmark datasets.GSVA gene set variation analysis, an analysis of microarray and RNA-seq data gene sets under parameter-free and unsupervised conditions.GSVA enables the enrichment of a gene - sample data matrix (GSVA converts a gene-sample data matrix (microarray data, FPKM, RPKM, etc.) into a gene-set-sample matrix. Based on this matrix, the enrichment of gene sets (e.g. KEGG pathway) in individual samples can be further analysed. As GSVA results in a gene-set-sample enrichment matrix, there is more freedom for downstream analysis than with other gene-set enrichment methods.

The GSEA database of EMT (Epithelial Mesenchymal Transition) pathways was scored for gene set enrichment using

the AddModuleScore function of the Seruat package and divided into two groups: EMT-high level and EMT-low level. GO analysis, KEGG analysis and GSEA-GO were performed using the R package "clusterProfiler" (17)(version 4.0.5), with a false discovery rate (FDR) < 0.05 to determine significant enrichment. Gene Onto Logy (GO) is a public database built by the Gene Onto Logy Consortium, which contains annotated information on the properties of species genes and related products, with the aim of standardising annotation information on the function of biological gene products. The Kyoto Encyclopedia of Genes and Genomes (KEGG), developed by the Kanehisa Laboratory in collaboration with the Institute of Chemistry of Kyoto University, integrates information on genomes, biological pathways, chemicals and system functions from the Human Genome Project. KEGG is a systematic knowledge base that uses specific algorithms to access and collate the results of existing experiments. The database can be divided into four main modules: Systems Information, Genomic Information, Chemical Information and Health Information. The KEGG Pathway database in the Systems Information module is a standard biological pathway database that is highly recognised in the field of bioinformatics research and can be used to explain biological processes within cells in a graphical form.

## 2.4 Construction of prognostic biomarkers for SOC based on EMT risk genes and functional validation of the model

The original dataset of RNAseq data in HTSeq-Counts format was converted to TPM format followed by Log2 transformation. First, we pre-processed the transformed data using the R language survival package and did one-to-one one-way Cox analysis of the EMT-associated gene datasets with significant expression differences. Secondly, as the traditional Cox regression model is only applicable when the number of covariates is smaller than the number of samples, when the number of covariates is larger than the number of samples, the model parameters will be difficult to be calculated. Also, there may be a high degree of similarity in gene expression. To reduce the appearance of overfitting of the results, we took Least absolute shrinkage and selection operator (LASSO) downscaling of the EMT-related genes that regressed significantly after the one-way Cox regression to screen out the EMT-related genes that were more correlated with survival outcomes. The major genes obtained by LASSO regression were combined with their linear coefficients to form an EMT-related risk score, and the median risk score differentiated the Lowrisk and Highrisk groups to establish a Riskscore score. The Kaplan-Meier survival curve analysis was used to further analyse the survival of the EMT risk scores obtained from the LASSO regressions in relation to their corresponding clinical survival times of patients. A forest plot was used to calculate the proportion of risk for each EMT-related gene included in the model. Box plots were also used to depict the mRNA expression of EMT-associated genes in patients in the low-risk versus high-risk groups of the EMT score. In this paper, EMT-related genes obtained from screening in the Cox risk assessment model and m RNAs that were differentially expressed in ovarian cancer and normal ovarian tissues were selected and put into a co-expression network for analysis. The Pearson correlation coefficients between two different EMT genes included in the model were calculated based on the gene expression values. In addition, a mutational demonstration of the TCGA-OV cohort was performed. The mutation data were visualised by using the R package "maftools" (version 2.12.0).

## 2.5 Immune infiltration association analysis and drug therapy sensitivity assessment of EMT risk model

There is an increasing emphasis on immunotherapy in the field of oncology treatment, and infiltrating immune cells in the tumour microenvironment are an important cellular component and have a potential role in relation to tumour cells that may influence tumour progression and patient prognosis survival. To fully understand whether key EMT genes are associated with immune activity, this paper uses the R software GSVA (18) package to assess the expression levels of target EMT genes in relation to the infiltration of immune cells in tumour tissue. The Single Sample Gene Set Enrichment AnalysisA (ss GSEA) algorithm is based on the principle of calculating an enrichment score for a given gene set for each sample. The ss GSEA in immune infiltration uses markers specific to each type of immune cell as a gene set to calculate an enrichment score for each type of immune cell in each sample, inferring the infiltration of immune cells in each sample. the Xcell algorithm allows conversion of gene expression profiles to enrichment scores for 64 immune and stromal cell types across samples. Differences in the composition of cell types across subjects can identify cellular targets of disease and suggest novel therapeutic strategies. Therefore, we calculated immune infiltration scores using 2 methods: ssGSEA, xCell algorithm, visualised with box line plots, heat maps and scatter plots, respectively. ssGSEA calculates enrichment scores for single samples and gene set pairs to determine the degree of immune infiltration. xCell quantifies the abundance of 67 immune cells using transcriptomic data. In addition, adjusting for these variants allows detection of true gene expression differences and improves interpretation of downstream analyses.

To clarify the relationship between EMT immune infiltration and tumour immunotherapy and drug sensitivity, we further evaluated the efficacy of EMT risk scores stratified with tumour drug resistance and multiple drug therapy sensitivity phenotypes. Sensitivity scores were calculated for drugs in the GDSC database based on the R package "oncoPredict", thus extending the clinical application of the EMT risk score.

## 2.6 GAS1 gene in epithelial mesenchymal transition regulates the development of SOC invasion - experimental validation at the cellular level

Firstly, we collected clinical samples of ovarian cancer by collecting 10 tumour samples and 10 paracancerous tissues. The

qPCR technique was used to obtain the transcript expression of GAS1 gene in tumour tissues and normal tissues. qPCR experiments were performed in a similar way as reported in previous studies.The further ovarian cancer cell level experiments were set up as follows.

### 2.6.1 Cell culture and gas1 knockdown validation

SKOV3 as well as 3AO cell lines were obtained from the ATCC repository. All 2 cell lines were cultured in McCoy's 5a medium supplemented with 10% fetal bovine serum (FBS). (GIBCO, Thermo Fisher, Carlsbad, CA) and 1% antibiotic-antimycotic (#15240062, Thermo Fisher). All cell lines were stored in a humidified incubator at 37°C containing 5% CO2. GAS1 stable knockdown clones of both cell lines were obtained by using Mission shRNA (knockdown [KD]1: TRCN0000084044, KD2: TRCN0000294078, Sigma). MISSION(R) pLKO.1-puro empty vector (SHC001, Sigma) was used as shRNA control (Ctl). For lentivirus generation, HEK293T cells were transfected with 4.05 μg of target plasmid, 0.45 μg of pCMV-VSV-G (#8584, Addgene) and 3.5 μg of pCMV delta R8.2 (#12263, Addgene) using Lipofectamine 2000 (Invitrogen) for 24 hours. Cells were incubated with virus-containing supernatant and 8 ug/mL of polyethylene glycol for 24 hours. Purimycin was screened for 2 weeks. For tail vein injection, GAS1 shCtl or knockout SKOV3 as well as 3AO cells were stably transduced with a lentiviral vector. Transduction was performed with a lentiviral vector carrying red fluorescent protein (RFP), followed by fluorescence-activated cell sorting (FACSAria II, BD Biosciences, San Jose, CA) and expanded *in vitro*. WB experiments were performed to analyse the transcriptional expression of GAS1 in control, SI-control, and SI-GAS1 groups.

### 2.6.2 CCK-8 test

Logarithmic growth stage cells were taken, digested with 0.25% trypsin, centrifuged, diluted into single cell suspensions with complete medium containing 10% fetal bovine serum, inoculated into 96-well plates, adjusted to 8 000 cells/well, incubated at 37 °C in a 5% CO2 incubator, and 20 μL of CCK8 solution was added to each well at 24 h, 48 h, 72 h and 96 h. After 2 h, the absorbance (A) value at 450 nm was measured by enzyme marker. The absorbance (A) values at 450 nm were measured on an enzyme marker after 2 h.

### 2.6.3 Colony formation assay

To assess colony formation, 100 shCtl or GAS1 knockdown cells were plated in triplicate on 6-well plates. 15 days later, cells were fixed in methanol for 2 min and stained with 0.2% (W/V) crystal violet for 30 min. Colonies were counted using GelCount™ (Oxford Optoronix, Abingdon, UK) and the values were normalised to the control.

### 2.6.4 Transwell experiment

Matrix gel was used to spread the plates, cell suspensions were made, walled cells were digested using trypsin, washed with PBS and then inoculated cells were resuspended in serum-free medium, the cell suspensions were added to Transwell chambers, incubated for 24 h and fixed, and finally stained and counted. Control, SI-control and SI-GAS1 invasion were analysed.

## 2.7 Statistical analysis

Statistical analysis was performed using R tools and GraphPad Prism8. Data were expressed as mean ± SD (standard deviation) from 3 independent measurements. Each experiment was repeated 3 times, and the $P < 0.05$ means statistically significant.

# 3 Results

## 3.1 Annotated analysis and downscaling clustering of single cell samples from GSE154600

First, we performed cell type annotation analysis and downlink clustering on four single-cell samples from the GSE154600 database. Major cell clustering and downlinked expression analysis were performed using the t-sne method. First, we performed a generalized analysis of the major cell types in the obtained SOC single-cell samples and color differentiated them according to the differences in the obtained cell types. Six main cell clusters were identified and labeled in the samples, and the major cell clusters were labeled in the form of (Figure 1A):T cell, Myeloid, Fibroblast, Epithelial Cell, Endothelial cell, and B cell. there were some differences in the expression distribution and content percentage of each cell cluster, but all of them had significant expression in the SOC single-cell The expression distribution and content of each cell group differed somewhat, but all of them were significantly expressed in the SOC single cell samples. T cells, Fibroblast and Myeloid were the main cells expressed in SOC, except for Endothelial and B cells, which were relatively less expressed. Further, we spatially annotated the respective spatial distribution as well as cell expression correlations of the four samples (Figure 1B). The results showed that Epithelial, Myeloid as well as T cells were predominantly expressed in case 1, Fibroblast and T cells were predominantly present in case 2, Epithelial and T cells were the predominantly expressed cells in patient 3, and the predominantly expressed cells in patient 4 were proximal to those in patient 3. EMT-specific markers that have been widely formalized annotate the spatial distribution of expression in the annotated SOC cell types. The main EMT markers included: EPCAM, LUM, RAMP2, COL3A1, CD79A, CD3D, CD8A, LYZ, CD68 (Figure 1C). expression of EPCAM overlaps with Epithelial cell enrichment, LUM is mainly located in patient 2 and Fibroblast, T cell expression regions, RAMP COL3A1 was present in a variety of cell types, including Fibroblast, T cells, etc. CD79A was present in SOC samples as a marker for B cells. CD3C, CD8A were expressed in the T cell region. CD68 were mainly associated with Meloid distribution. In Figure S1A, we performed a histogram analysis of the percentage of major cell types for the four SOC samples. For sample 1, T cells, Myeloid, and Epithelial were the predominant cell types. Myeloid expression was lacking in sample 2, and the relative

10.3389/fendo.2023.1196094

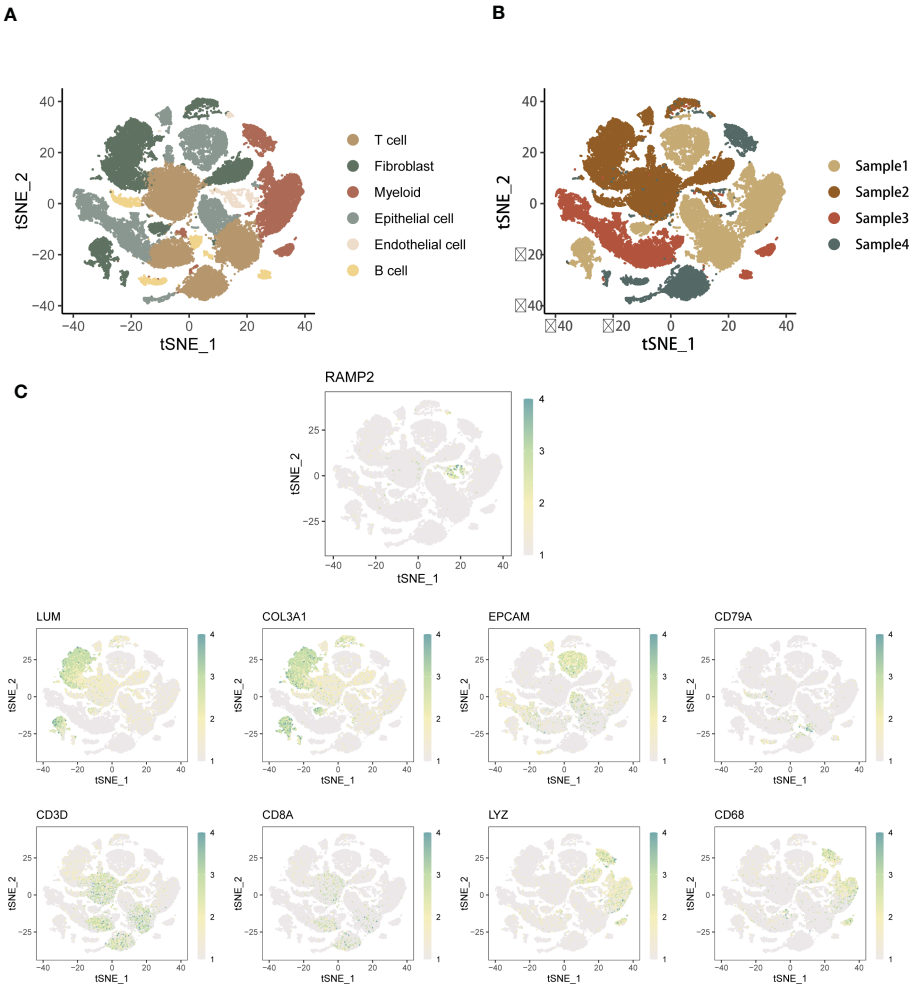



FIGURE 1
Annotated analysis and descending clusters of single cell samples from GSE154600. **(A)** scRNA-seq analysis of 4 soc samples from GSE154600, t-sne revealed cell clustering including Hepatocyte, T cell, Myeloid, Fibroblast, Epithelial Cell, Endothelial cell, and B cell; **(B)** Single cell sequencing Annotated maps of spatial distribution were obtained for four SOC patients; **(C)** Annotated maps of the spatial distribution of expression in annotated SOC cell types by widely formalized EMT-specific markers. The main EMT markers included were: EPCAM, LUM, RAMP2, COL3A1, CD79A, CD3D, CD8A, LYZ, CD68 (blue-green represents high expression, white represents low expression).

expression was similar in samples 3 and 4, with Epithelial, and T cells occupying a larger expression.

## 3.2 Signaling communication and pathway enrichment analysis of major annotated cell types of EMT-related SOC

The above study gave us a preliminary understanding of the major cell types and annotation of SOC single cell samples, with immune cells occupying the major cell expression and functional distribution weights. Therefore, we further describe the SOC tumor pathways and biological functions associated with the analysis of EMT expression. Fibroblast is the most active and interacting cell type in SOC samples, and its signals are mainly directed to Myeloid and Endothelial cells. The communication from Endothelial cells was mainly focused on Myeloid, while less signals from B cells occurred in Epithelial. Further, GSVA scoring was used to reveal the correlation between the expression of different cell types and the main tumor pathways (Figures S1C, D). T cells positively correlated with ALLOGRAFT-REJECTION expression, and Fibroblast was mainly enriched in Epithelial mesenchymal transition. Myeloid is positively co-expressed with multiple functional pathways, including E2F_TARGETS, inflammatory response and fatty acid metabolism. B cells were mainly enriched in Epithelial mesenchymal transition. This shows a significant correlation between the differential expression of EMT in different cell types further demonstrating the relationship between the enrichment of tumor functional pathways.

To understand the spatial distribution of expression of different SOC annotated cell types, we also used the AddModuleScore function of the Seruat package to score the gene set enrichment of the EMT pathway in the GSEA database into two groups, EMT-high level and EMT-low level. Spatial annotation was seen (Figure 2A), Fibroblast was mainly EMT-high level, T cell was mainly EMT-low level group, and other cells were a mixture of the 2

10.3389/fendo.2023.1196094

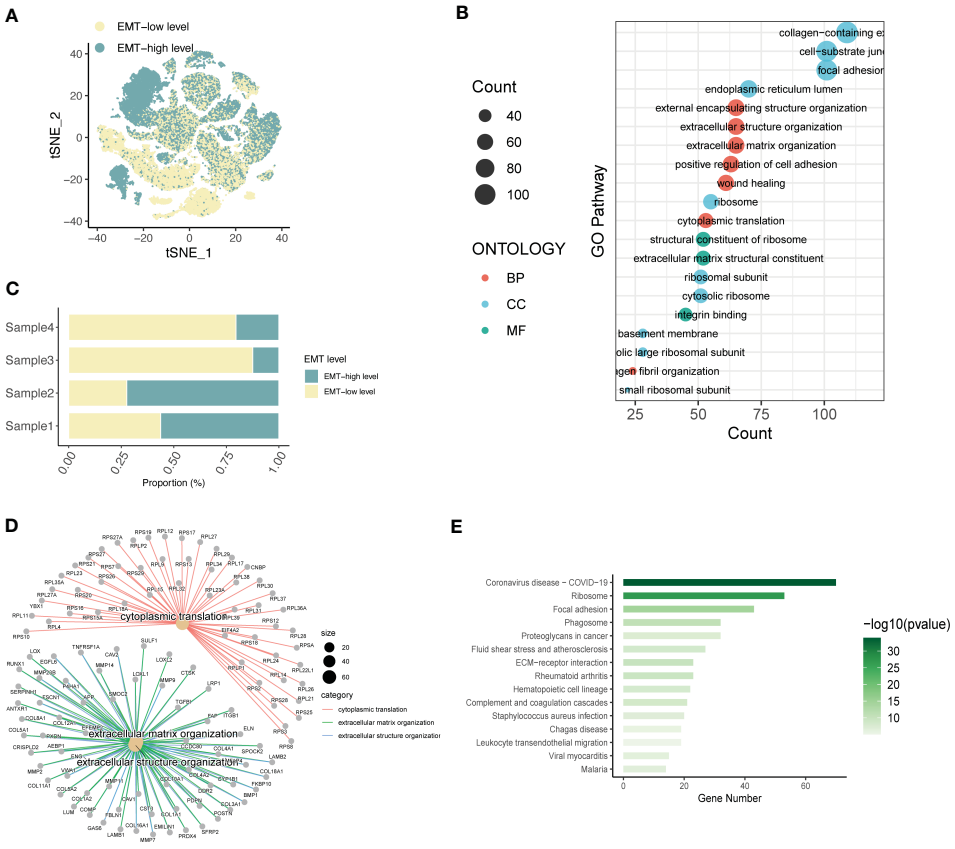



FIGURE 2
Expression annotation and biological pathway enrichment analysis of EMT-related genes. **(A)** Single-cell spatial distribution maps of EMT signature-high and EMT signature-low groups obtained from gene set enrichment scoring; **(B)** Percentage of EMT-high and EMT-Low expression in four SOC single-cell data samples; **(C)** GO analysis depicting the major enrichment pathways mediating tumor infiltration by the major acting cell types in LIHC and **(D)** Depiction of extracellular matrix network analysis; **(E)** Histogram of KEGG analysis showing the enrichment of tumour pathways associated with high EMT expression.

groups. It is speculated that this may be related to the role of different cell types in the EMT process of SOC. Fibroblasts are the main function-playing role of EMT, while T cells are involved in the anti-fibrotic process and tumor immune response, and thus have lower expression of related genes. In addition, we analyzed the EMT gene expression in four SOC patients using a hierarchical bar graph (Figure 2B). Patient 1 and patient 2 were predominantly EMT-high, while patient 3 and patient 4 had predominant EMT-Low occupant expression.

Based on this, the main functional pathways and biological mechanisms by which EMT plays a role in SOC are the next major thing we urgently need to understand, so we performed KEGG analysis and GO analysis on this basis. GO analysis revealed (Figure 2C) that the differential genes of ECM were mainly enriched in external encapsulating structure organization, extracellular structure organization, extracellular matrix organizatio, positive regulation of cell adhesion, wound healing; cell components are mainly enriched in collagen-containing, cell-substrate junction, focal adhesio, endoplasmic reticulum lumen, ribosome; molecular functional enrichment is expressed in structural constituent of ribosomal In addition, in Figure 2D, we also performed the analysis of extracellular matrix components and related gene expression. KEGG analysis further demonstrated the main relevant functions and enrichment pathways of EMT in SOC development. The main components include Ribosome,Focal adhesion,Phagosome,Proteoglycans in cancer,Fluid shear stress and atherosclerosis.

## 3.3 Construction of SOC prognostic biomarkers for EMT genetic risk model and functional assessment

Through the previous studies on signaling communication and tumor pathway networks, we found that the major gene types of EMT may be closely associated with the development of SOC and invasive metastasis. Considering that there is a lack of EMT-related prognostic risk markers for SOC, we further analyzed the prognostic value of ECM-related genes for SOC. Key risk genes for the major differential genes of ECM were searched and prognostic models were composed. Using the TCGA-OV cohort as the training group, the cohort selected 7 genes of prognostic significance from 22 genes according to Lasso regression and constructed prognostic models. Multi-factorial Cox regression and LASSO analysis were performed on 50 prognosis-related

10.3389/fendo.2023.1196094

EMT risk genes (Figures 3A, B), and 7 EMT risk genes with significant correlation with prognosis of ovarian cancer patients were obtained ($P < 0.01$), which constituted a prognostic prediction model for ovarian cancer. Meanwhile, the Risk Score (RS) of each patient was calculated in this paper. Then, patients were divided into high-risk and low-risk groups according to the median of risk score, and the forest plot of risk score and risk genes was drawn (Figure 3C). Among them, SEINC2, GAS1 and EMP1 were significantly positively expressed genes, while several other risk genes were mainly negatively expressed. The formula was: risk score = SERINC2*0.08956304 - CXCR4 *0.109059452 + GAS1*0.15692187+ EMP1*0.13645030 - IFI27*0.04696752 - CD48*0.05994817 - LYAR*0.16842919. Figure 3D shows that both in the training group of TCGA-OV and in the independent validation group of 173 ovarian cancer patients in GSE53963. The grouping curves based on EMT risk marker expression demonstrated significant prognostic stratification and survival predictive power. In addition, we analyzed the expression of seven EMT risk genes individually, in the EMT-High and EMT-Low groups. All seven expressed genes possessed significant differences between the two groups (Figure 3E). In addition, we examined the influence of the above model genes on the clinical prognosis of ovarian cancer, and found that the expressions of SERINC2, GAS1 and EMP1 were negatively correlated with clinical outcome, while the expressions of LYAR, IFI27, CXCR4 and CD48 were positively correlated with clinical outcome (Figure S2). Using Spearman correlation analysis, we described the correlation between EMT risk score and the expression of multiple tumor pathways and immune factors (Figure 3F). CXCL10, LAG3, CCL5, IFNG, CD274, and CD40 showed a significant negative correlation with EMT score expression.CD276, TNFSF4, CX3CL1, TNFSF9, and TGFB1, on the other hand, showed a significant positive correlation with the expression of EMT risk genes.

In addition, we analyzed clinical indicators and risk scores by multifactorial COX regression and established an integrated model based on EMT. to meet the practical decision-making needs of clinical visibility and multifactorial integrated analysis. The final model consisting of Stage, age, and Riskscore was finally screened by COX analysis and incorporated, as shown in Figure 4A. Figure 4B shows the main expression correlations of the 7 EMT risk genes. positive correlation of SERINC2, EMP1 expression is strong. high correlation of CXCR4, IF127, LYAR expression. gas1 is mainly correlated with EMP1 and CD48. We also analyzed the main locations of the major genes in the EMT markers on human

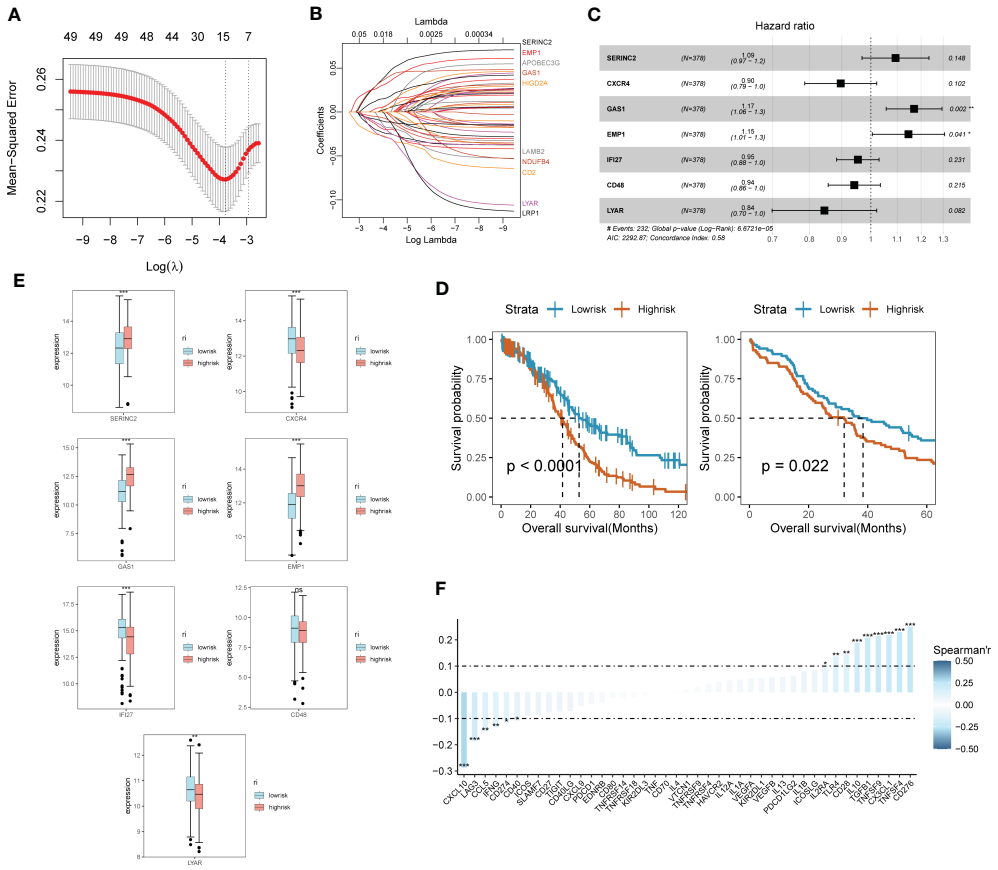



FIGURE 3
SOC survival model construction and functional validation of prognostic stratification for EMT risk markers. **(A)** Feature screening process curves from LASSO regression; **(B)** LASSO regression reveals feature screening of SOC risk model; **(C)** Forest plot demonstrating major EMT risk genes; **(D)** Prognostic value of EMT survival biomarkers assessed by survival curve Kaplan-Meier analysis in TCGA-OV training group and GEO validation; **(E)** Seven major EMT risk genes with EMT-HIGH and EMT-LOW expression box line plots; **(F)** Spearman expression correlation plots reveal the correlation between the expression of immune factors and EMT risk genes.

10.3389/fendo.2023.1196094

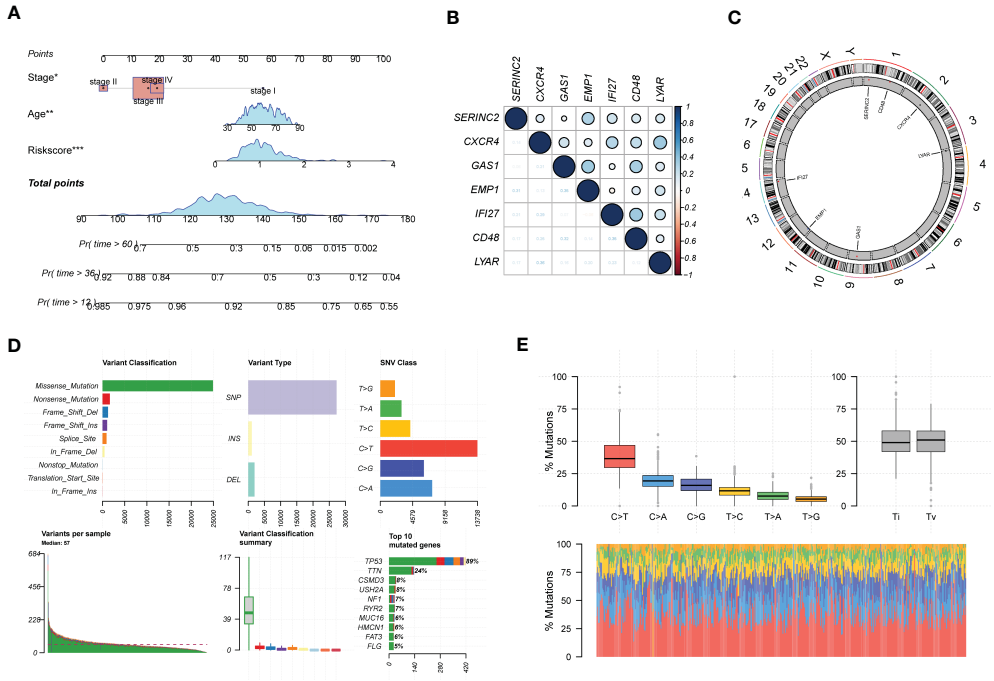



FIGURE 4
In-depth profiling and mutation analysis of EMT risk biomarkers. **(A)** Prognostic prediction exhibition of SOC risk markers consisting of age, Stage, and EMT Riskscore; **(B)** Correlation heat map demonstrating the expression correlation of seven EMT risk genes; **(C)** Chromosome distribution map revealing the chromosomal expression locations of seven EMT genes; **(D)** Chromosomal mutations and gene mutations depicted; **(E)** Base pair mutations and box line plot of mutation probabilities. *p < 0.05, **p < 0.01, ***p < 0.001.

chromosomes (Figure 4C). SERINC2 and CD48 were located on chromosome 1, CXCR4 on chromosome 2, LYAR on chromosome 4, RAC1 on chromosome 7, and GAS1 on chromosome 9.EMP1 and IF127 were located on chromosomes 12 and 14, respectively. In addition, we also describe the mutations in the chromosomes of EMT-related genes in Figures 4D, E. Missense mutations and NONSENSE mutations are the major mutation forms. And the main mutated genes were TP53 and TTN.

## 3.4 Association assessment of SOC immune infiltration with EMT model expression

The results of gene pathway enrichment and immunological analyses initially indicated that the regulatory role of EMT in SOC showed significant correlation mainly with immune pathway-related expression. Therefore, deeper immune infiltration analysis was used to characterize the impact of EMT-related risk genes in SOC in association with immune infiltration and regulation of immune cell expression. We chose the ssGSEA, xCell algorithm to calculate the immune infiltration score and visualize it in multiple forms. First of all, in the immune cell enrichment score calculated by ssGSEA we found (Figure 5A). Eosinophil, Mast cells, NK cells, T follicular helper cells, and Th1 possessed significant expression in the EMT high risk group. In Figure 5B, we calculated the correlation between the distribution of multiple tumor immune cells and found that the tumor distribution of multiple immune cells possessed significant specificity-related characteristics. To describe the direct association of EMT with tumor immune infiltration, we calculated and depicted the correlation between EMT risk score and immune cell expression by linear correlation scatter plots (Figures 5C, D). The results showed that activated CD4 T cells, activated CD8 T cells possessed a significant negative linear correlation with EMT risk score. eosinophil, Immature dendritic cells, Mast Cell, macrophages, NK cells and NK T cells showed a positive correlation with risk score expression. In Figures S3A, S3B, we also depicted the gene mutations in the high-risk and low-risk groups by chromosome analysis. TP53, TTN, and CSMD3 were the major mutation types in both the high- and low-risk EMT groups. XCell method further demonstrated the expression correlation between EMT risk grouping and tumor immune infiltration (Figures S3C, S3D).

In addition, to understand the correlation between the expression of EMT risk genes and the efficacy of tumor immunotherapy therapies, we also calculated sensitivity scores for the drugs in the GDSC database based on the R package “oncoPredict”. Figure 6A shows the correlation between EMT risk genes and the sensitivity of various immunotherapeutic drugs. In Figures 6B, C, we further clarify the correlation between the expression of EMT risk genes and treatment sensitivity, TAF1_5496_1732, and ML323_1629 all had higher treatment sensitivity scores in the high-risk group than in patients in the low-risk group for EMT. AGI-6780_1634, I-BRD9_1928, Pevonedistat_1529, and OF-1_1853 also had better treatment outcomes in the high-risk group for EMT.

10.3389/fendo.2023.1196094

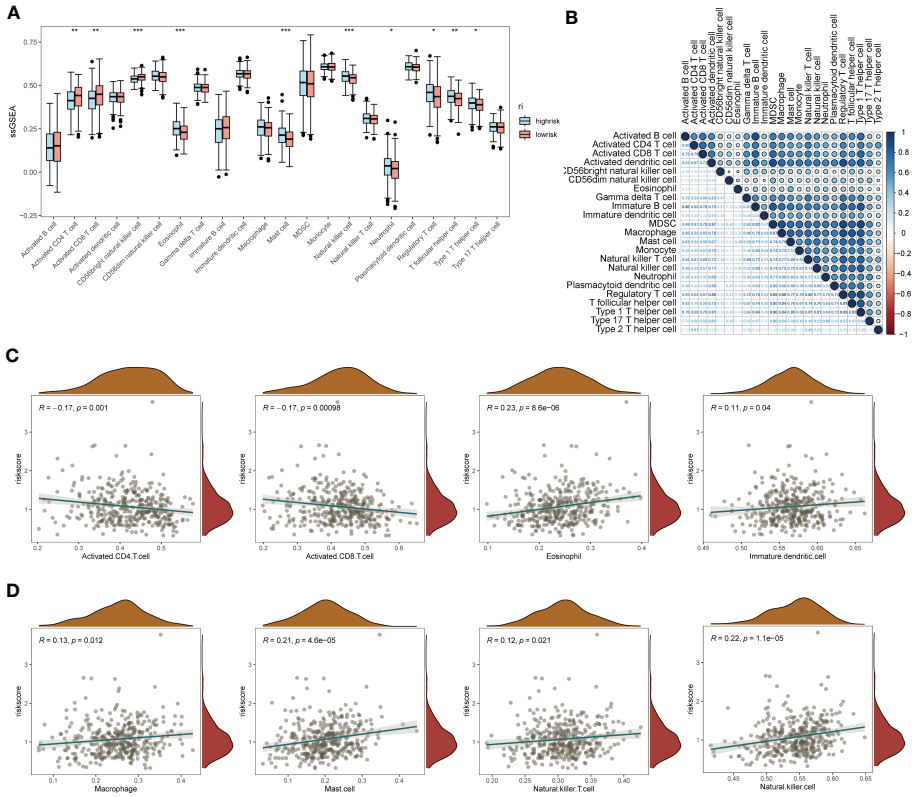



FIGURE 5
ssGSEA immuno-infiltration analysis of the EMT risk gene model. **(A)** Box line plot of ssGSEA revealing the differential box line plot of expression of major immune cells and immune pathways in the low and high risk groups for EMT; **(B)** Correlation heat map revealing the correlation of expression of major immune cells; **(C)** Scatter plot of linear correlation of expression of several major immune cells with EMT risk score; **(D)** Scatter plot of linear correlation of expression of several additional major immune cells with EMT risk model. *p < 0.05, **p < 0.01, ***p < 0.001.

## 3.5 GAS1 in epithelial mesenchymal transition regulates the development of SOC invasion - experimental validation at the cellular level

For the obtained tissue samples from ovarian cancer and paracancer, qPCR showed that GAS1 was significantly highly expressed in the tumour samples, with significant differences in GAS1 transcript expression between the two groups (Figure 7A). This firmly established the need for subsequent cellular experiments to clarify the specific effects of GAS1 on the regulation of invasion and metastasis in ovarian cancer cells.

Subsequent cellular experiments were then carried out. We first verified the successful knockdown of the GAS1 gene in the strips by WB experiments. WB strips and WB quantitative analysis (Figures 7B, C) showed that GAS1 expression was significantly reduced in the SI-GAS1 group compared to the control and SI-control. On this basis, we assessed the effect of GAS1 gene affecting the proliferation of ovarian cancer cell lines by colony assay. Figure 7D shows the results of the CCK-8 experiments at different time periods for the cell survival rates of the two ovarian cancer cell lines. Contrast control, SI-control. The survival rate of ovarian cancer cells in the SI-GAS1 group decreased significantly as the culture time increased. Especially after 96h of culture, which suggests that GAS1 is a key gene for ovarian cancer cells to maintain viability. Figure 7E shows that pancreatic cancer cell colony formation was significantly reduced in the SI-GAS1 group. This suggests that genes associated with epithelial mesenchymal transition can regulate SOC aggregation and chemotaxis. Similarly, we also clarified this differential expression relationship using bar graphs (Figure 7F). Transwell assays showed that knockdown of the GAS1 gene significantly inhibited the proliferation, invasion and migration of ovarian cancer cells (Figures 7G, H). This suggests that epithelial mesenchymal transition and its associated genes profoundly influence the development of ovarian cancer. The flow chart of analyses was showed in Figure 8.

# 4 Discussions

Ovarian cancer has the highest mortality rate among all gynecologic malignancies and is on the rise. GLOBOCAN 2020 (19) showed 313,959 new cases of ovarian cancer and 207,252 new deaths worldwide in 2020, with 5-year survival rates below 40% (20) and mortality rates as high as 60%. Despite great progress in surgery and drug therapy, complete/partial remission is difficult to maintain in ovarian cancer patients, and exploring reliable biomarkers and precise molecular mechanisms is crucial for early diagnosis, treatment and prognosis of OC. In recent years, rapid advances in bioinformatics have enabled gene microarrays and sequencing data to provide a convenient

10.3389/fendo.2023.1196094

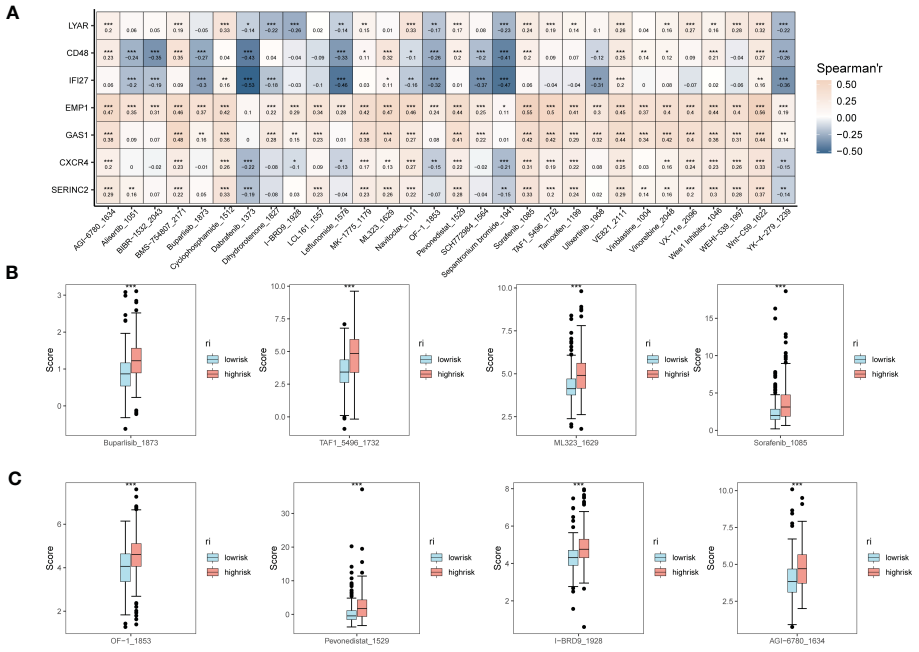



FIGURE 6
Expression correlation assessment of tumor treatment sensitivity and EMT risk score. **(A)** Correlation heat map demonstrating the therapeutic sensitivity of seven EMT risk genes with multiple chemotherapeutic agents; **(B)** Box plots of therapeutic sensitivity of Buparlisib_1873, Sorafenib_1085, TAF1_5496_1732, ML323_1629 in EMT-HIGH and EMT-LOW; **(C)** AGI-6780_ 1634, I-BRD9_1928, Pevonedistat_1529, OF-1_1853 box-line plots of treatment sensitivity in EMT-HIGH and EMT-LOW. *p < 0.05, **p < 0.01, ***p < 0.001.

and comprehensive platform for exploring general genetic alterations in tumors, identifying DEGs, and elucidating their molecular mechanisms for diagnosis, treatment, and prognosis (21, 22). Wu et al (23) found that Sialyltransferase ST3GAL1 promotes cell migration, invasion and TGF-β1-induced EMT, and confers ovarian cancer with paclitaxel resistance. Xu et al (24) revealed the four-EMT gene model used to predict the outcome of patients with HGSOC, the ability of mCAF to enhance the invasion of ovarian cancer cells, and the potential therapeutic value of anti-Tigit therapy through the transcriptome results of single cell sequencing analysis. Currently, biomarkers are not sufficiently studied by investigators and the results of DEGs are inconsistent; therefore, reanalysis of relevant database data may provide new ideas for current therapeutic studies in OC (25) to address the issues of prognosis, drug resistance, and recurrence of ovarian cancer. In this study, based on previous studies on the role of epithelial mesenchymal transition in SOC, EMT risk genes were identified and their spatial distribution was depicted by single-cell sequencing analysis. The main associations between EMT risk genes and tumor immunity were described by KEGG analysis and GO analysis. Meanwhile, the main functional role of EMT in SOC development was clarified by immune infiltration analysis and prognostic model construction. This lays the foundation for future construction of EMT-related risk markers and the study of their functional characteristics in the immune infiltration and pathway regulation of SOC.

Clinical work has revealed a high degree of heterogeneity in the growth, invasive metastasis, chemoresistance and other behaviors of ovarian cancer, suggesting that ovarian cancer is not a single disease but a group of diseases with different molecular phenotypes, pathogenesis and prognosis.In 2004, American pathologists proposed the doctrine of ovarian cancer dichotomy (26), which divided ovarian cancer into type I and type II ovarian cancer, followed by successive studies that found that the fallopian tube Subsequently, successive studies found the existence of lesions and precancerous lesions similar to high-grade plasmacytoma in the mucosa of the fallopian tubes, and therefore proposed the theory of tubal origin (27). Once the intraepithelial carcinoma of the fallopian tube is formed, the cells detach from the cilia to reach the ovarian surface and then form an invasive carcinoma. Migration of intraepithelial carcinoma cells from the fallopian tube to the ovary is a very important step in ovarian carcinogenesis. It has been shown that growth factors and hormones secreted by the ovary (28), such as TGFβ and activator A, have a role in inducing the migration of cancer cells to the ovarian surface. Activin A, which is released from the TGFβ superfamily in the follicular fluid during ovulation, can induce EMT and promote the migration of tubal epithelial cells and high-grade plasmacytoid ovarian cancer cells by activating PI3K/AKT and MEK/ERK pathways (29). In junctional plasmacytoid ovarian tumor cells, downregulation of p53 was found to promote the aggressiveness of junctional tumors by downregulating E-cadherin expression through the PI3K/AKT pathway (30). The main biological functions of EMT process in malignant tumors are to enhance cell motility and cellular drug resistance, and EMT has important research value in the development, clinical diagnosis and treatment of ovarian cancer. We hope to provide new reliable and specific therapeutic targets for ovarian cancer in the near future through in-depth study of EMT and key molecular nodes in the EMT-driven process. This also suggests the important role of constructing EMT-

10.3389/fendo.2023.1196094

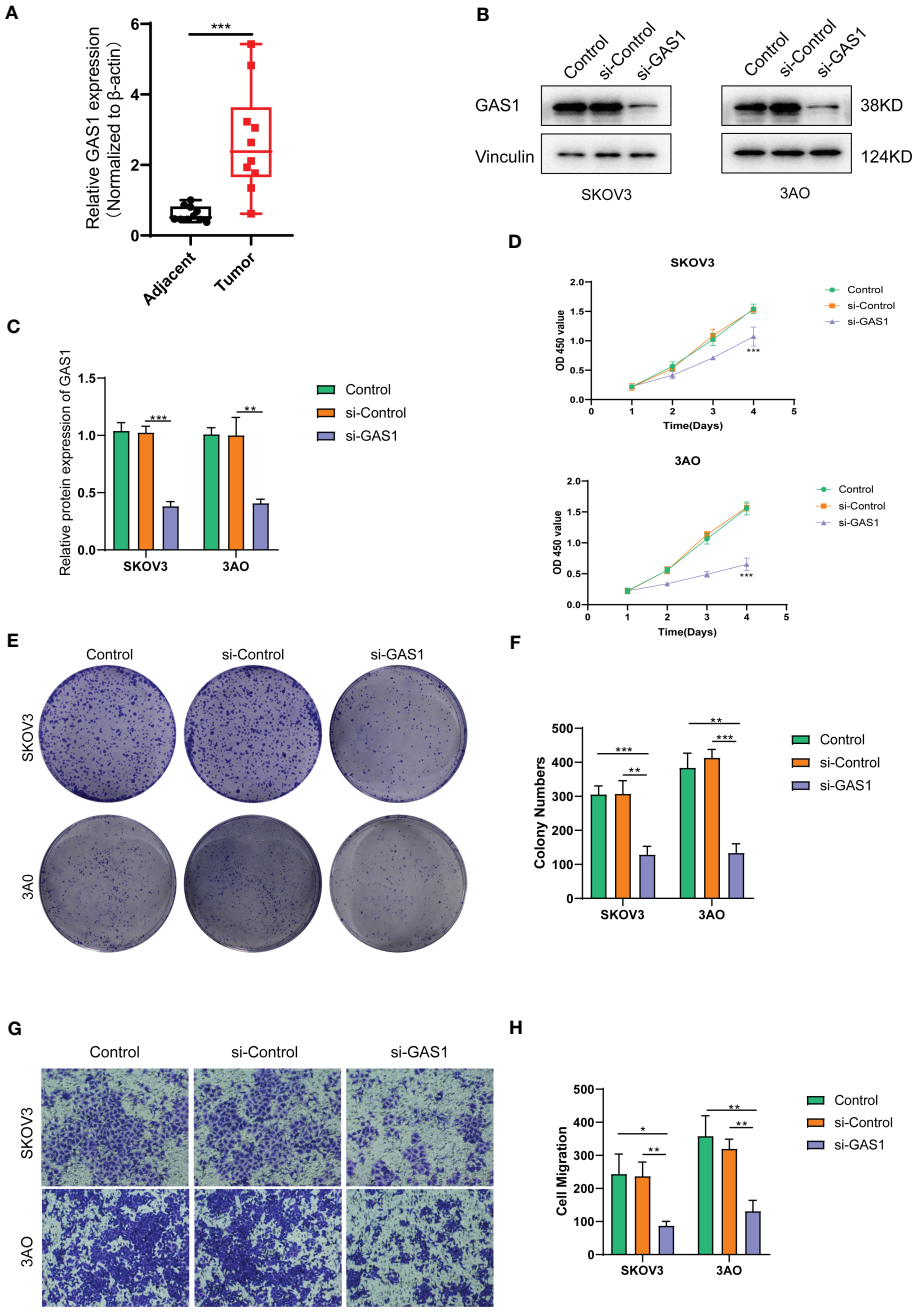



FIGURE 7
SOC validation of GAS1 gene regulation in epithelial mesenchymal transition at the cellular level. **(A)**. qPCR analysis of GAS1 expression histograms in 10 pairs of SOC tumor and paracancer samples; **(B)** WB bands demonstrating GAS1 gene expression in control, SI-control, and SI-GAS1; **(C)** Bar and bar graphs quantifying GAS1 gene expression in control, SI-control, and SI-GAS1; **(D)** CCK8 experiments revealed cell survival of two SOC cell lines in culture for 24h, 48h, 72h and 96h in control, SI-control and SI-GAS1; **(E)** Colony formation in control, SI-control, and SI-GAS1 demonstrated; **(F)** Control, SI-control, and SI- Colony formation histogram analysis of GAS1; **(G)** Transwell invasion assay comparing the effects of control, SI-control, and SI-GAS1 in two SOC cell lines; **(H)** Histogram analysis of Transwell assay for control, SI-control, and SI-GAS1. *p < 0.05, **p < 0.01, ***p < 0.001.

related risk markers for studying ovarian carcinogenesis and development.

The present study has several advantages and limitations. First, this paper is based on a multi-omics (single-cell sequencing data +Bulk sequencing data) model. No such systematic analysis has been conducted to explore the effect of EMT signaling pathway on ovarian cancer. while the prognostic model was constructed and validated using retrospective data from public databases, and more prospective data are needed to validate its clinical utility. Second, this study only included EMT-related models for prognostic modeling, which is difficult to avoid confounding factors, as there are many mutated prognostic genes in ovarian cancer that may be excluded. Further experiments will be conducted in the future to verify the relationship between EMT-related genes and tumor immunity.

Overall, this study constructed an EMT-based risk marker for SOC survival prediction and investigated its main functional characteristics

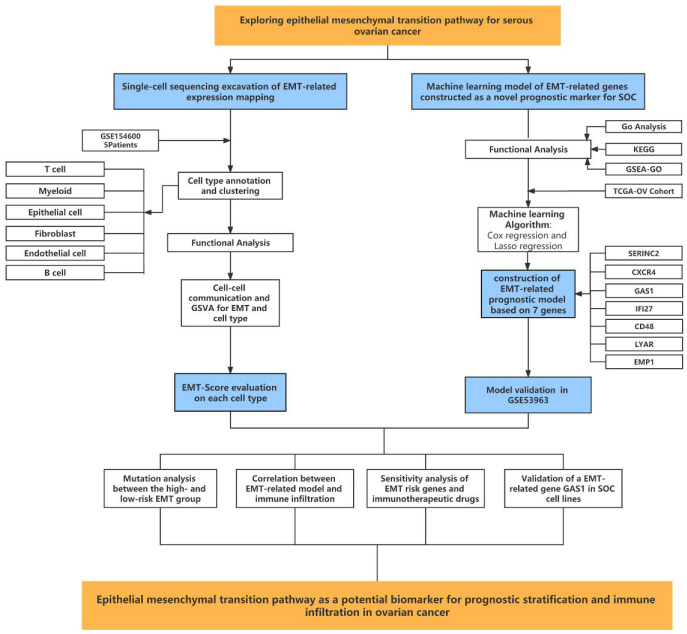



FIGURE 8
Flow chart showing the methodology of the study.

and the exact association between it and tumor immune infiltration. In-depth study of the molecular mechanism of EMT can help to understand ovarian cancer more deeply. Enriching the mechanistic network of EMT in epithelial ovarian cancer will potentially identify potential therapeutic targets for the invasive and metastatic properties of epithelial ovarian cancer, which will help us create new drug targets and intervene in ovarian cancer and provide a more reliable basis for effective treatment of SOC.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

## Ethics statement

The studies involving human participants were reviewed and approved by Cangzhou Central Hospital. The patients/participants provided their written informed consent to participate in this study. Its number is LCYJ: NO. 2019-090.

## Author contributions

QL designed the study. XX, JF performed data analysis. RY drafted the manuscript. QL and JX revised the manuscript. All authors read and approved the final manuscript.

## Acknowledgments

We sincerely appreciate all members who participated in data collection and analysis.

## Conflict of interest

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.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1196094/full#supplementary-material

# References

1. Huang J, Li L, Liu J, Yu J, Wu X, Xu Y, et al. Altered expression of lysosomal associated membrane protein 1 in esophageal squamous cell carcinoma. *Pathol Res Pract* (2017) 213(8):938–42. doi: 10.1016/j.prp.2017.05.008

2. Eisenhauer EA. Real-world evidence in the treatment of ovarian cancer. *Ann Oncol* (2017) 28(suppl_8):i61–5. doi: 10.1093/annonc/mdx443

3. Cho KR, Shih I. Ovarian cancer. *Annu Rev Pathol* (2009) 4:287–313. doi: 10.1146/annurev.pathol.4.110807.092246

4. Menon U, Karpinskyj C, Gentry-Maharaj A. Ovarian cancer prevention and screening. *OBSTET GYNECOL* (2018) 131(5):909–27. doi: 10.1097/AOG.0000000000002580

5. Morand S, Devanaboyina M, Staats H, Stanbery L, Nemunaitis J. Ovarian cancer immunotherapy and personalized medicine. *Int J Mol Sci* (2021) 22(12). doi: 10.3390/ijms22126532

6. Roett MA, Evans P. Ovarian cancer: an overview. *Am FAM PHYSICIAN* (2009) 80(6):609–16.

7. Moschetta M, George A, Kaye SB, Banerjee S. BRCA somatic mutations and epigenetic BRCA modifications in serous ovarian cancer. *Ann Oncol* (2016) 27 (8):1449–55. doi: 10.1093/annonc/mdw142

8. Bowtell DD, Böhm S, Ahmed AA, Aspuria P, Bast RCJ, Beral V, et al. Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. *Nat Rev Cancer* (2015) 15(11):668–79. doi: 10.1038/nrc4019

9. Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol* (2014) 15(3):178–96. doi: 10.1038/nrm3758

10. Dongre A, Weinberg RA. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. *Nat Rev Mol Cell Biol* (2019) 20 (2):69–84. doi: 10.1038/s41580-018-0080-4

11. Chen H, Liu H, Mao M, Tan Y, Mo X, Meng X, et al. Crosstalk between autophagy and epithelial-mesenchymal transition and its application in cancer therapy. *Mol Cancer* (2019) 18(1):101. doi: 10.1186/s12943-019-1030-2

12. Wang H, Mei Y, Luo C, Huang Q, Wang Z, Lu G, et al. Single-cell analyses reveal mechanisms of cancer stem cell maintenance and epithelial-mesenchymal transition in recurrent bladder cancer. *Clin Cancer Res* (2021) 27(22):6265–78. doi: 10.1158/1078-0432.CCR-20-4796

13. Lu W, Kang Y. Epithelial-mesenchymal plasticity in cancer progression and metastasis. *Dev Cell* (2019) 49(3):361–74. doi: 10.1016/j.devcel.2019.04.010

14. Dean M, Davis DA, Burdette JE. Activin a stimulates migration of the fallopian tube epithelium, an origin of high-grade serous ovarian cancer, through non-canonical signaling. *Cancer Lett* (2017) 391:114–24. doi: 10.1016/j.canlet.2017.01.011

15. Al Habyan S, Kalos C, Szymborski J, McCaffrey L. Multicellular detachment generates metastatic spheroids during intra-abdominal dissemination in epithelial ovarian cancer. *ONCOGENE* (2018) 37(37):5127–35. doi: 10.1038/s41388-018-0317-x

16. Rizvi I, Gurkan UA, Tasoglu S, Alagic N, Celli JP, Mensah LB, et al. Flow induces epithelial-mesenchymal transition, cellular heterogeneity and biomarker modulation in 3D ovarian cancer nodules. *P Natl Acad Sci USA* (2013) 110(22):E1974–83. doi: 10.1073/pnas.1216989110

17. Yu G, Wang L, Han Y, He Q. clusterProfiler: an r package for comparing biological themes among gene clusters. *Omics: J Integr Biol* (2012) 16(5):284–7. doi: 10.1089/omi.2011.0118

18. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinf* (2013) 14:7. doi: 10.1186/1471-2105-14-7

19. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660

20. Perales-Puchalt A, Perez-Sanz J, Payne KK, Svoronos N, Allegrezza MJ, Chaurio RA, et al. Frontline science: microbiota reconstitution restores intestinal integrity after cisplatin therapy. *J LEUKOCYTE Biol* (2018) 103(5):799–805. doi: 10.1002/JLB.5HI1117-446RR

21. Wang Y, Sun L, Chen S, Guo S, Yue T, Hou Q, et al. The administration of escherichia coli nissle 1917 ameliorates irinotecan-induced intestinal barrier dysfunction and gut microbial dysbiosis in mice. *Life Sci* (2019) 231:116529. doi: 10.1016/j.lfs.2019.06.004

22. Pflug N, Kluth S, Vehreschild JJ, Bahlo J, Tacke D, Biehl L, et al. Efficacy of antineoplastic treatment is associated with the use of antibiotics that modulate intestinal microbiota. *ONCOIMMUNOLOGY* (2016) 5(6):e1150399. doi: 10.1080/2162402X.2016.1150399

23. Wu X, Zhao J, Ruan Y, Sun L, Xu C, Jiang H. Sialyltransferase ST3GAL1 promotes cell migration, invasion, and TGF-β1-induced EMT and confers paclitaxel resistance in ovarian cancer. *Cell Death Dis* (2018) 9(11):1102. doi: 10.1038/s41419-018-1101-0

24. Xu J, Fang Y, Chen K, Li S, Tang S, Ren Y, et al. Single-cell RNA sequencing reveals the tissue architecture in human high-grade serous ovarian cancer. *Clin Cancer Res* (2022) 28(16):3590–602. doi: 10.1158/1078-0432.CCR-22-0296

25. Davar D, Dzutsev AK, McCulloch JA, Rodrigues RR, Chauvin J, Morrison RM, et al. Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients. ; (2021) pp:595–602. doi: 10.1126/science.abf3363

26. Lei JH, Lee M, Miao K, Huang Z, Yao Z, Zhang A, et al. Activation of FGFR2 signaling suppresses BRCA1 and drives triple-negative mammary tumorigenesis that is sensitive to immunotherapy. *Advanced Sci (Weinheim Baden-Wurttemberg Germany)* (2021) 8(21):e2100974. doi: 10.1002/advs.202100974

27. Tone AA. Taking the tube: from normal fallopian tube epithelium to ovarian high-grade serous carcinoma. *Clin OBSTET GYNECOL* (2017) 60(4):697–710. doi: 10.1097/GRF.0000000000000313

28. Hooda J, Novak M, Salomon MP, Matsuba C, Ramos RI, MacDuffie E, et al. Early loss of histone H2B monoubiquitylation alters chromatin accessibility and activates key immune pathways that facilitate progression of ovarian cancer. *Cancer Res* (2019) 79(4):760–72. doi: 10.1158/0008-5472.CAN-18-2297

29. Ullah R, Yin Q, Snell AH, Wan L. RAF-MEK-ERK pathway in cancer evolution and treatment. *Semin Cancer Biol* (2022) 85:123–54. doi: 10.1016/j.semcancer.2021.05.010

30. Chiu W, Huang Y, Tsai H, Chen C, Chang C, Huang S, et al. FOXM1 confers to epithelial-mesenchymal transition, stemness and chemoresistance in epithelial ovarian carcinoma cells. *Oncotarget* (2015) 6(4):2349–65. doi: 10.18632/oncotarget.2957

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 04 July 2023
DOI 10.3389/fendo.2023.1184190



OPEN ACCESS

EDITED BY
Prem P. Kushwaha,
Case Western Reserve University,
United States

REVIEWED BY
Raushan Kumar,
Allahabad University, India
Jad Ahmad Degheili,
Children's Hospital of Eastern Ontario
(CHEO), Canada

*CORRESPONDENCE
Hao lin Wang
haolinwang@cqmu.edu.cn
Ting ting Zhang
zhangting835566@126.com

†These authors have contributed
equally to this work

RECEIVED 11 March 2023
ACCEPTED 09 June 2023
PUBLISHED 04 July 2023

CITATION
Liu Xz, Duan Mj, Huang Hd, Zhang Y, Xiang Ty, Niu Wc, Zhou B, Wang Hl and Zhang Tt (2023) Predicting diabetic kidney disease for type 2 diabetes mellitus by machine learning in the real world: a multicenter retrospective study.
*Front. Endocrinol.* 14:1184190.
doi: 10.3389/fendo.2023.1184190



# Predicting diabetic kidney disease for type 2 diabetes mellitus by machine learning in the real world: a multicenter retrospective study

Xiao zhu Liu[1,2†], Minjie Duan[2,3†], Hao dong Huang[2,3], Yang Zhang[2,3], Tian yu Xiang[4], Wu ceng Niu[5], Bei Zhou[1], Hao lin Wang[3*] and Ting ting Zhang[6*]

[1]Department of Cardiology, the Second Affiliated Hospital of Chongqing Medical University, Chongqing, China, [2]Medical Data Science Academy, Chongqing Medical University, Chongqing, China, [3]College of Medical Informatics, Chongqing Medical University, Chongqing, China, [4]Information Center, The University-Town Hospital of Chongqing Medical University, Chongqing, China, [5]Department of Nuclear Medicine, Handan First Hospital, Hebei, China, [6]Department of Endocrinology, Fifth Medical Center of Chinese People's Liberation Army (PLA) Hospital, Beijing, China

**Objective:** Diabetic kidney disease (DKD) has been reported as a main microvascular complication of diabetes mellitus. Although renal biopsy is capable of distinguishing DKD from Non Diabetic kidney disease(NDKD), no gold standard has been validated to assess the development of DKD.This study aimed to build an auxiliary diagnosis model for type 2 Diabetic kidney disease (T2DKD) based on machine learning algorithms.

**Methods:** Clinical data on 3624 individuals with type 2 diabetes (T2DM) was gathered from January 1, 2019 to December 31, 2019 using a multi-center retrospective database. The data fell into a training set and a validation set at random at a ratio of 8:2. To identify critical clinical variables, the absolute shrinkage and selection operator with the lowest number was employed. Fifteen machine learning models were built to support the diagnosis of T2DKD, and the optimal model was selected in accordance with the area under the receiver operating characteristic curve (AUC) and accuracy. The model was improved with the use of Bayesian Optimization methods. The Shapley Additive explanations (SHAP) approach was used to illustrate prediction findings.

**Results:** DKD was diagnosed in 1856 (51.2 percent) of the 3624 individuals within the final cohort. As revealed by the SHAP findings, the Categorical Boosting (CatBoost) model achieved the optimal performance in the prediction of the risk of T2DKD, with an AUC of 0.86 based on the top 38 characteristics. The SHAP findings suggested that a simplified CatBoost model with an AUC of 0.84 was built in accordance with the top 12 characteristics. The more basic model features consisted of systolic blood pressure (SBP), creatinine (CREA), length of stay (LOS),

thrombin time (TT), Age, prothrombin time (PT), platelet large cell ratio (P-LCR), albumin (ALB), glucose (GLU), fibrinogen (FIB-C), red blood cell distribution width-standard deviation (RDW-SD), as well as hemoglobin A1C(HbA1C).

**Conclusion:** A machine learning-based model for the prediction of the risk of developing T2DKD was built, and its effectiveness was verified. The CatBoost model can contribute to the diagnosis of T2DKD. Clinicians could gain more insights into the outcomes if the ML model is made interpretable.

KEYWORDS

type 2 diabetes mellitus, diabetic kidney disease, machine learning, prediction, CatBoost model

# Introduction

Diabetes mellitus refers to a chronic epidemic metabolic disease with high blood glucose. The latest statistics from the International Diabetes Federation (IDF) suggested that approximately 463 million adults (aged 20-79 years) worldwide would have diabetes by 2019; the number of people with diabetes was estimated to reach 700 million by 2045 (1). Complications of diabetes have been found as the leading cause of death in diabetic patients (2), with 76.4% of diabetic patients reporting at least one complication (3). Diabetic kidney disease (DKD) has been reported as a main microvascular complication of diabetes mellitus, which is characterized by high prevalence, mortality, and treatment costs, but low awareness and poor prevention and treatment rates (4). In China, nearly 20-40% of diabetic patients suffer from DKD, while the awareness rate of DKD is lower than 20%, and the treatment rate is even lower than 50% (5).

The typical progression of DKD refers to an initial increase in urinary albumin excretion (called microalbuminuria), which is accompanied with progression to massive albuminuria and subsequent rapid decline in renal function. As a result, proteinuria has been considered the initial pathway for the progression of declining renal function from the traditional perspective (6). However, the above theory has been challenged since numerous patients with proteinuria have been found to return to normal albumin excretion rates either spontaneously or based on the integrated risk management with DKD (7–11). On that basis, the effectiveness of microalbuminuria as a traditional marker of DKD and the optimal opportunity for intervention are challenged since DKD is generally insidious during onset (12). Although renal biopsy is capable of distinguishing DKD from Diabetic kidney disease (NDKD), no gold standard has been validated to assess the development of DKD. Although increased screening frequency can avoid delayed diagnoses, this is not uniformly implemented. Furthermore, the prevention, early diagnosis and treatment of DKD take on a critical significance in reducing the incidence of cardiovascular events in diabetic patients and improving their survival and quality of life. Accordingly, there is an urgent need for a simple and convenient clinical tool to assess DKD in daily clinical practice.

Developing a risk scoring system based on simple predictors, i.e., clinical data, is considered a vital for monitoring and diagnosing DKD. Machine learning algorithm (ML) show significant advantages in processing a considerable number of data with high-dimensional properties and numerous cases. It is extensively employed for disease prediction (13). Machine learning algorithms can efficiently predict the DKD (14–17). Identification of risk factors for the progression of DKD to ESRD is expected to improve the prognosis by early detection and appropriate intervention (18). Most studies on predictive models for DKD have adopted a single classifier for statistical analysis, and most of them achieved small sample sizes. Under excessive samples and variables, the models will be prone to underfitting or overfitting, and the performance and efficiency could be enhanced. Most of the prediction models developed by foreign researchers apply to the white population, and they are likely to be less applicable to the Asian population (19, 20). Thus, it is of clinical significance in developing ancillary diagnostic models for DKD with the use of ML. However, few large-scale studies have investigated the use of machine learning analysis of clinical characteristics to predict DKD in the Chinese population. A retrospective cohort study was conducted, which involved the collection of clinical parameters and the application of machine learning models to differentiate between DKD and NDKD.

# Materials and methods

## Study population

The data originated from Chongqing Medical University's Medical Data Intelligence Platform(Yidu-Cloud (Beijing) Technology Co, Ltd, China), which consisted of the data from seven institutions and over 40 million electronic medical records (during 1 January 2010 to 31 May 2020) (21–23). Only the information from the first hospitalization was applied in the

event of subsequent hospitalizations. Xiaozhu Liu (Account Number: cy2014223346)and Minjie Duan (Account Number: MI2020111943) were permitted to access data directly on the platform system where all information is anonymous and has a unique identifying code to preserve privacy, while an informed consent from the patient is unnecessary. The local institutional ethics committee gave us their authorization.The inclusion and exclusion criteria below were employed for screening:

Inclusion criteria: (1) hospitalization for T2DM or T2DKD; (2) following the WHO 1999 diagnostic criteria for diabetes mellitus; (3) age >18 years; following the diagnostic criteria of the KDOQI US commentary on the 2012 KDIGO clinical practice guideline for Chronic kidney disease (24).

Exclusion criteria: (1) combination of other possible complications such as urinary tract infection, malignancy; (2) immune diseases (e.g., systemic lupus erythematosus and vasculitis); (3) Other endocrine diseases; (4) type 1 diabetes, gestational diabetes and other diabetes with unclear classification; (5) hospitalization days <1; (6) discharge against medical advice; (7) patients lost to follow-up or death during index hospitalization; and (8) patients with >25% of data missing.

In accordance with the inclusion and exclusion criteria, 3624 patients with T2DM were recruited, which consisted of 1856 patients with T2DKD (Figure 1). In this study, DKD was defined in accordance with the National Health Insurance Administration's definition of catastrophic illness ICD-9 and ICD-10 codes for DKD.

The definition of CKD in the 2012 KDIGO clinical practice guideline was adopted (24, 25).

## Data collection and data preprocessing

The latest literature on DKD was reviewed and combined with clinical experience to acquire relevant clinical data and laboratory characteristics (25–28). 53 clinical characteristics with missing values ≤ 25% were covered. Since most models cannot analyze data with missing values, Multivariate Imputation by Chained Equations (MICE) algorithm was used for data filling.

Baseline data were compared in patients with DKD and T2DKD from the first examination and test results after admission (Table S1)

## Model development and performance evaluation

The data set was randomly assigned to a training set (80%) and a validation set (20%) based on stratified random sampling. Our models were developed using the training set, and their performance was assessed using the validation set. The least absolute shrinkage and selection operator (LASSO) was employed for selecting the risk predictors to eliminate unnecessary and redundant information and increase the model's discriminative capacity. Finally, non-zero regression coefficient variables were selected to build the prediction models.

To select the prospective algorithms for our prediction models, we first analyzed the performance of 15 machine learning algorithms without hyper-parameters tuning. After that, an algorithm with the optimal performance was selected in accordance with the model's accuracy and the area under the receiver operating characteristic curve (AUC). PyCaret (version 2.3.3), an open source, low-code machine learning library in Python, was employed to perform the screening procedure. Second, the Bayesian Optimization approach with 10-fold cross validation was adopted for adjusting a prediction model based on the training set to find the ideal hyper-parameter configuration. The above algorithm is an efficient constrained global optimization tool, which was performed based on the functions of the bayes_opt Python package (version 1.2.0). AUC, accuracy and sensitivity were obtained to assess the models performance based on the independent validation set.

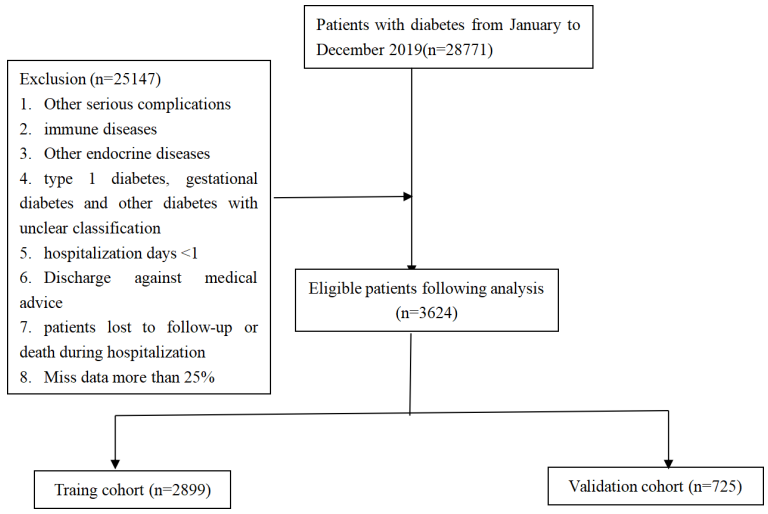



FIGURE 1
The flowchart of the study.

To reduce the black-box nature of machine learning and to allow clinicians to understand the results of the provided model, SHapley Additive exPlanations (SHAP) was adopted to interpret the model with the use of SHAP python package (version 0.39.0). The significance of input features was obtained with the use of a game-theoretic algorithm based on the independent validation set. It is noteworthy that all 38 variables would not always be available in clinical practice. Accordingly, the top 12 were taken from SHAP summary plots to build the simpler model, and the discriminative power was compared between the full model and simpler models.

## Statistical analysis

For baseline comparison of data sets, categorical variables were denoted as percentages, and Chi-square test or Fisher's exact test was performed for comparison between groups. Continuous variables were examined for normality using Kolmogorov-Smirnov test, and measures following normal distribution were denoted as mean ± standard deviation, and Student's t-test was used for comparison between groups, and measures not following normal distribution were denoted as median (interquartile range), and Mann-Whitney U rank sum test was performed for comparison between groups. R (version 4.0.2) was adopted for statistical analysis. A two-sided $P < 0.05$ was considered to achieve statistical significance.

# Results

## Patient characteristics

The data were assigned to a training set and a validation set at 8:2. The training set consisted of 2899 cases, including 1485 cases of T2DKD (51.2%) and 1414 cases of T2DM (48.8%); the validation set consisted of 725 cases, including 371 cases of T2DKD (51.2%) and 354 cases of T2DM (48.8%) (see Table S2 for details).

## Feature selection

Least absolute shrinkage and selection operator(LASSO) was employed to select the most significant features, so as to classify individuals diagnosed DKD. All features (a total of 53 variables) were included in the LASSO regression analysis and narrowed down to 38 features with non-zero β coefficients in the LASSO regression model. The above features were Sex, Smoke, Drink history, Age, length of stay (LOS), Systolic blood pressure (SBP), diastolic blood pressure (DBP), total protein (TP), albumin (ALB), gamma glutamyltransferase (GGT), alanine aminotransferase (ALT), alkaline phosphatase (ALP), total cholesterol (TC), triglyceride (TG), high-density lipoprotein cholesterol (HDL-C), phosphorus (P), glucose (GLU), apolipoprotein A1 (ApoA1), Hemoglobin A1C (HbA1C), creatinine (CREA), urea, uric acid (UA), fibrinogen (FIB-C), platelet count (PLT), prothrombin time (PT), thrombin time (TT), monocyte percentage (Mon%), basophil count (Bas), eosinophil count (Eos), neutrophil count (Neu), platelet large cell ratio (P-LCR), mean corpuscular volume (MCV), mean corpuscular hemoglobin concentration (MCHC), lymphocyte count (Lym), red blood cell distribution width-standard deviation (RDW-SD), hematocrit (HCT), platelet distribution width (PDW), mean platelet volume (MPV) (Figure 2A).

## Performance of models in predicting DKD

Figure 3 lists the predictive performance of 15 ML models after 10-fold cross validation for internal training. Almost all classic ML methods capable of conducting classification analysis were considered. The top six models consisted of CatBoost Classifier, Light Gradient Boosting Machine, extreme gradient Boosting, Extra Trees Classifier, Gradient Boosting Classifier, Random Forest Classifier, with AUC over 0.8. As revealed by the results, the CatBoost model indicated the maximum performance in predicting DKD risk with AUC and accuracy of 0.840 and 0.755, respectively. As a result, the CatBoost model was selected and optimized in the following step.

Bayesian optimization algorithm with 10-fold cross validation to select the optimal hyperparameter configuration for the CatBoost model. The optimized CatBoost model exhibited the optimal and the most stable performance, with an AUC of 0.861, an accuracy of 0.777, a sensitivity of 0.755 (Figure 4). To increase the transparency and usability in real clinical setting of the prediction model, 12 top features were selected to construct the simpler prediction model based on the SHAP values and clinical availability. The top 12 most significant features consisted of SBP, CREA, LOS, TT, Age, PT, PLCR, ALB, GLU, FIBC, RDWSD, HbA1c (Figure 2C). As depicted in Figure 4, the simpler CatBoost model showed slight worse performance (AUC: 0.840). In this study, our CatBoost model was illustrated using the SHAP method by Lundberg and Lee (29). We employ the Shap technique to gain a global interpretation of our reserved cohort as well as individual patient interpretations. The SHAP summary plots for the top 38 clinical characteristics contributing to our ML model's prediction of DKD development in our research are shown in Figures 2A, B. The SHAP summary plots for the top 12 clinical characteristics contributing to our ML model's prediction of developing DKD in our research are shown in Figures 2C, D.

Meanwhile, we show the SHAP explanation force diagram for two patients from the CatBoost model's validation set (Figure 5). Figure 5A depicts a patient who is 48 years old. This patient's anticipated risk of having DKD is significant, at 160 percent, in comparison with a baseline risk of 10%. (average prevalence of the validation cohort). Lower ALB of 29.5g/l, increased HbA1C of 15.1 percent, increased RDWSD of 52.7 mg/dl, prolonged LOS of 16 days, lower PLCR of 22.9 percent, and PT of 11.1 seconds were the characteristics found by the model for the prediction of a greater prevalence in this patient. The patient's age of 48 years and TT of 18.8 seconds help to mitigate the increased risk. Figure 5B presents another T2DM patient. This patient's anticipated risk of getting DKD was -146 percent, in comparison with a baseline risk of 10%. (average prevalence of the validation cohort). Normal SBP of

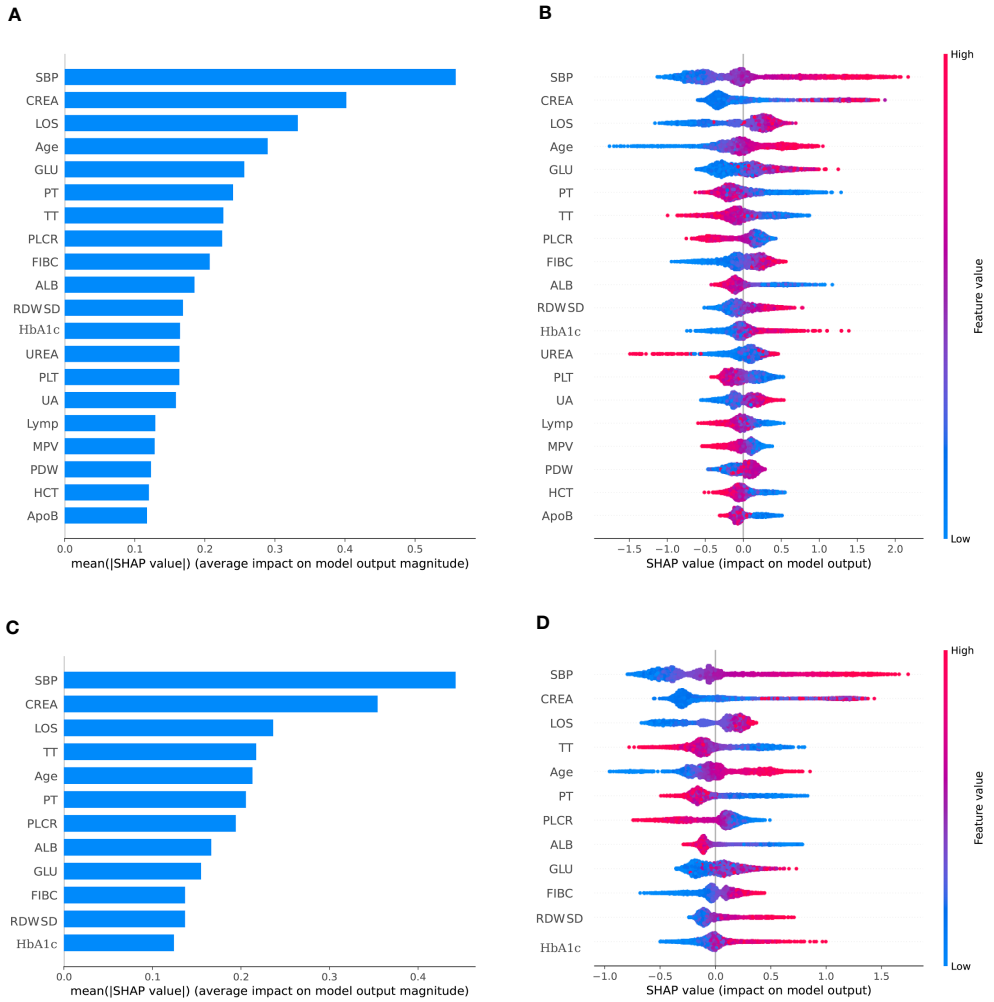
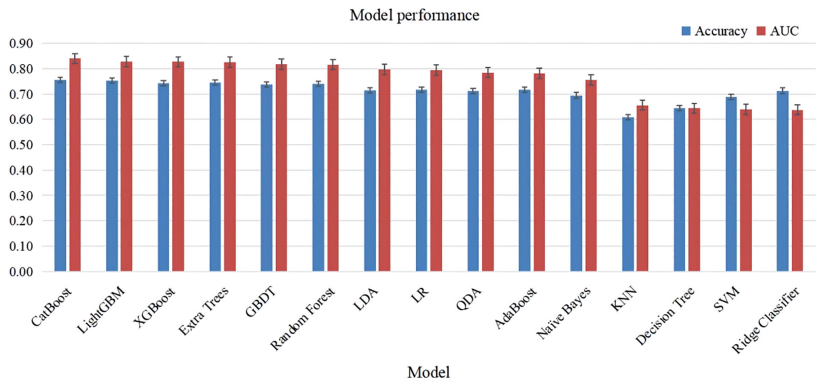



FIGURE 2
The SHAP summary plots for the CatBoost model. **(A)** depicts the most 38 effective characteristics on prediction (ranked from showing the highest to lowest importance). **(B)** depicts the distribution of the effects of 38 characteristics on the output of the model. **(C)** depicts the most 12 effective characteristics on prediction (ranked from showing the highest to lowest importance). **(D)** depicts the distribution of the effects of 12 characteristics on the output of the model.For numeric characteristics, the colors indicate the feature values: red for larger values and blue for smaller values. The thickness of the line is defined by the number of instances at a specific value, and it is made up of individual dots representing each DKD (e.g., most patients have a low risk of developing of DKD). A lower likelihood is indicated by a negative SHAP value (stretching to the left), while a higher probability is indicated by a positive SHAP value (reaching to the right). The gray dots reflect particular possible values for non-numeric characteristics such as main diagnosis, with select diagnoses considerably raising or decreasing the model's output, while the majority of diagnoses have just a little influence on prediction.

120mmHg, shorter LOS of 3 days, normal TT of 18.53 seconds, normal FIBC of 2.06, normal CREA of 42.6 umol/l, and normal RDWSD of 40.1 mg/dl were the parameters found using the model for the inhibition of DKD development. The lower risk was somewhat countered by a 12.8 percent HbA1C and a 22.9 percent PLCR.

## Discussion

T2DKD has been recognized as the major cause of end-stage renal failure (4). Its diagnosis is largely dependent on kidney biopsy, which is generally used to distinguish diabetic kidney disease from other kidney diseases. However, biopsy cannot be employed for early screening and diagnosis of T2DKD, thus resulting in missed diagnosis and misdiagnosis. The development of chronic albuminuria followed by a steady drop in GFR (classical phenotype of DKD) (24) has been adopted to diagnose DKD. Several studies have indicated the trajectories of renal function (i.e., changes in GFR and albuminuria with time) that diverge from this traditional phenotype over the past decade (30). Three non-classical DKD phenotypes have been identified, each of which are defined by albuminuria regression, fast GFR decrease, or the lack of proteinuria or albuminuria, respectively (31). Albuminuria has limitations in the prediction of the progression of DKD. The determination of albuminuria values is affected by a wide variety of factors (e.g., fever, strenuous exercise within 24h, menstruation, hyperglycemia and hypertension). For atypical DKD, albuminuria is not sufficiently specific for the diagnosis of DKD, and some studies have indicated that 30% of patients with albuminuria had




| Model | Accuracy | AUC |
|---|---|---|
| CatBoost Classifier | 0.755 | 0.840 |
| Light Gradient Boosting Machine | 0.754 | 0.828 |
| Extreme Gradient Boosting | 0.743 | 0.827 |
| Extra Trees Classifier | 0.745 | 0.825 |
| Gradient Boosting Classifier | 0.737 | 0.818 |
| Random Forest Classifier | 0.741 | 0.816 |
| Linear Discriminant Analysis | 0.714 | 0.797 |
| Logistic Regression | 0.717 | 0.795 |
| Quadratic Discriminant Analysis | 0.711 | 0.785 |
| Ada Boost Classifier | 0.717 | 0.782 |
| Naïve Bayes | 0.695 | 0.756 |
| K Neighbors Classifier | 0.609 | 0.656 |
| Decision Tree Classifier | 0.645 | 0.644 |
| SVM-Linear Kernel | 0.688 | 0.640 |
| Ridge Classifier | 0.713 | 0.638 |

Models are ordered according to their AUC.AUC area under receiver operating characteristic curve; CatBoost, Categorical Boosting; SVM, support vector machine.

FIGURE 3
Performance of different models in internal validation.

negative urine albumin within 10 years, and this phenomenon has been more significant in type 2 diabetes patients (32, 33). Urinary albumin excretion was influenced by many factors (34). It was recommended that the diagnosis of albuminuria requires three 24-h urine collections over a 3-month period, with at least two of the three results exceeding the threshold and not measured by urinary routine. Thus, the diagnosis of albuminuria as a basis for DKD should be based on a combination of multiple tests and long-term follow-up with glomerular filtration rate, and the cause of albuminuria should be excluded. Thus, the necessity of a simple and convenient clinical tool to assess DKD in daily clinical practice is highlighted.

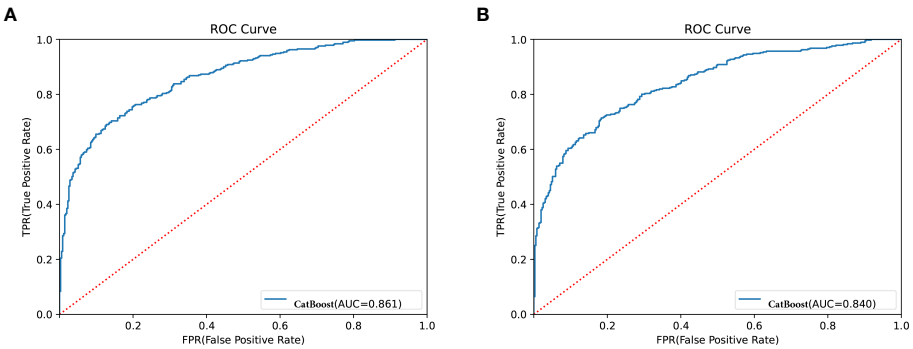



FIGURE 4
Receiver operator characteristic (ROC) curves for the CatBoost model. **(A)** Shows ROC for CatBoost with most 38 effective characteristics on prediction (ranked from most to least important). **(B)** Shows ROC for CatBoost with most 12 effective characteristics on prediction (ranked from most to least important).

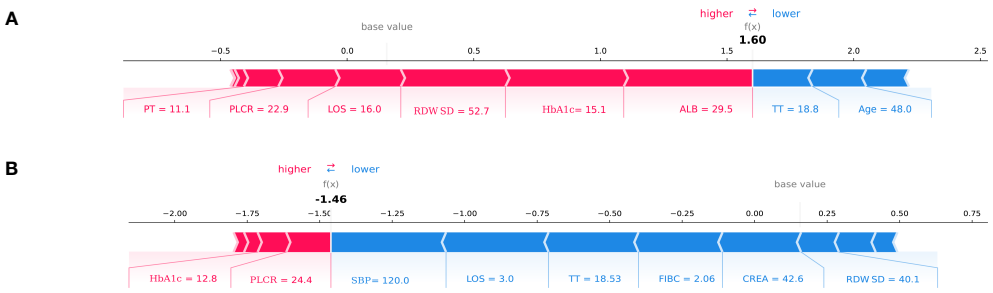



FIGURE 5
SHAP force plot for two patients of the held-out validation set. **(A)** patient at high risk of developing T2DKD; **(B)** patient at low risk of developing T2DKD. DKD, diabetic kidney disease; ALB, albumin; HbA1C, hemoglobin A1C; RDWSD, red blood cell distribution width-standard deviation; LOS, length of stay; PLCR, platelet large cell ratio; PT, prothrombin time; TT, thrombin time; SBP, Systolic blood pressure; FIBC,fibrinogen; CREA, creatine.

Accordingly, a multi-center retrospective study was conducted to analyze clinical indicators of T2DM and T2DKD based on real-world data, and adopted machine learning algorithms to investigate potential clinical and Laboratory risk factors for DKD among patients with T2DM. In this study, 15 ML models for ancillary diagnosis of T2DKD were initially developed in accordance with the clinical data from seven hospitals in China, and the efficacy of the 15 ML models was assessed. Meanwhile, we tried the CatBoost algorithms, which are seldom employed in medical studies. Our retrospective study showed that CatBoost is very effective for ancillary diagnosis of DKD. The patients' clinical and laboratory parameters were assessed with a CatBoost model, and key features correlated with an increased risk of DKD, (e.g., SBP, CREA, LOS, TT, Age, PT, PLCR, ALB, GLU, FIBC, RDWSD, as well as HbA1c) were identified.

In this study, the differential diagnosis model of T2DKD was built based on 15 machine learning algorithms, thus solving the nonlinear relationship between clinical features and diagnosis results. The CatBoost model with the highest diagnostic accuracy than the other 14 models, such as light gradient booting model, Extreme Gradient Boosting and so on, indicating a good predictive performance. With LASSO analysis, SBP, CREA, LOS, TT, Age, PT, PLCR, ALB, GLU, FIBC, RDWSD, HbA1c were the top 12 major influencing factors of the index importance, which achieved statistical significance in multivariate logistic regression analysis.

Existing studies suggested that SBP, CREA, Age, ALB, and GLU are factors for DKD. High SBP was reported with rapidly eGFR decline in the Atherosclerosis Risk in Communities (ARIC) study (35). As reported by Gross JL et al., hypertension increased the morbidity of patients hospitalized with kidney disease, and increased blood pressure was found as a major risk factor for DKD (36). Sasso FC et al. found in their study that arterial pressure is a relevant factor for the progression of DKD in patients with DM, accompanied by hypertension is highly susceptible to periglomerular microvascular changes leading to development of DKD (37). Viazzi F et al. investigated the clinical records of more than 30,000 patients with T2DM combined with hypertension over 4 years of follow-up. It was found that elevated long-term blood pressure variability predicted CKD in T2DM and (38). In the model built in this study, SBP was the primary predictor of DKD, consistent with previous studies mentioned above. As revealed by the analysis of the examination of renal function in patients with DKD hospitalized between 2015 and 2017, CREA achieved a high predictive value in the diagnosis of patients with DKD and could effectively assess the status of renal function in patients with DM (39). This is consistent with the findings of our study.

Radcliffe NJ et al. found a correlation between elevated age, early GFR decline and DKD progression, consistent with the results of the present study (40). Elley et al. demonstrated an independent relationship between higher age and increased risk of DKD progression (28). López-Revuelta K et al. also suggested that age could be a risk factor for DKD development, with a mean age of 58.3 years in terms of DKD (41). The above studies assessed changes in GFR in predominantly adult patients (28). Several cross-sectional studies have shown changes in P-LCR, PLT, and FIBC in DKD patients in comparison with normal, suggesting that the occurrence of DKD is closely related to abnormal platelet function (42–44).

The study by Zoppini G et al. followed more than 1,000 patients with DKD and found that HbA1c was a risk factor for the progression of DKD. A decrease in HbA1c significantly reduced the risk of complications in patients with DM. A decrease in Hb A1c from 10% to 9% was also found to have a greater impact on reducing the risk of complications than a decrease in Hb A1c from 7% to 6% (45). Yun KJ, et al. found HbA1c variability may affect the development and progression of DKD in their study (46). Visit-to-visit variability of HbA1c was an independent risk factor of microalbuminuria in association with oxidative stress among type 2 diabetes mellitus patients (47, 48). Meanwhile, observational studies have not consistently demonstrated a glucose threshold (49). In a referred population of established DKD, higher HbA1c was not associated with higher risk of ESKD or death (50). In addition, our study found that HbA1c also influences the progression of DKD, in agreement with most previous studies.In addition, our study found that LOS, TT, PT, RDWSD also influence of DKD progression, which has not been reported in the literature and deserves further study.

Previously, it was confirmed that metabolic syndrome(MetS) and associated components (abdominal obesity, elevated BG, elevated BP and lipid metabolic disorder) are strongly related to

CKD and a decreased estimated glomerular filtration rate (eGFR) (51–55). Over the 4-year follow-up period, Peijia L et al. found that MetS recovery was associated with a reduced risk of rapid eGFR decline in middle-aged and older adults, while MetS occurrence was not related to rapid eGFR decline. Recovery from MetS appeared to protect against a rapid decline in eGFR (56).

Due to the strong interpretability, logistic regression model is widely used to explore the risk factors of diseases. However, problems such as under-fitting, data missing, poor overall performance of the model are likely to occur in the process of modeling. In terms of the machine learning algorithms, this study has been considered the first to assess the risk of patients with DKD using the CatBoost model. As revealed by this study, the CatBoost model achieved great performance in the prediction of DKD. By analyzing clinical indicators of 1768 cases of type 2 diabetes and 1856 cases of type 2 diabetic kidney disease, this study applied the CatBoost model to the risk assessment of type 2 diabetic kidney disease for the first time, and analyzed the weight relationship of influencing factors. A good classification results was obtained (AUC=0.840).

This study had several limitations. First, although general clinical data and laboratory indexes were collected more comprehensively, some of the indexes were not covered in the model due to the missing values of ≥25%, and the significance of the correlation with type 2 diabetic kidney disease should be investigated in more detail when the volume of data is expanded.However, it was found through our data that some clinical parameters (cystatin-C, total 24-hour urine protein, duration of disease, etc.) are missing in many patients and many indicators cannot be generalized in primary care. Second, We found in the construction of the model that SBP was included as an important parameter in the prediction model, considering hypertension as an important confounding factor that is best analyzed in a stratified manner. Third, Due to data limitations, we were unable to select patients with a first diagnosis of T2DKD to model.Fourth, a cross-sectional study was conducted, and the results should be validated through a prospective study.

## Conclusions

To sum up, this retrospective study suggested that CatBoost could be highly effective in the early ancillary diagnosis of DKD. The importance of the model's correlation to type 2 diabetic kidney disease should be investigated in depth after the data volume is expanded. In subsequent research, a greater amount of data and more machine learning models will be adopted for modeling research, as an attempt to build a better risk assessment model.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the Chongqing Medical University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

XL, TZ, MD and HW conceived and designed the study. All authors contributed to the acquisition of data or analysis and interpretation of data. XL drafted the manuscript. MD drew the figures and tables.HH, YZ, TX, WN, and BZ revised the manuscript critically for essential intellectual content. All authors read and approved the final version to be published.

## Funding

This work was supported by the National Natural Science Foundation of China under Grant 72101040 and Intelligent Medicine Research Project of Chongqing Medical University under Grant ZHYXQNRC202201.

## Acknowledgments

We acknowledge Yidu Cloud Technology Inc.'s participation in this project.

## Conflict of interest

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.

## Publisher's note



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1184190/full#supplementary-material

# References


1. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the international diabetes federation diabetes atlas, 9th edition. *Diabetes Res Clin Pract* (2019) 157:107843. doi: 10.1016/j.diabres.2019.107843

2. An Y, Zhang P, Wang J, Gong Q, Gregg EW, Yang W, et al. Cardiovascular and all-cause mortality over a 23-year period among Chinese with newly diagnosed diabetes in the da Qing IGT and diabetes study. *Diabetes Care* (2015) 38(7):1365–71. doi: 10.2337/dc14-2498

3. Hu H, Sawhney M, Shi L, Duan S, Yu Y, Wu Z, et al. A systematic review of the direct economic burden of type 2 diabetes in china. *Diabetes therapy: research Treat Educ Diabetes Related Disord* (2015) 6(1):7–16. doi: 10.1007/s13300-015-0096-0

4. Cho NH, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, et al. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract* (2018) 138:271–81. doi: 10.1016/j.diabres.2021.109119

5. Zhang L, Wang F, Wang L, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, et al. Prevalence of chronic kidney disease in China: a cross-sectional survey. *Lancet (London England)* (2012) 379(9818):815–22. doi: 10.1016/S0140-6736(12)60033-6

6. Colhoun HM, Marcovecchio ML. Biomarkers of diabetic kidney disease. *Diabetologia* (2018) 61(5):996–1011. doi: 10.1007/s00125-018-4567-5

7. de Galan BE, Perkovic V, Ninomiya T, Pillai A, Patel A, Cass A, et al. Lowering blood pressure reduces renal events in type 2 diabetes. *J Am Soc Nephrol* (2009) 20 (4):883–92. doi: 10.1681/ASN.2008070667

8. de Zeeuw D, Remuzzi G, Parving HH, Keane WF, Zhang Z, Shahinfar S, et al. Proteinuria, a target for renoprotection in patients with type 2 diabetic nephropathy: lessons from RENAAL. *Kidney Int* (2004) 65(6):2309–20. doi: 10.1111/j.1523-1755.2004.00653.x

9. Araki S, Haneda M, Koya D, Hidaka H, Sugimoto T, Isono M, et al. Reduction in microalbuminuria as an integrated indicator for renal and cardiovascular risk reduction in patients with type 2 diabetes. *Diabetes* (2007) 56(6):1727–30. doi: 10.2337/db06-1646

10. Yokoyama H, Araki S, Haneda M, Matsushima M, Kawai K, Hirao K, et al. Chronic kidney disease categories and renal-cardiovascular outcomes in type 2 diabetes without prevalent cardiovascular disease: a prospective cohort study (JDDM25). *Diabetologia* (2012) 55(7):1911–8. doi: 10.1007/s00125-012-2536-y

11. Yokoyama H, Araki S, Honjo J, Matsushima M, Kawai K, Hirao K, et al. Association between remission of macroalbuminuria and preservation of renal function in patients with type 2 diabetes with overt proteinuria. *Diabetes Care* (2013) 36(10):3227–33. doi: 10.2337/dc13-0281

12. Mogensen CE. Microalbuminuria as a predictor of clinical diabetic nephropathy. *Kidney Int* (1987) 31(2):673–89. doi: 10.1038/ki.1987.50

13. Lan K, Wang DT, Fong S, Liu LS, Wong KKL, Dey N. A survey of data mining and deep learning in bioinformatics. *J Med Syst* (2018) 42(8):139. doi: 10.1007/s10916-018-1003-9

14. Allen A, Iqbal Z, Green-Saxena A, Hurtado M, Hoffman J, Mao Q, et al. Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus. *BMJ Open Diabetes Res Care* (2022) 10(1): e002560. doi: 10.1136/bmjdrc-2021-002560

15. David SK, Rafiullah M, Siddiqui K. Comparison of different machine learning techniques to predict diabetic kidney disease. *J Healthcare Eng* (2022) 2022:7378307. doi: 10.1155/2022/7378307

16. Makino M, Yoshimoto R, Ono M, Itoko T, Katsuki T, Koseki A, et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Sci Rep* (2019) 9(1):11862. doi: 10.1038/s41598-019-48263-5

17. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. *Health Inf Sci Syst* (2022) 8(1):7. doi: 10.1007/s13755-019-0095-z

18. Chan L, Nadkarni GN, Fleming F, McCullough JR, Connolly P, Mosoyan G, et al. Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease. *Diabetologia Vol* (2021) 64(7):1504–15. doi: 10.1007/s00125-021-05444-0

19. Viana LV, Gross JL, Camargo JL, Zelmanovitz T, da Costa Rocha EP, Azevedo MJ. Prediction of cardiovascular events, diabetic nephropathy, and mortality by albumin concentration in a spot urine sample in patients with type 2 diabetes. *J Diabetes its Complications Vol* (2012) 26(5):407–12. doi: 10.1016/j.jdiacomp.2012.04.014

20. Park SB, Kim SS, Kim IJ, Nam YJ, Ahn KH, Kim JH, et al. Variability in glycated albumin levels predicts the progression of diabetic nephropathy. *J Diabetes its Complications Vol* (2017) 31(6):1041–6. doi: 10.1016/j.jdiacomp.2017.01.014

21. Xu Q, Peng Y, Tan J, Zhao W, Yang M, Tian J. Prediction of atrial fibrillation in hospitalized elderly patients with coronary heart disease and type 2 diabetes mellitus using machine learning: a multicenter retrospective study. *Front Public Health* (2022) 10:842104. doi: 10.3389/fpubh.2022.842104

22. Tan J, Tang X, He Y, Xu X, Qiu D, Chen J, et al. In-patient expenditure between 2012 and 2020 concerning patients with liver cirrhosis in chongqing: a hospital-based multicenter retrospective study. *Front Public Health* (2022) 10:780704. doi: 10.3389/fpubh.2022.780704

23. Xu X, Wang H, Zhao W, Wang Y, Wang J, Qin B. Recompensation factors for patients with decompensated cirrhosis: a multicentre retrospective case-control study. *BMJ Open* (2021) 11(6):e043083. doi: 10.1136/bmjopen-2020-043083

24. National Kidney Foundation. KDOQI clinical practice guideline for diabetes and CKD: 2012 update. *Am J Kidney Dis* (2012) 60(5):850–86. doi: 10.1053/j.ajkd.2012.07.005

25. KDOQI. KDOQI clinical practice guidelines and clinical practice recommendations for diabetes and chronic kidney disease. *Am J Kidney Dis* (2007) 492 Suppl 2:S12–154. doi: 10.1053/j.ajkd.2006.12.005

26. Song X, Waitman LR, Yu AS, Robbins DC, Hu Y, Liu M. lLongitudinal risk prediction of chronic kidney disease in diabetic patients using a temporal-enhanced gradient boosting machine: retrospective cohort study. *JMIR Med Inf* (2020) 8(1): e15510. doi: 10.2196/15510

27. Macisaac RJ, Ekinci EI, Jerums G. Markers of and risk factors for the development and progression of diabetic kidney disease. *Am J Kidney Dis* (2014) 63 (2 Suppl 2):S39–62. doi: 10.1053/j.ajkd.2013.10.048

28. Elley CR, Robinson T, Moyes SA, Kenealy T, Collins J, Robinson E, et al. Derivation and validation of a renal risk score for people with type 2 diabetes. *Diabetes Care vol* (2013) 36(10):3113–20. doi: 10.2337/dc13-0190

29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Proc Adv Neural Inf Process Syst* (2017), 4768–77.

30. Afkarian M, Zelnick LR, Hall YN, Heagerty PJ, Tuttle K, Weiss NS, et al. Clinical manifestations of kidney disease among US adults with diabetes, 1988-2014. *JAMA vol* (2016) 316(6):602–10. doi: 10.1001/jama.2016.10924

31. Oshima M, Shimizu M, Yamanouchi M, Toyama T, Hara A, Furuichi K, et al. Trajectories of kidney function in diabetes: a clinicopathological update nature reviews. *Nephrol vol* (2021) 17(11):740–50. doi: 10.1038/s41581-021-00462-y

32. American Diabetes association. 16. diabetes advocacy: standards of medical care in diabetes-2019. *Diabetes Care* (2019) 42(Suppl 1):S182–3. doi: 10.2337/dc19-S016

33. Ekinci EI, Jerums G, Skene A, Crammer P, Power D, Cheong KY, et al. Renal structure in normoalbuminuric and albuminuric patients with type 2 diabetes and impaired renal function. *Diabetes Care* (2013) 36(11):3620–6. doi: 10.2337/dc12-2572

34. Tuttle KR, Bakris GL, Bilous RW, Chiang JL, de Boer IH, Goldstein-Fuchs J, et al. Diabetic kidney disease: a report from an ADA consensus conference. *Am J Kidney Dis* (2014) 64(4):510–33. doi: 10.1053/j.ajkd.2014.08.001

35. Warren B, Rebholz CM, Sang Y, Lee AK, Coresh J, Selvin E, et al. Diabetes and trajectories of estimated glomerular filtration rate: a prospective cohort analysis of the atherosclerosis risk in communities study. *Diabetes Care vol* (2018) 41:8. doi: 10.2337/dc18-0277

36. Gross JL, de Azevedo MJ, Silveiro SP, Jorge L , Canani LH, Caramori ML, Zelmanovitz T. Diabetic nephropathy: diagnosis, prevention, and treatment. *Diabetes Care vol* (2005) 28:1. doi: 10.2337/diacare.28.1.164

37. Sasso FC, De Nicola L, Carbonara O, Nasti R, Minutolo R, Salvatore T, et al. Cardiovascular risk factors and disease management in type 2 diabetic patients with diabetic nephropathy. *Diabetes Care* (2006) 29(3):498–503. doi: 10.2337/diacare.29.03.06.dc05-1776

38. Viazzi F, Bonino B, Mirijello A, Fioretto P, Giorda C, Ceriello A, et al. Long-term blood pressure variability and development of chronic kidney disease in type 2 diabetes. *J hypertension* (2019) 37(4):805–13. doi: 10.1097/HJH.0000000000001950

39. Rigalleau V, Lasseur C, Perlemoine C, Barthe N, Raffaitin C, Chauveau P, et al. Cockcroft-gault formula is biased by body weight in diabetic patients with renal impairment. *Metabolism: Clin Exp* (2006) 55(1):108–12. doi: 10.1016/j.metabol.2005.07.014

40. Radcliffe NJ, Seah JM, Clarke M, MacIsaac RJ, Jerums G, Ekinci EI. Clinical predictive factors in diabetic kidney disease progression. *J Diabetes Invest* (2017) 8 (1):6–18. doi: 10.1111/jdi.12533

41. López-Revuelta K, Galdo PP, Stanescu R, Parejo L, Guerrero C, Pérez-Fernández E. Silent diabetic nephropathy. *World J Nephrol* (2014) 3(1):6–15. doi: 10.5527/wjn.v3.i1.6

42. Doshi SM, Friedman AN. Diagnosis and management of type 2 diabetic kidney disease. *Clin J Am Soc Nephrol* (2017) 12(8):1366–73. doi: 10.2215/CJN.11111016

43. American Diabetes Association. Standards of medical care in diabetes-2016 abridged for primary care providers. *Clin Diabetes* (2016) 34(1):3–21. doi: 10.2337/diaclin.34.1.3

44. Yu M, Xie R, Zhang Y, Liang H, Hou L, Yu C, et al. Phosphatidylserine on microparticles and associated cells contributes to the hypercoagulable state in diabetic kidney disease. *Nephrology dialysis Transplant* (2018) 33(12):2115–27. doi: 10.1093/ndt/gfy027

45. Zoppini G, Targher G, Chonchol M, Ortalda V, Negri C, Stoico V, et al. Predictors of estimated GFR decline in patients with type 2 diabetes and preserved kidney function. *Clin J Am Soc Nephrol* (2012) 7(3):401–8. doi: 10.2215/CJN.07650711

46. Yun KJ, Kim HJ, Kim MK, Kwon HS, Baek KH, Roh YJ, et al. Risk factors for the development and progression of diabetic kidney disease in patients with type 2 diabetes mellitus and advanced diabetic retinopathy. *Diabetes Metab J* (2016) 40(6):473–81. doi: 10.4093/dmj.2016.40.6.473

47. Yan Y, Kondo N, Oniki K, Watanabe H, Imafuku T, Sakamoto Y, et al. Predictive ability of visit-to-Visit variability of HbA1c measurements for the development of diabetic kidney disease: a retrospective longitudinal observational study. *J Diabetes Res* (2022) 2022:6934188. doi: 10.1155/2022/6934188

48. Ceriello A, De Cosmo S, Rossi MC, Lucisano G, Genovese S, Pontremoli R, et al. Variability in HbA1c, blood pressure, lipid parameters and serum uric acid, and risk of development of chronic kidney disease in type 2 diabetes. *Diabetes Obes Metab* (2017) 19(11):1570–8. doi: 10.1111/dom.12976

49. MacIsaac RJ, Jerums G, Ekinci EI. Effects of glycaemic management on diabetic kidney disease. *World J Diabetes* (2017) 8(5):172–86. doi: 10.4239/wjd.v8.i5.172

50. Limkunakul C, de Boer IH, Kestenbaum BR, Himmelfarb J, Ikizler TA, Robinson-Cohen C. The association of glycated hemoglobin with mortality and ESKD among persons with diabetes and chronic kidney disease. *J Diabetes its Complications* (2019) 33(4):296–301. doi: 10.1016/j.jdiacomp.2018.12.010

51. Xie K, Bao L, Jiang X, Ye Z, Bing J, Dong Y, et al. The association of metabolic syndrome components and chronic kidney disease in patients with hypertension. *Lipids Health Dis* (2019) 18(1):229. doi: 10.1186/s12944-019-1121-5

52. Viazzi F, Piscitelli P, Giorda C, Ceriello A, Genovese S, Russo G, et al. Metabolic syndrome, serum uric acid and renal risk in patients with T2D. *PloS One* (2017) 12(4): e0176058. doi: 10.1371/journal.pone.0176058

53. Chen J, Kong X, Jia X, Li W, Wang Z, Cui M, et al. Association between metabolic syndrome and chronic kidney disease in a Chinese urban population. *Clinica chimica acta; Int J Clin Chem* (2017) 470:103–8. doi: 10.1016/j.cca.2017.05.012

54. Thomas G, Sehgal AR, Kashyap SR, Srinivas TR, Kirwan JP, Navaneethan SD. Metabolic syndrome and kidney disease: a systematic review and meta-analysis. *Clin J Am Soc Nephrol* (2011) 6(10):2364–73. doi: 10.2215/CJN.02180311

55. Chang IH, Han JH, Myung SC, Kwak KW, Kim TH, Park SW, et al. Association between metabolic syndrome and chronic kidney disease in the Korean population. *Nephrol (Carlton Vic.)* (2009) 14(3):321–6. doi: 10.1111/j.1440-1797.2009.01091.x

56. Liu P, Tang L, Fang J, Chen C, Liu X. Association between recovery/occurrence of metabolic syndrome and rapid estimated glomerular filtration rate decline in middle-aged and older populations: evidence from the China health and retirement longitudinal study. *BMJ Open* (2022) 12(10):e059504. doi: 10.1136/bmjopen-2021-059504

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 18 July 2023
DOI 10.3389/fendo.2023.1200002



OPEN ACCESS

EDITED BY
Qiuming Yao,
Fudan University, China

REVIEWED BY
Chenggang Li,
Nankai University, China
Huiyuan Pang,
Renmin Hospital of Wuhan University, China

*CORRESPONDENCE
Zhijian Wang
wzjnfyy@163.com

RECEIVED 04 April 2023
ACCEPTED 20 June 2023
PUBLISHED 18 July 2023

CITATION
Tong Y, Sun Q, Shao X and Wang Z (2023) Effect of vaginal microbiota on pregnancy outcomes of women from Northern China who conceived after IVF. *Front. Endocrinol.* 14:1200002. doi: 10.3389/fendo.2023.1200002



# Effect of vaginal microbiota on pregnancy outcomes of women from Northern China who conceived after IVF

Yu Tong[1,2], Qiang Sun[2], Xiaoguang Shao[2] and Zhijian Wang[1*]

[1]Department of Obstetrics and Gynecology, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, China, [2]Department of Obstetrics and Gynecology, Dalian Municipal Women and Children's Medical Group, Dalian, Liaoning, China

**Objective:** This study aimed to investigate the correlation between vaginal microbiota and pregnancy outcomes of women who achieved pregnancy *via in vitro* fertilization (IVF) in Northern China, and to determine a biomarker for evaluation of the risk of preterm births in these women.

**Methods:** In total, 19 women from Northern China women who conceived after IVF and 6 women who conceived naturally were recruited in this study. The vaginal samples of the healthy participants were collected throughout pregnancy, that is, during the first, second, and third trimesters. The V3–V4 region of 16S rRNA was used to analyze the vaginal microbiome, and the bioinformatic analysis was performed using QIIME Alpha and Beta diversity analysis.

**Results:** Either IVF group or Natural conception group, bacterial community diversities and total species number of vagnal samples from who delivered at term were significantly higher than those who delivered before term. Low abundance of vaginal bacteria indicates an increased risk of preterm delivery. Further, more abundant vaginal bacteria was found in first trimesters instead of the next two trimesters. Vignal samples collected during first trimester showed richer differences and more predictive value for pregnancy outcoes. In addition, the diversity of the vaginal bacterial community decreased as the gestational age increased, in all samples. *Alloscardovia* was only found in participants who conceived after IVF, and the percentage of *Alloscardovia* in viginal samples of normal delivery group is much higher than the samples from preterm delivery group.*Vobrio* specifically colonized in vagina of pregnant woman in AFT group (those who conceived after IVF (A), first trimester (F), and delivered at term (T)) and *Sporosarcina* was detected only in women with AFT and AST (those who conceived after IVF (A), second trimester (S), and delivered at term (T)). These data indicates that *Alloscardovia, Vobrio* and *Sporosarcina* have great potential in predicting pregnancy outcomes who pregnanted by vitro fertilization

**Conclusions:** Vaginal microbiota were more stable in women who conceived naturally and those who carried pregnancy to term. *Oceanobacillus* might act as a positive biomarker, whereas *Sulfurospirillum* and *Propionispira* may act as negative biomarkers for the risk of preterm birth.

KEYWORDS

pregnancy, vaginal microbiome, preterm birth, *in vitro* fertilization (IVF), pregnancy outcomes

# 1 Introduction

Preterm birth, defined as delivery of the infant before 37 weeks of gestation, is a major contributor to infant death and neonatal mortality, which is a global public health concern (1–3). Human reproduction is an inefficient process, and couples opt for conception through *in vitro* fertilization (IVF) for several reasons such as higher maternal age and failure to conceive naturally. However, IVF (4, 5) is associated with increased rates of multiple gestations and an increased incidence of preterm birth (6, 7).

The World Health Organization has reported that approximately 15 million babies are "born too early" (8–10). The risk of preterm birth is multifactorial, and reproductive tract infection is considered a major triggering factor (11). The costs associated with preterm births have been estimated to rise to $26 billion since the year 2005 (12); thus, it is important to determine an efficient approach to detect the risk of preterm birth during gestation.

Vaginal microbiomes have been proposed to play some roles in risk evaluation and disease diagnosis in women (13). Van der Wijgert et al. reviewed studies reported on molecular vaginal microbiota (VMB) for 6 year, and found that molecular techniques such as sequencing, PCR(polymerase chain reaction), DNA fingerprinting, and DNA hybridization are effective to characterize the VMB, and that lactobacilli-dominated VMB are associated with a healthy vaginal microenvironment (14, 15). Bacterial vaginosis, described as a polybacterial dysbiosis, is a risk factor for preterm births (16–19). A previous study showed that abnormal vaginal microorganisms can negatively affect the pregnancy rate among women who conceive after IVF (18), and that VMB present on the day of embryo transfer significantly affect the pregnancy outcome of IVF (live birth/no live birth) (20), indicating that some VMB might act as positive or negative biomarkers for the risk of preterm births.

It is difficult to identify patients with a risk of preterm birth among those who undergo IVF (17). The microbial composition in the vagina of women who conceive following IVF is still unclear, and whether differences exist in the vaginal bacterial communities during the first, second, and third trimesters in women who conceive naturally and those who conceive *via* IVF is poorly understood. Vaginal bacteria vary among women from different races and ethnicities (21, 22). Previous studies in this regard have mostly focused on women from Europe and America. Few studies focusing on African populations have been reported (23, 24), mainly in Nigeria; limited studies have been conducted on women in China.

In this study, we investigated whether the composition of VMB among women in Northern China who conceived naturally differed from the composition among women who conceived after IVF, as well as whether the composition of VMB among women who subsequently delivered before term differed from that among women who delivered at term. Through this study, we aimed to unveil the changes in VMB in the first, second, and third trimesters of pregnancy to determine the most crucial period and the key biomarkers of vaginal bacteria.

# 2 Materials and methods

## 2.1 Sampling information

The study participants were women from Dalian, a northern coastal city in China. Participants were selected from women who underwent routine prenatal examination at the Dalian Women's and Children's Medical Center between 2017 and 2019. The mothers were enrolled at the hospital admission for delivery and provided written informed consent for participation. The study protocol was approved by the Dalian Women's and Children's Medical Center (approval number 20160021). All operations and evaluations were performed by the same doctor. In total, 19 women had conceived after IVF after a long protocol of conventional solutions of fresh-cycle IVF and conventional luteal support after 1 week; 6 of these delivered before term, whereas 13 delivered at term. Six pregnant women had visited the obstetrics and gynecology clinic for a routine pregnancy test during the same period; three of these women delivered before term, whereas three delivered at term.

The inclusion criteria for this study were as follows: (1) No symptoms of vaginitis, single live births, and primipara; (2) abstinence from sex for 2 weeks, no antibiotic use, no medication history related to the vulva and vagina for 1 month prior to the study, and no history of systemic use of hormone drugs and immunosuppressants; (3) no history of pregnancy complications such as vaginitis, gestational hypertension disease, gestational diabetes mellitus, or pregnancy with abnormal thyroid function; and (4) no history of smoking, drinking, or drug consumption.

Vaginal swab specimens were collected throughout the prenatal check-ups of the healthy participants in all three trimesters of

pregnancy. The swabs were obtained at routine prenatal visits, and by rotating a sterilized swab five times along the vaginal lumen in a circular motion. Speculum was not used. No occurrences of premature rupture of membranes was observed, and no participant was administered intravenous antibiotics during delivery. The participants were divided into those who conceived after IVF and those who conceived naturally (controls), and according to the mode of delivery, as full term (≥37 weeks) and preterm (<37 weeks). The collected swabs were placed into sterile tubes by trained research staff and stored at −80°C for DNA extraction.

The vaginal swabs were collected at least in triplicate at three time points: first trimester (F, 10–13 + 6 weeks), second trimester (S, 20–27 + 6 weeks), and third trimester (T, 28–33 + 6 weeks) of pregnancy. The presence of specific vaginal bacterial communities in any single gestational period in participants who delivered before term (P, six participants in IVF pregnancy group, three participants in natural pregnancy group) was compared with that in participants who delivered at term (T, 13 participants in IVF pregnancy group, 3 participants in natural pregnancy group).

## 2.2 IVF protocol

The standard long IVF regimen was used. A gonadotropin-releasing hormone (GnRH) agonist regimen was the main protocol used in this study, and all patients underwent embryo transfer in fresh cycles. Controlled ovarian hyper-stimulation, oocyte retrieval, and embryo transfer were carried out. Participants were treated with downregulation from the mid-luteal phase of the previous cycle. When the pituitary reached desensitization, recombinant FSH was begun at 150–225 IU/day. Human chorionic gonadotropin (hCG) was given (4000–10,000 IU) once two or more follicles had reached a size of 18 mm. Oocytes were extracted 34–36 h after the hCG trigger, and this was followed by intracytoplasmic sperm injection (ICSI). Participants were injected progesterone (lot NO. 1220507 Shanghai General Pharma, China) at 40 mg/day 48 h after oocytes fertilization. The progesterone supplementation was continued until 10 weeks of gestation after pregnancy was achieved.

## 2.3 Genomic DNA extraction and 16S rDNA gene sequencing analysis

Genomic DNA was extracted using cetyltrimethylammonium bromide; the concentration and purity of DNA were detected by 1% agarose gel electrophoresis. The appropriate amount of sample was placed into the tube and diluted with sterile water to 1 ng/μL. High-throughput sequencing technology was used to sequence the V3–V4 region of the 16S rRNA gene, which was amplified used a specific primer with the barcode 314F-806R (V3V4 primers: 314F-5′ CCTAYGGGRBGCASCAG 806R5′ GGACTACNNGGGTATCTAAT) Phusion® High-Fidelity PCR Master Mix (New England Biolabs) was used for PCR amplification according to the selection of the sequencing region.

## 2.4 PCR product purification

PCR products were mixed with the same volume of 1× loading buffer and detected on 2% agarose gel electrophoresis. Sample strips of 400–450 bp were chosen for further experiments. The PCR products were purified using Qiagen Gel Extraction Kit (Qiagen, Hilden, Germany).

## 2.5 Library preparation and sequencing

TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, San Diego, CA, USA) was used to generate sequencing libraries, and their quality was analyzed using the Qubit@ 2.0 Fluorometer (Life Technologies, Thermo Fisher Scientific, Waltham, MA, USA) and Agilent Bioanalyzer 2100 system. The library was sequenced on Illumina HiSeq2500 platform, and finally, 250-bp paired-end reads were generated.

## 2.6 Sequencing data analyses

The reads were merged by FLASH to obtain raw tags (V1.2.7) (25), and quality filtering was performed to obtain high-quality clean tags (26) by QIIME (V1.7.0) (27). The tags were compared with the Gold database, and finally, the effective tags were obtained using UCHIME algorithm (28, 29). Uparse software (Uparse v7.0.1001) was used to analyze the sequences (30), which were assigned to the same operational taxonomic units (OTUs) once the similarities were more than 97%. The Silva Database (31) was used based on the Ribosomal Databases Project classifier (Version 2.2) algorithm to annotate taxonomic information (32). MUSCLE software (Version 3.8.31) was used to align multiple sequences, to further analyze the phylogenetic relationships of OTUs, and to determine the dominant species in the different groups (33). Based on further information about the abundance of normalized OTUs, we subsequently analyzed their alpha/beta diversity.

## 2.7 Alpha and beta diversity analysis

Alpha diversity was exhibited by six indices, namely Observed-species, Chao1, Shannon, Simpson, ACE, and Good's Coverage, which were calculated with QIIME (Version 1.7.0) and displayed using R software (Version 2.15.3). Beta diversity was calculated by QIIME software (Version 1.7.0) and R software (Version 2.15.3). Species analysis with significant differences between groups was performed using Student's *t*-test and mapping between groups was done using R software (Version 2.15.3).

## 2.8 Date availability statement

We Uploaded the raw 16S rRNA gene sequencing data to the National Center for Biotechnology Information (accession no. PRJNA728871).

10.3389/fendo.2023.1200002

# 3 Results

## 3.1 Basic participant and specimen characteristics

The details of the participants are stated in Table 1. The average age of participants who conceived *via* IVF(A)was 31.42 ± 10.23 years, and that of those who conceived naturally (B) was 31.50 ± 4.52 years (P > 0.05). The mean BMI of the participants in Group A was 26.57 ± 2.03 kg/m$^2$ and that of participants in Group B was 27.48 ± 0.95 kg/m$^2$ (P > 0.05); The mean fetal weight in Group A was 3157.89 ± 628.55 g and that in Group B was 2991.67 ± 452.12 g (P > 0.05).

The details of the specimens collected from the participants who conceived *via* IVF (Group A, 19 participants) and those who conceived naturally (Group B, 6 participants) are stated in Table 2. The participants were further divided into women who delivered before term (P) and those who delivered at term (T).

16S rRNA of each sample was analyzed using high-throughput sequencing technology. The rarefaction curves tended to approach saturation in all samples except for sample BFT (women who conceived naturally (B), first trimester (F), and delivered at term (T))(Figure 1); Good's Coverage revealed that 99%–100% of the species were detected in all samples, suggesting that the data were suitable for further analyses. The details are shown in Table 3.

## 3.2 Bacterial characteristics changed throughout pregnancy

Among women who conceived *via* IVF (A) or naturally (B), the number of species and diversities in the bacterial community in those who delivered at term (T) were significantly higher than those in women who delivered before term (P) throughout pregnancy (F, first trimester; S, second trimester; T, third trimester), regardless of the pregnancy pattern. The numbers of the observed species were as follows: AFT > AFP, AST > ASP, ATT > ATP, BFT > BFP, BST > BSP, and BTT > BTP (Figure 2), indicating that preterm delivery can be predicted by the diversity of vaginal bacterial communities, regardless of whether the pregnancy is natural or *via* IVF. In addition, the bacterial communities could be more easily detected in early gestation than in the next two trimesters; the number of bacteria sharply declined during late gestation, and AFT > AST > ATT, AFP > ASP > ATP, BFT > BST > BTT, BFP > BSP > BTP (Figure 2).

Among women who conceived naturally as well as in those who conceived after IVF, and in those who delivered at term (T) and before term (P), the vaginal bacterial communities in the first trimester were higher than those in the other two trimesters. This indicated that the first trimester was a critical period during pregnancy, and the diversity of vaginal bacterial communities decreased with increasing gestational age in all samples (Figure 2).

The results indicated that conception after IVF could be associated with significantly decreased number of VMB in the first trimester, and that reduced microbial diversity in the vagina could be associated with an increased risk of preterm birth.

## 3.3 Vaginal bacterial diversity in different samples

At the genus level, *Lactobacillus* was dominant in all the samples (>93%), which was consistent with previous reports (34, 35). However, we found some differences among these samples (Figure 3) with regard to bacterial diversity. In the group of participants who conceived after IVF, the vaginal bacterial diversity in those who delivered before term was lower in the first and second trimesters, compared with participants who delivered at

TABLE 1 Basic information of participants.

| | Group A (n=19) | Group B (n=6) | P-value |
|---|---|---|---|
| Age (y) | 31.42 ± 10.23 | 31.50 ± 4.52 | P>0.05 |
| BMI | 26.57 ± 2.03 | 27.48 ± 0.95 | P>0.05 |
| Gestation (wk) | 37.84 ± 2.54 | 33.75 ± 2.61 | P>0.05 |
| Fetal weight (g) | 2486.00 ± 628.55 | 2991.67 ± 452.12 | P>0.05 |

Group A means who conceived via IVF; Group B who conceived naturally.

TABLE 2 Participant groups along with their abbreviations*.

| Pregnancy mode | First word | Second word | | | Delivery mode | Third word |
|---|---|---|---|---|---|---|
| | | First trimester | Second trimester | Third trimester | | |
| IVF | A | F | S | T | Term | T |
| Natural | B | Preterm | P | | | |

Designation of the bigger group contains only the first and the third word.

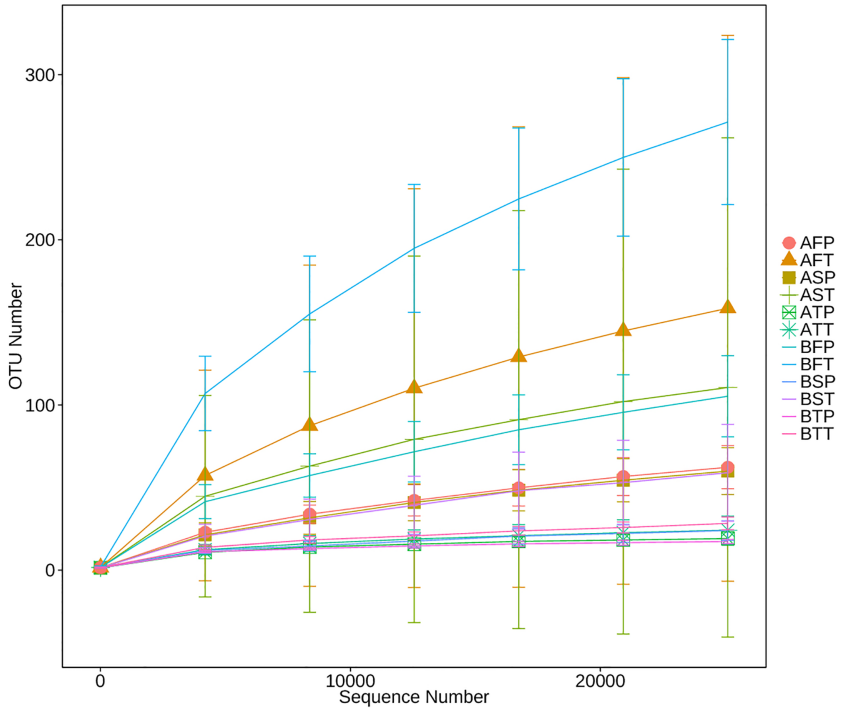



FIGURE 1
Rarefaction analysis of different samples. Rarefaction curves of operational taxonomic units across different samples, which indicated the saturation plateau of the different samples. Samples were designated by three letters. For the first letter, **(A)** indicates IVF pregnancy, **(B)** indicates natural pregnancy; for the second letter, F indicates early gestation, S indicates mid-gestation, T indicates late gestation; for the third letter, P indicates preterm delivery, T indicates term delivery.

term; however, there were no differences in the bacterial diversity in the third trimester between the groups. In the group of participants who conceived naturally, the vaginal bacterial diversity in those who delivered before term was lower in the first trimester, compared with participants who delivered at term; however, there were no differences in vaginal bacterial diversity in the second and third trimesters between the groups.

In all women who conceived naturally, regardless of whether they delivered at term or before term, bacterial diversity was obvious in the first trimester. Thus, detection of vaginal bacteria might be helpful to monitor the risk of preterm birth in pregnant women. We found that among the top 30 genera, *Lactobacillus* was dominant in all samples, *Alloscardovia* was found only in participants who conceived after IVF, no *Bacillus* was found in

TABLE 3 Basic information of different samples.

| Group | Observed species | Shannon | Simpson | Chao1 | ACE | Good's coverage |
|---|---|---|---|---|---|---|
| AFP | 62 | 0.546 | 0.181 | 117.174 | 142.635 | 0.999 |
| AFT | 158 | 0.617 | 0.135 | 270.588 | 302.016 | 0.997 |
| ASP | 60 | 0.468 | 0.121 | 99.868 | 111.116 | 0.999 |
| AST | 110 | 0.555 | 0.122 | 182.460 | 199.882 | 0.998 |
| ATP | 19 | 0.468 | 0.153 | 21.338 | 24.101 | 1.000 |
| ATT | 24 | 0.589 | 0.205 | 33.044 | 35.261 | 1.000 |
| BFP | 105 | 0.575 | 0.119 | 186.236 | 240.106 | 0.998 |
| BFT | 271 | 0.993 | 0.177 | 447.082 | 479.467 | 0.995 |
| BSP | 24 | 0.424 | 0.117 | 36.133 | 40.399 | 1.000 |
| BST | 59 | 0.426 | 0.111 | 95.988 | 114.137 | 0.999 |
| BTP | 17 | 0.449 | 0.123 | 20.500 | 22.917 | 1.000 |
| BTT | 28 | 0.550 | 0.182 | 41.028 | 47.717 | 1.000 |

10.3389/fendo.2023.1200002

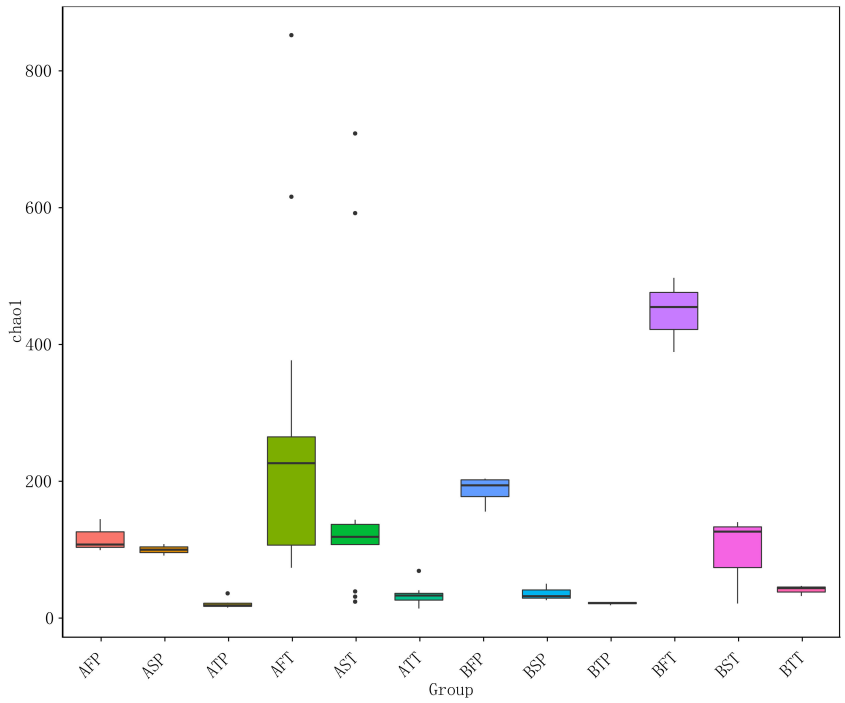



FIGURE 2
Chao 1 diversity index of different samples. For the first letter, A indicates IVF pregnancy, B indicates natural pregnancy; for the second letter, F indicates first trimester, S indicates second trimester, T indicates third trimester; for the third letter, P indicates preterm delivery, T indicates term delivery.

the third trimester in all samples, *Vibrio* appeared only in AFT, *Sulfurospirillum* and *Propionispira* were detected only in BFT, *Sporosarcina* was detected only in AFT and AST, *Lactococcus* was found only in BFT and BFP, and *Oceanobacillus* was found only in BFP. In women who conceive after IVF, *Vibrio* and *Sporosarcina* might act as negative biomarkers for the risk of preterm delivery. As per our research, in women who conceive naturally, *Oceanobacillus* can act as a positive biomarker, and *Sulfurospirillum* and *Propionispira* can act as negative biomarkers for the risk of preterm delivery.

## 3.4 Bacterial diversity is important in naturally pregnant women who deliver at term

We further analyzed the differences of vaginal bacterial diversities among the different groups: between participants who conceived after IVF (A) and those who conceived naturally (B), and between women who delivered before term (P) and those who delivered at term (T) (Figure 4). We observed that group BT had the highest number of observed species, and although groups BP and AT had a similar number

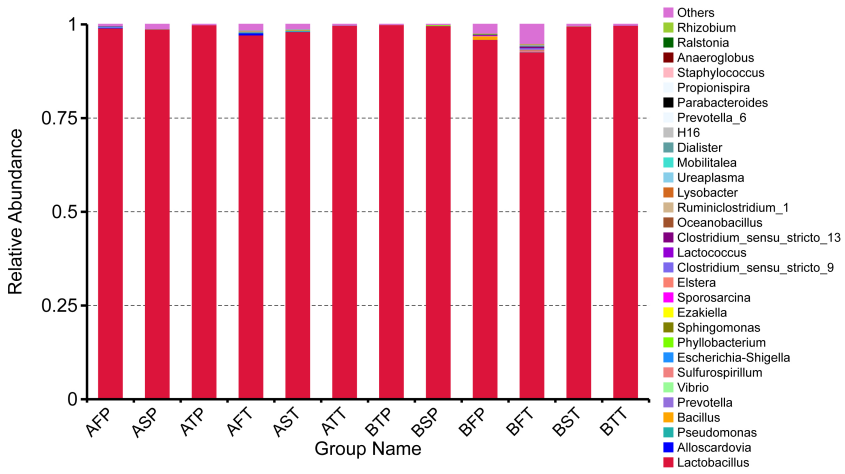



FIGURE 3
Relative abundance of top 30 genera in different samples. The distribution of the top 30 genera of all samples was analyzed.

of species, group AT had higher diversity than group BP did, and group AP had the lowest number of observed species and diversity among the four groups, indicating that bacterial diversity might play an important role in a healthy pregnancy. *Firmicutes* was dominant (98.501%) in all the samples, followed by *Proteobacteria* (0.133%), *Actinobacteria* (0.052%), and *Bacteroidetes* (0.050%). *Actinobacteria* appeared only in groups AT and AP (mostly in group AT), and *Sulfurospirillum* was found only in group BT. *Vibrio* was found only in group AT (Figure 5). The diversities and numbers of microorganisms were important for further outcome of pregnancy, especially for patients who conceived after IVF. Hence, standard protocols should be established to support a shift of vaginal microbiota during IVF therapy.

## 3.5 Differential species analysis in different groups

Student's *t*-test was performed to determine the significant difference between the bacterial species in all the groups ($P < 0.05$), and the results showed significant differences between groups AP and AT. *Clostridia* were a significant differential species and were dominant in group AT; between groups BP and BT, unidentified *Actinobacteria* were a significant differential species and were dominant in group BT; between groups AT and BT, *Lactobacillus equicursoris* was a significant differential species and was dominant in group AT (Figure 6). Thus, *Clostridia* and *Lactobacillus equicursoris* can act as positive biomarkers for preterm birth in patients who conceive after IVF.

Between groups P and T, 35 bacterial species were found only in group P (Figure 7), and the annotation results indicated that they were novel species and may act as biomarkers for the risk of preterm birth.

# 4 Discussion

Preterm birth is a major contributor to infant death and neonatal mortality, which is a global concern, and researchers have mostly focused on finding an efficient approach to prevent preterm delivery (36, 37). Currently, IVF is an effective method for conception in women who fail to conceive naturally; however, IVF is associated with an increased risk of preterm birth. Dynamic changes in the VM of women who become pregnant through IVF are still poorly understood. As the importance of microorganisms for human health is well known (38–40), vaginal bacteria have been extensively studied in recent years, with an aim to determine their roles in pregnant women (13, 41–44). Previous researchers have mostly focused on the association between vaginal bacteria and preterm birth among American, African, and European women (21, 22, 45, 46), and little is known about the relationship between the vaginal microbiome and delivery pattern among women from Northern China who conceive *via* IVF.

In this study, using high-throughput sequencing technology (41, 47), we characterized the vaginal microbiome of women from Northern China who conceived *via* IVF and those who conceived naturally. We focused on the relationships between VMB and pregnancy after IVF, especially their effect on pregnancy outcome, to determine whether any taxa distinguished women who conceive after IVF from those who conceive naturally, and women who deliver before term from those who deliver at term.

We found that the community richness and diversity of VMB was lesser in women who conceived after IVF than in those who conceived naturally throughout pregnancy, regardless of the delivery pattern, suggesting that the VMB were more stable in naturally pregnant women than in those who were pregnant after IVF. The number of observed species was lesser in women who conceived *via* IVF and delivered before term than in women who delivered at term, suggesting that the vaginal bacterial communities associated with term birth were more stable than those associated with preterm birth. The diversity and richness of the vaginal bacterial communities diversity decreased in each trimester (first trimester > second trimester > third trimester), indicating that early pregnancy is a significant period for the vaginal bacterial community, and that the fluctuation of vaginal bacteria in the first trimester may be a biomarker for preterm birth.

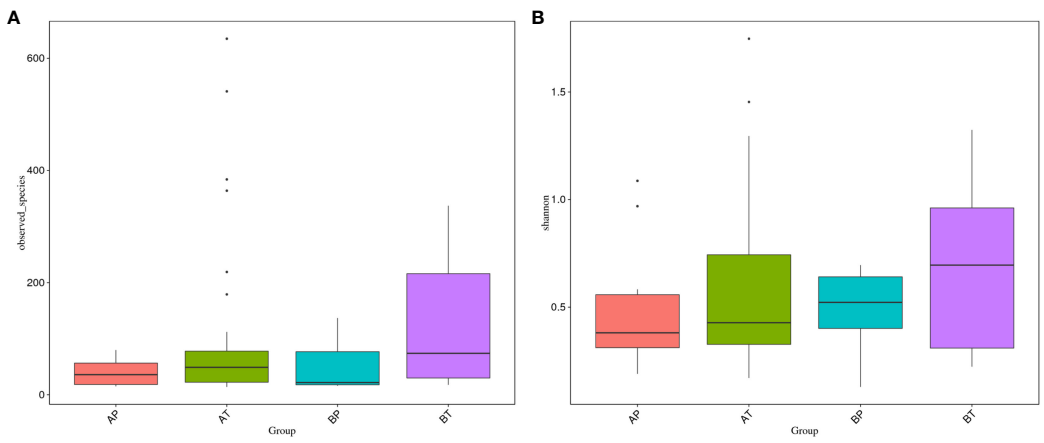



FIGURE 4
Vaginal bacteria diversities among different groups. **(A)** Women who were pregnant after IVF; **(B)** naturally pregnant women; P, pregnant women who delivered before term; T, pregnant women who delivered at term.

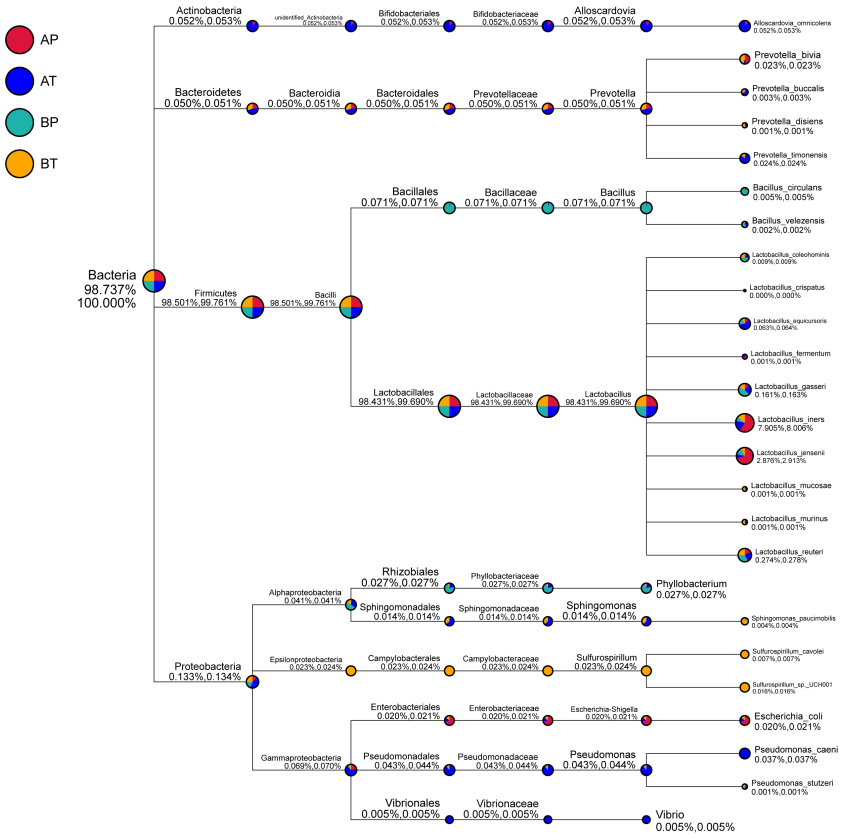



FIGURE 5
Taxonomic tree of specific species in each group. A, women who were pregnant after IVF; B, naturally pregnant women; P, pregnant women who delivered preterm; T, pregnant women who delivered at term.

A previous study reported that the genus *Lactobacillus* was predominant in the vagina of naturally pregnant women (42), and lower vaginal levels of *Lactobacillus crispatus* indicate a higher possibility of premature birth. Fettweis (48) reported Shannon Diversity is higher in women who deliver preterm. In that study, they used the data from women of African ancestry. Subsequently confirmed in other studies, the Vaginal microbiome profiles of women of African, European and Asian ancestry differ significantly. In Richard W. Hyman (43) study from Stanford University, they showed Chao I analysis is significantly distinguished by race/ethnicity. Black people are much higher than Hispanic, Caucasian and Asian. Chao I value of Caucasian and Asian are close and much lower than Black. And When they measure Shannon Diversity Index for Caucasian, they found Shannon Diversity is higher in woman who deliver at term which is consistent with us. We think microbiota diversity is influenced by many factors included ancestry and area.

Consistent with this observation, our data also showed that this genus was dominant in women who were pregnant after IVF. To determine the specific taxa in women who were pregnant after IVF, the details of vaginal bacteria were analyzed; we observed that *Alloscardovia* was found only in women who were pregnant after IVF, *Vibrio* was found only in AFT, and *Sporosarcina* was detected only in AFT and AST. Multiple group analysis showed that the number of observed species and diversity were in the order BT > AT > BP > AP, indicating that VMB play important roles in a healthy pregnancy. *Actinobacteria* were detected only in groups AT and AP (especially in group AT), *Vibrio* was found only in group AT, and *Sulfurospirillum* was found only in group BT. *Alloscardovia* was specifically detected in women who conceived *via* IVF and could act as a positive biomarker, whereas *Vibrio* and *Sporosarcina* could act as negative biomarkers for the risk of preterm birth. *Oceanobacillus* might act as a positive biomarker, whereas *Sulfurospirillum* and *Propionispira* might act as negative biomarkers for the risk of preterm birth in women who conceive naturally.

Oceanobacillus, a member of the Baciliaceae, was isolated from deep-sea sediments by Lu et al. (49) using O. iheyensis as a model species and was gram positive. Studies have shown that colonizing vaginas in pregnant mice causes vaginal inflammation and may alter cervix functions and integrity (50). Previous work has validated the breakdown of the cervical epithelial barrier due to inflammatory damage (51).

Kumar M and Hočevar K (52, 53) identified that increased bacterial diversity/BV was positively associated with PTB.in Hočevar K' study of shown that the predominant bacterial families were Lactobacillaceae (77.9%), in our finding that Lactobacillaceae dominant in all the samples(>93%), that mean there are great differences in the results of vaginal microecology between the two populations. And the data of Kumar'reaserch are more broadly applicable to Asian women in ethnically-diverse populations in high income nations.

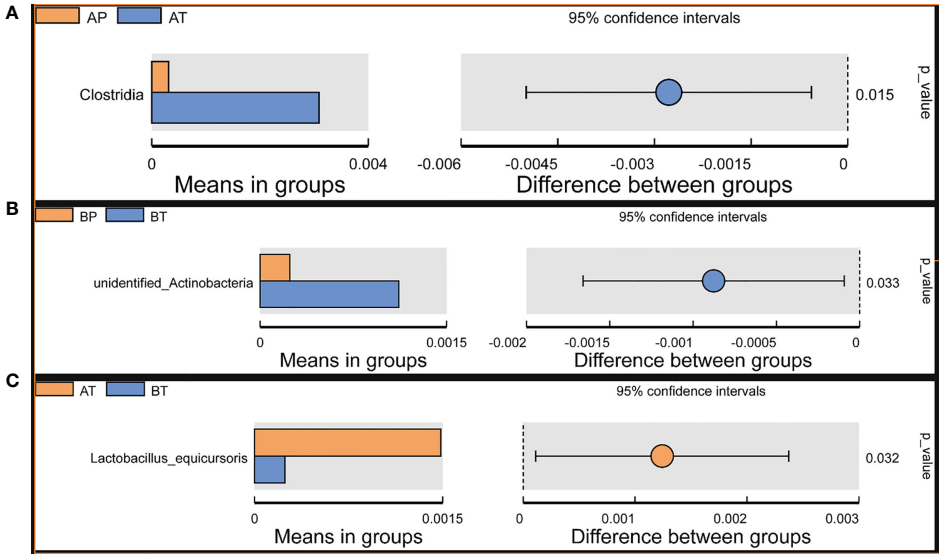



FIGURE 6
T-test analyses of different microbes between the groups. T-test analysis of different microbes between the groups. **(A)** AP and AT; **(B)** BP and BT; **(C)** AT and BT.

The main reason for the different research results may be that all our samples were after IVF and some vaginal operations were performed before pregnancy.

So far the characteristics of vaginal Microecology in pregnant women after IVF have not been described.

19 women had conceived after IVF after a long protocol of conventional solutions of fresh-cycle IVF, our sample pregnant women are completely infertile due to male factors, avoiding the changes of vaginal microecology caused by changes in drug and hormone levels. However, before becoming pregnancy, these

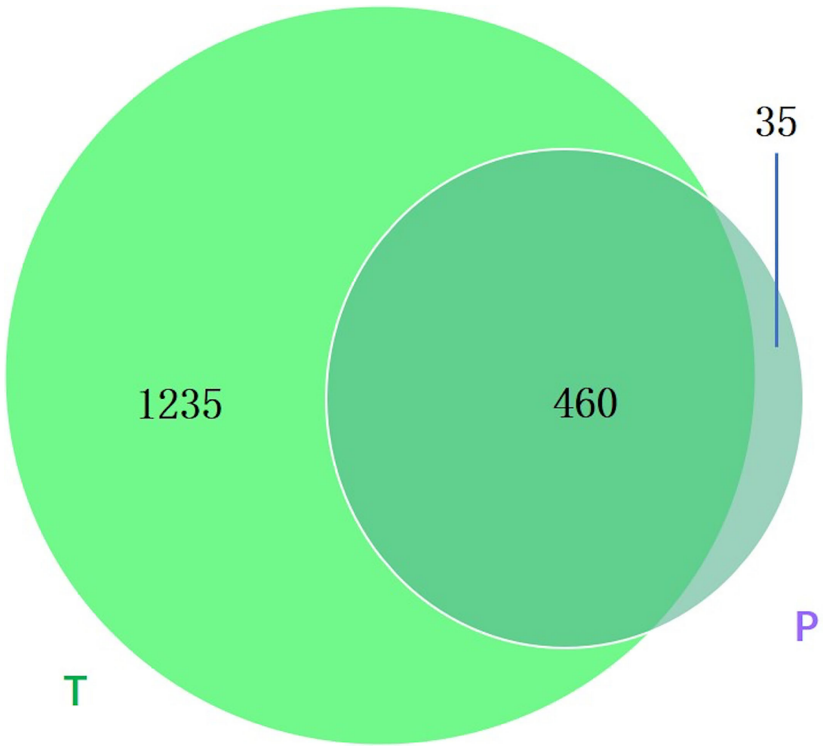



FIGURE 7
Venn diagram showing the number of microbes in different groups. The sharing numbers indicate the same microbes between the two groups, and the sole number represents microbes that appeared in only one group. P indicates preterm delivery and T indicates term delivery.

pregnant women will have repeated vaginal ultrasonography, vaginal flushing and transvaginal fornix oocyte collection, which will destroy the integrity of vaginal flora, These reasons may be the reasons for the differences in vaginal flora samples. Because our samples are all male factors, this also avoids the impact of changes in maternal hormone levels on vaginal flora.

Before pregnancy, vaginal flora structure was destroyed after vaginal flushing. So after pregnancy, vaginal flora structure also means reestablishment. For this reason perhaps the number of species and diversities in the bacterial community in those who delivered at before term throughout pregnancy were lower.

These results are currently observed in our research. The specific mechanism is not very clear. We will study it in the next research.Second, due to the small number of pregnant women with simple male factor and insufficient sample size, the results may also be different. We will increase sample size in the next research.

We recruited 19 volunteers in IVF group and 6 volunteers in Natural conception group and collected vaginal samples from each volunteer in three periods. These samples are very precious and can effectively remind us of the relationship between vaginal bacteria and pregnancy outcomes. In our results, we found vaginal samples collected during first trimester showed richer differences and more predictive value for pregnancy outcomes. In addition, these data indicates that Alloscardovia, Vobrio and Sporosarcina have great potential in predicting pregnancy outcomes who pregnant by vitro fertilization. While Oceanobacillus, Sulfurospirillum and Propionispira have great potential in predicting the risk of preterm birth in women who conceive naturally.On the other hand, the sample size in our research is not enough to conclude these bacteria strains we suggested could be biomarker for preterm birth. And the mechanism of these bacteria strains on preterm birth is unclear. But our data is a powerful indicator.Although the exact causal relationship remains to be determined, our results confirm an association between some bacteria and preterm birth. However, the richness, diversity, and stability of the microbiome may be important during pregnancy. Our finding is consistent with previous studies (54, 55).

In summary, our study showed that compared with naturally pregnant women, women who conceive *via* IVF and those who deliver before term show lesser richness, diversity, and stability of VMB. Therefore, standard protocols should be established and used to support a shift of VMB.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA728871.

## Ethics statement

The studies involving human participants were reviewed and approved by Dalian Women's and Children's Medical Center (approval number 20160021). The patients/participants provided their written informed consent to participate in this study.

## Author contributions

YT: responsible for the experimental concept, data collection, drafting papers; QS: responsible for the data collection, drafting papers; XS: Responsible for the important revision of the thesis; ZW: Responsible for the important revision of the thesis and for the important revision of the thesis. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by National Key R&D Program of China (No. 2022YFC2704504), and Dalian Science and Technology Innovation Fund (No. 2019J13SN83).

## Acknowledgments

The authors extend their sincere gratitude to all medical workers for their enthusiastic help to assist in samples collection in our hospital.

## Conflict of interest

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.

## Publisher's note



## References

1. Witkin SS. The vaginal microbiome, vaginal anti-microbial defence mechanisms and the clinical challenge of reducing infection-related preterm birth. *BJOG: Int J Obstetrics Gynaecol* (2015) 122(2):213–8. doi: 10.1111/1471-0528.13115

2. Beck S, Wojdyla D, Say L, Betran AP, Merialdi M, Requejo JH, et al. The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity. *Bull World Health Organ* (2010) 88:31–8. doi: 10.2471/BLT.08.062554

3. Goldenberg RL, McClure EM. The epidemiology of preterm birth. In: *Preterm birth: prevention and management* (2010). p. 22–38.

4. Doyle JO, Richter KS, Lim J, Stillman RJ, Graham JR, Tucker MJ. Successful elective and medically indicated oocyte vitrification and warming for autologous *in vitro* fertilization, with predicted birth probabilities for fertility preservation according to number of cryopreserved oocytes and age at retrieval. *Fertil Steril* (2016) 105(2):459–66.e2. doi: 10.1016/j.fertnstert.2015.10.026

5. Simon L, Zini A, Dyachenko A, Ciampi A, Carrell DT. A systematic review and meta-analysis to determine the effect of sperm DNA damage on *in vitro* fertilization and intracytoplasmic sperm injection outcome. *Asian J androl* (2017) 19(1):80. doi: 10.4103/1008-682X.182822

6. Saldeen P, Sundström P. Would legislation imposing single embryo transfer be a feasible way to reduce the rate of multiple pregnancies after IVF treatment? *Hum Reprod* (2005) 20(1):4–8. doi: 10.1093/humrep/deh610

7. Sunkara SK, La Marca A, Seed PT, Khalaf Y. Increased risk of preterm birth and low birthweight with very high number of oocytes following IVF: an analysis of 65 868 singleton live birth outcomes. *Hum Reprod* (2015) 30(6):1473–80. doi: 10.1093/humrep/dev076

8. Blencowe H, Cousens S, Chou D, Oestergaard M, Say L, Moller AB, et al. Born too soon: the global epidemiology of 15 million preterm births. *Reprod Health* (2013) 10(1): S2. doi: 10.1186/1742-4755-10-S1-S2

9. Hamilton BE, Hoyert DL, Martin JA, Strobino DM, Guyer B. Annual summary of vital statistics: 2010–2011. *Pediatrics* (2013) 131(3):2012–3769. peds. doi: 10.1542/peds.2012-3769

10. Osterman MJK, Kochanek KD, MacDorman MF, Strobino DM, Guyer B. Annual summary of vital statistics: 2012–2013. *Pediatrics* (2015) 135(6):2015–0434. doi: 10.1542/peds.2015-0434

11. Andrews WW, Hauth JC, Goldenberg RL. Infection and preterm birth. *Am J perinatol* (2000) 17(07):357–66. doi: 10.1055/s-2000-13448

12. Behrman RE, Butler AS. *Preterm birth: causes, consequences, and prevention*. Washington, DC: National Academies Press (2007).

13. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci* (2011) 108(Supplement 1):4680–7. doi: 10.1073/pnas.1002611107

14. van der Wijgert JHHM, Borgdorff H, Verhelst R, Crucitti T, Francis S, Verstraelen H, et al. The vaginal microbiota: what have we learned after a decade of molecular characterization. *PloS One* (2014) 9(8):e105998. doi: 10.1371/journal.pone.0105998

15. Jakobsson T, Forsum U. Lactobacillus iners: a marker of changes in the vaginal flora. *J Clin Microbiol* (2007) 45(9):3145–5. doi: 10.1128/JCM.00558-07

16. Leitich H, Bodner-Adler B, Brunbauer M, Kaider A, Egarter C, Husslein P. Bacterial vaginosis as a risk factor for preterm delivery: a meta-analysis. *Am J Obstetrics Gynecol* (2003) 189(1):139–47. doi: 10.1067/mob.2003.339

17. Romero R, Chaiworapongsa T, Kuivaniemi H, Tromp G. Bacterial vaginosis, the inflammatory response and the risk of preterm birth: a role for genetic epidemiology in the prevention of preterm birth. *Am J Obstetrics Gynecol* (2004) 190(6):1509–19. doi: 10.1016/j.ajog.2004.01.002

18. Klebanoff MA, Hillier SL, Nugent RP, MacPherson CA, Hauth JC, Carey JC, et al. Is bacterial vaginosis a stronger risk factor for preterm birth when it is diagnosed earlier in gestation? *Am J Obstetrics Gynecol* (2005) 192(2):470–7. doi: 10.1016/j.ajog.2004.07.017

19. Donders GG, Van Calsteren C, Bellen G, Reybrouck R, Van den Bosch T, Riphagen I, et al. Association between abnormal vaginal flora and cervical length as risk factors for preterm birth. *Ultrasound Obstetrics Gynecol: Off J Int Soc Ultrasound Obstetrics Gynecol* (2010). doi: 10.1002/uog.7568

20. Hyman RW, Herndon CN, Jiang H, Palm C, Fukushima M, Bernstein D, et al. The dynamics of the vaginal microbiome during infertility therapy with *in vitro* fertilization-embryo transfer. *J Assisted Reprod Genet* (2012) 29(2):105–15. doi: 10.1007/s10815-011-9694-6

21. Fettweis JM, Brooks JP, Serrano MG, Sheth NU, Girerd PH, Edwards DJ, et al. Differences in vaginal microbiome in African American women versus women of European ancestry. *Microbiology* (2014) 160(10):2272–82. doi: 10.1099/mic.0.081034-0

22. Stout MJ, Zhou Y, Wylie KM, Tarr PI, Macones GA, Tuuli MG. Early pregnancy vaginal microbiome trends and preterm birth. *Am J Obstetrics gynecol* (2017) 217 (3):356. e1–356. e18. doi: 10.1016/j.ajog.2017.05.030

23. Adaobi CO, Nneka RA, Chinyere CE, Oguejiofor CB. Comparative abundance and functional biomarkers of the vaginal and gut microbiome of Nigerian women with bacterial vaginosis: a study with 16S rRNA metagenomics. *J Lab Med* (2019) 29):1–26.

24. Anukam KC, Agbakoba NR, Okoli AC, Oguejiofor CB. Vaginal bacteriome of Nigerian women in health and disease: a study with 16S rRNA metagenomics. *Trop J Obstetrics Gynaecol* (2019) 36(1):96–104. doi: 10.4103/TJOG.TJOG_67_18

25. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* (2011) 27(21):2957–63. doi: 10.1093/bioinformatics/btr507

26. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. *Nat Methods* (2013) 10(1):57. doi: 10.1038/nmeth.2276

27. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* (2010) 7(5):335. doi: 10.1038/nmeth.f.303

28. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* (2011) 27(16):2194–200. doi: 10.1093/bioinformatics/btr381

29. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* (2011) 21:494–504. doi: 10.1101/gr.112730.110

30. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* (2013) 10(10):996. doi: 10.1038/nmeth.2604

31. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* (2006) 72(7):5069–72. doi: 10.1128/AEM.03006-05

32. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* (2007) 73(16):5261–7. doi: 10.1128/AEM.00062-07

33. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* (2004) 32(5):1792–7. doi: 10.1093/nar/gkh340

34. Verstraelen H, Verhelst R, Claeys G, De Backer E, Temmerman M, Vaneechoutte M. Longitudinal analysis of the vaginal microflora in pregnancy suggests that l. crispatus promotes the stability of the normal vaginal microflora and that *L. gasseri* and/or *L. iners* are more conducive to the occurrence of abnormal vaginal microflora. *BMC Microbiol* (2009) 9(1):116. doi: 10.1186/1471-2180-9-116

35. Hernández-Rodríguez C, Romero-González R, Albani-Campanario M, Figueroa-Damián R, Meraz-Cruz N, Hernández-Guerrero C. Vaginal microbiota of healthy pregnant Mexican women is constituted by four lactobacillus species and several vaginosis-associated bacteria. *Infect Dis Obstetrics Gynecol* (2011) 2011:851485. doi: 10.1155/2011/851485

36. Meis PJ, Goldenberg RL, Mercer B, Moawad A, Das A, McNellis D, et al. The preterm prediction study: significance of vaginal infections. *Am J Obstetrics Gynecol* (1995) 173(4):1231–5. doi: 10.1016/0002-9378(95)91360-2

37. Carey JC, Klebanoff MA, Hauth JC, Hillier SL, Thom EA, Ernest JM, et al. Metronidazole to prevent preterm delivery in pregnant women with asymptomatic bacterial vaginosis. *New Engl J Med* (2000) 342(8):534–40. doi: 10.1056/NEJM200002243420802

38. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* (2012) 486(7402):207. doi: 10.1038/nature11234

39. Zhou Y, Gao H, Mihindukulasuriya KA, La Rosa PS, Wylie KM, Vishnivetskaya T, et al. Biogeography of the ecosystems of the healthy human body. *Genome Biol* (2013) 14(1):R1. doi: 10.1186/gb-2013-14-1-r1

40. Ma B, Forney LJ, Ravel J. Vaginal microbiome: rethinking health and disease. *Annu Rev Microbiol* (2012) 66:371–89. doi: 10.1146/annurev-micro-092611-150157

41. Aagaard K, Versalovic J, Petrosino J, Segata N, Mistretta TA, Coarfa C, et al. 73: metagenomic-based approach to a comprehensive characterization of the vaginal microbiome signature in pregnancy. *Am J Obstetrics Gynecol* (2011) 204(1):S42. doi: 10.1016/j.ajog.2010.10.087

42. Romero R, Hassan SS, Gajer P, Tarca AL, Fadrosh DW, Nikita L, et al. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome* (2014) 2(1):4. doi: 10.1186/2049-2618-2-10

43. Hyman RW, Fukushima M, Jiang H, Fung E, Rand L, Johnson B, et al. Diversity of the vaginal microbiome correlates with preterm birth. *Reprod Sci* (2014) 21(1):32–40. doi: 10.1177/1933719113488838

44. Romero R, Hassan SS, Gajer P, Tarca AL, Fadrosh DW, Bieda J, et al. The vaginal microbiota of pregnant women who subsequently have spontaneous preterm labor and delivery and those with a normal delivery at term. *Microbiome* (2014) 2(1):18. doi: 10.1186/2049-2618-2-18

45. Hillier SL, Krohn MA, Cassen E, Easterling TR, Rabe LK, Eschenbach DA. The role of bacterial vaginosis and vaginal bacteria in amniotic fluid infection in women in preterm labor with intact fetal membranes. *Clin Infect Dis* (1995) Suppl 2:S276–8. doi: 10.1093/clinids/20.Supplement_2.S276

46. Nelson DB, Hanlon A, Nachamkin I, Haggerty C, Mastrogiannis DS, Liu C, et al. Early pregnancy changes in bacterial vaginosis-associated bacteria and preterm delivery. *Paediatric Perinatal Epidemiol* (2014) 28(2):88–96. doi: 10.1111/ppe.12106

47. Pallen MJ, Loman NJ, Penn CW. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr Opin Microbiol* (2010) 13 (5):625–31. doi: 10.1016/j.mib.2010.08.003

48. Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, et al. The vaginal microbiome and preterm birth. *Nat Med* (2019) 25(6):1012–21. doi: 10.1038/s41591-019-0450-2

49. Lu J, Nogi Y, Takami H. Oceanobacillus iheyensis gen. nov., sp. nov., a deep-sea extremely halotolerant and alkaliphilic species isolated from a depth of 1050 m on the iheya ridge. *FEMS Microbiol Lett* (2001) 205(2):291–7. doi: 10.1111/j.1574-6968.2001.tb10963.x

50. Sierra LJ, Brown AG, Barilá GO, Anton L, Barnum CE, Shetye SS, et al. Colonization of the cervicovaginal space with gardnerella vaginalis leads to local inflammation and cervical remodeling in pregnant mice. *PloS One* (2018) 13(1): e0191524. doi: 10.1371/journal.pone.0191524

51. Nold C, Anton L, Brown A, Elovitz M. Inflammation promotes a cytokine response and disrupts the cervical epithelial barrier: a possible mechanism of premature cervical remodeling and preterm birth. *Am J Obstet Gynecol* (2012) 206(3):208.e1–208.e2087. doi: 10.1016/j.ajog.2011.12.036

52. Kumar M, Murugesan S, Singh P, Saadaoui M, Elhag DA, Terranegra A, et al. Vaginal microbiota and cytokine levels predict preterm delivery in Asian women. *Front Cell Infect Microbiol* (2021) 11:639665. doi: 10.3389/fcimb.2021.639665

53. Hočevar K, Maver A, Vidmar Šimic M, Hodžić A, Haslberger A, Premru Seršen T, et al. Vaginal microbiome signature is associated with spontaneous preterm delivery. *Front Med (Lausanne)* (2019) 6:201. doi: 10.3389/fmed.2019.00201

54. Freitas AC, Bocking A, Hill JE, Money DM; VOGUE Research Group. Increased richness and diversity of the vaginal microbiota and spontaneous preterm birth. *Microbiome* (2018) 6(1):117. doi: 10.1186/s40168-018-0502-8

55. Haahr T, Jensen JS, Thomsen L, Duus L, Rygaard K, Humaidan P. Abnormal vaginal microbiota may be associated with poor reproductive outcomes: a prospective study in IVF patients. *Hum Reprod* (2016) 31(4):795–803. doi: 10.1093/humrep/dew026

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 19 July 2023
DOI 10.3389/fendo.2023.1174911



OPEN ACCESS

EDITED BY
Wenjie Shi,
Otto von Guericke University Magdeburg, Germany

REVIEWED BY
Guangying Cui,
Zhengzhou University, China
Liang Yu,
Second Affiliated Hospital of Harbin Medical University, China

*CORRESPONDENCE
Xin Lian
jessicashengjing@hotmail.com

RECEIVED 27 February 2023
ACCEPTED 21 June 2023
PUBLISHED 19 July 2023

CITATION
Zhang X, Li X, Wang C, Wang S, Zhuang Y, Liu B and Lian X (2023) Identification of markers for predicting prognosis and endocrine metabolism in nasopharyngeal carcinoma by miRNA–mRNA network mining and machine learning.
*Front. Endocrinol.* 14:1174911.
doi: 10.3389/fendo.2023.1174911



# Identification of markers for predicting prognosis and endocrine metabolism in nasopharyngeal carcinoma by miRNA–mRNA network mining and machine learning

Xixia Zhang[1], Xiao Li[2], Caixia Wang[2], Shuang Wang[2], Yuan Zhuang[2], Bing Liu[1] and Xin Lian[1]*

[1]Department of Otolaryngology Head and Neck Surgery, Shengjing Hospital of China Medical University, Shenyang, China, [2]Department Obstetrics and Gynecology, Shengjing Hospital of China Medical University, Shenyang, China

**Background:** Nasopharyngeal cancer (NPC) has a high incidence in Southern China and Asia, and its survival is extremely poor in advanced patients. MiRNAs play critical roles in regulating gene expression and serve as therapeutic targets in cancer. This study sought to disclose key miRNAs and target genes responsible for NPC prognosis and endocrine metabolism.

**Materials and methods:** Three datasets (GSE32960, GSE70970, and GSE102349) of NPC samples came from Gene Expression Omnibus (GEO). Limma and WGCNA were applied to identify key prognostic miRNAs. There were 12 types of miRNA tools implemented to study potential target genes (mRNAs) of miRNAs. Univariate Cox regression and stepAIC were introduced to construct risk models. Pearson analysis was conducted to analyze the correlation between endocrine metabolism and RiskScore. Single-sample gene set enrichment analysis (ssGSEA), MCP-counter, and ESTIMATE were performed for immune analysis. The response to immunotherapy was predicted by TIDE and SubMap analyses.

**Results:** Two key miRNAs (miR-142-3p and miR-93) were closely involved in NPC prognosis. The expression of the two miRNAs was dysregulated in NPC cell lines. A total of 125 potential target genes of the key miRNAs were screened, and they were enriched in autophagy and mitophagy pathways. Five target genes (E2F1, KCNJ8, SUCO, HECTD1, and KIF23) were identified to construct a prognostic model, which was used to divide patients into high group and low group. RiskScore was negatively correlated with most endocrine-related genes and pathways. The low-risk group manifested higher immune infiltration, anticancer response, more activated immune-related pathways, and higher response to immunotherapy than the high-risk group.

**Conclusions:** This study revealed two key miRNAs that were highly contributable to NPC prognosis. We delineated the specific links between key miRNAs and prognostic mRNAs with miRNA–mRNA networks. The effectiveness of the five-gene model in predicting NPC prognosis as well as endocrine metabolism provided a guidance for personalized immunotherapy in NPC patients.

KEYWORDS

nasopharyngeal cancer, micro RNAs, miRNA-mRNA network, immunotherapy, immune checkpoint blockade, risk model, endocrine

## Introduction

Nasopharyngeal cancer (NPC) is an epithelial malignancy, which has discrepant occurrence in different regions and countries. The etiology of NPC is multiple and remains incompletely understood, but most cases are closely linked to Epstein–Barr virus (EBV) infection (1). In Western countries, the incidence rate is relatively low with a ratio of 1:100,000 each year. However, in the regions of Southern China and Asia, the annual incidence elevates to 25–50 cases per 100,000 (2), which accounts for approximately 70% new cases worldwide (3). The incidence also strikingly increases in the recent decades in China. The age-standardized incidence rate (ASIR) was 3.3/100,000 in 1990, whereas it reached 5.7/100,000 in 2019. The rising incidence rate was especially startling in men, with ASIR of 4.3/100,000 to 8.6/100,000 from 1990 to 2019 (4). The distant metastasis contributes to an extremely poor prognosis in NPC patients, and the patients of stages III and IV have a 5-year survival rate less than 10%.

With intensive usage of modulated treatment including chemotherapy and radiotherapy, the control of NPC metastasis reaches a satisfactory outcome and the 5-year survival rate drastically improves (5–7). Nevertheless, the prognosis and treatment efficiency were awfully unfavorable in the NPC patients of late stages. It is still a great challenge for clinicians to control and lessen the metastasis and recurrence of advanced NPC patients. In the recent years, immunotherapy reaches a milestone in clinical cancer therapy, not excluding in NPC. Immune checkpoint blockade (ICB) therapy is one of the promising strategies to increase antitumor activity in NPC. Studies have shown that NPC patients expresses high expression levels of programmed death protein 1 (PD-1) and programmed death ligand 1 (PD-L1) that are associated with poor outcomes and recurrence (8, 9). The blockade of PD-1/PD-L1 expression can recover the ability of cytotoxic lymphocytes exerting anticancer response. Lines of clinical trials of ICB therapy have demonstrated the positive response to anti-PD-1/PD-L1 drugs such as pembrolizumab in phase 1 and 2 studies (10, 11). In the phase 2 study, of 44 enrolled NPC patients, eight patients reached a partial response and one patient reached a complete response, showing an objective response rate (ORR) of 20.5% (11). As we know, in the tumor microenvironment, the expression of PD-1 on the surface of tumor cells and the combination of PD-L1 on the surface of tumor-infiltrating lymphocytes can inhibit the activity of T cells, lead to the loss of function of effector factors TNF-a, IFN-gamma, and IL-2, and inhibit cytolysis function through granulozyme B and perforin, and ultimately promote immune escape (12, 13). Obviously, still a high proportion of NPC patients showed a negative response to ICB therapy, which may result from their disadvantageous tumor microenvironment. Therefore, in order to raise the treatment accuracy, understanding the mechanism of immune evasion and developing molecular biomarkers is critically needed.

Over the last few decades, it has become clear that endocrines are also involved in regulating tumor cells and that cancer cells themselves abnormally express and respond to many hormones (14). Both leptin and its receptor have been reported in cancer biopsy specimens, indicating autocrine and/or paracrine roles in tumorigenesis (15). Several hypothalamic hormones have been implicated in a variety of human cancers, including growth hormone-releasing hormone, luteinizing hormone-releasing hormone, somatostatin, and bombotin (16, 17). It is becoming increasingly clear that many human cancer cells are sensitive to a variety of hormones and that they themselves express many hormones that play an important role in the development and progression of cancer.

Non-coding RNAs play essential roles in gene modulation and pathway regulation. Some microRNAs (miRNAs) were unveiled to participate in NPC invasion and metastasis, immune escape, and resistance to chemotherapy and radiotherapy (18), endowing miRNAs possible to serve as potential therapeutic targets (19). For example, miR-26a was found to have an anticancer effect and overexpressing miR-26a could inhibit the metastatic feature in NPC cells (20). MiR-663 targeting WAF1/CIP1 promotes the proliferation and tumorigenesis in NPC cells (21). The crosstalk between extracellular microRNAs and the tumor microenvironment has also been profoundly parsed by previous studies (22, 23).

**Abbreviations:** ASIR, age-standardized incidence rate; ceRNA, competing endogenous RNA; CR, complete response; DEmiRNAs, differentially expressed miRNAs; DFS, disease-free survival; DDR, DNA damage repair; EBV, Epstein–Barr virus; GEO, Gene Expression Omnibus; GO, Gene Ontology; HR, hazard ratio; ICB, immune checkpoint blockade; KEGG, Kyoto Encyclopedia of Genes and Genomes; MFS, metastasis-free survival; miRNAs, micro RNAs; MDSCs, myeloid-derived suppressor cells; NPC, nasopharyngeal cancer; PR, partial response; PD-L1, programmed death ligand 1; PD-1, programmed death protein 1; ROC, receiver operating characteristic; ssGSEA, single sample gene set enrichment analysis; TOM, topological overlap matrix; WGCNA, weighted correlation network analysis.

Consequently, our study tried to mine the key miRNAs that had a prognostic value in NPC. We identified two potentially key miRNAs (has-miR-142-3p and has-miR-93) closely involved in NPC prognosis. By building competing endogenous RNA (ceRNA) networks using multiple miRNA tools, we determined 125 potential target genes (mRNAs) and screened five key genes contributable for the prognostic model. The five-gene prognostic model manifested a satisfactory performance in prognosis prediction and estimating the response to immunotherapy and chemotherapy.

# Materials and methods

## Data source and preprocessing

GSE32960, GSE70970, and GSE102349 datasets containing the expression profiles of NPC samples were accessed from the Gene Expression Omnibus (GEO) database (24), where GSE32960 and GSE70970 include miRNA expression data and GSE102349 includes mRNA expression data. We screened the samples of GEO datasets according to the following criteria: 1) remove samples without survival time and survival status; 2) convert probes into gene symbols; 3) remove a probe matching to multiple genes; 4) select the averaged expression of a gene with multiple probes; 5) for miRNA data, only human samples were remained. After preprocessing, 312 NPC and 18 normal samples were remained in the GSE32960 dataset; 253 NPC and 10 normal samples were remained in the GSE70970 dataset; and 88 NPC samples were remained in the GSE102349 dataset.

## Identification of NPC-associated differentially expressed miRNAs

In the GSE32960 dataset, differentially expressed miRNAs (DEmiRNAs) were screened by Limma R package (25) from NPC and normal samples under conditions of $P < 0.05$ and $|\log2(\text{fold change, FC})| > 1.2$. Gene modules were identified by weighted correlation network analysis (WGCNA) (26). Firstly, samples were clustered and the co-expression network was constructed. A scale-free network was ensured under scale-free R2 = 0.85, and soft threshold (power) = 3 was determined. The co-expression network was then converted to the adjacent matrix and was further converted to the topological overlap matrix (TOM). Subsequently, we clustered genes by the dynamic cutting method and average-linkage hierarchical clustering based on TOM. Eigengenes were calculated for each gene, and gene modules were clustered and merged under parameters of deepSplit = 2, and minModuleSize = 60, height = 0.25. The module–trait relationships were assessed by Pearson correlation analysis. By overlapping the miRNAs in the NPC-associated gene modules and DEmiRNAs, NPC-associated miRNAs were determined.

## Construction of a miRNA-based risk model

Random grouping of samples from the GSE32960 dataset into training group and test group at a ratio of 3:2 was performed. Two-group differences were assessed using Student's t test. Univariate Cox regression analysis screened NPC-associated miRNAs from the training group. MiRNAs with $P < 0.01$ were selected as prognostic miRNAs. Multivariate analysis (stepAIC) was introduced to measure the coefficients of prognostic miRNAs. Then, the miRNA-based risk model was defined as: risk score = $\Sigma$(coef i*expression i), where coef indicates the coefficients of miRNAs and i represents miRNAs.

## Evaluation and optimization of the risk model

Each sample obtained a risk score calculated by the risk model. The median risk score was employed in dividing samples into low-risk and high-risk groups. Receiver operating characteristic (ROC) curve analysis was used to predict the efficiency of the risk model in predicting overall survival through timeROC R package (27). The prognosis difference of two risk groups was studied by Kaplan–Meier survival analysis. Univariate and multivariate Cox regression models were used to analyze the hazard ratio of risk type. A nomogram was used to optimize the clinical use of the risk model with the rms package.

## Construction of a mRNA-based prognostic model

First of all, the potential target genes of miRNAs were predicted by different online tools and software including microT (28), miRanda (29), mircode (30), miRDB (31), miRmap (32), miRtarbase (33), PicTar (34), PITA (35), TargetMiner (36), TargetScan (37), RNA22 (38), and starbase (39). The target genes predicted by at least six tools were remained as key potential target genes. The WebGestaltR package (40) was utilized to annotate the enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Gene Ontology (GO) terms. Then, we used the key target genes to establish the mRNA-based prognostic model. Prognostic target genes in 70% samples of the GSE102349 dataset were screened by univariate Cox regression analysis. Random sampling with 70% samples in the GSE102349 dataset was performed for 1,000 times corresponding with univariate analysis. The top five frequent target genes from the results of 1,000 times of univariate analysis were selected as the final target genes for constructing the mRNA-based prognostic model (defined by the same formula with the miRNA risk model).

## Endocrine metabolism analysis

There were 31 SECRETORY_PATHWAYS screened from the KEGG PATHWAY Database (https://www.genome.jp/kegg/pathway.html). SECRETORY_PATHWAYS scores in the GSE102349 dataset were calculated by ssGSEA (41). There were 26 SECRETORY-related genes obtained from KEGG (https://www.genome.jp/dbget-bin/www_bfind_sub?mode=bfind&max_hit=1000&locale=en&serv=kegg&dbkey=genes&keywords=

Secretory&page=1). Pearson analysis was conducted to determine the correlation between SECRETORY_PATHWAYS scores/SECRETORY-related genes and RiskScore.

## Analysis of immune characteristics

We used ssGSEA (41) to estimate the immune cell proportion in two risk groups. The gene sets of 22 immune-related cells, innate immune response, and adaptive immune response were obtained from Charoentong et al. (42). The ESTIMATE algorithm (43) was employed to measure immune cell infiltration and stromal cell infiltration. The ESTIMATE score represents the combined score of immune score and stromal score. MCP-counter (44) was used to evaluate 10 immune-related cells including monocytic lineage, CD3 + T cells, NK cells, CD8+ T cells, endothelial cells, cytotoxic lymphocytes, B lymphocytes, neutrophils, fibroblasts, and myeloid dendritic cells based on the expression matrix. The immune checkpoint genes were downloaded from a previous study (42).

## Assessment of biological pathways

Biological pathways were accessed from "h.all.v7.4.symbols.gmt" downloaded from the Molecular Signatures Database (MSigDB) (45). There were 13 tumor-related genes obtained from a previous research (46), which are classic cancer pathways, involved in the development and progression of cancer, including DNA damage repair (DDR), epithelial–mesenchymal transition (EMT), cell cycle, mismatch repair, CD8 T effector, FGFR3, nucleotide excision repair, base excision repair, DNA replication, homologous recombination, and WNT target. ssGSEA was performed to determine the enrichment score of biological pathways. The relation of risk score with pathways was inspected with Pearson correlation analysis.

## Predicting the response to immunotherapy and chemotherapy

We conducted SubMap analysis (47) to compare the expression profiles between GSE102349 and IMvigor210. IMvigor210 contains the expression profiles of patients with metastatic urothelial carcinoma treated by PD-L1 inhibitors (48). The higher similarity of samples in GSE102349 with complete response (CR) and partial response (PR) groups in IMvigor210 suggested that the samples were more sensitive to anti-PD-L1 treatment. The TIDE (http://tide.dfci.harvard.edu/) algorithm (49) was employed to predict the escape and response to immune checkpoint inhibitors, according to the score of T-cell dysfunction and exclusion, the enrichment of immunosuppressive cells. The sensitivity of two risk groups to chemotherapeutic drugs was estimated using the pRRophetic R package (50).

The performance of the prognostic model was further examined in immunotherapy datasets including IMvigor210, GSE135222, and GSE78220. GSE135222 contains non-small cell lung cancer patients treated with immune checkpoint inhibitors (51). GSE78220 includes patients suffering from metastatic melanoma treated by anti-programmed cell death protein 1 (anti-PD-1) therapy.

## Cell culture

After resuscitating NPC cells and normal cells, they were placed in 1,640 and 5A cell medium containing 10% fetal bovine serum, respectively, at 37°C, 5% $CO_2$ concentration, and constant temperature to around 80%–90%, then passed and spread on a plate.

## RT-qPCR

Total RNA was extracted from cells by RT-qPCR, and cDNA was synthesized by reverse transcription. After reverse transcription, samples were added according to the experimental instructions. The reaction conditions of RT-qPCR were 95°C for 30 s. 95°C, 5 s; 60°C, 30 s, 40 cycles. The mRNA expressions of miR-142-3p and miR-93 in the experimental group and control group were analyzed.

## Statistical analysis

The statistical methods used in this study were performed in R software (v4.2.0). The Sangerbox platform (52) was used to provide an assistant in data analysis. The log-rank test was applied in survival analysis. Difference between two groups was examined by the Wilcoxon test. The Kruskal–Walls test was used to test the difference among over two groups. $P < 0.05$ was determined as statistically significant.

# Results

## Identification of DEmiRNAs related to NPC based on WGCNA

To begin with, we assessed the expression difference of miRNAs between normal and tumor samples in the GSE32960 dataset. The results showed that 332 miRNAs were differentially expressed between normal and tumor groups, including 168 upregulated and 164 downregulated miRNAs in tumor groups (Figure 1A; Table S1). Then, we applied WGCNA to cluster samples and dig out key gene modules based on DEmiRNAs. To ensure the co-expression network to be a scale-free network, the Pearson coefficient was selected as 0.85 and the power of the soft threshold was determined as 3 to construct an adjacency matrix (Figures 1B–D). Next, a topology overlap matrix (TOM) was generated based on the adjacency matrix and the genes were divided into different modules using the Dynamic Tree Cut algorithm (Figure 1E). Four modules were finally determined after merging the adjacent modules. Then, we analyzed the correlation of four modules with different sample groups. As a

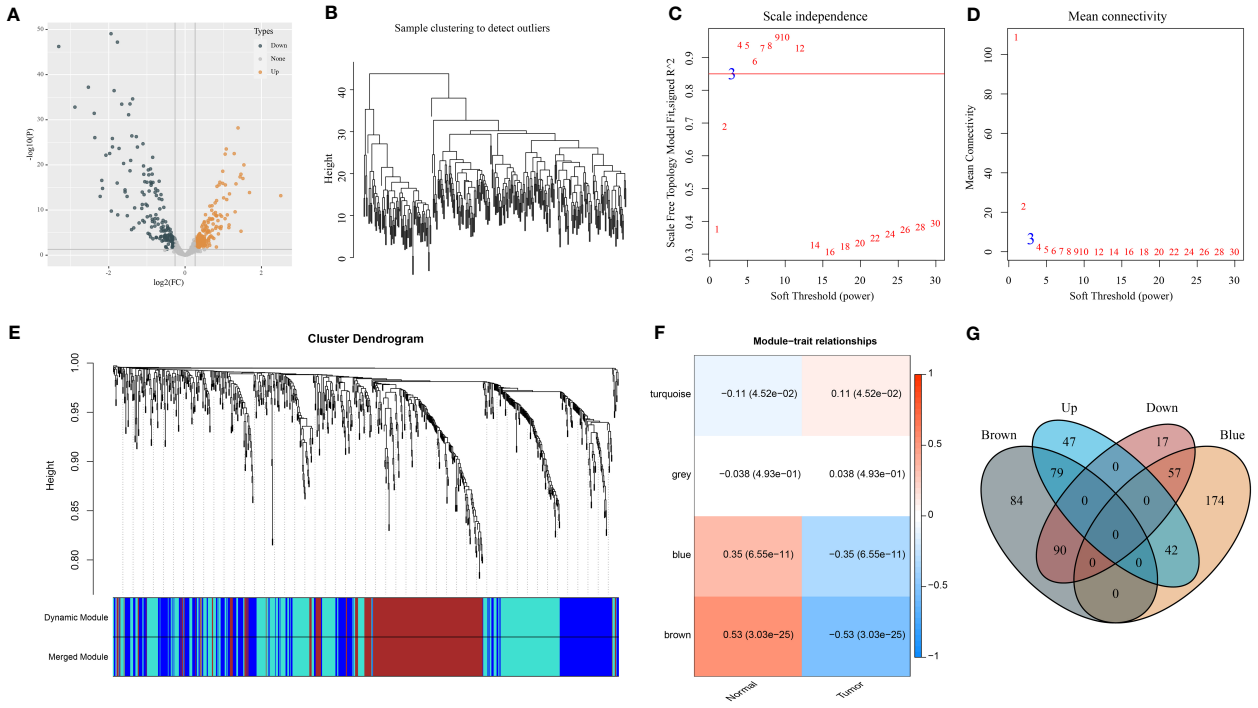



FIGURE 1
Identification of NPC-related DEmiRNAs in the GSE32960 dataset. **(A)** Volcano plot of 332 DEmiRNAs. **(B)** Clustering dendrogram of 312 samples. **(C, D)** Under different soft thresholds (power), we analyzed scale independence and mean connectivity. **(E)** Clustering dendrogram of DEmiRNAs based on TOM and dynamic cut methodology. **(F)** The relationships of modules with normal and tumor groups. **(G)** Venn plot of DEmiRNAs and miRNAs of brown and blue modules.

result, brown and blue modules were found to be evidently associated with sample groups and they manifested opposite correlations with two groups (Figure 1F). Specifically, blue and brown modules were negatively correlated with the tumor group (R = -0.35 and -0.53, respectively). The Venn plot was constructed to describe the intersection between DEmiRNAs and miRNAs of blue and brown modules (Figure 1G). There were 99 DEmiRNAs including 42 upregulated and 55 downregulated found in the blue module. There were 169 DEmiRNAs including 79 upregulated and 90 downregulated found in the brown module. The above total 268 DEmiRNAs were determined as potential miRNAs associated with NPC.

## Establishment and verification of a miRNA-related risk model

We randomly divided 312 NPC samples of the GSE32960 dataset at a ratio of 3:2 into the training and test groups (Table S2). We used the training group to screen prognostic miRNAs from 268 DEmiRNAs according to the univariate Cox regression model. The analysis identified two miRNAs (hsa-miR-142-3p and hsa-miR-93) that were significantly associated with the overall survival ($P < 0.01$). Based on the multivariate coefficients of two miRNAs by stepAIC analysis, we established the risk model defined as follows (Figure S1A).

Risk score = -0.835*(hsa-miR-142-3p) + 0.85*(hsa-miR-93)

Moreover, the RT-qPCR analysis results showed that miR-142-3p expression was downregulated and miR-93 was upregulated in five NPC cell lines in comparison with normal NP69 cells (Figures S1B, C).

To validate the performance of the risk model, we calculated the risk score for each tumor sample in the training and test groups. The AUC for 1-, 3-, and 5-year survival derived from the ROC curve analysis was 0.56, 0.70, and 0.70 in the training group and 0.96, 0.71, and 0.71 in the test group, respectively (Figures 2A, B). Furthermore, based on the median risk score, tumor samples were classified into low-risk and high-risk groups. The two risk groups had distinct overall survival (OS) in the training and test groups, as shown by Kaplan–Meier survival analysis ($P = 0.00035$ and $P = 0.013$, respectively, Figures 2A, B).

In the total GSE32960 dataset, the risk model still showed good performance in predicting overall survival (Figure 2C). In addition, we evaluated the effectiveness of the risk model in different survival times including recurrence-free survival (RFS), disease-free survival (DFS), and metastasis-free survival (MFS). High-risk and low-risk groups exhibited differential prognosis of DFS and MFS ($P < 0.0001$) but showed no significant difference in RFS classification ($P = 0.063$, Figure 2C). In another independent dataset (GSE70970), the risk model also showed an effect classification for both OS and DFS ($P = 0.017$ and $P = 0.022$, respectively) (Figure 2D). Moreover, we verified the effectiveness of the risk model in the groups of different clinical characteristics. In addition to the female group and N0 group, the risk model was sufficient to distinguish samples into

10.3389/fendo.2023.1174911

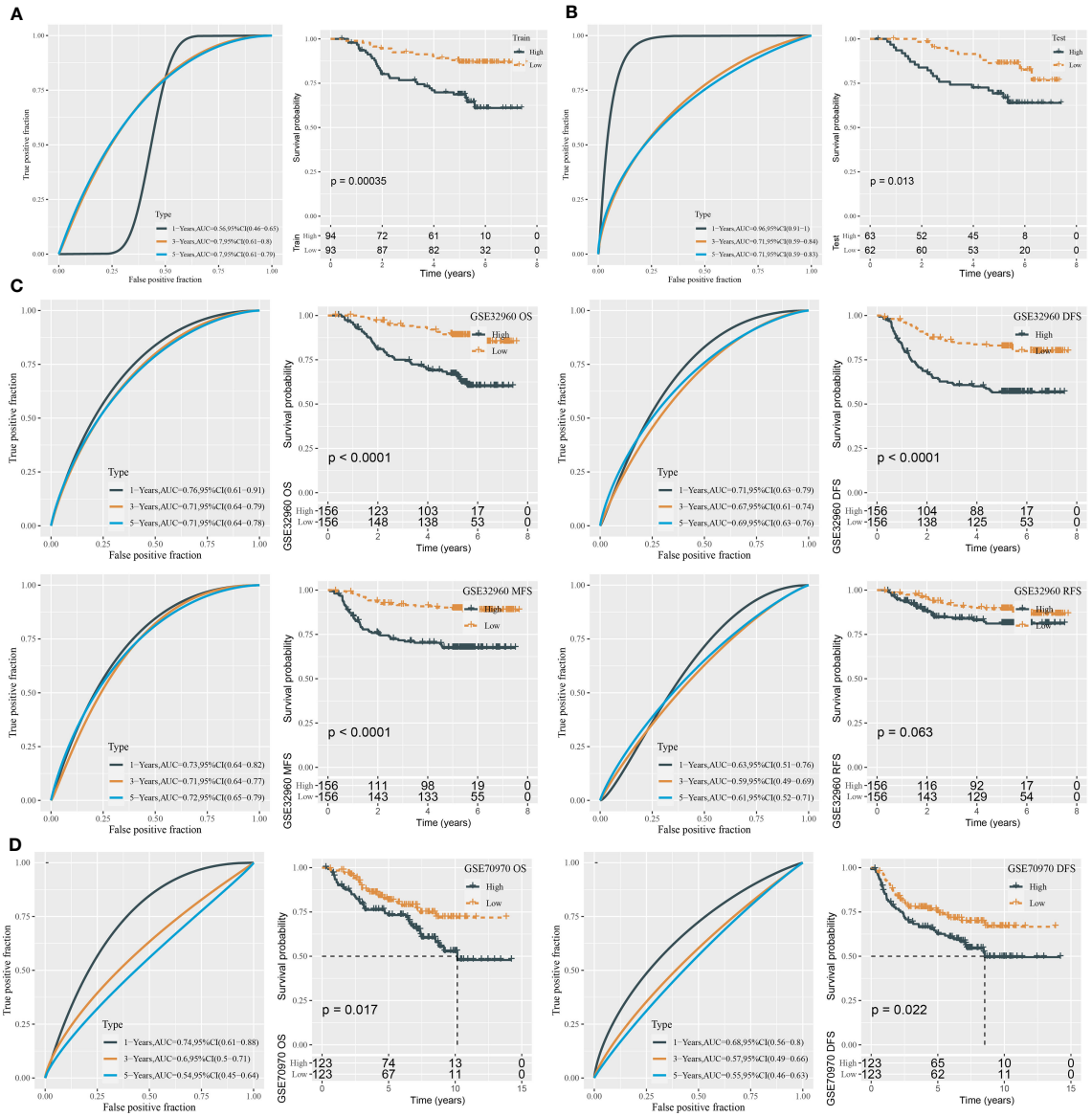



FIGURE 2
ROC analysis and survival analysis for evaluating the performance of the miRNA risk model. **(A)** ROC curves and survival curves in the training group. **(B)** ROC curves and survival curves in the test group. **(C)** ROC curves and survival curves of DFS, OS, MFS, and RFS in the GSE32960 dataset. **(D)** ROC curves and survival curves of OS and DFS in the GSE70970 dataset.

different risk levels in the groups of age >45, age ≤45, male, T1–T2, T3–T4, N1–N3, I–II, and III–IV (Figure 3).

## Boosting the prediction efficiency of the miRNA-related risk model by constructing a nomogram

Using univariate and multivariate Cox regression analyses, we evaluated the relation of clinical features and risk type with prognosis. The result displayed that only gender and risk type were independent risk factors in both univariate and multivariate analyses (Figures 4A, B). Gender had a hazard ratio (HR) of 2.2 in both univariate and multivariate analyses (P = 0.019 and P = 0.02, respectively). Risk type had HR of 3.7 (P = 1.1e-6) and 3.6 (P = 2.1e-6) in univariate and multivariate analyses, respectively. Therefore, we included gender and risk score to construct the nomogram (Figure 4C). Compared with gender, risk score contributed the more total points in the nomogram. Calibration curve analysis suggested that the predicted 1-, 3-, and 5-year survival was highly accordant with the observed survival (Figure 4D). In addition, decision curve analysis was implemented to evaluate the benefit that patients may obtain from gender, risk score, and nomogram. As a result, nomogram performed a more favorable performance

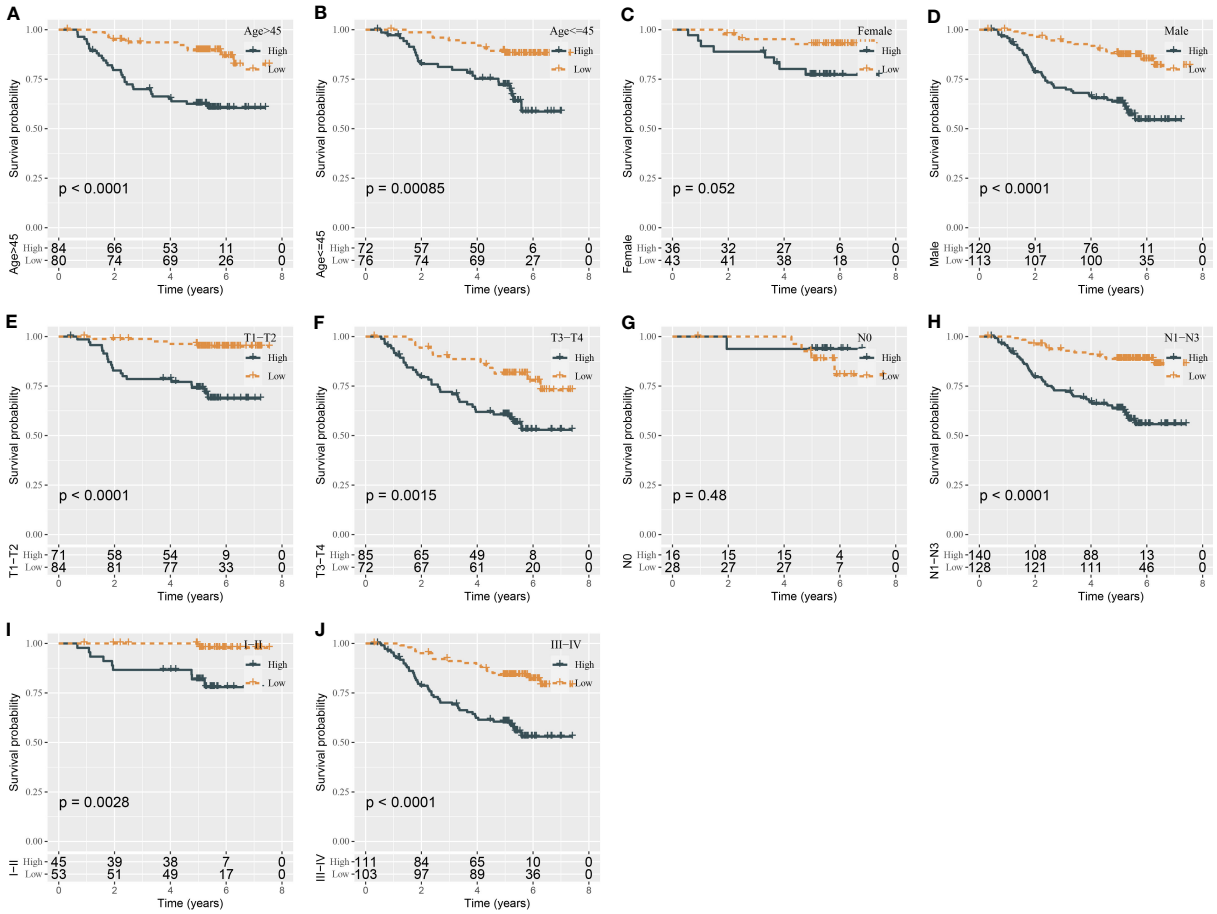



FIGURE 3
Kaplan–Meier survival analysis of two risk groups in the samples with different clinical characteristics **(A–J)**.

than gender and risk score (Figure 4E). Consequently, the nomogram based on gender and risk score was more efficient to predict the prognosis of NPC patients.

## Construction of miRNA–mRNA competing endogenous RNA networks

In the above sections, we identified two key miRNAs (hsa-miR-142-3p and hsa-miR-93) that had a close relation with NPC prognosis. To understand the potential molecular mechanism of the miRNAs, we applied 12 tools (starbase, PITA, TargetMiner, miRanda, microT, miRmap, miRtarbase, mircode, TargetScan, RNA22, miRDB, PicTar) to predict the potential targets of two miRNAs. The results output 39 target genes of hsa-miR-142-3p and 86 target genes of hsa-miR-93, which were visualized in ceRNA networks (Figure 5A). KEGG analysis demonstrated the involvement of these target genes in mitophagy and autophagy (Figure 5B). The top 10 enriched terms of cellular component and molecular function, as well as biological process, were visualized using GO function analysis (Figures 5C–E). For example, biological process terms of positive regulation of amyloid–beta metabolic process, amyloid–beta formation, and negative regulation of phosphatase activity were enriched (Figure 5C). Cellular component terms of clathrin-coated pit and trans-Golgi network membrane were enriched (Figure 5D). Molecular function terms of clathrin adaptor activity and clathrin heavy chain binding were enriched (Figure 5E).

## Establishing a mRNA prognostic model based on key target genes of hsa-miR-142-3p and hsa-miR-93

We predicted a total of 125 potential target genes of hsa-miR-142-3p and hsa-miR-93. Then, we used the univariate Cox regression model to analyze the relation between target genes and overall survival. Random sampling for 1,000 times from the samples of the GSE102349 dataset was performed. The target genes closely related to overall survival were remained ($P < 0.05$) and were ranked by the occurring frequency from 1,000-times analysis. The top five frequent target genes were selected as key target genes for establishing the mRNA

10.3389/fendo.2023.1174911

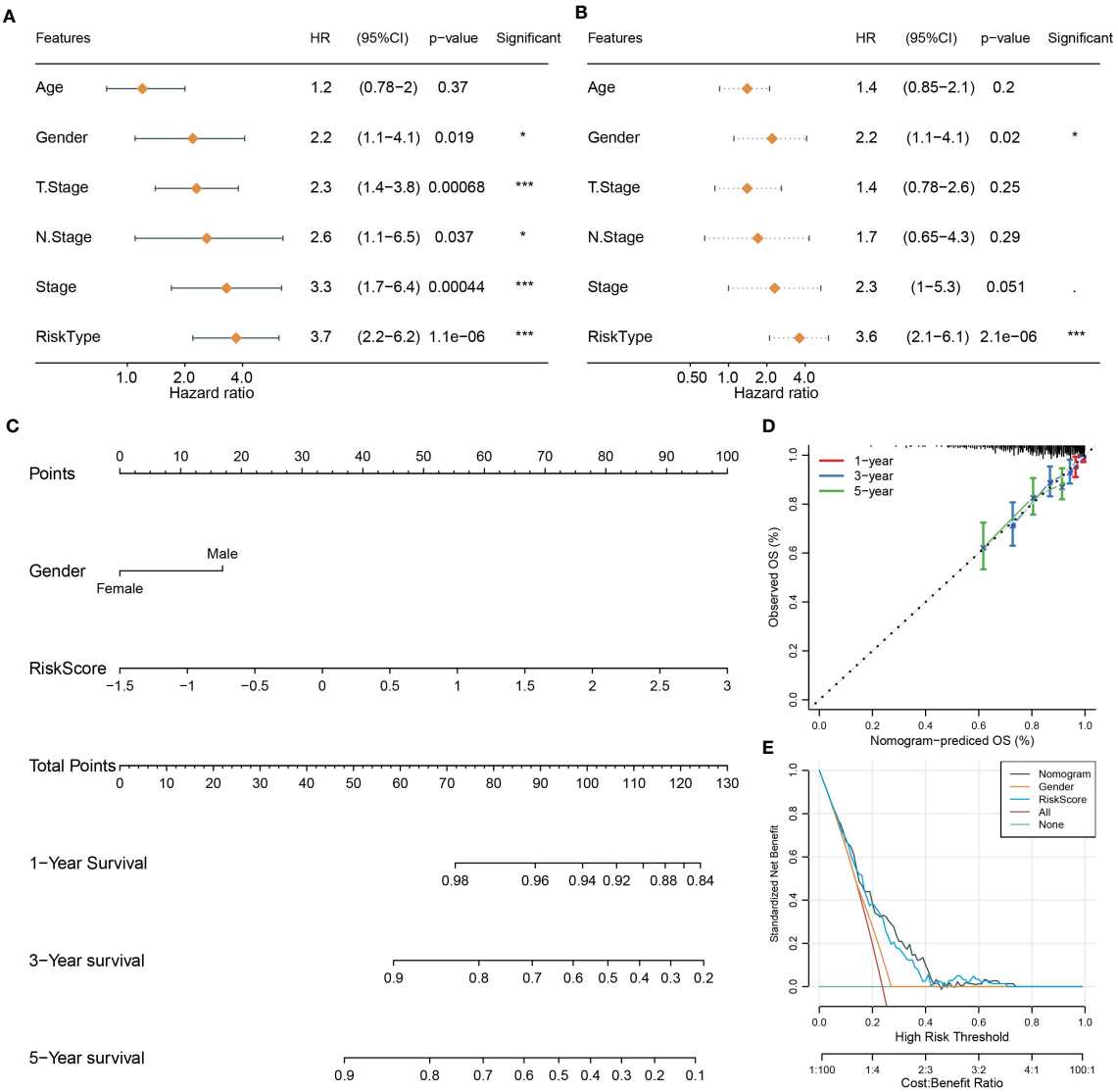



FIGURE 4
Constructing a nomogram based on clinical features and risk score. **(A, B)** Univariate **(A)** and multivariate Cox regression analyses of clinical features and risk type. **(C)** The nomogram based on gender and risk score for predicting 1-, 3-, and 5-year survival. **(D)** Calibration curve of 1-, 3-, and 5-year survival. **(E)** DCA of nomogram, gender, and risk score. *P < 0.05, ***P < 0.001.

prognosticmodel (Figure 6A). The coefficients of five target genes were calculated by multivariate analysis. Finally, the prognostic model was defined as the following: mRNA risk score = 1.333* E2F1 - 1.766*KCNJ8 + 1.075*SUCO - 1.030* HECTD1 - 0.340*KIF23.

The risk score for each sample in the GSE102349 dataset was determined with the mRNA prognostic model. ROC curve analysis showed that the model was efficient in predicting 1-, 3-, and 5-year survival, with AUCs of 0.87, 0.81, and 0.79, respectively (Figure 6B). The median risk score value was used in classifying NPC samples into two groups, high-risk and low-risk groups. Kaplan–Meier survival curves of two risk groups showed that they had an apparently different prognosis (P = 0.0018, Figure 6B). Hence, the five target genes could be validated by the above results to be closely involved in the prognosis.

## Immune characteristics of high- and low-risk groups

The tumor microenvironment plays a central role in antitumor response and immunotherapeutic response. We used several tools to assess the immune cell component, as well as immune and stromal infiltration, and analyzed immune checkpoint genes for their expression levels. Two risk groups showed a significantly different immune microenvironment. We estimated an ssGSEA enrichment score of 28 immune-related cells and found that 25 of them had a differential enrichment score, such as myeloid-derived suppressor cells (MDSCs), activated CD8 T cells, regulatory T cells, natural killer cells, activated B cells, and macrophages (Figure 7A). Most immune cells showed a higher abundance in the group with a

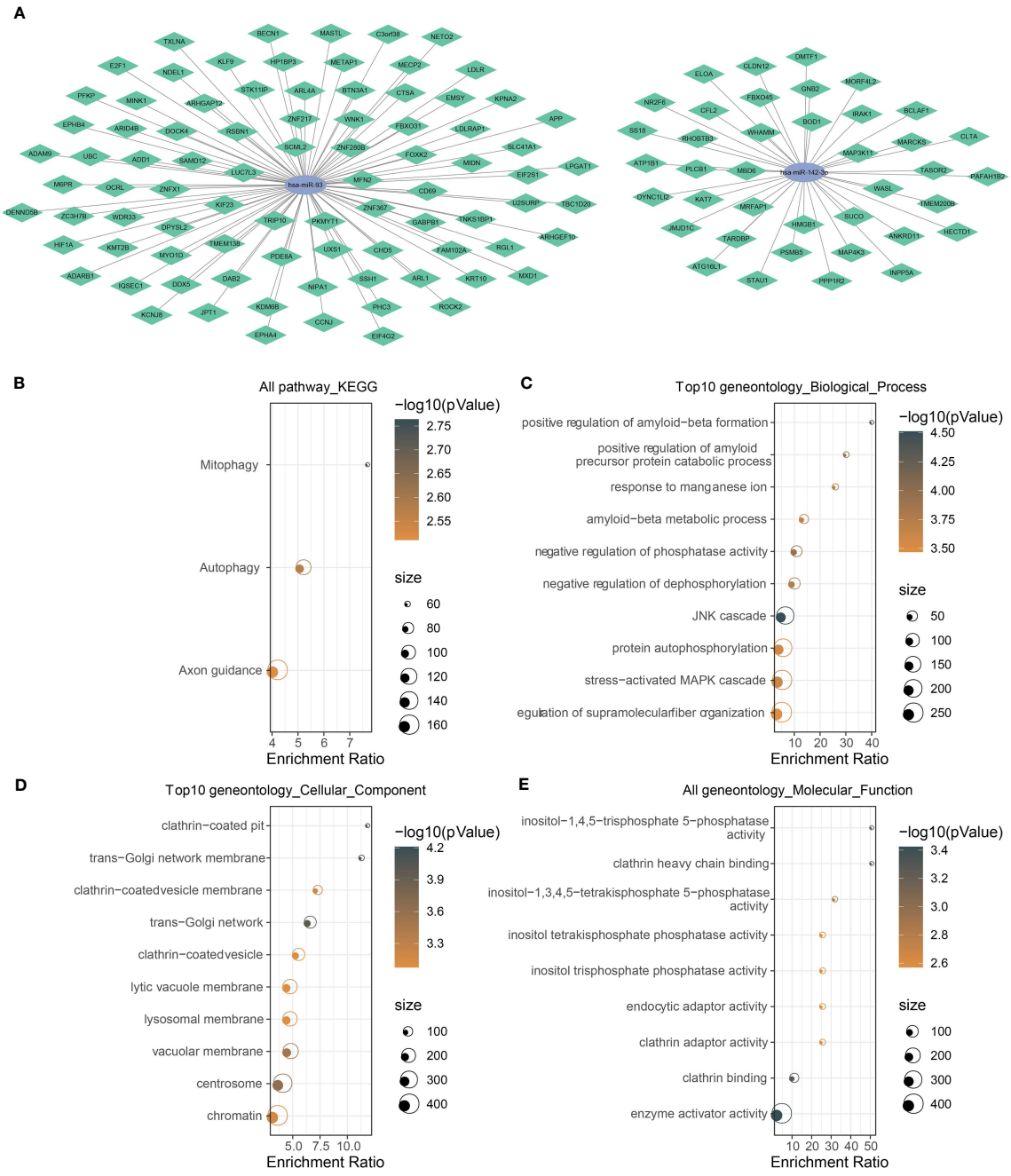



FIGURE 5
Analysis of the target genes of has-miR-142-3p and has-miR-93. **(A)** The mRNA−miRNA ceRNA networks. Green rhombus indicates target mRNAs, and ellipse indicates miRNAs. **(B)** KEGG and **(C−E)** GO functional analyses of potential target mRNAs. The color of dots indicate the significance of P values, and the dot size indicates the gene counts.

low risk. In addition, ssGSEA also revealed higher enrichment of both adaptive and innate immune response scores in the low-risk group ($P < 0.0001$, Figure 7B). The ESTIMATE algorithm was used to evaluate stromal and immune infiltration of two groups, and not surprisingly, the low-risk group showed both higher immune score and stromal score than the high-risk group (Figure 7C, $P < 0.0001$). Moreover, MCP-counter was employed to dig out similar results with ssGSEA. A total of 10 types of immune-related cells all showed a higher enrichment score in the low-risk group (Figure 7D). Immune checkpoint expression levels also had an extreme difference in two risk groups that most of immune checkpoints, such as CTLA4, TIGIT, LAG3, and PCDC1, were more highly expressed in the low-risk group than in the high-risk group (Figure 7E). Immune checkpoints' differential expression may

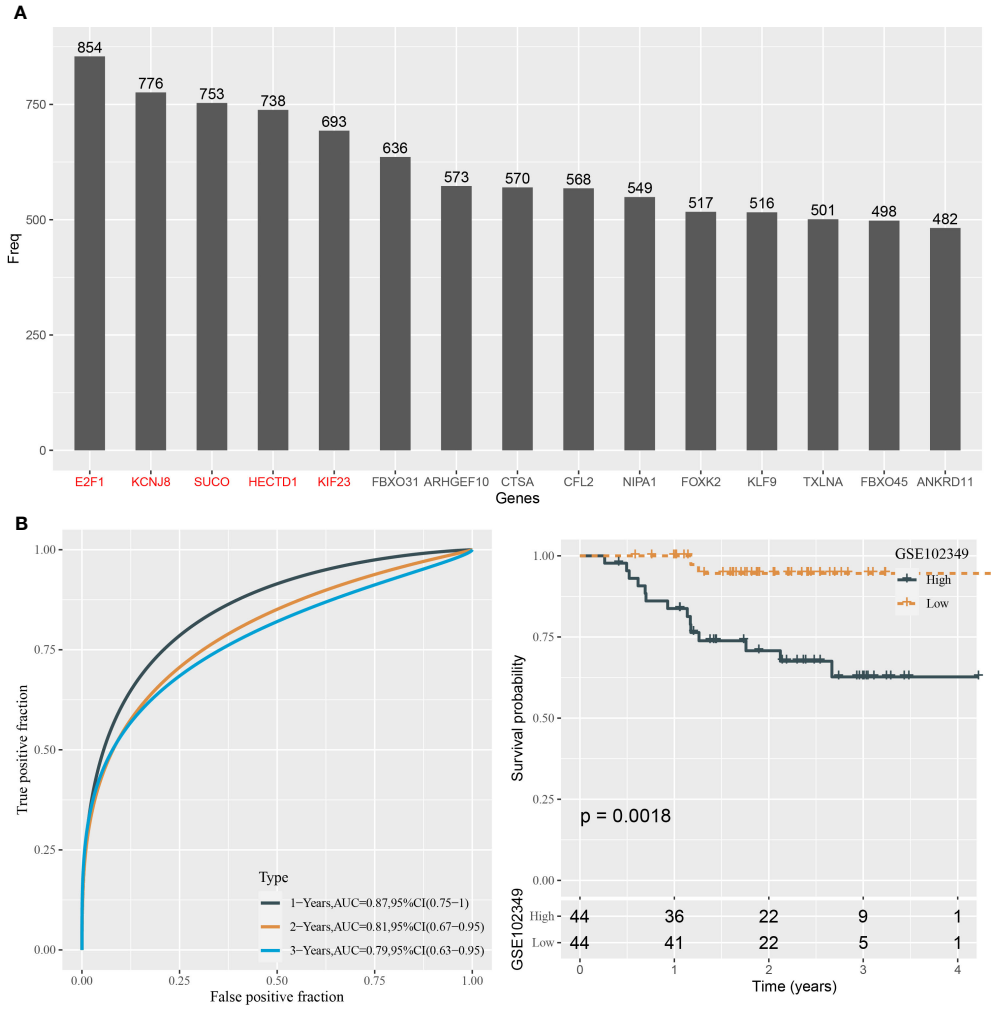



FIGURE 6
Construction of the prognostic model based on the target genes in the GSE102349 dataset. **(A)** The top 15 target genes associated with prognosis from 1,000-times random sampling. **(B)** ROC analysis and survival analysis of the five-gene prognostic model.

contribute to different antitumor immune responses. The distinct immune microenvironment suggested that the five prognostic genes may play critical roles in immune modulation.

## Analysis of biological pathways, immunotherapeutic response, and drug sensitivity in two risk groups

To further understand the molecular mechanisms contributing different prognoses of two risk groups, the enrichment score of pathways from the "h.all.v7.4.symbols.gmt" file was calculated using ssGSEA. There were 30 pathways differentially activated in the two risk groups (Figure 8A). Immune-related pathways, for instance, interferon alpha response, interferon gamma response, inflammation response, IL2-STAT5 signaling, IL6-JAK-STAT3 signaling, and complement, were significantly more activated in the low-risk group, which was consistent with the result of immune analysis. Cell cycle-related pathways such as E2F targets, MYC target V2, MYC target V1, and G2M checkpoint were less enriched in the low-risk group. In addition, some oncogenic pathways were more activated in the high-risk group, such as Wnt signaling and Hedgehog signaling. Notably, apoptosis was more enriched in the low-risk group, which was associated with good prognosis. Risk score also manifested significant correlations with the above pathways (Figure S2).

Gene sets of 13 tumor-related pathways were obtained from a previous study, and their enrichment scores were calculated using ssGSEA. Pearson correlation analysis uncovered a negative association of risk score with DNA repair-related pathways including DDR (R = -0.41), base excision repair (R = -0.25), nucleotide excision repair (R = -0.33), homologous recombination (R = -0.33), and mismatch repair (R = -0.33) (Figure 8B).

Next, we evaluated the response of two risk groups to chemotherapy and immunotherapy. The similarity of expression profiles between GSE102349 and IMvigor210 (treated by PD-L1 inhibitors) datasets was shown by SubMap analysis. Higher similarity between two datasets indicates higher sensitivity to PD-L1 inhibitors. The results presented that the low-risk group had a higher similarity with CR and PR groups ($P < 0.05$, Figure 8C), implying that the low-risk group could obtain more benefit from

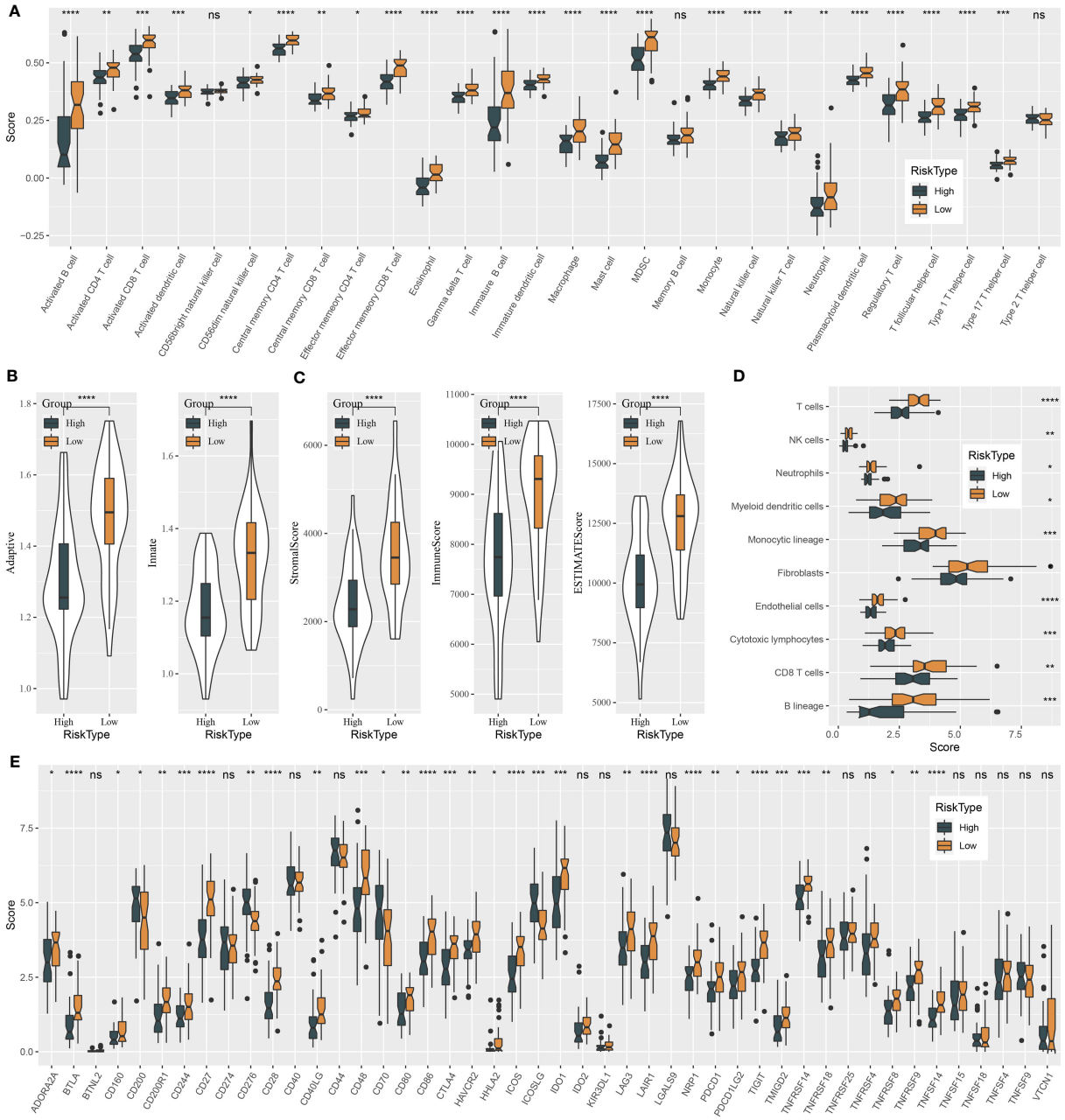



FIGURE 7
Immune characteristics of high-risk and low-risk groups in the GSE102349 dataset. **(A)** The ssGSEA score of 22 immune-related cells in two risk groups. **(B)** The ssGSEA score of adaptive and innate immune response in two risk groups. **(C)** ESTIMATE analysis of immune infiltration and stromal infiltration. **(D)** MCP-counter analysis for estimating the enrichment of 10 immune-related cells. **(E)** Expressions of immune checkpoint genes in two risk groups. ns, not significant. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.

immunotherapy. Furthermore, the TIDE algorithm was implemented to predict immune escape to immune checkpoint inhibitors. Two risk groups showed no significant difference of TIDE score (Figure 8D). However, the low-risk group exhibited more severe T-cell dysfunction than the high-risk group in which enrichment of MDSCs and M2 tumor-associated macrophages (TAM) was higher, contributing to its higher T-cell exclusion. In the response to chemotherapeutic drugs, we screened a total of 103 drugs including 46 drugs such as YM155 and vinorelbine sensitive to high-risk groups and 57 drugs such as sunitinib and temsirolimus sensitive to the low-risk group (Figure 8E).

## The analysis of endocrine metabolism

Abnormal endocrine metabolism is one of the complications of tumor treatment. As shown in Figure 9A, endocrine-related genes, such as MON1B, SCAMP2, and FAM20A, were negatively

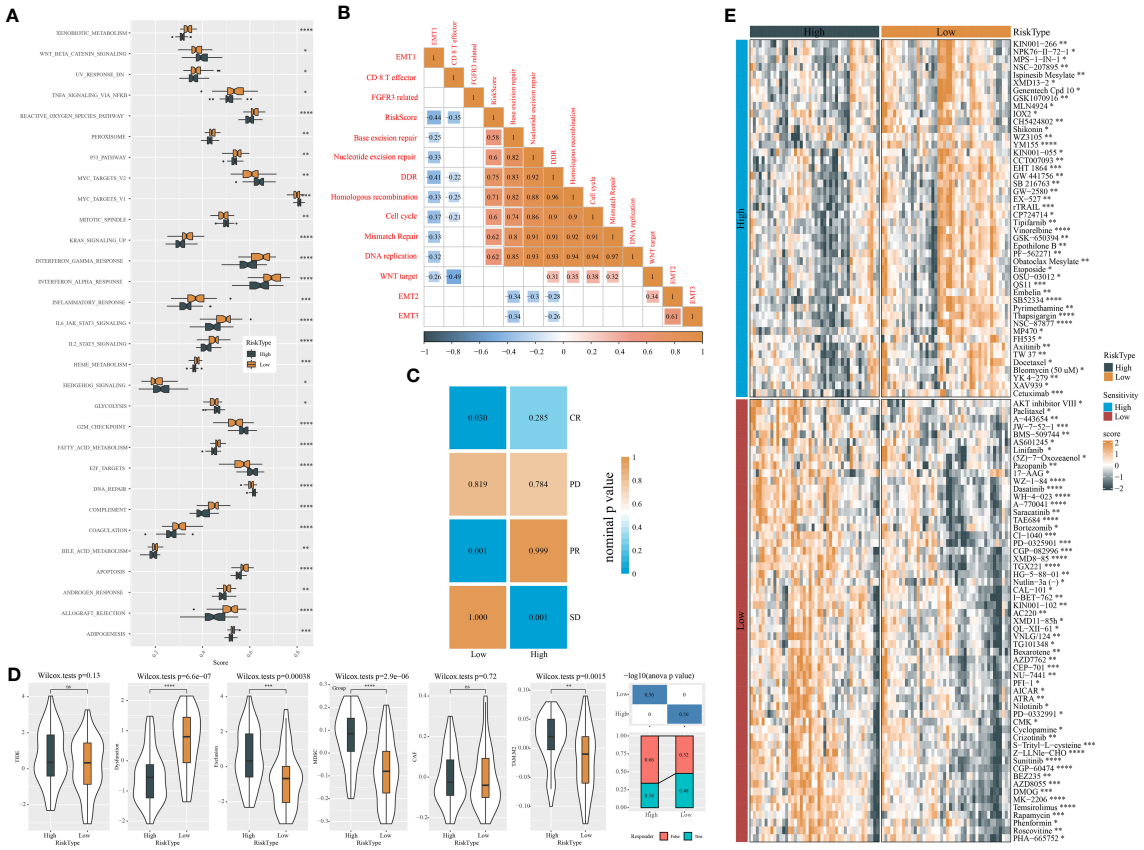


FIGURE 8
Analysis of biological pathways, response to immunotherapy, and drug sensitivity of two risk groups in the GSE102349 dataset. **(A)** Screening differentially enriched pathways between two risk groups based on ssGSEA. **(B)** Correlation analysis of risk score with 13 tumor-related pathways. Orange and blue indicate positive and negative correlations, respectively. **(C)** SubMap analysis of expression data of GSE102349 and IMvigor210 (anti-PD-L1 treatment) datasets. **(D)** TIDE analysis of two risk groups for assessing immune escape and response to immunotherapy. **(E)** The sensitivity to chemotherapeutic drugs predicted by the pRRophetic package. *P < 0.05, **P < 0.01, ***P < 0.001, ****P<0.0001.

associated with RiskScore; SCAMP3 was positively correlated with RiskScore. Pearson analysis between endocrine pathways and RiskScore showed that most endocrine pathways were negatively related to RiskScore (Figure 9B).

## The performance of the mRNA prognostic model in immunotherapy datasets

To further examine the exhibition of the five-gene prognostic model, we introduced three independent datasets containing patients administrated by immunotherapy (IMvigor210, GSE135222, and GSE78220). Using the same calculation method, we measured the risk score of each patient in three datasets. In the IMvigor210 dataset, the high-risk group displayed apparently worse prognosis than the low-risk group (P < 0.0001, Figure 10A). The prediction efficiency was evaluated by ROC, with AUCs of 0.61, 0.66, and 0.64 at 0.5-, 1-, and 1.5-year survival, respectively (Figure 10A). In addition, we analyzed the proportion of different response groups in two risk groups. The low-risk group had a higher proportion of CR and PR groups (11% and 20%, respectively), compared with the high-risk group (6% and 9%, respectively). The CR and PR groups also showed a lower risk score than PD and SD groups. In the GSE135222 and GSE78220 datasets, we observed correspondent results (Figures 10B, C). The GSE135222 dataset showed AUCs of 0.82 and 0.85 at 0.5- and 1-year survival, respectively. The low-risk group exhibited a markedly higher percentage of CR/PR patients compared with the high-risk group (50% versus 8% in low- versus high-risk groups). The GSE78220 dataset showed AUCs of 0.74 and 0.71 at 1- and 2-year survival, respectively. Expectedly, the CR and PR groups were more accumulated in the low-risk group (21% and 43%) compared with the high-risk group (8% and 31%). Moreover, the CR/PR group showed a lower risk score than the PD/SD group in both GSE135222 and GSE78220 datasets, but the difference was not significant in the GSE78220 dataset. The above observation suggested that patients with a low risk score were more sensitive to immunotherapy and could attain longer survival.

# Discussion

MiRNAs are considered as potential therapeutic targets for NPC treatment, and till now abundant miRNAs have been unveiled to be

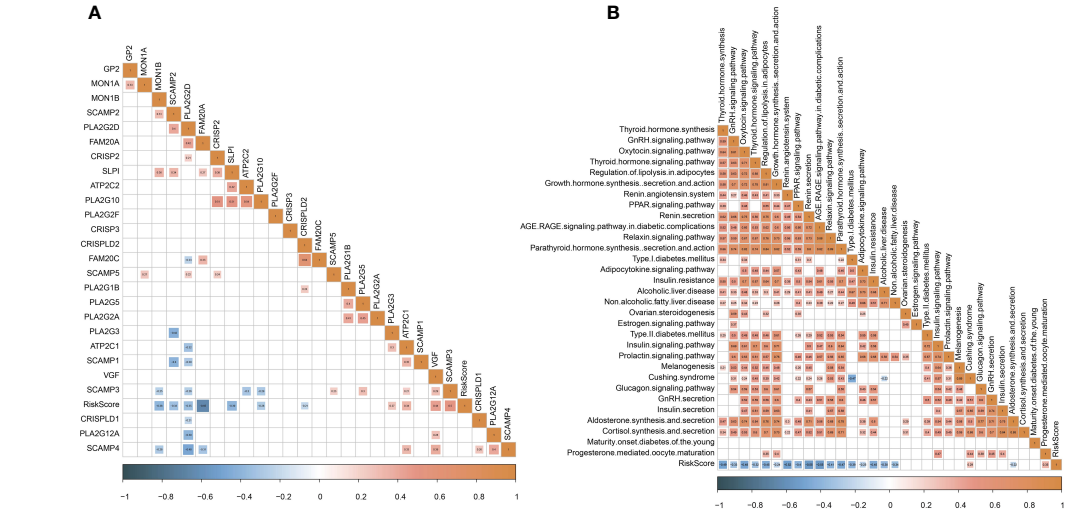



FIGURE 9
Analysis of endocrine metabolism. **(A)** Pearson analysis between endocrine-related genes and RiskScore. **(B)** Pearson analysis between endocrine pathways and RiskScore.

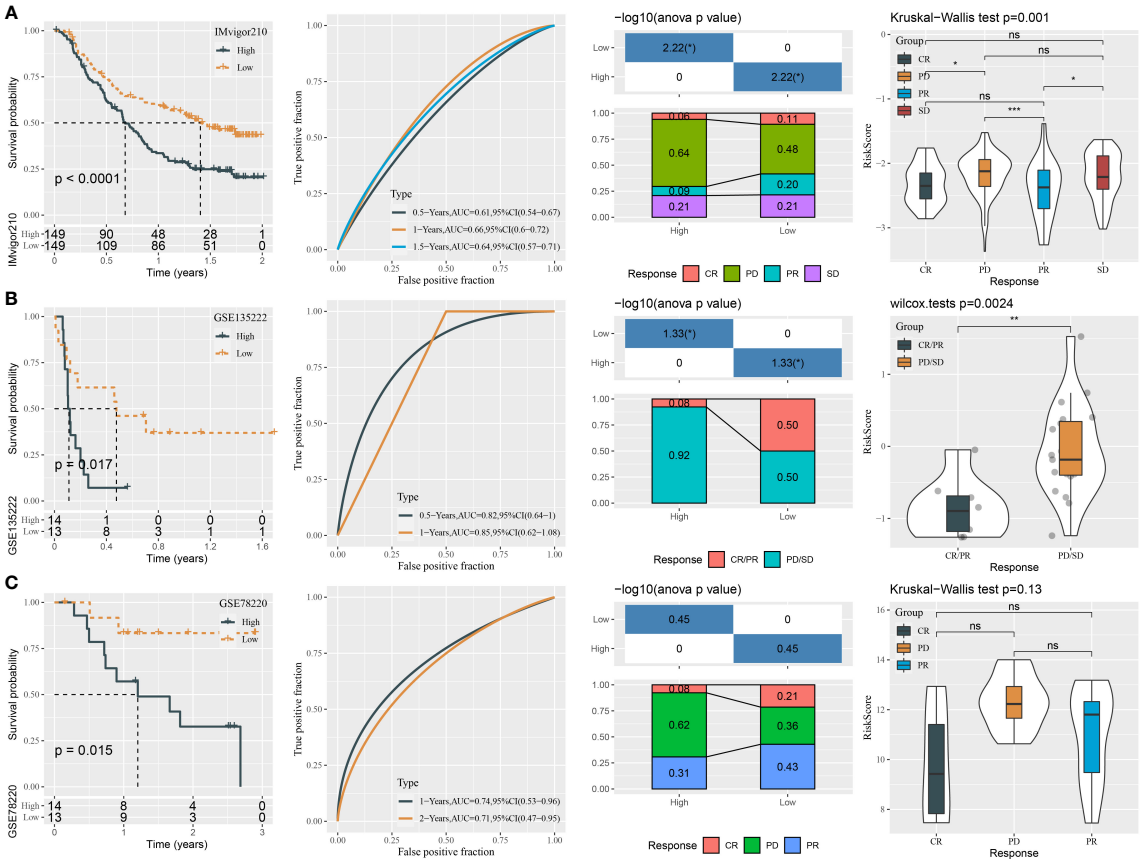



FIGURE 10
The performance of the mRNA prognostic model in three independent immunotherapy datasets. **(A–C)** ROC analysis, survival analysis, the distribution responder groups in two risk groups, and the risk score of responder groups in the IMvigor210 **(A)**, GSE135222 **(B)**, and GSE78220 **(C)** datasets. ns P>0.05, *P < 0.05, **P < 0.01, ***P < 0.001.

dysregulated in NPC patients (19). MiRNAs serve as importantly regulatory roles in gene expression and pathway modulation. In this study, we dug out the potential key miRNAs and mRNAs that were probably responsible for NPC development and progression. We interpreted the association between key miRNAs and mRNAs and decoded the relationships of prognostic miRNA-related mRNAs with tumor microenvironment, immunotherapy, and functional pathways.

To begin with, we deciphered the miRNA expression data and screened 332 aberrantly expressed miRNAs in NPC samples compared with normal samples. By using WGCNA, we further identified two key miRNAs (hsa-miR-142-3p and hsa-miR-93) strongly related to NPC prognosis and phenotype. We constructed a miRNA risk model based on hsa-miR-142-3p and hsa-miR-93. The risk model exhibited an intensive relation with NPC prognosis in two independent datasets (GSE32960 and GSE70970). Patients with a high risk showed significantly worse OS and DFS than the low-risk group. In addition, the risk model was also effective to predict OS in NPC samples with different clinical characteristics including ages, T stage, N1–N3 stage, and AJCC stage I–IV. These results demonstrated that hsa-miR-142-3p and hsa-miR-93 were highly responsible for NPC development.

Moreover, RiskScore was negatively correlated with endocrine genes, especially FAM20A, which had been important for endocrine-related tumors (53). The FAM20 family of kinases is a newly discovered class of secreted kinases that are capable of phosphorylating secreted proteins and proteoglycans (54). FAM20A may play a more complex role in gliomas, as correlations between FAM20A genes and low-grade gliomas have been found (55). Combining the known literature and the results of this study, it is suggested that endocrine gene FAM20A may be closely related to NPC.

The two key miRNAs have been reported in the contribution of other cancer types. For example, in esophageal squamous cell carcinoma (ESCA), hsa-miR-142-3p was identified as a prognostic biomarker (56). Non-small cell lung cancer (NSCLC) cells could be promoted by overexpression of miR-142-3p *via* interfering TGFβR1 expression (57). However, Dong et al. revealed that miR-142-3p suppressed the growth of human cervical cancer cells by attenuating HMGB1 expression levels (58). Moreover, Sharma et al. excavated that miR-142-3p functioned as a tumor-suppressive miRNA through modulating the expression of HMGA1, A2, B1, and B3 in human cervical cancer (59). miR-142-3p in different cancer types showed a discord in expression that miR-142-3p was suggested to play complicated roles (both promotive and suppressive) by interacting with specific pathways and genes in different cancers. In our study, the miR-142-3p level was significantly decreased in NPC samples, implying that high-expressed miR-142-3p may facilitate the progression of NPC.

The role of hsa-miR-93 has also been uncovered in different cancer types. For instance, a miRNA microarray result showed that miR-93 was downregulated in human colon cancer stem cells and overexpressing miR-93 strikingly inhibited cell proliferation and colony formation (60). In triple-negative breast cancer cells, cell migratory capability and invasive potential could be weakened by overexpression of mature miR-93-5p possibly by targeting WNK1 (61). In uterine cancer, the high miRNA-93 expression group had an evidently higher survival rate than the low miRNA-93 expression group (62). The results of the above studies were accordant with our study that miRNA-93 expression was elevated in the NPC group compared with the normal group.

To clarify the potential mechanisms of the two miRNAs in NPC, their potential target genes (mRNAs) were predicted by utilizing 12 different tools to build ceRNA networks. As a result, we confirmed 39 target genes of miR-142-3p and 86 target genes of miR-93. KEGG analysis revealed that these target genes were significantly involved in autophagy and mitophagy. Autophagy occurs under stressful situations such as the presence of abnormal proteins and nutrient deprivation, which degrades cellular proteins and organelles to provide precursors for recycling (63). Some clinical trials have presented that inhibiting autophagy has feasible benefits in multiple cancer types such as glioblastoma, melanoma, and pancreatic cancer (64). Autophagy and mitophagy are demonstrated to contribute to the reprogramming of cancer metabolism that is a major challenge for anticancer therapy (65). Therefore, we supposed that maybe one of the mechanisms of miR-142-3p and miR-93 in NPC development was their participation in autophagy and mitophagy process responsible for cancer metabolism.

In order to distinguish key target genes of the two miRNAs, we applied random sampling and univariate Cox regression on 125 potential target genes. As a consequence, we confirmed five target genes that had prognostic effects on NPC, namely, E2F1, KCNJ8, SUCO, HECTD1, and KIF23, where SUCO and HECTD1 are the targets of miR-142-3p and E2F1, KCNJ8, and KIF23 are the targets of miR-93. E2F1 has been widely reported to regulate cell cycle and cell death and has a significant role in multiple cancer types. E2F1 target pathways are considered as important targets for cancer treatment (66). Limited studies of cancer have been found related to the other four genes. Based on these five target genes, we further established a prognostic model. The five-gene prognostic model manifested substantial performance in predicting 1-, 2-, and 3-year survival with AUCs of 0.87, 0.81, and 0.79 respectively. According to the model, high-risk and low-risk groups were defined with disparate prognosis.

We further investigated the differences of the tumor microenvironment and functional pathways between two risk groups for excavating the biological influence of the five target genes in NPC. Two risk groups had distinct immune cell infiltration. Anticancer immune cells, for instance, activated B cells, NK cells, dendritic cells, and CD8 T cells, were evidently higher in the low-risk group compared with the high-risk group, which led to the stronger anticancer response and clearance of cancer cells. The high-risk group showed both higher innate and adaptive immune response than the low-risk group. Analysis of biological pathways unveiled that the low-risk group displayed higher activation of immune-related pathways, for instance, IL2-STAT5 signaling, interferon alpha response, interferon gamma response, IL6-JAK-STAT3 signaling, inflammation response, and complement. Importantly, DNA repair pathways, cell cycle-related pathways, and apoptosis were also more enriched in the low-risk group, supporting that the two key miRNAs may have an interaction with these target genes. However, the specific association between the miRNAs and five target genes should be further validated in future experiments.

The different proportion of tumor-infiltrating immune cells has a profound effect on both cancer prognosis and immunotherapy (67). SubMap analysis predicted that the low-risk group was more sensitive to ICB therapy than the high-risk group, which may result from the higher expression of critical immune checkpoints such as CTLA4, IDO1, LAG3, and PDCD1 (PD-1) in the low-risk group. Anti-PD-1/PD-L1 therapy has shown some favorable outcomes in clinical trials of NPC (68). Our five-gene model can help clinicians to better select the patients sensitive to ICB therapy and raise the efficiency of immunotherapy.

## Conclusions

In conclusion, this study identified two key miRNAs (miR-142-3p and miR-93) and predicted their potential key target genes. MiR-142-3p and miR-93 contributed to NPC survival possibly through regulating autophagy pathways. In addition, we confirmed five prognostic target genes (E2F1, KCNJ8, SUCO, HECTD1, and KIF23) and constructed a five-gene prognostic model. The model was effective to predicting NPC prognosis and could provide a guidance for personalized immunotherapy in NPC patients.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

## Author contributions

All authors contributed to this present work. XXZ, XiaL and CW designed the study, SW acquired the data. YZ and BL drafted the manuscript. XinL revised the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by Key Projects of Intergovernmental Cooperation in National Key R&D Programs (Grant no. 2022YFE0131800), Natural Science Foundation of Liaoning Province (Grant no. 2022-MS-235), and Project of Shenyang Science and Technology Bureau (Grant no. RC210316).

## Conflict of interest

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.

## Publisher's note



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1174911/full#supplementary-material

## References

1. Chan KCA, Woo JKS, King A, Zee BCY, Lam WKJ, Chan SL, et al. Analysis of plasma Epstein-Barr virus DNA to screen for nasopharyngeal cancer. *New Engl J Med* (2017) 377(6):513–22. doi: 10.1056/NEJMoa1701717

2. Jain A, Chia WK, Toh HC. Immunotherapy for nasopharyngeal cancer-a review. *Chin Clin Oncol* (2016) 5(2):22. doi: 10.21037/cco.2016.03.08

3. Chen YP, Chan ATC, Le QT, Blanchard P, Sun Y, Ma J. Nasopharyngeal carcinoma. *Lancet (London England).* (2019) 394(10192):64–80. doi: 10.1016/S0140-6736(19)30956-0

4. Bai R, Sun J, Xu Y, Sun Z, Zhao X. Incidence and mortality trends of nasopharynx cancer from 1990 to 2019 in China: an age-period-cohort analysis. *BMC Public Health* (2022) 22(1):1351. doi: 10.1186/s12889-022-13688-7

5. Bhattacharyya T, Babu G, Kainickal CT. Current role of chemotherapy in nonmetastatic nasopharyngeal cancer. *J Oncol* (2018) 2018:3725837. doi: 10.1155/2018/3725837

6. Verma V, Allen PK, Simone CB2nd, Gay HA, Lin SH. Addition of definitive radiotherapy to chemotherapy in patients with newly diagnosed metastatic nasopharyngeal cancer. *J Natl Compr Cancer Network JNCCN.* (2017) 15(11):1383–91. doi: 10.6004/jnccn.2017.7001

7. Tseng M, Ho F, Leong YH, Wong LC, Tham IW, Cheo T, et al. Emerging radiotherapy technologies and trends in nasopharyngeal cancer. *Cancer Commun (London England).* (2020) 40(9):395–405. doi: 10.1002/cac2.12082

8. Zhang J, Fang W, Qin T, Yang Y, Hong S, Liang W, et al. Co-Expression of PD-1 and PD-L1 predicts poor outcome in nasopharyngeal carcinoma. *Med Oncol (Northwood London England).* (2015) 32(3):86. doi: 10.1007/s12032-015-0501-6

9. Zhou Y, Miao J, Wu H, Tang H, Kuang J, Zhou X, et al. PD-1 and PD-L1 expression in 132 recurrent nasopharyngeal carcinoma: the correlation with anemia and outcomes. *Oncotarget* (2017) 8(31):51210–23. doi: 10.18632/oncotarget.17214

10. Hsu C, Lee SH, Ejadi S, Even C, Cohen RB, Le Tourneau C, et al. Safety and antitumor activity of pembrolizumab in patients with programmed death-ligand 1-positive nasopharyngeal carcinoma: results of the KEYNOTE-028 study. *J Clin Oncol* (2017) 35(36):4050–6. doi: 10.1200/JCO.2017.73.3675

11. Ma BBY, Lim WT, Goh BC, Hui EP, Lo KW, Pettinger A, et al. Antitumor activity of nivolumab in recurrent and metastatic nasopharyngeal carcinoma: an international, multicenter study of the Mayo clinic phase 2 consortium (NCI-9742). *J Clin Oncol* (2018) 36(14):1412–8. doi: 10.1200/JCO.2017.77.0388

12. Chen G, Huang AC, Zhang W, Zhang G, Wu M, Xu W, et al. Exosomal PD-L1 contributes to immunosuppression and is associated with anti-PD-1 response. *Nature* (2018) 560(7718):382–6. doi: 10.1038/s41586-018-0392-8

13. Garcia-Diaz A, Shin DS, Moreno BH, Saco J, Escuin-Ordinas H, Rodriguez GA, et al. Interferon receptor signaling pathways regulating PD-L1 and PD-L2 expression. *Cell Rep* (2017) 19(6):1189–201. doi: 10.1016/j.celrep.2017.04.031

14. Holly JM, Perks CM. Cancer as an endocrine problem. *Best Pract Res Clin Endocrinol Metab* (2008) 22(4):539–50. doi: 10.1016/j.beem.2008.07.007

15. Cascio S, Bartella V, Auriemma A, Johannes GJ, Russo A, Giordano A, et al. Mechanism of leptin expression in breast cancer cells: role of hypoxia-inducible factor-1alpha. *Oncogene* (2008) 27(4):540–7. doi: 10.1038/sj.onc.1210660

16. Schally AV, Varga JL, Engel JB. Antagonists of growth-hormone-releasing hormone: an emerging new therapy for cancer. *Nat Clin Pract Endocrinol Metab* (2008) 4(1):33–43. doi: 10.1038/ncpendmet0677

17. Schally AV. New approaches to the therapy of various tumors based on peptide analogues. *Hormone Metab Res = Hormon- und Stoffwechselforschung = Hormones metabolisme.* (2008) 40(5):315–22. doi: 10.1055/s-2008-1073142

18. Eichmüller SB, Osen W, Mandelboim O, Seliger B. Immune modulatory microRNAs involved in tumor attack and tumor immune escape. *J Natl Cancer Institute* (2017) 109(10). doi: 10.1093/jnci/djx034

19. Wang S, Claret FX, Wu W. MicroRNAs as therapeutic targets in nasopharyngeal carcinoma. *Front Oncol* (2019) 9:756. doi: 10.3389/fonc.2019.00756

20. Yu L, Lu J, Zhang B, Liu X, Wang L, Li SY, et al. miR-26a inhibits invasion and metastasis of nasopharyngeal cancer by targeting EZH2. *Oncol Letters.* (2013) 5 (4):1223–8. doi: 10.3892/ol.2013.1173

21. Yi C, Wang Q, Wang L, Huang Y, Li L, Liu L, et al. MiR-663, a microRNA targeting p21(WAF1/CIP1), promotes the proliferation and tumorigenesis of nasopharyngeal carcinoma. *Oncogene* (2012) 31(41):4421–33. doi: 10.1038/onc.2011.629

22. Bell E, Taylor MA. Functional roles for exosomal MicroRNAs in the tumour microenvironment. *Comput Struct Biotechnol J* (2017) 15:8–13. doi: 10.1016/j.csbj.2016.10.005

23. Liao C, Liu H, Luo X. The emerging roles of exosomal miRNAs in nasopharyngeal carcinoma. *Am J Cancer Res* (2021) 11(6):2508–20.

24. Clough E, Barrett T. The gene expression omnibus database. *Methods Mol Biol (Clifton NJ).* (2016) 1418:93–110. doi: 10.1007/978-1-4939-3578-9_5

25. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* (2015) 43(7):e47. doi: 10.1093/nar/gkv007

26. Langfelder P, Horvath S. WGCNA: an r package for weighted correlation network analysis. *BMC Bioinf* (2008) 9:559. doi: 10.1186/1471-2105-9-559

27. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med* (2013) 32(30):5381–97. doi: 10.1002/sim.5958

28. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, et al. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res* (2013) 41(Web Server issue):W169–73. doi: 10.1093/nar/gkt393

29. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. *PloS Biol* (2004) 2(11):e363. doi: 10.1371/journal.pbio.0020363

30. Jeggari A, Marks DS, Larsson E. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinf (Oxford England).* (2012) 28 (15):2062–3. doi: 10.1093/bioinformatics/bts344

31. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res* (2020) 48(D1):D127–d31. doi: 10.1093/nar/gkz757

32. Vejnar CE, Blum M, Zdobnov EM. miRmap web: comprehensive microRNA target prediction online. *Nucleic Acids Res* (2013) 41(Web Server issue):W165–8. doi: 10.1093/nar/gkt430

33. Huang HY, Lin YC, Li J, Huang KY, Shrestha S, Hong HC, et al. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res* (2020) 48(D1):D148–d54. doi: 10.1093/nar/gkz896

34. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. *Nat Genet* (2005) 37(5):495–500. doi: 10.1038/ng1536

35. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. *Nat Genet* (2007) 39(10):1278–84. doi: 10.1038/ng2135

36. Bandyopadhyay S, Mitra R. TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinf (Oxford England).* (2009) 25(20):2625–31. doi: 10.1093/bioinformatics/btp503

37. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* (2015) 4:e05005. doi: 10.7554/eLife.05005

38. Loher P, Rigoutsos I. Interactive exploration of RNA22 microRNA target predictions. *Bioinf (Oxford England).* (2012) 28(24):3322–3. doi: 10.1093/bioinformatics/bts615

39. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-seq data. *Nucleic Acids Res* (2014) 42(Database issue):D92–7. doi: 10.1093/nar/gkt1248

40. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* (2019) 47(W1):W199–w205. doi: 10.1093/nar/gkz401

41. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinf* (2013) 14:7. doi: 10.1186/1471-2105-14-7

42. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep* (2017) 18 (1):248–62. doi: 10.1016/j.celrep.2016.12.019

43. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* (2013) 4:2612. doi: 10.1038/ncomms3612

44. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol* (2016) 17(1):218. doi: 10.1186/s13059-016-1070-5

45. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Systems.* (2015) 1(6):417–25. doi: 10.1016/j.cels.2015.12.004

46. Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y, et al. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* (2018) 554(7693):544–8. doi: 10.1038/nature25501

47. Roh W, Chen PL, Reuben A, Spencer CN, Prieto PA, Miller JP, et al. Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci Trans Med* (2017) 9(379):eaah3560. doi: 10.1126/scitranslmed.aah3560

48. Balar AV, Galsky MD, Rosenberg JE, Powles T, Petrylak DP, Bellmunt J, et al. Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial. *Lancet (London England).* (2017) 389(10064):67–76. doi: 10.1016/S0140-6736(16)32455-2

49. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med* (2018) 24(10):1550–8. doi: 10.1038/s41591-018-0136-1

50. Geeleher P, Cox N, Huang RS. pRRophetic: an r package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PloS One* (2014) 9(9):e107468. doi: 10.1371/journal.pone.0107468

51. Kim JY, Choi JK, Jung H. Genome-wide methylation patterns predict clinical benefit of immunotherapy in lung cancer. *Clin Epigenetics.* (2020) 12(1):119. doi: 10.1186/s13148-020-00907-4

52. Shen W, Song Z, Xiao Z, Huang M, Shen D, Gao P, et al. Sangerbox: a comprehensive, interaction-friendly clinical bioinformatics analysis platform. *iMeta* (2022) 1(3):e36. doi: 10.1002/imt2.36

53. Wells SAJr. Progress in endocrine neoplasia. *Clin Cancer Res* (2016) 22 (20):4981–8. doi: 10.1158/1078-0432.CCR-16-0384

54. Zhang H, Zhu Q, Cui J, Wang Y, Chen MJ, Guo X, et al. Structure and evolution of the Fam20 kinases. *Nat Commun* (2018) 9(1):1218. doi: 10.1038/s41467-018-03615-z

55. Feng J, Zhou J, Zhao L, Wang X, Ma D, Xu B, et al. Fam20C overexpression predicts poor outcomes and is a diagnostic biomarker in lower-grade glioma. *Front Genet* (2021) 12:757014. doi: 10.3389/fgene.2021.757014

56. Lin RJ, Xiao DW, Liao LD, Chen T, Xie ZF, Huang WZ, et al. MiR-142-3p as a potential prognostic biomarker for esophageal squamous cell carcinoma. *J Surg Oncol* (2012) 105(2):175–82. doi: 10.1002/jso.22066

57. Lei Z, Xu G, Wang L, Yang H, Liu X, Zhao J, et al. MiR-142-3p represses TGF-β-induced growth inhibition through repression of TGFβR1 in non-small cell lung cancer. *FASEB J* (2014) 28(6):2696–704. doi: 10.1096/fj.13-247288

58. Dong H, Song J. miR-142-3p reduces the viability of human cervical cancer cells by negatively regulating the cytoplasmic localization of HMGB1. *Exp Ther Med* (2021) 21(3):212. doi: 10.3892/etm.2021.9644

59. Sharma P, Yadav P, Jain RP, Bera AK, Karunagaran D. miR-142-3p simultaneously targets HMGA1, HMGA2, HMGB1, and HMGB3 and inhibits tumorigenic properties and in-vivo metastatic potential of human cervical cancer cells. *Life Sci* (2022) 291:120268. doi: 10.1016/j.lfs.2021.120268

60. Yu XF, Zou J, Bao ZJ, Dong J. miR-93 suppresses proliferation and colony formation of human colon cancer stem cells. *World J Gastroenterology.* (2011) 17 (42):4711–7. doi: 10.3748/wjg.v17.i42.4711

61. Shyamasundar S, Lim JP, Bay BH. miR-93 inhibits the invasive potential of triple-negative breast cancer cells *in vitro via* protein kinase WNK1. *Int J Oncol* (2016) 49(6):2629–36. doi: 10.3892/ijo.2016.3761

62. Fang S, Gao M, Xiong S, Chen Q, Zhang H. Expression of serum hsa-miR-93 in uterine cancer and its clinical significance. *Oncol Letters.* (2018) 15(6):9896–900. doi: 10.3892/ol.2018.8553

63. Yun CW, Lee SH. The roles of autophagy in cancer. *Int J Mol Sci* (2018) 19 (11):3466. doi: 10.3390/ijms19113466

64. Levy JMM, Towers CG, Thorburn A. Targeting autophagy in cancer. *Nat Rev Cancer.* (2017) 17(9):528–42. doi: 10.1038/nrc.2017.53

65. Ferro F, Servais S, Besson P, Roger S, Dumas JF, Brisson L. Autophagy and mitophagy in cancer metabolic remodelling. *Semin Cell Dev Biol* (2020) 98:129–38. doi: 10.1016/j.semcdb.2019.05.029

66. Huang Y, Chen R, Zhou J. E2F1 and NF-κB: key mediators of inflammation-associated cancers and potential therapeutic targets. *Curr Cancer Drug Targets.* (2016) 16(9):765–72. doi: 10.2174/1568009616666160216130755

67. Lu J, Chen XM, Huang HR, Zhao FP, Wang F, Liu X, et al. Detailed analysis of inflammatory cell infiltration and the prognostic impact on nasopharyngeal carcinoma. *Head Neck.* (2018) 40(6):1245–53. doi: 10.1002/hed.25104

68. Adkins DR, Haddad RI. Clinical trial data of anti-PD-1/PD-L1 therapy for recurrent or metastatic nasopharyngeal carcinoma: a review. *Cancer Treat Rev* (2022) 109:102428. doi: 10.1016/j.ctrv.2022.102428

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 31 July 2023
DOI 10.3389/fendo.2023.1201755



OPEN ACCESS

EDITED BY
Wenjie Shi,
Otto von Guericke University Magdeburg, Germany

REVIEWED BY
Ziheng Wang,
University of Macau, China
Liang Yu,
Second Affiliated Hospital of Harbin Medical University, China

*CORRESPONDENCE
Qiang Wang
wangqiang1983@zju.edu.cn

RECEIVED 07 April 2023
ACCEPTED 04 July 2023
PUBLISHED 31 July 2023

CITATION
Xu Y, Chen X, Liu N, Chu Z and Wang Q (2023) Identification of fibroblast-related genes based on single-cell and machine learning to predict the prognosis and endocrine metabolism of pancreatic cancer.
*Front. Endocrinol.* 14:1201755.
doi: 10.3389/fendo.2023.1201755



# Identification of fibroblast-related genes based on single-cell and machine learning to predict the prognosis and endocrine metabolism of pancreatic cancer

Yinghua Xu[1], Xionghuan Chen[2,3], Nan Liu[1], Zhong Chu[1] and Qiang Wang[1]*

[1]Department of Translational Medicine and Clinical Research, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China, [2]Department of General Surgery, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China, [3]Department of Trauma Surgery, Tiantai People's Hospital of Zhejiang Province, Taizhou, China

**Background:** Single-cell sequencing technology has become an indispensable tool in tumor mechanism and heterogeneity studies. Pancreatic adenocarcinoma (PAAD) lacks early specific symptoms, and comprehensive bioinformatics analysis for PAAD contributes to the developmental mechanisms.

**Methods:** We performed dimensionality reduction analysis on the single-cell sequencing data GSE165399 of PAAD to obtain the specific cell clusters. We then obtained cell cluster-associated gene modules by weighted co-expression network analysis and identified tumorigenesis-associated cell clusters and gene modules in PAAD by trajectory analysis. Tumor-associated genes of PAAD were intersected with cell cluster marker genes and within the signature module to obtain genes associated with PAAD occurrence to construct a prognostic risk assessment tool by the COX model. The performance of the model was assessed by the Kaplan–Meier (K-M) curve and the receiver operating characteristic (ROC) curve. The score of endocrine pathways was assessed by ssGSEA analysis.

**Results:** The PAAD single-cell dataset GSE165399 was filtered and downscaled, and finally, 17 cell subgroups were filtered and 17 cell clusters were labeled. WGCNA analysis revealed that the brown module was most associated with tumorigenesis. Among them, the brown module was significantly associated with C11 and C14 cell clusters. C11 and C14 cell clusters belonged to fibroblast and circulating fetal cells, respectively, and trajectory analysis showed low heterogeneity for fibroblast and extremely high heterogeneity for circulating fetal cells. Next, through differential analysis, we found that genes within the C11 cluster were highly associated with tumorigenesis. Finally, we constructed the RiskScore system, and K-M curves and ROC curves revealed that RiskScore possessed objective clinical prognostic potential and demonstrated consistent

robustness in multiple datasets. The low-risk group presented a higher endocrine metabolism and lower immune infiltrate state.

**Conclusion:** We identified prognostic models consisting of APOL1, BHLHE40, CLMP, GNG12, LOX, LY6E, MYL12B, RND3, SOX4, and RiskScore showed promising clinical value. RiskScore possibly carries a credible clinical prognostic potential for PAAD.

KEYWORDS

single-cell sequencing, pancreatic adenocarcinoma, tumorigenesis, RiskScore, prognosis, endocrine metabolism

## Introduction

Although pancreatic adenocarcinoma (PAAD) is a relatively low-incidence cancer, it is a highly lethal tumor (1). The deficiency of specific early symptoms of PAAD and the fact that the majority of patients are experiencing advanced progression or organ metastases contribute to PAAD being a high-mortality cancer (2). Frustratingly, radiotherapy, as well as chemotherapy, were not effective options in the treatment of PAAD, and surgical resection was currently the best option for most patients, but the prognosis was graded poorly, with an overall 5-year survival (OS) rate of less than 10% (3–5). Several studies have shown that prognosis in a variety of cancers, including PAAD, can be predicted using carbohydrate antigen 19-9 (CA 19-9) and carcinoembryonic antigen (CEA) (6, 7). However, they lack specificity and sensitivity for PAAD (8). To address the clinical pain point that PAAD prognosis was difficult to assess, it was imperative to develop effective prognostic tools to achieve patient prognostic risk assessment as well as personalized and precise treatment.

The pancreas has two functions: endocrine and exocrine. The exocrine glands of the pancreas are composed of acinar cells and duct cells. Previous studies have believed that pancreatic ductal adenocarcinoma (PDAC) originates from ductal cells because of tumor histological similarity to ductal morphology (9). On the other hand, pancreatic endocrine tumors are caused by endocrine cells (10). With the emergence of single-cell RNA sequencing (scRNA-Seq) technology, exploring deeper molecular mechanisms of life from cellular genetic material, functional heterogeneity, and the identification of specific cell subtypes emerged as mainstream research directions (11, 12). scRNA-seq maps the gene expression patterns of each cell and decodes its intercellular signaling network. This unbiased characterization provides clear insights into the entire tumor ecosystem, such as the mechanisms of intra- and intertumor heterogeneity and the tumor microenvironment (13). Tumors lead to an individualized prognosis and variable therapeutic responses due to their heterogeneity, and single-cell technologies showed powerful functions in revealing the molecular mechanisms of cancer through the precise analysis of specific cells or cell clusters (14–16). For example, Wang et al. developed a lung cancer artificial intelligence detector using scRNA-Seq data from early lung cancer, which showed great specificity in the early detection of lung cancer and large-scale early screening of high-risk populations (17). Li et al. identified proinvasive cancer-associated fibroblast subtypes in patients with poor prognosis for gastric cancer, and inhibition of these cell subsets contributed to creating an activated immune tumor microenvironment (TME) (18). These studies demonstrated the ability to integrate scRNA-Seq data to deepen insights into cancer.

Fibroblast growth in pancreatic cancer (PDAC) tumors is known as a tumor suppressor (19, 20). Cancer-associated fibroblasts (CAFs) are a collective term for these cells. CAFs may play a role in the development and progression of PDAC and the response to treatment (21, 22). CAFs are an important stromal component, secreting growth factors, inflammation mediators, and extracellular matrix (ECM) proteins that facilitate tumor growth, resistance to therapy, and immune rejection (23).

Machine learning is a branch of artificial intelligence that focuses on making predictions by using mathematical algorithms to identify patterns in data. Deep learning is a branch of machine learning that focuses on making predictions using multi-layered neural network algorithms inspired by the neural structure of the brain. In contrast to other machine learning methods, such as logistic regression, deep learning's neural network architecture enables models to scale exponentially as the amount and dimension of data grow (24). Machine learning algorithms to help with cancer detection (identifying the presence of cancer) and diagnosis (characterizing cancer) have become increasingly common (25, 26). In this study, we integrated scRNA-Seq data from three different samples from the Gene Expression Omnibus (GEO) database to identify cell clusters associated with PAAD occurrence. RNA-Seq data from PAAD

**Abbreviations:** PAAD, pancreatic adenocarcinoma; K-M, Kaplan–Meier; ROC, receiver operating characteristic; scRNA-Seq, single-cell RNA sequencing; TME, tumor microenvironment; GEO, Gene Expression Omnibus; TCGA, the Cancer Genome Atlas; GTEx, genotype-tissue expression; WGCNA, weighted gene correlation network analysis; LASSO, least absolute shrinkage and selection operator; UCSC Xena, University of California Santa Cruz; PCA, principal component analysis; FC, fold change; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes; ssGSEA, single-sample geneset enrichment analysis; CAFs, cancer-associated fibroblasts; BLCA, bladder urothelial carcinoma.

samples and normal samples from the Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) databases were subsequently identified by weighted gene correlation network analysis (WGCNA), which identified cellular clusters associated with gene modules, and we screened prognostic genes associated with PAAD occurrence by the univariate COX model and the least absolute shrinkage and selection operator (LASSO) COX model to construct a prognostic risk assessment system for PAAD.

# Materials and methods

## Data acquisition

The scRNA-Seq data (registration number: GSE165399) were downloaded from the GEO (https://www.ncbi.nlm.nih.gov/geo/) database, containing three samples, and the sample information is presented in Table 1. Four PAAD patient sequencing datasets were also downloaded (registration numbers: GSE28735, tumor samples: 42; GSE57495, tumor samples: 63; GSE62452, tumor samples: 64; and GSE85916, tumor samples: 79). RNA-Seq data (TCGA-PAAD, tumor samples: 177) from the PAAD sequencing project were downloaded from TCGA database (https://portal.gdc.cancer.gov/), as well as clinical information for the 177 samples. Normal pancreatic samples were downloaded from the GTEx (https://www.gtexportal.org/home/) database. Finally, RNA-seq data from the IGGC-AU sequencing project were downloaded from the University of California Santa Cruz (UCSC Xena, https://xena.ucsc.edu/).

## scRNA-Seq data pre-processing

The scRNA-Seq data of the GSE165399 cohort samples were processed utilizing the Seurat package (27). First, the genes that were expressed in all three cells were screened, and the number of genes expressed in each cell was greater than 250. The PercentageFeatureSet function was employed to calculate the percentage of mitochondria and rRNA and to ensure that each cell expressed more than 500 and less than 7,000 genes with less than 30% mitochondrial content. Also, the number of UMI in each cell was ensured to be no less than 500.

## scRNA-Seq data clustering and dimension reduction

Initially, the samples were merged by the merge function in the Seurat package, and the merged data were normalized by log normalization. High-variability genes were then detected by the FindVariableFeatures function (based on the variance stabilization transformation (vst) to identify variable features). All genes were scaled with the ScaleData function and subjected to principal component analysis (PCA) with the RunPCA function. We then performed cell clustering analysis (set resolution = 0.3) by selecting dim = 40 and identifying specific cell clusters in PAAD by the FindNeighbors and FindClusters functions. Next, with the top 40 principal components selected, we operated the UMAP program to further reduce the dimensionality. Finally, we screened marker genes in cell clusters by |logfold change (FC)| = 0.35 and Minpct = 0.3 (minimum expression ratio of differential genes) via the FindAllMarkers function.

## RNA-Seq data processing

RNA-Seq data were processed on the SangerBox website, a comprehensive online bioinformatics analysis website (28). Samples without follow-up information were removed from the TCGA-PAAD cohort, FPKM data were transformed into TPM data, and normal pancreatic samples from the UCSC Xena were subsequently merged, and the merged cohort was recorded as TCGA _GTEx-PAAD (tumor: 177, normal: 167, gene number: 24210). Normal samples, samples with missing follow-up information in the GSE28735, GSE57495, GSE62452, and GSE85916 cohorts were excluded.

## Annotation of cell clusters

The cell marker genes for human cells were selected from the official cell marker website (http://biocc.hrbmu.edu.cn/CellMarker/) for the pancreas, pancreatic acinar tissue, peripheral blood, and blood corresponding tissues. The enricher function in the clusterProfiler package (29) was provided for cell cluster annotation.

## Monocle trajectory analysis

Monocle (version 2.18.0) is used to infer the developmental trajectory of subpopulations of cells. Cells were isolated from the Seurat object and transferred into the SingleCellExperiment format (follow the official tutorial for trajectory analysis: http://cole-trapnell-lab.github.io/monocle-release/docs/#constructing-single-cell-trajectories). The Monocle object is built from the SingleCellExperiment format using the new cell dataset function in Monocle.

TABLE 1 Clinical information for samples in the GSE165399 cohort.

| Sample ID | Sample tissues | Age | Gender |
|---|---|---|---|
| GSM5032771 | Intraductal papillary mucinous neoplasm | 74 | Male |
| GSM5032772 | Pancreatic adenosquamous carcinoma | 59 | Male |
| GSM5032773 | normal pancreas sample | 50 | Male |

## PAAD-related cell cluster abundance analysis

Based on marker genes in cell clusters, we computed the relative abundance of cell subpopulations in tumor and normal tissues in the TCGA_GTEx-PAAD cohort using the CIBERSORT method (30).

## WGCNA analysis

To identify key genes for tumorigenesis, we performed WGCNA analysis on samples in the TCGA_GTEx-PAAD cohort. Cluster analysis was first performed on 177 tumor samples and 167 normal samples to further calculate the Pearson's correlation between each gene, followed by constructing co-expression networks and forming gene modules using the WGCNA package (31). Subsequently, the Pearson's correlation analysis was performed with each gene module using the first principal component (ME) of the cell subpopulation to identify the key gene modules for PAAD occurrence. The Monocle3 package was also performed to analyze cellular pseudo-temporal trajectories (32).

## Enrichment analysis

To explore the biological functions as well as signaling pathways involved in genes within the key modules of PAAD occurrence, we performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) functional enrichment analyses using the clusterProfiler package at $p< 0.05$ and FDR< 0.05 as thresholds to screen the most significantly enriched molecular functions and signaling pathways.

## Cell communication analysis

In multicellular organisms, the basic vital activities of life depend on cell–cell interactions as a contribution to the coordination of their behavior. The communication between cells relies mainly on multisubunit protein complexes. Based on this, we used the cell chart package (33) to analyze the number of interacting ligands between all cell subpopulations as well as changes in the strength of the interaction.

## Screening for PAAD-generating genes

Differential analysis was conducted via the limma package (34) to obtain tumor-associated differentially expressed genes (DEGs) in tumor and normal tissues in the TCGA_GTEx-PAAD cohort, which were subsequently intersected with cellular subpopulations associated with PAAD occurrence, and genes within the signature module were taken to obtain key genes for PAAD occurrence.

## Prognostic model construction, evaluation, and validation

Univariate COX models were generated for the expression matrix of tumor samples from the TCGA-PAAD cohort in combination with patient survival status and survival time to identify genes affecting PAAD survival ($p< 0.01$). Models with a large number of genes were not conducive to clinical test manipulation, so we constructed LASSO COX models based on the above genes employing the glmnet package (35, 36) and removed genes with high similarity in the models by introducing the penalty parameter lambda in 10-fold crossvalidation. The resulting genes were the PAAD prognostic signature genes. Based on the regression coefficients in the LASSO COX model and the expression levels of individual genes, we constructed the RiskScore for the PAAD prognostic risk assessment tool, which was calculated by the following equation.

$$\text{RiskScore} = \sum \beta_i * \text{Exp}\ i$$

where β was the regression coefficient normalized by the *Z*-score for each gene in the LASSO COX model, and Exp represented the gene expression data.

The RiskScore of tumor samples in the TCGA-PAAD, GSE28735, GSE57495, GSE62452, GSE85916, and IGGC-AU cohorts was determined according to the formula, and the high RiskScore group and low RiskScore group were classified based on RiskScore = 0 as the threshold. Kaplan–Meier (K-M) survival curves were plotted to assess the prognostic differences between the two groups, and receiver operating characteristic curve (ROC) was developed to assess the performance of RiskScore in predicting PAAD prognosis.

## Potential associations between RiskScore and clinical features

Patients in the TCGA-PAAD cohort were grouped according to clinical features, and the RiskScore of patients in each subgroup was counted. The Wilcox test was conducted to calculate the statistical difference between the two groups, and the Kruskal–Wallis test was conducted to calculate the statistical difference among the four groups. $P< 0.05$ was considered to be significantly distinct.

## Gene set enrichment analysis

To explore the biological pathways that existed differently between the high RiskScore and low RiskScore groups, single sample gene set enrichment analysis (ssGSEA) was performed on the high RiskScore samples and low RiskScore samples in the TCGA-PAAD cohort, and the ssGSEA scores of pathways were analyzed for the Pearson's correlation with RiskScore, and pathways with $r > 0.5$ were considered potentially regulated pathways by RiskScore.

## Endocrine metabolism analysis

The KEGG website provided 34 secretory genes. Genes in SECRETORY_PATHWAY were obtained from the GSEA website (https://www.gsea-msigdb.org/gsea/index.jsp) to calculate the score using ssGSEA.

## Statistical analysis

This study was performed using R software (version 4.1.1) for data analysis. For all statistical analyses, bilateral $p< 0.05$ was considered statistically significant.

# Results

## Dimensionality reduction and clustering of scRNA-Seq data

Initially, the scRNA-Seq data were filtered to retain the genes that were expressed in GSM5032771, GSM5032772, and GSM5032773 (Supplementary Figures S1, S2). The filtered data were combined, and the highly variable genes in the samples were filtered by the FindVariableFeatures function. The volcano figure showed the highly variable genes in the samples and marked the top 20 highly labeled genes (Supplementary Figure S3). All genes were scaled using the ScaleData function, and PCA downscaling was performed to find the anchor points (Supplementary Figure S4). By cluster analysis, we obtained 17 subgroups and showed their distribution characteristics in the sample (Figure 1A), and we further selected the top 40 principal components to further downscale by UMAP to obtain 17 cell clusters (Figure 1B). We used the FindAllMarkers function to screen marker genes in 17 clusters by |logFC| = 0.35, Minpct = 0.3 (minimum expression proportion of difference genes) with corrected $p< 0.05$, and Figure 1C demonstrates the top five marker gene expression levels in 17 cell clusters.

## Annotation of 17 cell clusters

The marker genes of the human pancreas, pancreatic acinar tissue, peripheral blood, and blood tissues were annotated by the enricher function of the clusterProfiler package for 17 cell clusters. The annotation information of each cell cluster is shown in Table 2. We found the presence of multiple small clusters in four cell subgroups, including B cell with two clusters, C4 and C10; cancer cell with two clusters, C8 and C15; CD1C–CD141-dendritic cell with three clusters, C0, C1, and C9; and fibroblast with C2 and C11 clusters. Furthermore, we analyzed the differential expression of marker genes in each cell cluster, and we found that C0 specifically expressed CLEC4E, C1 specifically expressed MRC1, C2 subpopulation specifically expressed GAS1, C4 specifically expressed FCMR, C8 specifically expressed DEFB1, C9 specifically expressed MTND1P23, C10 specifically expressed TCL1A, C11 specifically expressed the *HSPB6* gene, and C15 specifically expressed RGS5 (Figure 2A). In addition, we found higher abundances of C0, C1, C2, C4, C6, C8, C9, C10, C11, C13, C14, and C15 in tumor tissues (Figure 2B).

## PAAD-related gene module identification

To identify tumor-associated gene modules in PAAD, we performed a WGCNA analysis. After the samples were clustered to construct a scale-free network, we found that the co-expression network conformed to the scale-free network at the soft threshold $\beta = 7$, when the scale-free $R^2$ was 0.85 (Figures 3A–C). A total of six gene modules were generated, among which the brown module (gene number: 4811) was highly correlated with the C11 ($r = 0.8$, $p =$ 5e−79) and C14 ($r = 0.86$, $p =$ 8e−104) cluster (Figure 3D). To explore the biological functions of genes within the brown module, we performed GO and KEGG enrichment analyses. We found that these genes were mainly involved in biological processes like angiogenesis and blood vessel morphogenesis; they may also be involved in extracellular matrix and extracellular matrix that contains collagen, adherens junctions between cells and their substrates, and focal adhesion sites; they are also closely related to SMAD binding, extracellular matrix structural constituents, and cell adhesion molecule-binding functions (Figures 4A–C). We also revealed that these genes are actively involved in signaling pathways such as focal adhesion and regulation of the actin cytoskeleton (Figure 4D). Our results suggested that genes within the brown module were intimately associated with intercellular signaling transitions.

## Trajectory analysis of critical cell cluster

The WCGNA analysis revealed a significant correlation between genes within the brown module and tumorigenesis, while the module was highly correlated with the C11 and C14 clusters. The two clusters belong to fibroblast and circulating fetal cells, respectively, where fibroblast cells are characterized by two clusters, C11 and C2. We suggested that the two clusters might be critical clusters for tumorigenesis. We then performed cell trajectory analysis of the critical cluster by Monocle. From the cell differentiation trajectory, C11 and C2, which were also fibroblast cells, showed the same differentiation trend, basically at the tail end of the state 1 branch (Figures 5A, B), while the heterogeneity of circulating fetal cell cells was extremely high (Figure 5C).

## Cell communication analysis

To better investigate how the C11 cluster communicates with other cell clusters, we performed a cellular communication analysis. Figure 6A shows the interactions and intensity changes between 17 cell clusters, and the results indicate a high correlation between cells. Subsequently, we extracted the ligand–receptor information of each subpopulation to communicate with each other, and we found that

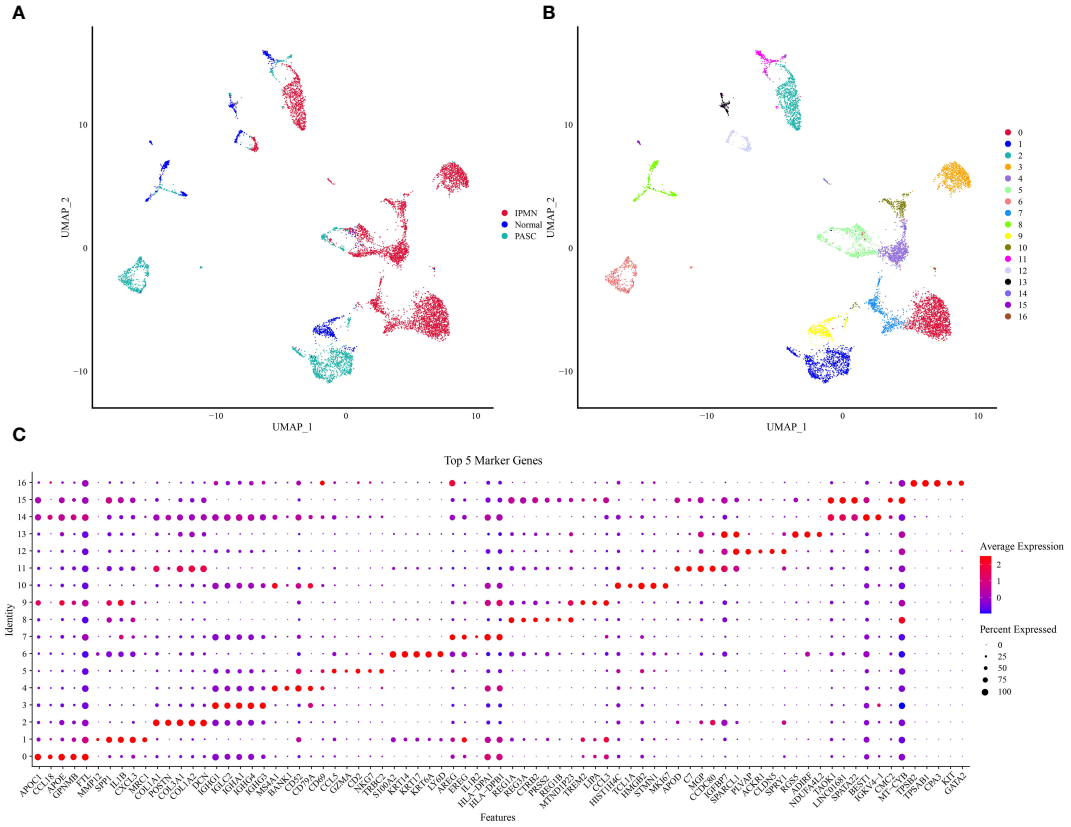



FIGURE 1
UMAP downscaling analysis of scRNA data. **(A)** Distribution of cell subpopulations in GSM5032771, GSM5032772, and GSM5032773 samples. **(B)** Distribution of 17 cell clusters. **(C)** Heatmap of top 50 gene expression in 17 cell clusters.

TABLE 2 Annotation information for 17 cell clusters.

| Seraut_cluster | Cell_type |
|---|---|
| C0 | CD1C-CD141- dendritic cell |
| C1 | CD1C-CD141- dendritic cell |
| C2 | Fibroblast |
| C3 | Plasmacytoid dendritic cell |
| C4 | B cell |
| C5 | T cell |
| C6 | Epithelial cell |
| C7 | CD1C+_B dendritic cell |
| C8 | Cancer cell |
| C9 | CD1C-CD141- dendritic cell |
| C10 | B cell |
| C11 | Fibroblast |
| C12 | Endothelial cell |
| C13 | CD141+CLEC9A+ dendritic cell |
| C14 | Circulating fetal cell |
| C15 | Cancer cell |
| C16 | Basophil |

*(Continued)*

TABLE 2 Continued

C11 and C14 influenced another cluster through some ligand receptors, and the C11 subpopulation influenced other cell cluster by acting on cell surface receptors through LAMC1 (Figure 6B). In addition, we also found some novel pairing relationships, such as LAMA4-CD44 and FN1-SDC4, and these results suggested that the C11 subpopulation played a great role in the development of PAAD.

## Tumorigenesis gene screening

Differential analysis of tumor samples and normal samples in TCGA_GTEx-PAAD identified 3,864 DEGs in tumor tissues, of which 2,008 upregulated DEGs and 1,856 downregulated DEGs were identified (Figure 7A). Furthermore, Venn diagrams were drawn to identify overlapping genes in DEGs, brown module genes, and marker genes in C11 and C14 cell clusters. The C11 cluster, DEGS, and brown module genes contained 107 overlapping genes

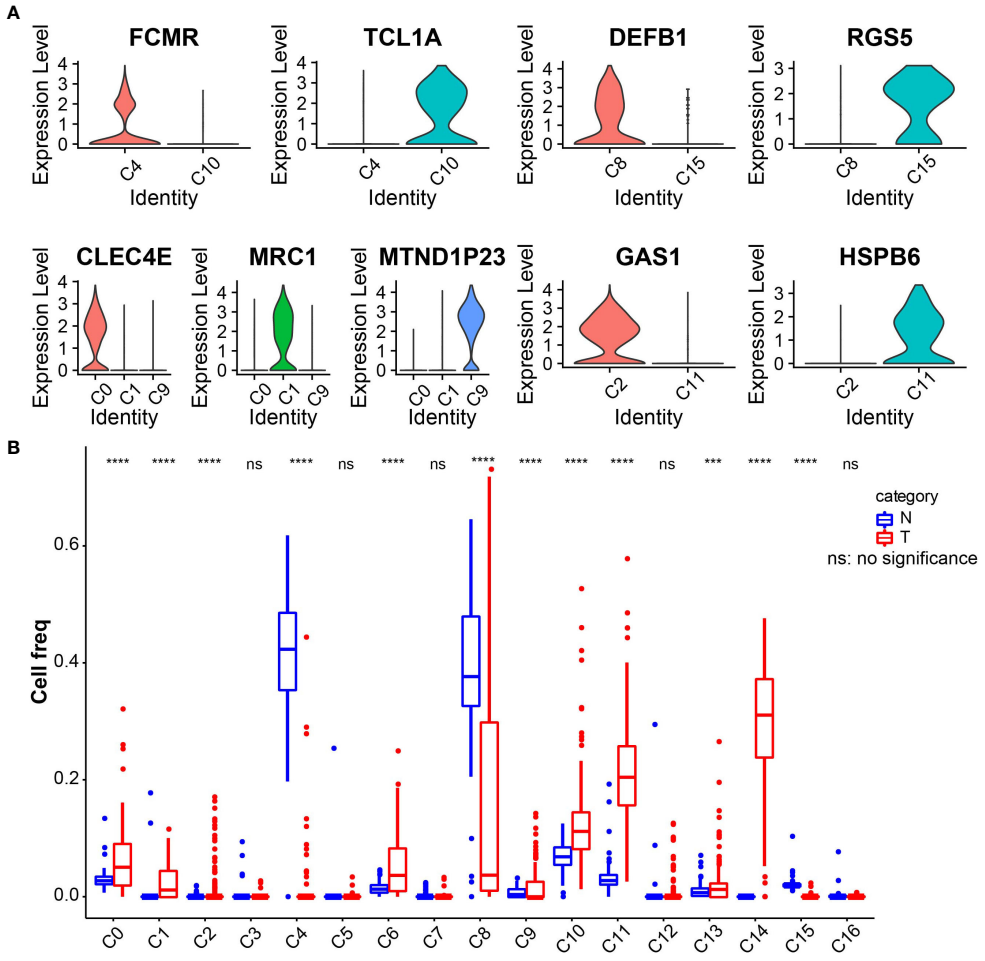



FIGURE 2
**(A)** Violin plot of expression of characteristic genes in cell clusters. **(B)** Boxplot of the abundance of 17 cell clusters in tumor and normal tissues in the TCGA_GTEx-PAAD cohort. ***p<0.001, ****p<0.0001.

(Figures 7B, C), while the C14 cluster, DEGs, and brown module genes contained one overlapping gene (Figures 7D, E). Our results indicated that the overlapping genes were all highly expressed in tumor tissues. Only one overlapping gene was present in marker genes in the C14 cluster; therefore, we concluded that genes in the C11 cluster might be pivotal genes in PAAD tumorigenesis.

## PAAD clinical prognostic model

The univariate COX model found 24 prognostic genes that were significantly associated with PAAD prognosis. It was well known that multigene models were unfavorable for clinical detection, so we employed the LASSO COX model to compress the number of genes in the model and remove the genes with high similarity. Based on 10-fold crossvalidation to select the best penalty parameter lambda, we found that the model was optimal at lambda = 0.0269, so we selected nine genes (APOL, BHLHE40, CLMP, GNG12, LOX, LY6E, MYL12B, RND3, SOX4) at lambda = 0.0269 as the target genes of the next procedure (Figures 8A, B). Based on the regression coefficients and gene expression levels, we constructed a clinical prognosis assessment system for PAAD patients with RiskScore = 0.128 * APOL1 + 0.153 * BHLHE40 – 0.552 * CLMP – 0.363 * GNG12 + 0.528 * LOX – 0.202 * LY6E – 0.202 * MYL12B + 0.051 * RND3 + 1.003 * SOX4. Patients were classified into the high RiskScore group ($N$ = 108) and low RiskScore group ($N$ = 68) by RiskScore Z-score normalized to 0 as the grouping threshold for the sample. We identified that patients in the high RiskScore group had a worse prognosis and a higher death rate in the TCGA-PAAD cohort (Figures 8C, D). The AUC values for RiskScore to predict 1-, 3-, and 5-year survival in PAAD were 0.67, 0.76, and 0.77, respectively (Figure 8E).

## Validation of RiskScore

To better assess the robustness of RiskScore, the prognostic value of RiskScore was evaluated in the external datasets GSE28735, GSE57495, GSE62452, GSE85916, and ICGC-AU. We found that the OS of high RiskScore in the five datasets was significantly worse than that of the low-risk group ($p$< 0.05), and the 1-, 3-, and 5-year survival rates of RiskScore-predicting PAAD were all above 0.6 (Figure 9).

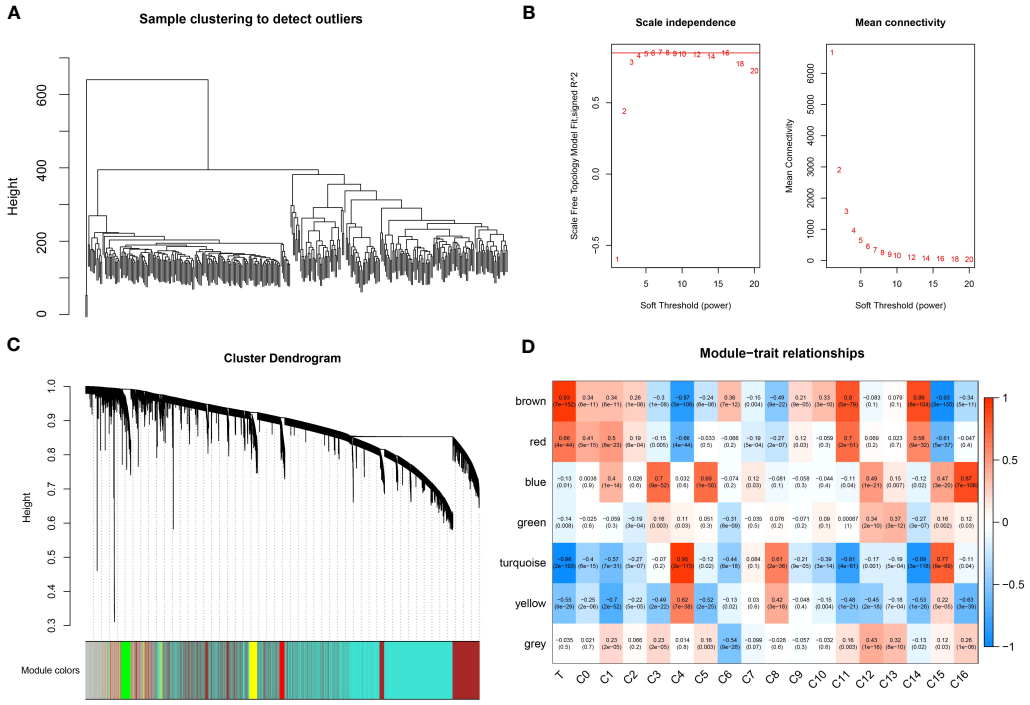



FIGURE 3
WGCNA analysis. **(A)** Sample clustering map. **(B)** Soft threshold β selection in a scale-free network. **(C)** Gene modules. **(D)** Correlation heatmap.

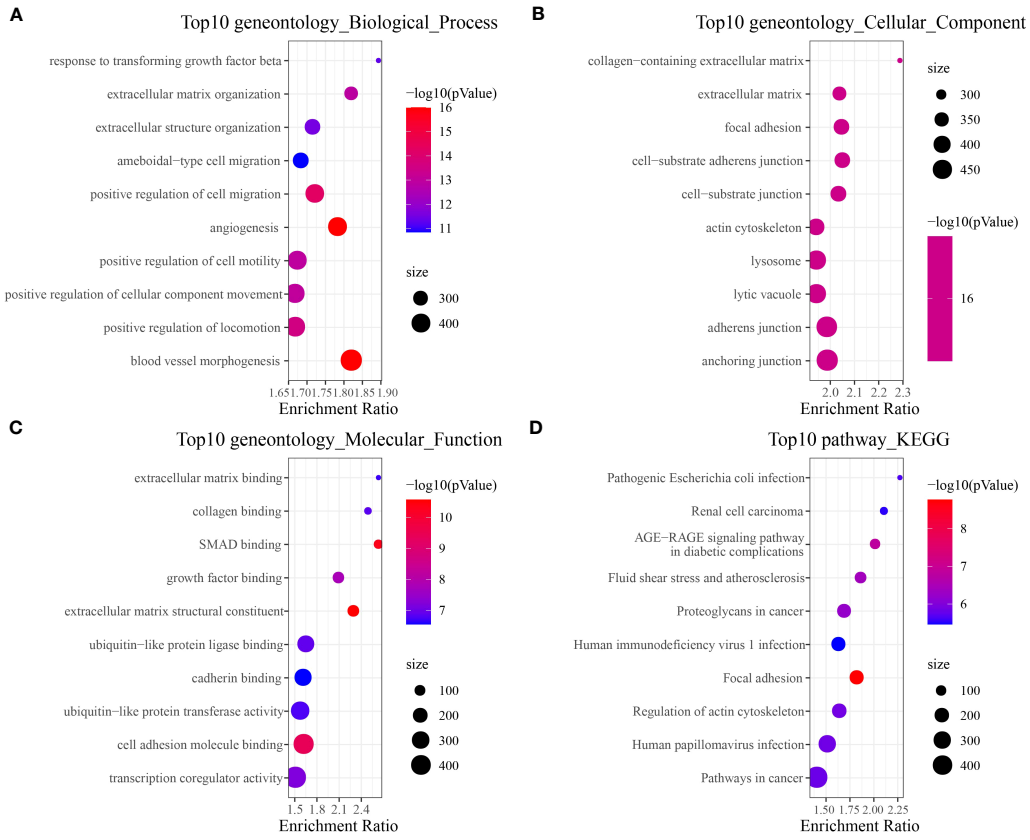



FIGURE 4
Brown module gene function enrichment analysis. **(A)** Biological process. **(B)** Cellular component. **(C)** Molecular function. **(D)** KEGG.

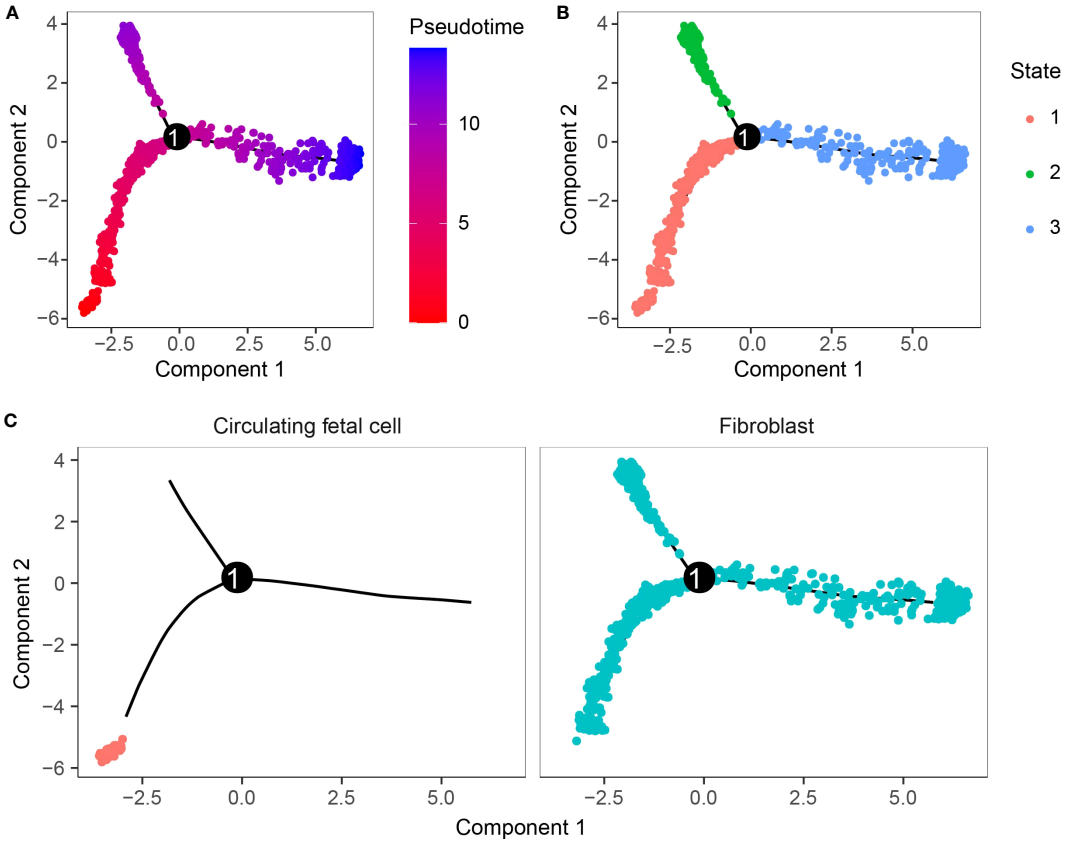



FIGURE 5
Cell trajectory analysis. **(A)** Pseudo-time measurement of developmental time. **(B)** Two cell subpopulations could differentiate into three branches. **(C)** Differentiation trajectory of cell clusters.

## Association between RiskScore and clinical features of PAAD

Clinical features, as traditional prognostic elements, were associated with the survival rate of cancer patients. In this study, we counted the distribution of RiskScore in patients with different clinical feature subgroups. We found a significant difference between RiskScore and T stage, N stage, and stages I–IV ($p< 0.05$), and the overall trend of increasing RiskScore with increasing stage. There was no significant difference between RiskScore and gender, M, stage, and age (Figure 10).

## Gene set enrichment analysis

To further investigate the relationship between RiskScore and biological function in different samples, ssGSEA enrichment analysis was performed for patients in the high and low RiskScore groups in the TCGA-PAAD cohort. Also, the Pearson's correlation analysis was performed between the ssGSEA score of each pathway and RiskScore; a total of 48 KEGG pathways were significantly correlated with RiskScore (correlation ≥ 0.5), among which six KEGG pathways were significantly negatively correlated with RiskScore, containing KEGG_RNA_ POLYMERASE, KEGG_PARKINSONS_DISEASE, KEGG_OXIDATIVE_PHOSPHORYLATION, KEGG_CARDIAC_MUSCLE_CONTRACTION, KEGG_GLYCOSYLPHOSPHATIDYLINOSITOL_GPI_ANCHOR_BIOSYNTHESIS, and KEGG_PROTEIN_EXPORT. RiskScore was strongly and positively correlated with 42 KEGG pathways (Figure 11A). Subsequent cluster analysis of the samples according to each KEGG pathway revealed that KEGG_BASAL_TRANSCRIPTION_FACTORS and KEGG_PROGESTERONE_MEDIATED_OOCYTE_MATURATION pathways increased with higher RiskScore scores (Figure 11B).

## Immune microenvironment analysis

To clarify the relationship between RiskScore and patients' immune microenvironment, we first used ESTIMATE to evaluate immune infiltration. The high-risk group had a higher StromalScore and ESTIMATEScore (Figure 12A). CIBERSORT analysis showed that the low-risk group had significantly enriched T_cells_CD8, NK_cells_activated, and B_cells_naive, and the high-risk group had significantly enriched Macrophages_M2 (Figure 12B). MCP-counter, TIMER, and EPIC analyses suggested that the high-risk group had higher immune infiltration (Figures 12C–E).

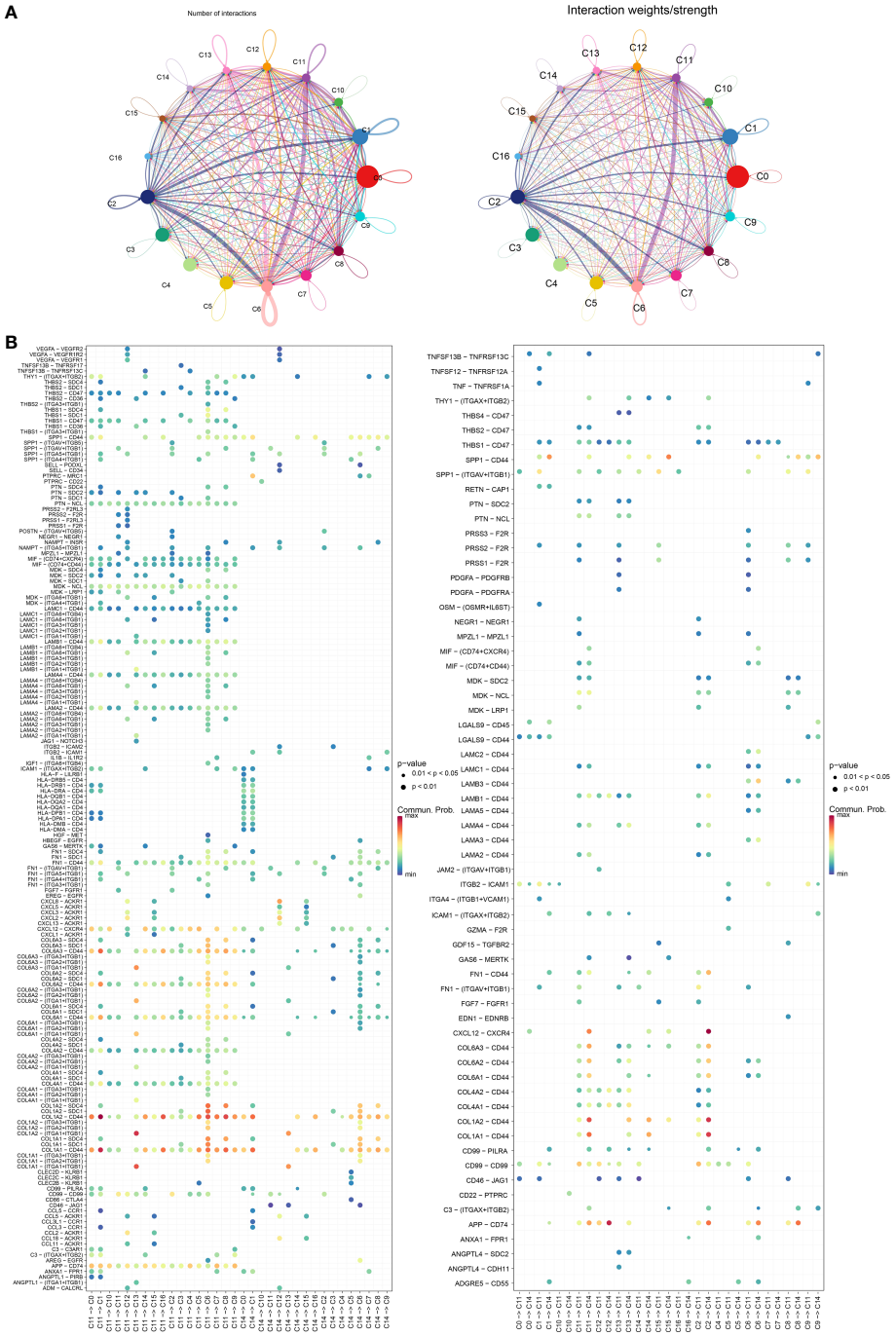



FIGURE 6
**(A)** Graph of changes in the number of receptors and ligands as well as intensity in cellular communication of 17 cell clusters. **(B)** Bubble diagram of receptors and ligands of C11 and C14 cell clusters with other cell clusters.

## Endocrine metabolism analysis

The 13 of 34 endocrine-related gene expressions differed in the high- and low-risk groups (Figure 13A). The low RiskScore group had a higher secretory pathway score. In addition, a negative phenomenon was observed between the secretory pathway score and the RiskScore (Figure 13B). Seven genes from the prognosis model were negatively correlated with the secretory pathway score (Figure 13C).

## Construction of nomogram

Univariate and multivariate regression analyses showed that age and RiskScore were independent prognostic factors (Figures 14A, B). We next combined age and RiskScore to build a nomogram, which could predict the prognosis of pancreatic cancer patients (Figure 14C). The nomogram shows that the 1- and 3-year prognosis lines are close to the 45° standard line, indicating good predictive performance (Figure 14D). The decision curve analysis

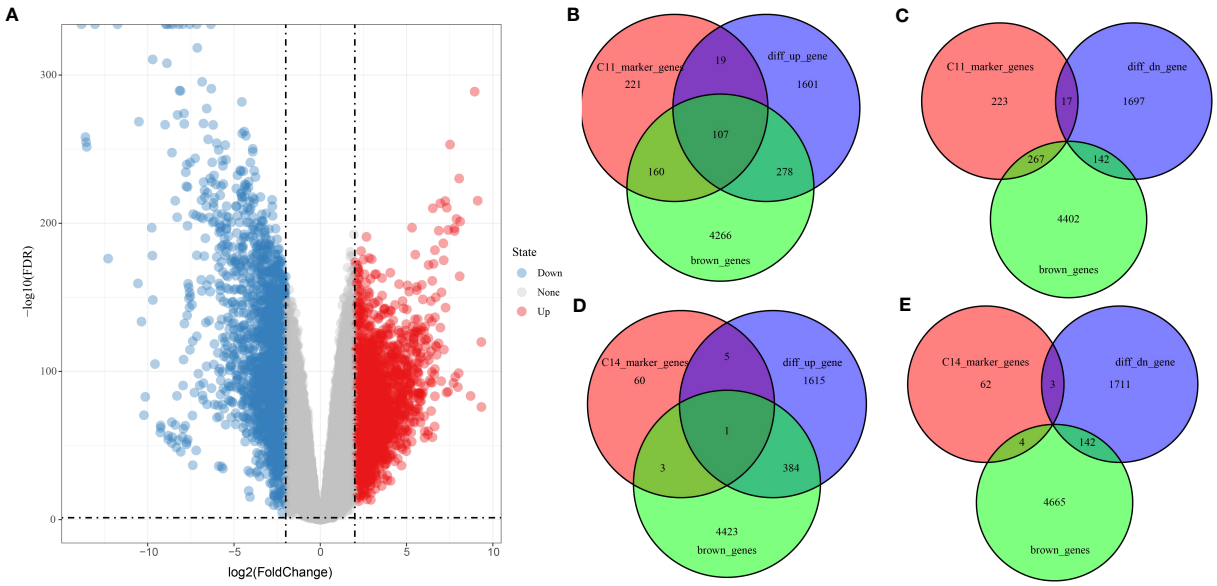



FIGURE 7
Screening of genes related to PAAD occurrence. **(A)** Volcano plot of DEGs between tumor and normal tissues in TCGA_GTEx-PAAD cohort. **(B, C)** Wayne plots of genes within C11 cell clusters, up- and downregulated DEGs, and brown modules. **(D, E)** Wayne plots of genes within C14 cell clusters, up- and downregulated DEGs, and brown modules.

(DCA) was employed to further confirm the clinical effectiveness of the nomogram, followed by RiskScore and age (Figure 14E).

## Discussion

In this study, we integrated PAAD scRNA-Seq data as well as RNA-Seq data to construct a promising prognostic tool (RiskScore) using genes associated with PAAD tumorigenesis in fibroblast and validated the generalizability of RiskScore with multiple datasets. We also explored the correlation between KEGG pathways significantly associated with RiskScore and clinical features.

Recent years have demonstrated that scRNA-Seq sequencing technology displays powerful advantages in probing the mechanism of tumorigenesis. Firstly, we identified 17 cell clusters with specific marker expression in intraductal papillary mucinous neoplasm, pancreatic adenosquamous carcinoma, and normal pancreas samples. In our study, we indicated that the C11 subpopulation belongs to fibroblast at the stage of tumor development. CAFs were the most abundant components of the tumor microenvironment and were heterogeneous, playing a pro- or anticancer role in different individual settings (23, 37). CAFs positively influenced cancer progression in tumors by mimicking or dominating the extracellular matrix (ECM) and thus remodeling the ECM structure. For one, the remodeled ECM structure served as a physical barrier for the infiltration of immune cells with killing functions, enhancing tumor killing, and for another, the ECM served as a structural scaffold for the interaction between tumor cells and stromal cells in the TME, promoting cardiac angiogenesis to regulate tumor metastasis (38). In this study, we also identified C11 subpopulation-related gene modules mainly associated with components or biological processes such as intercellular information exchange. Thus, it was the close communication between CAFs and tumor cells that might be responsible for the development of PAAD.

In this study, we also found that the C11 cluster specifically expresses HSPB6, which is currently focused on bladder urothelial carcinoma (BLCA). High HSPB6 expression was the critical factor for BLCA cell migration, and elevated HSPB6 expression inhibited BLCA cell migration (39). In contrast, the results of cell communication analysis demonstrated that the C11 cluster could be influenced by other cells by interacting with cell surface receptors via LAMC1. LAMC1 secretion was associated with the formation of inflammatory CAFs in esophageal squamous cell carcinoma, and upregulation of LAMC1 expression promoted CXCL1 secretion, which stimulated inflammatory CAFs via CXCR2-pSTAT3 and thus promoted tumor progression (40). Trajectory analysis showed consistent differentiation trends between the C11 and C2 clusters in fibroblasts, but C2 was not a tumorigenesis-associated cell cluster, and the distinction was that the specifically characterized genes were different, whereas the mechanism of HSPB6 in PAAD was unknown and its function in fibroblast was unclear. Our findings provided a new potential mechanism by which the C11 cluster-specific expression of HSPB6 may promote PAAD development.

We constructed the RiskScore tool to attempt to assess the prognosis of PAAD patients. We calculated the RiskScore based on the formula, and PAAD patients were divided into a high RiskScore group and a low RiskScore group. The results indicated that the RiskScore demonstrated a good prognostic value, and patients in the high RiskScore group had a worse prognosis. This result was validated in all five external datasets. We also performed ssGSEA

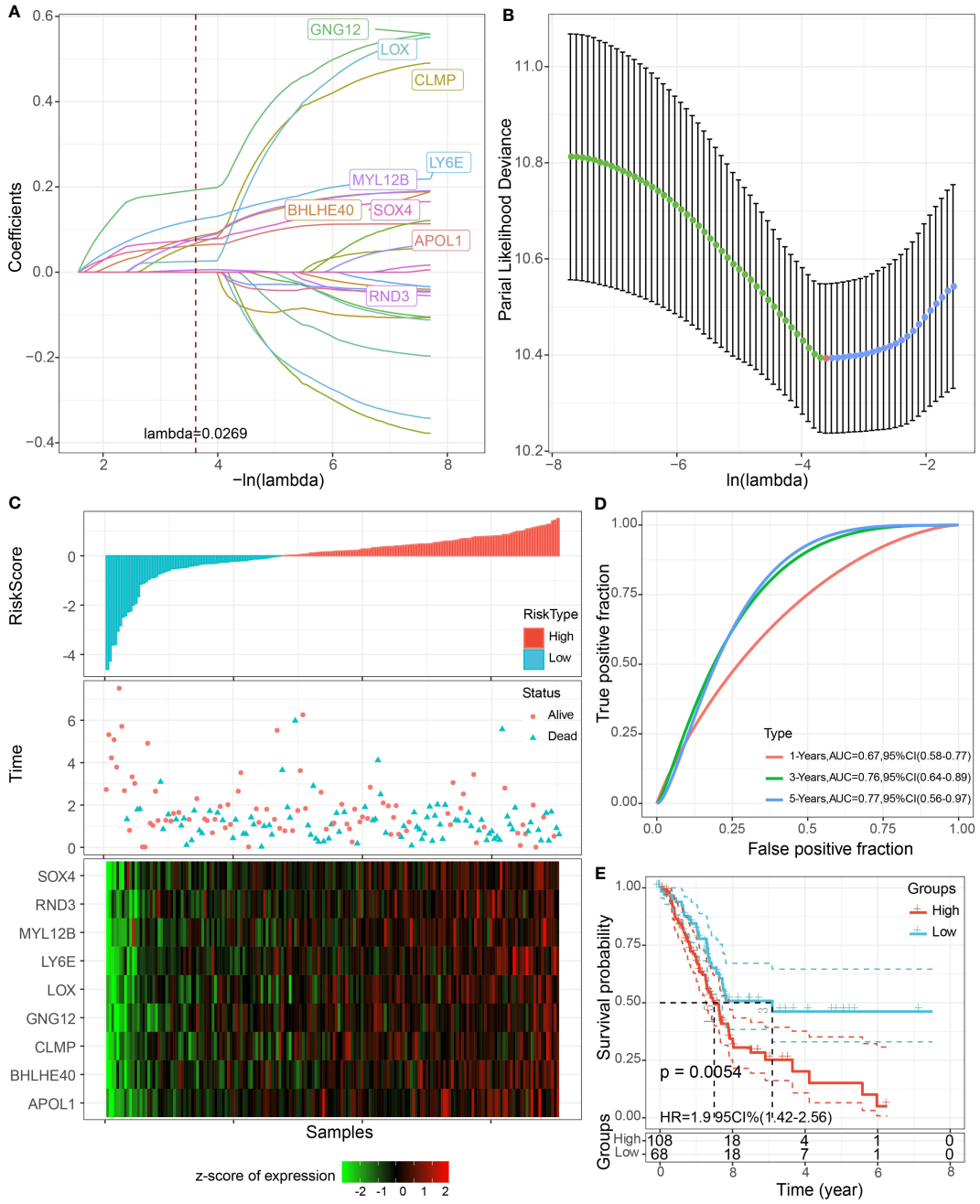



FIGURE 8
LASSO COX model construction. **(A)** Trajectory plot of independent variables with lambda. **(B)** The confidence interval of lambda. **(C)** Scatter plot of RiskScore distribution, survival status, and nine-gene expression heatmap of patients in TCGA-PAAD cohort. **(D)** ROC curves. **(E)** K-M curves of patients in high and low RiskScore groups.

analysis on samples from the high and low RiskScore groups, and basal transcription factors and progesterone-mediated oocyte maturation pathways were the characteristic pathways in the high RiskScore group. Moreover, RiskScore is negatively correlated with endocrine pathways, and the high-risk group had an enhanced immune infiltration status.

RiskScore consisted of APOL1, BHLHE40, CLMP, GNG12, LOX, LY6E, MYL12B, RND3, and SOX4, all of which were PAAD prognosis-associated genes. APOL1 was observed to be a critical enzyme in lipid functioning and metabolic processes and was found to be aberrantly highly expressed in hepatocellular carcinoma, small-cell lung cancer, and bladder cancer (41–44). Recent studies indicated that APOL1 exhibited oncogenic effects in PAAD, inhibiting PAAD cell apoptosis and promoting tumor cell proliferation through activation of the NOTCH1 signaling pathway (45), which was the first study reporting APOL1 function in PAAD.

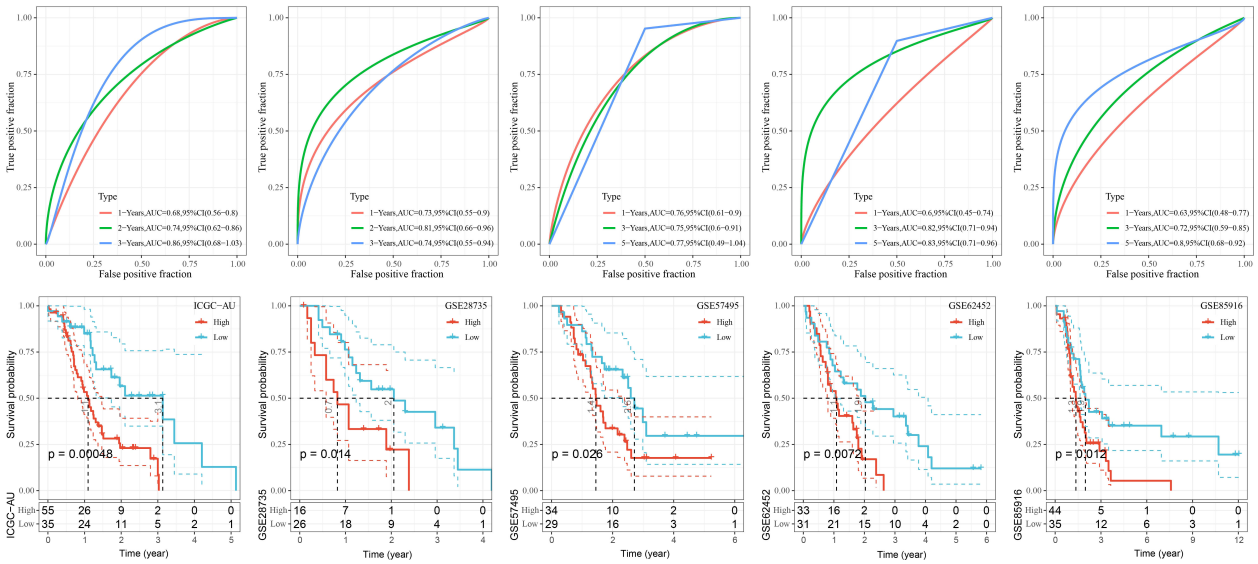



FIGURE 9
K-M curves as well as ROC curves for patients in the high and low RiskScore groups in the GSE28735, GSE57495, GSE62452, and GSE85916, ICGC-AU cohorts.

Overexpression of BHLHE40 caused the differentiation of tumor-associated neutrophils into a protumor subpopulation (TAN-1) and enhanced tumor immune suppression (46). CLMP was the central immune-related gene in colon cancer, associated with the inflammatory response, KRAS signaling pathway, and T-cell infiltration (47). Upregulation of pro-oncogenic MiR-106b-5p expression influenced survival outcomes in invasive breast cancer via suppression of GNG12 (48). LOX family genes were remodeling agents of hypoxia-induced ECM and were also pivotal inducers of chemotherapeutic drug resistance (49). The remaining genes were

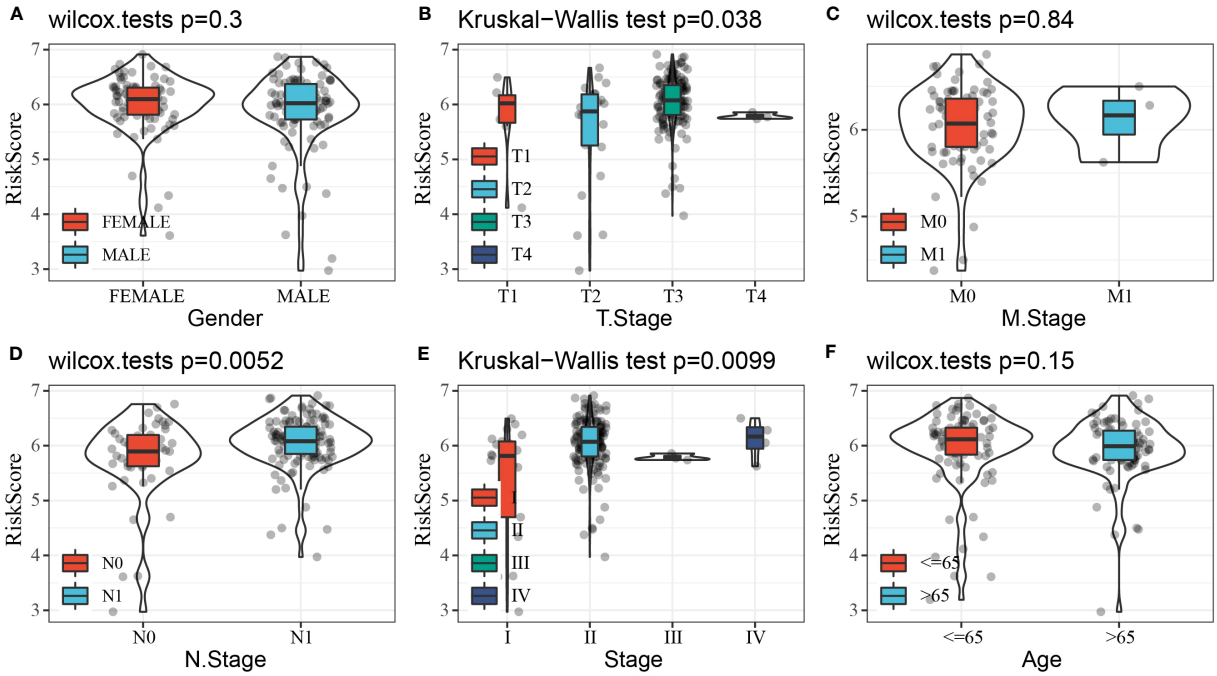



FIGURE 10
Distribution of RiskScore in subgroups of clinical features. **(A)** Gender. **(B)** T stage. **(C)** M stage. **(D)** N stage. **(E)** Stage. **(F)** Age.

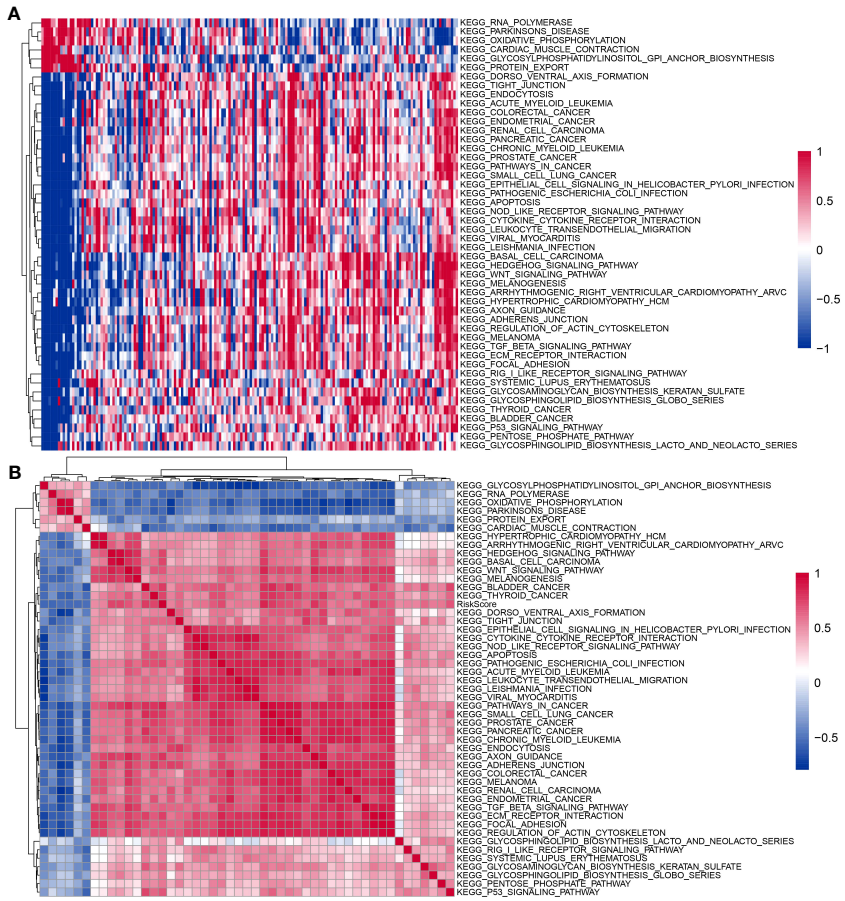



FIGURE 11
KEGG pathways affected by RiskScore. **(A)** Heatmap of clustering between 48 KEGG pathways and RiskScore. **(B)** Heatmap of KEGG pathways with changes in RiskScore.

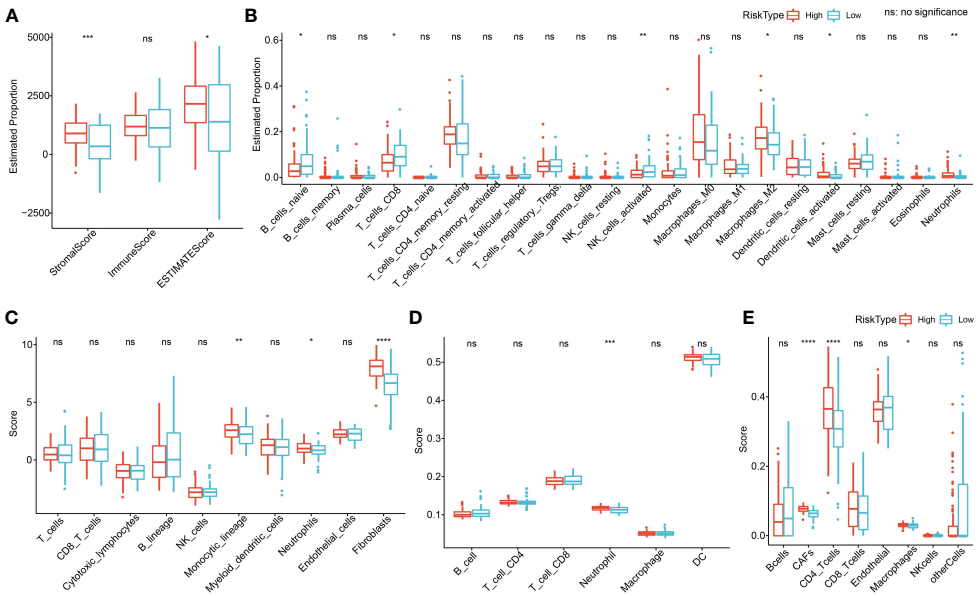



FIGURE 12
Immune microenvironment analysis. **(A)** ESTIMATE analysis. **(B)** CIBERSORT analysis shows a difference of 22 immune cells between high and low groups. **(C)** Using MCP-counter, we found 10 immune cell differences between high and low groups. **(D)** Using TIMER, six immune cell differences were found between high and low groups. **(E)** Using EPIC analysis, we found eight immune cell differences between high and low groups. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001.

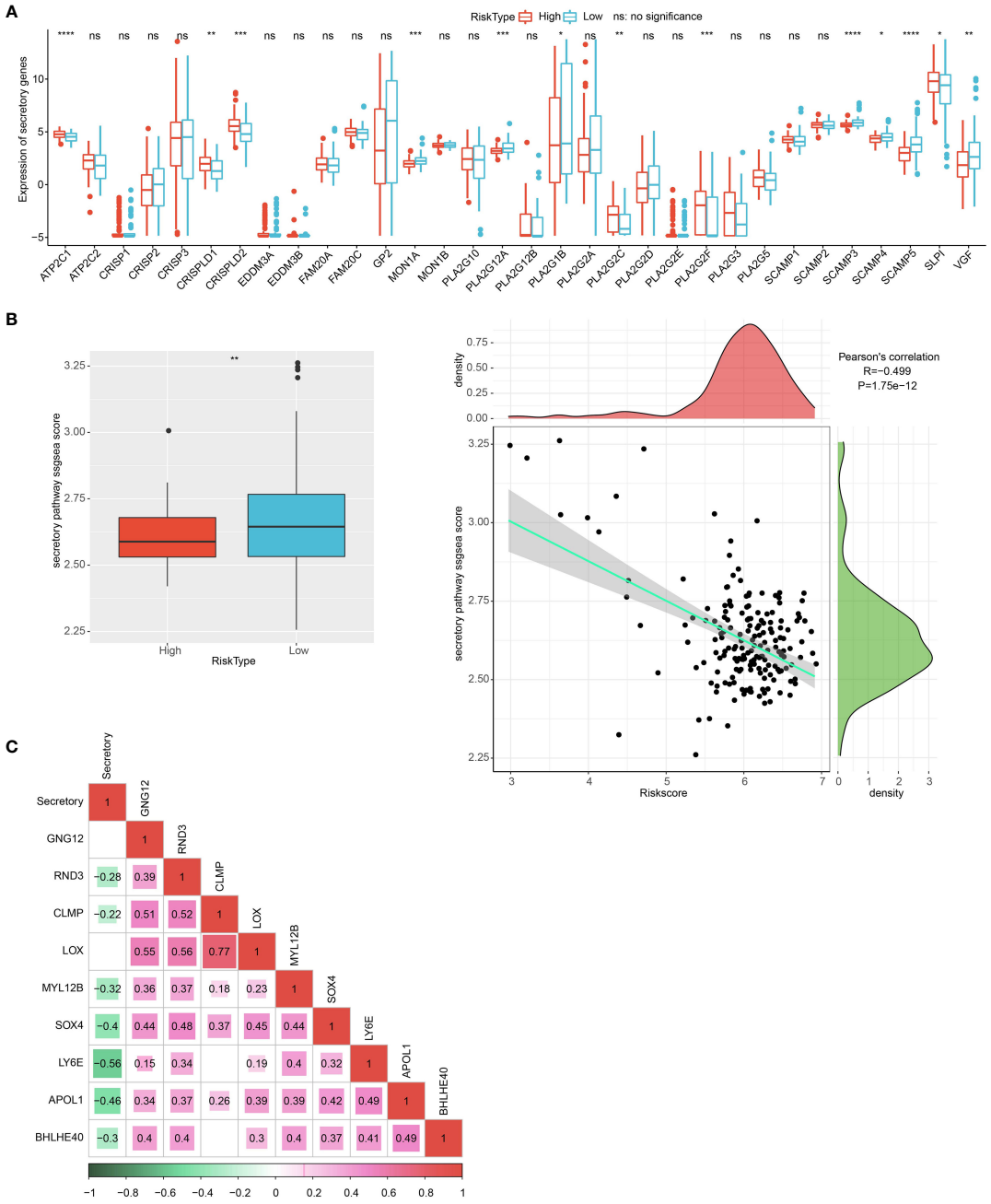



FIGURE 13
Endocrine metabolism analysis. **(A)** Difference of 34 endocrine-related genes between high and low group. **(B)** The difference in endocrine pathway scores between high and low groups. **(C)** The correlation analysis between genes in the prognosis model and the endocrine pathway score. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001.

also found to positively influence cancer progression (50–53). CLMP, GNG12, LOX, LY6E, MYL12B, and SOX4 were reported for the first time as prognostic signature genes for PAAD, and the mechanisms of how they regulate PAAD occurrence deserve further investigation.

Our study defined the critical cell cluster during PAAD genesis, which might promote tumor progression through frequent communication with tumor cells. In addition, we constructed a robust prognostic tool that demonstrated good robustness in predicting PAAD prognosis. However, this study was a

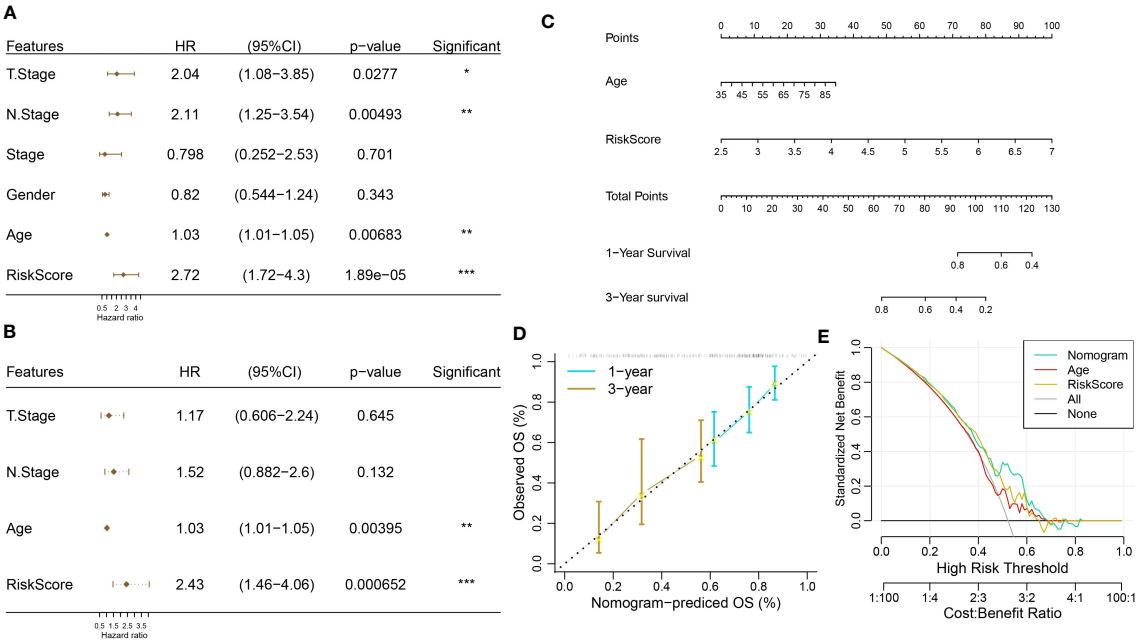



FIGURE 14
Construction of a nomogram. **(A, B)** Univariate and multivariate Cox regression analyses. **(C)** Construction of nomogram using age and RiskScore. **(D)** Calibration curve analysis. **(E)** Decision curve analysis (DCA).

comprehensive bioinformatic analysis conducted with public databases, and the molecular mechanisms of the C11 cluster and PAAD prognostic genes still remain to be further confirmed by relevant experiments as well as clinical trials.

## Conclusion

In conclusion, we identified the C11 cluster in fibroblasts that specifically expressed HSPB6 as the essential cluster for PAAD development and constructed a nine-gene prognostic model through tumor-associated PAAD prognostic genes in the C11 subpopulation. RiskScore might carry a credible clinical prognostic potential for PAAD.

## Data availability statement

All data generated or analyzed during this study are included in this published article.

## Author contributions

All authors contributed to this present work: YX and XC designed the study, and NL acquired the data. QW drafted the manuscript, and YX revised the manuscript. All authors read and approved the manuscript.

## Conflict of interest

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.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1201755/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Gene quality control map in the sample before filtering.

SUPPLEMENTARY FIGURE 2
Gene quality control plot in the sample after filtering.

SUPPLEMENTARY FIGURE 3
Distribution of high variant genes and non-high variant genes, the left panel showed the distribution of high variant genes and the right panel showed the distribution of non-high variant genes.

SUPPLEMENTARY FIGURE 4
PCA downscaling to find anchor points The left figure showed the anchor points in GSM5032771, GSM5032772, GSM5032773, and the right figure showed all the principal components in PCA analysis.

# References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660

2. Kamisawa T, Wood LD, Itoi T, Takaori K. Pancreatic cancer. *Lancet* (2016) 388 (10039):73–85. doi: 10.1016/S0140-6736(16)00141-0

3. Mizrahi JD, Surana R, Valle JW, Shroff RT. Pancreatic cancer. *Lancet* (2020) 395 (10242):2008–20. doi: 10.1016/S0140-6736(20)30974-0

4. Yamasaki A, Yanai K, Onishi H. Hypoxia and pancreatic ductal adenocarcinoma. *Cancer Lett* (2020) 484:9–15. doi: 10.1016/j.canlet.2020.04.018

5. Pasqualetti F, Sainato A, Morganti R, Laliscia C, Vasile E, Gonnelli A, et al. Adjuvant radiotherapy in patients with pancreatic adenocarcinoma. Is it still appealing in clinical trials? A meta-analysis and review of the literature. *Anticancer Res* (2021) 41 (10):4697–704. doi: 10.21873/anticanres.15283

6. Stojkovic Lalosevic M, Stankovic S, Stojkovic M, Markovic V, Dimitrijevic I, Lalosevic J, et al. Can preoperative CEA and CA19-9 serum concentrations suggest metastatic disease in colorectal cancer patients? *Hellenic J Nucl Med* (2017) 20(1):41–5. doi: 10.1967/s002449910505

7. Zhou G, Liu X, Wang X, Jin D, Chen Y, Li G, et al. Combination of preoperative CEA and CA19-9 improves prediction outcomes in patients with resectable pancreatic adenocarcinoma: results from a large follow-up cohort. *OncoTargets Ther* (2017) 10:1199–206. doi: 10.2147/OTT.S116136

8. Zhu L, Xue HD, Liu W, Wang X, Sui X, Wang Q, et al. Enhancing pancreatic mass with normal serum CA19-9: key MDCT features to characterize pancreatic neuroendocrine tumours from its mimics. *La Radiologia medica* (2017) 122(5):337–44. doi: 10.1007/s11547-017-0734-x

9. Storz P. Acinar cell plasticity and development of pancreatic ductal adenocarcinoma. *Nat Rev Gastroenterol hepatol* (2017) 14(5):296–304. doi: 10.1038/nrgastro.2017.12

10. Asa SL. Pancreatic endocrine tumors. *Modern Pathol* (2011) 24 Suppl 2:S66–77. doi: 10.1038/modpathol.2010.127

11. Navin NE. The first five years of single-cell cancer genomics and beyond. *Genome Res* (2015) 25(10):1499–507. doi: 10.1101/gr.191098.115

12. Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* (2017) 541(7637):331–8. doi: 10.1038/nature21350

13. Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat Rev Cancer* (2017) 17(9):557–69. doi: 10.1038/nrc.2017.58

14. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res* (2019) 29(9):725–38. doi: 10.1038/s41422-019-0195-y

15. Moncada R, Barkley D, Wagner F, Chiodin M, Devlin JC, Baron M, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat Biotechnol* (2020) 38 (3):333–42. doi: 10.1038/s41587-019-0392-8

16. Lu J, Chen Y, Zhang X, Guo J, Xu K, Li L. A novel prognostic model based on single-cell RNA sequencing data for hepatocellular carcinoma. *Cancer Cell Int* (2022) 22(1):38. doi: 10.1186/s12935-022-02469-2

17. Wang G, Qiu M, Xing X, Zhou J, Yao H, Li M, et al. Lung cancer scRNA-seq and lipidomics reveal aberrant lipid metabolism for early-stage diagnosis. *Sci Transl Med* (2022) 14(630):eabk2756. doi: 10.1126/scitranslmed.abk2756

18. Li X, Sun Z, Peng G, Xiao Y, Guo J, Wu B, et al. Single-cell RNA sequencing reveals a pro-invasive cancer-associated fibroblast subgroup associated with poor clinical outcomes in patients with gastric cancer. *Theranostics* (2022) 12(2):620–38. doi: 10.7150/thno.60540

19. Chen Y, McAndrews KM, Kalluri R. Clinical and therapeutic relevance of cancer-associated fibroblasts. *Nat Rev Clin Oncol* (2021) 18(12):792–804. doi: 10.1038/s41571-021-00546-5

20. LeBleu VS, Kalluri R. A peek into cancer-associated fibroblasts: origins, functions and translational impact. *Dis Models Mech* (2018) 11(4):dmm029447. doi: 10.1242/dmm.029447

21. Hosein AN, Brekken RA, Maitra A. Pancreatic cancer stroma: an update on therapeutic targeting strategies. *Nat Rev Gastroenterol hepatol* (2020) 17(8):487–505. doi: 10.1038/s41575-020-0300-1

22. Sahai E, Astsaturov I, Cukierman E, DeNardo DG, Egeblad M, Evans RM, et al. A framework for advancing our understanding of cancer-associated fibroblasts. *Nat Rev Cancer* (2020) 20(3):174–86. doi: 10.1038/s41568-019-0238-1

23. Kalluri R. The biology and function of fibroblasts in cancer. *Nat Rev Cancer* (2016) 16(9):582–98. doi: 10.1038/nrc.2016.73

24. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med* (2021) 13(1):152. doi: 10.1186/s13073-021-00968-x

25. Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell* (2023) 186(8):1772–91. doi: 10.1016/j.cell.2023.01.035

26. Tharwat M, Sakr NA, El-Sappagh S, Soliman H, Kwak KS, Elmogy M. Colon cancer diagnosis based on machine learning and deep learning: modalities and analysis techniques. *Sensors (Basel Switzerland)* (2022) 22(23):9250. doi: 10.3390/s22239250

27. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* (2015) 161(5):1202–14. doi: 10.1016/j.cell.2015.05.002

28. Shen W, Song Z, Zhong X, Huang M, Shen D, Gao P, et al. Sangerbox: A comprehensive, interaction-friendly clinical bioinformatics analysis platform. *iMeta* (2022) 1(3):e36. doi: 10.1002/imt2.36

29. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* (2021) 2(3):100141. doi: 10.1016/j.xinn.2021.100141

30. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* (2015) 12 (5):453–7. doi: 10.1038/nmeth.3337

31. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf* (2008) 9:559. doi: 10.1186/1471-2105-9-559

32. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* (2019) 566(7745):496–502. doi: 10.1038/s41586-019-0969-x

33. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and analysis of cell-cell communication using CellChat. *Nat Commun* (2021) 12(1):1088. doi: 10.1038/s41467-021-21246-9

34. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* (2015) 43(7):e47. doi: 10.1093/nar/gkv007

35. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Software* (2010) 33(1):1–22. doi: 10.18637/jss.v033.i01

36. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for cox's proportional hazards model via coordinate descent. *J Stat Software* (2011) 39(5):1–13. doi: 10.18637/jss.v039.i05

37. Mueller MM, Fusenig NE. Friends or foes - bipolar effects of the tumour stroma in cancer. *Nat Rev Cancer* (2004) 4(11):839–49. doi: 10.1038/nrc1477

38. Yang X, Lin Y, Shi Y, Li B, Liu W, Yin W, et al. FAP promotes immunosuppression by cancer-associated fibroblasts in the tumor microenvironment via STAT3-CCL2 signaling. *Cancer Res* (2016) 76(14):4124–35. doi: 10.1158/0008-5472.CAN-15-2973

39. Chen S, Huang H, Yao J, Pan L, Ma H. Heat shock protein B6 potently increases non-small cell lung cancer growth. *Mol Med Rep* (2014) 10(2):677–82. doi: 10.3892/mmr.2014.2240

40. Fang L, Che Y, Zhang C, Huang J, Lei Y, Lu Z, et al. LAMC1 upregulation via TGFbeta induces inflammatory cancer-associated fibroblasts in esophageal squamous cell carcinoma via NF-kappaB-CXCL1-STAT3. *Mol Oncol* (2021) 15(11):3125–46. doi: 10.1002/1878-0261.13053

41. Thomson R, Genovese G, Canon C, Kovacsics D, Higgins MK, Carrington M, et al. Evolution of the primate trypanolytic factor APOL1. *Proc Natl Acad Sci USA* (2014) 111(20):E2130–9. doi: 10.1073/pnas.1400699111

42. Shi J, Yang H, Duan X, Li L, Sun L, Li Q, et al. Apolipoproteins as differentiating and predictive markers for assessing clinical outcomes in patients with small cell lung cancer. *Yonsei Med J* (2016) 57(3):549–56. doi: 10.3349/ymj.2016.57.3.549

43. Bharali D, Banerjee BD, Bharadwaj M, Husain SA, Kar P. Expression analysis of apolipoproteins AI & AIV in hepatocellular carcinoma: A protein-based hepatocellular carcinoma-associated study. *Indian J Med Res* (2018) 147(4):361–8. doi: 10.4103/ijmr.IJMR_1358_16

44. Ma XL, Gao XH, Gong ZJ, Wu J, Tian L, Zhang CY, et al. Apolipoprotein A1: a novel serum biomarker for predicting the prognosis of hepatocellular carcinoma after curative resection. *Oncotarget* (2016) 7(43):70654–68. doi: 10.18632/oncotarget.12203

45. Lin J, Xu Z, Xie J, Deng X, Jiang L, Chen H, et al. Oncogene APOL1 promotes proliferation and inhibits apoptosis via activating NOTCH1 signaling pathway in pancreatic cancer. *Cell Death Dis* (2021) 12(8):760. doi: 10.1038/s41419-021-03985-1

46. Wang L, Liu Y, Dai Y, Tang X, Yin T, Wang C, et al. Single-cell RNA-seq analysis reveals BHLHE40-driven pro-tumour neutrophils with hyperactivated glycolysis in pancreatic tumour microenvironment. *Gut* (2022) 72(5):958-71. doi: 10.1136/gutjnl-2021-326070

47. Wang X, Duanmu J, Fu X, Li T, Jiang Q. Analyzing and validating the prognostic value and mechanism of colon cancer immune microenvironment. *J Transl Med* (2020) 18(1):324. doi: 10.1186/s12967-020-02491-w

48. Farre PL, Duca RB, Massillo C, Dalton GN, Grana KD, Gardner K, et al. MiR-106b-5p: A master regulator of potential biomarkers for breast cancer aggressiveness and prognosis. *Int J Mol Sci* (2021) 22(20):11135. doi: 10.3390/ijms222011135

49. Saatci O, Kaymak A, Raza U, Ersan PG, Akbulut O, Banister CE, et al. Targeting lysyl oxidase (LOX) overcomes chemotherapy resistance in triple negative breast cancer. *Nat Commun* (2020) 11(1):2416. doi: 10.1038/s41467-020-16199-4

50. AlHossiny M, Luo L, Frazier WR, Steiner N, Gusev Y, Kallakury B, et al. Ly6E/K signaling to TGFbeta promotes breast cancer progression, immune escape, and drug resistance. *Cancer Res* (2016) 76(11):3376–86. doi: 10.1158/0008-5472.CAN-15-2654

51. Dabrowska M, Skoneczny M, Rode W. Functional gene expression profile underlying methotrexate-induced senescence in human colon cancer cells. *Tumour Biol* (2011) 32(5):965–76. doi: 10.1007/s13277-011-0198-x

52. Wu N, Zheng F, Li N, Han Y, Xiong XQ, Wang JJ, et al. RND3 attenuates oxidative stress and vascular remodeling in spontaneously hypertensive rat via inhibiting ROCK1 signaling. *Redox Biol* (2021) 48:102204. doi: 10.1016/j.redox.2021.102204

53. Good CR, Aznar MA, Kuramitsu S, Samareh P, Agarwal S, Donahue G, et al. An NK-like CAR T cell transition in CAR T cell dysfunction. *Cell* (2021) 184(25):6081–100 e26. doi: 10.1016/j.cell.2021.11.016

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 17 August 2023
DOI 10.3389/fendo.2023.1170526



OPEN ACCESS

EDITED BY
Wenjie Shi,
Otto von Guericke University Magdeburg, Germany

REVIEWED BY
Wei Li,
Shenzhen Longhua District Central Hospital, China
Zigang Zhao,
Hainan Hospital of PLA General Hospital, China

*CORRESPONDENCE
Lei Liang
leiliang@jnu.edu.cn
Xuhui Zhang
zhangxh@gd2h.org.cn

†These authors have contributed equally to this work

RECEIVED 21 February 2023
ACCEPTED 04 July 2023
PUBLISHED 17 August 2023

CITATION
Zeng X, Sun L, Ling X, Jiang Y, Shen J, Liang L and Zhang X (2023) Comprehensive analysis identifies novel targets of gemcitabine to improve chemotherapy treatment strategies for colorectal cancer.
*Front. Endocrinol.* 14:1170526.
doi: 10.3389/fendo.2023.1170526



# Comprehensive analysis identifies novel targets of gemcitabine to improve chemotherapy treatment strategies for colorectal cancer

Xinxin Zeng[1†], Liyue Sun[1†], Xiaomei Ling[2†], Yuying Jiang[1,3], Ju Shen[1,3], Lei Liang[4*] and Xuhui Zhang[1*]

[1]Second Department of Oncology, The Affiliated Guangdong Second Provincial General Hospital of Jinan University, Guangzhou, China, [2]Medical Research Center, The Affiliated Guangdong Second Provincial General Hospital of Jinan University, Guangzhou, China, [3]Department of Radiation Oncology, Guangdong Medical University, Zhanjiang, China, [4]Guangdong Engineering Research Center of Chinese Medicine & Disease Susceptibility, School of Traditional Chinese Medicine, Jinan University, Guangzhou, China

**Background:** Gemcitabine (GEM) is a second-line anticancer drug of choice for some colorectal cancer (CRC) patients, and GEM inability to be commonly available in the clinic due to the lack of clarity of the exact action targets.

**Methods:** The half maximal inhibitory concentration (IC50) of GEM treatment for 42 CRC cell lines were accessed from the Genomics of Drug sensitivity in Cancer (GDSC) database. High-throughput sequencing data of CRC patients were captured in The Cancer Genome Atlas (TCGA) and Weighted correlation network analysis (WGCNA) was conducted. Pearson correlations were derived for GEM potency-related genes. Differential analysis was conducted in the TCGA cohort to obtain CRC development-related genes (CDRGs), and univariate COX model analysis was performed on CDRGs overlapping with GEM potency-related genes to obtain CDRGs affecting CRC prognosis. Hub genes affecting GEM potency were identified by Spearman correlation.

**Results:** CALB2 and GPX3 were identified as potential targets for GEM treatment of CRC *via* prognostic analysis, which we also observed to be elevated with elevated clinical stage in CRC patients. The enhanced expression of CALB2 and GPX3 genes identified in the pathway analysis might inhibit the body metabolism as well as activate immune and inflammation related pathways. In addition, we found that CALB2 and GPX3 could also be considered as prognostic biomarkers in pan-cancer. Finally, we found that CALB2 and GPX3 were remarkably associated with the drug sensitivity of MG-132, Dasatinib, Shikonin, Midostaurin, MS-275, and Z-LNle-CHO, which were expected to be the drugs of choice for GEM combination.

**Conclusion:** CALB2 and GPX3 represent prognostic biomarkers for CRC and they might be potential action targets for GEM. Our study offered innovative ideas for GEM administration strategies.

KEYWORDS

colorectal cancer, chemotherapy, gemcitabine, combination drug, CALB2, GPX3

10.3389/fendo.2023.1170526

# Introduction

Genetic alterations resulting from somatic mutations or gene fusions contributed to colorectal cancer (CRC) being a highly heterogeneous cancer due to the coexistence of multiple pathogenic mechanisms (1). Over million people developed CRC and hundreds of thousands of CRC patients died yearly (2). The reason for the high mortality rate of CRC was that most patients already had metastases when diagnosed (2). Currently, surgical resection was the dominant treatment option for CRC, and chemotherapy was generally considered for patients with local metastases, but tumor heterogeneity caused some CRC patients to develop chemotherapy resistance (3). Identifying effective treatment modalities is crucial to improve survival in CRC.

Gemcitabine (GEM) was second-line resistant drug, and high resistance limited the applicability of GEM in the clinic (4). GEM could hardly be treated as first-line chemotherapeutic agent due to enzymatic deamination, low clearance and high resistance (4). Recent report by Chocry et al. (5) illustrated that GEM was a potential alternative drug when CRC patients developed Oxaliplatin resistance. GEM remained an alternative option for CRC patients. In recent years, several studies have focused on the administration of GEM to tumor cells using nanotechnology to enhance the efficacy of GEM (6). It was evident that current drug delivery strategies and the absence of an exact drug target were the major limiting factors for GEM. Studies suggested that novel drug delivery modalities could assist GEM as cancer-targeted drugs, but related studies were still exploring (7). However, studies focusing on marker genes identification from GEM-related genes in CRC remain limited.

RNA sequencing (RNA-Seq) is a high-throughput sequencing technology used to study transcriptomes, enabling more accurate quantification of gene expression levels, both to identify novel transcript sequences and for differential expression studies (8). In this study, based on multiple databases, we investigated the hub genes of GEM acting in CRC. we comprehensively explored the association between hub genes and multiple cancer prognosis, tumor-infiltrating immune cells (TIIC) in the tumor microenvironment (TME), and chemotherapeutic drug sensitivity. Our study explored the potential of GEM as CRC-targeting agent and the potential contribution of these hub genes in CRC prognosis.

# Materials and methods

## Data acquisition and pre-processing

Transcriptome high-throughput sequencing datasets of CRC patients as well as normal tissues and corresponding clinical phenotype data were obtained from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) website. The gene expression was showed as log2(TPM+1). Tumor simples whose survival time was absence and less than 0 days of survival were removed, and 432 tumor samples as well as 41 normal tissue samples were retained. The microarray sequencing datasets GSE17536, GSE17537, GSE17538, GSE39582 of CRC patients were loaded from the GENE EXPRESSION OMNIBUS (GEO, https://www.ncbi.nlm.nih.gov/geo/) website. Normal tissue samples, samples with missing clinical follow-up information, and survival time less than 0 days were excluded in 4 cohorts. 177, 55, 232, and 573 tumor samples were retained in GSE17536, GSE17537, GSE17538, and GSE39582, respectively, for follow-up studies. The clinical information was showed in Table 1. The half maximal inhibitory concentration (IC50) data for 42 CRC cell lines treated with GEM were accessed from the Genomics of Drug sensitivity in Cancer database (GDSC, https://www.cancerrxgene.org/).

## WGCNA Analysis

In this study, Weighted correlation network analysis (WGCNA) was performed on genes in the TCGA dataset by referring to

TABLE 1 The clinical information of TCGA dataset.

| TCGA | | |
|---|---|---|
| Gender | | |
| | male | 232 |
| | female | 200 |
| T stage | | |
| | T1 | 11 |
| | T2 | 75 |
| | T3 | 295 |
| | T4 | 50 |
| | Unknown | 1 |
| N stage | | |
| | N0 | 253 |
| | N1 | 102 |
| | N2 | 77 |
| M stage | | |
| | M0 | 319 |
| | M1 | 61 |
| | Unknown | 52 |
| Stage | | |
| | I | 73 |
| | II | 164 |
| | III | 123 |
| | IV | 61 |
| | Unknown | 11 |
| Status | | |
| | Alive | 339 |
| | Dead | 93 |
| Age | <=65 | 183 |
| | >65 | 249 |

the method of Langfelder et al. (9) using WGCNA R package (9). The parameters were set: correlation coefficient > 0.85, minimum number of module genes > 50. After merging similar gene modules, principal component analysis (PCA) was performed on the final gene modules, the first principal component of each module was analyzed as Module eigengene E with IC50 values of GEM for pearson correlation analysis to determine the gene modules affecting GEM potency, and the GEM potency-related genes within the modules were included for subsequent analysis.

## Identification of CRC development-related genes

In the TCGA cohort, differential analysis was performed using the limma package (10) to identify CRC development-related genes (CDRGs) in tumor tissues using normal tissues as controls. CDRGs were intersected with GEM potency-related genes to obtain candidate CDRGs affecting GEM potency. univariate COX model analysis based on the expression matrix of these CDRGs and the survival information of CRC patients in the TCGA cohort was performed to identify candidate hub genes associated with CRC survival. Finally, based on the expression levels of candidate hub genes, Spearman correlation between them and IC50 values of GEM was assessed to determine the hub genes of GEM for CRC treatment.

## Prognostic impact of hub genes on CRC

In the TCGA, GSE17536, GSE17537, GSE17538, and GSE39582 cohorts, CRC patients were clustered into high and low expression groups using the survminer code package (https://rpkgs.datanovia.com/survminer/index.html) to determine the optimal group cut-off values, and Kaplan-Meier (K-M) survival curves were plotted for patients in the high- and low-expression groups using survminer R package (11).

## Association between hub genes and CRC clinical phenotypes

In the TCGA cohort, the expression levels of hub genes were compared among patients in the Stage, TNM. Stage subgroups to explore the association between hub genes and CRC clinical phenotypes.

## Gene set variation analysis

In the TCGA dataset, we performed Gene Set Variation Analysis (GSVA) using the GSVA code package (12) to resolve CALB2 and GPX3 regulated pathways. To calculate potential connections between CALB2 and GPX3 and their regulatory pathways, we performed spearman correlation analysis between CALB2 and GPX3 expression levels and GSVA scores of the pathways to mine the pathways markedly associated with CALB2 and GPX3.

## Connection between CALB2 and GPX3 and TIIC

In the TCGA dataset, we used the Estimation of Stromal and Immune cells in Malignant Tumours using Expression data (ESTIMATE) algorithm (13) to assess the TIICs in TME of CRC patients with ImmuneScore, StromalScore of stromal cells, and ESTIMATEScore. the CIBERSORT algorithm (14) was utilized to assess the relative infiltration scores of 22 TIICs in TME and to calculate the spearman correlation between CALB2 and GPX3 and TIICs. Further, 28 signatures in pan-cancer that could predict Checkpoint Blockade response were captured from the research of Charoentong et al. (15) and ssGSEA Score was calculated (16). Finally, the correlation between CALB2 and GPX3 and the 28 signatures capable of predicting Checkpoint Blockade response was assessed before using the mantel test and pearson correlation.

## Prognostic utility of CALB2 and GPX3 in pan-cancer

The expression profiles of 32 cancers were downloaded from Sangerbox (http://vip.sangerbox.com) (17) and the expression levels of CALB2 and GPX3 in tumor tissues and normal tissues were assessed using the wilcox test. The survival time and survival status of patients with 32 cancers in TCGA were extracted from the study of Liu et al. (18), and the prognostic role of CALB2 and GPX3 was assessed by plotting K-M survival curves for the groups using the optimal group cut-off values obtained from the survminer code package.

## Pharmaceutical sensitivity analysis of CALB2 and GPX3

In the TCGA cohort, we utilized the pRRophetic code package (19) to predict the IC50 for 51 chemotherapeutic agents in the high- and low-expression groups of CRC patients with CALB2 and GPX3. p-values of the IC50 for the drugs were examined by wilcox.test and histograms were plotted. We screened the spearman correlation between the three groups of drugs with the largest and smallest IC50 and CALB2 and GPX3.

## Statistical analysis

All statistical analyses were performed by R software (version 3.62). Wilcoxon nonparametric rank sum test was used to analyze the differences, and a P-value < 0.05 was considered significant unless otherwise specified.

# Results

## Identification of GEM potency-related genes

The workflow was showed in Figure S1. Firstly, we extracted the IC50 data of 42 CRC cell lines in response to GEM from the GDSC2 database (Figure 1A). Based on the dynamic shear tree algorithm, 16 gene modules were identified *via* WGCNA analysis by selecting a soft threshold β=4 to construct a scale-free network (Figures 1B, C). The GEM drug IC50 data of 42 CRC cells were considered as clinical data, and Pearson correlation analysis was performed with the first principal component Module eigengene E of the 16 gene modules to select the most relevant gene modules for GEM efficacy. We found that genes within the blue and magenta modules were remarkably negatively correlated with GEM efficacy (Figure 1D), and this result suggested that genes within these two modules might be potential target genes for GEM treatment of CRC. Therefore, blue and magenta intramodule genes were selected for subsequent study.

## Hub genes influencing GEM potency

To further identify hub genes for GEM potency, we identified CDRGs in the TCGA dataset. 2664 CDRGs were identified by differential analysis, including 1416 up-regulated CDRGs and 1248 down-regulated CDRGs (Figure 2A). Subsequently, candidate CDRGs affecting GEM potency were identified by Venn diagram analysis. we intersected the CDRGs in TCGA with genes within the blue and magenta modules, respectively. there were 56 up-regulated basal CDRGs and 19 down-regulated CDRGs in the blue module, and 34 up-regulated CDRGs and 11 down-regulated CDRGs in the magenta module (Figure 2B). These 120 CDRGs may be potential hub genes affecting the potency of GEM. we demonstrated the expression levels of 120-CDRGs in 42 CRC cells by heat map (Figure 2C). 9 CDRGs associated with CRC prognosis was identified by univariate COX regression model (Figure 2D). Finally, Spearman correlation analysis based on the expression levels of the 9-CDRGs with the IC50 values of GEM was conducted. We identified C4orf19, GPX3, C20orf27, AADAT and CALB2 as hub genes affecting the potency of GEM, with C4orf19, GPX3 and C20orf27 showing remarkable positive correlation with IC50 of GEM, and AADAT and CALB2 showing remarkable negative correlation with IC50 of GEM (Figure 2E). Overall, these results suggest that C4orf19, GPX3, C20orf27, AADAT, and CALB2 might be the candidate hub genes for GEM treatment of CRC.

## Correlation of 5-hub genes with CRC prognosis and clinical information

We found that high expression of C4orf19 and AADAT resulted the promising prognosis of CRC patients, and low expression of GPX3, C20orf27 and CALB2 resulted to better prognosis of CRC patients (Figure 3A). Subsequently, we further validated the relationship between 5-hub genes and CRC prognosis in four external GEO datasets (GSE17536, GSE17537, GSE17538,

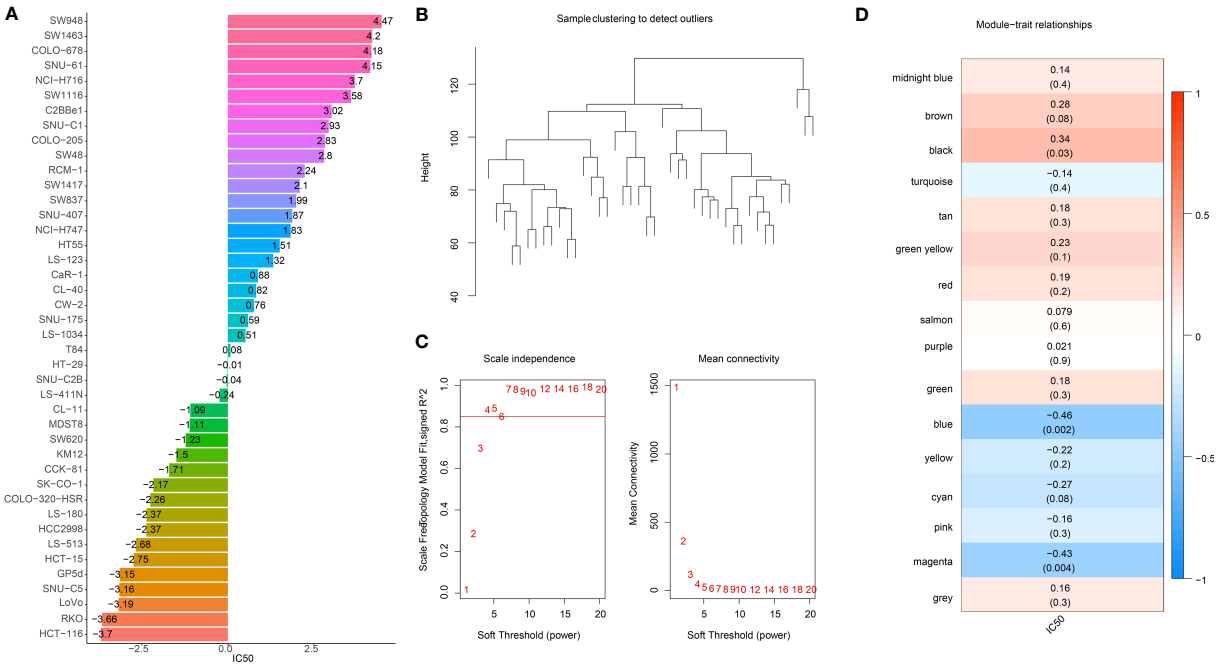



FIGURE 1
Identification of GEM potency-related genes. **(A)** IC50 of Gemcitabine-treated CRC cells **(B)** Clustering tree of TCGA samples **(C)** Scale-free network analysis **(D)** Heatmap of IC50 of Gemcitabine with pearson correlation of Module eigengene E of gene module.

10.3389/fendo.2023.1170526

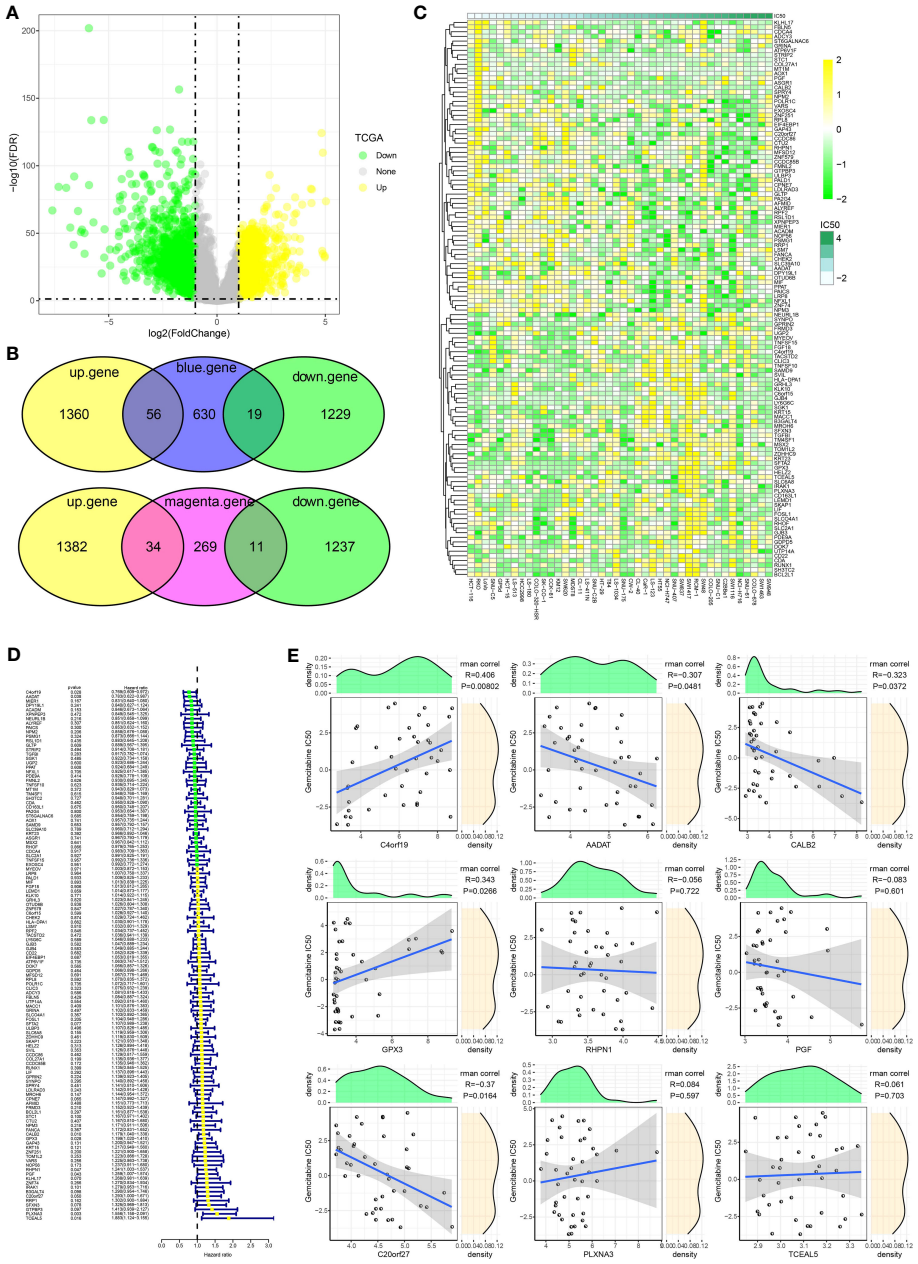



FIGURE 2
Hub genes influencing GEM potency. **(A)** Volcano map of CDRGs **(B)** Venn diagram of potential hub genes influencing the potency of GEM **(C)** Heatmap of 120-CDRGs expression **(D)** Forest plot of univariate COX model for 120-CDRGs **(E)** Correlation analysis between CDRGs and IC50 values of Gemcitabine.

GSE39582). We found that CALB2 and GPX3 showed concordance in the four datasets, and patients in the high expression group had markedly poor prognosis (Figures 3B–E, $p<0.05$). In combination with the TCGA dataset, CALB2 and GPX3 might be hub genes for GEM treatment to CRC. To further investigate the potential association between CALB2 and GPX3 expression and Stage, TNM. Stage, we found that CALB2 and GPX3 expression increased with Stage, T. Stage, and N. Stage staging (Figure 4A). The expression of GPX3 increased with Stage, N. Stage (Figure 4B). The ridge analysis of CALB2 and GPX3 in 5 dataset was presented in Figure S1. those findings indicated that the 5 hub genes were closely associated with development of CRC.

## Biological pathways involved in CALB2 and GPX3

To further resolve the pathways potentially regulated by CALB2 and GPX3 in the TCGA dataset, we compared the pathways with significantly enriched pathways in tumor tissues and paraneoplastic tissues by GSVA method and calculated the GSVA scores of all pathways. We found that 172 KEGG pathways were significantly different in tumor tissues and paracancerous tissues, and heatmap was presented to show the GSVA enrichment fractions of 172 differential pathways in tumor tissues (Figure 5A). Accumulating studies indicated that tumor

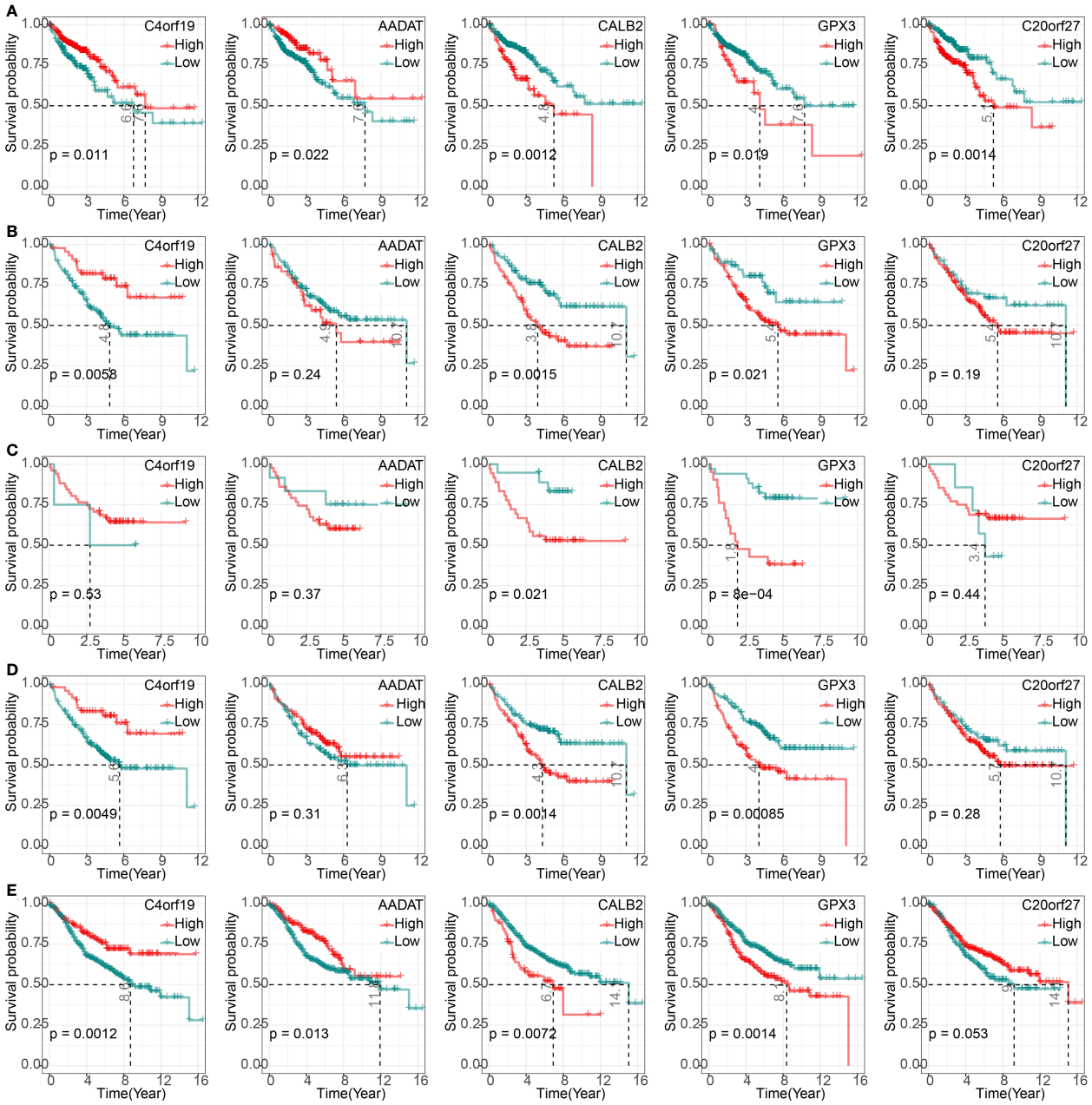



FIGURE 3
Correlation of 5-hub genes with CRC prognosis. **(A–E)** K-M survival curves of 5-hub genes in TCGA, GSE17536, GSE17537, GSE17538, and GSE39582 cohorts.

development was closely related to metabolic and signaling pathways in the organism (20, 21), we extracted 172 metabolic and signaling-related pathways in the organism among KEGG pathways and calculated their correlations with CALB2 and GPX3. We found that CALB2 was significantly associated with 27 METABOLISM Pathways and 14 SIGNALING Pathways, respectively, and GPX3 was significantly associated with 26 METABOLISM Pathways and 17 SIGNALING Pathways, respectively (Figures 5B, C). Those results revealed that the enhanced expression of CALB2 and GPX3 might inhibit the organism metabolism. Among the signaling pathways, it was found that the enhanced expression of CALB2 and GPX3 would activate immune and inflammation-related pathways.

## Correlation between CALB2 and GPX3 and immune microenvironment

To investigate the potential connection between CALB2 and GPX3 and immunity, we evaluated the correlation between CALB2 and GPX3 and immune cell infiltration scores in TME. First, we evaluated immune cell scores in TME of CRC patients in the TCGA

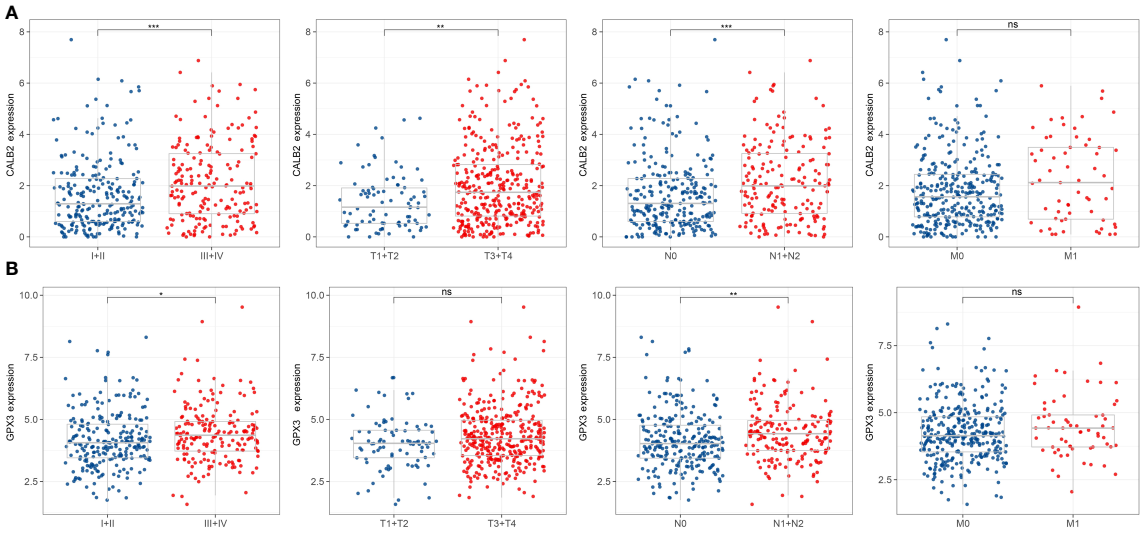



FIGURE 4
Correlation of 5-hub genes with CRC clinical information. ns, p>0.05, *p<0.05, **p<0.01, ***p<0.001. **(A, B)** Expression levels of CALB2 and GPX3 in clinical subgroups.

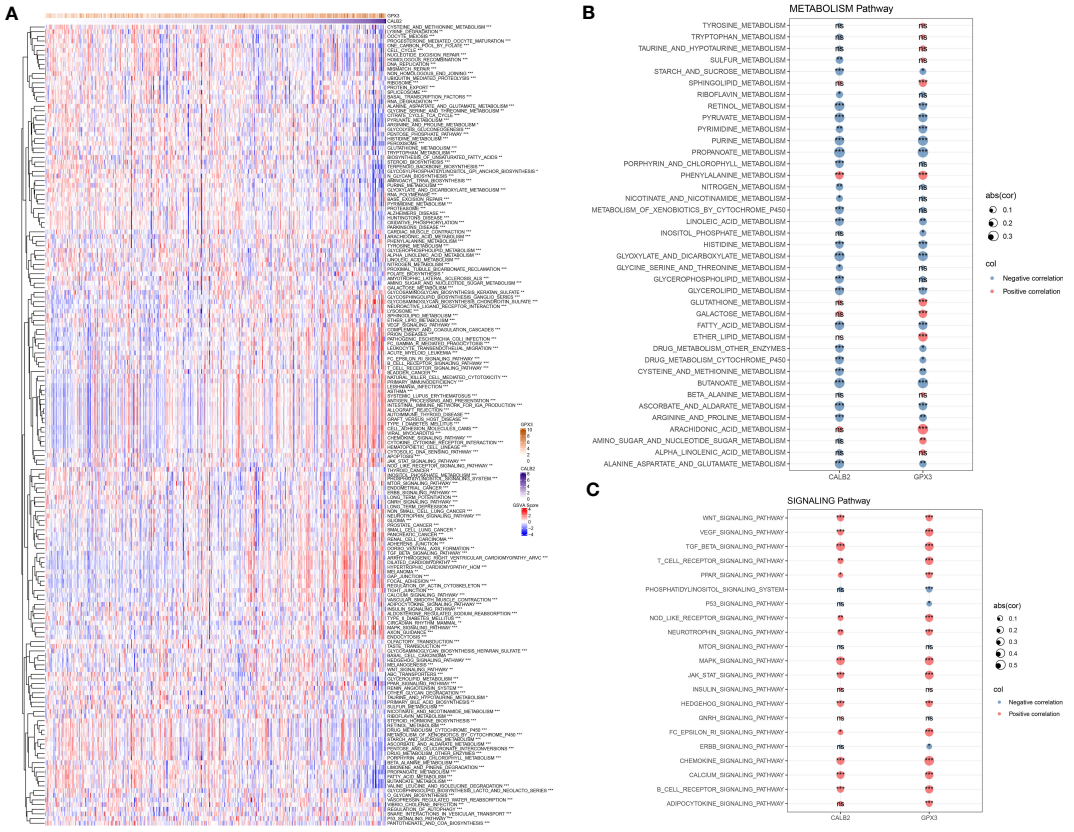



FIGURE 5
Biological pathways involved in CALB2 and GPX3. **(A)** Heatmap of GSVA results for CALB2 and GPX3 **(B)** Bubble plot of metabolism-related pathway enrichment score correlation analysis with CALB2 and GPX3 expression **(C)** Bubble plot of correlation analysis of signaling pathway enrichment score with CALB2 and GPX3 expression. ns p>0.05, *p<0.05, **p<0.01, ***p<0.001.

10.3389/fendo.2023.1170526

cohort by ESTIMATE and CIBERSORT algorithms and found correlations between CALB2 and GPX3 expression and immune scores by spearman correlation analysis. We found that the expression levels of CALB2 and GPX3 were significantly correlated with the StromalScore, ImmuneScore, and ESTIMATEScore of CRC ($p<0.05$) (Figures 6A, B). Except for the infiltration score of Dendritic cells resting, the infiltration scores of the remaining 21 immune cells showed concordance with the expression of CALB2 and GPX3 (Figure 6C). Finally, we calculated the ssGSEA enrichment scores of 28 signatures that could predict Checkpoint Blockade response. The results of the mantel test and pearson correlation showed that CALB2 and GPX3 expression were significantly correlated with most signatures that could predict the checkpoint Blockade response (Figure 6D). Immunocyte analysis implied that CALB2 and GPX3 had obviously correlated 7 immune cells (Figure S2). These results indicated that the immune function of CRC patients was enhanced with the increasing expression of CALB2 and GPX3 genes.

## The role of CALB2 and GPX3 in pan-cancer

To further analyze the prognostic value of CALB2 and GPX3 in pan-cancer, we compared the expression levels of CALB2 and GPX3 in 32 tumor tissues and paraneoplastic tissues from TCGA and GTEx data. The results showed that CALB2 and GPX3 were highly expressed in most tumor tissues (Figures 7A, B). Next, the K-M survival curves demonstrated the prognostic status in the high- and low-expression groups of CALB2 and GPX3 in 26 cancers. We found that high CALB2 expression was associated with poorer prognosis in GBM, OV, LUAD, BLCA, PAAD, KIRP, CESC, STAD,

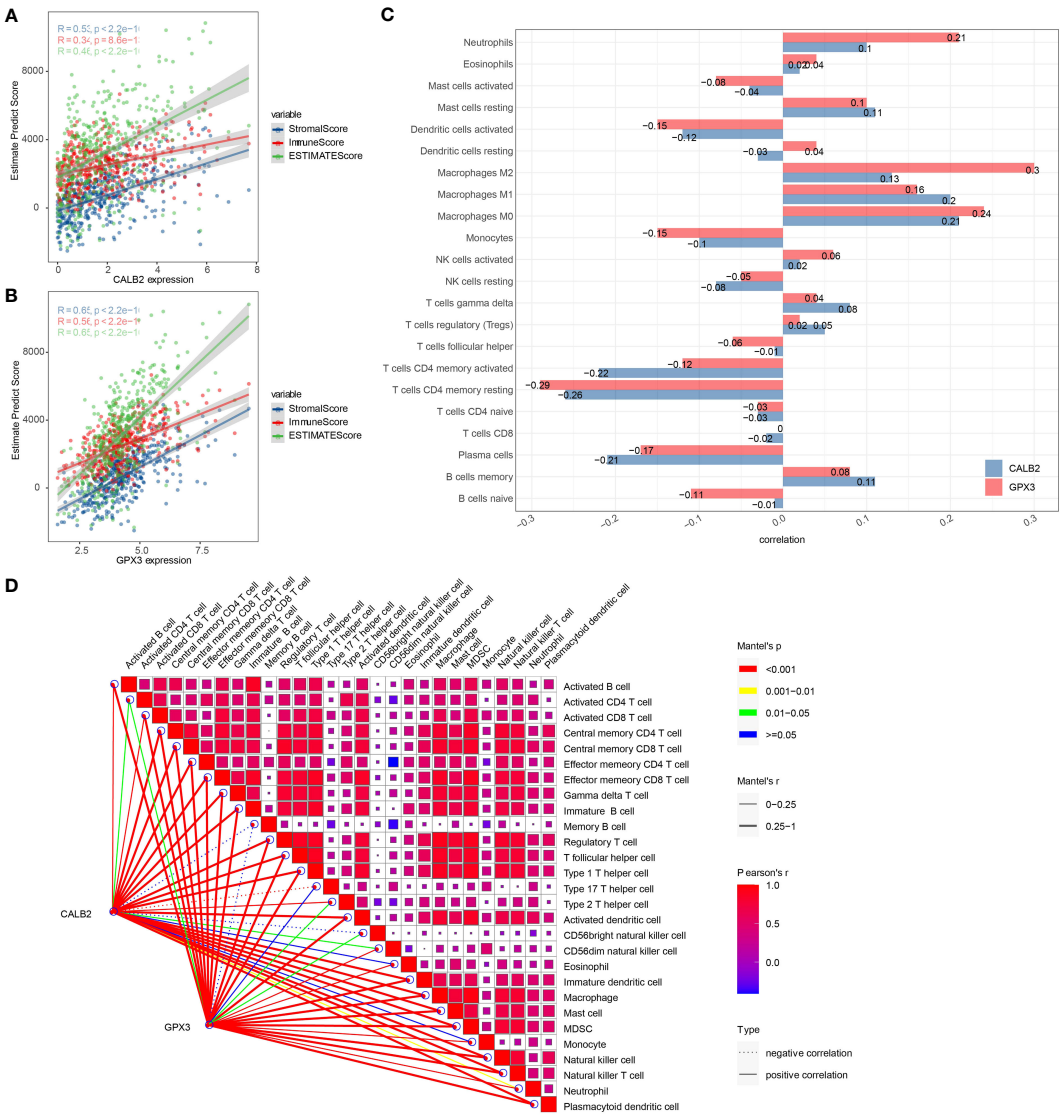



FIGURE 6
Correlation between CALB2 and GPX3 and immune microenvironment. **(A)** Scatter plot of CALB2 expression correlation with immune and stromal scores **(B)** Scatter plot of GPX3 expression correlation with immune and stromal scores **(C)** Histogram of the correlation between CALB2 and GPX3 expression and 22 TIICs. **(D)** Pearson analysis between 28 immune cell score and CALB2/GPX3.

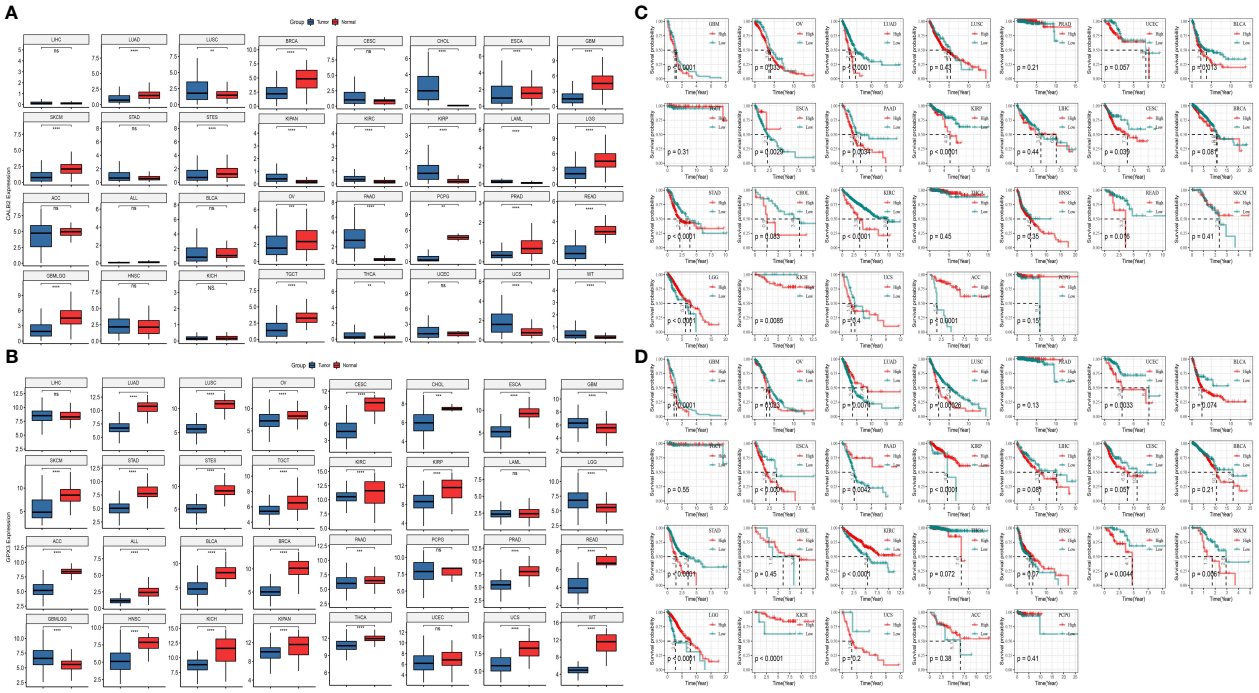


FIGURE 7
The role of CALB2 and GPX3 in pan-cancer. **(A, B)** Expression of CALB2 and GPX3 in pan-cancer **(C, D)** K-M survival curves of CALB2 and GPX3 in multiple cancer types.

CHOL, KIRC, READ, and KICH ($p<0.05$) (Figure 7C). High CALB2 expression was associated with more favorable prognosis in ESCA, LGG, and ACC patients ($p<0.05$) (Figure 7C). High GPX3 expression was associated with less favorable prognosis in GBM, OV, LUSC, UCEC, ESCA, LIHC, STAD, READ, and SKCM ($p<0.05$) (Figure 7D). GPX3 High expression was associated with favorable prognosis in LUAD, PAAD, KIRP, LGG, and KICH ($p<0.05$) (Figure 7D). These results demonstrated that the expression of CALB2 and GPX3 was intimately related to the prognosis of various cancer types.

## Chemotherapy drug sensitivity

We compared the IC50 data of 51 chemotherapeutic agents in the high and low expression groups of CALB2 and GPX3 (Figures 8A, B). The three groups with the largest and smallest IC50 values were selected for correlation analysis with CALB2 and GPX3 expression. We found that CALB2 and GPX3 expression were consistent with the drug response trend of MG-132, Dasatinib, Shikonin, Midostaurin, MS-275, and Z-LNle-CHO. The expressions of CALB2 and GPX3 were positively correlated with the IC50 of MG-132, Dasatinib, Shikonin, MS-275 and Z-LLNle-CHO. The expressions of CALB2 and GPX3 were significantly negatively correlated with the IC50 in Midostaurin (Figure 8C). These results suggested that GEM combined with MG-132, Dasatinib, Shikonin, Midostaurin, MS-275, and Z-LLNle-CHO might treat CRC through the action of CALB2 and GPX3.

# Discussion

CRC was characterized with high tumor heterogeneity, tumor mutation-resistant cells increasing the challenge of chemotherapy treatment, and new targets are crucial in CRC treatment (3). In this study, comprehensive bioinformatics analysis identified CALB2 and GPX3 which functioned as important targets in GEM treatment of CRC.

CALB2 encodes a Ca2+ binding protein that was intimately associated with cancer (22). a study by Bertschy et al. determined that CALB2 was expressed mainly in nervous system cells or ovarian cells (23). Further studies demonstrated that CALB2 was specifically expressed in CRC and mesotheliomas, which was considered as well as the diagnostic biomarker for CRC and mesotheliomas (23–27). In a recent study, Ojasalu et al. (28) demonstrated through *In-vitro* assays that CALB2 silencing inhibits ovarian high-grade plasmacytoma (HGSC) cell adhesion, which in turn caused peritoneal spread, and notably, that high CALB2 expression contributed to poor prognosis of HGSC.GEM was primarily subject to enzymatic deamination, low clearance, and drug resistance and It was currently intended primarily as alternative second-line therapeutic agent to 5-FU for the treatment of multiple cancers (4). At the cellular level, GEM is internalized *via* nucleic acid transporters. It is subsequently phosphorylated by dioxycytidine kinase (DCK). The stepwise phosphorylation leads to the formation of GEM-triphosphate, which is incorporated into cellular DNA, thereby inhibiting nuclear replication (7). Recent research concluded that GEM held

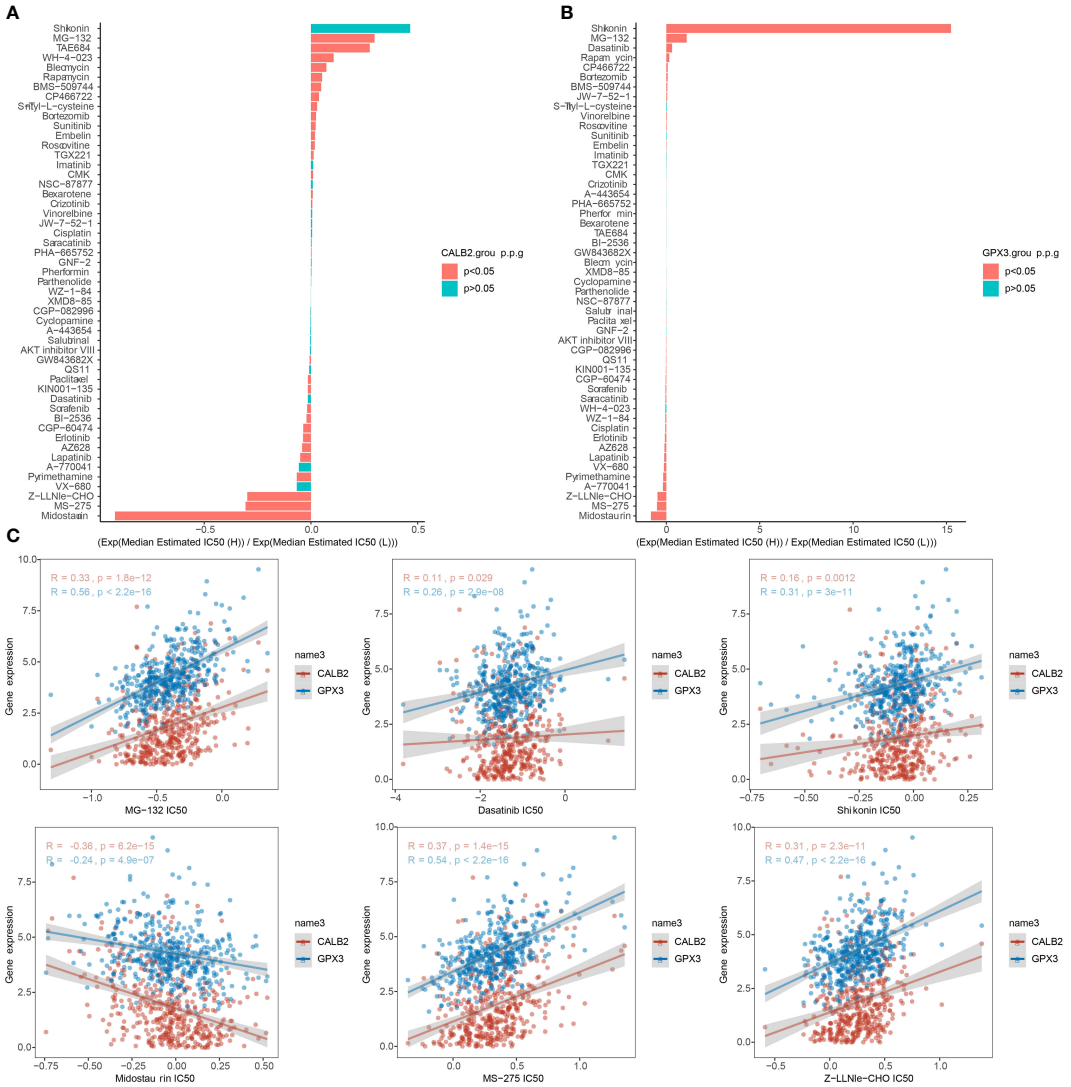



FIGURE 8
Chemotherapy drug sensitivity. **(A, B)** Histogram of CALB2 and GPX3 drug sensitivity analysis **(C)** Correlation of CALB2 and GPX3 with IC50 of chemotherapeutic agents.

promise as tumor-targeting agent by optimizing the mode of drug delivery action (7). Stevenson et al. (29) established that CALB2 in CRC responded to 5-FU regulation and that expression of down-regulated CALB2 induced death of CRC cells. Numerous investigations proved that CALB2 was the key gene in CRC development as well as treatment. Excitingly, the present report identified CALB2 as the possible hub gene affecting the potency of GEM through bioinformatics approaches. our pan-cancer analysis similarly established that CALB2 probably served as prognostic biomarker for multiple cancer species, thus illustrating that GEM-CALB2 might be promising for novel therapeutic modalities in cancer.

GPX3 transcription was regulated by selenium (5) and peroxisome proliferator-activated receptor γ (PPARγ), which protected cells against reactive oxygen species (ROS) accumulation (30–32). The finding in this study that GPX3 low expression caused poor prognosis in CRC was also demonstrated in earlier studies. The findings of Barrett et al. (33) found accelerated tumor accretion and significantly higher number of tumor cells in GPX3-deficient COAD mice, which also exhibited macrophage tendency to M2 polarization, enhanced expression of inflammatory factors, and over-activation of WNT signaling pathway.GPX3 in COAD mice exhibited immunomodulatory effects limiting the development of enteritis-associated cancers. Another investigation confirmed that downregulation of GPX3 expression led to increased H2O2 levels in TME and promoted tumor malignancy (34). Enrichment analysis in this study revealed that GPX3 was closely associated mainly with immune and inflammation-related pathways and that increased GPX3 expression could inhibit the metabolic response of the organism. Our study was consistent with the results of previous studies. In addition, Ji et al. (35) found that the administration of GEM induced ROS generation in HCC and activated Ets2 to

upregulate CD13 expression, and the activated expression of CD13 induced GEM resistance by activating NRF1 to upregulate GPX3 expression to clear intracellular ROS levels in HCC. This showed that GPX3 was closely associated with GEM potency, which was further confirmed by our study.

For CRC treatment, combination drug treatment modalities were feasible strategies (36). A recent report suggested that the combination of drugs could appropriately prolong the survival of CRC patients compared to chemotherapy alone (2). There was no exact effective targeted therapy for patients with high variability (2). The evaluation indexes of drug sensitivity generally include Area Under the Curve (AUC), Half maximal inhibitory concentration (IC50), Half maximal effective concentration (EC50) and Maximal effect level (Amax) (37–39). But IC50 is by far the most used. Therefore, tapping the exact therapeutic target is an urgent issue for CRC treatment. In this study, CALB2 and GPX3 expression were found to be consistent with the drug response trends of MG-132, Dasatinib, Shikonin, Midostaurin, MS-275, and Z-LNle-CHO, and CALB2 and GPX3 were potential pharmacodynamic targets of GEM. We hypothesized that the combination of GEM with MG-132, Dasatinib, Shikonin, Midostaurin, MS-275, and Z-LLNle-CHO may target CALB2 or GPX3 for CRC. These results demonstrated that CALB2 and GPX3 might be hub genes for GEM action.

Although this investigation integrated several databases to explore the hub genes affecting the potency of GEM, there were still shortcomings in this study. First, the integrated bioinformatics results provided that CALB2 and GPX3 were possible hub genes for GEM action, but there was no *In-vitro* cellular assay or *in-vivo* assay to validate this result, and subsequent wet experiments needed to be designed to further validate our results. Second, we determined that CALB2 and GPX3 enhanced immune function in CRC patients, but we did not conduct in-depth studies to explore the molecular mechanisms involved. Subsequent studies will focus on the specific regulatory mechanisms of GEM on CALB2 and GPX3 as well as a large sample multicenter prospective study to explore the effects of GEM combination with targeted therapies on CRC, leading to the development of novel therapeutic tools. Overall, this study revealed that CALB2 and GPX3 are potential target genes for GEM action.

## Conclusion

CALB2 and GPX3 served as biomarkers of CRC prognosis and as potential target genes for GEM. Our study provided new thought for the development of novel combination drug-targeted therapies for CRC.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

## Author contributions

All authors contributed to this present work. XXZ and LL designed the study, LS acquired the data. XL and YJ drafted the manuscript, JS and XHZ revised the manuscript. All authors read and approved the manuscript.

## Funding

This work was supported by Natural Science Foundation of China (No.81973718), Guangdong Natural Science Foundation (No.2021A1515011297), Guangzhou Science and Technology Plan Projects (No.202201020483 & No.202201010786) and the Foundation of Guangdong Second Provincial General Hospital (No.3DA2021015).

## Conflict of interest

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.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1170526/full#supplementary-material

## References

1. Xu H, Liu L, Li W, Zou D, Yu J, Wang L, et al. Transcription factors in colorectal cancer: molecular mechanism and therapeutic implications. *Oncogene* (2021) 40 (9):1555–69. doi: 10.1038/s41388-020-01587-3

2. Biller LH, Schrag D. Diagnosis and treatment of metastatic colorectal cancer: A review. *JAMA* (2021) 325(7):669–85. doi: 10.1001/jama.2021.0106

3. Russo M, Crisafulli G, Sogari A, Reilly NM, Arena S, Lamba S, et al. Adaptive mutability of colorectal cancers in response to targeted therapies. *Science* (2019) 366 (6472):1473–80. doi: 10.1126/science.aav4474

4. Miao H, Chen X, Luan Y. Small molecular gemcitabine prodrugs for cancer therapy. *Curr Med Chem* (2020) 27(33):5562–82. doi: 10.2174/0929867326666190816230650

5. Chocry M, Leloup L, Parat F, Messe M, Pagano A, Kovacic H. Gemcitabine: an alternative treatment for oxaliplatin-resistant colorectal cancer. *Cancers (Basel)* (2022) 14(23):5894. doi: 10.3390/cancers14235894

6. Guo B, Wei J, Wang J, Sun Y, Yuan J, Zhong Z, et al. CD44-targeting hydrophobic phosphorylated gemcitabine prodrug nanotherapeutics augment lung cancer therapy. *Acta Biomater* (2022) 145:200–9. doi: 10.1016/j.actbio.2022.04.016

7. Pandit B, Royzen M. Recent development of prodrugs of gemcitabine. *Genes (Basel)* (2022) 13(3):466. doi: 10.3390/genes13030466

8. Patterson J, Carpenter EJ, Zhu Z, An D, Liang X, Geng C, et al. Impact of sequencing depth and technology on *de novo* RNA-Seq assembly. *BMC Genomics* (2019) 20(1):604. doi: 10.1186/s12864-019-5965-x

9. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf* (2008) 9:559. doi: 10.1186/1471-2105-9-559

10. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* (2015) 43(7):e47. doi: 10.1093/nar/gkv007

11. Wang S, Su W, Zhong C, Yang T, Chen W, Chen G, et al. An eight-circRNA assessment model for predicting biochemical recurrence in prostate cancer. *Front Cell Dev Biol* (2020) 8:599494. doi: 10.3389/fcell.2020.599494

12. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinf* (2013) 14:7. doi: 10.1186/1471-2105-14-7

13. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* (2013) 4:2612. doi: 10.1038/ncomms3612

14. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol Biol* (2018) 1711:243–59. doi: 10.1007/978-1-4939-7493-1_12

15. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep* (2017) 18 (1):248–62. doi: 10.1016/j.celrep.2016.12.019

16. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* (2009) 462(7269):108–12. doi: 10.1038/nature08460

17. Shen W, Song Z, Zhong X, Huang M, Shen D, Gao P, et al. Sangerbox: A comprehensive, interaction-friendly clinical bioinformatics analysis platform. *iMeta* (2022) 1(3):e36. doi: 10.1002/imt2.36

18. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* (2018) 173(2):400–16.e11. doi: 10.1016/j.cell.2018.02.052

19. Geeleher P, Cox N, Huang RS. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PloS One* (2014) 9(9):e107468. doi: 10.1371/journal.pone.0107468

20. Bergers G, Fendt SM. The metabolism of cancer cells during metastasis. *Nat Rev Cancer* (2021) 21(3):162–80. doi: 10.1038/s41568-020-00320-2

21. Jiang C, Zhang N, Hu X, Wang H. Tumor-associated exosomes promote lung cancer metastasis through multiple mechanisms. *Mol Cancer* (2021) 20(1):117. doi: 10.1186/s12943-021-01411-w

22. Schwaller B, Durussel I, Jermann D, Herrmann B, Cox JA. Comparison of the Ca2+-binding properties of human recombinant calretinin-22k and calretinin. *J Biol Chem* (1997) 272(47):29663–71. doi: 10.1074/jbc.272.47.29663

23. Bertschy S, Genton CY, Gotzos V. Selective immunocytochemical localisation of calretinin in the human ovary. *Histochem Cell Biol* (1998) 109(1):59–66. doi: 10.1007/s004180050202

24. Gotzos V, Wintergerst ES, Musy JP, Spichtin HP, Genton CY. Selective distribution of calretinin in adenocarcinomas of the human colon and adjacent tissues. *Am J Surg Pathol* (1999) 23(6):701–11. doi: 10.1097/00000478-199906000-00010

25. Gotzos V, Schwaller B, Gander JC, Bustos-Castillo M, Celio MR. Heterogeneity of expression of the calcium-binding protein calretinin in human colonic cancer cell lines. *Anticancer Res* (1996) 16(6B):3491–8.

26. Doglioni C, Dei Tos AP, Laurino L, Iuzzolino P, Chiarelli C, Celio MR, et al. Calretinin: a novel immunocytochemical marker for mesothelioma. *Am J Surg Pathol* (1996) 20(9):1037–46. doi: 10.1097/00000478-199609000-00001

27. Chu AY, Litzky LA, Pasha TL, Acs G, Zhang PJ. Utility of D2-40, a novel mesothelial marker, in the diagnosis of malignant mesothelioma. *Mod Pathol* (2005) 18 (1):105–10. doi: 10.1038/modpathol.3800259

28. Ojasalu K, Brehm C, Hartung K, Nischak M, Finkernagel F, Rexin P, et al. Upregulation of mesothelial genes in ovarian carcinoma cells is associated with an unfavorable clinical outcome and the promotion of cancer cell adhesion. *Mol Oncol* (2020) 14(9):2142–62. doi: 10.1002/1878-0261.12749

29. Stevenson L, Allen WL, Proutski I, Stewart G, Johnston L, McCloskey K, et al. Calbindin 2 (CALB2) regulates 5-fluorouracil sensitivity in colorectal cancer by modulating the intrinsic apoptotic pathway. *PloS One* (2011) 6(5):e20276. doi: 10.1371/journal.pone.0020276

30. Ottaviano FG, Tang SS, Handy DE, Loscalzo J. Regulation of the extracellular antioxidant selenoprotein plasma glutathione peroxidase (GPx-3) in mammalian cells. *Mol Cell Biochem* (2009) 327(1-2):111–26. doi: 10.1007/s11010-009-0049-x

31. Reddy AT, Lakshmi SP, Banno A, Reddy RC. Role of GPx3 in PPARgamma-induced protection against COPD-associated oxidative stress. *Free Radic Biol Med* (2018) 126:350–7. doi: 10.1016/j.freeradbiomed.2018.08.014

32. Chung SS, Kim M, Youn BS, Lee NS, Park JW, Lee IK, et al. Glutathione peroxidase 3 mediates the antioxidant effect of peroxisome proliferator-activated receptor gamma in human skeletal muscle cells. *Mol Cell Biol* (2009) 29(1):20–30. doi: 10.1128/MCB.00544-08

33. Barrett CW, Ning W, Chen X, Smith JJ, Washington MK, Hill KE, et al. Tumor suppressor function of the plasma glutathione peroxidase gpx3 in colitis-associated carcinoma. *Cancer Res* (2013) 73(3):1245–55. doi: 10.1158/0008-5472.CAN-12-3150

34. Moloney JN, Cotter TG. ROS signalling in the biology of cancer. *Semin Cell Dev Biol* (2018) 80:50–64. doi: 10.1016/j.semcdb.2017.05.023

35. Ji S, Ma Y, Xing X, Ge B, Li Y, Xu X, et al. Suppression of CD13 enhances the cytotoxic effect of chemotherapeutic drugs in hepatocellular carcinoma cells. *Front Pharmacol* (2021) 12:660377. doi: 10.3389/fphar.2021.660377

36. Kim JH. Chemotherapy for colorectal cancer in the elderly. *World J Gastroenterol* (2015) 21(17):5158–66. doi: 10.3748/wjg.v21.i17.5158

37. Sebaugh JL. Guidelines for accurate EC50/IC50 estimation. *Pharm statistics* (2011) 10(2):128–34. doi: 10.1002/pst.426

38. Dalton BR, Rajakumar I, Langevin A, Ondro C, Sabuda D, Griener TP, et al. Vancomycin area under the curve to minimum inhibitory concentration ratio predicting clinical outcome: a systematic review and meta-analysis with pooled sensitivity and specificity. *Clin Microbiol infection Off Publ Eur Soc Clin Microbiol Infect Diseases* (2020) 26(4):436–46. doi: 10.1016/j.cmi.2019.10.029

39. Winding A, Modrzyński JJ, Christensen JH, Brandt KK, Mayer P. Soil bacteria and protists show different sensitivity to polycyclic aromatic hydrocarbons at controlled chemical activity. *FEMS Microbiol Lett* (2019) 366(17):fnz214. doi: 10.1093/femsle/fnz214

10.3389/fendo.2023.1170526

# Glossary

| | |
|---|---|
| CRC | colorectal cancer |
| GEM | Gemcitabine |
| TIIC | tumor-infiltrating immune cells |
| TME | tumor microenvironment |
| TCGA | The Cancer Genome Atlas |
| GEO | GENE EXPRESSION OMNIBUS |
| IC50 | half maximal inhibitory concentration |
| GDSC | Genomics of Drug Sensitibity in Cancer |
| WGCNA | Weighted correlation network analysis |
| PCA | principal component analysis |
| CDRGs | CRC development-related genes |
| K-M | Kaplan-Meier |
| GSVA | Gene Set Variation Analysis |
| ESTIMATE | Estimation of STromal and Immune cells in MAlignant Tumours using Expression data |
| HGSC | high-grade plasmacytoma |
| 5-FU | 5-fluorouracil |
| PPARγ | peroxisome proliferator-activated receptor γ |
| ROS | reactive oxygen species. |

frontiers | Frontiers in Endocrinology

TYPE Original Research
PUBLISHED 22 November 2023
DOI 10.3389/fendo.2023.1165178



OPEN ACCESS

EDITED BY
Prem P. Kushwaha,
Case Western Reserve University,
United States

REVIEWED BY
Jianbo Chang,
Peking Union Medical College and Chinese Academy of Medical Sciences, China
Hai-Yang Wang,
Jining First People's Hospital Affiliated to Shandong First Medical University, China
Liang Pan,
People's Hospital of Deyang City, China
Qiang He,
Sichuan University, China

*CORRESPONDENCE
Wenle Li
drlee0910@163.com
Liangqun Rong
rongliangqun@163.com
Haosheng Wang
Dr_haosheng@163.com

†These authors have contributed equally to this work

RECEIVED 13 February 2023
ACCEPTED 21 April 2023
PUBLISHED 22 November 2023

CITATION
Wang K, Jiang Q, Gao M, Wei X, Xu C, Yin C, Liu H, Gu R, Wang H, Li W and Rong L (2023) A clinical prediction model based on interpretable machine learning algorithms for prolonged hospital stay in acute ischemic stroke patients: a real-world study.
*Front. Endocrinol.* 14:1165178.
doi: 10.3389/fendo.2023.1165178



# A clinical prediction model based on interpretable machine learning algorithms for prolonged hospital stay in acute ischemic stroke patients: a real-world study

Kai Wang[1,2†], Qianmei Jiang[3†], Murong Gao[4†], Xiu'e Wei[1,2], Chan Xu[5], Chengliang Yin[6], Haiyan Liu[1,2], Renjun Gu[7], Haosheng Wang[7,8*], Wenle Li[2,9*] and Liangqun Rong[1,2*]

[1]Department of Neurology, The Second Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu, China, [2]Key Laboratory of Neurological Diseases, The Second Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu, China, [3]Department of General Practice, Xindu District People's Hospital of Chengdu, Chengdu, Sichuan, China, [4]Department of Rehabilitation, Beijing Rehabilitation Hospital Affiliated to Capital Medical University, Beijing, China, [5]Department of Dermatology, Xianyang Central Hospital, Xianyang, China, [6]Faculty of Medicine, Macau University of Science and Technology, Macau, Macao SAR, China, [7]School of Chinese Medicine and School of Integrated Chinese and Western Medicine, Nanjing University of Chinese Medicine, Nanjing, China, [8]State Key Laboratory of Pharmaceutical Biotechnology, Division of Sports Medicine and Adult Reconstructive Surgery, Department of Orthopedic Surgery, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, Jiangsu, China, [9]The State Key Laboratory of Molecular Vaccinology and Molecular Diagnostics and Center for Molecular Imaging and Translational Medicine, School of Public Health, Xiamen University, Xiamen, China

**Objective:** Acute ischemic stroke (AIS) brings an increasingly heavier economic burden nowadays. Prolonged length of stay (LOS) is a vital factor in healthcare expenditures. The aim of this study was to predict prolonged LOS in AIS patients based on an interpretable machine learning algorithm.

**Methods:** We enrolled AIS patients in our hospital from August 2017 to July 2019, and divided them into the "prolonged LOS" group and the "no prolonged LOS" group. Prolonged LOS was defined as hospitalization for more than 7 days. The least absolute shrinkage and selection operator (LASSO) regression was applied to reduce the dimensionality of the data. We compared the predictive capacity of extended LOS in eight different machine learning algorithms. SHapley Additive exPlanations (SHAP) values were used to interpret the outcome, and the most optimal model was assessed by discrimination, calibration, and clinical utility.

**Results:** Prolonged LOS developed in 149 (22.0%) of the 677 eligible patients. In eight machine learning algorithms, prolonged LOS was best predicted by the Gaussian naive Bayes (GNB) model, which had a striking area under the curve (AUC) of 0.878 ± 0.007 in the training set and 0.857 ± 0.039 in the validation set. The variables sorted by the gap values showed that the strongest predictors were pneumonia, dysphagia, thrombectomy, and stroke severity. High net benefits were observed at 0%–76% threshold probabilities, while good agreement was found between the observed and predicted probabilities.

**Conclusions:** The model using the GNB algorithm proved excellent for predicting prolonged LOS in AIS patients. This simple model of prolonged hospitalization could help adjust policies and better utilize resources.

KEYWORDS

prolonged hospital stay, stroke, machine learning, prediction model, SHAP (SHapley Additive exPlanations)

# Introduction

With acute ischemic stroke (AIS) being the first leading cause of disability and the second leading cause of mortality worldwide, economic burden remains a prominent issue in clinical practice (1). Length of stay (LOS) is a vital factor of overwhelmed healthcare cost expenditures. Pellico-Lopez et al. (2) found that 15.8% of the total cost of stroke cases depended on the cost of prolonged stay. Reducing unnecessary hospital stays is important to relieve insurance stress, especially under the policy of diagnosis-related groups (DRGs) payment. Therefore, it is essential that the risk model of prolonged LOS be analyzed to relieve economic burden and optimize the discharge plan for patients with AIS.

The average LOS following stroke onset varied according to time and country. In the United States, the LOS for stroke hospitalizations decreased from 2004 to 2018, according to the data survey of 8 million stroke patients (unadjusted: 6.3 days in 2004 vs. 5.6 days in 2018; adjusted: 7.6 days in 2004 vs. 5.4 days in 2018) (3). A *post-hoc* analysis (4) based on information from multiple sources in China found that the median and IQR of LOS for AIS was 10.0 (7.0–13.0) days. Hao et al. (5) found that malnutrition estimated by the CONUT score on admission could increase LOS in elderly AIS patients. Moreover, Neale et al. (6) found that stroke patients receiving an early supported discharge model of care spent fewer days in hospital and incurred less cost. In addition, the mode of treatment could also be related to the LOS after a stroke. Intravenous tissue plasminogen activator (IV-tPA) was associated with an increase in LOS in stroke patients treated with endovascular treatment within 4.5 h (7).

Only a few articles have currently established risk models for predicting the length of hospital stay in stroke patients. Koton et al. (8) evaluated the performance of the prolonged length of stay (PLOS) score in the cohort of stroke, and concluded that the PLOS score could be clinically useful in different healthcare systems. However, they only included patients from 2002 to 2007, and the treatments for stroke have developed dramatically in recent years. Nowadays, artificial intelligence is able to deduce from voluminous datasets and to incorporate nonlinear interactions among a large set of predictors (9–11). For machine learning predicting prolonged LOS in AIS, Kurtz et al. (12) accurately predicted the LOS of patients admitted to the ICU with stroke through machine learning methods, but they did not include stroke-specific data, such as the National Institutes of Health Stroke Scale (NIHSS) score or neuroimaging findings. Yang et al. (13) found that the artificial neural network model achieved adequate discriminative power for predicting prolonged LOS after AIS and identified crucial factors associated with a prolonged hospital stay. However, they did not include pneumonia or another important onset symptom of stroke, which proved to be strong influencing factors of LOS in AIS patients.

As a result, we set out to gather extensive stroke-specific data and create a scientific risk model based on an interpretable machine learning algorithm to predict prolonged hospital LOS in AIS patients. This simple model of prolonged hospitalization could help adjust policies and better utilize resources.

# Methods

## Participant selection

This study continuously enrolled AIS patients who were admitted to the Department of Neurology at the Second Affiliated Hospital of Xuzhou Medical University between August 2017 and July 2019 (Figure 1). The inclusion criteria were as follows: (1) age ≥ 18 years; (2) a diagnosis of AIS (14, 15)and within 24 h of onset (16, 17). The exclusion criteria were as follows: (1) patients who needed to be transferred from one department (or hospital) to another; (2) patients who had in-hospital strokes; (3) patients who had transient ischemic attack; and (4) patients who were unable to extract complete data. This flowchart indicated that our hospital managed about a total of 1,354 patients from August 2017 and July 2019, of whom 745 (55%) AIS participants had complete data (Figure 1). Of these 745 patients, 68 patients were those who needed to be transferred from one department (or hospital) to another/those who had in-hospital strokes, leaving a final cohort of 677 patients. Retrospective review of medical health records for this study was approved by our Institutional Review Board. Owing to the retrospective nature of this study, written informed consent was waived (Number: 2020081603). Moreover, the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statements were followed for all data analysis and reporting (18).

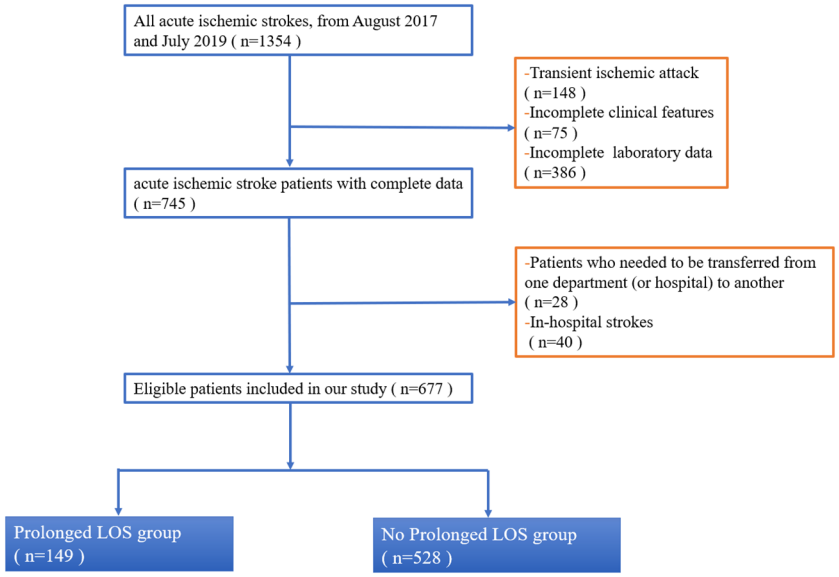



FIGURE 1
Flowchart of inclusion and exclusion of study patients. This flowchart indicated that our hospital managed about total 1,354 patients from August 2017 and July 2019, of which 745 (55%) AIS participants had complete data (Figure 1). Of these 745 patients, 68 patients were those who needed to be transferred from one department (or hospital) to another/those who had in-hospital strokes, leaving a final cohort of 677 patients. Abbreviation: LOS, length of stay.

## Data collection and definitions

The primary outcome was the prediction of a prolonged LOS for AIS patients, which was defined as more than 7 days of hospitalization. The LOS was measured from the admission day to the death or discharge day. This definition was similar to previous studies on LOS in stroke patients (8, 19, 20). The main clinical data included the following categories: baseline demographics, clinical features, and laboratory data. For baseline demographics, systolic blood pressure (SBP) and diastolic blood pressure (DBP) were tested on the right hand and extracted from the nursing record sheet on admission. For clinical features, stroke severity was divided into "mild" (NIHSS score < 8) and "moderate to severe" (NIHSS score ≥9), which was similar to previous clinical trials (21–23). Sato et al. (22) found that the optimal cutoff score of the baseline NIHSS for the favorable outcome was 8 for patients with anterior circulation stroke (sensitivity, 80%; specificity, 82%). The pneumonia in our study referred to those with development of pneumonia within 72 h after hospitalization (24). We diagnosed pneumonia by the CDC criteria because it was the most commonly used (25). The dysphagia was defined as abnormal swallowing physiology of the upper aerodigestive tract and as detected from clinician testing including screening, clinical bedside, or instrumental tests (26). The thrombolysis, thrombectomy, antiplatelets, anticoagulation, statins, and proton pump inhibitors were also collected from medical records. Treatment methods for AIS were followed by the 2019 American Heart Association/American Stroke Association (AHA/ASA) guideline (27). For laboratory data, they were extracted from blood test results on admission.

## Machine learning algorithm and data analysis

Continuous data were presented as median and interquartile range (IQR), and the Mann–Whitney *U*-test was used for statistical comparison between two groups. Categorical data were described as proportions, and the chi-squared or Fisher's exact test was used for comparison between two groups. The least absolute shrinkage and selection operator (LASSO) regression was applied to reduce the dimensionality of the data. In total, we utilized eight different machine learning algorithms, including the extreme gradient boosting (XGB) classifier, logistic regression, the light gradient boosting machine (LGBM) classifier, the AdaBoost classifier, Gaussian naive Bayes (GNB), complement naive Bayes (Complement NB), the multilayered perceptron (MLP) classifier, and the support vector (SVC) classifier. The hyperparameter settings for eight different machine learning algorithms used in our study are listed in Supplementary Table 1. For the XGB classifier, learning rate was set as 0.001, and the reg lambda was 0.01. Max depth and min child weight were set as 2. The area under the receiver operating characteristic (ROC) curve of the model was calculated by 10 bootstrapping resamples. For each bootstrap resample, the validation set (135 cases) accounted for 20% of the total sample, and the training set (542 cases) accounted for 80% of the total sample. After selecting the best model classifiers for this dataset, we exploited SHapley Additive exPlanations (SHAP) values to interpret the outcomes of the classifiers, which was a unified approach that connected cooperative game theory with local explanations to explain the output of any machine learning model. In addition, the decision curve analysis (DCA) was

applied to present the net benefits at various threshold probabilities. A calibration plot was used to investigate the degree of agreement between two groups.

# Results

## Patient characteristics

A total of 677 patients remained for evaluation of the machine learning algorithms to predict prolonged LOS in AIS patients, among whom prolonged LOS was detected in 22.0% ($n$ = 149). The average of LOS in all 677 participants was 10.78 ± 4.69 days. The baseline and clinical characteristics between the two groups are compared in Table 1. Longer LOS was linked to elevated levels of brain natriuretic peptide (BNP), S100-β, and neuron-specific enolase (NSE). Moreover, the prolonged LOS group was more likely to suffer from dysphagia, pneumonia, and a moderate-to-severe stroke. As for treatment, the prolonged LOS group had more frequent use of thrombolysis, thrombectomy, anticoagulation, and proton pump inhibitors (PPIs). Then, least absolute shrinkage and selection operator (LASSO) regression was used to reduce the number of factors with an optimal λ of 0.002. The candidate characteristics were narrowed down to the following 28 features with nonzero coefficients: age, gender, diastolic blood pressure, anterior or posterior stroke, side of hemisphere, stroke lesion, single or multiple lesions, cholesterol, triglyceride, low-density lipoprotein (LDL), glycosylated hemoglobin (HbA1c), homocysteine (HCY), uric acid (UA), myoglobin (MB), and fibrinogen. The coefficients of characteristics selected by LASSO regression are illustrated in Figure 2.

## Development and validation of models

As shown in Table 2, the GNB model with all characteristics had a striking AUROC of 0.878 ± 0.007 in the training set and 0.857 ± 0.039 in the validation set, while the other seven representative models had the highest AUROC of 0.875 ± 0.014 in the training set and 0.837 ± 0.031 in the validation set. For the GNB model, the sensitivities were 0.818 (training sets) and 0.804 (validation sets), while the specificities were 0.814 (training sets) and 0.816 (validation sets). The cross-reference between the full names and abbreviations in our manuscript is shown in Supplementary Table 2. The forest plot of each AUROC of eight models is depicted in Figure 3. Figures 4A, B present the comparison of AUROC between the GNB model and the other seven models, respectively, in the training and validation sets. The learning curve of the GNB model is displayed in Figure 5. Obviously, the GNB model significantly outperformed the other seven models in both

TABLE 1 The baseline and clinical characteristics in prolonged LOS patients and no prolonged LOS patients.

| Variables | Category | All patients ($n$ = 677) | No prolonged LOS ($n$ = 528) | Prolonged LOS ($n$ = 149) | Statistical value | $p$ |
|---|---|---|---|---|---|---|
| Demographics | | | | | | |
| Age | NA | 57 [45,68] | 57 [45,68] | 58 [47,68] | −1.210 | 0.226 |
| Gender | Female | 279 (41.21) | 209 (39.58) | 70 (46.98) | 2.624 | 0.105 |
| | Male | 398 (58.79) | 319 (60.42) | 79 (53.02) | | |
| SBP | NA | 143 [132,156] | 142 [132,155] | 147 [135,159] | −2.302 | 0.021 |
| DBP | NA | 87 [74,97] | 86 [73,96] | 88 [76,98] | −1.251 | 0.211 |
| Clinical features | | | | | | |
| Stroke severity | Mild | 385 (56.87) | 339 (64.21) | 46 (30.87) | 52.637 | <0.001 |
| | Moderate to severe | 292 (43.13) | 189 (35.80) | 103 (69.13) | | |
| Dysphagia | No | 525 (77.55) | 465 (88.07) | 60 (40.27) | 152.495 | <0.001 |
| | Yes | 152 (22.45) | 63 (11.93) | 89 (59.73) | | |
| Stroke distribution | Anterior | 270 (39.88) | 208 (39.39) | 62 (41.61) | 0.737 | 0.692 |
| | Posterior | 252 (37.22) | 201 (38.07) | 51 (34.23) | | |
| | Both | 155 (22.90) | 119 (22.54) | 36 (24.16) | | |
| Side of hemisphere | Left | 283 (41.80) | 223 (42.24) | 60 (40.27) | 0.462 | 0.794 |
| | Right | 270 (39.88) | 207 (39.21) | 63 (42.28) | | |
| | Both | 124 (18.32) | 98 (18.56) | 26 (17.45) | | |

*(Continued)*

10.3389/fendo.2023.1165178

TABLE 1 Continued

| Variables | Category | All patients (*n* = 677) | No prolonged LOS (*n* = 528) | Prolonged LOS (*n* = 149) | Statistical value | *p* |
|---|---|---|---|---|---|---|
| Site of stroke lesion | Cortex | 155 (22.90) | 109 (20.64) | 46 (30.87) | 9.095 | 0.059 |
| | Cortex-subcortex | 155 (22.90) | 125 (23.67) | 30 (20.13) | | |
| | Subcortex | 186 (27.47) | 151 (28.60) | 35 (23.49) | | |
| | Brainstem | 104 (15.36) | 86 (16.29) | 18 (12.08) | | |
| | Cerebellum | 77 (11.37) | 57 (10.80) | 20 (13.42) | | |
| Number of stroke lesions | Single | 470 (69.42) | 372 (70.46) | 98 (65.77) | 1.200 | 0.273 |
| | Multiple | 207 (30.58) | 156 (29.55) | 51 (34.23) | | |
| Thrombolysis | No | 473 (69.87) | 385 (72.92) | 88 (59.06) | 10.598 | 0.001 |
| | Yes | 204 (30.13) | 143 (27.08) | 61 (40.94) | | |
| Thrombectomy | No | 644 (95.13) | 525 (99.43) | 119 (79.87) | 95.943 | <0.001 |
| | Yes | 33 (4.87) | 3 (0.57) | 30 (20.13) | | |
| Antiplatelet | No | 122 (18.02) | 101 (19.13) | 21 (14.09) | 1.994 | 0.158 |
| | Yes | 555 (81.98) | 427 (80.87) | 128 (85.91) | | |
| Anticoagulation | No | 576 (85.08) | 467 (88.45) | 109 (73.15) | 21.411 | <0.001 |
| | Yes | 101 (14.92) | 61 (11.55) | 40 (26.85) | | |
| Statin | No | 103 (15.21) | 84 (15.91) | 19 (12.75) | 0.898 | 0.343 |
| | Yes | 574 (84.79) | 444 (84.09) | 130 (87.25) | | |
| PPI | No | 535 (79.03) | 462 (87.50) | 73 (48.99) | 103.954 | <0.001 |
| | Yes | 142 (20.98) | 66 (12.50) | 76 (51.01) | | |
| Pneumonia | No | 512 (75.63) | 473 (89.58) | 39 (26.17) | 253.486 | <0.001 |
| | Yes | 165 (24.37) | 55 (10.42) | 110 (73.83) | | |
| Laboratory data | | | | | | |
| S-100β | NA | 275 [224,290] | 273 [221,288] | 281 [237,297] | −3.057 | 0.002 |
| NSE | NA | 16.24 [12.69,18.60] | 15.73 [12.61,18.43] | 17.61 [14.05,18.92] | −3.196 | 0.001 |
| BNP | NA | 93 [73,162] | 89 [73,158] | 103 [77,168] | −2.213 | 0.027 |
| D-dimer | NA | 174 [133,221] | 174 [134,219] | 175 [132,224] | −0.239 | 0.812 |
| FIB | NA | 4.35 [3.96,4.75] | 4.35 [3.95,4.71] | 4.440 [4.04,4.79] | −1.488 | 0.137 |
| CRP | NA | 12.56 [7.72,17.63] | 12.21 [7.63,17.19] | 13.90 [8.11,19.06] | −1.956 | 0.050 |
| MB | NA | 97.66 [75.12,147.84] | 98.77 [76.43,147.84] | 94.85 [72.32,144.45] | 0.864 | 0.388 |
| UA | NA | 349.80 [309.80,408.10] | 353.20 [310.50,408.50] | 343 [307.80,406.30] | 0.595 | 0.552 |
| HCY | NA | 15.77 [12.74,19.37] | 16.12 [12.54,19.31] | 15.51 [13.04,20.01] | −0.525 | 0.600 |
| HbA1c | NA | 5.60 [5.30,5.90] | 5.60 [5.40,5.90] | 5.60 [5.30,6.00] | 0.462 | 0.643 |
| FBG | NA | 5.28 [4.63,5.83] | 5.28 [4.67,5.83] | 5.22 [4.55,5.79] | 1.099 | 0.272 |
| LDL | NA | 4.75 [4.33,4.94] | 4.75 [4.31,4.95] | 4.75 [4.37,4.90] | 0.131 | 0.896 |
| Triglyceride | NA | 2.17 [1.93,2.37] | 2.19 [1.93,2.37] | 2.16 [1.93,2.37] | 0.568 | 0.570 |
| Cholesterol | NA | 5.33 [4.43,6.15] | 5.33 [4.40,6.11] | 5.40 [4.54,6.25] | −1.075 | 0.283 |

LOS, length of stay; DBP, diastolic blood pressure; PPI, proton pump inhibitor; NSE, neuron-specific enolase; BNP, b-type natriuretic peptide; FIB, fibrinogen; CRP, C-reaction protein; MB, myoglobin; UA, uric acid; HCY, homocysteine; HbA1c, glycosylated hemoglobin; FBG, fasting blood glucose; LDL, low density lipoprotein; NA, not available.

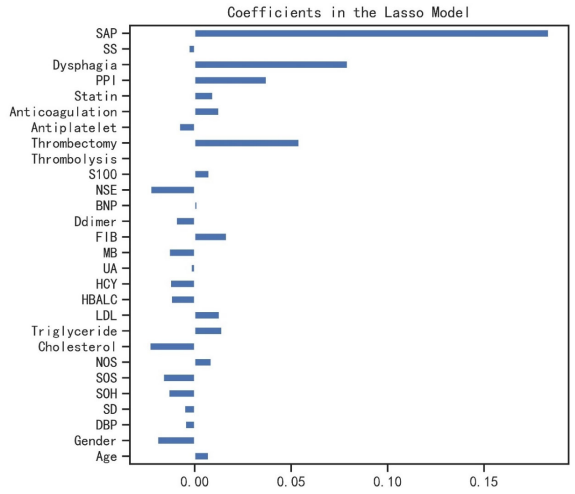



FIGURE 2
The coefficients of characteristics selected by LASSO regression. LASSO, least absolute shrinkage and selection operator; SAP, stroke-associated pneumonia; SS, stroke severity; PPI, proton pump inhibitor; NSE, neuron-specific enolase; BNP, b-type natriuretic peptide; FIB, fibrinogen; MB, myoglobin; UA, uric acid; HCY, homocysteine; HbA1c, glycosylated hemoglobin; LDL, low-density lipoprotein; NOS, number of stroke lesions; SOS, site of stroke lesion; SOH, side of hemisphere; SD, stroke distribution; DBP, diastolic blood pressure.

the training and validation sets. Despite the narrow gap, De Long's test showed that the difference between the GNB and XGB model remained significant ($p$ = 0.04).

## SHAP values depending on variables

The SHAP values for the GNB model and the importance of the variables sorted by the gap values are shown in Figures 6A, B. Red bars indicated an increase in the probability of prolonged LOS, whereas blue bars demonstrated a decrease in the probability of prolonged LOS for AIS patients. As Figure 6B shows, pneumonia, dysphagia, thrombectomy, and stroke severity all substantially increased the probability of prolonged LOS. In addition, we performed a decision curve analysis (Figure 7A) and a calibration plot (Figure 7B) to illustrate the performance of the GNB model. High net benefits could be observed in 0%–76% threshold probabilities, while good agreement could be found between the observed and predicted probabilities of prolonged LOS.

# Discussion

This study generated a simple clinical risk model that can be used to determine patients at increased risk of prolonged LOS. Our risk model had a promising AUC of 0.878 and 0.857 in the training

TABLE 2 The predictive capacity of eight different machine learning algorithms.

| | Model | AUC (SD) | Accuracy (SD) | Sensitivity (SD) | Specificity (SD) | PPV (SD) | NPV (SD) | Kappa (SD) |
|---|---|---|---|---|---|---|---|---|
| | XGB | 0.863 (0.011) | 0.862 (0.009) | 0.799 (0.027) | 0.845 (0.033) | 0.671 (0.045) | 0.923 (0.008) | 0.609 (0.021) |
| | logistic | 0.875 (0.014) | 0.837 (0.019) | 0.752 (0.042) | 0.863 (0.032) | 0.608 (0.046) | 0.924 (0.009) | 0.561 (0.034) |
| | LGBM | 0.817 (0.009) | 0.782 (0.008) | 0.739 (0.016) | 0.895 (0.007) | NA | 0.782 (0.008) | 0.000 (0.000) |
| Train | AdaBoost | 0.817 (0.009) | 0.782 (0.008) | 0.739 (0.016) | 0.895 (0.007) | NA | 0.782 (0.008) | 0.000 (0.000) |
| set | GNB | 0.878 (0.007) | 0.813 (0.020) | 0.818 (0.030) | 0.814 (0.030) | 0.551 (0.039) | 0.939 (0.007) | 0.533 (0.037) |
| | CNB | 0.706 (0.028) | 0.613 (0.075) | 0.747 (0.145) | 0.577 (0.133) | 0.337 (0.036) | 0.896 (0.031) | 0.222 (0.046) |
| | MLP | 0.519 (0.045) | 0.626 (0.147) | 0.401 (0.287) | 0.690 (0.266) | 0.314 (0.098) | 0.809 (0.018) | 0.072 (0.045) |
| | SVM | 0.503 (0.033) | 0.658 (0.112) | 0.304 (0.227) | 0.761 (0.207) | 0.274 (0.051) | 0.798 (0.019) | 0.052 (0.030) |
| | XGB | 0.837 (0.031) | 0.862 (0.023) | 0.759 (0.052) | 0.877 (0.052) | 0.682 (0.091) | 0.917 (0.019) | 0.606 (0.090) |
| | logistic | 0.833 (0.035) | 0.813 (0.034) | 0.750 (0.080) | 0.840 (0.102) | 0.575 (0.105) | 0.906 (0.022) | 0.501 (0.095) |
| | LGBM | 0.815 (0.040) | 0.774 (0.031) | 0.730 (0.073) | 0.900 (0.028) | NA | 0.774 (0.031) | 0.000 (0.000) |
| Validation | AdaBoost | 0.815 (0.040) | 0.774 (0.031) | 0.730 (0.073) | 0.900 (0.028) | NA | 0.774 (0.031) | 0.000 (0.000) |
| set | GNB | 0.857 (0.039) | 0.791 (0.036) | 0.804 (0.035) | 0.816 (0.075) | 0.527 (0.098) | 0.926 (0.022) | 0.487 (0.098) |
| | CNB | 0.680 (0.053) | 0.582 (0.073) | 0.740 (0.173) | 0.609 (0.181) | 0.316 (0.047) | 0.862 (0.052) | 0.166 (0.037) |
| | MLP | 0.515 (0.028) | 0.599 (0.157) | 0.463 (0.308) | 0.680 (0.300) | 0.274 (0.145) | 0.787 (0.039) | 0.039 (0.050) |
| | SVM | 0.498 (0.059) | 0.636 (0.142) | 0.559 (0.338) | 0.560 (0.330) | 0.243 (0.122) | 0.772 (0.040) | 0.014 (0.064) |

AUC, area under the curve; SD, standard deviation; PPV, positive predictive value; NPV, negative predictive value; XGB, extreme gradient boosting; LGBM, light gradient boosting machine; GNB, Gaussian naive bayes; CNB, complement naive Bayes; MLP, multilayered perceptron; SVM, support vector machine; NA, not available.

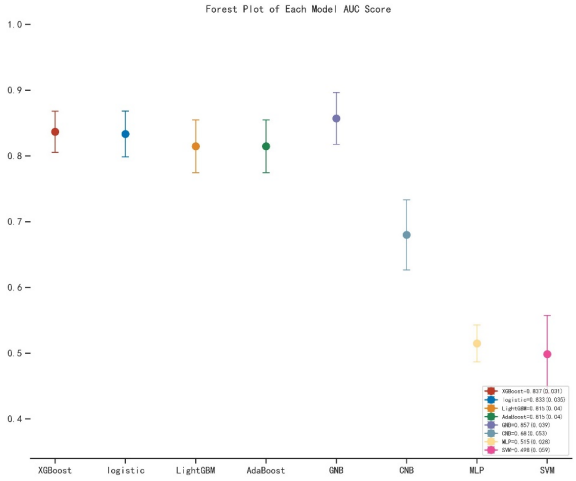



FIGURE 3
The forest plot of the each AUROC of eight models. AUROC, area under the receiver operating characteristic curve; XGB, extreme gradient boosting; LGBM, light gradient boosting machine; GNB, Gaussian naive Bayes; CNB, complement naive Bayes; MLP, multilayered perceptron; SVM, support vector machine.

and validation sets, respectively. The main outcomes of the current study were that pneumonia, dysphagia, thrombectomy, and stroke severity were the strongest clinical parameters for prolonged LOS following AIS after recursive feature elimination. Moreover, the artificial intelligence algorithms developed by these parameters showed excellent model performance on discrimination, calibration, and decision curve analysis. The strengths of our clinical risk score included the use of simple demographic and common biochemical parameters, and we collected enough candidate variables to develop this model. To our knowledge, this is the first study to predict prolonged LOS for common AIS patients based on an interpretable machine learning algorithm. The difference from previous studies was that we developed an integrated machine learning model with high performance, which could help adjust the policies to better utilize resources, especially under the DRG payment policy and the increasingly serious aging problem in the global world.

Su et al. (28) included 129,444 patients with AIS and found that the inpatient cost was $1,020 ($742–$1,545) in China. In an attempt to decrease patients' risk of prolonged LOS following AIS, previous retrospective studies have identified some factors. Many studies define prolonged LOS as more than 7 days (8, 19, 20). However, when it comes to patients with severe strokes or those admitted to an intensive care unit, some studies define it as more than 30 days (12, 29). Common factors affecting stroke hospitalization duration included quality of care, hospital-acquired infection, stroke severity and type, level of consciousness, history of heart failure and atrial fibrillation, and receiving reperfusion therapy (19, 29–33). Interestingly, during adolescence, low stress resilience, underweight, and higher systolic blood pressure were associated

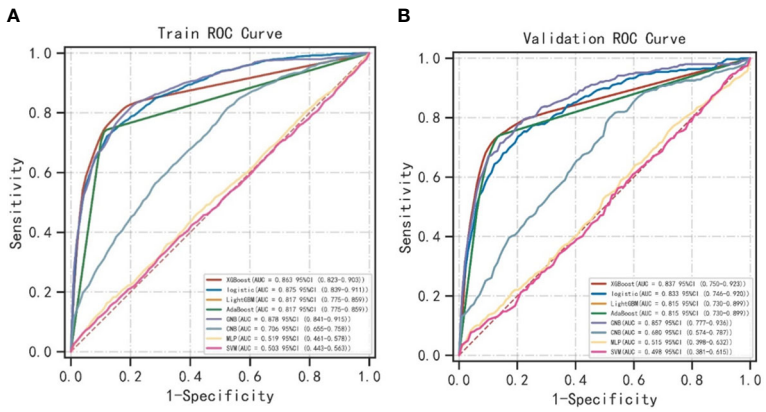



FIGURE 4
The comparison of AUROC between the GNB model and the other seven models. **(A)** The comparison of AUROC between the GNB model and the other seven models in the training sets. **(B)** The comparison of AUROC between the GNB model and the other seven models in the validation sets. ROC, area under the receiver operating characteristic curve; XGB, extreme gradient boosting; LGBM, light gradient boosting machine; GNB, Gaussian naive Bayes; CNB, complement naive Bayes; MLP, multilayered perceptron; SVM, support vector machine.

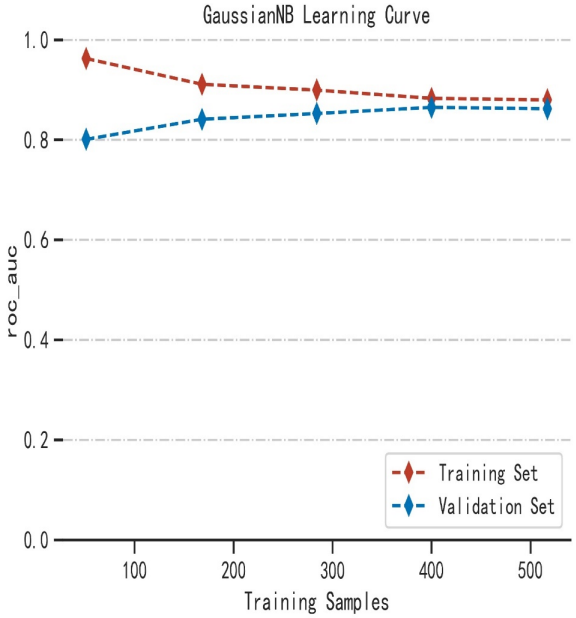



FIGURE 5
The learning curve of the GNB model.

with longer hospital stays in AIS, with adjusted relative hazard ratios of 1.46, 1.41, and 1.01, respectively (34), whereas these prior studies did not show the weight of each parameter on the probability of prolonged LOS. An interpretable machine learning algorithm has the ability to analyze big datasets with high accuracy through automated analysis of non-linear relationships between numerous variables (35). Machine learning algorithms apply various statistical methods from past experience to select useful patterns in large and complex datasets, which involves extreme gradient boosting (XGB) classifier, GNB, SVC classifier, and so on (36). Raizada et al. (37) concluded the advantages and limitations of different algorithms and found that GNB produced results that were statistically robust and were replicates across two independent datasets. An additional advantage of GNB classifiers was that GNB produced an accuracy similar to more sophisticated classifiers but with a substantial gain in speed (38). Therefore, we selected the GNB model from eight different machine learning algorithms that showed excellent performance in predicting prolonged LOS in AIS patients.

In this study, pneumonia, dysphagia, thrombectomy, and stroke severity were the leading clinical parameters in our interpretable machine learning algorithm. Pneumonia is an early complication of stroke and usually leads to prolonged LOS. The prevalence of pneumonia in patients with dysphagia after stroke was reported to

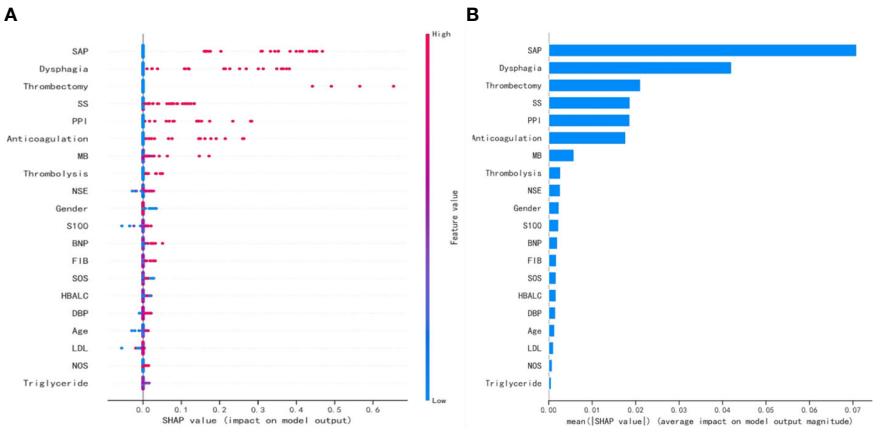



FIGURE 6
The SHAP values for the GNB model and the importance ranking of the variables. **(A)** The SHAP values for the GNB model. **(B)** The importance of the variables sorted by the gap values. SHAP, SHapley Additive exPlanations; GNB, Gaussian naive Bayes; SAP, stroke-associated pneumonia; SS, stroke severity; PPI, proton pump inhibitor; MB, myoglobin; NSE, neuron-specific enolase; BNP, b-type natriuretic peptide; FIB, fibrinogen; SOS, site of stroke lesion; HbA1c, glycosylated hemoglobin; DBP, diastolic blood pressure; LDL, low-density lipoprotein; NOS, number of stroke lesions.

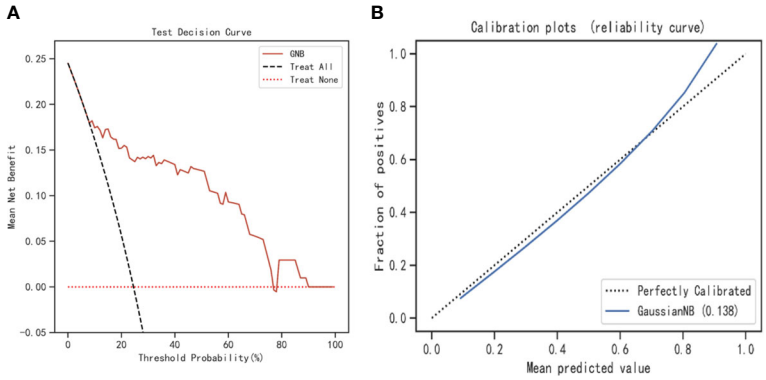



FIGURE 7
The decision curve analysis and calibration plot to illustrate the performance of the GNB model. **(A)** The decision curve analysis for the GNB model. **(B)** Calibration plot for the GNB model. GNB, Gaussian naive Bayes.

range from 7% to 33%, and the prevalence of dysphagia has been reported as between 28% and 65% (39, 40). Aspiration without a cough, known as "silent aspiration," further increased the incidence of pneumonia to 54% (40). A systematic review of stroke-associated pneumonia reported that the overall incidence of pneumonia ranged from 0% to 23.6% (41), which was a little lower than the incidence in our study. In our study, the incidence of pneumonia in all participants is 24.37%. It may be because of the varied definitions and diagnosis criteria of stroke-associated pneumonia. The Centers for Disease Control and Prevention (CDC) criteria (25), the PISCES SAP diagnostic criteria (42), and the combination of the clinical symptoms and auxiliary examination results criteria were all used to diagnose stroke-associated pneumonia in previous studies (41). In our study, we diagnosed pneumonia by the CDC criteria because it was the most commonly used, using clinical (lung auscultation and percussion, presence of fever, and purulent tracheal secretion), microbiological (tracheal specimens and blood cultures), and chest radiography findings. For dysphagia, the incidence in all participants was 22.45%, while in the "prolonged LOS group", it was 59.73%, and in the "no prolonged LOS group", it was 11.93% (Table 1). The incidence of dysphagia varied greatly between studies (ranged from 20% to 80%), depending on the definition of dysphagia, which can range from failing a dysphagia screen, to prescribed diet modifications, to measures of physiology on an instrumented swallowing study (26, 41, 43). Ogawa et al. (40) found that patients who underwent a flexible endoscopic evaluation of swallowing and received optimal nutritional intervention were more likely to have a shorter hospital stay ($p$ = 0.005). The complications of dysphagia include the consequences of modifications to dietary intake: compromised nutrition and hydration, prolonged LOS, and reduced quality of life. As a result, the optimal treatments and measures for dysphagia should be performed. Many studies have investigated a variety of interventions, including therapist-delivered, behavioral, acupuncture, and electrical or magnetic stimulation to treat dysphagia (39). As for stroke severity, it was the most consistent factor among the factors contributing to LOS in AIS patients, and those who received reperfusion therapy were more likely to have prolonged LOS, which was similar to the previous study (29). Patients with more severe strokes may require more intensive medical care, including medication treatment and rehabilitation. Thrombectomy is a procedure used to remove a blood clot from a blood vessel, and is typically used in the treatment of acute ischemic stroke. While thrombectomy can be effective in reducing the severity of stroke and improving patient outcomes, it is also a relatively invasive procedure that can carry some risks and complications. As a result, patients who undergo thrombectomy may require longer hospital stays than those who do not. In summary, both thrombectomy and stroke severity are independent risk factors for prolonged LOS following AIS.

Our study has several limitations. First, its retrospective study design and only including patients from one single tertiary central hospital may limit the generalizability of the machine learning algorithm in clinical practice. Second, owing to the availability of the data, we were not able to consider more detailed factors, such as specific steps of reperfusion therapy, infarction or penumbra volume, and the collateral circulation status. More valuable and dynamic predictors could improve the performance. Third, some special reasons that might affect hospitalization time, such as economic stress or medical disputes, were not analyzed. Fourth, the sample size and certain bias limited the predictive ability of the model. We just internally validated our interpretable machine learning algorithms by bootstrap resample and multi-center large-sample studies are warranted to verify this conclusion in the future.

## Conclusion

We developed a model for predicting the prolonged LOS for AIS patients using the GNB algorithm. This model included 20 potential clinical factors and performed well in terms of discrimination, calibration, and clinical utility, but it needs to be validated in larger multicenter cohorts. In this model, pneumonia, dysphagia, thrombectomy, and stroke severity might be strong predictors of prolonged LOS. We explained these main variables

and analyzed the effects of their changing trends on prolonged LOS. Timely prevention and intervention for complications, as well as high quality standard of care, may be prospects worthy of clinicians' promising efforts.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of the Second Affiliated Hospital of Xuzhou Medical University. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

WL, LR, and HW completed the study design. KW and WL performed the study, and collected and analyzed the data. QJ and WL drafted the manuscript. LR, XW, KW, and HL provided the expert consultations and suggestions. MG, RG, CX, and CY conceived the study, participated in its design and coordination, and helped to embellish language. All authors contributed to the article and approved the submitted version.

## Funding

This study was supported by: Scientific Research Project of Jiangsu Health Committee (No.H2019054), the Xuzhou Science and Technology Planning Project (No. KC21220) and Science and Technology Development Fund of Affiliated Hospital of Xuzhou Medical University (No.XYFY202250), Shaanxi Provincial Health and Health Research Fund Project (2022E006).

## Conflict of interest

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.

## Publisher's note



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fendo.2023.1165178/full#supplementary-material

SUPPLEMENTARY TABLE 1
The hyperparameter settings for eight different machine learning algorithms. Abbreviations: XGB, extreme gradient boosting; LGBM, light gradient boosting machine; GNB, Gaussian naive Bayes; CNB, complement naive Bayes; MLP, multilayered perceptron; SVM, support vector machine.

SUPPLEMENTARY TABLE 2
The cross-reference between the full names and abbreviations in our manuscript.

## References

1. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. Heart disease and stroke statistics-2022 update: a report from the American heart association. *Circulation.* (2022) 145(8):e153–639. doi: 10.1161/CIR.0000000000001052

2. Pellico-Lopez A, Fernandez-Feito A, Cantarero D, Herrero-Montes M, Cayón-de Las Cuevas J, Parás-Bravo P, et al. Cost of stay and characteristics of patients with stroke and delayed discharge for non-clinical reasons. *Sci Rep Jun 27* (2022) 12 (1):10854. doi: 10.1038/s41598-022-14502-5

3. Salah HM, Minhas AMK, Khan MS, Khan SU, Ambrosy AP, Blumer V, et al. Trends in hospitalizations for heart failure, acute myocardial infarction, and stroke in the united states from 2004 to 2018. *Am Heart J* (2022) 243:103–9. doi: 10.1016/j.ahj.2021.09.009

4. Wang YJ, Li ZX, Gu HQ, Zhai Y, Zhou Q, Jiang Y, et al. China Stroke statistics: an update on the 2019 report from the national center for healthcare quality management in neurological diseases, China national clinical research center for neurological diseases, the Chinese stroke association, national center for chronic and non-communicable disease control and prevention, Chinese center for disease control and prevention and institute for global neuroscience and stroke collaborations. *Stroke Vasc Neurol* (2022) 7(5):415–50. doi: 10.1136/svn-2021-001374

5. Hao R, Qi X, Xia X, Wang L, Li X. Malnutrition on admission increases the in-hospital mortality and length of stay in elder adults with acute ischemic stroke. *J Clin Lab Anal* (2022) 36(1):e24132. doi: 10.1002/jcla.24132

6. Neale S, Leach K, Steinfort S, Hitch D. Costs and length of stay associated with early supported discharge for moderate and severe stroke survivors. *J Stroke Cerebrovasc Dis* (2020) 29(8):104996. doi: 10.1016/j.jstrokecerebrovasdis.2020.104996

7. Hassan AE, Ringheanu VM, Preston L, Tekle W, Qureshi AI. IV tPA is associated with increase in rates of intracerebral hemorrhage and length of stay in patients with acute stroke treated with endovascular treatment within 4.5 hours: should we bypass IV tPA in large vessel occlusion? *J neurointerv Surg* (2021) 13(2):114–8. doi: 10.1136/neurintsurg-2020-016045

8. Koton S, Luengo-Fernandez R, Mehta Z, Rothwell PM. Independent validation of the prolonged length of stay score. *Neuroepidemiology.* (2010) 35(4):263–6. doi: 10.1159/000320241

9. Hey T, Butler K, Jackson S, Thiyagalingam J. Machine learning and big scientific data. *Philos Trans A Math Phys Eng Sci* (2020) 378(2166):20190054. doi: 10.1098/rsta.2019.0054

10. Zhang W, Bao Z, Jiang S, He J. An artificial neural network-based algorithm for evaluation of fatigue crack propagation considering nonlinear damage accumulation. *Materials (Basel).* (2016) 9(6):484. doi: 10.3390/ma9060483

11. Kulkarni H, Thangam M, Amin AP. Artificial neural network-based prediction of prolonged length of stay and need for post-acute care in acute coronary syndrome patients undergoing percutaneous coronary intervention. *Eur J Clin Invest.* (2021) 51 (3):e13406. doi: 10.1111/eci.13406

12. Kurtz P, Peres IT, Soares M, Salluh JIF, Bozza FA. Hospital length of stay and 30-day mortality prediction in stroke: a machine learning analysis of 17,000 ICU admissions in Brazil. *Neurocrit Care* (2022) 37(Suppl 2):313–21. doi: 10.1007/s12028-022-01486-3

13. Yang CC, Bamodu OA, Chan L, Chen JH, Hong CT, Huang YT, et al. Risk factor identification and prediction models for prolonged length of stay in hospital after acute ischemic stroke using artificial neural networks. *Front Neurol* (2023) 14:1085178. doi: 10.3389/fneur.2023.1085178

14. Feske SK. Ischemic stroke. *Am J Med* (2021) 134(12):1457–64. doi: 10.1016/j.amjmed.2021.07.027

15. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American heart Association/American stroke association. *Stroke.* (2019) 50(12):e344–418. doi: 10.1161/STR.0000000000000211

16. Zivin JA, Sehra R, Shoshoo A, Albers GW, Bornstein NM, Dahlof B, et al. NeuroThera(R) efficacy and safety trial-3 (NEST-3): a double-blind, randomized, sham-controlled, parallel group, multicenter, pivotal study to assess the safety and efficacy of transcranial laser therapy with the NeuroThera(R) laser system for the treatment of acute ischemic stroke within 24 h of stroke onset. *Int J Stroke.* (2014) 9 (7):950–5. doi: 10.1111/j.1747-4949.2012.00896.x

17. Renner CJ, Kasner SE, Bath PM, Bahouth MN, Committee VAS. Stroke outcome related to initial volume status and diuretic use. *J Am Heart Assoc* (2022) 11(24): e026903. doi: 10.1161/JAHA.122.026903

18. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Ann Intern Med* (2015) 162(10):735–6. doi: 10.7326/L15-5093-2

19. Koton S, Bornstein NM, Tsabari R, Tanne D, Investigators N. Derivation and validation of the prolonged length of stay score in acute stroke patients. *Neurology.* (2010) 74(19):1511–6. doi: 10.1212/WNL.0b013e3181dd4dc5

20. Saposnik G, Webster F, O'Callaghan C, Hachinski V. Optimizing discharge planning: clinical predictors of longer stay after recombinant tissue plasminogen activator for acute stroke. *Stroke.* (2005) 36(1):147–50. doi: 10.1161/01.STR.0000150492.12838.66

21. Yaghi S, Harik SI, Hinduja A, Bianchi N, Johnson DM, Keyrouz SG. Post t-PA transfer to hub improves outcome of moderate to severe ischemic stroke patients. *J Telemed Telecare.* (2015) 21(7):396–9. doi: 10.1177/1357633X15577531

22. Sato S, Toyoda K, Uehara T, Toratani N, Yokota C, Moriwaki H, et al. Baseline NIH stroke scale score predicting outcome in anterior and posterior circulation strokes. *Neurology* (2008) 70(24 Pt 2):2371–7. doi: 10.1212/01.wnl.0000304346.14354.0b

23. de Celis-Ruiz E, Fuentes B, Alonso de Lecinana M, Gutiérrez-Fernández M, Borobia AM, Gutiérrez-Zúñiga R, et al. Final results of allogeneic adipose tissue-derived mesenchymal stem cells in acute ischemic stroke (AMASCIS): a phase II, randomized, double-blind, placebo-controlled, single-center, pilot clinical trial. *Cell Transplant.* (2022) 31:9636897221083863. doi: 10.1177/09636897221083863

24. Maeshima S, Osawa A, Hayashi T, Tanahashi N. Elderly age, bilateral lesions, and severe neurological deficit are correlated with stroke-associated pneumonia. *J Stroke Cerebrovasc Dis* (2014) 23(3):484–9. doi: 10.1016/j.jstrokecerebrovasdis.2013.04.004

25. Garner JS, Jarvis WR, Emori TG, Horan TC, Hughes JM. CDC Definitions for nosocomial infections, 1988. *Am J Infect Control.* (1988) 16(3):128–40. doi: 10.1016/0196-6553(88)90053-3

26. Martino R, Foley N, Bhogal S, Diamant N, Speechley M, Teasell R. Dysphagia after stroke: incidence, diagnosis, and pulmonary complications. *Stroke.* (2005) 36 (12):2756–63. doi: 10.1161/01.STR.0000190056.76543.eb

27. Warner JJ, Harrington RA, Sacco RL, Elkind MSV. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke. *Stroke.* (2019) 50(12):3331–2. doi: 10.1161/STROKEAHA.119.027708

28. Su M, Pan D, Zhao Y, Chen C, Wang X, Lu W, et al. The direct and indirect effects of length of hospital stay on the costs of inpatients with stroke in ningxia, China, between 2015 and 2020: a retrospective study using quantile regression and structural equation models. *Front Public Health* (2022) 10:881273. doi: 10.3389/fpubh.2022.881273

29. Lin KH, Lin HJ, Yeh PS. Determinants of prolonged length of hospital stay in patients with severe acute ischemic stroke. *J Clin Med* (2022) 11(12):3457. doi: 10.3390/jcm11123457

30. Svendsen ML, Ehlers LH, Andersen G, Johnsen SP. Quality of care and length of hospital stay among patients with stroke. *Med Care* (2009) 47(5):575–82. doi: 10.1097/MLR.0b013e318195f852

31. Borghans I, Hekkert KD, den Ouden L, Cihangir S, Vesseur J, Kool RB, et al. Unexpectedly long hospital stays as an indicator of risk of unsafe care: an exploratory study. *BMJ Open* (2014) 4(6):e004773. doi: 10.1136/bmjopen-2013-004773

32. Beckers V, De Smedt A, Van Hooff RJ, De Raedt S, Van Dyck R, Putman K, et al. Prediction of hospitalization duration for acute stroke in Belgium. *Acta Neurol Belg.* (2012) 112(1):19–25. doi: 10.1007/s13760-012-0026-0

33. Spratt N, Wang Y, Levi C, Ng K, Evans M, Fisher J. A prospective study of predictors of prolonged hospital stay and disability after stroke. *J Clin Neurosci* (2003) 10(6):665–9. doi: 10.1016/j.jocn.2002.12.001

34. Bergh C, Udumyan R, Appelros P, Fall K, Montgomery S. Determinants in adolescence of stroke-related hospital stay duration in men: a national cohort study. *Stroke.* (2016) 47(9):2416–8. doi: 10.1161/STROKEAHA.116.014265

35. Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. *Curr Rev Musculoskelet Med* (2020) 13(1):69–76. doi: 10.1007/s12178-020-09600-8

36. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak.* (2019) 19(1):281. doi: 10.1186/s12911-019-1004-8

37. Raizada RD, Lee YS. Smoothness without smoothing: why Gaussian naive bayes is not naive for multi-subject searchlight studies. *PloS One* (2013) 8(7):e69566. doi: 10.1371/journal.pone.0069566

38. Ontivero-Ortega M, Lage-Castellanos A, Valente G, Goebel R, Valdes-Sosa M. Fast Gaussian naive bayes for searchlight classification analysis. *Neuroimage.* (2017) 163:471–9. doi: 10.1016/j.neuroimage.2017.09.001

39. Cohen DL, Roffe C, Beavan J, Blackett B, Fairfield CA, Hamdy S, et al. Post-stroke dysphagia: a review and design considerations for future trials. *Int J Stroke.* (2016) 11(4):399–411. doi: 10.1177/1747493016639057

40. Ogawa Y, Inagawa M, Kimura M, Iida T, Hirai A, Yoshida T, et al. Nutritional intervention after an early assessment by a flexible endoscopic evaluation of swallowing is associated with a shorter hospital stay for patients with acute cerebral infarction: a retrospective study. *Asia Pac J Clin Nutr* (2021) 30(2):199–205. doi: 10.6133/apjcn.202106_30(2).0003

41. Eltringham SA, Kilner K, Gee M, Sage K, Bray BD, Pownall S, et al. Impact of dysphagia assessment and management on risk of stroke-associated pneumonia: a systematic review. *Cerebrovasc Dis* (2018) 46(3-4):99–107. doi: 10.1159/000492730

42. Smith CJ, Kishore AK, Vail A, Chamorro A, Garau J, Hopkins SJ, et al. Diagnosis of stroke-associated pneumonia: recommendations from the pneumonia in stroke consensus group. *Stroke.* (2015) 46(8):2335–40. doi: 10.1161/STROKEAHA.115.009617

43. Jones CA, Colletti CM, Ding MC. Post-stroke dysphagia: recent insights and unanswered questions. *Curr Neurol Neurosci Rep* (2020) 20(12):61. doi: 10.1007/s11910-020-01081-z

# Frontiers in Endocrinology

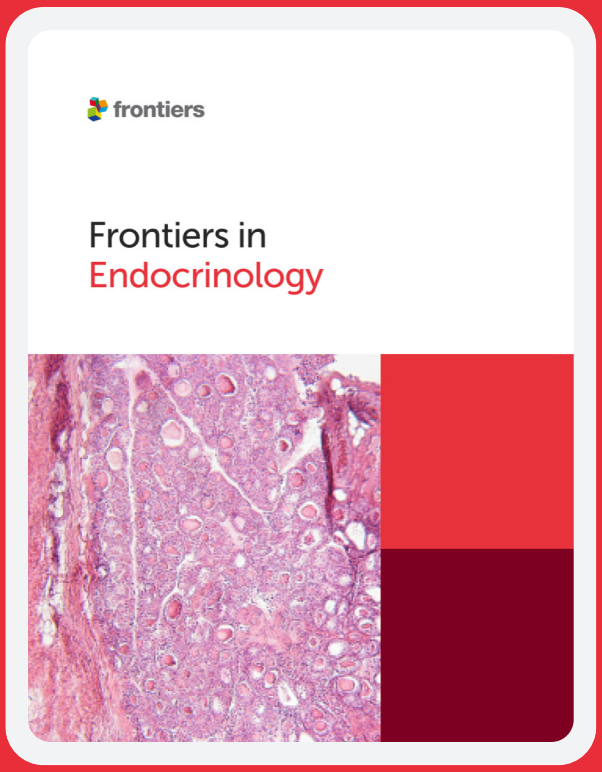
**Explores the endocrine system to find new therapies for key health issues**

The second most-cited endocrinology and metabolism journal, which advances our understanding of the endocrine system. It uncovers new therapies for prevalent health issues such as obesity, diabetes, reproduction, and aging.

---

## Discover the latest Research Topics

See more →




**Frontiers**

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

**Contact us**

+41 (0)21 510 17 00
frontiersin.org/about/contact