frontiers | Research Topics

Applications of RNA-seq in cancer and tumor research

Edited by
Jidong Lang, Taoyang Wu, Tao Huang, Junlin Xu
and William C. Cho

Published in
Frontiers in Genetics
Frontiers in Oncology

ISSN 1664-8714
ISBN 978-2-8325-4022-0
DOI 10.3389/978-2-8325-4022-0

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

Applications of RNA-seq in cancer and tumor research

Topic editors
Jidong Lang — Qitan Technology Co., Ltd., China
Taoyang Wu — University of East Anglia, United Kingdom
Tao Huang — Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences (CAS), China
Junlin Xu — Hunan University, China
William C. Cho — QEH, Hong Kong, SAR China

Citation
Lang, J., Wu, T., Huang, T., Xu, J., Cho, W. C., eds. (2023). *Applications of RNA-seq in cancer and tumor research*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-4022-0

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 14 November 2023
DOI 10.3389/fgene.2023.1331576

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Fangqing Zhao,
Beijing Institutes of Life Science (CAS),
China

*CORRESPONDENCE
Jidong Lang,
✉ langjidong@hotmail.com

RECEIVED 01 November 2023
ACCEPTED 07 November 2023
PUBLISHED 14 November 2023

CITATION
Lang J, Cho WC, Huang T, Wu T and Xu J (2023), Editorial: Applications of RNA-seq in cancer and tumor research.
Front. Genet. 14:1331576.
doi: 10.3389/fgene.2023.1331576

Editorial: Applications of RNA-seq in cancer and tumor research

Jidong Lang[1]*, William C. Cho[2], Tao Huang[3], Taoyang Wu[4] and Junlin Xu[5]

[1]Department of Bioinformatics, Qitan Technology (Beijing) Co., Ltd., Beijing, China, [2]Department of Clinical Oncology, Queen Elizabeth Hospital, Hong Kong SAR, China, [3]Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences (CAS), Shanghai, China, [4]School of Computing Science, University of East Anglia, Norwich, United Kingdom, [5]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

KEYWORDS

RNA sequencing, cancer research, next-generation sequencing, biomarker, machine learning

Editorial on the Research Topic
Applications of RNA-seq in cancer and tumor research

Over the past decade, RNA sequencing, commonly referred to as RNA-Seq, has emerged as a powerful and significant tool, leading to remarkable advancements in the fields of cancer and tumor research. Presently, RNA-Seq is extensively employed in molecular biology, playing a pivotal role in enhancing our comprehension of genome functions, particularly those relevant to cancer research (Stark et al., 2019). Notably, it has become an indispensable tool for conducting whole-transcriptome analysis, enabling the study of differential gene expression and differential splicing of mRNAs. Transcript isoform expression and usage, a key source of variation between healthy and cancerous or diseased tissues in various medical conditions, can be effectively investigated using this technique (Gonzalez and McGraw, 2009). Undoubtedly, the advent of sequencing technologies, such as next-generation sequencing (NGS) and nanopore sequencing, has facilitated comprehensive transcriptome analysis, leading to significant breakthroughs in cancer and tumor research. These technologies have enabled the examination of single-cell gene expression, translation, RNA structure, and spatial transcriptomics. Nanopore sequencing stands out for its ability to deliver full-length transcripts accurately and to identify and quantify multiple isoforms, making it particularly valuable for cancer research. This technology has been successfully applied in the study of various cancer types, including leukaemia, breast cancer, ovarian cancer, and lung cancer (de Jong et al., 2017; Minervini et al., 2017; Suzuki et al., 2017). Some studies have even suggested that RNA-Seq has the potential to revolutionize the analysis of eukaryotic transcriptomes (Wang et al., 2009). Its ability to investigate diverse aspects of RNA biology in cancer and tumors is critical for developing a functional understanding of the genome, studying development processes, and identifying molecular dysregulation underlying cancer and other diseases. Consequently, RNA-Seq has already assumed a vital role in practical clinical applications (Byron et al., 2016; Haque et al., 2017). In this Research Topic, we have compiled 11 papers that highlight several frontiers in the role of RNA-Seq in cancer and tumor research.

Du et al. focused on investigating the genomic effects of high-dose single-shot radiotherapy with the aim of providing a dynamic map for non-small cell lung cancer (NSCLC). The authors employed whole-transcriptome sequencing to elucidate molecular-level changes in A549 and H1299 cell lines exposed to 10 Gy X-rays at different time points, comparing them to a no radiation group, and found dynamic changes following radiation therapy within 48 h. Their findings emphasized key molecules and pathways involved in NSCLC after high-dose single-shot radiotherapy. This study contributes to enriching the content of radiobiology in precision radiation oncology.

Jin et al. utilized a published single-cell transcriptomics profile to deconvolute the abundance of cell types among paired plasma samples from colorectal cancer patients who underwent tumor-ablative surgery. Their objective was to identify the tissue-specific contributions of circulating cell-free RNA (cfRNA) transcriptomic profiles. Furthermore, they validated differentially expressed cfRNAs using RNA-Seq. The authors observed a significant decrease in the transcriptomic component from intestinal secretory cells in post-surgical cfRNA samples. They also found consistent expression of *HPGD*, *PACS1*, and *TDP2* between cfRNA and tissue samples, indicating the potential of these markers for minimal residual disease (MRD) testing, which involves profiling remnants cancer cells after or during treatment.

Song et al. identified key genes associated with cuproptosis and ferroptosis (*POR*, *SLC7A5* and *STAT3*) involved in sepsis-induced cardiomyopathy (SIC). Additionally, they explored therapeutic drug candidates. This work holds promise for the development of treatments for SIC.

Nousiainen et al. conducted RNA-Seq analysis on xenografts and immortalized cell lines to gain insights into the pathobiology of hepatoblastoma (HB). Through protein-protein interaction analysis, they identified ubiquitination as a key dysregulated pathway in HB. The study also revealed the potential prognostic utility of *UBE2C* in HB and highlighted the ubiquitin pathway as a potential therapeutic target of the disease.

Zhu et al. provided a comprehensive summary of the main methods for detecting circulating tumor DNA (ctDNA), including PCR-based sequencing and NGS, along with their respective advantages and disadvantages. Additionally, the authors reviewed the significance of ctDNA analysis in guiding adjuvant therapy and predicting relapse in lung, breast, and colon cancers, among others. Despite the existing challenges in MRD detection, the feasibility of ctDNA as a detection method and the revolutionary potential of ctDNA-based liquid biopsies offer a promising approach to cancer monitoring.

Xie et al. developed a prognostic risk model and identified immune ferroptosis-related genes with independent prognostic value using procedural algorithm analysis. Their findings demonstrated significant correlations between immune scores, immune checkpoints, and chemotherapeutic agents with prognostic models. These features were subsequently considered as independent prognostic factors for predicting overall survival (OS) and clinical treatment response in breast cancer patients. This study provides a better understanding of the contribution of immune ferroptosis-related genes in breast cancer and highlights their potential as prognostic markers and therapeutic targets.

Wang et al. employed consensus clustering to identify two disulfidptosis-molecular subtypes in breast cancer, with differing OS outcomes. Subsequently, the authors developed a prognostic signature based on differentially expressed genes related to disulfidptosis, which demonstrated improved predictive capabilities for patient survival and provided preliminary insights into the relationship between the risk model and the immune landscape. This study offers valuable prognostic predictions for breast cancer patients, with prognostic signatures closely associated with the tumor microenvironment, potentially informing clinical treatment decisions.

Niu et al. proposed a microRNA (miRNA) and small molecule association prediction model, named GCNNMMA, by integrating graph neural networks and convolutional neural networks. This model inspired by ensemble learning, demonstrated superior cross-validation results compared to other comparative models, suggesting the effectiveness of GCNNMMA in mining the relationship between small molecule drugs and disease-relevant miRNAs. GCNNMMA holds promise as a valuable tool for exploring the associations between small molecules and miRNAs in disease contexts.

Li et al. developed a novel ensemble model, called autoencoder-assisted graph convolutional neural network (AE-GCN), that combined autoencoder and graph convolutional neural network techniques to identify accurate and fine-grained spatial domains. In cancer datasets, AE-GCN successfully identified disease-related spatial domains, revealing more heterogeneity than traditional histological annotations. Moreover, AE-GCN facilitated the discovery of novel differentially expressed genes with significant prognostic relevance. This study demonstrates the ability of AE-GCN to unveil complex spatial patterns from spatially resolved transcriptomics data.

Chen et al. addressed the lack of a specialized database focusing on alternative splicing events (ASEs) in esophageal squamous cell carcinoma (ESCC) and the underrepresentation of long non-coding RNAs (lncRNAs) in ESCC molecular mechanisms with the development of a database, called DASES. DASES provides comprehensive insights into ASEs in ESCC, encompassing both lncRNAs and mRNAs, thereby enhancing the understanding of ESCC molecular mechanisms and serving as a valuable resource for the ESCC research community.

Su et al. introduced a machine learning-based method, called LDAenDL, which utilizes an ensemble of deep neural networks and LightGBM, to detect potential lncRNA biomarkers for lung cancer and neuroblastoma. The authors demonstrated that LDAenDL outperformed classical LDA prediction methods, and identified new potential biomarkers for these diseases. The application of LDAenDL may facilitate the development of targeted therapies for lung cancer and neuroblastoma.

In summary, these papers demonstrate the diverse applications of RNA-Seq in cancer and tumor research. The studies utilize RNA-Seq to identify differentially expressed genes, explore molecular mechanisms, and identify potential therapeutic targets in various types of cancer. The findings contribute to our understanding of cancer biology and highlight the potential of RNA-Seq in improving cancer diagnosis, prognosis, and treatment.

Author contributions

JL: Conceptualization, Investigation, Project administration, Writing–original draft. WC: Investigation, Writing–review and editing. TH: Writing–review and editing. TW: Writing–review and editing. JX: Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We would like to express our gratitude to the authors for their valuable contributions to this Research Topic, as well as to the reviewers who kindly agreed to assess the individual submissions.

Conflict of interest

Author JL was employed by the company Qitan Technology (Beijing) Co., Ltd.

The remaining 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

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.

References

Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D., and Craig, D. W. (2016). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* 17, 257–271. doi:10.1038/nrg.2016.10

de Jong, L. C., Cree, S., Lattimore, V., Wiggins, G. A. R., Spurdle, A. B., kConFab, I., et al. (2017). Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. *Breast Cancer Res.* 19, 127. doi:10.1186/s13058-017-0919-1

Gonzalez, E., and McGraw, T. E. (2009). The Akt kinases: isoform specificity in metabolism and cancer. *Cell. Cycle* 8, 2502–2508. doi:10.4161/cc.8.16.9335

Haque, A., Engel, J., Teichmann, S. A., and Lonnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* 9, 75. doi:10.1186/s13073-017-0467-4

Minervini, C. F., Cumbo, C., Orsini, P., Anelli, L., Zagaria, A., Impera, L., et al. (2017). Mutational analysis in BCR-ABL1 positive leukemia by deep sequencing based on nanopore MinION technology. *Exp. Mol. Pathol.* 103, 33–37. doi:10.1016/j.yexmp.2017.06.007

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656. doi:10.1038/s41576-019-0150-2

Suzuki, A., Suzuki, M., Mizushima-Sugano, J., Frith, M. C., Makalowski, W., Kohno, T., et al. (2017). Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer. *DNA Res.* 24, 585–596. doi:10.1093/dnares/dsx027

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi:10.1038/nrg2484

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 03 March 2023
DOI 10.3389/fgene.2023.1126236

Check for updates

OPEN ACCESS

EDITED BY
Junlin Xu,
Hunan University, China

REVIEWED BY
Houwu Gong,
Hunan University, China
Narasimha Kumar Karanam,
University of Texas Southwestern Medical Center, United States

*CORRESPONDENCE
Tiankui Qiao,
✉ qiaotiankui@fudan.edu.cn

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 17 December 2022
ACCEPTED 13 February 2023
PUBLISHED 03 March 2023

CITATION
Du Y, Zheng Y, Yu K, Zhan C and Qiao T (2023), Genome-wide analyses of lung cancer after single high-dose radiation at five time points (2, 6, 12, 24, and 48 h). *Front. Genet.* 14:1126236.
doi: 10.3389/fgene.2023.1126236

Genome-wide analyses of lung cancer after single high-dose radiation at five time points (2, 6, 12, 24, and 48 h)

Yajing Du[1], Yunna Zheng[2], Kaiwen Yu[3], Cheng Zhan[4] and Tiankui Qiao[1]*

[1]Center for Tumor Diagnosis and Therapy, Jinshan Hospital, Fudan University, Shanghai, China, [2]OE Biotech Co., Ltd., Shanghai, China, [3]Shanghai Chest Hospital, Shanghai Jiaotong University, Shanghai, China, [4]Department of Thoracic Surgery, Zhongshan Hospital, Fudan University, Shanghai, China

Background: An increasing number of clinicians are experimenting with high-dose radiation. This study focuses on the genomic effects of high-dose single-shot radiotherapy and aims to provide a dynamic map for non-small cell lung cancer (NSCLC).

Methods: We used whole-transcriptome sequencing to understand the evolution at molecular levels in A549 and H1299 exposed to 10 Gy X-rays at different times (2, 6, 12, 24, and 48 h) in comparison with the no radiation group. Ingenuity pathway analysis, ceRNA analysis, enrichment analysis, and cell cycle experiments are performed for molecular analyses and function analyses.

Results: Whole-transcriptome sequencing of NSCLC showed a significant dynamic change after radiotherapy within 48 h. MiR-219-1-3p and miR-221-3p, miR-503-5p, hsa-miR-455-5p, hsa-miR-29-3p, and hsa-miR-339-5p were in the core of the ceRNA related to time change. GO and KEGG analyses of the top 30 mRNA included DNA repair, autophagy, apoptosis, and ferroptosis pathways. Regulation of the cell cycle-related transcription factor E2F1 might have a key role in the early stage of radiotherapy (2.6 h) and in the later stage of autophagy (24 and 48 h). Functions involving different genes/proteins over multiple periods implied a dose of 10 Gy was related to the kidney and liver pathway. Radiation-induced cell cycle arrest at the G2/M phase was evident at 24 h. We also observed the increased expression of CCNB1 at 24 h in PCR and WB experiments.

Conclusion: Our transcriptomic and experimental analyses showed a dynamic change after radiation therapy in 48 h and highlighted the key molecules and pathways in NSCLC after high-dose single-shot radiotherapy.

KEYWORDS

whole-transcriptome sequencing, non-small cell lung cancer, radiobiology, precision radiotherapy, bioinformatics

1 Introduction

Lung cancer (LC) accounted for the world's highest mortality rate and second-highest incidence rate in 2022 (Siegel et al., 2022). Radiotherapy (RT) can cure about 40% of cancers (De Ruysscher et al., 2019), which has bright therapeutic prospects for patients.

Precision radiotherapy aims to optimize outcomes and minimize toxicity to patients (Joseph and Vijayakumar, 2020). Most researchers are currently studying the balance of the dose (Scott et al., 2021) or the spatial depth per fraction to decrease side effects. By using artificial intelligence, dose distributions can be predicted based on the anatomy of a patient and calculated more quickly (Hosny et al., 2018; Huynh et al., 2020; Luk et al., 2022; Teuwen et al., 2022). In clinical practice, doctors usually adapt 24 h or 48 h/fraction (fx). For a high dose (such as 10–12 Gy), the total time of five fractions can range from 1.5 to 2 weeks (Chmura et al., 2021).The hours of the fraction are not accurate, and few studies discussed suitable hour of fraction, involving less dynamic changes.

In recent years, precision radiotherapy applied high-dose therapy (Burkoň et al., 2022; Chairmadurai et al., 2022; DLP et al., 2022; Milic et al., 2022; Sidaway, 2022; Tadimalla et al., 2022). Stereotactic body radiotherapy (SBRT) has the characteristics of high tumor dose distribution in the irradiation center and a rapid drop of extradural dose. For lung cancers that are early-stage and inoperable, this is the standard radiation therapy (Lo et al., 2010; Timmerman et al., 2014). The efficacy and toxicity of stereotactic body radiotherapy in patients with centrally located non-small cell lung cancer (10–12 Gy/fraction) were studied (Chmura et al., 2021).

Our study designed groups after radiation for 2, 6, 12, 24, and 48 h to reveal the characteristics of different time periods to discover the suitable interval time for multi-fractions and explore the dynamic change of a gene caused by radiation in single-fraction therapy. Meanwhile, we used the whole-transcriptome sequencing method to learn radiobiology from the perspective of a genome. This enriches the radiobiological content of high-dose radiation therapy, providing biological basics for treatment of SBRT and suggesting new possible molecular methods for combined targeted therapy and chemotherapy.

2 Materials and methods

2.1 Ethics statement

The Ethics Committees of Jinshan Hospital of Fudan University exempted the study because no personal information is included in the study.

2.2 Transcriptome sequencing sample preparation

For the present study, NSCLC cells (A549 and H1299) in a six-well plate at 40% density were divided into no radiation and radiation groups. The radiation group was split into five time points, with two repeats per group.

The radiation group was exposed to a single high dose (Trilogy linear accelerator, 6 MV X-ray radiation, absorption dose rate of 600 cGy/min, once, 10 Gy dose). The cells were washed with PBS twice. TRIzol was added to lysis cells at 2, 6, 12, 24, and 48 h after radiation with the no radiation group. The whole transcriptome was sequenced in a total of 24 samples.

2.3 Cell cycle assays

A549 and H1299 were collected at 2, 6, 12, 24, and 48 h after radiation and fixed in 70% ethyl alcohol at −20°C overnight together with the no radiation group. For 15 min, they were incubated in 0.5 mL PI/RNase Staining Buffer (BD Biosciences, Franklin, NJ, United States) after three washes with PBS. Flow cytometry was used to analyze the cell fractions (Beckman Coulter or BD Biosciences in the United States).

2.4 Western blot detection

Cells were washed twice with PBS and lysed at 4°C for 30 min. Purities were selected by centrifugation at 15,000*g, at 4°C for 20 min, 10% SDS-PAGE was used to separate proteins, and a nitrocellulose filter was used for transfer. All samples were evenly transferred and incubated in a closed solution for 2 hours at room temperature using a stained filter. Anti-CCNB1 was diluted at 1: 1000 for 12 h, washed twice with PBS and TBST, and then exposed to the filter. The filter was incubated with the secondary antibody, at 1:1000 for 1 h, and then washed with TBST. In addition, anti-β-actin antibodies were used as an internal reference.

2.5 Real-time fluorescence quantitative polymerase chain reaction detection

The RNA Purification Kit (Yishan Biotechnology Company, Shanghai, China) and the 5x Reverse Transcriptase Master Mix (Takara, Osaka, Japan) were used to obtain cDNA. The primers were as follows: β-actin, forward 5′-TGACGTGGACATCCGCAAAG-3′, reverse 5′-CTGGAAGGTGGACAGCGAGG-3′; CCNB1, forward 5′-AATAAGGCGAAGATCAACATGGC-3′, reverse 5′-TTTGTTACCAATGTCCCCAAGAG-3′.

2.6 Bioinformatics analysis

2.6.1 Differential mRNA, miRNA, circRNA, and lncRNA

The original data were standardized. The mRNA, miRNA, circRNA, and lncRNA of the 2-, 6-, 12-, 24-, and 48-h treatment groups and the no radiation group were analyzed by the DESeq package of the R language software. |log2 (fold change) | >1 and $p < 0.05$ were set as the criteria for intergroup differences.

The no radiation group of A549 and H1299 was also analyzed with the DESeq package to obtain the differential mRNA, miRNA, circRNA, and lncRNA, named as NCdiffmRNA, NCdiffmiRNA, NCdiffcircRNA, and NCdifflncRNA, respectively. |log2 (fold

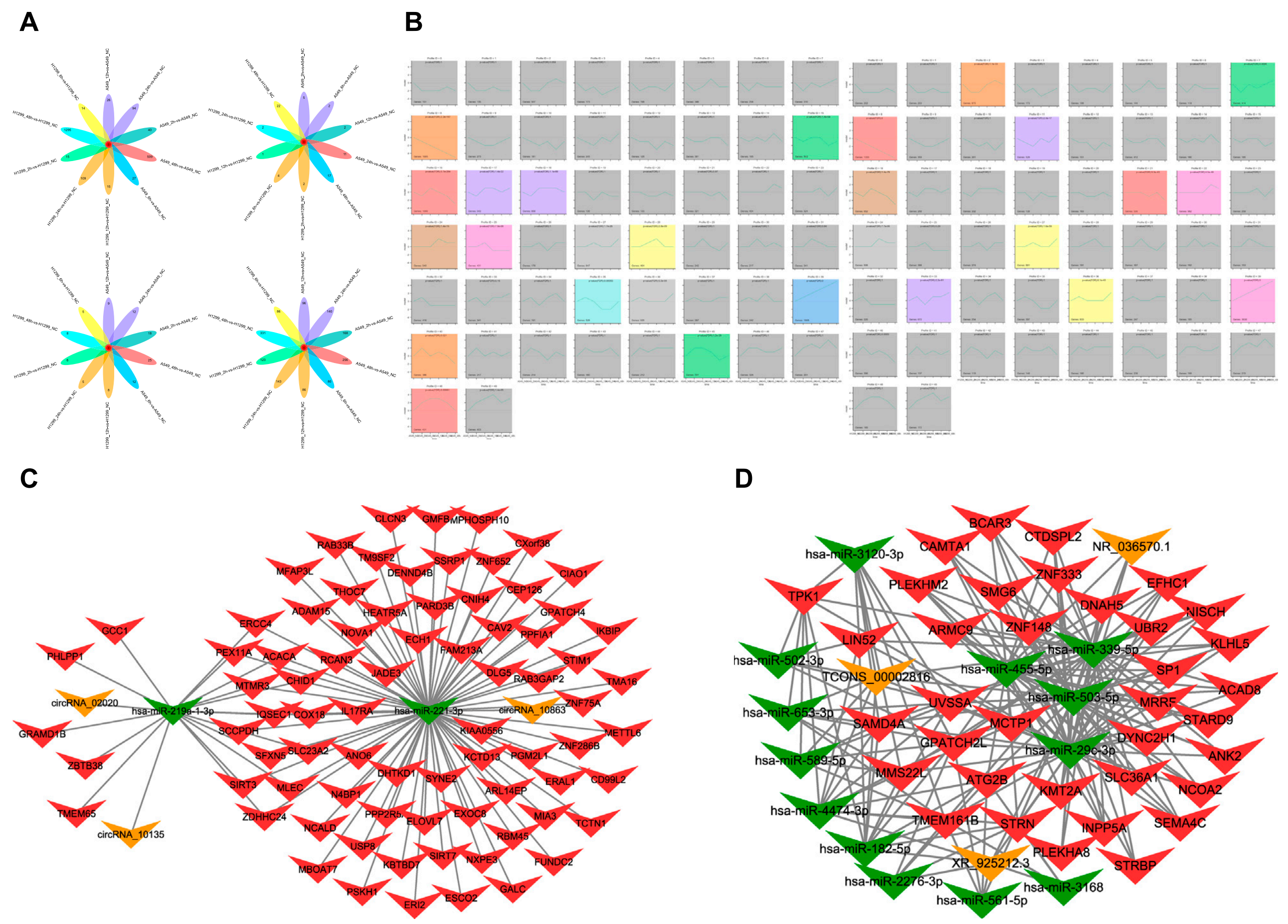


FIGURE 1
(A) Diagram of intersections in different mRNA, miRNA, lncRNA, and circRNA after radiation of A549 and H1299 cells in five time points (2, 6, 12, 24, and 48 h) compared with non-irradiated cells. **(B)** Trend chart of mRNA after radiotherapy in A549 and H1299 cells in STEM analyses (a screenshot of the main interface window of the STEM is found in Figure 1B. In this window, each box corresponds to one of the model temporal expression profiles. The data were sampled at five time points 2, 6, 12, 24, and 48 h. The number at the top of a profile box is the profile ID number. The colored profiles had a statistically significant number of genes assigned). **(C)** Significant ceRNA network related to five time points (2, 6, 12, 24, and 48 h) after radiation therapy in A549 and H1299 cells: mRNA–miRNA–circRNA network. **(D)** Significant ceRNA network related to five time points (2, 6, 12, 24, 48 h) after radiation therapy in A549 and H1299 cells: mRNA–miRNA–lncRNA network.

change) | >1 and $p < 0.05$ were set as the criteria. These NCdiffRNAs (NCdiffmRNA, NCdiffmiRNA, NCdiffcircRNA, and NCdifflncRNA) represent the difference caused by the cell line. A549 is an epithelial cell isolated from the lungs of a 58-year-old white male with carcinoma. H1299 is isolated from the lungs of a 43-year-old white male patient with carcinoma.

2.6.2 Short Time-series Expression Miner

The analysis samples were analyzed with the Short Time-series Expression Miner (STEM) (Ernst and Bar-Joseph, 2006) in the order of [“A549_NC”—> “A549_2 h”—> “A549_6 h”—> “A549_12 h”—> “A549_24 h”—> “A549_48 h”]. The *p*-value was corrected by the false discovery rate method, and the significant modules with *p*-value less than 0.05 were selected. A total of 16 significant modules of A549 were screened in 50 modules. A total of 39 significant modules were also screened in the order of ["H1299_NC”—> “H1299_2 h”—> “H1299_6 h”—> “H1299_12 h”—> “H1299_24 h”—> “H1299_48 h”]. The trend map and clustering heatmap of the significant module in A549 and H1299 were drawn (Figure 1B). These mRNAs in significant modules related to time after radiation were recorded as STEM genes.

The intersection between the NCdiffmRNA and STEM gene was taken, named as diffmRNA. DiffmRNA represented differences after radiation in 48 h caused by the two cell lines. The remaining common gene was named commonmRNA. CommonmRNA represents common genes of non-small cell lung cancer, regardless of the differences caused by the two cell lines after radiation in 48 h. By applying the same process for NCdiffcircRNA, NCdifflncRNA, and NCdiffmiRNA, we got commoncircRNA, commonlncRNA, and commonmiRNA, respectively.

2.6.3 Ingenuity pathway analysis

CommonmRNA at 2, 6, 12, 24, and 48 h was analyzed by Ingenuity pathway analysis (IPA, http://www.ingenuity.com). The setting is shown in Supplementary Materials (S2–S6).

2.6.4 CeRNA analysis and enrichment analysis

Pearson r was used to calculate the correlation of 24 samples. MiRNA–mRNA relationship pairs were screened (the absolute correlation coefficient value greater than or equal to 0.60, and the *p*-value was less than or equal to 0.05). According to the mechanism of action of miRNA and mRNA, the relationship pairs of negative regulation were screened, and miRNA–mRNA relationship pairs were screened. The miRanda program was used to predict the binding between these miRNA–mRNA sequences, using the default parameter of miRanda v3.3a. Finally, pairs of miRNA–mRNA relationships were obtained. Pairs of miRNA–circRNA relationships were also obtained by the same way.

For these predicted relationships, the MuTaME method was performed to get a ceRNA score (Tay et al., 2011). At the same time, the *p*-value corresponding to the ceRNA relationship was calculated in combination with the hypergeometric distribution, and the smaller the *p*-value, the more significant these miRNAs shared between the two ceRNAs (mRNA and target).

MRNA–circRNA relationship pairs was screened by Pearson r (the absolute correlation coefficient value greater than or equal to 0.60, and the *p*-value was less than or equal to 0.05). According to the role of mRNA–circRNA in the ceRNA relationship, the relationship between mRNA and circRNA with positive correlation was screened, and the results of the ceRNA score were calculated and the two intersected. Then, the ceRNA results helped build the ceRNA network.

GO and KEGG pathway analyses were performed on the mRNA in the ceRNA network. The top 30 mRNAs in the RNA score in the mRNA–miRNA–circRNA network and the mRNA–miRNA–lncRNA network are used for pathway enrichment by GO and KEGG analyses, separately. CeRNA analysis and enrichment analysis of mRNA gene sets helped obtain key regulatory network molecules and key pathways that may be caused by radiation in 48 h.

2.7 Statistical analysis

Line charts and histograms were produced by GraphPad 7.0. Bioinformatics analysis was carried out using the R language (Version 4.0.0). The gray value of protein bands was analyzed by ImageJ software, and statistical analysis was carried out using SPSS 24.0. Also, the comparison of two sets of disordered variables was *t*-tested; the categorical variables were chi-squared. The bilateral $p < 0.05$ was statistically significant.

3 Results

3.1 Whole-transcriptome sequencing of NSCLC cells

A flowchart is shown in Supplementary Materials S1. The differential mRNA of the no radiation group of A549 and H1299 (NCdiffmRNA) has 5452 genes, which is related to the intrinsic difference between the two cell lines. A total of 2755 genes were upregulated and 2697 downregulated in NCdiffmRNA. The STEM intersection gene is composed of 5076 genes, which is associated with time change after radiation within 48 h. Intersecting diffmRNA has 576 genes, related to the intrinsic difference of two cell lines and time change after radiation. CommonmRNA has 4509 genes, which is related to the time change after radiation within 48 h in non-small cell lung cancer.

The whole-transcriptome sequencing results of A549 and H1299 showed that the intersections of different mRNA at different times are 0 compared with non-irradiated cells. The differential mRNAs at 2, 6, 12, 24, and 48 h of A549, compared with the no radiation group, are 40 genes, 27 genes, 26 genes, 84 genes, and 509 genes, respectively. The differential mRNAs at 2, 6, 12, 24, and 48 h of H1299, compared with the no radiation group, are 15 genes, 14 genes, 15 genes, 109 genes, and 1295 genes, respectively. The results of the intersection of differential miRNAs, lncRNAs, and circRNAs are 0, suggesting that the genome is in a significant dynamic change within 48 h after radiation in NSCLC (Figure 1A).

The differential lncRNAs at 2, 6, 12, 24, and 48 h in A549 are 168 lncRNAs, 80 lncRNAs, 96 lncRNAs, 140 lncRNAs, and 256 lncRNAs, respectively. The differential lncRNAs at 2, 6, 12, 24, and 48 h in H1299 are 123 lncRNAs, 86 lncRNAs, 86 lncRNAs, 143 lncRNAs, and 331 lncRNAs, respectively. The results of the difference between miRNAs and circRNAs are shown in Figure 1A.

3.2 The STEM analysis of temporal trends of mRNAs

A total of 16 significant modules were screened in A549, and 13 significant modules were screened in H1299. We plotted a meaningful module trend map (Figure 1B) and took the union of genes in 39 modules. Finally, we got 5076 genes in 39 modules related to the temporal changes in gene expression after radiation in NSCLC.

3.3 CeRNA analyses and enrichment analyses

The ceRNA score is used to obtain the mRNA–miRNA–lncRNA network (Figure 1C), and hsa-miR-503-5p, hsa-miR-455-5p, hsa-miR-29c-3p, and hsa-miR-339-5p are located at the core of ceRNA. GO analysis showed that evident biological processes include DNA repair, negative regulation of G2/M transition of the mitotic cell cycle, protein polyubiquitination, ER to Golgi vesicle-mediated transport, and intracellular protein transport. KEGG analyses showed that evident pathways include autophagy, ferroptosis, endocytosis, purine metabolism, neurotrophin signaling pathway, and insulin signaling pathway (Figures 2A, B). Similarly, in the mRNA–miRNA–circRNA network (Figure 1D), miR-219-1-3p and miR-221-3p are in the core. GO analysis showed that evident biological processes include the intra-Golgi vesicle-mediated transport, positive regulation of the canonical Wnt signaling pathway, negative regulation of Arp2/3 complex-mediated actin nucleation, the SCF-dependent proteasomal ubiquitin-dependent protein catabolic process, and regulation of the Arp2/3 complex-mediated actin

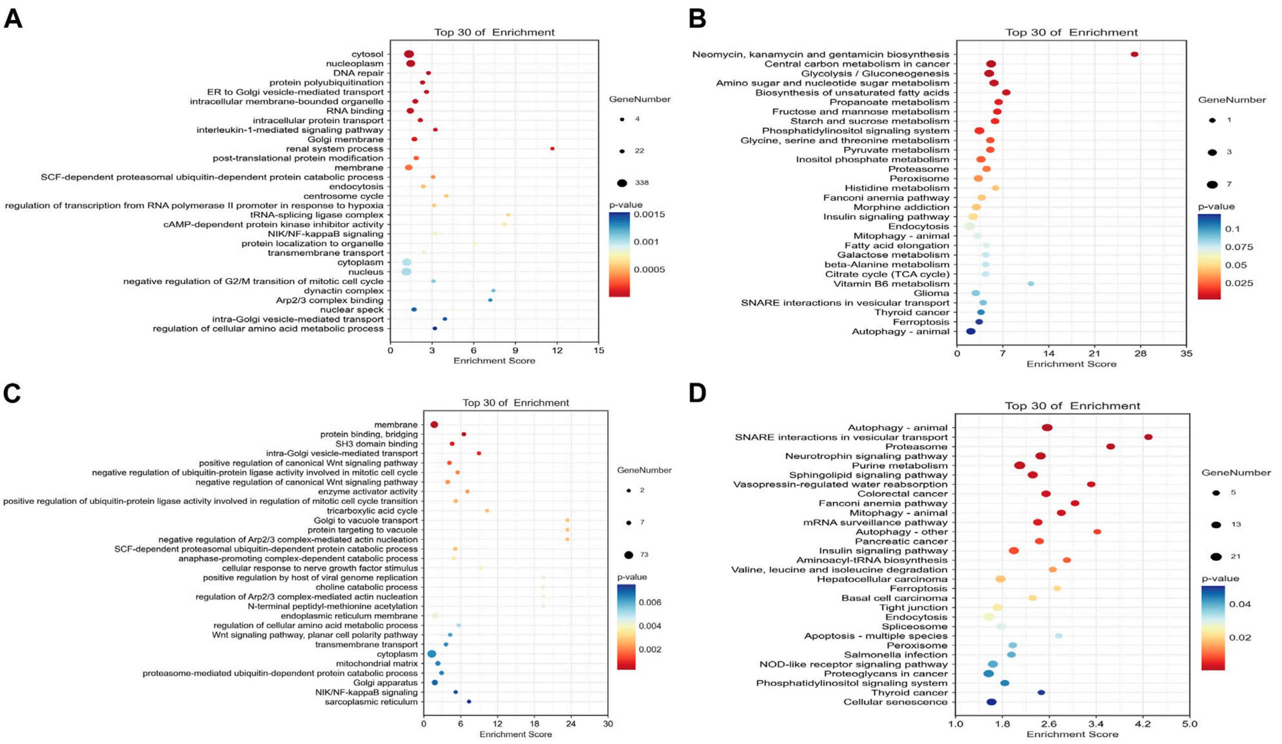


FIGURE 2
Dot plots of the top 30 mRNAs in the ceRNA network. **(A)** Significantly different pathways from GO analysis in the mRNA–miRNA–circRNA network. **(B)** Significantly different pathways from KEGG analyses in the mRNA–miRNA–circRNA network. **(C)** Significantly different pathways from GO analysis in the mRNA–miRNA–lncRNA network. **(D)** Significantly different pathways from KEGG analysis in the mRNA–miRNA–lncRNA network.

nucleation. KEGG analyses showed evident pathways including SNARE interactions in vesicular transport, ferroptosis, autophagy in animals, and apoptosis (Figures 2C, D).

3.4 Ingenuity pathway analysis

A graphical summary (Figure 3) showed that E2F1 regulation occupies a key position at 2 h after radiation, damage repair of DNA at 6 h accounts for the core, and autophagy occupies the core at 12–48 h in A549. In H1299, E2F1 regulation within 2–6 h after radiation occupies the core position, and cellular changes within 12–24 h are mainly related to metabolism; autophagy occupies the core position at 48 h.

The analysis of causal pathways (Figure 4A) shows evident pathways and significant changes caused by radiation including CREB signaling in neurons and synaptogenesis signaling pathway, cardiac hypertrophy signaling (enhanced), insulin secretion signaling pathway, G-protein coupled receptor signaling, hepatic fibrosis signaling pathway, and pulmonary fibrosis idiopathic signaling pathway.

Multi-time analysis of toxic pathways (Tox functions) is shown in Figure 4B. The results show that evident pathways are cell death of kidney cell lines, cell death of kidney cells, apoptosis of kidney cell lines, inflammation of the liver, increased activation of alkaline phosphatase, cell proliferation of kidney cell lines, proliferation of hepatic stellate cells, and apoptosis of hepatocytes, which suggests liver and kidney death, or the damage caused by radiation with 10 Gy is more evident.

Diseases/biological functions involving different genes/proteins over multiple periods are shown in Table 1. The top 10 pathways are infection of cells, transport of molecules, viral infection, migration of cells and cell movement, infection of tumor cell lines, metabolism of carbohydrate, synthesis of carbohydrate, infection by the RNA virus, and protein kinase cascade.

At 2 h after radiotherapy, the top canonical pathways in NSCLC mainly include the BAG2 signaling pathway, the FAT10 signaling pathway, and inhibition of ARE-mediated mRNAs. Top diseases and biofunctions mainly include cancer, organismal injury and abnormalities, endocrine system disorders, and gastrointestinal diseases. Molecular and cellular functions focus on DNA replication, recombination and repair, and cell death and survival (Supplementary Materials S2). At 6 h after radiotherapy, top canonical pathways in NSCLC mainly include the BAG2 signaling pathway and the FAT10 signaling pathway. The results of diseases and biofunctions and molecular and cellular functions are similar with those of NSCLC at 2 h (Supplementary Materials S3).

At 12 h after radiotherapy, top canonical pathways in NSCLC mainly include the CLEAR signaling pathway and melatonin signaling. Top canonical pathways in NSCLC include neurological diseases, compared with the results at 2 h and 6 h.

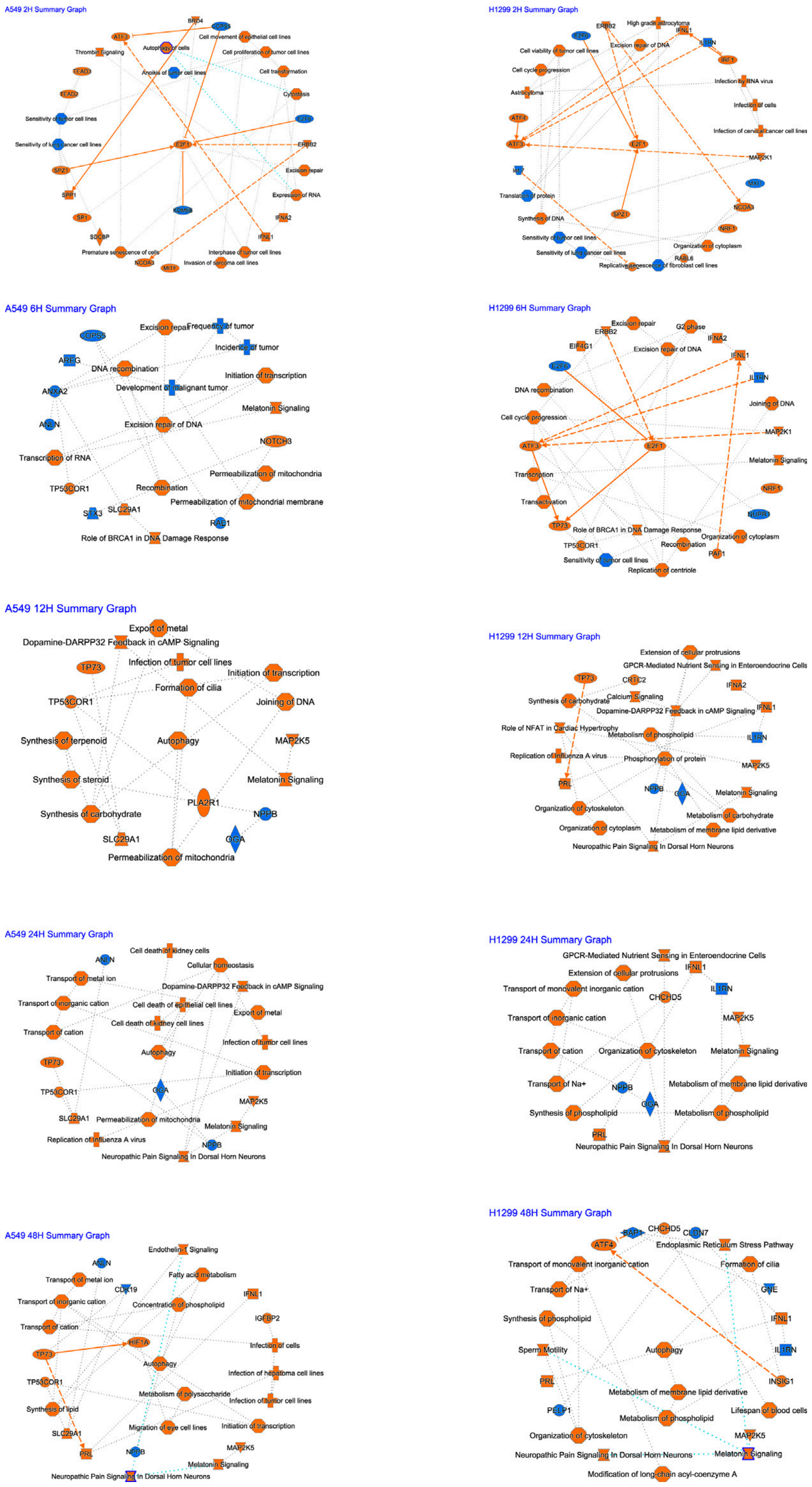

A549 2H Summary Graph
H1299 2H Summary Graph
A549 6H Summary Graph
Excision repair
Frequency of tumor
Incidence of tumor
DNA recombination
Initiation of transcription
Development of malignant tumor
Melatonin Signaling
Excision repair of DNA
NOTCH3
Transcription of RNA
Permeabilization of mitochondria
TP53COR1
Recombination
Permeabilization of mitochondrial membrane
SLC29A1
Role of BRCA1 in DNA Damage Response
H1299 6H Summary Graph
A549 12H Summary Graph
Export of metal
Dopamine-DARPP32 Feedback in cAMP Signaling
TP73
Infection of tumor cell lines
Initiation of transcription
Formation of cilia
TP53COR1
Joining of DNA
Synthesis of terpenoid
Autophagy
MAP2K5
Synthesis of steroid
Melatonin Signaling
Synthesis of carbohydrate
PLA2R1
NPPB
SLC29A1
GNA
Permeabilization of mitochondria
H1299 12H Summary Graph
A549 24H Summary Graph
H1299 24H Summary Graph
GPCR-Mediated Nutrient Sensing in Enteroendocrine Cells
Extension of cellular protrusions
IFNL1
Transport of monovalent inorganic cation
CHCHD5
IL1RN
Transport of inorganic cation
MAP2K5
Transport of cation
Organization of cytoskeleton
Melatonin Signaling
Transport of Na+
NPPB
Metabolism of membrane lipid derivative
GNA
Synthesis of phospholipid
Metabolism of phospholipid
PRL
Neuropathic Pain Signaling In Dorsal Horn Neurons
A549 48H Summary Graph
H1299 48H Summary Graph

FIGURE 3
Summary of IPA functions for radiotherapy of A549 and H1299 cells at five time points (2, 6, 12, 24, and 48 h).

Molecular and cellular functions focus on cellular assembly and organization, cell cycle, and carbohydrate metabolism (Supplementary Materials S4).

At 24 h after radiotherapy, top canonical pathways in NSCLC mainly include the super pathway of cholesterol biosynthesis, cholesterol biosynthesis I, and cholesterol biosynthesis II (*via* 24,25-dihydrolanosterol). Top canonical pathways are the same with the pathways at 12 h. Molecular and cellular functions mainly focus on the metabolism (Supplementary Materials S5).

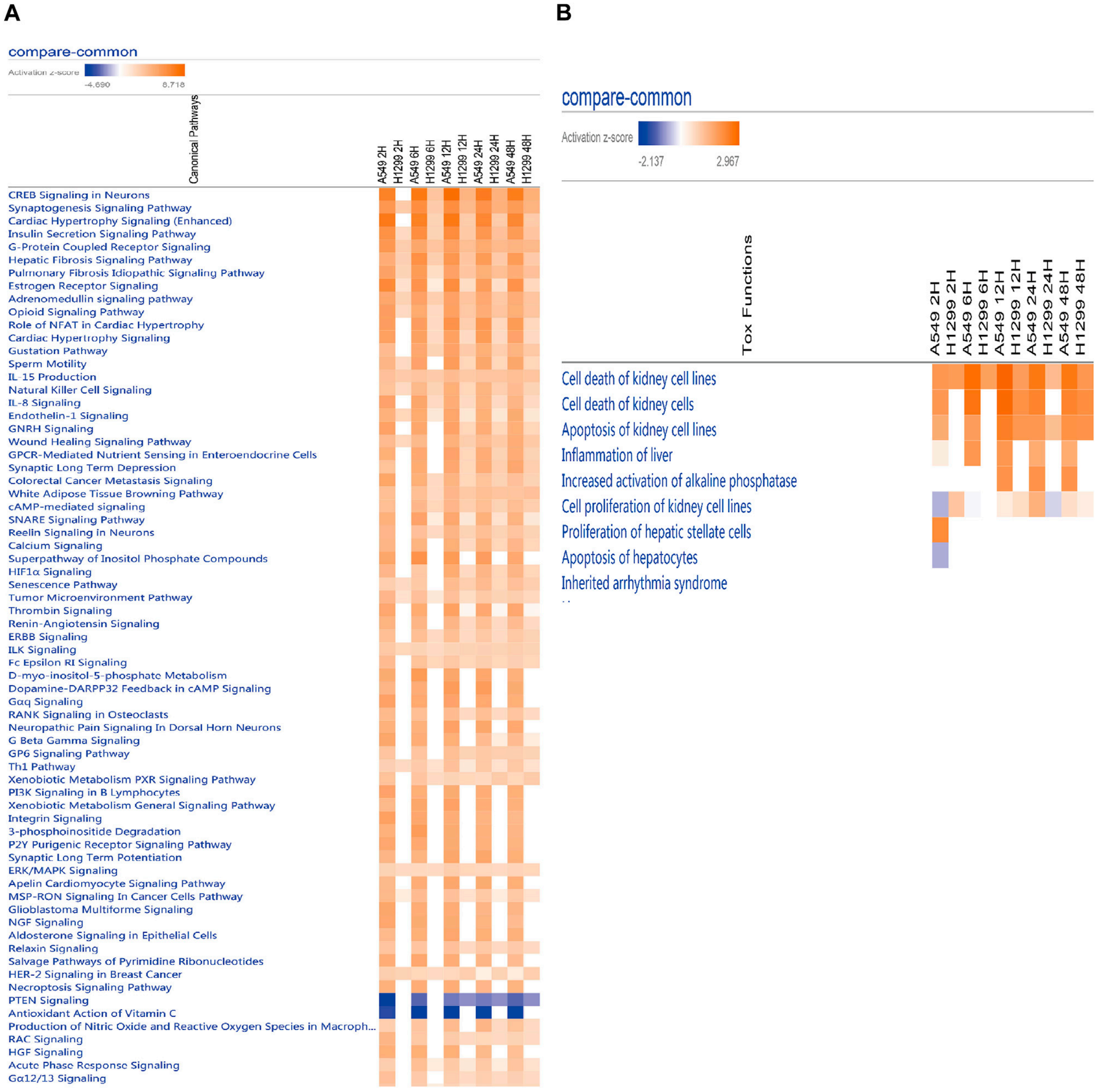


FIGURE 4
(A) Heatmap of the classical pathway trend of A549 and H1299 cells at five time points (2, 6, 12, 24, and 48 h) in radiotherapy predicted by IPA. (B) Heatmap of the toxicity pathway trend of A549 and H1299 cells at five time points (2, 6, 12, 24, and 48 h) in radiotherapy predicted by IPA.

At 48 h, top canonical pathways in NSCLC mainly include the CLEAR signaling pathway. Top canonical pathways are the same with the pathways at 12 and 24 h. Molecular and cellular functions mainly focus on the metabolism (Supplementary Materials S6).

3.5 Cell cycle analysis

The proportion of each cycle phase is shown in A, B, and C in Figure 5. Compared with the non-radiation group, the G2/M stage arrest of post-radiation NSCLC gradually worsened, peaking at 24 h, and decreased progressively at 48 h. Changes in cyclin B1 (CCNB1) showed a similar trend. The PCR results show a maximum value was reached at 24 h, see Figure 5D, and the WB results also show the expression of CCNB1 reached a maximum at 24 h (Figures 5E, F).

4 Discussion

Precision medicine is becoming a new direction for cancer treatment.

TABLE 1 Top 30 diseases and functions with the most significant changes at five time points (2, 6, 12, 24, and 48 h) after radiation in NSCLC.

Diseases and biofunctions	A549 2 h	H1299 2 h	A549 6 h	H1299 6 h	A549 12 h	H1299 12 h	A549 24 h	H1299 24 h	A549 48 h	H1299 48 h
Infection of cells	10.645	N/A	11.243	N/A	N/A	N/A	9.581	N/A	8.77	N/A
Transport of molecules	5.234	2.422	N/A	3.109	5.675	3.297	5.069	3.144	5.302	3.683
Viral infection	11.546	N/A	N/A	N/A	N/A	N/A	9.767	N/A	9.214	N/A
Migration of cells	N/A	4.655	N/A	3.543	N/A	5.121	N/A	5.455	N/A	5.818
Cell movement	N/A	4.822	N/A	N/A	N/A	5.644	N/A	5.904	N/A	6.204
Infection of tumor cell lines	N/A	N/A	N/A	N/A	7.873	N/A	7.721	N/A	6.915	N/A
Metabolism of carbohydrate	4.247	N/A	4.093	N/A	4.419	1.067	4.081	N/A	4.316	N/A
Synthesis of carbohydrate	4.292	N/A	4.141	N/A	4.464	N/A	4.129	N/A	4.362	N/A
Infection by the RNA virus	10.769	N/A	N/A	N/A	N/A	N/A	10.216	N/A	N/A	N/A
Protein kinase cascade	4.53	2	N/A	N/A	4.199	N/A	4.222	N/A	4.721	N/A
Fatty acid metabolism	N/A	2.923	N/A	3.088	N/A	3.508	N/A	3.508	3.36	3.232
Synthesis of lipids	N/A	1.387	N/A	2.402	4.985	2.013	4.047	N/A	4.555	N/A
Invasion of cells	N/A	3.745	N/A	N/A	N/A	5.101	N/A	4.972	N/A	5.376
Cell movement of tumor cell lines	N/A	4.09	N/A	N/A	N/A	4.446	N/A	4.549	N/A	4.982
Migration of tumor cell lines	N/A	4.356	N/A	N/A	N/A	4.159	N/A	4.266	N/A	4.859
Extracranial solid tumors	2.991	−1.342	2.104	−1.067	1.918	−1.633	2.217	−1.134	1.453	−1.195
Migration of endothelial cells	N/A	2.722	N/A	2.926	N/A	3.223	N/A	3.505	N/A	3.74
Oxidation of lipids	N/A	N/A	N/A	N/A	3.404	2.407	3.112	2.407	3.101	N/A
Invasion of tumor cell lines	N/A	3.732	N/A	N/A	N/A	5.013	N/A	N/A	N/A	5.315
Cell proliferation of tumor cell lines	6.878	N/A	N/A	N/A	N/A	2.362	N/A	2.064	N/A	2.314
Malignant solid tumors	1.787	−1.195	1.28	−1.698	1.295	−1.195	1.143	−1.195	1.058	−1.51
Autophagy	3.156	N/A	1.378	N/A	2.22	N/A	2.942	N/A	3.067	N/A
Metabolism of polyunsaturated fatty acids	N/A	2.582	N/A	2.433	N/A	2.582	N/A	2.582	N/A	2.582
Organization of the cytoplasm	N/A	N/A	N/A	N/A	3.967	N/A	3.911	N/A	3.949	N/A
Cellular homeostasis	3.896	N/A	N/A	N/A	N/A	N/A	3.96	N/A	3.62	N/A
Solid tumors	2.954	N/A	1.883	N/A	1.967	N/A	2.191	N/A	2.151	N/A
Cell death of tumor cell lines	−4.481	−1.916	N/A	N/A	N/A	−1.469	N/A	−1.442	N/A	−1.77
Metabolism of membrane lipid derivatives	2.564	N/A	N/A	N/A	2.947	N/A	2.641	N/A	2.918	N/A
Replication of Influenza A virus	N/A	N/A	N/A	N/A	N/A	N/A	5.534	N/A	5.439	N/A
Synthesis of polysaccharides	2.596	N/A	N/A	N/A	2.578	N/A	2.377	1	2.399	N/A

Personalized and precise management relies heavily on developing new technologies for next-generation sequencing and data processing of radiobiological information (Yang et al., 2020).

In this study, whole-transcriptome sequencing was used to comprehensively detect molecular changes of NSCLC in different periods after radiation, providing a dynamic molecular process map for precision radiotherapy.

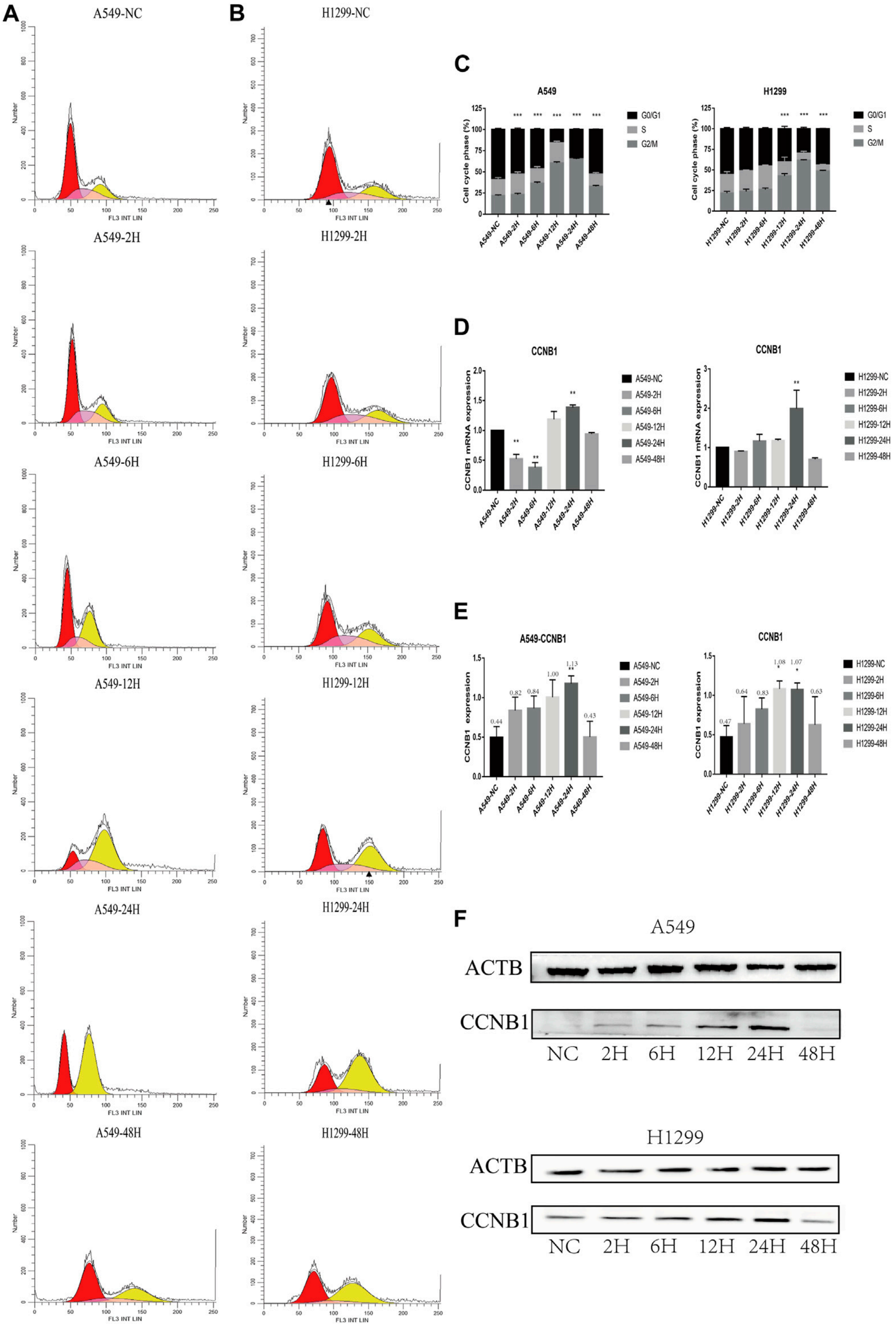


FIGURE 5
Cell cycle and cell cycle-related protein expression after radiation in NSCLC cells. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$. **(A)** Cell cycle analysis of A549 at 2, 6, 12, 24, and 48 h after radiation. **(B)** Cell cycle analysis of H1299 at 2, 6, 12, 24, and 48 h after radiation. **(C)** Statistical analysis of different phases of the cell cycle at 2, 6, 12, 24, and 48 h after radiation in A549 and H1299 cell lines. **(D)** PCR statistical results of CCNB1 at 2, 6, 12, 24, and 48 h after radiation in A549 and H1299 cell lines. **(E)** Western blot statistical results of cyclin B1 at 2, 6, 12, 24, and 48 h after radiation in A549 and H1299 cell lines (quantitative data are shown as means). **(F)** Western blot results of cyclin B1 at 2, 6, 12, 24, and 48 h after radiation in A549 and H1299 cell lines, compared with the non-radiation control group.

Data from short time-series expressions can be analyzed using two methods. The first employs methods that do not take advantage of the sequential information in time-series data. The second method was primarily designed for a longer time series, ignoring the temporal dependency among successive time points. The Short Time-series Expression Miner was designed for short time-series microarray gene expression data. It also has the advantage of visualization capabilities and integration with GO (Ernst and Bar-Joseph, 2006).

Time trend analysis obtained significant dynamic changes in mRNA, miRNA, lncRNA, and circRNA gene sets. According to the ceRNA analysis of RNAs related to time, we found the main regulatory networks and key molecules of post-radiation in NSCLC. These can provide new ideas for post-radiation molecular regulation mechanism research and seeking to target molecular therapies for NSCLC. For example, in the ceRNA network, miR-219-1-3p, which occupies the core, negatively regulates MUC4 and has a tumor-suppressive effect in pancreatic cancer (Chae et al., 2017). Related studies have found that miR-219-1-3p inhibits proliferation and weakens cell migration (Lahdaoui et al., 2015). MiR-221-3p downregulates the proto-oncogene MDM2, reversing paclitaxel resistance in non-small cell carcinoma and inducing apoptosis (Ni et al., 2021).

The results of GO and KEGG include DNA repair, negative regulation of G2/M transition of mitotic cell cycle, regulation of the autophagosome assembly, DNA replication, autophagy, ferroptosis, apoptosis, glucose metabolism, and insulin pathways. It broadens the content of radiobiology and the study of intersecting fields, providing new insights for combining radiation and drugs to improve radiotherapy efficacy.

IPA implies NSCLC cells started DNA damage and repair mainly in the early phase (2–6 h) after radiation, and E2F1 may play an important role in this early response phase. The cells started autophagy mainly in the later stages (24–48 h). These findings significantly enrich the content of radiobiology at various periods and help us get the key molecular or pathway or function to respond to radiotherapy at a specific time slot. Additionally, the molecules we are familiar with may regulate other pathways under radiotherapy conditions, which open up our perspective of molecular biology. For example, it is acknowledged that E2F1 is related to the cell cycle (Schuldt, 2011). In recent years, RB/E2F1 has been the main regulator of cancer cell metabolism in advanced diseases. It promotes the synthesis of antioxidant glutathione after RB loss, regulates redox metabolism, and reveals the protective effect of therapeutic intervention on reactive oxygen species (Mandigo et al., 2021). E2F1 may also be associated with the metabolism after radiotherapy by combining IPA, GO, and KEGG results, but it needs to be verified experimentally.

Toxic pathways after radiation mainly focus on hepatic and renal pathways. The in-depth understanding of the molecular and pathophysiology of radiation organs needs further study (Wang and Tepper, 2021).

The G2-phase arrest plays a role in cell survival after irradiation (Hwang and Muschel, 1998). Cells at this stage are sensitive to radiation therapy. Some studies discuss the potential use of G2/M cell cycle checkpoint inhibitors to enhance tumor control rates (Hellmann and Rhomberg, 1991; Löbrich and Jeggo, 2007; Dillon et al., 2014). Our results suggest that 24 h is proper for radiation therapy to maximize the effect of killing tumor cells. Some studies showed that A549 under dose 2 Gy at G2 / M phase arrest the most at 72 h (Yang et al., 2015). Our results suggest that 24 h may be best for radiation therapy in larger doses (10 Gy), guiding the practice of clinical radiation, combination chemotherapeutic drugs, and radiotherapy sensitizers.

There are few papers that compare the changes in transcriptome induced by low-dose radiation with those induced by high-dose SBRT radiation. Research about chronic low-dose radiation exposure in a zebrafish model found that radiation exposure resulted in transcriptomic perturbations in wound healing, immune response, lipid metabolism and absorption, and fibrogenic pathways (Cahill et al., 2023). Genomic and transcriptomic results of SBRT showed that in patients with renal cell carcinoma, pathways including G2/M checkpoint, mitotic spindle, and E2F targets were significant (Zengin et al., 2023). These results are consistent with our results.

Tumor treating fields (TTFields) is a new modality of cancer treatment. The treatment is based on transdermally transmitting alternating current (AC) electric fields at 100–400 kHz to tumors with two orthogonal transducer arrays (Moser et al., 2022). It can cause DNA damage and replication stress (Karanam et al., 2020). Our results can be combined with those of tumor treating fields to provide a biological basis for the timing of tumor treating fields after SBRT for non-small cell lung cancer. Compared with a low dose, our results would provide more economical ways to apply to the TTF. At the same time, our research further screens and models the common non-small cell lung cancer genes, which can achieve individualized treatment for patients with high matching genes with our gene set.

There were some flaws in the experiment. Our selection of genes common to non-small cell lung cancer needs to be verified. The genes and networks that change in each period need to be further explored. We did not perform animal experiments and lacked clinical samples of radiation therapy to verify whether the results we found were related to radiation. Further research is needed in the future.

5 Conclusion

Our transcriptomic and experimental analyses provide the dynamic change of radiation therapy in NSCLC, enriching the content of radiobiology in precision radiation oncology.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://www.ncbi.nlm.nih.gov/ [Accession number PRJNA916274]. Due to disk damage, data of files (A549_2h2.R1.fastq.gz A549_2h2.R2.fastq.gz. H1299_NC2.R1.fastq.gz H1299_NC2.R2.fastq.gz A549_48h2.R1.fastq.gz, A549_NC1.R1.fastq.gz A549_NC1.R2.fastq.gz H1299_2h2.R1.fastq.gz) are unavailable. Further enquiries can be directed to the corresponding author.

Author contributions

Conception, design, data curation, and formal analysis: YD, YZ, CZ, and TQ. Funding acquisition: TQ. Writing—original draft: YD and KY. Writing—review and editing: YD and CZ. Final approval of the manuscript and submission. All authors read and approved the final manuscript.

Funding

This work was supported by the Key Subject Construction Program of the Shanghai Health Administrative Authority (ZK 2019B30).

Acknowledgments

The authors thank Yawen Song (OE Biotech Co., Ltd.) for help in uploading the data. They are also thankful to Danting Zhang and Haoyi Chen (Shanghai Jiao Tong University, China) for help in IPA; thankful for Guoxiong Xu's (Center for Tumor Diagnosis and Therapy, Jinshan Hospital, Fudan University, Shanghai 201508, China) guidance in writing the original draft; thankful to Muyang He (Shanghai Medical College, Fudan University, Shanghai 200032, China) for help in references; thankful to Wenbin Kong, Rong Yang, Jianzhong Yang, Huizhong He, Zhicheng Yang, Wei Fan, Xuan Li, Sujuan Yuan, Wei Chen, and Xiaoqun Liu (Center for Tumor Diagnosis and Therapy, Jinshan Hospital, Fudan University, Shanghai 201508, China) for help in radiation therapy and laboratory management; and thankful to Yiwei Huang, Hao Niu, Ming Li, Yuansheng Zheng, Zhengyang Hu, Xing Jin, Mengnan Zhao, Jiaqi Liang, Guoshu Bi, Yunyi Bian, Zhencong Chen, Tao Lu, and Qihai Sui (Zhongshan Hospital, Fudan University, Shanghai 200032, China) for help in data analysis and laboratory management.

Conflict of interest

Author YZ was employed by OE Biotech Co., Ltd.

The remaining 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/fgene.2023.1126236/full#supplementary-material

SUPPLEMENT MATERIALS S1
Flowchart of genome-wide analysis of human non-small cell lung cancer after radiation

SUPPLEMENT MATERIALS S2
IPA results of A549 and H1299 after radiation at 2 h

SUPPLEMENT MATERIALS S3
IPA results of A549 and H1299 after radiation at 6 h

SUPPLEMENT MATERIALS S4
IPA results of A549 and H1299 after radiation at 12 h

SUPPLEMENT MATERIALS S5
IPA results of A549 and H1299 after radiation at 24 h

SUPPLEMENT MATERIALS S6
IPA results of A549 and H1299 after radiation at 48 h

References

Burkoň, P., Trna, J., Slávik, M., Němeček, R., Kazda, T., Pospíšil, P., et al. (2022). Stereotactic body radiotherapy (SBRT) of pancreatic cancer-A critical review and practical consideration. *Biomedicines* 10 (10), 2480. doi:10.3390/biomedicines10102480

Cahill, T., da Silveira, W. A., Renaud, L., Wang, H., Williamson, T., Chung, D., et al. (2023). Investigating the effects of chronic low-dose radiation exposure in the liver of a hypothermic zebrafish model. *Sci. Rep.* 13 (1), 918. doi:10.1038/s41598-022-26976-4

Chae, Y. K., Pan, A. P., Davis, A. A., Patel, S. P., Carneiro, B. A., Kurzrock, R., et al. (2017). Path toward precision oncology: Review of targeted therapy studies and tools to aid in defining "actionability" of a molecular lesion and patient management support. *Mol. Cancer Ther.* 16 (12), 2645–2655. doi:10.1158/1535-7163.MCT-17-0597

Chairmadurai, A., Jain, S. K., Jain, A., and Prakash, H. (2022). Rapid Arc-SBRT: Non-Invasive immune adjuvant for advanced stage non-small cell lung carcinoma. *Anticancer Agents Med. Chem.* 22 (2), 202–205. doi:10.2174/1871520621666210322105641

Chmura, S., Winter, K. A., Robinson, C., Pisansky, T. M., Borges, V., Al-Hallaq, H., et al. (2021). Evaluation of safety of stereotactic body radiotherapy for the treatment of patients with multiple metastases: Findings from the NRG-br001 phase 1 trial. *JAMA Oncol.* 7 (6), 845–852. doi:10.1001/jamaoncol.2021.0687

De Ruysscher, D., Niedermann, G., Burnet, N. G., Siva, S., Lee, A. W. M., and Hegi-Johnson, F. (2019). Radiotherapy toxicity. *Nat. Rev. Dis. Prim.* 5 (1), 13. doi:10.1038/s41572-019-0064-5

Dillon, M. T., Good, J. S., and Harrington, K. J. (2014). Selective targeting of the G2/M cell cycle checkpoint to improve the therapeutic index of radiotherapy. *Clin. Oncol. R. Coll. Radiol.* 26 (5), 257–265. doi:10.1016/j.clon.2014.01.009

Dlp, C., Latorre, R. G., and Fuentes, R. (2022). SBRT in localized renal carcinoma: A review of the literature. *Anticancer Res.* 42 (2), 667–674. doi:10.21873/anticanres.15525

Ernst, J., and Bar-Joseph, Z. (2006). Stem: A tool for the analysis of short time series gene expression data. *BMC Bioinforma.* 7, 191. doi:10.1186/1471-2105-7-191

Hellmann, K., and Rhomberg, W. (1991). Radiotherapeutic enhancement by razoxane. *Cancer Treat. Rev.* 18 (4), 225–240. doi:10.1016/0305-7372(91)90014-q

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. (2018). Artificial intelligence in radiology. *Nat. Rev. Cancer* 18 (8), 500–510. doi:10.1038/s41568-018-0016-5

Huynh, E., Hosny, A., Guthier, C., Bitterman, D. S., Petit, S. F., Haas-Kogan, D. A., et al. (2020). Artificial intelligence in radiation oncology. *Nat. Rev. Clin. Oncol.* 17 (12), 771–781. doi:10.1038/s41571-020-0417-8

Hwang, A., and Muschel, R. J. (1998). Radiation and the G2 phase of the cell cycle. *Radiat. Res.* 150 (5), S52–S59. doi:10.2307/3579808

Joseph, S. A., and Vijayakumar, S. (2020). Radiobiotherapy and radiobiomedicine-two novel paradigms in radiation medicine. *Int. J. Radiat. Oncol. Biol. Phys.* 108 (1), 326–327. doi:10.1016/j.ijrobp.2020.05.033

Karanam, N. K., Ding, L., Aroumougame, A., and Story, M. D. (2020). Tumor treating fields cause replication stress and interfere with DNA replication fork maintenance: Implications for cancer therapy. *Transl. Res.* 217, 33–46. doi:10.1016/j.trsl.2019.10.003

Lahdaoui, F., Delpu, Y., Vincent, A., Renaud, F., Messager, M., Duchêne, B., et al. (2015). miR-219-1-3p is a negative regulator of the mucin MUC4 expression and is a tumor suppressor in pancreatic cancer. *Oncogene* 34 (6), 780–788. doi:10.1038/onc.2014.11

Lo, S. S., Fakiris, A. J., Chang, E. L., Mayr, N. A., Wang, J. Z., Papiez, L., et al. (2010). Stereotactic body radiation therapy: A novel treatment modality. *Nat. Rev. Clin. Oncol.* 7 (1), 44–54. doi:10.1038/nrclinonc.2009.188

Löbrich, M., and Jeggo, P. A. (2007). The impact of a negligent G2/M checkpoint on genomic instability and cancer induction. *Nat. Rev. Cancer* 7 (11), 861–869. doi:10.1038/nrc2248

Luk, S. M. H., Ford, E. C., Phillips, M. H., and Kalet, A. M. (2022). Improving the quality of care in radiation oncology using artificial intelligence. *Clin. Oncol. R. Coll. Radiol.* 34 (2), 89–98. doi:10.1016/j.clon.2021.11.011

Mandigo, A. C., Yuan, W., Xu, K., Gallagher, P., Pang, A., Guan, Y. F., et al. (2021). RB/E2F1 as a master regulator of cancer cell metabolism in advanced disease. *Cancer Discov.* 11 (9), 2334–2353. doi:10.1158/2159-8290.CD-20-1114

Milic, M., Mondini, M., and Deutsch, E. (2022). How to improve SBRT outcomes in NSCLC: From pre-clinical modeling to successful clinical translation. *Cancers (Basel)* 14 (7), 1705. doi:10.3390/cancers14071705

Moser, J. C., Salvador, E., Deniz, K., Swanson, K., Tuszynski, J., Carlson, K. W., et al. (2022). The mechanisms of action of tumor treating fields. *Cancer Res.* 82 (20), 3650–3658. doi:10.1158/0008-5472.CAN-22-0887

Ni, L., Xu, J., Zhao, F., Dai, X., Tao, J., Pan, J., et al. (2021). MiR-221-3p-mediated downregulation of MDM2 reverses the paclitaxel resistance of non-small cell lung cancer *in vitro* and *in vivo*. *Eur. J. Pharmacol.* 899, 174054. doi:10.1016/j.ejphar.2021.174054

Schuldt, A. (2011). Cell cycle: E2F1 ensures the endocycle. *Nat. Rev. Mol. Cell Biol.* 12 (12), 768. doi:10.1038/nrm3232

Scott, J. G., Sedor, G., Ellsworth, P., Scarborough, J. A., Ahmed, K. A., Oliver, D. E., et al. (2021). Pan-cancer prediction of radiotherapy benefit using genomic-adjusted radiation dose (GARD): A cohort-based pooled analysis. *Lancet Oncol.* 22 (9), 1221–1229. doi:10.1016/S1470-2045(21)00347-8

Sidaway, P. (2022). SBRT feasible for oligometastatic RCC. *Nat. Rev. Clin. Oncol.* 19 (1), 6. doi:10.1038/s41571-021-00582-1

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2022). Cancer statistics, 2022. *CA Cancer J. Clin.* 72 (1), 7–33. doi:10.3322/caac.21708

Tadimalla, S., Wang, W., and Haworth, A. (2022). Role of functional MRI in liver SBRT: Current use and future directions. *Cancers (Basel).* 14 (23), 5860. doi:10.3390/cancers14235860

Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S. M., Ala, U., et al. (2011). Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell* 147 (2), 344–357. doi:10.1016/j.cell.2011.09.029

Teuwen, J., Gouw, Z. A. R., and Sonke, J. J. (2022). Artificial intelligence for image registration in radiation oncology. *Semin. Radiat. Oncol.* 32 (4), 330–342. doi:10.1016/j.semradonc.2022.06.003

Timmerman, R. D., Herman, J., and Cho, L. C. (2014). Emergence of stereotactic body radiation therapy and its impact on current and future clinical practice. *J. Clin. Oncol.* 32 (26), 2847–2854. doi:10.1200/JCO.2014.55.4675

Wang, K., and Tepper, J. E. (2021). Radiation therapy-associated toxicity: Etiology, management, and prevention. *CA Cancer J. Clin.* 71 (5), 437–454. doi:10.3322/caac.21689

Yang, S., Xu, J., Shao, W., Geng, C., Li, J., Guo, F., et al. (2015). Radiation-induced bystander effects in A549 cells exposed to 6 MV X-rays. *Cell Biochem. Biophys.* 72 (3), 877–882. doi:10.1007/s12013-015-0555-2

Yang, W. C., Hsu, F. M., and Yang, P. C. (2020). Precision radiotherapy for non-small cell lung cancer. *J. Biomed. Sci.* 27 (1), 82. doi:10.1186/s12929-020-00676-5

Zengin, Z. B., Govindarajan, A., Salgia, N., Sayegh, N., Tripathi, N., Muddasani, R., et al. (2023). Genomic and transcriptomic predictors of response from stereotactic body radiation therapy in patients with oligoprogressive renal cell carcinoma. *Eur. Urol. Oncol.* doi:10.1016/j.euo.2022.11.006

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 23 March 2023
DOI 10.3389/fgene.2023.1170737

OPEN ACCESS

EDITED BY
Tao Huang,
Shanghai Institute of Nutrition
and Health (CAS), China

REVIEWED BY
Hao Lin,
University of Electronic Science and
Technology of China, China
Lei Deng,
Central South University, China

*CORRESPONDENCE
Ying Deng,
✉ ldlrrr@163.com
Huadong Zhu,
✉ zhuhuadong1970@126.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to RNA,
a section of the journal
Frontiers in Genetics

RECEIVED 21 February 2023
ACCEPTED 14 March 2023
PUBLISHED 23 March 2023

CITATION
Song J, Ren K, Zhang D, Lv X, Sun L, Deng Y and Zhu H (2023), A novel signature combing cuproptosis- and ferroptosis-related genes in sepsis-induced cardiomyopathy.
Front. Genet. 14:1170737.
doi: 10.3389/fgene.2023.1170737

A novel signature combing cuproptosis- and ferroptosis-related genes in sepsis-induced cardiomyopathy

Juanjuan Song[1†], Kairui Ren[2†], Dexin Zhang[1], Xinpeng Lv[1], Lin Sun[1], Ying Deng[1]* and Huadong Zhu[2]*

[1]Department of Emergency, The Second Affiliated Hospital of Harbin Medical University, Harbin, China, [2]Department of Emergency, Peking Union Medical College Hospital, Chinese Academy of Medical Science, Beijing, China

Objective: Cardiac dysfunction caused by sepsis, usually termed sepsis-induced cardiomyopathy (SIC), is one of the most serious complications of sepsis, and ferroptosis can play a key role in this disease. In this study, we identified key cuproptosis- and ferroptosis-related genes involved in SIC and further explored drug candidates for the treatment of SIC.

Methods: The GSE79962 gene expression profile of SIC patients was downloaded from the Gene Expression Omnibus database (GEO). The data was used to identify differentially expressed genes (DEGs) and to perform weighted correlation network analysis (WGCNA). Furthermore, Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were conducted. Then, gene set enrichment analysis (GSEA) was applied to further analyze pathway regulation, with an adjusted p-value <0.05 and a false discovery rate (FDR) <0.25. Ferroptosis-related genes were obtained from the FerrDb V2 database, and cuproptosis-related genes were obtained from the literature. We constructed a novel signature (CRF) by combing cuproptosis-related genes with ferroptosis-related genes using the STRING website. The SIC hub genes were obtained by overlapping DEGs, WGCNA-based hub genes and CRF genes, and receiver operating characteristic (ROC) curve analysis was used to determine the diagnostic value of hub genes. A transcription factor-microRNA-hub gene network was also constructed based on the miRnet database. Finally, potential therapeutic compounds for SIC were predicted based on the Drug Gene Interaction Database.

Results: We identified 173 DEGs in SIC patients. Four hub modules and 411 hub genes were identified by WGCNA. A total of 144 genes were found in the CRF. Then, POR, SLC7A5 and STAT3 were identified as intersecting hub genes and their diagnostic values were confirmed with ROC curves. Drug screening identified 15 candidates for SIC treatment.

Conclusion: We revealed that the cuproptosis- and ferroptosis-related genes, POR, SLC7A5 and STAT3, were significantly correlated with SIC and we also predicted therapeutic drugs for these targets. The findings from this study will make contributions to the development of treatments for SIC.

KEYWORDS

cuproptosis, ferroptosis, sepsis-induced cardiomyopathy, signature, candidates

Introduction

Sepsis is a dysregulated host response to infection that can cause life-threatening organ dysfunction (Singer et al., 2016). It is a leading cause of mortality and critical illness worldwide. A recent study estimated that the number of sepsis cases and deaths is twice as high as previously thought (Rudd et al., 2020). Sepsis-induced cardiomyopathy (SIC) is a common and well-elucidated complication of sepsis and is associated with higher mortality rates in patients with sepsis (Hanumanthu et al., 2021). Myocardial dysfunction is characterized by cellular abnormalities, circulating mediators and instrumental parameters. However, the lack of a consensus definition and uncertainties of the pathophysiology of SIC make it difficult to identify and validate biomarkers of the disease. In Addition, the cytokine storm also makes it difficult to identify cytokine biomarkers of SIC. Bioinformatic analyses have the potential to decipher these complex signals. Tumor necrosis factor, Jak-signal transducer and activator of transcription (STAT), and hypoxia-inducible transcription factor-1, and their interactions are increasingly recognized as main factors in sepsis cardiomyopathy (Chen et al., 2020).

Ferroptosis is a newly identified iron-dependent form of cell death that is different from other forms of cell death (Yan et al., 2021) and is involved in the development of cardiomyopathy. Downregulating HO-1 expression and iron concentration can reduce ferroptosis, thereby attenuating myocardial cell injury in sepsis (Wang et al., 2020). Ferritinophagy-mediated ferroptosis is a critical mechanism contributing to sepsis-induced cardiac injury (Ning et al., 2020) and targeting ferroptosis in cardiomyocytes may be a therapeutic strategy for preventing sepsis. We therefore aimed to use bioinformatic technology to quickly identify ferroptosis-related genes in SIC. This information can then be used for the early diagnosis of SIC and the development of new treatments of the disease.

Similar to iron, copper is also an essential micronutrient. Cells exhibit cytotoxicity when the intracellular concentration of copper ions exceeds the homeostatic threshold. Copper induces cell death by targeting lipoylated TCA cycle proteins. This leads to the aggregation of fatty acylated proteins and the loss of iron-sulfur cluster proteins, which in turn triggers proteotoxic stress and ultimately cell death (Tsvetkov et al., 2022). Cuproptosis is associated with various disease conditions, including Wilson's disease, neurodegenerative diseases, cancer (Li et al., 2022) and heart diseases (Chen et al., 2022). Copper levels are also closely related to the morbidity and mortality of cardiovascular diseases (Cui et al., 2022; Liu and Miao, 2022).

Cuproptosis- and ferroptosis-related regulatory mechanisms are expected to be novel targets for SIC treatment. However, whether cuproptosis-related genes combined with ferroptosis-related genes can be used for diagnosis and to predict responses to immunotherapy and drug sensitivity in SIC have not been addressed. This study analyzed the difference in gene expression between non-failing hearts and SIC hearts in GEO database, and also performed WGCNA. Then, the function of DEGs was determined using GO and KEGG analysis. In addition, GSEA was applied to further analyze pathway regulation. In addition, we constructed a novel signature (CRF) by combing cuproptosis-related genes with ferroptosis-related genes using the STRING website for predicting diagnosis. Besides, SIC hub genes were obtained by overlapping DEGs, WGCNA-based hub genes and CRF genes, and receiver operating characteristic (ROC) curve analysis was used to determine the diagnostic value of hub genes. A transcription factor-microRNA-hub gene network was also constructed based on the miRnet database. Finally, potential therapeutic compounds for SIC were predicted based on the Drug Gene Interaction Database, which provided a theoretical basis for clinical treatment of SIC.

Materials and methods

Data resource

The GSE79962 gene expression profile was downloaded from the GEO database (http://www.ncbi.nih.gov/geo) using the GEOquery package of R software (version 4.2.1). The chip platform for GSE79962 was GPL6244 (Affymetrix Human Gene 1.0 ST Array), which consists of 20 SIC human heart tissue samples and 11 healthy human heart tissue samples. The ferroptosis-related genes were obtained from the GeneCards database (https://www.genecards.org/). All data are publicly available.

Identification of differentially expressed genes

Raw data were downloaded as MINiML files from the Gene Expression Omnibus (GEO) database. Probes were converted to gene symbols according to the platform annotation information of the normalized data. Probes with more than one gene were eliminated, and the average of genes corresponding to more than one probe was calculated. The limma package in R software (version 4.2.1) was used to study differentially expressed genes. The adjusted *p*-value was determined to correct the false positive results in the GEO datasets. "Adjusted $p < 0.05$ and log(fold change) > 1 or log(fold change) < −1" were defined as the threshold for the differential expression of genes. The data for the listed DEGs were processed, and heatmaps and volcano plots were drawn using ComplexHeatmap and ggplot2 R packages.

Functional and pathway enrichment analysis

To further confirm the underlying function of potential targets, the data were analyzed by functional enrichment. Gene Ontology (GO) is a widely used tool for annotating genes with functions, especially molecular function (MF), biological pathways (BP), and cellular components (CC). Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis is a practical resource for studying gene functions and associated high-level genome functional information. To better understand mRNA involvement in pathogenesis, ClusterProfiler package (version: 3.18.0) in R was employed to analyze the GO function of potential targets and enrichment in the KEGG pathway. The R software package, pheatmap, was used to draw heatmaps. Then, gene set enrichment analysis (GSEA) was applied for further analysis of pathway regulation. We used KEGG rest API (https://www.kegg.jp/kegg/rest/keggapi.html) to obtain the latest gene annotation. Enrichment analysis was performed using the R package, clusterProfiler (version 3.14.3). In this analysis, the minimum gene set was 5 and the maximum gene set was 5,000. A p-value <0.05, false discovery rate (FDR) <0.25 and |NES| >1 indicated a significantly enriched term.

Weighted gene co-expression network analysis (WGCNA)

Human SIC heart and healthy heart tissue samples from the GSE79962 dataset were analyzed by the WGCNA R 1.70-3 package. First, Pearson's correlation matrices and the average linkage method were both performed for all pair-wise genes. Then, a weighted adjacency matrix was constructed using a power function, β was a soft-thresholding parameter that could emphasize strong correlations between genes and penalize weak correlations. Meanwhile, a soft threshold was reasonably selected as the degree of scale independence reached 0.8. After choosing the power of $\beta = 5$, the adjacency was transformed into a topological overlap matrix (TOM), which could measure the network connectivity of a gene defined as the sum of its adjacency with all other genes for network Gene ration, and the corresponding dissimilarity (1-TOM) was calculated. To classify genes with similar expression profiles into gene modules, average linkage hierarchical clustering was conducted according to the TOM-based dissimilarity measure with a minimum size (gene group) of 30 for the genes dendrogram. The modules that correlated the most with the clinical traits were identified as SIC-related modules. All functions of hub genes with gene significance (GS) >0.2 and module membership (MM) >0.8 were analyzed by GO enrichment.

Construction and validation of a cuproptosis- and ferroptosis-related gene signature

First, we obtained ferroptosis-related genes were from the FerrDb V2 database (http://www.zhounan.org/ferrdb/current/). Then, cuproptosis-related genes were obtained from the literature. The obtained cuproptosis-related and ferroptosis-related genes were inputed into the STRING website (https://string-db.org/), and the minimum required interaction score was set to 0.9 to obtain iron genes related to copper genes. Thus, we constructed a novel signature (CRF) by combing cuproptosis-related genes with ferroptosis-related genes.

Intersection genes and venn analysis

A Venn diagram drawing tool (http://bioinformatics.psb.ugent.be/webtools/Venn/) was used to generate Venn diagrams of DEGs, WGCNA-based hub genes and CRF genes. Intersection genes were included in subsequent analyses.

Identification of hub genes based on receiver operating characteristic (ROC) curve analysis

The diagnostic values of intersection genes for SIC were detected by ROC curve and area under the ROC curve (AUC) analysis using the pROC R 1.17.0.1 package.

Construction of a transcription factor (TF)-microRNA-hub gene network

microRNAs (miRNAs) and TFs related to intersection genes were screened for based on the miRNet2/0 online database (https://www.mirnet.ca/). TFs and miRNAs related to intersection genes were identified and added to the network using Cytoscape software (version 3.8.2).

Screening the drug-gene interaction database (DGIdb) for potential therapeutic drugs for SIC

DGIdb (https://www.dgidb.org/) was used as a drug–gene interaction database to screen for drug–gene interactions and information from papers, databases, and web resources. Therapeutic drugs for intersection genes were identified based on the DGIdb.

Results

Differential gene expression analysis

The gene expression dataset, GSE79962, contained data from 20 SIC samples, and 11 normal myocardial tissue samples. As shown in Figure 1A, data normalization and cross comparability were assessed. Using the limma package in R software (version 4.2.1) for differential expression analysis, with adjusted $p < 0.05$ and |log2 FC| >1 as filtering conditions, we found that 173 genes were differentially expressed in the myocardial tissue of patients with SIC cardiomyopathy compared with normal myocardial tissue. Sixty-seven DEGs were upregulated and 106 were downregulated. The ferroptosis-related genes (NOX4, HMOX1, POR, SAT1, etc.) were highly expressed in sepsis-induced cardiomyopathy model. Recent studies showed that NOX4 was characterized in the cardiovascular system, HMOX1 can be induced by sepsis, and POR was selected as the central gene of SIC. The above results are consistent with our research. Clustering analysis of these DEGs was performed, as shown in a volcano plot (Figure 1B). The heatmap for the dataset indicated better clustering of samples and higher confidence (Figure 1C).

Functional pathway enrichment analysis of DEGs in SIC

We performed KEGG and GO enrichment analyses on the up and downregulated DEGs in the GSE79962 dataset. The results showed that there were significant differences among the functions of DEGs. The upregulated DEGs were mainly enriched in pathways of muscle tissue development, regulation of small molecule metabolic process, regulation of actin cytoskeleton, leukocyte transendothelial migration, and AMPK signaling pathway (Figures 2A, B). The downregulated DEGs were mainly enriched in pathways of response to interleukin 1, regulation of peptidase activity, positive regulation of cytokine production, neutrophil activation involved in immune response, MAPK signaling

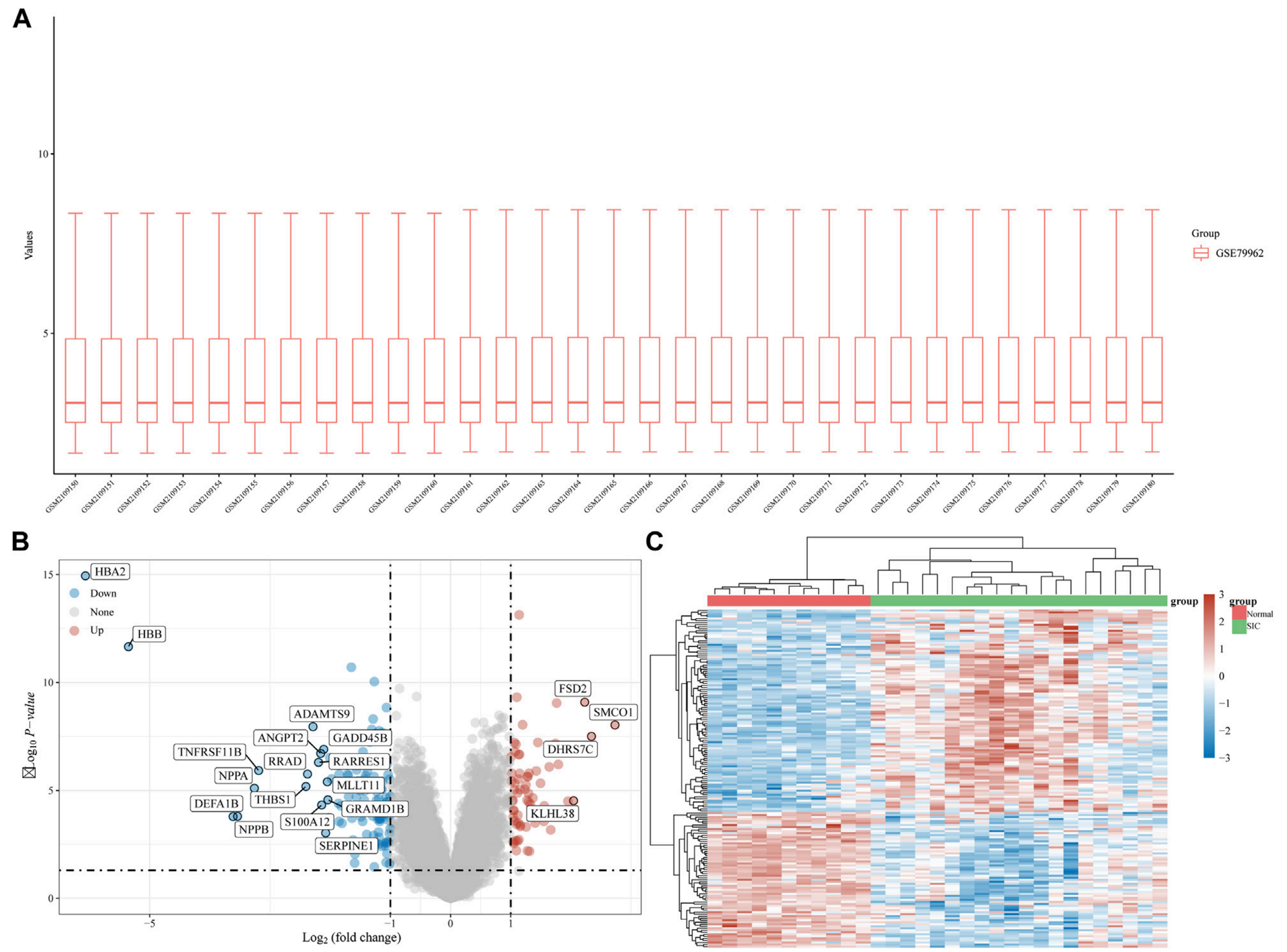


FIGURE 1
DEG analysis. **(A)** Boxplot diagram of the DEGs in the GSE79962 dataset. **(B)** Volcano plot of the DEGs in the GSE79962 dataset. **(C)** Heatmap of the DEGs in the GSE79962 dataset.

pathway, JAK-STAT signaling pathway, and cytokine-cytokine receptor interaction (Figures 2C, D). As we all know, AMPK has a strong regulatory effect on cellular energy balance, metabolic homeostasis, inflammatory response, oxidative stress and myocardial cell survival, and is closely related to the pathogenesis of septic cardiomyopathy. MAPK signaling pathway and JAK/STAT signaling is an important pathway for the signal transduction of several key cytokines in the pathogenesis of sepsis, which can transcribe and modulate the host immune response. p38-MAPK in the MAPK family is involved in SIC signaling and apoptosis mechanism. Application of clinically used JAK/STAT inhibitors, tofacitinib and baricitinib, fully prevented IFNγ-induced cardiomyopathy, confirming the critical roles of this signaling pathway in inflammatory cardiac disease. Our results are consistent with previous studies.

GSEA showed that compared with control samples, the identified KEGG pathways were Huntington's disease, Alzheimer's disease, oxidative phosphorylation, citrate cycle TCA cycle, Parkinson's disease, cardiac muscle contraction, valine leucine and isoleucine degradation, fatty acid metabolism and peroxisome (Figure 3A). The identified hallmark gene sets were oxidative phosphorylation, fatty acid metabolism, adipogenesis, UV response up, estrogen response early, apoptosis, androgen response, hypoxia, inflammatory response, estrogen response late, IL6-JAK-STAT3 signaling, P53 pathway, IL2-STAT5 signaling, unfolded protein response, TNFα signaling *via* NF-κB, and TGFβ signaling (Figure 3B). The human phenotype ontologies identified were arrhythmia and mitochondrion (Figures 3C, D).

Hub modules and genes identified by WGCNA

A total of 22,828 genes were derived from the 31 samples of the GSE79962 dataset. These genes were used to construct a co-expression network. The cluster analysis results of the samples are shown in Figure 4. Clustering trees for each dataset were established and no outliers were found (Figure 4A). Soft threshold was reasonably selected as the degree of scale independence reached 0.8. The scale-free fit index and mean connectivity were calculated and the power of $\beta = 5$ (scale free R2 = 0.87) was selected (Figure 4B). The minimum number of genes per

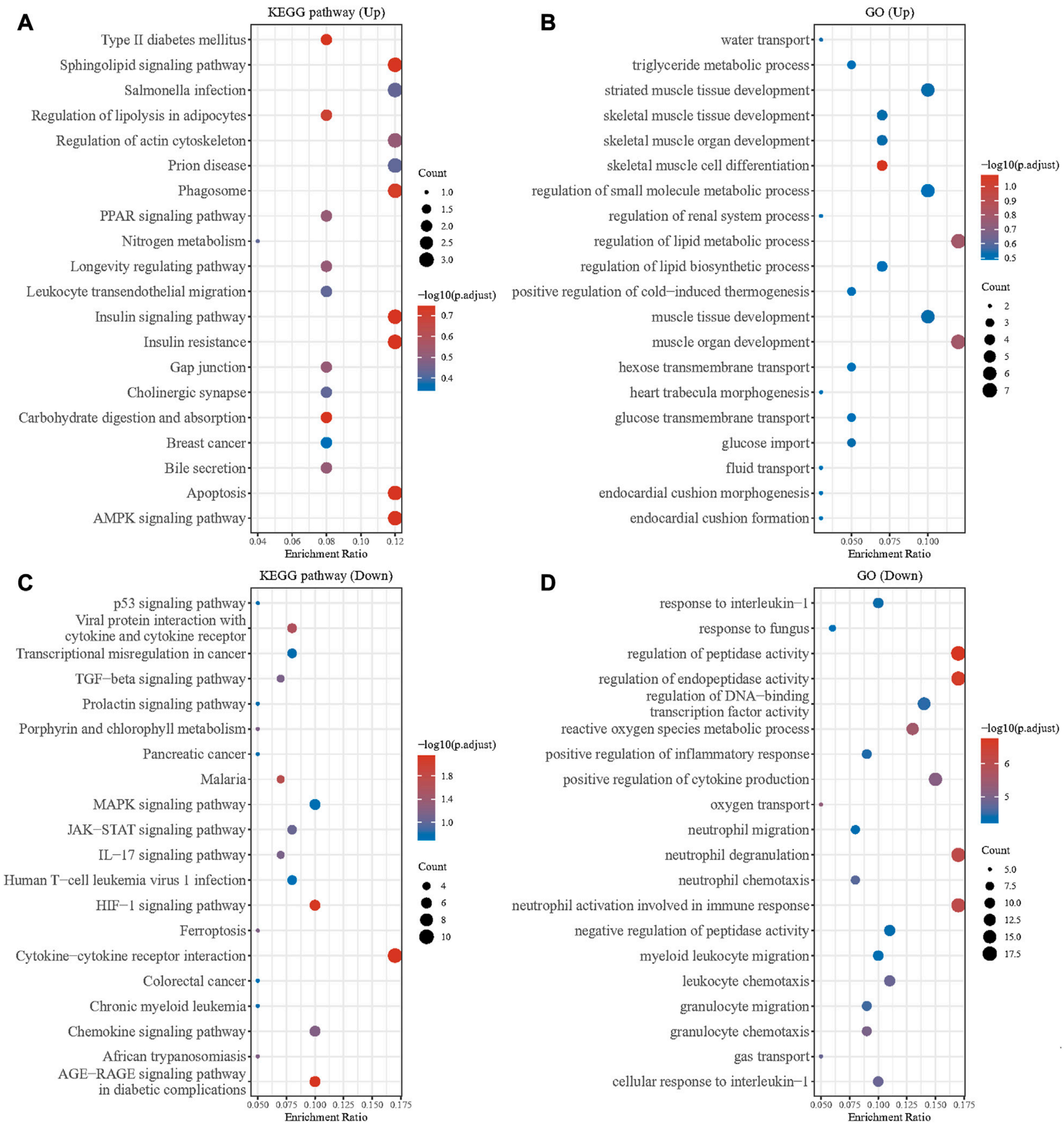


FIGURE 2
GO and KEGG enrichment analysis of DEGs. **(A)** GO enrichment analysis of the upregulated DEGs. **(B)** KEGG enrichment analysis of the upregulated DEGs. **(C)** GO enrichment analysis of the downregulated DEGs. **(D)** KEGG enrichment analysis of the downregulated DEGs.

module was set to 30 according to the criteria of the dynamic tree-cutting algorithm. The final 37 transcriptional modules represented by different colors were identified (Figure 4C). The adjacencies of modules in the network are shown in Figure 4D. To correlate the modules with sample information, we analyzed the data according to the heatmap of module-clinical trait correlations, thereby correlating data for the clinical traits (Figure 4E). The black, floral white, magenta, and pale violet red 3 modules, which were identified as the hub modules associated with clinical traits, were used to explore the correlation between module membership (MM) and gene significance (GS) to identify the hub genes in SIC (Figures 4F–I). Furthermore, we demonstrated 411 hub genes were respectively identified from the above four modules with MM >0.8 and GS >0.2.

Selection of intersection hub genes and their functions in SIC

Based on the FerrDb V2 database, we obtained 612 ferroptosis-related genes. We also identified

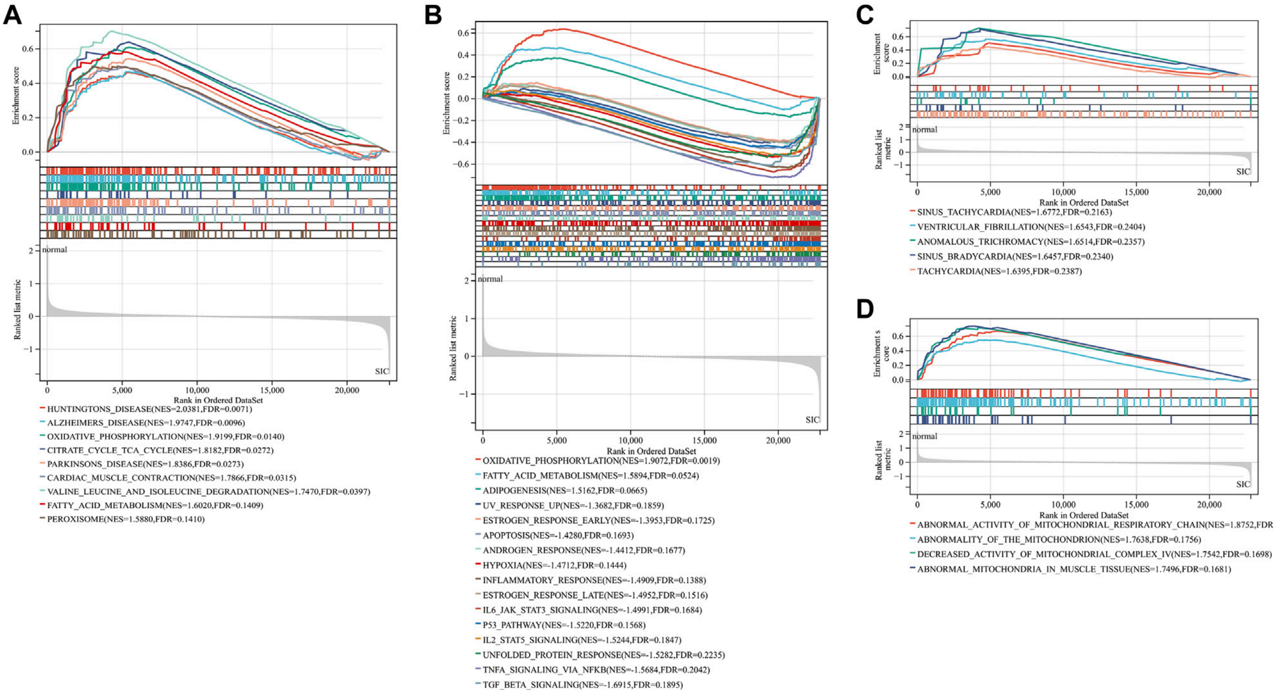


FIGURE 3
Gene set enrichment analysis (GSEA) for GSE79962. **(A)** KEGG **(B)** hallmark gene sets **(C)** arrhythmia **(D)** mitochondrion.

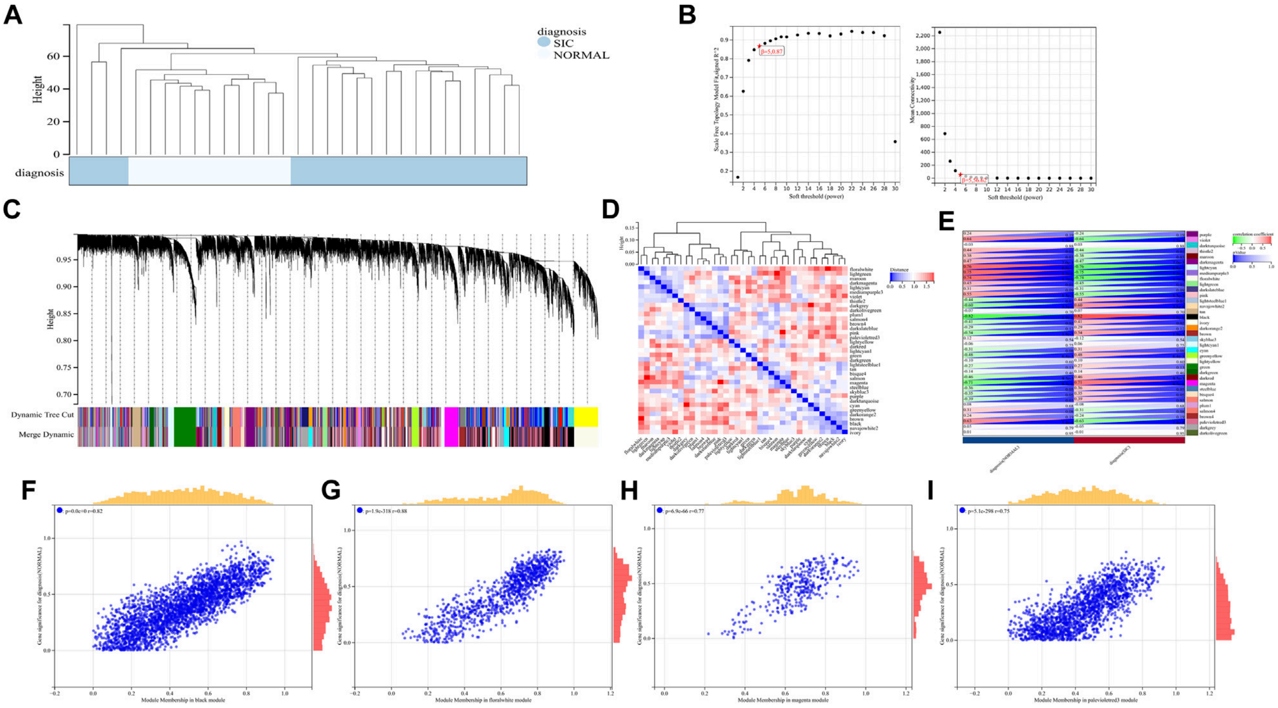


FIGURE 4
(A–I) WGCNA analysis of GSE79962.

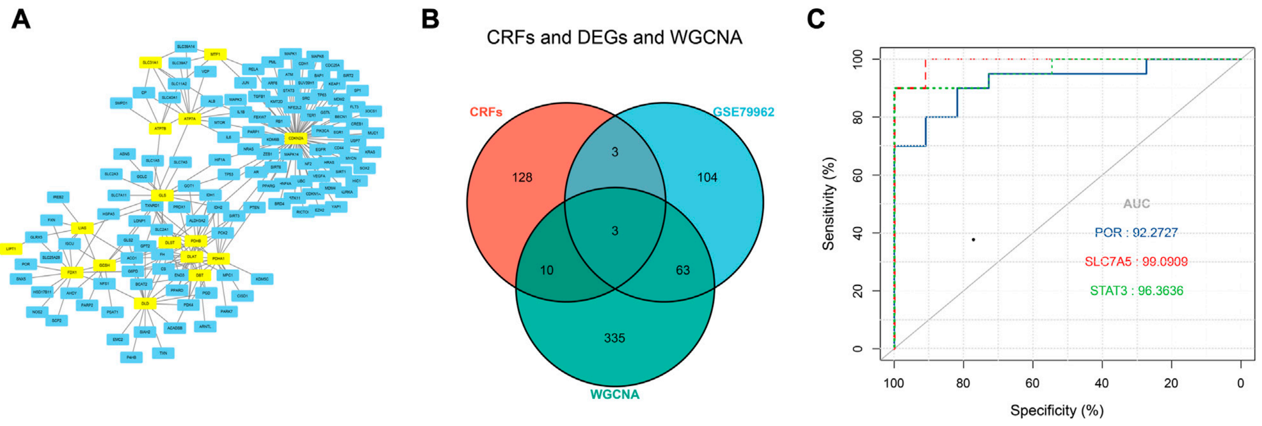


FIGURE 5
(A) The correlation between cuproptosis-related genes and ferroptosis-related genes (CRFs). **(B)** Venn diagram of CRFs related DEGs. **(C)** Quantification of ROC curves values of AUC for POR, SLC7A5, STAT3.

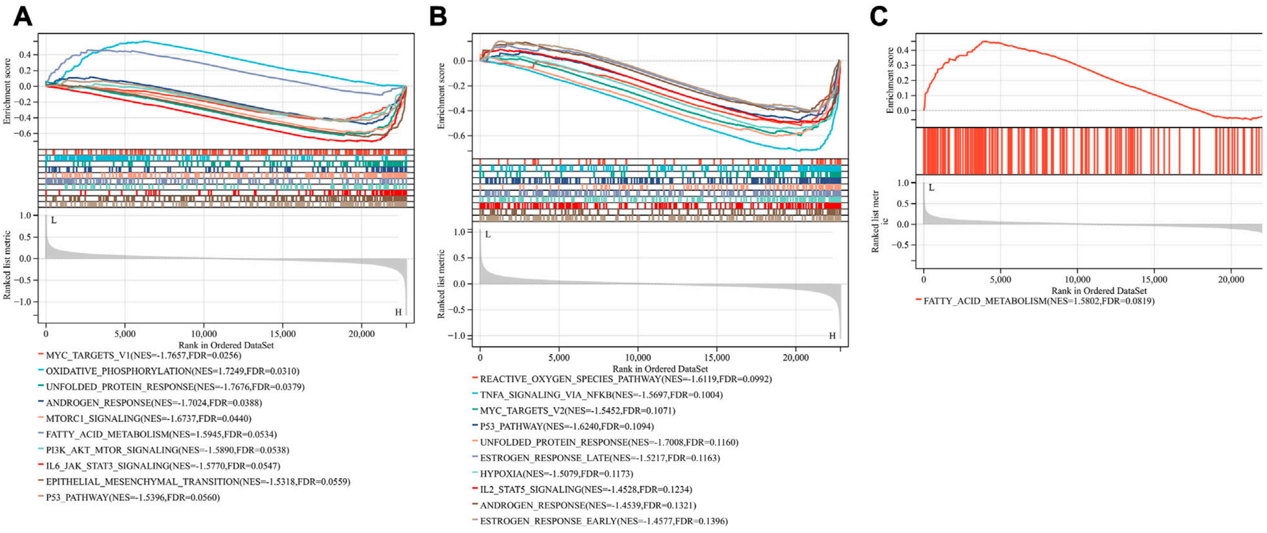


FIGURE 6
Gene set enrichment analysis (GSEA). The pathway related to three genes **(A)** POR **(B)** SLC7A5 **(C)** STAT3.

16 cuproptosis-related genes from the literature (Li et al., 2022; Tsvetkov et al., 2022). A protein-protein interaction network was created using the STRING database to further explore relationships among these genes. We identified 128 ferroptosis-related genes to be closely associated with cuproptosis-related genes. Therefore, we constructed a novel signature (CRF) by combing cuproptosis-related genes with ferroptosis-related genes. The protein-protein interaction network was constructed based on the STRING online database and visualized using Cytoscape software (Figure 5A). By taking the intersection of the DEGs, WGCNA-based hub genes and CRF genes, three overlapping genes (POR, SLC7A5, and STAT3) were identified for SIC (Figure 5B). The diagnostic values of the three genes were confirmed by ROC curve and AUC analysis. As shown in Figure 5C, the AUC values of POR, SLC7A5 and STAT3 produced diagnosis powers for SIC of 0.922727, 0.990909, and 0.963636, respectively. Subsequent GSEA showed that the hallmark gene sets of the three genes were for fatty acid metabolism (Figures 6A–C).

Construction of a TF-miRNA-hub gene network for SIC

We further investigated the regulatory mechanism of these three genes in SIC. The target miRNAs and TFs of the three genes were identified and then the TF–miRNA-hub gene network was constructed based on miRnet. Finally, a

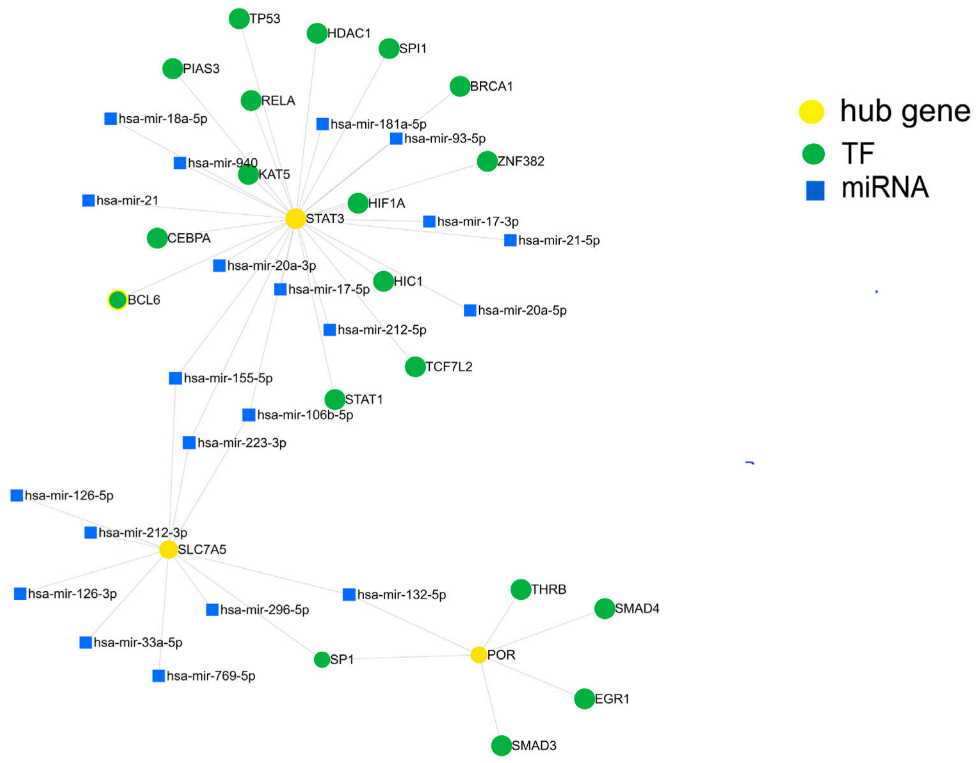


FIGURE 7
Construction of the TF-miRNA-hub gene network in sepsis-induced cardiomyopathy based on miRnet.

TABLE 1 The potential compounds of three genes were identified using DGIdb.

Gene	Drug	match_type	Sources
POR	NICOTINE	Definite	PharmGKB
POR	MIDAZOLAM	Definite	PharmGKB
POR	CYCLOSPORINE	Definite	PharmGKB
POR	ZIDOVUDINE	Definite	PharmGKB
POR	SIROLIMUS	Definite	PharmGKB
POR	ATORVASTATIN	Definite	PharmGKB
POR	SUNITINIB	Definite	PharmGKB
POR	TACROLIMUS	Definite	PharmGKB
SLC7A5	MELPHALAN	Definite	PharmGKB
STAT3	ACITRETIN	Definite	TTD
STAT3	PYRIMETHAMINE	Definite	DTC
STAT3	DIGITOXIN	Definite	DTC
STAT3	NICLOSAMIDE	Definite	DTC
STAT3	DIGOXIN	Definite	DTC
STAT3	OUABAIN	Definite	DTC

TF–miRNA-hub gene network, which included the three genes, 19 TFs, and 21 miRNAs, was constructed with 45 edges (Figure 7).

Screening for SIC therapeutic drugs

Potential therapeutic compounds for SIC associated with the three hub genes were screened for using DGIdb (Table 1). We identified 15 drugs as potential therapeutic compounds for SIC.

Discussion

Sepsis has become one of the top ten causes of death in both developed and developing countries (Reinhart et al., 2017). Sepsis-induced cardiomyopathy, which is common and closely associated with higher mortality, has been the focus of attention. Although intensive efforts have been made to understand the molecular mechanism of sepsis-induced cardiomyopathy, a precise definition and prognostic parameters remain uncertain. Although biomarkers were added to the physiological parameters of sepsis-induced cardiomyopathy, their release was observed to be generally inconsistent with the severity of the disease (Hollenberg and Singer, 2021).

Programmed cell death is critical for organ development, tissue homeostasis, as well as the prevention of tissue injury and tumorigenesis (Fuchs and Hermann, 2011). As a newly recognized form of programmed cell death, ferroptosis is closely related to the pathogenesis of a large variety of diseases, such as cancer (Lei et al., 2022), cardiovascular disease (Fang et al., 2019), Parkinson's disease (Bruce et al., 2016), chronic obstructive pulmonary disease (Yoshida et al., 2019), and autoimmune hepatitis (Zhu et al., 2021). Recent studies have shown that ferroptosis is closely related to the occurrence of sepsis and plays a crucial role in sepsis organ damage (Liu et al., 2022a). Vital roles of ferroptosis in the pathogenesis of SIC were also identified and ferritinophagy-mediated ferroptosis is involved in sepsis-induced cardiac injury (Ning et al., 2020). Both ferroptosis and cuproptosis are associated with mitochondria and are involved in the progression of a number of malignant tumors. Cuproptosis is also closely related to cardiovascular diseases. In this study, we screened cuproptosis- and ferroptosis-related genes using a bioinformatics approach. Previous studies have only focused on WCGNA to identify key genes; however, we applied text mining and WCGNA and, thereby, identified three hub genes (POR, SLC7A5, and STAT3). The diagnostic values of the three genes for SIC were confirmed using ROC curves.

The three hub genes are all associated with ferroptosis, their connection with cuproptosis in the pathophysiology of SIC is unknown. Cytochrome p450 oxidoreductase (POR) encodes an oxidoreductase that is indispensable for metabolism (Sugishima et al., 2019). The reactive oxygen species (ROS) that initiate ferroptosis come from a variety of sources, including iron-mediated Fenton reactions, mitochondrial ROS, and membrane-associated ROS driven by the NOX protein family. Polyunsaturated fatty acid-containing phospholipids are the main substrates of lipid peroxidation in ferroptosis, which is positively regulated by POR (Liu et al., 2022b). In a recent bioinformatics analysis of sepsis-induced cardiomyopathy, POR was selected as the central gene and its expression level was higher than that of the control group. GSEA then demonstrated POR to have a close relationship with cardiac metabolism, necroptosis and apoptosis of cells in SIC (Li et al., 2021). Solute carrier family 7 member 5 (SLC7A5), also known as L-type amino acid transporter (LAT1) (Galluccio et al., 2013), is a sodium-independent high-affinity amino acid transporter. SLC7A5 together with SLC3A2 mediate cellular uptake of the large neutral amino acids, phenylalanine, tyrosine, leucine, and tryptophan (Mastroberardino et al., 1998). SLC7A5 may affect the development of many diseases by regulating ferroptosis (Mao and Ma, 2022). Signal Transducer and Activator of Transcription 3 (STAT3) is a member of the STAT protein family. It can trigger transcription of a variety of genes in response to cytokines, which play a key role in many cellular processes, such as cell growth, apoptosis and ferroptosis. Accumulating evidence indicates STAT3 to be a converging point of multiple inflammatory response pathways in sepsis pathophysiology (Lei et al., 2021). This indicates that these genes are promising targets for drug development. Regulating fatty acid metabolism can sensitize cells to ferroptosis. In our study, GSEA showed that the hallmark gene sets of the three genes were for fatty acid metabolism. Therefore, we speculate that fatty acid metabolism maybe also involved in cuproptosis in SIC patients, but further experiments are needed to confirm this.

In addition to single protein-expressing genes, whole pathway networks may be deregulated in SIC. This may be mediated by miRNAs. miRNAs are associated with the pathophysiological process of many diseases (Formosa et al., 2022) and are involved in the occurrence and development of SIC (Beltrán-García et al., 2021). We therefore built a TF-miRNA-hub gene network depending on the shared dataset and published literature. We identified 19 TFs and 21 miRNAs as the master regulators of the resulting gene regulatory network that have the largest connectivity with the three co-expressed genes associated with SIC. Finally, by screening DGIdb, target therapeutic compounds for the hub genes were identified.

There are some limitations to this study. We only extracted data from databases and did not validate these data with animal experiments or clinical specimens. The screening results of this study were relatively accurate, which provides theoretical support for clinical drug development.

Conclusion

In summary, this study used bioinformatics methods to identify hub genes and pathways involved in sepsis-induced cardiomyopathy and revealed the potential role of ferroptosis and cuproptosis. Our findings indicated 15 drugs as candidates for sepsis-induced cardiomyopathy therapy. Further studies are needed to explore the causal relationship between ferroptosis and cuproptosis and sepsis-induced cardiomyopathy and to provide prognostic markers. Overall, our analysis provides a workflow for predicting biomarkers and drug targets, which can be widely used in other diseases.

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

YD contributed to the study design. JS conducted the literature search, acquired the data and wrote the article. KR performed data analysis. JS and LS drafted the article. DZ and XL revised the article. YD and HZ gave the final approval of the version to be submitted.

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.

References

Beltrán-García, Jesús, Osca-Verdegal, Rebeca, Nácher-Sendra, Elena, Cardona-Monzonis, A., Sanchis-Gomar, F., Carbonell, N., et al. (2021). Role of non-coding RNAs as biomarkers of deleterious cardiovascular effects in sepsis. *Prog. Cardiovasc Dis.* 68, 70–77. doi:10.1016/j.pcad.2021.07.005

Bruce, Do Van, Gouel, Flore, Jonneaux, Aurélie, Timmerman, K., Gele, P., Petrault, M., et al. (2016). Ferroptosis, a newly characterized form of cell death in Parkinson's disease that is regulated by PKC. *Neurobiol. Dis.* 94, 169–178. doi:10.1016/j.nbd.2016.05.011

Chen, Liyun, Min, Junxia, and Wang, Fudi (2022). Copper homeostasis and cuproptosis in health and disease. *Signal Transduct. Target Ther.* 7 (1), 378. doi:10.1038/s41392-022-01229-y

Chen, Mengwei, Kong, Chengqi, Zheng, Zhiyuan, and Yin, Li (2020). Identification of biomarkers associated with septic cardiomyopathy based on bioinformatics analyses. *J. Comput. Biol.* 27 (1), 69–80. doi:10.1089/cmb.2019.0181

Cui, Xiangning, Wang, Yan, Han, Liu, Shi, M., and Wang, J. (2022). The molecular mechanisms of defective copper metabolism in diabetic cardiomyopathy. *Oxid. Med. Cell Longev.* 2022, 5418376. doi:10.1155/2022/5418376

Fang, X., Wang, H., Han, D., Xie, E., Yang, X., Wei, J., et al. (2019). Ferroptosis as a target for protection against cardiomyopathy. *Proc. Natl. Acad. Sci. U. S. A.* 116 (7), 2672–2680. doi:10.1073/pnas.1821022116

Formosa, Amanda, Turgeon, Paul, and ClaudiaDos Santos, C. (2022). Role of miRNA dysregulation in sepsis. *Mol. Med.* 28 (1), 99. doi:10.1186/s10020-022-00527-z

Fuchs, Yaron, and Hermann, Steller (2011). Programmed cell death in animal development and disease. *Cell* 147 (4), 1640–1658. doi:10.1016/j.cell.2011.11.045

Galluccio, M., Pingitore, P., Scalise, M., and Indiveri, C. (2013). Cloning, large scale over-expression in *E. coli* and purification of the components of the human LAT 1 (SLC7A5) amino acid transporter. *Protein J.* 32 (6), 442–448. doi:10.1007/s10930-013-9503-4

Hanumanthu, B. K. J., Nair, A. S., Katamreddy, A., Gilbert, J. S., You, J. Y., Offor, O. L., et al. (2021). Sepsis-induced cardiomyopathy is associated with higher mortality rates in patients with sepsis. *Acute Crit. Care* 36 (3), 215–222. doi:10.4266/acc.2021.00234

Hollenberg, Steven M., and Singer, Mervyn (2021). Pathophysiology of sepsis-induced cardiomyopathy. *Nat. Rev. Cardiol.* 18 (6), 424–434. doi:10.1038/s41569-020-00492-2

Lei, Guang, Zhuang, Li, and Gan, Boyi (2022). Targeting ferroptosis as a vulnerability in cancer. *Nat. Rev. Cancer* 22 (7), 381–396. doi:10.1038/s41568-022-00459-0

Lei, Wangrui, Liu, Dianxiao, Sun, Meng, Lu, C., Yang, W., Wang, C., et al. (2021). Targeting STAT3: A crucial modulator of sepsis. *J. Cell Physiol.* 236 (11), 7814–7831. doi:10.1002/jcp.30394

Li, Juexing, Zhou, Lei, Li, Zhenhua, Yang, S., Tang, L., and Gong, H. (2021). Identification of crucial genes and infiltrating immune cells underlying sepsis-induced cardiomyopathy via weighted gene Co-expression network analysis. *Front. Genet.* 12, 812509. doi:10.3389/fgene.2021.812509

Li, Zhi, Zhang, Hua, Wang, Xixi, Wang, Q., Xue, J., Shi, Y., et al. (2022). Identification of cuproptosis-related subtypes, characterization of tumor microenvironment infiltration, and development of a prognosis model in breast cancer. *Front. Immunol.* 13, 996836. doi:10.3389/fimmu.2022.996836

Liu, Jiao, Kang, Rui, and Tang, Daolin (2022). Signaling pathways and defense mechanisms of ferroptosis. *FEBS J.* 289 (22), 7038–7050. doi:10.1111/febs.16059

Liu, Yanting, Tan, Sichuang, Wu, Yongbin, and Tan, Sipin (2022). The emerging role of ferroptosis in sepsis. *DNA Cell Biol.* 41 (4), 368–380. doi:10.1089/dna.2021.1072

Liu, Yun, and Miao, Ji (2022). An emerging role of defective copper metabolism in heart disease. *Nutrients* 14 (3), 700. doi:10.3390/nu14030700

Mao, Jingyi, and Ma, Xin (2022). Bioinformatics identification of ferroptosis-associated biomarkers and therapeutic compounds in psoriasis. *J. Oncol.* 2022, 3818216. doi:10.1155/2022/3818216

Mastroberardino, L., Spindler, B., Pfeiffer, R., Skelly, P. J., Loffing, J., Shoemaker, C. B., et al. (1998). Amino-acid transport by heterodimers of 4F2hc/CD98 and members of a permease family. *Nature* 395 (6699), 288–291. doi:10.1038/26246

Ning, Li, Wang, Wei, Zhou, Heng, Wu, Qingqing, Duan, Mingxia, Liu, C., et al. (2020). Ferritinophagy-mediated ferroptosis is involved in sepsis-induced cardiac injury. *Free Radic. Biol. Med.* 160, 303–318. doi:10.1016/j.freeradbiomed.2020.08.009

Reinhart, Konrad, Daniels, Ron, Kissoon, Niranjan, Machado, Flavia R., Schachter, R. D., and Finfer, S. (2017). Recognizing sepsis as a global health priority—A WHO resolution. *N. Engl. J. Med.* 377 (5), 414–417. doi:10.1056/nejmp1707170

Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., et al. (2020). Global, regional, and national sepsis incidence and mortality, 1990-2017: Analysis for the global burden of disease study. *Lancet* 395 (10219), 200–211. doi:10.1016/S0140-6736(19)32989-7

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., et al. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* 315 (8), 801–810. doi:10.1001/jama.2016.0287

Sugishima, Masakazu, Sato, Hideaki, Wada, Kei, and Yamamoto, Ken (2019). Crystal structure of a NADPH-cytochrome P450 oxidoreductase (CYPOR) and heme oxygenase 1 fusion protein implies a conformational change in CYPOR upon NADPH/NADP+ binding. *FEBS Lett.* 593 (8), 868–875. doi:10.1002/1873-3468.13360

Tsvetkov, P., Coy, S., Petrova, B., Dreishpoon, M., Verma, A., Abdusamad, M., et al. (2022). Copper induces cell death by targeting lipoylated TCA cycle proteins. *Science* 375 (6586), 1254–1261. doi:10.1126/science.abf0529

Wang, Chunyan, Yuan, Wenlin, Hu, Anmin, Lin, Juan, Xia, Zhengyuan, Yang, C. F., et al. (2020). Dexmedetomidine alleviated sepsis-induced myocardial ferroptosis and septic heart injury. *Mol. Med. Rep.* 22 (1), 175–184. doi:10.3892/mmr.2020.11114

Yan, H. F., Zou, T., Tuo, Q. Z., Xu, S., Li, H., Belaidi, A. A., et al. (2021). Ferroptosis: Mechanisms and links with diseases. *Signal Transduct. Target Ther.* 6 (1), 49. doi:10.1038/s41392-020-00428-9

Yoshida, Masahiro, Minagawa, Shunsuke, Araya, Jun, Sakamoto, T., Hara, H., Tsubouchi, K., et al. (2019). Involvement of cigarette smoke-induced epithelial cell ferroptosis in COPD pathogenesis. *Nat. Commun.* 10 (1), 3145. doi:10.1038/s41467-019-10991-7

Zhu, Lujian, Chen, Dazhi, Zhu, Yin, Pan, T., Xia, D., Cai, T., et al. (2021). GPX4-Regulated ferroptosis mediates S100-induced experimental autoimmune hepatitis associated with the Nrf2/HO-1 signaling pathway. *Oxid. Med. Cell Longev.* 2021, 6551069. doi:10.1155/2021/6551069

TYPE Original Research
PUBLISHED 05 April 2023
DOI 10.3389/fonc.2023.1134445

Cell-free circulating tumor RNAs in plasma as the potential prognostic biomarkers in colorectal cancer

Nana Jin[1†], Chau-Ming Kan[2†], Xiao Meng Pei[3], Wing Lam Cheung[3], Simon Siu Man Ng[4], Heong Ting Wong[5], Hennie Yuk-Lin Cheng[3], Wing Wa Leung[4], Yee Ni Wong[4], Hin Fung Tsang[2], Amanda Kit Ching Chan[6], Yin Kwan Evelyn Wong[2], William Chi Shing Cho[7], John Kwok Cheung Chan[6], William Chi Shing Tai[3], Ting-Fung Chan[8,9], Sze Chuen Cesar Wong[3*], Aldrin Kay-Yuen Yim[1*] and Allen Chi-Shing Yu[1*]

[1]R&D, Codex Genetics Limited, Hong Kong, Hong Kong SAR, China, [2]Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China, [3]Department of Applied Biology & Chemical Technology, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China, [4]Department of Surgery, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China, [5]Department of Pathology, Kiang Wu Hospital, Macau, Macau SAR, China, [6]Department of Pathology, Queen Elizabeth Hospital, Hong Kong, Hong Kong SAR, China, [7]Department of Clinical Oncology, Queen Elizabeth Hospital, Hong Kong, Hong Kong SAR, China, [8]School of Life Sciences, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China, [9]State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China

Background: Cell free RNA (cfRNA) contains transcript fragments from multiple cell types, making it useful for cancer detection in clinical settings. However, the pathophysiological origins of cfRNAs in plasma from colorectal cancer (CRC) patients remain unclear.

Methods: To identify the tissue-specific contributions of cfRNAs transcriptomic profile, we used a published single-cell transcriptomics profile to deconvolute cell type abundance among paired plasma samples from CRC patients who underwent tumor-ablative surgery. We further validated the differentially expressed cfRNAs in 5 pairs of CRC tumor samples and adjacent tissue samples as well as 3 additional CRC tumor samples using RNA-sequencing.

Results: The transcriptomic component from intestinal secretory cells was significantly decreased in the in-house post-surgical cfRNA. The *HPGD*, *PACS1*, and *TDP2* expression was consistent across cfRNA and tissue samples. Using the Cancer Genome Atlas (TCGA) CRC datasets, we were able to classify the patients into two groups with significantly different survival outcomes.

Conclusions: The three-gene signature holds promise in applying minimal residual disease (MRD) testing, which involves profiling remnants of cancer

OPEN ACCESS

EDITED BY
Chiara Bazzichetto,
Regina Elena National Cancer Institute (IRCCS), Italy

REVIEWED BY
Serena Bonin,
University of Trieste, Italy
Andreas Herbst,
LMU Munich University Hospital, Germany

*CORRESPONDENCE
Allen Chi-Shing Yu
allenyu@codexgenetics.com
Aldrin Kay-Yuen Yim
aldrinyim@codexgenetics.com
Sze Chuen Cesar Wong
cesar.wong@polyu.edu.hk

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Gastrointestinal Cancers: Colorectal Cancer, a section of the journal Frontiers in Oncology

RECEIVED 30 December 2022
ACCEPTED 24 March 2023
PUBLISHED 05 April 2023

CITATION
Jin N, Kan C-M, Pei XM, Cheung WL, Ng SSM, Wong HT, Cheng HY-L, Leung WW, Wong YN, Tsang HF, Chan AKC, Wong YKE, Cho WCS, Chan JKC, Tai WCS, Chan T-F, Wong SCC, Yim AK-Y and Yu AC-S (2023) Cell-free circulating tumor RNAs in plasma as the potential prognostic biomarkers in colorectal cancer. *Front. Oncol.* 13:1134445. doi: 10.3389/fonc.2023.1134445

cells after or during treatment. Biomarkers identified in the present study need to be validated in a larger cohort of samples in order to ascertain their possible use in early diagnosis of CRC.

KEYWORDS

cell-free circulating tumor RNAs, colorectal (colon) cancer, CRC prognostic biomarkers, RNA sequencing (RNA-seq), transcriptome (RNA-seq)

1 Introduction

Colorectal cancer (CRC) is the third leading cause of cancer-related mortality and morbidity in the world[1] (1–4). One of the major factors affecting the survival of patients with CRC is the high frequency of recurrence after curative surgery, which is estimated to be 22.5% at 5 years. Approximately 11% of patients survive for 5 years after recurrence (5). Even though advances in cancer therapy have been made in recent decades, metastatic cancer and recurrence still pose a serious threat to the survival of CRC patients (6). Therefore, the identification of post-treatment biomarkers that reflect the potential of CRC recurrence is required to improve the survival of patients.

Genomic alterations associated with oncogenic drivers have traditionally been detected with invasive tissue biopsy, which is highly dependent on the amount of tumor tissue recovered in the biopsy and the initial analysis of the tissue for diagnosis (7). Liquid biopsy, through the use of circulating tumor molecules isolated from blood, has shown to be a promising minimally-invasive approach to detect, monitor, and evaluate the genetic profile of cancer patients (8). Currently, tumor-derived circulating cell-free DNA (ctDNA) analysis has been shown to predict cancer progression. However, there is only a limited amount of ctDNA shed into the circulation, and have different characteristics from patient to patient, which is hard to determine the tumor tissue of origin in cancer patients (9, 10). Although the circulating cfDNA methylation approach in plasma was effective in detecting and localizing cancer with higher specificity (11, 12), these methods may be ineffective without extensive deep sequencing coverage, and their sensitivity and specificity may not be adequate (9, 10). According to our previous study, cfRNA could serve as a potential diagnostic biomarker for patients with colorectal adenoma (13, 14). Therefore, additional circulating cell-free RNA (cfRNA) biomarkers may be required to complement detection by ctDNA to detect cancer, especially at the earliest stages or monitoring the outcome of surgery (10).

Plasma cfRNA is released from cells through active secretion, necrosis, and apoptosis (15, 16). Plasma cfRNA can reflect localized tumor sites as well as systemic tumor responses (17). In this study, we have performed a comprehensive profiling of the transcriptome in both pre-surgical and post-surgical cfRNAs, as well as the paired CRC tumor samples and CRC tumor-adjacent samples, in order to examine the mutational landscape in cfRNAs upon removal of tumor tissue. We deconvolved the relative abundance of cell types in plasma samples using published single-cell RNA-seq datasets and examined whether tissue after surgical might lead to a decrease in the ratio of intestinal cell-associated RNAs in plasma. Novel cfRNA expression biomarkers that showed consistent gene expression changes across in-house plasma samples, tissue samples, and the CRC samples in TCGA were identified. Survival analysis was used to evaluate the prognostic performance of these potential biomarkers and quantitative reverse transcription polymerase chain reaction (qRT-PCR) was conducted to validate these biomarkers in plasma from an independent cohort of 36 cancer patients. The biomarkers we identified could play an important role in the early diagnosis and prognosis of CRC.

2 Materials and methods

2.1 Subject recruitment

A total of 45 CRC patients were recruited from the Prince of Wales Hospital (PWH) between May 2020 and January 2022 with the approval from the joint Chinese University of Hong Kong- New Territories East Cluster Clinical Research Ethics Committee (CUHK-NTEC CREC; Ref No: 2019.542). Only individuals unrelated to each other were included. Diagnosis of CRC was based on the histological confirmation of colon adenocarcinoma. Patients with hereditary CRC and inflammatory bowel disease were excluded in this study. Each patient was invited to donate tissues (CRC tumor samples and CRC tumor-adjacent samples) and blood (pre-surgery on the day before surgery and post-surgery on the 5th-7th day after surgery) for research purposes with written informed consent before the operation. After the surgical removal of the tumor, the tissues were immediately preserved in RNAlater^{TM} Stabilization Solution (Cat# AM7020, Thermo Fisher Scientific, USA) at 4°C overnight in order to make sure the RNAlater can penetrate into the tissue. Then the tissues were stored at -80°C. The tumor-adjacent samples were cut 3 to 4 cm from the tumor. Plasma isolation was performed within 3 hours after the anti-coagulated blood collection using the VACUETTE® TUBE 2 ml K2E K2EDTA (Cat#454024, Greiner Bio-one, Austria). The blood was firstly centrifuged for 1,600 g, 10 minutes at 4 °C. The upper layer plasma without disturbing the buffy coat was collected to the other tube, then re-centrifuged for 16,000 g, 4 °C for 10 minutes to remove residual cell

[1] https://seer.cancer.gov/statfacts/html/colorect.html

pellet. After that, plasma was collected and preserved by 2 ml TRIzol^{TM} LS Reagent (Cat#10296028, Thermo Fisher Scientific, USA) before storage at −80 °C.

2.2 Extraction of cfRNAs from blood

Eight pairs of pre- and post-surgical cfRNA that were prepared for sequencing were extracted from 2-4 ml plasma by using 10ml TRIzol^{TM} LS Reagent (Cat#10296028,Thermo Fisher Scientific, USA). The cfRNA was extracted by using QIAamp cfRNA/cfDNA extraction kit (Cat#55184, Qiagen, Germany) following the manufacturer's instruction and eluted in 30ul water. The RNA quality was assessed by the TapeStation using High sensitivity RNA assay (Cat#5067-5579, Agilent, USA). The RNA quantity was measured by Qubit^{TM} RNA High Sensitivity (HS) (Cat# Q32852, Invitrogen^{TM}, USA) (Supplementary Table S1).

2.3 Total RNA extraction from tissues

The tissues were shredded by a homogenizer. CRC tumor samples and CRC tumor-adjacent samples from eight patients that were prepared for sequencing were extracted from the AllPrep DNA/RNA kit (Qiagen). The RNA quality was assessed by the TapeStation, using High sensitivity RNA assay (Cat#5067-5579, Agilent, USA). The RINs for all tissue RNA were> 2. The RNA quantity was measured by Qubit^{TM} RNA High Sensitivity (HS) (Cat# Q32852, Invitrogen^{TM}, USA).

2.4 Ribosomal RNA (rRNA) depletion and library construction for tissue RNA

rRNA depletion was performed on the extracted total RNAs from tissue and subsequent library prep following the NEBNext$^{®}$ rRNA Depletion Kit v2 (Human/Mouse/Rat) (Cat#7400L, New England BioLabs, England)'s protocol, which depletes both mitochondrial (12S and 16S) and cytoplasmic (5S, 5.8S, 18S, and 28S) rRNA species. cDNA synthesis was performed by using Maxima First Strand cDNA Synthesis Kit for RT-qPCR, with dsDNase (Cat#1671, Thermo Scientific^{TM}, USA). End-repair, A tailing, adaptor ligation, and library amplification were performed by using the KAPA HyperPlus kit (Cat#KK8512, Rocha, USA). Completed libraries were quantified by each library by Qubit^{TM} 1X dsDNA High Sensitivity (HS) assay kit (Cat#Q33231, Invitrogen^{TM}, USA) and the insert size estimation was measured by TapeStation, using D1000 ScreenTape assay (Cat#5067-5582, Agilent, USA).

2.5 Library construction for plasma cfRNA

In order to compare the genetic composition of cfRNA before and after surgery, rRNA depletion was not performed in cfRNA as part of the whole transcriptome study (10).

cfRNAs were converted to the cDNA by using SMARTer$^{®}$ Universal Low Input RNA Kit for Sequencing (Cat#634940, Takara Bio, Japan). End-repair, A tailing, adaptor ligation, and library amplification were performed according to the protocol of the NEBNext$^{®}$ Ultra^{TM} II DNA Library Prep Kit for Illumina$^{®}$ (Cat# E7645S, New England BioLabs, England). Completed libraries were quantified by each library by Qubit^{TM} 1X dsDNA High Sensitivity (HS) assay kit (Cat#Q33231, Invitrogen^{TM}, USA) and the insert size estimation was measured by TapeStation, using D1000 ScreenTape assay (Cat#5067-5582, Agilent, USA).

2.6 RNA sequencing

The Illumina sequencing adaptors were ligated onto the fragments. Constructed libraries were sequenced (300 cycles) using Illumina NextSeq550 (Illumina Inc), according to the manufacturer's instructions. The Binary Base Call (BCL) files were converted to FASTQ files using the Illumina BCL Convert (v3.7.5). Raw-seq reads quality was assessed using FastQC (v0.11.9)[2] (18). Adapters and low-quality bases (Q<20 in 4bp sliding window) were trimmed using fastp (v0.20.1) (19). Specifically, seven bases SMARTer adapter from both ends of the reads will be trimmed for plasma cfRNA only. Clean RNA-seq reads were then mapped to the human genome from the Genome Reference Consortium (GRCh38) using STAR aligner (v2.7.7a) with the 2-pass mode (20). Alignments were quantitated using HTSeq (v0.13.5) (21) overlapping with the annotations in GENCODE human release 35. The definition of the biotypes was referenced to GENCODE[3] (22). Gene expression estimation in terms of Fragments Per Kilobase of transcript per Million mapped reads (FPKM) and differential expression analysis was performed by the R (v4.0.5)/Bioconductor package DESeq2 (v1.30.1) (23). Reactome Pathway Database (24) annotation was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID v2021) (25).

2.7 Public dataset collections

2.7.1 TCGA dataset

Gene expression data and the corresponding clinical information of 453 patients with CRC (colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ))[4] were downloaded from the TCGA data portal (26), including 453 CRC tumor samples and 42 CRC tumor-adjacent samples. The identification of the differentially expressed genes (DEGs) was performed using the R/Bioconductor package DESeq2 (v1.30.1).

[2] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[3] https://www.gencodegenes.org/pages/biotypes.html

[4] https://portal.gdc.cancer.gov

2.7.2 CRC single-cell datasets

Single-cell 3' mRNA sequencing data from 23 colorectal cancer patients with the annotation to the cell types including B cell, epithelial cell, mast cell, myeloid cell, stromal cell, and T cell were downloaded from the Gene Expression Omnibus (GEO) database (GSE132465) (27).

2.7.3 Tabula Sapiens

Tabula Sapiens version 1.0 was used to determine the origin of cells of plasma transcriptome. Tabula Sapiens is a human cell atlas of nearly 500,000 cells from 24 organs. The single cell signature used in CIBERSORTx referred to the deconvolution of cell-free RNA tutorial (28) (https://github.com/sevahn/deconvolution/tree/master/deconvolve_cfrna_tutorial).

2.7.4 Cell type abundance determination

The single-cell datasets were used to deconvolute the cell type proportion of bulk tissues and plasma using CIBERSORTx (29). The top 1,000 variated genes in CRC single-cell dataset and Tabula Sapiens dataset were used as the single-cell signatures. All parameters were set as default, except for the permutation was set as 1,000 in the cell fraction imputation step.

2.7.5 Reference-guided *de novo* assemblies

Reference-guided *de novo* assemblies were assembled and quantitated using StringTie (v2.1.4) (30) after mapping to the human genome from the GRCh38 using HISAT2 (v2.2.1) (31), overlapping with the annotations in GENCODE human release 35. Gffcompare (v0.11.2) (32) was used to compare with the reference annotation. The transcripts with the classification code of i, x, y, and u were defined as novel transcripts, otherwise, the transcripts were defined as known transcripts. Differential expression analyses were performed by the R/Bioconductor package ballgown (v2.22.0) (33). The transcripts that (1) not overlapped with regulatory regions in its 5kb upstream and downstream regions from the transcription start site, and (2) with abs(log2Fold-Change) of the expression less than 1 were filtered out as transcripts with low confidence. Regulatory regions were obtained from ORegAnno (v3.0) (34). CPC 2.0 (35) was used to predict the coding potential for the assembled transcripts. AnnoLnc2 (36) was used to predict the expression of the novel transcripts in human samples. We used lncPro to predict the interaction between novel transcripts and proteins (37).

2.8 qRT-PCR validation

Plasma samples from 36 patients were used to validate the expression of the candidate genes. The cfRNA was extracted from 1-4.5 ml TRIzol^{TM} LS Reagent (Cat#10296028, Thermo Fisher Scientific, USA) preserved plasma by using miRNeasy Serum/Plasma Kit (Cat#217184, Qiagen, Germany). RNA quantity was measured by Qubit^{TM} RNA High Sensitivity (HS) (Cat# Q32852, Invitrogen^{TM}, USA). A majority of the extracted RNAs were below the limit of detection (LOD) of the Qubit^{TM} RNA High Sensitivity (HS) (Cat# Q32852, Invitrogen^{TM}, USA) (LOD<10ng) (Supplementary Table S2). Reverse transcription reactions were performed following the manufacturer's instructions using PrimeScript RT Master Mix (Takara) in 10 μL reactions. Otherwise, 30ng RNA was input for reverse transcription.

The primers (Supplementary Table S3) for the candidate genes were designed based on the gene sequences gained from the GeneBank, National Centre for Biotechnology Information, NCBI and validated for the absence of self and cross dimers, secondary structures as well as primer efficiency and specificity. Melting curve plot of RT-PCR products showed that no unspecific amplification was detected (Supplementary Figure S1).

qRT-PCR assays were performed using the SsoAdvanced Universal SYBR Green Supermix (Cat# 1725270, Bio-Rad, USA) in ABi ViiA7 Real-Time PCR System (ThermoFisher Scientific) in a 20 μL reaction volume according to the manufacturer's instructions. The thermal cycling condition was 30 seconds at 95° C for initial activation, followed by 45 cycles of 15 seconds at 95°C and 60 seconds at 60°C.

GAPDH was demonstrated as useful housekeeping gene to normalize the data, in order to determine the relative target gene expression in cfRNA samples (38). The gene expression was normalized to *GAPDH* among the same patient by delta-delta Ct method as following. The expression level of GAPDH was detected as stable among samples (Supplementary Figure S2).

$$\Delta Ct = Ct(PACS1/HPGD/TDP2) - Ct(GAPDH)$$

$$\Delta\Delta Ct = \Delta Ct - \Delta Ct(pre-surgical \quad cfRNA)$$

$$\text{Fold change expression} \; = 2^{-\Delta\Delta Ct}$$

2.9 Survival analysis

451 TCGA CRC samples were split into training and test datasets: 70% of samples of the dataset were randomly selected as training dataset (N=315) and 30% as test dataset (N=136). Gene expression was standardized by removing the mean and scaling to unit variance before analysis. We generated a protective score for each sample – the accumulative weighted gene expression of the *HPGD*, *PACS1*, and *TDP2* by the first principal component. Linear regression was used to fine-tune the protective score. Then samples with a protective score>0.5 were classified as a low-risk group, otherwise as a high-risk group. AUC was used to evaluate the model performance. Survival curves were estimated by the Kaplan-Meier method and compared with a log-rank test.

2.10 Statistics

The correlation between gene expression in CRC tumor samples vs TCGA tumor samples and pre-surgical cfRNA vs CRC tumor samples was described using the linear regression model. The significance of the overlapping between significantly upregulated genes in pre-surgical plasma and upregulated protein-coding genes in TCGA tumor samples was described using the hypergeometric test. P values from the Wilcoxon rank-sum method indicated significance levels for differences in cell type proportion across sample groups. Gene expression detected using qRT-PCR was compared between post- and pre-surgical cfRNAs using the paired T-test. The error bars represented mean ± standard deviation (SD).

We used the Python library SciPy (v1.5.2) to perform the statistical analysis. We used adjusted p-value< 0.001 and abs (log2Fold-change) >1 to identify DEGs in in-house CRC tumor samples and CRC tumor-adjacent samples, as well as TCGA tumor samples and tumor-adjacent samples; p-value<0.05 to identify DEGs and DETs in pre- and post-surgical cfRNAs. Gene expression detected using qRT-PCR was compared between post- and pre-surgical cfRNAs using the paired T-test.

2.11 Study approval

Each patient was invited to donate tissues (CRC tumor samples and CRC tumor-adjacent samples) and blood (pre-surgical and post-surgical cfRNA) for research purposes with written informed consent before the operation. This study was approved by the joint Chinese University of Hong Kong- New Territories East Cluster Clinical Research Ethics Committee (CUHK-NTEC CREC; Ref No: 2019.542).

2.12 Data availability

The raw RNA-seq data of the plasma and tissue samples in this study are available in the NCBI Sequence Read Archive (SRA) database under the accession code PRJNA891435.

3 Results

A total of 45 rectal and colon adenocarcinoma patients were recruited in this study (62.2% men; age: 70.5 ± 8.8 years, Supplementary Table S4). cfRNA-seq was performed for 8 patients with matching pre- and post-surgical plasma samples (pre-surgical cfRNA and post-surgical cfRNA) (75% men; 71.6 ± 7.0 years), and bulk RNA-seq was performed using the 5 pairs of CRC tumor samples and CRC tumor-adjacent samples as well as 3 additional CRC tumor samples (87.5% men; age 70.5 ± 5.7 years, Figure 1A). The remaining 36 plasma samples (58.3% men; age 70.1 ± 9.3 years) were used in downstream qRT-PCR validation for the biomarkers discovered in this study.

3.1 Genetic characterization of plasma cell-free and tissue transcriptome

To characterize the expression landscape of CRC, RNA-seq was performed using the entire yield of extracted cfRNAs and tissue RNAs (see Methods). We systematically profiled the genetic

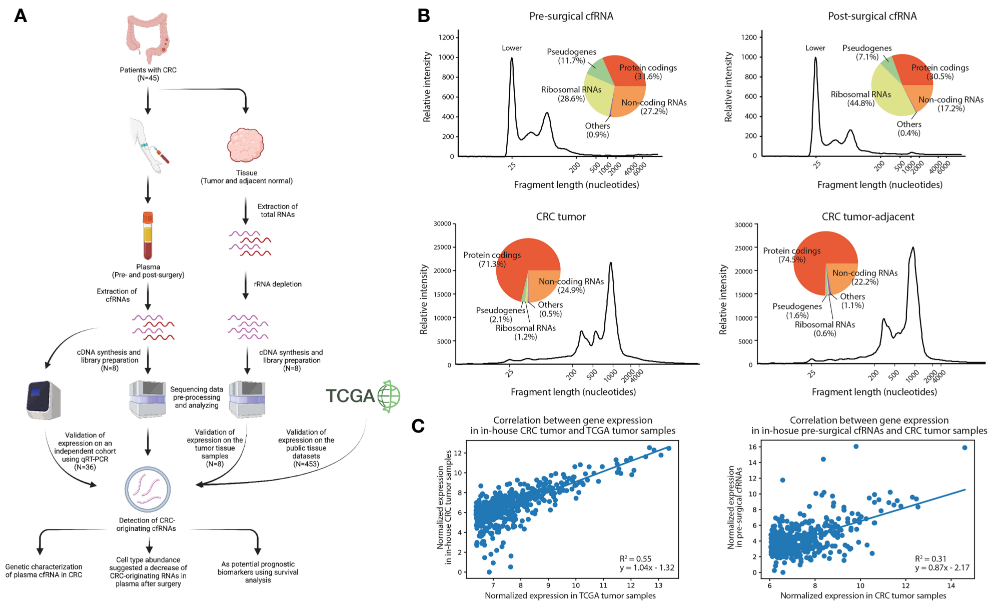


FIGURE 1
Analytical characterization of cell-free RNA and tissue transcriptome. **(A)** Experimental design of the study, Created with BioRender.com. **(B)** Relative intensity across different fragment lengths of plasma and tissue transcriptomes in a patient with CRC. Pie charts showed the percentage of gene expression in each biotype. **(C)** Correlation between gene expression of the top expressed genes in in-house tumor samples and TCGA datasets, and plasma and tissue datasets. The top 500 expressed genes in TCGA and in-house tissue were selected, respectively.

composition of the plasma cell-free and tissue transcriptome (Figure 1B; Supplementary Tables S5, S6). By comparing pre-surgical and post-surgical cfRNAs, we identified that the percentage of noncoding RNAs decreased significantly (*T-test*: p-value=1.42e-03) after surgery (39, 40), while the percentage of rRNAs increased significantly (T-test: p-value=1.20e-02). The genetic composition of tissue, however, has no significant variation between in-house CRC tumor samples and CRC tumor-adjacent samples as expected (Supplementary Table S6). This suggests that surgical removal of CRC tissue samples may affect the corresponding cfRNA abundance in plasma.

To examine the level of concordance between in-house CRC tumor sample RNA-seq profiles and published CRC RNA-seq data, we compared the gene expression level between 8 in-house CRC tumor samples and 453 TCGA CRC tumor samples (see Methods). A positive correlation (R^2 = 0.55, p-value=1.53e-89) was observed in the top 500 expressed genes in the TCGA dataset (Figure 1C). Interestingly, the expression between in-house CRC tumor samples and pre-surgical cfRNA was also positively correlated (R^2 = 0.31, p-value=9.05e-43). The correlation coefficient is higher than it between in-house CRC tumor-adjacent samples and pre-surgical cfRNA (R^2 = 0.23, p-value=1.17e-30), as well as between in-house CRC tumor samples and post-surgical cfRNA (R^2 = 0.27, p-value=4.67e-36; Supplementary Figure S3). The concordance between the in-house CRC tumor samples and pre-surgical cfRNA leads us to the hypothesis that the patients' cfRNA could be derived from subpopulations of cells within the tumor (41), additional analysis is therefore necessary to delineate the tissue origin of cfRNA in plasma as to identify biomarkers for CRC in blood.

3.2 Cell type abundance suggested a decrease of intestinal cell-originating RNAs in plasma after surgery

Given the correlation between in-house CRC tumor samples and pre-surgical cfRNA, we hypothesize that a portion of the cfRNA in plasma could be originating from the cancer tissue. We performed a single-cell deconvolution analysis to predict the relative ratio of contributing cell types based on their specific expression signatures. Firstly, we used CIBERSORTx to predict the cell type proportion of all in-house CRC tumor and CRC tumor-adjacent samples using the published CRC single-cell RNA-seq dataset (GSE132465). A marginal increase in myeloid cells was observed in CRC tumor samples than in CRC tumor-adjacent samples (Figure 2A; log2Fold-change=7.35; p-value=5.4e-02), consistent with the role of myeloid cells in providing growth factors and metabolites for tumor growth (42). B Cells, however, were depleted in the tumor samples when compared to CRC tumor-

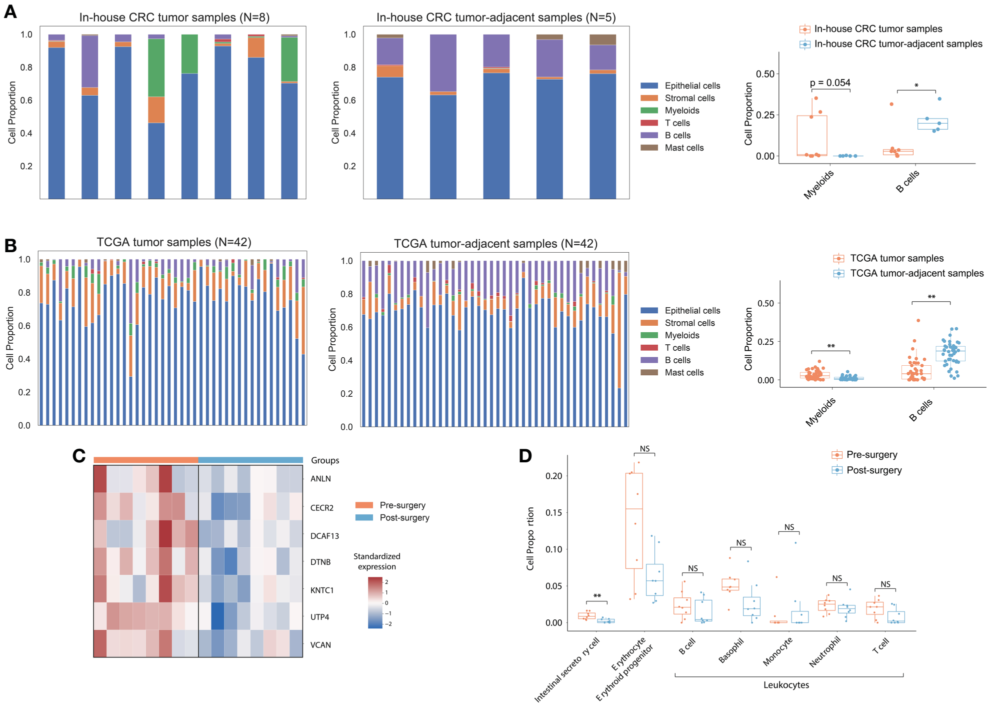


FIGURE 2
Cell type proportion in tissue and cfRNA. **(A)** Deconvoluted cell type proportion of in-house bulk CRC tumor samples and CRC tumor-adjacent samples. Box plot showed myeloid and B cell fraction distribution from CRC patients. **(B)** Deconvoluted cell type proportion of 42 TCGA bulk tumor and normal samples. Each stacked bar of the left and middle panels represented matched tumor and normal tissue from a single participant. Box plots showed myeloid and B cell fraction distribution from 42 CRC patients in TCGA. **(C)** Heatmap of expression for CRC tissue-highly expressed genes in plasma samples. **(D)** Box plots showed intestinal secretory cells, erythrocytes, erythroid progenitors, and leukocyte fraction distribution from pre- and post-surgical plasma samples. P values from the Wilcoxon rank-sum method indicated significance levels for differences in cell type proportion across sample groups. "**" represented P<0.01; "*" represented P<0.05. NS, not significant.

adjacent samples (log2Fold-change=-1.91; p-value=2.3e-02), which is expected for the inhibition role of B cells in tumor development (43). These results were also observed in the TCGA dataset (Figure 2B; p-value=2.6e-05 in myeloid cells; p-value=1.1e-08 in B cells) and showed high consistency with the findings from the single-cell RNA-seq data (27).

After demonstrating consistent cell-type specific expression signature between the in-house tissue sample RNA-seq data and TCGA dataset, we hypothesize that the proportion of cfRNA secreted by intestinal cells should be decreased in post-surgical cfRNA samples. Principal component analysis (PCA) based on the top 500 DEGs showed discriminating between the pre- and post-surgical plasma samples (Supplementary Figure S4). We detected 50 significantly upregulated genes that expressed across all the samples (FPKM>1) in pre-surgical cfRNA samples, 7 of them overlapped with the 2,379 upregulated protein-coding genes in TCGA tumor samples (Figure 2C; Hypergeometric test: p-value=3.63e-03). To determine if the up-regulated genes in pre-surgical cfRNA could be contributed by the intestinal-related cell, we further used the comprehensive human single-cell atlas - Tabula Sapiens (44) to deconvolute the cellular composition of the plasma samples (10). Consistent with the hypothesis, the proportion of intestinal secretory cells was significantly decreased after the surgery (p-value=7.2e-03) when compared to pre-surgical plasma samples (Figure 2D). We observed an insignificant change in the expression signature of erythrocyte, erythroid progenitor, and leukocytes between pre- and post-surgical samples (Figure 2D), which agrees with a previous study that demonstrated relatively stable expression of these cell types in plasma (28). Taken together, cfRNAs can reflect the intestinal tumor load, which has the potential to be the non-invasive biomarkers for CRC.

3.3 Identification of CRC non-invasive differential expression (DE) biomarkers

To identify potential blood-based DE biomarkers for CRC patients, we further performed a *de novo* assembly-based DE analysis to identify transcripts, potentially novel transcripts, that show consistent DE pattern across pre- and post-surgical cfRNA samples, as well as in-house CRC tumor samples and CRC tumor-adjacent samples (Figure 3A).

A total of 106,802 transcripts were assembled, the average length of which is 453 bp (see Methods, Supplementary Figure S5). We detected 409 differentially expressed transcripts (DETs) with more than 1 exon in the assembled transcripts. 268 out of the 409 transcripts were found to be known transcripts – overlapping with the GENCODE human release 35, and the remaining 141 transcripts were defined as novel transcripts (see Methods). Among the known transcripts, *RNU2-1* which was previously shown to be released from tissue to plasma among CRC patients (45), has been shown to be decreased in the post-surgical cfRNA (Supplementary Figure S6A; p-value=0.04, log2Fold-Change=-0.55). A lowered expression was also observed in in-house CRC tumor-adjacent samples (log2Fold-Change=-0.93) and TCGA tumor-adjacent samples (log2Fold-Change=-0.81; Supplementary Figure S6A). By selecting highly confident novel transcripts based on fold differences and TSS proximity (34) (see Methods), 10 transcripts were further shortlisted (Supplementary Table S7). Interestingly, we identified a significant decrease of the novel *MCF2L-intronic-AS* in post-surgical cfRNA (Supplementary Figures S6B, C; p-value=8.31e-03, log2Fold-Change=-1.03) located within the intronic region at antisense strand of *MCF2L*. This novel transcript was predicted as a non-coding transcript with a coding probability of 0.03 by using CPC 2.0 (35) and shown to be expressed in colon adenocarcinoma cell lines by AnnoLnc2 (36). The transcript was predicted to interact with a common set of proteins as *MCF2L-AS1* – a known antisense non-coding RNA of *MCF2L*. *MCF2L-AS1* showed distinctly higher expression in CRC compared to matched normal specimens (46), and its deficiency dramatically impeded cell proliferation, invasion, and migration capacities of CRC (47). *MCF2L-intronic-AS* may serve the consistent role as *MCF2L-AS1* according to interacting with the common proteins. In sum, these significantly depleted post-surgical cfRNAs could be contributed by the reduced intestinal secretory cells after surgical removal of the CRC tissue.

To identify genes with the same DE patterns in (i) CRC tumor tissue and tumor-adjacent tissue and (ii) pre- and post-surgical cfRNA samples (Figure 3A), we further performed DE analysis between CRC tumor and tumor-adjacent and identified 1,942 DE genes. Among these 1,942 genes, 11 genes were shown to be overlapping with the 409 DETs in cfRNAs. *CDCA7*, *CELSR3*, *PACS1*, *SNTB1*, and *TBC1D31* showed consistent upregulation in CRC tumor samples and pre-surgical cfRNA, while *GFI1B*, *HPGD*, *SH3BGRL2*, *SIAE*, *PKHD1L1*, and *TDP2* showed downregulation in CRC tumor samples and pre-surgical cfRNA. We further prioritize these genes based on their biomolecular functioning using Reactome Pathway Database (24). Only *HPGD*, *PACS1*, and *TDP2* showed involvement in biological pathways, including metabolism, infection, and DNA repair-related pathways (Supplementary Table S8).

3.4 Independent external validation of *HPGD, PACS1* and *TDP2* expression showed high concordance in CRC

We set out to validate the expression of the three cfRNA biomarkers – *HPGD*, *PACS1*, and *TDP2* identified in our in-house cfRNA and CRC tissue samples using an independent cohort of pre- and post-surgical cfRNA samples (N=36) and published TCGA CRC tumor and tumor-adjacent samples (N=453). *HPGD* has a significantly lower expression in in-house CRC tumor samples (p-value=4.63e-07, log2Fold-Change=-2.70) and pre-surgical cfRNA (p-value=1.67e-02, log2Fold-Change=-0.74). The loss expression of *HPGD* was reported in several colorectal carcinoma cell lines (48) and microscopic colon adenomas (49). We also observed a similarly low *HPGD* expression in TCGA tumor samples (p-value=6.67e-38, log2Fold-Change=-2.86) and the independent in-house pre-surgical cfRNA cohort (p-value=2.25e-02, log2Fold-Change=-0.58) (Figure 3B;

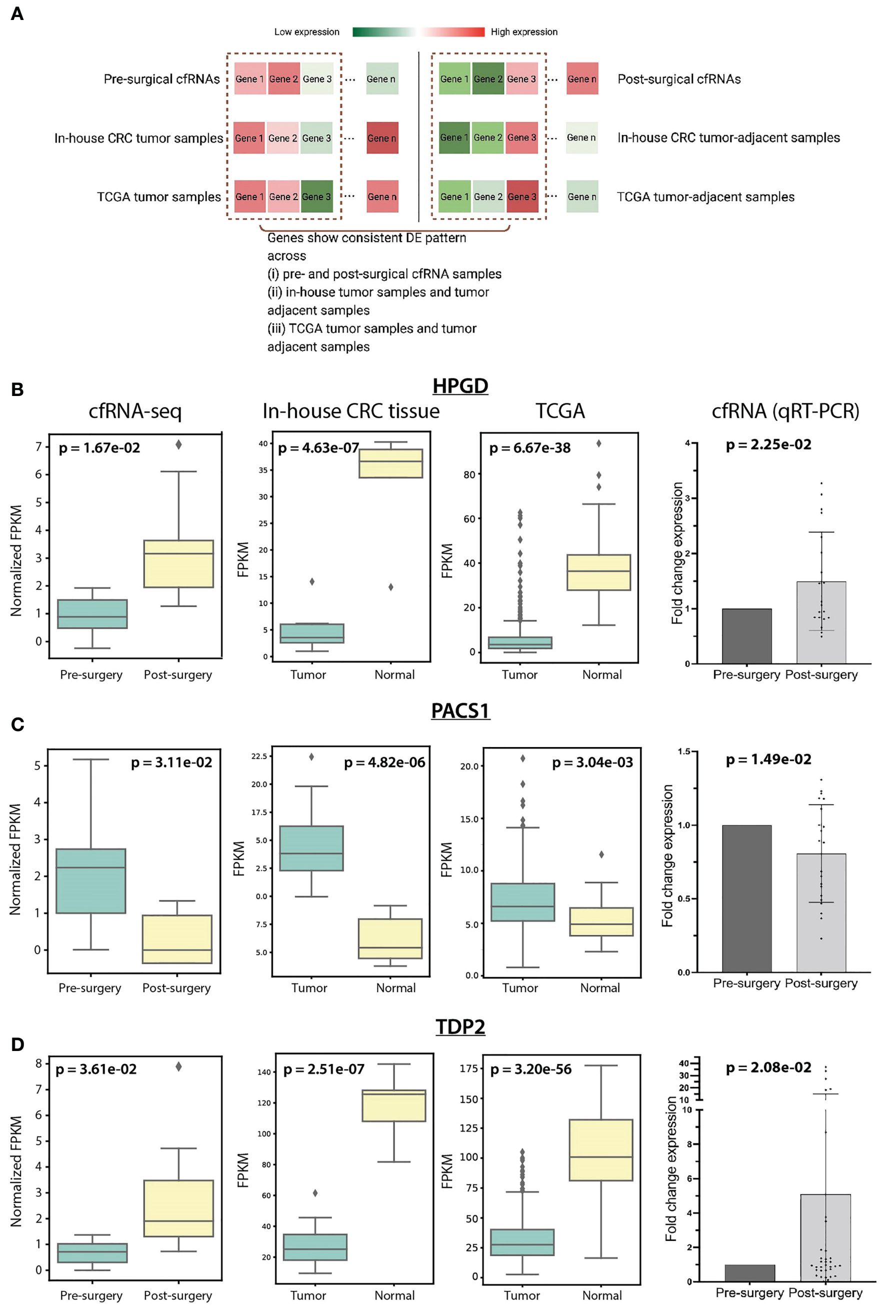


FIGURE 3
(A) Schematic diagram showed the consistent DEGs identification across cfRNAs, in-house CRC tissue samples, and TCGA samples. Gene expression of *HPGD* **(B)**, *PACS1* **(C)**, and *TDP2* **(D)** in cfRNAs, in-house CRC tissue samples, TCGA samples, and in-house qRT-PCR cfRNA data across sample groups.

Supplementary Table S9). Especially, the 9 out of 36 patients with N0 stage in the independent in-house cfRNA cohort showed a lower expression in pre-surgical cfRNA (p-value=2.53e-02, log2Fold-Change=-0.61; Supplementary Figure S7), implying a role of *HPGD* in early detection of CRC. The expression of *PACS1* and *TDP2* was also examined in both the TCGA CRC RNA-seq data and the independent pre- and post-surgical cfRNA cohort. *PACS1* expression is shown to be consistently higher in both in-house and TCGA CRC tumor samples, as well as pre-surgical cfRNA (Figure 3C). *TDP2*, however, is shown to be lowly expressed in the in-house tumor samples, TCGA CRC tumor samples, and pre-surgical cfRNA (Figure 3D). In summary, these results confirmed the monitoring potential of *HPGD*, *PACS1*, and *TDP2* in individuals with CRC.

3.5 Detection of survival outcome difference in TCGA CRC patients based on the linear combination of *HPGD*, *PACS1*, and *TDP2* expression

We next explored whether the expression of *HPGD*, *PACS1* and *TDP2* can guide the patient classification based on their survival time. We used a linear regression model to investigate the association between the survival time of TCGA CRC patients and the expression of *HPGD*, *PACS1*, and *TDP2* (see Methods). In order to evaluate the fitted model's accuracy in predicting the risk for CRC patients, we randomly split the TCGA CRC dataset into training (N=315) and test datasets (N=136) and used the receiver operating characteristic (ROC) and the area under the curve (AUC) to assess the model performance (see Methods). The AUC for the training dataset is 0.838 and 0.831 for the test dataset, indicating the good performance of the model (Figure 4A). *HPGD* (beta coefficients = -0.05, 95% confidence interval (CI): -0.09 to -0.02, p = 1.25e-03) and *PACS1* (beta coefficients=-0.06, 95% CI: -0.09 to -0.03, p = 6.34e-05) were identified as significant risk factors in the model, while *TDP2* (beta coefficients = 0.15, 95% CI: 0.11 to 0.19, p=3.61e-13) as a significant protective factor (Figure 4B). The linear combination of *HPGD*, *PACS1* and *TDP2* expression was used to assess patient survival probability (see Methods). A significant difference was detected for the training dataset (Log-rank p-value: 4.75e-02) and test dataset (Log-rank p-value: 2.95e-02) (Figure 4C). The median survival time for a low-risk group (N=110) in the test dataset was 1.65 years compared to 1.36 years for the high-risk group (N=26) (Figure 4C). Taken together, the linear combination of *HPGD*, *PACS1*, and *TDP2* expression showed an association with the survival probability of the CRC patient, suggesting the prognostic ability of these potential biomarkers.

4 Discussion

Identifying blood-based prognostic markers for minimally invasive cancer detection has been a major focus in the diagnostic area. ctDNA profiling is now being routinely applied clinically for both companion diagnosis and screening for minimal residual disease (MRD) among cancer patients. However, the detection of ctDNA for MRD is challenging as only a minute amount of ctDNA are present in blood at earlier cancer stages, especially in post-surgical setting (7, 10, 50). The cfDNA concentration may fall below the detection limit of the NGS-based ctDNA test, resulting in a very low or even zero mutation allele frequency (MAF) for the mutations (51). More importantly, it is difficult to determine the tumor tissue of origin (TOO) in cancer patients and differentiate informative cfDNA mutations from benign variants such as clonal hematopoiesis (7). Therefore through the amplification of tumor-derived RNA signal, we have shown that the detection of the expressed cfRNA in blood is technically feasible and may help circumvent the existing limitation in ctDNA detection, which will increase cancer detection sensitivity (7, 10). To our knowledge, this is the first study that compared the plasma transcriptomes derived both pre-operatively and post-operatively. Together with paired transcriptome derived from tumor tissues and adjacent normal tissues from CRC patients, we investigated the transcriptional landscape in both blood and tissue upon the surgical removal of CRC tissue.

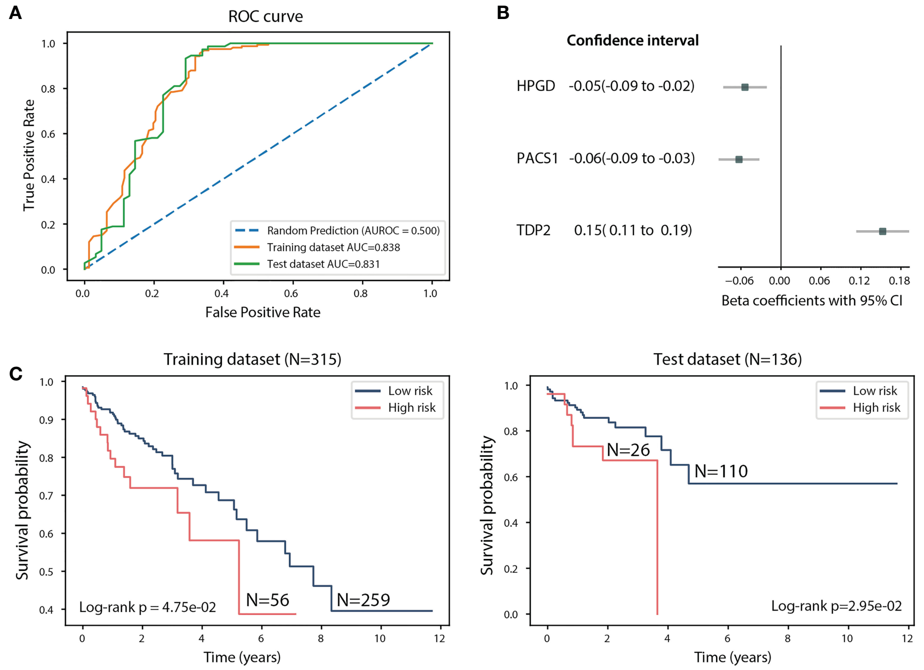


FIGURE 4
Assessment of the linear regression model using the 451 TCGA CRC samples. **(A)** ROC curve of the training and test datasets. **(B)** Beta coefficients and 95% CI of *HPGD*, *PACS1*, and *TDP2*. **(C)** Kaplan-Meier estimates of overall survival in the training and test datasets according to the linear combination of *HPGD*, *PACS1*, and *TDP2* expression.

Previous studies have shown that the cellular components within the tumor immune microenvironment (TIM) are important regulators of primary tumor progression, organ-specific metastasis, as well as a therapeutic response (52, 53). By using published CRC and healthy individual single-cell RNA-seq profiles, we showed that the CRC tumor micro-environment has a marked surge of immune cells, including both myeloid cells and B cells. This agrees with the finding that tumor-infiltrating cells play a critical role in tumor development and treatment response (53), myeloid cells were also previously found to be abundantly present within the TIM among immune cells (54). Interestingly, when examining the cell type contribution among the cfRNA transcriptome profiles, we detected more intestinal secretory cell signatures in pre-surgical cfRNA than post-surgical cfRNA, which have only been reported in the CRC tumor tissue in the previous study (55).

Three significant cfRNA biomarkers *HPGD*, *PACS1*, and *TDP2* were identified through our comprehensive analysis and qRT-PCR validation experiments. The reduction of *HPGD* promotes the expression of *COX-2*, including Ras-activated protein kinase (MAPK) and extracellular signal-regulated kinase (ERK) (56, 57), phosphoinositide 3-kinase (PI3K)–Akt signaling, epidermal growth factor receptor (EGFR) (58) and Wnt/β-catenin (59). The *PACS-1* promotes chromatin organization by increasing the acetylation of chromatin (60) and its deficiency results in replication stress and gross chromosomal aberrations (56). *TDP2* is a DNA repair enzyme that regulates DNA topology by creating double-strand breakage with free 5' phosphate for re-ligation (61–63). Since *HPGD* is a tumor suppressor gene, as expected, it is freshly expressed in the normal colonic mucosa (59). Interestingly, *HPGD* showed as a significant risk factor in the linear regression model when its expression combined with *PACS1* and *TDP2* expression. Importantly, the model based on the expression of the three genes showed a high AUC (>0.83) of the ROC curve in both training and test datasets. Taken together, linear combination of *HPGD*, *PACS1*, and *TDP2* expression was associated with survival probability, which provides support evidence to potential prognostic biomarkers for CRC. Surprisingly, our research identified a significant decline in *MCF2L-intronic-AS* expression following surgery, which is identical to *MCF2L-AS1* expression. Because potentially interact with the common proteins, *MCF2L-intronic-AS* may play a role in regulating the progression of CRC, which may include promoting cell proliferation, migration, invasion, epithelial-to-mesenchymal transition (EMT), and cell apoptosis (46, 47, 64). There are no studies that have reported the presence of *MCF2L-intronic-AS* in plasma; therefore, further investigations must be conducted to validate the dysregulation pattern of *MCF2L-intronic-AS*.

In the genetic characterization analysis of plasma cfRNA, upregulated expression was observed in rRNA in post-surgical cfRNA samples when compared to pre-surgical cfRNA samples. Two mitochondrially encoded ribosomal RNAs, *MT-RNR2* and *MT-RNR1* are dominant for the increasing expression in the post-surgical cfRNA samples (MT-RNR2: log2FC=2.60, p-value=1.41e-03; MT-RNR1: log2FC=2.73, p-value=1.06e-03), which may play an important role on aiding in the repair of damage during surgery (65). Meanwhile, although no noncoding RNA was observed as a dominant one in the reduction in post-surgical cfRNAs, non-coding RNAs have been reported as drivers of malignant transformation that promote the development of cancers (66). On the other hand, some CRC biomarkers identified from previous studies, such as *CTNNB1* (14), *S100A4* (67), and *EPAS1* (68) were also detected in this study with similar dysregulation patterns, but there was an insufficient sample size in this study that led to these biomarkers being statistically insignificant. While this study shows encouraging results and suggests that the adoption of cfRNA could be useful in a monitoring operation response, future studies with a larger number of replicates per condition should be performed. We acknowledge as a limitation of the present study the small sample size related to cfRNA analysis which did not allow associating the candidate biomarkers to CRC stages as well as investigating on their impact in MRD detection. In conclusion, *HPGD*, *PACS1*, and *TDP2* in CRC plasma samples were demonstrated as potential prognosis biomarkers of CRC, we hope that our results will enable future studies in incorporating cfRNA in the detection, monitoring, and diagnosis of premalignant CRC.

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.

Ethics statement

The studies involving human participants were reviewed and approved by The joint Chinese University of Hong Kong- New Territories East Cluster Clinical Research Ethics Committee (CUHK-NTEC CREC; Ref No: 2019.542). The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

AC-SY, AK-YY, and SCCW designed and supervised the project; NJ performed the bioinformatic analysis; C-MK conducted experiments; XP, WLC, HY-LC, and YKEW performed the qPCR validation; NJ and C-MK wrote the manuscript with input from all authors; SN, WL, and YNW recruited patients and collected samples with consent; HW, HT, AC, WCSC, JC, T-FC, and WST contributed to patient enrolment, provide resource or provide patient samples and scientific advice. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Lim Peng Suan Charitable Trust Research Grant (grant number ZH5G), Research Grants Council HK and HK Innovation and Technology Fund University-Industry Collaborative Programme (grants number UIM/354 and RGCQ71P).

Conflict of interest

AC-SY, AK-YY and NJ were employee of Codex Genetics Limited.

The remaining 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/fonc.2023.1134445/full#supplementary-material

References

1. Sagaert X, Vanstapel A, Verbeek S. Tumor heterogeneity in colorectal cancer: What do we know so far? *Pathobiology* (2018) 85(1-2):72–84. doi: 10.1159/000486721

2. Keum N, Giovannucci E. Global burden of colorectal cancer: Emerging trends, risk factors and prevention strategies. *Nat Rev Gastroenterol Hepatol* (2019) 16 (12):713–32. doi: 10.1038/s41575-019-0189-8

3. Binefa G, Rodríguez-Moranta F, Teule A, Medina-Hayas M. Colorectal cancer: From prevention to personalized medicine. *World J Gastroenterol* (2014) 20(22):6786–808. doi: 10.3748/wjg.v20.i22.6786

4. National Cancer Institute. *Cancer stat facts: Colorectal cancer* (2021). Available at: https://seer.cancer.gov/statfacts/html/colorect.html.

5. Farhat W, Azzaza M, Mizouni A, Ammar H, ben Ltaifa M, Lagha S, et al. Factors predicting recurrence after curative resection for rectal cancer: A 16-year study. *World J Surg Oncol* (2019) 17(1):173. doi: 10.1186/s12957-019-1718-1

6. Bhullar DS, Barriuso J, Mullamitha S, Saunders MP, O'Dwyer ST, Aziz O. Biomarker concordance between primary colorectal cancer and its metastases. *EBioMedicine* (2019) 40:363–74. doi: 10.1016/j.ebiom.2019.01.050

7. Raez LE, Danenberg K, Sumarriva D, Usher J, Sands J, Castrellon A, et al. Using cfrna as a tool to evaluate clinical treatment outcomes in patients with metastatic lung cancers and other tumors. *Cancer Drug Resistance* (2021) 4(4):1061–71. doi: 10.20517/cdr.2021.78

8. Lone SN, Nisar S, Masoodi T, Singh M, Rizwan A, Hashem S, et al. Liquid biopsy: A step closer to transform diagnosis, prognosis and future of cancer treatments. *Mol Cancer* (2022) 21(1):79. doi: 10.1186/s12943-022-01543-7

9. Roskams-Hieter B, Kim HJ, Anur P, Wagner JT, Callahan R, Spiliotopoulos E, et al. Plasma cell-free rna profiling distinguishes cancers from pre-malignant conditions in solid and hematologic malignancies. *NPJ Precis Oncol* (2022) 6(1):28. doi: 10.1038/s41698-022-00270-y

10. Larson MH, Pan W, Kim HJ, Mauntz RE, Stuart SM, Pimentel M, et al. A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nat Commun* (2021) 12(1):2357. doi: 10.1038/s41467-021-22444-1

11. Xu R-h, Wei W, Krawczyk M, Wang W, Luo H, Flagg K, et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat materials* (2017) 16(11):1155–61. doi: 10.1038/nmat4997

12. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun* (2018) 9(1):1–12. doi: 10.1038/s41467-018-07466-6

13. Xue VW, Cheung MT, Chan PT, Luk LLY, Lee VH, Au TC, et al. Non-invasive potential circulating mrna markers for colorectal adenoma using targeted sequencing. *Sci Rep* (2019) 9(1):12943. doi: 10.1038/s41598-019-49445-x

14. Wong SCC, Lo SFE, Cheung MT, Ng KOE, Tse CW, Lai BSP, et al. Quantification of plasma B-catenin mrna in colorectal cancer and adenoma patients. *Clin Cancer Res* (2004) 10(5):1613–7. doi: 10.1158/1078-0432.CCR-1168-3

15. Kopreski MS, Benko FA, Gocke CD. Circulating rna as a tumor marker: Detection of 5t4 mrna in breast and lung cancer patient serum. *Ann New York Acad Sci* (2001) 945(1):172–8. doi: 10.1111/j.1749-6632.2001.tb03882.x

16. Sunakawa Y, Usher JL, Jaimes YS, Tsuji A, Shiozawa M, Watanabe T, et al. Clinical verification of circulating tumor rna (Ctrna) as novel pretreatment predictor and tool for quantitative monitoring of treatment response in metastatic colorectal cancer (Mcrc): A biomarker study of the deeper trial. *J Clin Oncol* (2019) 37(15_suppl): TPS3621–TPS. doi: 10.1200/JCO.2019.37.15_suppl.TPS3621

17. Chen S, Jin Y, Wang S, Xing S, Wu Y, Tao Y, et al. Cancer type classification using plasma cell-free rnas derived from human and microbes. *eLife* (2022) 11:e75181. doi: 10.7554/eLife.75181

18. *Fastqc: A quality control tool for high throughput sequence data*. Babraham Bioinformatics (2010). Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

19. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-One fastq preprocessor. *Bioinformatics* (2018) 34(17):i884–i90. doi: 10.1093/bioinformatics/bty560

20. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. Star: Ultrafast universal rna-seq aligner. *Bioinformatics* (2013) 29(1):15–21. doi: 10.1093/bioinformatics/bts635

21. Anders S, Pyl PT, Huber W. Htseq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* (2015) 31(2):166–9. doi: 10.1093/bioinformatics/btu638

22. *Gene/Transcript biotypes in gencode & ensembl*. Gencode (2022). Available at: https://www.gencodegenes.org/pages/biotypes.html.

23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with Deseq2. *Genome Biol* (2014) 15(12):550. doi: 10.1186/s13059-014-0550-8

24. Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, et al. Reactome pathway analysis: A high-performance in-memory approach. *BMC Bioinf* (2017) 18(1):142. doi: 10.1186/s12859-017-1559-2

25. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. David: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* (2022) 50(W1):W216–21. doi: 10.1093/nar/gkac194

26. *Harmonized cancer datasets genomic data commons data portal*. National Cancer Institute (2022). Available at: https://portal.gdc.cancer.gov.

27. Lee HO, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat Genet* (2020) 52(6):594–603. doi: 10.1038/s41588-020-0636-z

28. Vorperian SK, Moufarrej MN, Quake SR. Cell types of origin of the cell-free transcriptome. *Nat Biotechnol* (2022) 40(6):855–61. doi: 10.1038/s41587-021-01188-9

29. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* (2019) 37(7):773–82. doi: 10.1038/s41587-019-0114-2

30. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nat Biotechnol* (2015) 33(3):290–5. doi: 10.1038/nbt.3122

31. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with Hisat2 and hisat-genotype. *Nat Biotechnol* (2019) 37 (8):907–15. doi: 10.1038/s41587-019-0201-4

32. Pertea G, Pertea M. Gff utilities: Gffread and gffcompare. *F1000Res* (2020) 9:304. doi: 10.12688/f1000research.23297.2

33. Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol* (2015) 33(3):243–6. doi: 10.1038/nbt.3172

34. Lesurf R, Cotto KC, Wang G, Griffith M, Kasaian K, Jones SJ, et al. Oreganno 3.0: A community-driven resource for curated regulatory annotation. *Nucleic Acids Res* (2016) 44(D1):D126–32. doi: 10.1093/nar/gkv1203

35. Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, et al. Cpc2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res* (2017) 45(W1):W12–w6. doi: 10.1093/nar/gkx428

36. Ke L, Yang DC, Wang Y, Ding Y, Gao G. Annolnc2: The one-stop portal to systematically annotate novel lncrnas for human and mouse. *Nucleic Acids Res* (2020) 48(W1):W230–w8. doi: 10.1093/nar/gkaa368

37. Lu Q, Ren S, Lu M, Zhang Y, Zhu D, Zhang X, et al. Computational prediction of associations between long non-coding rnas and proteins. *BMC Genomics* (2013) 14:651. doi: 10.1186/1471-2164-14-651

38. Zhao W, Song M, Zhang J, Kuerban M, Wang H. Combined identification of long non-coding rna Ccat1 and hotair in serum as an effective screening for colorectal carcinoma. *Int J Clin Exp Pathol* (2015) 8(11):14131–40.

39. Wang J, Song YX, Ma B, Wang JJ, Sun JX, Chen XW, et al. Regulatory roles of non-coding rnas in colorectal cancer. *Int J Mol Sci* (2015) 16(8):19886–919. doi: 10.3390/ijms160819886

40. Cheung KWE, S-yR C, Lee LTC, Lee NLE, Tsang HF, Cheng YT, et al. The potential of circulating cell free rna as a biomarker in cancer. *Expert Rev Mol Diagnostics* (2019) 19(7):579–90. doi: 10.1080/14737159.2019.1633307

41. Vong JSL, Ji L, Heung MMS, Cheng SH, Wong J, Lai PBS, et al. Single cell and plasma rna sequencing for rna liquid biopsy for hepatocellular carcinoma. *Clin Chem* (2021) 67(11):1492–502. doi: 10.1093/clinchem/hvab116

42. Chaib M, Chauhan SC, Makowski L. Friend or foe? recent strategies to target myeloid cells in cancer. *Front Cell Dev Biol* (2020) 8:351. doi: 10.3389/fcell.2020.00351

43. Yuen GJ, Demissie E, Pillai S. B lymphocytes and cancer: A love-hate relationship. *Trends Cancer* (2016) 2(12):747–57. doi: 10.1016/j.trecan.2016.10.010

44. Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR, Salzman J, et al. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Sci (New York NY)* (2022) 376(6594):eabl4896. doi: 10.1126/science.abl4896

45. Baraniskin A, Nöpel-Dünnebacke S, Ahrens M, Jensen SG, Zöllner H, Maghnouj A, et al. Circulating U2 small nuclear rna fragments as a novel diagnostic biomarker for pancreatic and colorectal adenocarcinoma. *Int J Cancer* (2013) 132(2): E48–57. doi: 10.1002/ijc.27791

46. Huang FK, Zheng CY, Huang LK, Lin CQ, Zhou JF, Wang JX. Long non-coding rna Mcf2l-As1 promotes the aggressiveness of colorectal cancer by sponging mir-874-3p and thereby up-regulating Ccne1. *J Gene Med* (2021) 23(1):e3285. doi: 10.1002/jgm.3285

47. Zhang Z, Yang W, Li N, Chen X, Ma F, Yang J, et al. Lncrna Mcf2l-As1 aggravates proliferation, invasion and glycolysis of colorectal cancer cells *Via* the crosstalk with mir-874-3p/Foxm1 signaling axis. *Carcinogenesis* (2021) 42(2):263–71. doi: 10.1093/carcin/bgaa093

48. Backlund MG, Mann JR, Holla VR, Buchanan FG, Tai HH, Musiek ES, et al. 15-hydroxyprostaglandin dehydrogenase is down-regulated in colorectal cancer. *J Biol Chem* (2005) 280(5):3217–23. doi: 10.1074/jbc.M411221200

49. Myung SJ, Rerko RM, Yan M, Platzer P, Guda K, Dotson A, et al. 15-hydroxyprostaglandin dehydrogenase is an in vivo suppressor of colon tumorigenesis. *Proc Natl Acad Sci U.S.A.* (2006) 103(32):12098–102. doi: 10.1073/pnas.0603235103

50. Larribère L, Martens UM. Advantages and challenges of using ctdna ngs to assess the presence of minimal residual disease (Mrd) in solid tumors. *Cancers (Basel)* (2021) 13(22):5698. doi: 10.3390/cancers13225698

51. Schraa SJ, van Rooijen KL, Koopman M, Vink GR, Fijneman RJA. Cell-free circulating (Tumor) DNA before surgery as a prognostic factor in non-metastatic colorectal cancer: A systematic review. *Cancers (Basel)* (2022) 14(9):2218. doi: 10.3390/cancers14092218

52. Ma BB, Lui VW, Poon FF, Wong SC, To KF, Wong E, et al. Preclinical activity of gefitinib in non-keratinizing nasopharyngeal carcinoma cell lines and biomarkers of response. *Invest New Drugs* (2010) 28(3):326–33. doi: 10.1007/s10637-009-9316-7

53. Wang W, Zhong Y, Zhuang Z, Xie J, Lu Y, Huang C, et al. Multiregion single-cell sequencing reveals the transcriptional landscape of the immune microenvironment of colorectal cancer. *Clin Trans Med* (2021) 11(1):e253. doi: 10.1002/ctm2.253

54. Schupp J, Krebs FK, Zimmer N, Trzeciak E, Schuppan D, Tuettenberg A. Targeting myeloid cells in the tumor sustaining microenvironment. *Cell Immunol* (2019) 343:103713. doi: 10.1016/j.cellimm.2017.10.013

55. Zhang GL, Pan LL, Huang T, Wang JH. The transcriptome difference between colorectal tumor and normal tissues revealed by single-cell sequencing. *J Cancer* (2019) 10(23):5883–90. doi: 10.7150/jca.32267

56. Mani C, Tripathi K, Luan S, Clark DW, Andrews JF, Vindigni A, et al. The multifunctional protein pacs-1 is required for Hdac2- and Hdac3-dependent chromatin maturation and genomic stability. *Oncogene* (2020) 39(12):2583–96. doi: 10.1038/s41388-020-1167-x

57. Wan L, Molloy SS, Thomas L, Liu G, Xiang Y, Rybak SL, et al. Pacs-1 defines a novel gene family of cytosolic sorting proteins required for trans-golgi network localization. *Cell* (1998) 94(2):205–16. doi: 10.1016/s0092-8674(00)81420-8

58. Brasacchio D, Busuttil R, Noori T, Johnstone R, Boussioutas A, Trapani J. Down-regulation of a pro-apoptotic pathway regulated by Pcaf/Ada3 in early stage gastric cancer. *Cell Death Dis* (2018) 9:442. doi: 10.1038/s41419-018-0470-8

59. Castellone MD, Teramoto H, Williams BO, Druey KM, Gutkind JS. Prostaglandin E2 promotes colon cancer cell growth through a gs-Axin-Beta-Catenin signaling axis. *Sci (New York NY)* (2005) 310(5753):1504–10. doi: 10.1126/science.1116221

60. Wells CE, Bhaskara S, Stengel KR, Zhao Y, Sirbu B, Chagot B, et al. Inhibition of histone deacetylase 3 causes replication stress in cutaneous T cell lymphoma. *PloS One* (2013) 8(7):e68915. doi: 10.1371/journal.pone.0068915

61. Shu J, Cui D, Ma Y, Xiong X, Sun Y, Zhao Y. Scfβ-Trcp-Mediated degradation of Top2β promotes cancer cell survival in response to chemotherapeutic drugs targeting topoisomerase ii. *Oncogenesis* (2020) 9(2):8. doi: 10.1038/s41389-020-0196-1

62. Kiselev E, Ravji A, Kankanala J, Xie J, Wang Z, Pommier Y. Novel deazaflavin tyrosyl-DNA phosphodiesterase 2 (Tdp2) inhibitors. *DNA Repair* (2020) 85:102747. doi: 10.1016/j.dnarep.2019.102747

63. Schellenberg MJ, Perera L, Strom CN, Waters CA, Monian B, Appel CD, et al. Reversal of DNA damage induced topoisomerase 2 DNA–protein crosslinks by Tdp2. *Nucleic Acids Res* (2016) 44(8):3829–44. doi: 10.1093/nar/gkw228

64. Kong W, Li H, Xie L, Cui G, Gu W, Zhang H, et al. Lncrna Mcf2l-As1 aggravates the malignant development of colorectal cancer *Via* targeting mir-105-5p/Rab22a axis. *BMC Cancer* (2021) 21(1):1069. doi: 10.1186/s12885-021-08668-w

65. Druzhyna NM, Wilson GL, LeDoux SP. Mitochondrial DNA repair in aging and disease. *Mech Ageing Dev* (2008) 129(7-8):383–90. doi: 10.1016/j.mad.2008.03.002

66. Slack FJ, Chinnaiyan AM. The role of non-coding rnas in oncology. *Cell* (2019) 179(5):1033–55. doi: 10.1016/j.cell.2019.10.017

67. Stein U, Burock S, Herrmann P, Wendler I, Niederstrasser M, Wernecke KD, et al. Diagnostic and prognostic value of metastasis inducer S100a4 transcripts in plasma of colon, rectal, and gastric cancer patients. *J Mol Diagn* (2011) 13(2):189–98. doi: 10.1016/j.jmoldx.2010.10.002

68. Collado M, Garcia V, Garcia JM, Alonso I, Lombardia L, Diaz-Uriarte R, et al. Genomic profiling of circulating plasma rna for the analysis of cancer. *Clin Chem* (2007) 53(10):1860–3. doi: 10.1373/clinchem.2007.089201

TYPE Original Research
PUBLISHED 12 April 2023
DOI 10.3389/fgene.2023.1173159

Identification of novel immune ferroptosis-related genes associated with clinical and prognostic features in breast cancer

Zhenlan Xie[1], Jialin Li[2], Chen Liu[1], Tie Zhao[1] and Yixiang Xing[1]*

[1]Department of Pathology, Tongling People's Hospital, Tongling, China, [2]Tongling Vocational and Technical College, Tongling, China

Check for updates

OPEN ACCESS

EDITED BY
Junlin Xu,
Hunan University, China

REVIEWED BY
Lihong Peng,
Hunan University of Technology, China
Lijuan Zhu,
Shijiazhuang Tiedao University, China

*CORRESPONDENCE
Yixiang Xing,
✉ xyq.35@163.com

SPECIALTY SECTION
This article was submitted to RNA,
a section of the journal
Frontiers in Genetics

RECEIVED 24 February 2023
ACCEPTED 28 March 2023
PUBLISHED 12 April 2023

CITATION
Xie Z, Li J, Liu C, Zhao T and Xing Y (2023), Identification of novel immune ferroptosis-related genes associated with clinical and prognostic features in breast cancer.
Front. Genet. 14:1173159.
doi: 10.3389/fgene.2023.1173159

Introduction: Breast cancer is the most common form of cancer among women, it is critical to identify potential targets and prognostic biomarkers. Ferroptosis combined with immunity shows a pivotal role in a variety of tumors, which provides new opportunities to detect and treat breast cancer.

Methods: Our first step was to combine multiple datasets to search for immune ferroptosis-related mRNAs. In the next step, risk signatures were created using Least Absolute Shrinkage and Selection Operator (LASSO). After that, based on the results of the multivariate Cox analysis, we created a prognostic nomogram and validated the model's accuracy. Finally, functional enrichment analysis, single sample gene set enrichment analysis (ssGSEA), immunity and drug sensitivity correlation analysis were performed to explore the possible mechanisms by which these immune ferroptosis associated mRNAs affect BRCA survival.

Results: An immune ferroptosis signature (IFRSig) consisting of 5 mRNAs was constructed and showed excellent predictability in the training and validation cohorts. A correlation analysis revealed that clinical characteristics were closely related to risk characteristics. Our nomogram model, which we created by combining risk characteristics and clinical parameters, was proven to be accurate at predicting BRCA prognosis. Further, we divided patients into lowrisk and high-risk groups based on the expression of the model-related genes. Compared with low-risk group, high-risk group showed lower levels of immune cell infiltration, immune-related functions, and immune checkpoints molecules, which may associate with the poor prognosis.

Discussion: The IFRSig could be used to predict overall survival (OS) and treatment response in BRCA patients and could be viewed as an independent prognostic factor. The findings in this study shed light on the role of immune ferroptosis in the progression of BRCA.

KEYWORDS

immune, breast cancer, prognosis, biomarkers, ferroptosis

1 Introduction

With the highest incidence of all female malignant tumors worldwide, BRCA is the most prevalent malignant tumor in women (Siegel et al., 2021). Although great progress has been made in the therapeutic effect of BRCA (Siegel et al., 2023), sadly, there are still no reliable diagnostic tools or markers for determining the prognosis of BRCA patients (Islam et al., 2020; Pupa et al., 2021; Sindhu et al., 2021). Until now, tumor lymph node metastasis (TNM) stage has been used to predict BRCA prognosis and treatment response. However, due to tumor heterogeneity, BRCA patients with the same TNM stage showed different prognosis and treatment response. Therefore, it is important to combine other useful indicators to predict prognosis and treatment response.

As opposed to apoptosis, necrosis, and autophagy, ferroptosis was a type of programmed cell death dependent on iron (Hadian and Stockwell, 2020; Fardi et al., 2021; Jiang et al., 2021). The classical mode of regulation of ferroptosis was through the neutralization of lipid peroxides by glutathione peroxidase 4 (GPX-4) (Yang and Stockwell, 2016; Ding et al., 2020; He et al., 2020). There was growing evidence that ferroptosis causes hypersensitivity reactions in cancer cells with a higher degree of malignancy, particularly those with intrinsic or acquired drug resistance (Hangauer et al., 2017; Viswanathan et al., 2017). In addition, ferroptosis influences the effectiveness of cancer immunotherapy and was associated with T cell-mediated antitumor immunity (Wang et al., 2019). Additionally, it had been demonstrated that immune modulation of the tumor microenvironment (TME) could facilitate ferroptosis, which in turn increases the immunogenicity of the TME, enhancing the immune modulation response (Zhang et al., 2019). It was anticipated that immunotherapy will had synergistic effects through ferroptosis, promoting tumor control, in combination with ferroptosis-promoting modalities like radiation therapy and targeted therapy (Lang et al., 2019; Chen et al., 2021). There was a close relationship between tumor cells, the immune microenvironment, and ferroptosis (Lang et al., 2019; Jiang et al., 2021). In addition, studies had found that ferroptosis intervention could effectively improve immunosuppression (Gao et al., 2015; Sun et al., 2016; Alavian and Ghasemi, 2021). In conclusion, the important role of immunity and ferroptosis might provide a new direction for predicting prognosis and treatment response of breast cancer.

The goal of this research was to develop new survival predictive risk signatures and to explore the prognostic role of immune ferroptosis-related mRNAs in BRCA. Firstly, we combined multiple datasets to screen mRNAs associated with prognosis. The risk features for BRCA prognosis prediction were then constructed by LASSO regression analysis. At the same time, the total samples were divided into training cohort and validation cohort according to the ratio of 1: 1. Then, by combining this feature with other clinical parameters, a nomogram was created to predict 1-, 3-, and 5-year survival. Ultimately, we explored the relationship between risk characteristics and underlying biological function, immunity, and drug susceptibility.

2 Materials and methods

2.1 Transcriptome data acquisition and model building

In this research, we downloaded the transcriptome Fregments Per Kilobaseper Million (FPKM) of breast cancer patients from the TCGA database (https://portal.gdc.cancer.gov/). The RNAseq data in FPKM format was converted into transcripts per millionreads (TPM) format and log2 conversion was performed. Transcriptome data was organized and ENSG numbers were converted to symbolic IDs. The research was carried out in accordance with the Helsinki Declaration (revised 2013).

The ImmPort database (https://www.immport.org./home) and the GeneCard database (https://www.genecards.org/) were used to obtain 17,500 human immune-related genes (IRGs). A total of 398 ferroptosis-related genes (FRGs) were downloaded through the FerrDb database (http://www.datjar.com) and literature (Song et al., 2021). Two gene sets were crossed with differentially expressed genes to obtain co-expressed genes (IFR-DEG), and the cutoff conditions were set as log2 fold change (logFC) < 1, *p*-value <0. 05. Then, univariate Cox regression analysis was performed, and the total samples were divided into training cohort and validation cohort according to the ratio of 1: 1. The training cohort builds a risk model based on LASSO-Cox regression analysis. The formula for calculating the risk score was as follows: Risk score = βgene1×exprgene1+βgene2×exprgene2+.+βgenen×exprgenen. At the same time, to reduce the dimensionality of the nomogram, we used an unsupervised learning algorithm called principal component analysis (PCA), which allowed us to visualize the spatial distribution of samples.

2.2 Gene correlation, gene network and functional enrichment analysis

Gene correlation analysis was performed by Spearman analysis and visualized with the ggplot2 package. Model-related genes were submitted through GeneMANIA (http://www.genemania.org), which analyzed and displayed genes that perform similar functions—representing protein expression and inheritance in the network. Genes were enriched by Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways to investigate the potential biological functions of interacting proteins and model gene co-expression in pan-cancer, GO enrichment analysis including molecular function (MF), cellular component (CC) and biological process (BP). Both GO and KEGG analyses were performed by the R package ClusterProfiler. Then high and low risk differential genes were also analyzed by GO and KEGG.

2.3 Model validation

We divided patients into high-risk and low-risk groups, and then we generated heatmaps associated with prognosis based on the median risk score. To determine differences in survival between

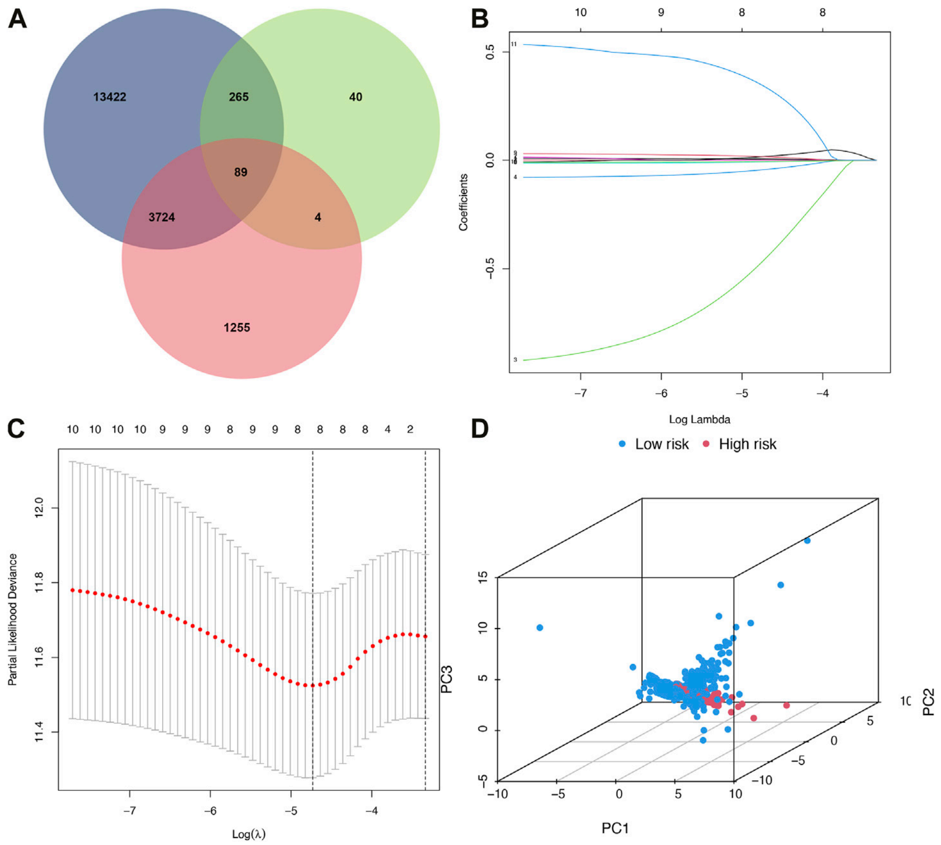


FIGURE 1
Build a risk model. **(A)** Venn diagram. Blue represented immunity genes, green represented ferroptosis genes, and red represented differential genes. **(B)** Distribution of LASSO regression coefficients for crossed genes. **(C)** LASSO deviation profile of crossed genes. **(D)** PCA plot of high and low-risk group. PCA, principal component analysis.

high- and low-risk groups, we plotted Kaplan-Meier survival curves, distributions of survival status, and distributions of risk scores. Finally, the predictions of the risk scoring model were further validated by applying the "timeROC" package to plot the ROC curves of the training and validation groups.

2.4 Independent prognostic analysis and nomogram construction

Nomograms were constructed by combining relevant clinical factors and risk scores obtained with risk scores (we used the R packages: "rms", "foreign" and "survival"). An evaluation of the model's discriminative ability was then carried out by drawing a calibration curve.

2.5 The relationship between risk score and immune cell infiltration

We calculated immune stages using single-sample gene set enrichment analysis (ssGSEA) (He et al., 2018). In exploring the relationship between risk score values and immune-infiltrating cells, we used Spearman's rank correlation analysis.

2.6 Immune microenvironment, immune checkpoints, immune escape

A stromal score, an immune score, an estimated score, and a tumor purity were calculated using transcriptome profiles from UCECs. In the high-evolution and low-evolution groups of hub genes, we compared stroma scores, immune scores, estimated scores, and tumor purity using Limma and ggpubr packages. In addition, tumor immune escape mechanisms in different risk groups were analyzed using the TIDE algorithm.

2.7 Gene mutation analysis

In the gene mutation analysis, the number and quality of gene mutations in two subgroups (Maftools package) of BRCA patients were analyzed. In addition, we also analyzed the relationship between tumor mutational burden (TMB) and risk score subgroups using Student's t-test.

2.8 Predicting response to chemotherapy

To elucidate the role of signatures in clinical treatment, IC50 values of commonly used chemotherapeutics were evaluated

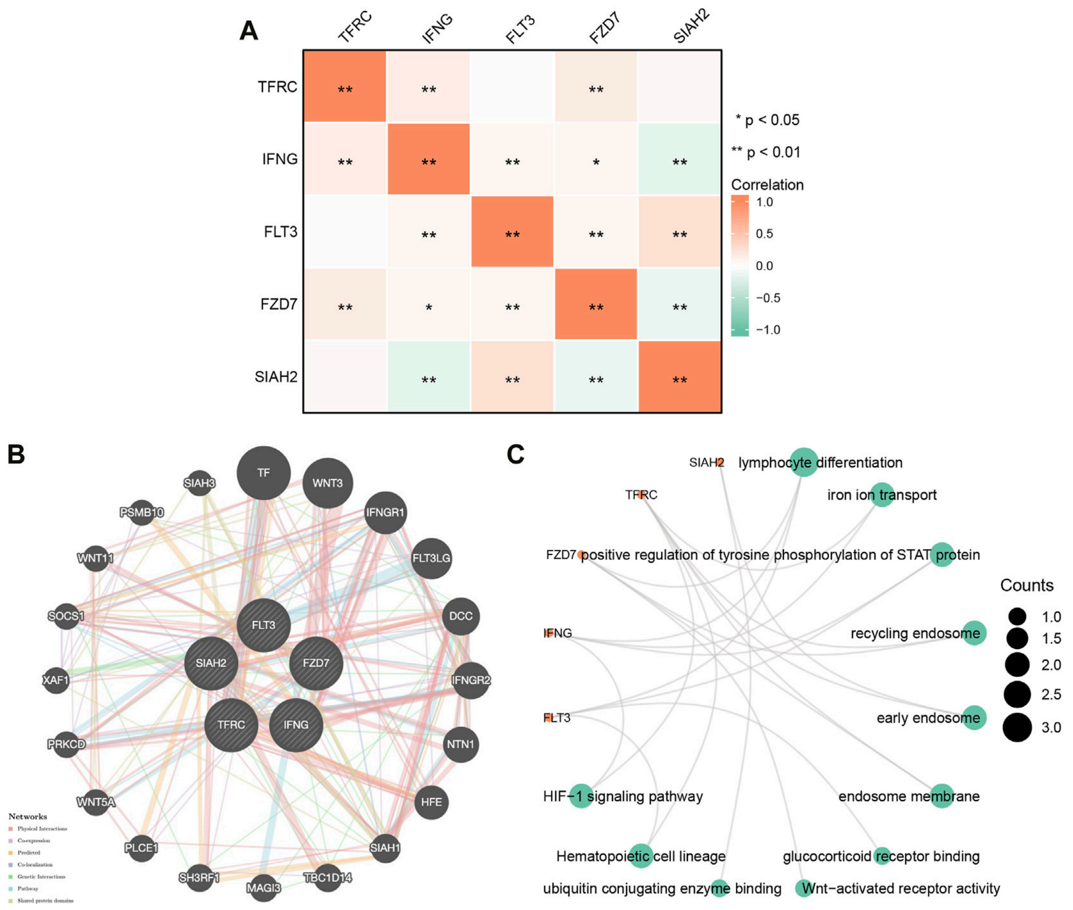


FIGURE 2
(A) Gene correlation network diagram of prognostic model. **(B)** Model -related gene network plotted using GeneMANIA. **(C)** Model gene enrichment analysis in pan-cancer: GO and KEGG.

using high-throughput sequencing data of BRCA in TCGA. In this study, the Wilcoxon signed-rank test was used to compare the differences between the two groups, and pRRophetic and ggplot2 were used for the visualization of the results.

2.9 Statistical analysis

R software (version 4.1.2) was used for statistical analysis. For data processing, the Perl programming language is used. Prognostic significance was determined using multivariate Cox regression analysis. PCA was also performed using R's ggplot2 package. The survival difference between the two groups was analyzed by Kaplan-Meier curve and logrank test was used. Gene correlations, risk scores and correlations between immune cells and immune genes were analyzed using Spearman's correlation coefficient test. When $p < 0.05$, the difference was statistically significant.

3 Results

3.1 Construction of a prognostic risk model for differentially expressed genes related to immune ferroptosis

89 co-expressed genes were discovered by combining 17,500 immune-related genes, 398 ferroptosis-related genes, and 5072 BRCA differentially expressed genes (Figure 1A). A predictive model of immune ferroptosis-related risk was then constructed using lasso regression (Figure 1B). The risk score formula was: riskscore= (0.008*TFRC) + (−1.042*IFNG) + (−0.064*FLT3) + (−0.016*FZD7) + (−0.009 *SIAH2) (Figure 1C). Patients were divided into high- and low-risk groups based on the median risk score (50%). The results of PCA validated the differential expression of high- and low-risk groups in BRCA patients (Figure 1D).

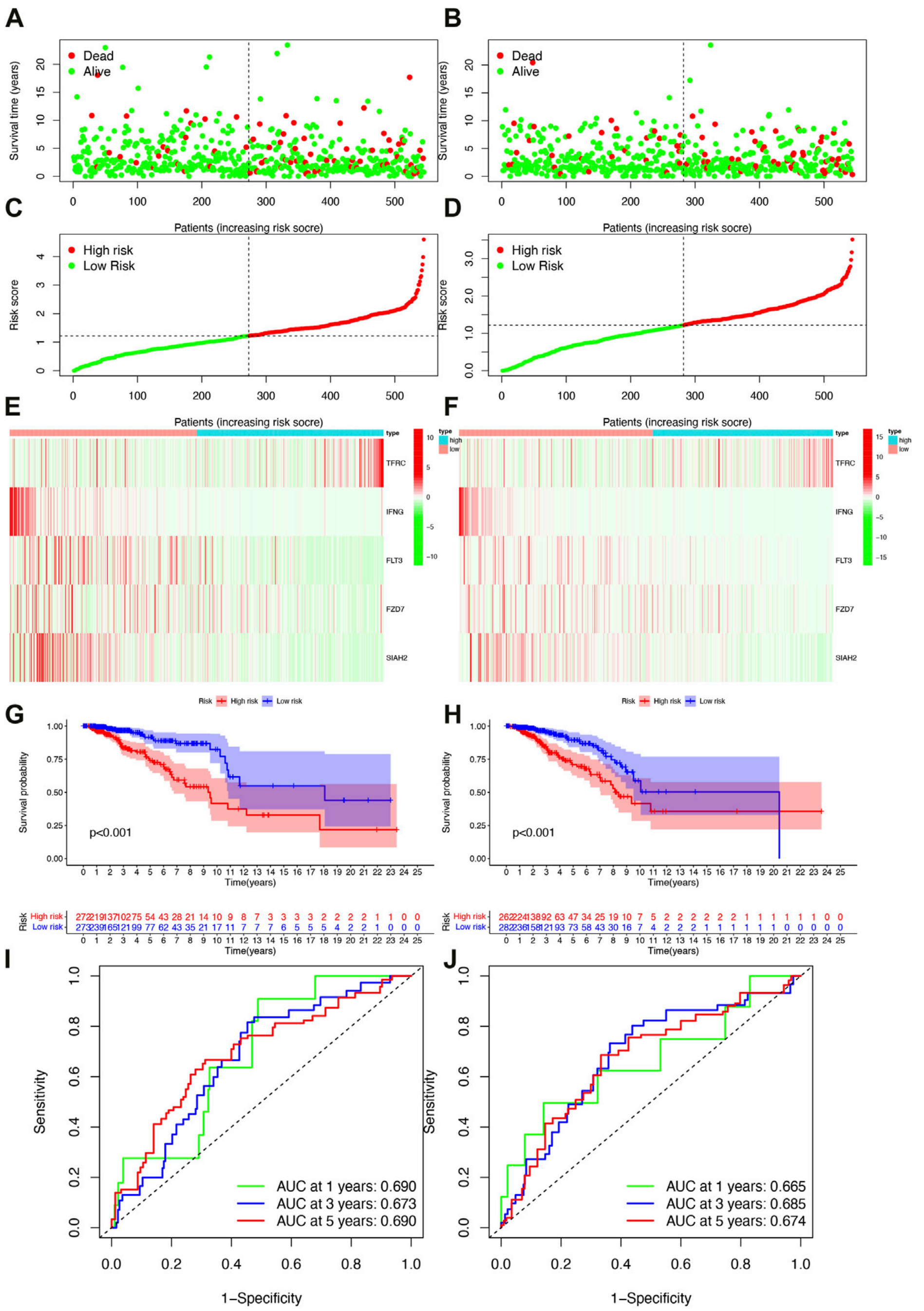


FIGURE 3
Survival analysis of patients in both the training and validation cohorts. **(A)** Distribution plots of survival times in the training cohort. **(B)** Distribution plots of survival times in the validation cohorts. **(C)** Scatter plots of risk scoresin the training cohort. **(D)** Scatter plots of risk scores in the validation cohorts. **(E)** Gene expression levels in the training cohort. **(F)** Gene expression levels in the validation cohorts. **(G)** Overall survival (OS) in the training cohort. **(H)** Overall survival (OS) in the validation cohorts. **(I)** Time-dependent ROC curves in the training cohort. **(J)** Time-dependent ROC curves in the validation cohorts.

3.2 Model gene correlation and functional enrichment analysis

To explore potential relationships of model genes, we examined correlations between Model-related genes using Spearman correlation analysis. As shown in Figure 2A, FLT3 negatively correlated with SIAH2, while TFRC positively correlated with IFNG and FZD7; SIAH2 positively correlated with FZD7 and SIAH2; TFRC positively correlated with FZD7 and SIAH2; FLT3 positively correlated with FZD7 and SIAH2; FZD7 was negatively correlated with SIAH2.

We constructed gene-gene networks through GeneMANIA to explore gene interactions. Figure 2B shows 20 nodes around the central node of the Model-related genes, which were genes related to

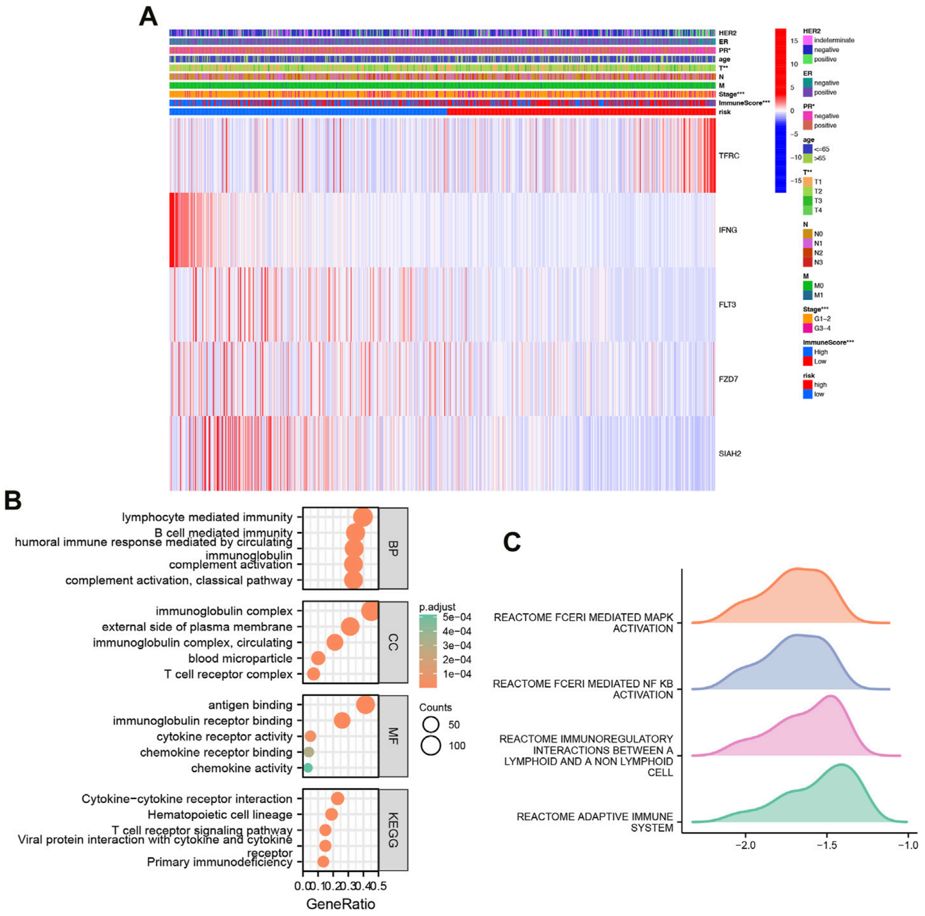


FIGURE 4
Heatmap and GO/KEGG pathway enrichment analysis. **(A)** Clinically relevant heatmap. A heatmap based on data on the clinicopathological characteristics of the patients was created based on the risk characteristics associated with prognosis. **p* < 0.05, ***p* < 0.01, ****p* < 0.001. **(B)** Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis of high and low risk differential genes. **(C)** GSEA analysis of high and low risk differential genes.

the genes model based on physical interactions, co-expression, predictions, co-localization, genetic interactions, pathways, and shared domains. Among them, TF, WNT3, IFNGR1, and FLT3LG were ranked in the top. Regarding model-related genes GO and KEGG enrichment analysis, as shown in Figure 2C, in BP, the regulation of lymphocyte differentiation, ironion transport, positive regulation of tyrosine phosphorylation of STAT rotein was dominant. MF was significantly enriched in glucocorticoid eceptor binding, Wnt-activated receptor activity, ubiquitin conjugating genzyme binding. In CC, they were mainly located in recycling endosome, early endosome, endosome membrane. A KEGG enrichment analysis indicated that model-related genes were associated with the hemotopoietic cell lineage, significantly associated with the HIF-1 signaling pathway.

3Validation of a prognostic risk model for differentially expressed genes related to immune ferroptosis.

Based on the median risk scores of the training and validation cohorts and the test cohorts, all patients were divided into high- or low-risk groups with each group accounting for 50%. As the risk score increased, so did the number of patient deaths (Figures 3A–D). The risk model-related genes expression level between high- and low-risk groups were shown in Figures 3E, F. In both trainning and validation cohorts, OS was significantly different (Figures 3G, H, $p < 0.001$). Low risk patients had better clinical outcomes than high risk patients in each cohort, which was consistent with both groups' results. The survival time ability of IFRSig was assessed using a time-dependent ROC curve. The areas under the curve (AUC) at 1, 3, and 5 years were 0.690, 0.673, and 0.690 for the training cohort (Figure 3I) and 0.665, 0.685, and 0.674 for the validation cohort (Figure 3J), respectively. The results of all studies suggest that IFRSig can accurately predict OS.

3.3 Heatmap and GO/KEGG pathway enrichment analysis

Based on clinical features, we created heatmap to compare the expression relationship of prognostic model-related genes between high-risk and low-risk subgroups, and the status of HER2, ER, PR, age, T, N, M, stage, immune score were shown as patient annotations (Figure 4A).

Classification analysis revealed that GO: BP was mainly concentrated in classical pathway, humoral immune response mediated by circulating immunoglobulin, complement activation,

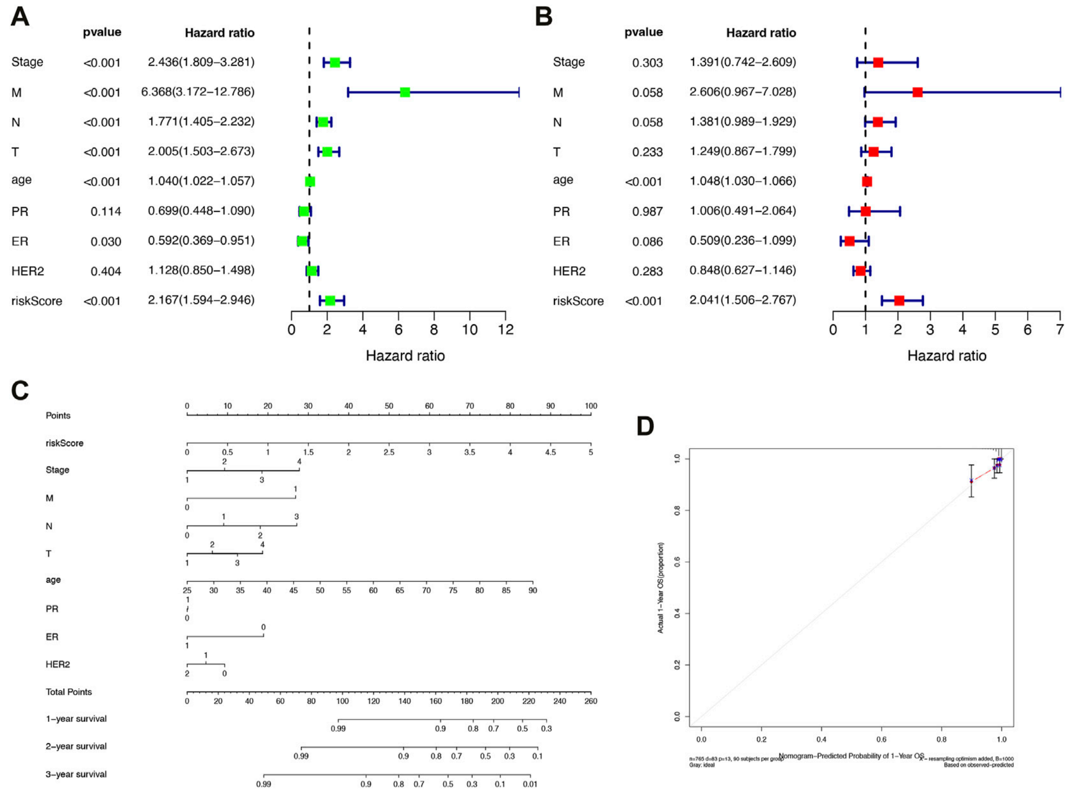


FIGURE 5
Construction of independent prognostic factors and nomogram. **(A)** Univariate Cox regression analysis. **(B)** Multivariate Cox regression analysis. **(C)** Survival nomogram based on the total TCGA cohort. **(D)** Calibration curves for predicting 1, 3, and 5-year survival of BRCA patients in the TCGA cohort. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

B cell mediated immunity, lymphocyte mediated immunity; CC was mainly concentrated in immunoglobulin complex, immunoglobulin complex, circulating, external side of plasma membrane, blood microparticle, T cell receptor complex; MF was mainly concentrated in immunoglobulin receptor binding, cytokine receptor activity, chemokine receptor binding, chemokine activity. Importantly, KEGG was mainly enriched in Hematopoietic cell lineage, Primary immunodeficiency, Cytokine-Cytokine receptor interaction, Viral protein interaction with cytokine and cytokine receptor, T cell receptor signaling Pathway (Figure 4B).

Further GSEA, we found that high and low risk were mainly enriched in REACTOME_FCERI_MEDIATED_MAPK_ACTIVATION, REACTOME_FCERI_MEDIATED_NF_KB_ACTIVATION, REACTOME_IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYMPHOID_CELL, REACTOME_ADAPTIVE_IMMUNE_SYSTEM (Figure 4C).

3.4 Independent prognostic factors and nomogram construction

In the TCGA cohort, univariate Cox regression analysis showed stage, M, N, T, age, ER, and risk score were significantly associated with OS; while multivariate Cox regression analysis showed that age ($p < 0.001$) and risk score ($p < 0.001$) were significantly associated with OS (Figures 5A, B). The results showed that IFRSig was an independent prognostic factor for BRCA.

We constructed a nomogram based on risk scores and other clinicopathological covariates for calculating individualized cancer risk scores (Figure 5C). According to calibration plots, the prognostic nomogram for 1-, 3-, and 5-year OS was in agreement with the diagonal lines (Figure 5D). The outcomes demonstrated that the nomogram created by IFRSig has a good level of prognostic accuracy for BRCA patients.

4 Immune characteristics

Tumor immune cell compositions played a major role in response to immunotherapy but the heterogeneity and dynamics of immune infiltrates in human cancer lesions remained poorly characterized. In BRCA samples, we assessed the immune infiltrating profile of immune infiltrating cells to better understand the complex crosstalk between IFRSig and immune signatures (Figure 6C). Moreover, we investigated the relationship between immune infiltrating cells and immune function as well as IFRSig, and immune infiltrating cells and immune function were found to be lower in high-risk individuals than in low-risk individuals (Figures 6A, B).

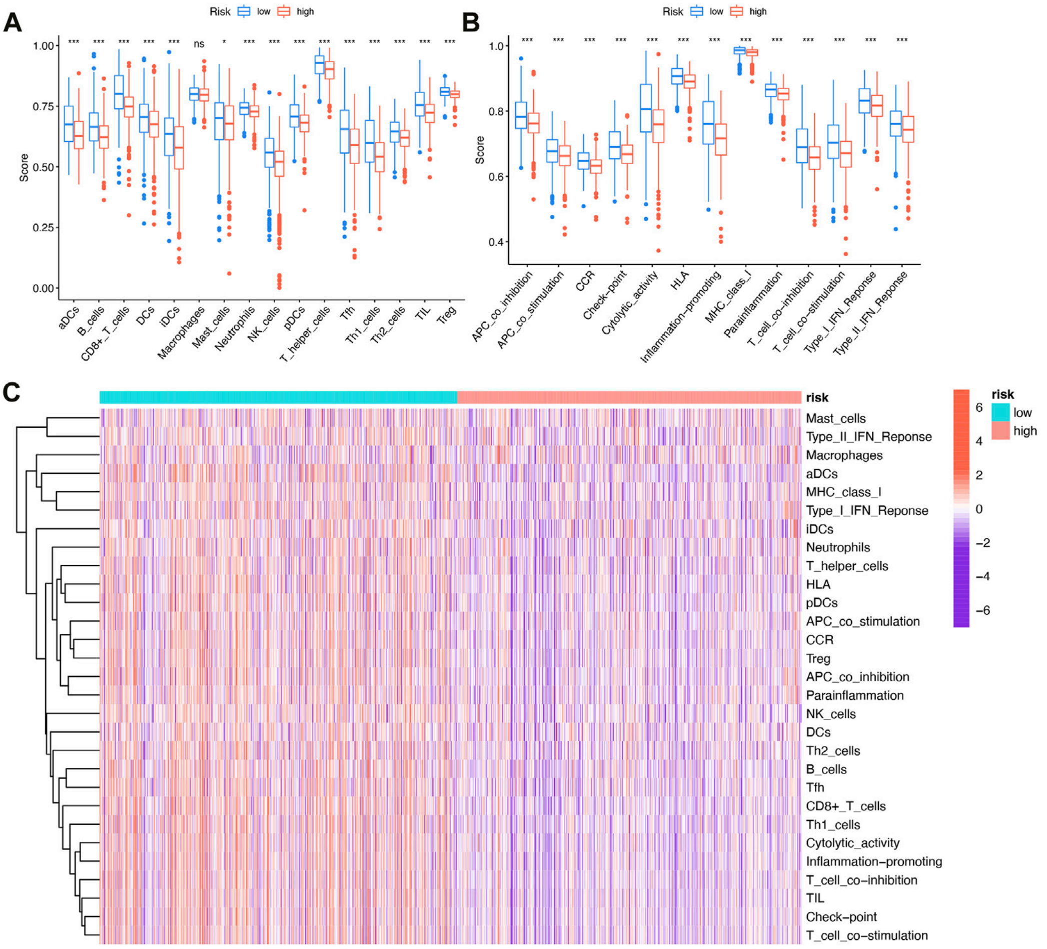


FIGURE 6
Relationship with immune infiltration **(A)** Boxplot of association between IFRSig and immune cell lineage; **(B)** Boxplot of association between IFRSig and immune function; ANOVA used as significance test, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. **(C)** Immune correlation heatmap.

4.1 The immune microenvironment, immune checkpoints, immune escape

Tumor microenvironments, as their name suggests, contained the necessary conditions for tumor cells to proliferate and metastasize. Tumor progression was influenced by immune cells, tumor cells, stromal cells, as well as a variety of active molecules. Figure 7A showed that high-risk patients have lower immune and ESTIMATE (Estimation of STromal and Immune cells in MAlignant Tumour tissues using Expression data) scores.

Our study compared the expression values of immune checkpoints molecules in patients with different IFRSigs. As shown in Figure 7B, the bar plot shows that the expression of immune checkpoints molecules were significantly lower in the high-risk score group than in the low-risk score group, except NRP1 and CD276. These findings imply that high-risk group may not benefit from anti-PD1/PD-L1/CTLA ICI immunotherapy, but from anti-NRP1/CD276.

As well, the Tumor Immune Dysfunction and Exclusion (TIDE) algorithm could predict how immune checkpoint inhibitors would react with different subgroups. Results showed that high-risk group dysfunction and TIDE scores were lower, and exclusion was higher than low-risk group exclusions (Figures 7E–H).

4.2 The association of immune ferroptosis-related mRNA signatures withTMB

It was reported that in many cancer types, including bread cancer, patients with higher tumor burden mutations (TMB) had lower survival rates. On the contrary, patients treated by ICI, with higher TMB generally associated with longer survival (Godenick, 1995).

Accordingly, we speculated that TMB might have a non-negligible relationship between prognosis risk score and TMB. Therefore, we analyzed and displayed the distribution of genetic mutations among high-risk and low-risk score subtypes. A total of 84.19% of low-risk BRCA samples had genetic mutations (Figure 8A), while 84.43% were mutated in the high-risk group (Figure 8B), indicating that samples from the high-risk group had a higher probability of gene mutation. A comprehensive landscape of somatic variation showed mutational patterns and clinical features

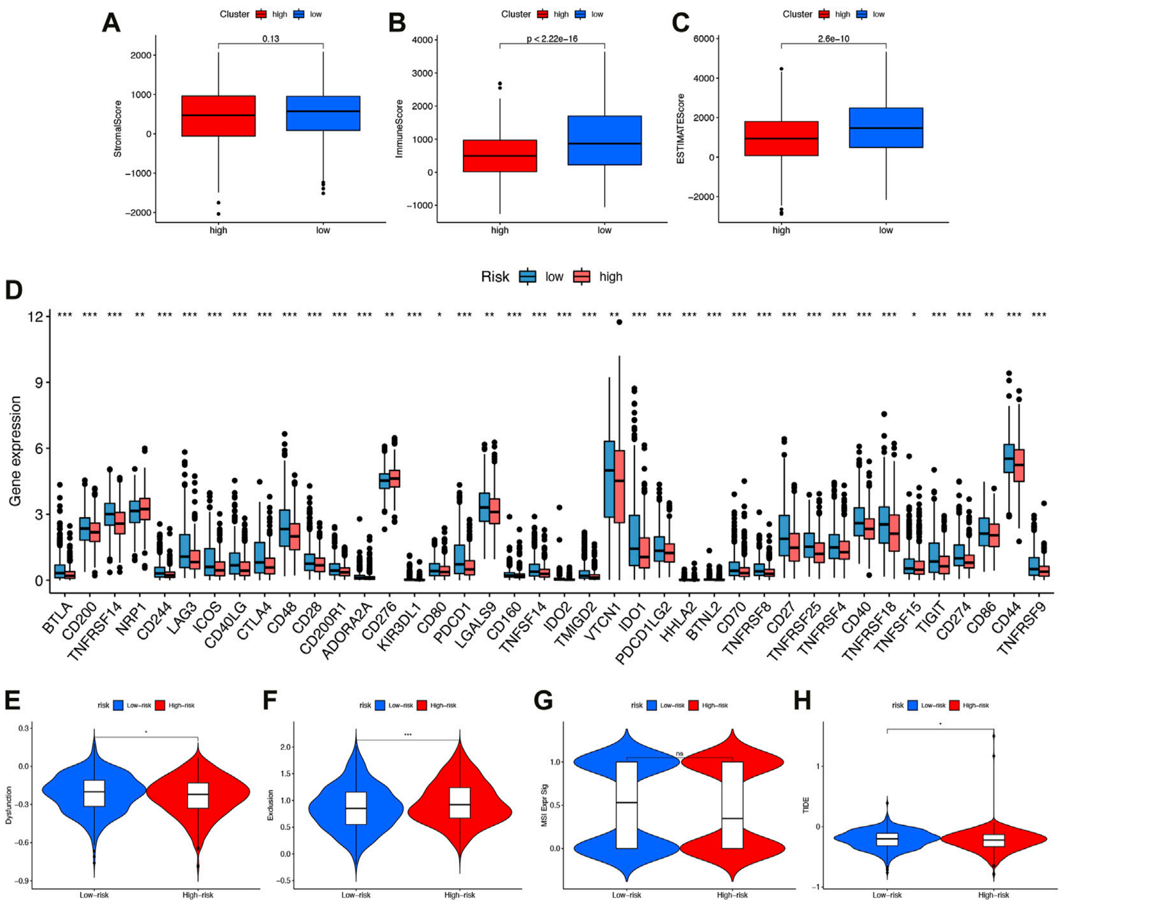


FIGURE 7
The immune microenvironment, immune checkpoints, immune escape **(A–C)** Comparison of interstitial scores, immune scores, and ESTIMATE scoresin high-risk and low-risk subgroups. **(D)** Boxplot showed association between IFRSig and immune checkpoints. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. **(E–H)**Immune escape. **(E)** Dysfunction **(F)** Exclusion **(G)** MSIExprsig **(H)** TIDE score in different risk-groups.

of the top 15 most frequently changed driver genes. There were 6 genes in the low-risk group with the highest mutation frequency, including PIK3CA (35%), TP53 (59%), TTN (17%), CDH1 (12%), GATA3 (11%), MUC16 (11%). TP53 (34%), PIK3CA (33%), TTN (18%), CDH1 (12%), GATA3 (11%), MAP3K1 (10%) and other genes had the top 6 mutation frequencies in the high-risk group. A number of anticancer genes, including TP53, had a relatively high mutation rate among high-risk individuals (34% compared to 30%), MUC16 had a relatively low mutation rate in the high-risk group (9% vs 11%).

A higher level of TMB was found in the high-risk subgroup compared to the low-risk group ($p = 0.028$, Figure 8C). Patients were then assigned to different subtypes on the TMB score. There was a significant correlation between high TMB values and short overall survival ($p = 0.018$, Figure 8D). Moreover, we validated that risk score and TMB could predict BRCA prognosis without immunotherapy synergistically. As shown by the stratified survival curves, TMB status did not interfere with the risk score prognostic prediction performance. In low and high TMB status subtypes, risk score subgroups were significantly different from each other in terms of prognosis ($p < 0.001$, Figure 8E).

4.3 Drug sensitivity

The sensitivity to chemotherapeutic drugs was also anticipated to better direct clinical practice because chemotherapy was a significant therapeutic approach. The IC50 of commonly used chemotherapy drugs (Bleomycin, Bryostatin, Doxorubicin, Cisplatin, Gemcitabine, Gefitinib, Imatinib, Vinorelbine) in BRCA patients in the high-risk group and low-risk group were calculated and compared by pRRophetic analysis (Figures 9A–H). In this study, it was determined that patients with a higher risk score might benefit more from chemotherapy including Bleomycin, Cisplatin, Doxorubicin, Gefitinib, Gemcitabine and Vinorelbine), while patients with a lower risk score might benefit more from chemotherapy including Bryostatin and Imatinib.

5 Discussion

Molecular heterogeneity, high recurrence and mortality rates, and a serious threat to women's health make BRCA one of the most complex cancer types (Saatci et al., 2021). Early detection of BRCA

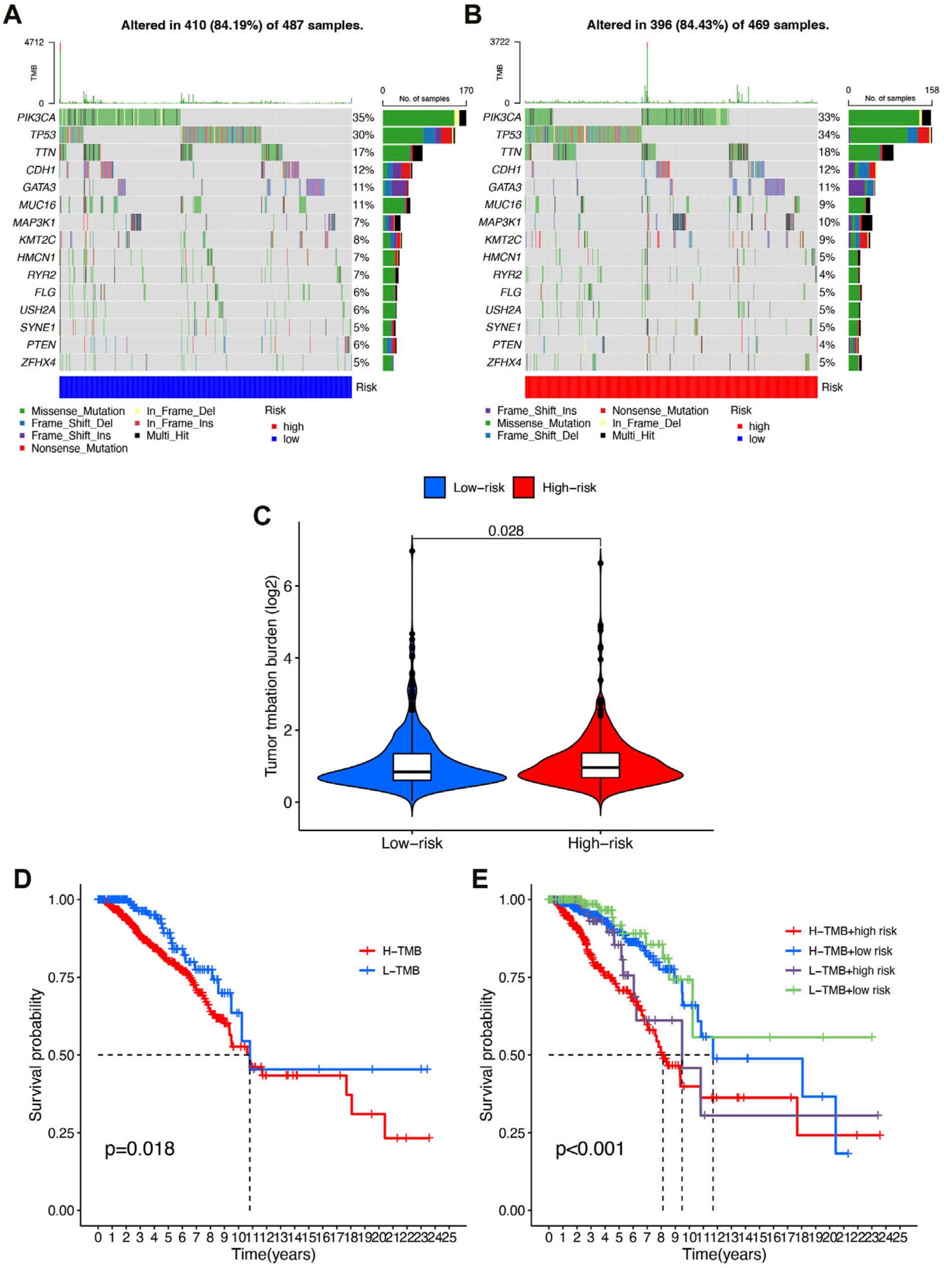


FIGURE 8
Correlation of risk score with TMB. **(A)** Oncoprint of the somatic mutational landscape of the low-risk group. **(B)** Oncoprint of the somatic mutational landscape of the high-risk group. **(C)** TMB differences between patients in low/high risk score subgroups. **(D)** Kaplan-Meier curves of high and low TMBgroups. **(E)** Kaplan-Meier Q19 curves of patients stratified by TMB and risk score.

is essential for effective treatment and an improved prognosis because BRCA has a poor prognosis, which has serious implications for human health and socioeconomics (Winters et al., 2017; Wang et al., 2018). Therefore, finding influential molecular markers, assessing BRCA tumor immunoreactivity, and establishing convincing prognostic models are critical for personalizing BRCA therapy.

There was a synergistic relationship between immunity and ferroptosis in tumors, according to the results of previous studies (Hong et al., 2021; Xu et al., 2021; Yang et al., 2021). In the TME, macrophages could convert from M2 to M1, making more H2O2 available for the Fenton reaction, leading to ferroptosis of tumor cells (Zanganeh et al., 2016). Another study found that activated CD8^{+} T cells release IFN- to prevent cystine from being absorbed by the body's systems, which caused tumor cells to ferroptose through lipid peroxidation (Shao et al., 2021). When tumor cells undergo ferroptosis, tumor antigens are released, resulting in the

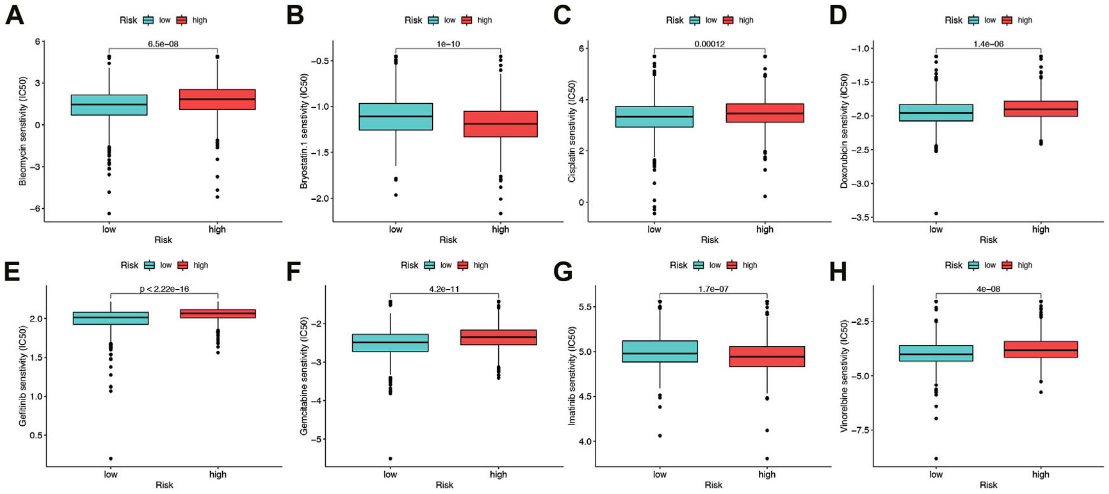


FIGURE 9
Drug sensitivity **(A–H)** Half maximal inhibitory concentration (IC50) of 8 common chemotherapeutic drugs (Bleomycin, Bryostatin, Cisplatin, Doxorubicin, Gefitinib, Gemcitabine, Imatinib, Vinorelbine). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

production of immunogenic TMEs that enhance the response to immune regulation (Lu et al., 2021).

As part of this study, we performed co-expression analyses of breast cancer-related and immune ferroptosis-related genes using the TCGA. After performing a lasso regression analysis, 89 co-expressed immune ferroptosis-related DEGs were collected in order to create prognostic risk models, which could be used for both prognostic and therapeutic purposes. A high-risk and low-risk IFRSig group was created for the cancer samples. In nomograms and prognostic risk models, IFRSig was the key factor. We demonstrated a satisfactory correlation between IFRSig and clinical outcomes, indicating the IFRSig was a useful risk factor for predicting clinical outcomes. To ascertain the effectiveness of the treatment, we examined the sensitivity and resistance to chemotherapeutic drugs.

An TME consists of a complex network of tumor cells within an extremely complex internal environment formed by tumor stromal cells and their secreted active factors, as well as vascular and lymphatic networks, and extracellular matrix (Xiang et al., 2022), of which immune cells and stromal cells were the most common non-tumor cells in TME.

In addition to targeting immunogenic tumor mutations, autologous tumor-infiltrating lymphocytes (TILs) and immune checkpoint inhibitors (ICIs) could help to promote tumor growth (Bu et al., 2021; Kirtane et al., 2021), and antibodies that target PD-1, PD-L1, and CTLA-4 could be used as ICB drugs for the treatment of a variety of cancers (Han et al., 2020; Archilla-Ortega et al., 2022). Thus, we examined how risk subgroups and IC expression relate and found that high-risk patients express more NRP1 and CD276 but less CTLA-4 and PCDC1. This finding suggested that NRP1 and CD276 could be used for targeted immunotherapy for BRCA high-risk patients. Studies have shown that targeting CD276 might reduce cancer stem cell (CSC) immune escape in neck squamous cell carcinoma (HNSCC) (Wang et al., 2021). In conclusion, risk models could be employed to choose immunotherapy that was more appropriate and to forecast how well it will work for BRCA patients.

Overall, we constructed a prognostic risk signature with many advantages, but it still has some limitations. Because of tumor heterogeneity, we needed to validate our risk profile across different cohorts, and it was necessary to validate our risk profile in clinical trials. Despite the fact that our signature was still reliable because we had proven its superiority in terms of survival, tumor-infiltrating immune cells, clinicopathological features, signaling pathways, ICs, and potential small molecule drugs. Upon receiving more information and larger clinical sample sizes, our team will continue to examine and validate the risk profile.

As a result, we developed IFRSig, which was closely related to BRCA prognosis, which along with immunological features could be used to better predict clinical treatment response in patients with BRCA.

6 Conclusion

Our study established a prognostic risk model and identified immune ferroptosis-related genes with independent prognostic value using procedural algorithm analysis. Immune scores, immune checkpoints, and chemotherapeutic agents all showed significant correlations with prognostic models, which were then regarded as an independent prognostic feature to predict OS and clinical treatment response in BRCA patients. In this study, we gained a better understanding of how immune ferroptosis-related genes contribute to BRCA occurrence and progression.

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

Author YX conceived designed the study. ZX and JL analyzed data. ZX wrote the manuscript. YX, CL, and TZ reviewed the manuscript All authors have read and approved this manuscript.

Funding

This study was supported by Tongling Municipal Health Commission Health research (2021) No. 12.

Acknowledgments

The authors thank reviewers for helpful comments on 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.

References

Alavian, F., and Ghasemi, S. (2021). The effectiveness of nanoparticles on gene therapy for glioblastoma cells apoptosis: A systematic review. *Curr. Gene Ther.* 21 (3), 230–245. doi:10.2174/1566523221666210224110454

Archilla-Ortega, A., Domuro, C., Martin-Liberal, J., and Muñoz, P. (2022). Blockade of novel immune checkpoints and new therapeutic combinations to boost antitumor immunity. *J. Exp. Clin. Cancer Res.* 41 (1), 62. doi:10.1186/s13046-022-02264-x

Bu, X., Juneja, V. R., Reynolds, C. G., Mahoney, K. M., Bu, M. T., McGuire, K. A., et al. (2021). Monitoring PD-1 phosphorylation to evaluate PD-1 signaling during antitumor immune responses. *Cancer Immunol. Res.* 9 (12), 1465–1475. doi:10.1158/2326-6066.Cir-21-0493

Chen, X., Kang, R., Kroemer, G., and Tang, D. (2021). Broadening horizons: The role of ferroptosis in cancer. *Nat. Rev. Clin. Oncol.* 18 (5), 280–296. doi:10.1038/s41571-020-00462-0

Ding, C., Ding, X., Zheng, J., Wang, B., Li, Y., Xiang, H., et al. (2020). miR-182-5p and miR-378a-3p regulate ferroptosis in I/R-induced renal injury. *Cell Death Dis.* 11 (10), 929. doi:10.1038/s41419-020-03135-z

Fardi, M., Mohammadi, A., Baradaran, B., and Safaee, S. (2021). ZEB2 knock-down induces apoptosis in human myeloid leukemia HL-60 cells. *Curr. Gene Ther.* 21 (2), 149–159. doi:10.2174/1566523221999210120210017

Gao, M., Monian, P., Quadri, N., Ramasamy, R., and Jiang, X. (2015). Glutaminolysis and transferrin regulate ferroptosis. *Mol. Cell* 59 (2), 298–308. doi:10.1016/j.molcel.2015.06.011

Godenick, M. T. (1995). Learning environmental and occupational medicine as a resident. *Fam. Med.* 27 (4), 226.

Hadian, K., and Stockwell, B. R. (2020). SnapShot: Ferroptosis. *Cell* 181 (5), 1188–1188.e1. doi:10.1016/j.cell.2020.04.039

Han, Y., Liu, D., and Li, L. (2020). PD-1/PD-L1 pathway: Current researches in cancer. *Am. J. Cancer Res.* 10 (3), 727–742.

Hangauer, M. J., Viswanathan, V. S., Ryan, M. J., Bole, D., Eaton, J. K., Matov, A., et al. (2017). Drug-tolerant persister cancer cells are vulnerable to GPX4 inhibition. *Nature* 551 (7679), 247–250. doi:10.1038/nature24297

He, P., Hua, H., Tian, W., Zhu, H., Liu, Y., and Xu, X. (2020). Holly (Ilex latifolia thunb.) polyphenols extracts alleviate hepatic damage by regulating ferroptosis following diquat challenge in a piglet model. *Front. Nutr.* 7, 604328. doi:10.3389/fnut.2020.604328

He, Y., Jiang, Z., Chen, C., and Wang, X. (2018). Classification of triple-negative breast cancers based on Immunogenomic profiling. *J. Exp. Clin. Cancer Res.* 37 (1), 327. doi:10.1186/s13046-018-1002-1

Hong, Y., Lin, M., Ou, D., Huang, Z., and Shen, P. (2021). A novel ferroptosis-related 12-gene signature predicts clinical prognosis and reveals immune relevancy in clear cell renal cell carcinoma. *BMC Cancer* 21 (1), 831. doi:10.1186/s12885-021-08559-0

Islam, M. A., Kundu, S., and Hassan, R. (2020). Gene therapy approaches in an autoimmune demyelinating disease: Multiple sclerosis. *Curr. Gene Ther.* 19 (6), 376–385. doi:10.2174/1566523220666200306092556

Jiang, X., Stockwell, B. R., and Conrad, M. (2021). Ferroptosis: Mechanisms, biology and role in disease. *Nat. Rev. Mol. Cell Biol.* 22 (4), 266–282. doi:10.1038/s41580-020-00324-8

Kirtane, K., Elmariah, H., Chung, C. H., and Abate-Daga, D. (2021). Adoptive cellular therapy in solid tumor malignancies: Review of the literature and challenges ahead. *J. Immunother. Cancer* 9 (7), e002723. doi:10.1136/jitc-2021-002723

Lang, X., Green, M. D., Wang, W., Yu, J., Choi, J. E., Jiang, L., et al. (2019). Radiotherapy and immunotherapy promote tumoral lipid oxidation and ferroptosis via synergistic repression of SLC7A11. *Cancer Discov.* 9 (12), 1673–1685. doi:10.1158/2159-8290.Cd-19-0338

Lu, Y., Yang, Q., Su, Y., Ji, Y., Li, G., Yang, X., et al. (2021). MYCN mediates TFRC-dependent ferroptosis and reveals vulnerabilities in neuroblastoma. *Cell Death Dis.* 12 (6), 511. doi:10.1038/s41419-021-03790-w

Pupa, S. M., Ligorio, F., Cancila, V., Franceschini, A., Tripodo, C., Vernieri, C., et al. (2021). HER2 signaling and breast cancer stem cells: The bridge behind HER2-positive breast cancer aggressiveness and therapy refractoriness. *Cancers (Basel)* 13 (19), 4778. doi:10.3390/cancers13194778

Saatci, O., Huynh-Dam, K. T., and Sahin, O. (2021). Endocrine resistance in breast cancer: From molecular mechanisms to therapeutic strategies. *J. Mol. Med. Berl.* 99 (12), 1691–1710. doi:10.1007/s00109-021-02136-5

Shao, L., Yu, Q., Xia, R., Zhang, J., Gu, S., Yu, D., et al. (2021). B7-H3 on breast cancer cell MCF7 inhibits IFN-γ release from tumour-infiltrating T cells. *Pathol. Res. Pract.* 224, 153461. doi:10.1016/j.prp.2021.153461

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2021). Cancer statistics, 2021. *CA Cancer J. Clin.* 71 (1), 7–33. doi:10.3322/caac.21654

Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. (2023). Cancer statistics, 2023. *CA Cancer J. Clin.* 73 (1), 17–48. doi:10.3322/caac.21763

Sindhu, R. K., Madaan, P., Chandel, P., Akter, R., Adilakshmi, G., and Rahman, M. H. (2021). Therapeutic approaches for the management of autoimmune disorders via gene therapy: Prospects, challenges, and opportunities. *Curr. gene Ther.* 22, 245–261. doi:10.2174/1566523221666210916113609

Song, Y., Tian, S., Zhang, P., Zhang, N., Shen, Y., and Deng, J. (2021). Construction and validation of a novel ferroptosis-related prognostic model for acute myeloid leukemia. *Front. Genet.* 12, 708699. doi:10.3389/fgene.2021.708699

Sun, X., Niu, X., Chen, R., He, W., Chen, D., Kang, R., et al. (2016). Metallothionein-1G facilitates sorafenib resistance through inhibition of ferroptosis. *Hepatology* 64 (2), 488–500. doi:10.1002/hep.28574

Viswanathan, V. S., Ryan, M. J., Dhruv, H. D., Gill, S., Eichhoff, O. M., Seashore-Ludlow, B., et al. (2017). Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway. *Nature* 547 (7664), 453–457. doi:10.1038/nature23007

Wang, C., Li, Y., Jia, L., Kim, J. K., Li, J., Deng, P., et al. (2021). CD276 expression enables squamous cell carcinoma stem cells to evade immune surveillance. *Cell Stem Cell* 28 (9), 1597–1613.e7. doi:10.1016/j.stem.2021.04.011

Wang, N., Gu, Y., Li, L., Wang, F., Lv, P., Xiong, Y., et al. (2018). Circular RNA circMYO9B facilitates breast cancer cell proliferation and invasiveness via upregulating FOXP4 expression by sponging miR-4316. *Arch. Biochem. Biophys.* 653, 63–70. doi:10.1016/j.abb.2018.04.017

Wang, W., Green, M., Choi, J. E., Gijón, M., Kennedy, P. D., Johnson, J. K., et al. (2019). CD8(+) T cells regulate tumour ferroptosis during cancer immunotherapy. *Nature* 569 (7755), 270–274. doi:10.1038/s41586-019-1170-y

Winters, S., Martin, C., Murphy, D., and Shokar, N. K. (2017). Breast cancer epidemiology, prevention, and screening. *Prog. Mol. Biol. Transl. Sci.* 151, 1–32. doi:10.1016/bs.pmbts.2017.07.002

Xiang, Z., Zhao, J., Qu, J., Song, J., and Li, L. (2022). A multivariate-gated DNA nanodevice for spatioselective imaging of pro-metastatic targets in extracellular microenvironment. *Angew. Chem. Int. Ed. Engl.* 61 (4), e202111836. doi:10.1002/anie.202111836

Xu, C., Sun, S., Johnson, T., Qi, R., Zhang, S., Zhang, J., et al. (2021). The glutathione peroxidase Gpx4 prevents lipid peroxidation and ferroptosis to sustain Treg cell activation and suppression of antitumor immunity. *Cell Rep.* 35 (11), 109235. doi:10.1016/j.celrep.2021.109235

Yang, L., Li, C., Qin, Y., Zhang, G., Zhao, B., Wang, Z., et al. (2021). A novel prognostic model based on ferroptosis-related gene signature for bladder cancer. *Front. Oncol.* 11, 686044. doi:10.3389/fonc.2021.686044

Yang, W. S., and Stockwell, B. R. (2016). Ferroptosis: Death by lipid peroxidation. *Trends Cell Biol.* 26 (3), 165–176. doi:10.1016/j.tcb.2015.10.014

Zanganeh, S., Hutter, G., Spitler, R., Lenkov, O., Mahmoudi, M., Shaw, A., et al. (2016). Iron oxide nanoparticles inhibit tumour growth by inducing pro-inflammatory macrophage polarization in tumour tissues. *Nat. Nanotechnol.* 11 (11), 986–994. doi:10.1038/nnano.2016.168

Zhang, F., Li, F., Lu, G. H., Nie, W., Zhang, L., Lv, Y., et al. (2019). Engineering magnetosomes for ferroptosis/immunomodulation synergism in cancer. *ACS Nano* 13 (5), 5662–5673. doi:10.1021/acsnano.9b00892

TYPE Original Research
PUBLISHED 25 May 2023
DOI 10.3389/fgene.2023.1202409

Check for updates

OPEN ACCESS

EDITED BY
Tao Huang,
Chinese Academy of Sciences (CAS),
China

REVIEWED BY
Chengming Zhang,
The University of Tokyo, Japan
Xiaoping Liu,
University of Chinese Academy of
Science, China

*CORRESPONDENCE
Qianqian Shi,
✉ qqshi@mail.hzau.edu.cn

RECEIVED 08 April 2023
ACCEPTED 09 May 2023
PUBLISHED 25 May 2023

CITATION
Li X, Huang W, Xu X, Zhang H-Y and Shi Q (2023), Deciphering tissue heterogeneity from spatially resolved transcriptomics by the autoencoder-assisted graph convolutional neural network.
Front. Genet. 14:1202409.
doi: 10.3389/fgene.2023.1202409

Deciphering tissue heterogeneity from spatially resolved transcriptomics by the autoencoder-assisted graph convolutional neural network

Xinxing Li, Wendong Huang, Xuan Xu, Hong-Yu Zhang and Qianqian Shi*

Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China

Spatially resolved transcriptomics (SRT) provides an unprecedented opportunity to investigate the complex and heterogeneous tissue organization. However, it is challenging for a single model to learn an effective representation within and across spatial contexts. To solve the issue, we develop a novel ensemble model, AE-GCN (**a**uto**e**ncoder-assisted **g**raph **c**onvolutional neural **n**etwork), which combines the autoencoder (AE) and graph convolutional neural network (GCN), to identify accurate and fine-grained spatial domains. AE-GCN transfers the AE-specific representations to the corresponding GCN-specific layers and unifies these two types of deep neural networks for spatial clustering via the clustering-aware contrastive mechanism. In this way, AE-GCN accommodates the strengths of both AE and GCN for learning an effective representation. We validate the effectiveness of AE-GCN on spatial domain identification and data denoising using multiple SRT datasets generated from ST, 10x Visium, and Slide-seqV2 platforms. Particularly, in cancer datasets, AE-GCN identifies disease-related spatial domains, which reveal more heterogeneity than histological annotations, and facilitates the discovery of novel differentially expressed genes of high prognostic relevance. These results demonstrate the capacity of AE-GCN to unveil complex spatial patterns from SRT data.

KEYWORDS

spatially resolved transcriptomics, spatial domain identification, spatial information, graph convolutional neural network, autoencoder

Introduction

Spatially resolved transcriptomics (SRT) technologies, such as spatial transcriptomics (ST) (Ståhl et al., 2016), 10x Visium, and Slide-seqV2 (Stickels et al., 2021), can measure the transcript localization and abundance in the dissected tissue area, enabling novel insights into tissue development and tumor heterogeneity (Atta and Fan, 2021; Nasab et al., 2022). Their generated data (i.e., gene expression in tissue locations [spots] and spatial locational information) can be used to decipher the spatially functional regions and cellular architectures in tissues (Maniatis et al., 2021; Marx, 2021; Zeng et al., 2022). However, due to technical limitations (Xu et al., 2022), modeling and integrating the available SRT modalities for accurate spatial domain identification still remain challenging.

Currently, the spatial domain detection methods could be mainly divided into two categories: non-spatial and spatial clustering methods. Some non-spatial methods originally developed for single-cell RNA-sequencing (scRNA-seq) studies, e.g., Seurat (Butler et al., 2018) and Scanpy (Wolf et al., 2018), are also applied in SRT studies. They only utilize the expression profiles to cluster spots while often obtaining domains lacking in spatial continuity to some extent. To address such issues, spatial clustering approaches generally incorporate the additional spatial information into their models. For example, with the spatial prior, BayesSpace (Zhao et al., 2021) and HMRF (Dries et al., 2021) use the Markov random field model (or its variant) to encourage the spatially neighboring spots to belong to the same domain. SpaGCN (Hu et al., 2021) and SEDR (Fu et al., 2021) enable spatial clustering by learning the low-dimensional representation with graph constraints that represent the spatial dependency. STAGATE (Dong and Zhang, 2022) identifies spatial domains by adaptively learning the similarity of neighboring spots via attention mechanisms. Modeling the spatial dependency of gene expression fairly facilitates the discovery of spatial domains with spatial coherence.

Though these methods have provided useful information on the usage of expression profiles and spatial information, they usually depend on single models, which center on either expression data itself or spatially neighboring structure, thus probably resulting in the preferred usage of the focused data type. For example, the non-spatial clustering methods only models the gene expression itself, while the spatial clustering methods often take spatial neighbors prior as a hard constraint to ensure spatial clustering continuity, which may lead to over-smoothing of expression (Huang et al., 2018) and missing subtle spatial regions with a handful of spots. Thus, the rational combination of these different kinds of models can fairly generate more useful representations, enabling better spatial domain detection in SRT studies.

Here, we develop a novel combined model, AE-GCN (**a**uto**e**ncoder-assisted **g**raph **c**onvolutional neural **n**etwork), which combines the autoencoder (AE) and graph convolutional neural network (GCN), for accurate and fine-grained spatial domain identification. Specifically, AE-GCN relies on AE for learning expression data-based representations and GCN for spatial graph-constrained learning. AE-GCN orderly transfers the AE-specific representations to GCN-specific layers and unifies these two types of neural networks for spatial clustering via a clustering-aware contrastive mechanism. In this way, AE-GCN combines the advantages of the two models and takes full integration of the expression data and spatial information during the representation learning process.

We demonstrate the effectiveness of AE-GCN on spatial domain identification and data denoising using SRT datasets generated from ST, 10x Visium, and Slide-seqV2 platforms. In particular, it is validated in two cancer samples that AE-GCN can refine the spatial functional regions and discover novel cancer-associated genes. These results show that AE-GCN is capable of unveiling complex tissue architecture from SRT data.

Materials and methods

Overview of AE-GCN

AE-GCN is an integrative scheme that incorporates the AE and GCN learning processes, enabling tasks of spatial domain detection and data denoising (Figure 1).

Given the original expression $X^0 \in R^{M\times N}$ (where M and N, respectively, denote the number of genes and spots) and spatial coordinates, the spatially neighboring network $A \in R^{N\times N}$ and the enhanced expression data $X \in R^{M\times N}$ are computed as the input of the combined learning process (see Methods). On the enhanced expression data X, AE-GCN employs AE to learn the low-dimensional representation (i.e., AE-specific representation $H_1^{(l)}, l = 1, \ldots, L$, where L is the number of total layers in AE) in each layer. With the spatially neighboring network A, AE-GCN utilizes GCN to learn the graph-constrained representation in each layer (i.e., GCN-specific representation $H_2^{(l)}, l = 1, \ldots, B$, where B is the number of layers in GCN or the encoder of AE). Then, AE-GCN transfers the AE-specific representations from the encoder to the corresponding GCN layer, thus generating the combined representation Y. Additionally, AE-GCN proposes a clustering-aware contrastive module to make the combined representation appropriate for spatial clustering.

When the learning process reaches convergence, the low-dimensional representation (i.e., Y) of the last layer and the reconstructed expression data (i.e., X') can be used for downstream analytical tasks. The optimal representation enables AE-GCN to identify spatial domains interoperating with the Leiden method (Traag et al., 2019). The reconstructed expression data serve as the denoised profile, which overcomes the sparsity of the original data to improve differentially expressed gene identification (see Methods).

Spatially neighboring network construction and expression augmentation

Spatially neighboring network construction

Assume that there are original expression matrix X^0 and spatial locations in the SRT dataset. We first use spatial coordinates and Euclidean distance to calculate the distance between spots and then select the k-nearest spatial neighbors of each spot to participate in the subsequent process. In this work, we set $k = 10$ for ST and 10x Visium datasets and $k = 30$ for Slide-seqV2 datasets. Then, we perform principal component analysis (PCA) based on gene expression and select the top p PCs (i.e., $U \in R^{p\times N}$, default to 15) to calculate the similarity matrix $D \in R^{N\times N}$ between the center spot and its spatial neighbors using cosine metric:

$$D = \exp(2 - cosine\_dist(U)), D_{ii} = 0 \tag{1}$$

Then, the weighted adjacency matrix $A \in R^{N\times N}$ is obtained by normalizing the similarity matrix D:

$$A_{ij} = \frac{D_{ij}}{\sum_{i=0}^{N} D_{ij}} \tag{2}$$

Spatial expression augmentation

Limited by the transcript capture rate of SRT technologies, expression data are often sparse and noisy. AE-GCN generates the enhanced expression data X by borrowing the shared information from spatial neighborhood, which can correct low-quality measurements and strengthen local similarity:

$$X = X^0 + \alpha X^0 A \tag{3}$$

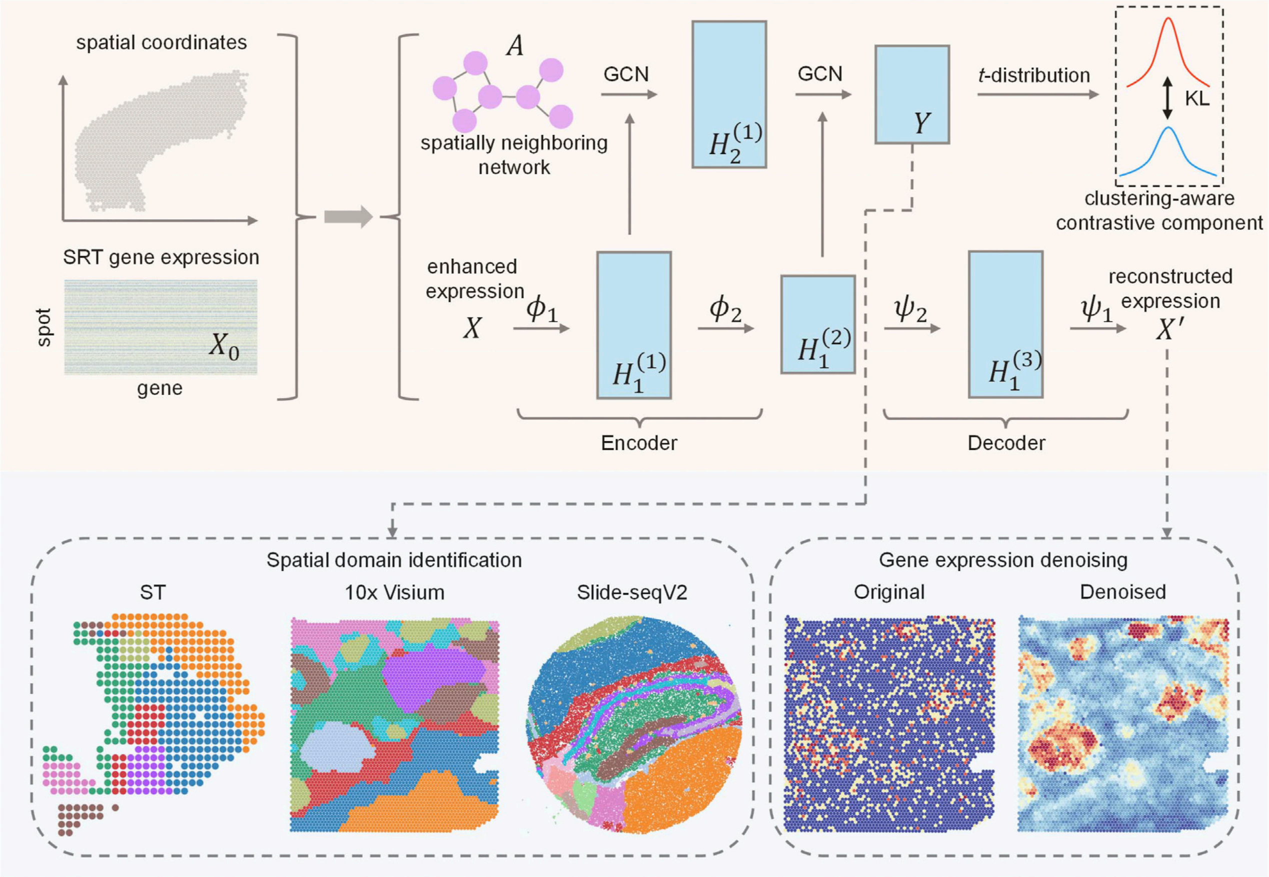


FIGURE 1
Schematic overview of AE-GCN and its potential applications. Given gene expression and spatial coordinates as input, AE-GCN first builds the spatially neighboring network *A* and enhances expression *X*. AE-GCN uses AE to learn representations from the enhanced expression and employs GCN to learn the representations of each layer from the spatially neighboring network *A*. Then, AE-GCN transfers the AE-specific representations from the encoder to the corresponding GCN-specific layer and learns the combined representation *Y*. To ensure effective training of the combined deep learning model for clustering, AE-GCN proposes a clustering-aware contrastive module based on the distribution of the representation *Y*. When AE-GCN reaches convergence, the latent combined representation *Y* enables AE-GCN to identify spatial domains for different platforms, i.e., ST, 10x Visium, and Slide-seqV2. The reconstructed expression data *X′* serves to denoise expression profiles.

where the tunable parameter α is flexibly set, and it controls the extent to aggregating expression information from neighboring spots.

AE component

We employ AE to learn the useful representations from the expression data itself and assume that there are B layers in the encoder and $(L-B)$ layers in the decoder. Specifically, the learned lth layer representation, $H_1^{(l)}$, can be obtained as follows:

$$H_1^{(l)} = \phi_l\left(W^{(l)}H_1^{(l-1)} + b^{(l)}\right), l = 1, \ldots, B \quad (4)$$

$$H_1^{(l)} = \psi_l\left(W^{(l)}H_1^{(l-1)} + b^{(l)}\right), l = B+1, \ldots, L \quad (5)$$

where ϕ_l and ψ_l are the activation functions of the lth layer in the respective encoder and decoder. $W^{(l)}$ and $b^{(l)}$ are the weight matrix and reconstruction error in the lth layer, respectively. For convenience, we denote the enhanced expression data X as $H_1^{(0)}$.

The output (i.e., $X' = H_1^{(L)}$) of the decoder part is obtained through the reconstruction of the input data (i.e., X) by minimizing the following loss function:

$$\mathcal{L}_{res} = \left\|X - X'\right\|_F^2 \quad (6)$$

GCN component

AE-specific representations, e.g., $H_1^{(1)}$, $H_1^{(2)}$, $\cdots$, $H_1^{(L)}$, can denoise data itself and extract valuable information from the data itself, which can effectively reflect expression variation but cannot guarantee the spatial smoothness of the identified domains. GCN can model the spatial structural dependency between spots, which is beneficial to improving the spatial smoothness of the identified domains. Thus, we then transfer AE-specific representations in the encoder into GCN-specific representations and use the GCN module to propagate these AE-specific representations for capturing a more complete and powerful representation. Thus, the GCN-learnable representations can accommodate two different kinds of information: gene expression values and spatial neighborhood structure. The representation learned by the lth layer of GCN, $H_2^{(l)}$, can be obtained as follows:

$$H_2^{(l)} = (1-\mu)\phi_l\left(W^{(l)}H_2^{(l-1)}\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\right) + \mu H_1^{(l)}\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}} \qquad (7)$$

where I denotes the identity diagonal matrix. $\tilde{A} = A + I$ and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ is the normalized adjacency matrix. μ is the balance coefficient and is often uniformly set to 0.5. Note that GCN and AE share weights.

Note that we denote the representation (i.e., $H_2^{(B)}$) of the last GCN layer as Y. The input of the first layer GCN can be obtained from the enhanced expression data X:

$$H_2^{(0)} = \phi_l\left(W^{(0)}X\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\right) \qquad (8)$$

Clustering-aware contrastive component

Although we have incorporated the encoder of AE into the neural network architecture of GCN to obtain the combined representation, this representation cannot be directly applied to the clustering problem. Herein, we propose a clustering-aware contrastive module to unify these two different deep learning models for effective spatial clustering.

Specifically, we use student's t-distribution to measure the probability of assigning the spot i to cluster j based on the combined latent representation Y as follows:

$$q_{ij} = \frac{\left(1 + \left\|y_i - \mu_j\right\|^2 / \rho\right)^{-\frac{\rho+1}{2}}}{\sum_{j'}\left(1 + \left\|y_i - \mu_{j'}\right\|^2 / \rho\right)^{-\frac{\rho+1}{2}}} \qquad (9)$$

where μ_j is the cluster center by K-means on learned representations. y_i is the ith column of Y. We regard $Q = [q_{ij}]$ as the distribution of the assignments of all samples. ρ is the degree of freedom of student's t-distribution.

To optimize the AE-GCN-learnable representation from the high-confidence assignment, we make data representation closer to cluster centers for improving the cluster cohesion. Hence, we calculate the target distribution P as follows:

$$p_{ij} = \frac{{q_{ij}}^2 / s_j}{\sum_{j'}{q_{ij'}}^2 / s_{j'}} \qquad (10)$$

where $s_j = \sum_i q_{ij}$ is the soft cluster frequency. Each assignment in Q is squared and normalized to produce the target distribution P, which makes the data representation surround the cluster centers closer and helps AE-GCN learn a better representation for the clustering task. By minimizing the KL (Kullback–Leibler) divergence loss between Q and P distributions, the target distribution P can help the AE-GCN learn a better representation for the clustering task, i.e., making the data representation surround the cluster centers closer, thus leading to the following loss function:

$$\mathcal{L}_{cl} = KL(P\|Q) = \sum_i\sum_j p_{ij} log\frac{p_{ij}}{q_{ij}} \qquad (11)$$

This design is regarded as a clustering-aware contrastive mechanism, where the P distribution supervises the updating of the distribution Q, and the target distribution P is calculated by the distribution Q in turn. Using this mechanism, AE-GCN can directly concentrate two different objectives: clustering objective and data reconstruction objective, in one loss function. Thus, the overall loss function of AE-GCN is

$$\mathcal{L}_{obj} = \mathcal{L}_{res} + \beta\mathcal{L}_{cl} \qquad (12)$$

where β denotes the tunable parameter $\beta > 0$ and can be flexibly set, which balances data reconstruction and clustering optimization.

Data collection and general preprocessing

The top 3,000 highly variable genes (HVGs) for 13 10x Visium datasets, one ST dataset, and one Slide-seqV2 dataset are selected using scanpy.pp.highly_variable_genes() from the Scanpy Python package. The log-transformation of the expression profiles is performed using scanpy.pp.log1p() on the original gene expression data.

Spatial domain detection and gene expression denoising

AE-GCN uses the combined latent representation Y to detect spatial domains by Leiden (Traag et al., 2019) algorithms implemented as scanpy.tl.leiden(). The parameter "resolution" can be adjusted to match the number of the manual annotations.

For the enhanced expression matrix X, AE-GCN aggregates the shared information between each spot and its surrounding neighbors by incorporating prior spatial information into gene expression, which is used to adjust expression values in each spot and enrich spatial local signals. For the reconstructed expression data X', AE-GCN uses AE and GCN to reconstruct the enhanced expression matrix X. By minimizing the reconstruction error, the reconstructed data X' can reflect both the spatial local signals and expression measurement global signals. Thus, AE-GCN uses the reconstructed expression data X' as the denoised profiles.

Performance evaluation

We use adjusted Rand index (ARI) (Hubert and Arabie, 1985) and cluster purity (i.e., Eq. 13) (Zhao et al., 2021) to quantify the accuracy of the identified spatial domain and the reference annotations from original publications.

$$\text{cluster purity} = \frac{1}{N}\sum_{c\in C}\max_{g\in G}\left|c \cap g\right| \qquad (13)$$

where C is denoted as the set of the spatial cluster set and G is regarded as the set of annotated groups. Due to cancer slices with rough annotations (e.g., IDC and PDAC cancer data), cluster purity

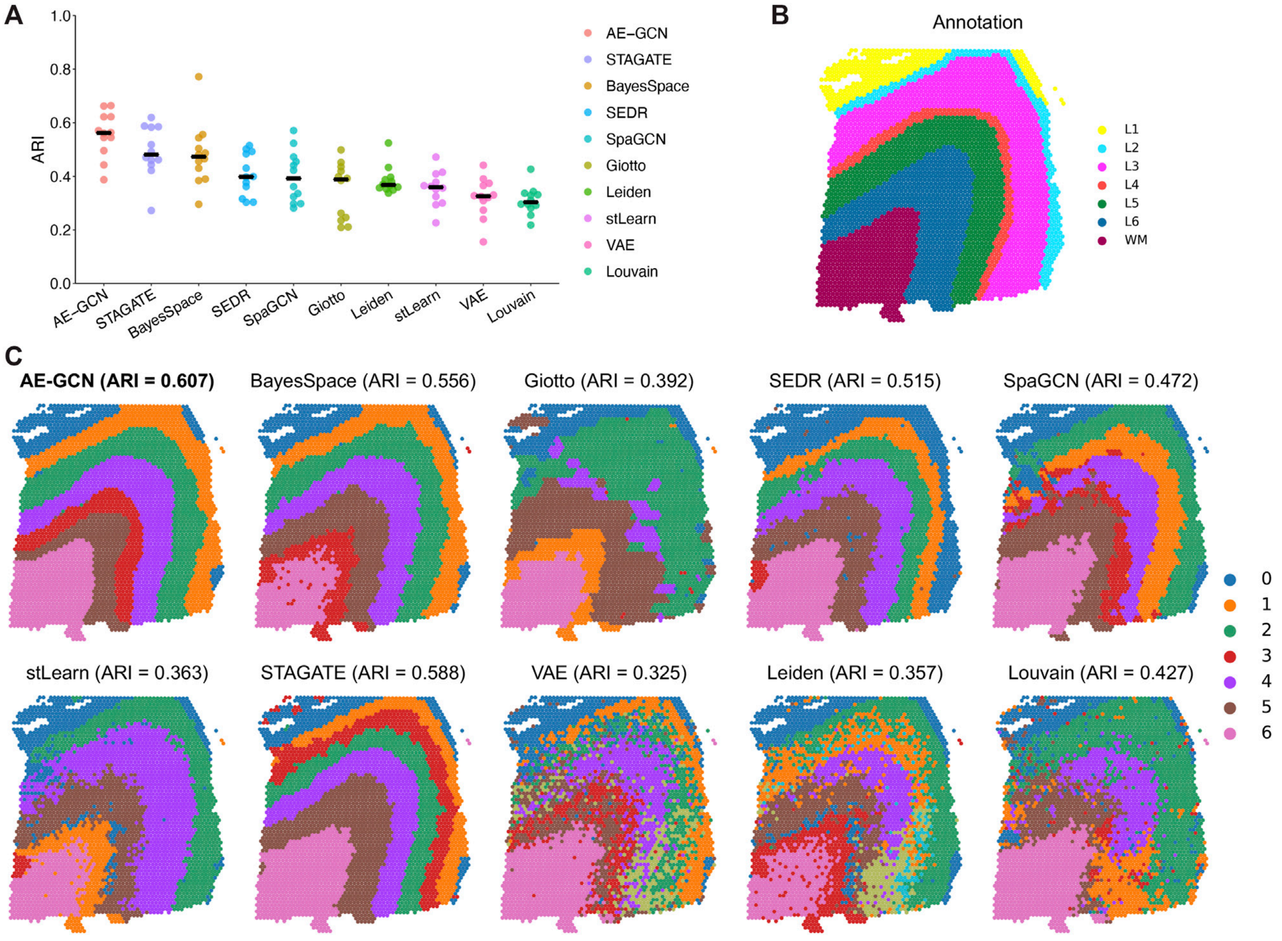


FIGURE 2
Benchmarking AE-GCN against state-of-the-art spatial domain detection methods. **(A)** Spatial clustering performance is compared using ARI on 12 manually annotated DLPFC datasets from spatialLIBD. The bold line represents the mean ARI value of each approach on all the datasets. **(B)** Slice 151673 with the manual annotation. **(C)** Comparative illustration of the identified spatial domain on slice 151673. The identified spatial domains of each method are distinguished by colors without strict correspondence.

is specifically used to evaluate the clustering performance on SRT cancer datasets (Zhao et al., 2021).

Survival analysis

We use bulk expression data with patient survival information to evaluate the prognostic significance of genes via the Kaplan–Meier plotter (Zwyea et al., 2021) in the IDC and PDAC cancer studies.

Results

Benchmarking AE-GCN against state-of-the-art methods

We evaluated the ability of AE-GCN to detect spatial domains using 12 human dorsolateral prefrontal cortex (DLPFC) slices generated using 10x Visium. The DLPFC dataset obtained from spatialLIBD (Pardo et al., 2022) is manually annotated as the layered regions by gene markers and cytoarchitecture. The annotations can be considered as the ground truth for benchmarking. Based on this dataset, we compared AE-GCN with the existing state-of-the-art methods, including six spatial clustering methods [i.e., BayesSpace (Zhao et al., 2021), Giotto (Dries et al., 2021), SEDR (Fu et al., 2021), SpaGCN (Hu et al., 2021), stLearn (Pham et al., 2020), and STAGATE (Dong and Zhang, 2022)] and three non-spatial algorithms [i.e., variational autoencoder (VAE) (Kingma and Welling, 2019), Leiden implemented in Scanpy (Wolf et al., 2018), and Louvain implemented in Seurat (Butler et al., 2018)]. The adjusted Rand index (ARI) is used to quantify the similarity between the manual labels and identified clusters, which ranges from 0 for poor consistency to 1 for identical clusters.

Generally, most of the spatial clustering methods performed better than non-spatial algorithms (Wilcoxon signed-rank test $P < 10^{-5}$, Figure 2A), which showed that the integration of spatial information is necessary to improve the spatial clustering performance. Strikingly, AE-GCN had the highest mean ARI (mean ARI = 0.561) and substantially performed better than the competing methods over the slices (Wilcoxon signed-rank test $P < 10^{-8}$, Figure 2A). Taking slice

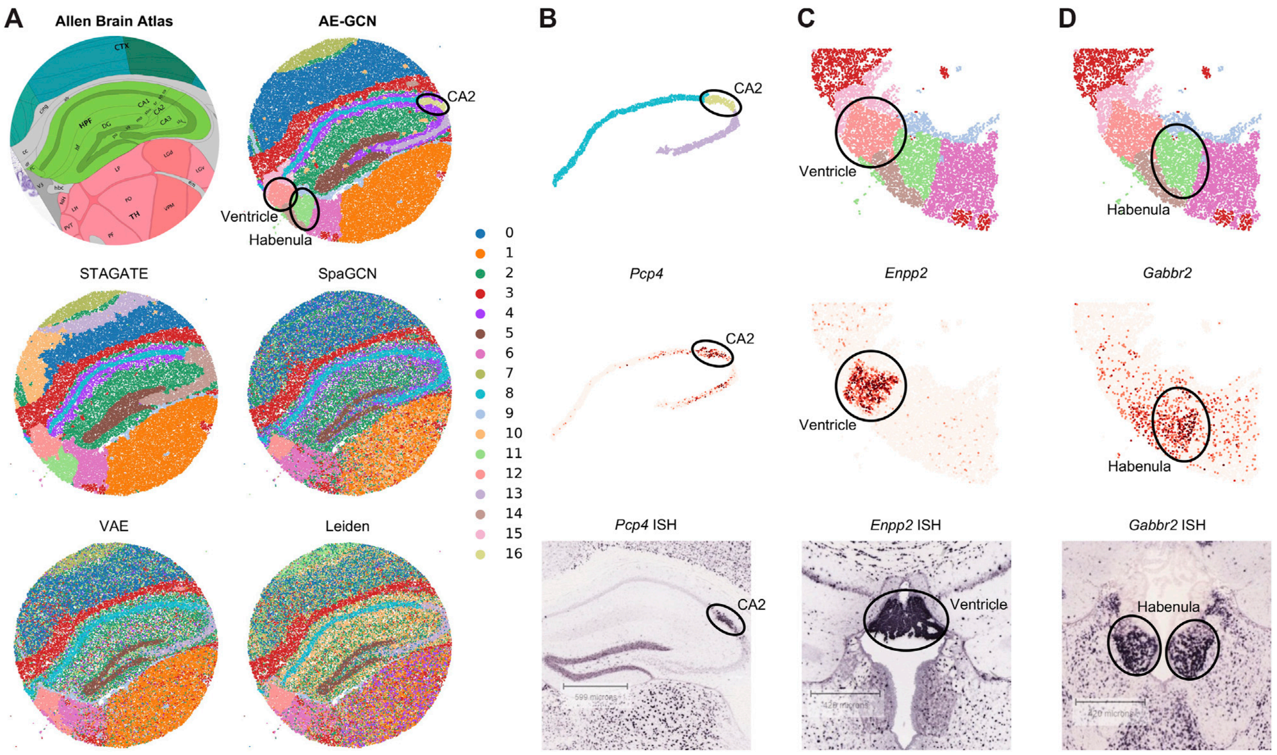


FIGURE 3
AE-GCN reveals the finer-grained anatomical regions on mouse hippocampus Slide-seqV2 data. **(A)** Corresponding anatomical diagram from the Allen Mouse Brain Atlas and spatial domains identified by each competed method. **(B–D)** CA2 and ventricle and habenula regions (at the top) from AE-GCN partitions are, respectively, validated by the known gene markers (i.e., *Pcp4*, *Enpp2*, and *Gabbr2*) from gene expression (at the middle) and ISH images (at the bottom). The ISH images of *Pcp4*, *Enpp2*, and *Gabbr2* are also obtained from the Allen Mouse Brain Atlas.

151673 as an example (Figure 2B), we found AE-GCN (ARI = 0.623), STAGATE (ARI = 0.588), BayesSpace (ARI = 0.556), and SEDR (ARI = 0.515) delineated the layered regions (Figure 2C). Notably, the partitions from AE-GCN (termed as a deep learning model-combined method) exhibited clearer and less noisy outcomes than those from single model-based methods (e.g., GCN-based SpaGCN and the VAE model).

AE-GCN reveals fine-grained anatomical regions on mouse hippocampus Slide-seqV2 data

To illustrate the effectiveness of AE-GCN on high-resolution SRT platforms, we applied AE-GCN to a mouse hippocampus Slide-seqV2 dataset (n = 41,786 spots). Slide-seqV2 can measure gene expression at near-cellular resolution (Stickels et al., 2021) but has lower number of transcripts per location/spot and higher dropouts than the 10x Visium platform. Thus, it poses more challenges for accurately distinguishing tissue structures from the data of high sparsity. To better validate the performance of AE-GCN, we also compared it with other domain detection methods and used the corresponding anatomical diagram from the Allen Mouse Brain Atlas (Sunkin et al., 2012) as the illustrative reference (Figure 3A).

Comparing with the reference, we found that AE-GCN and STAGATE can identify the spatially coherent domains compared to other involved methods. However, AE-GCN performed better to detect the fine-grained structures, such as the cornu ammonis 2 (CA2, AE-GCN domain 16), ventricle (AE-GCN domain 12), and habenula (AE-GCN domain 11) sections (Figure 3A). These sections are delineated with sharper boundaries and higher concordance with the anatomical annotation. We further isolated the focused regions and provided validations from other perspectives (Figures 3B–D). For the CA2 section, which is only detected by AE-GCN, the domain location showed good alignment with the marker gene expression (i.e., *Pcp4* (San Antonio et al., 2014)) and independent *in situ* hybridization (ISH) image (Figure 3B). For ventricle and habenula sections, AE-GCN domains are closer to the shapes of their respective marker expression (*Enpp2* for ventricle (Koike et al., 2006) and *Gabbr2* for habenula (De Beaurepaire, 2018)) or stained regions and match the anatomical shape well (Figures 3C, D). Thus, for higher-resolution SRT data, AE-GCN is capable of effectively unveiling the fine-grained anatomical functional regions.

AE-GCN accurately discerns tumor regions on human pancreatic ductal adenocarcinoma data

To illustrate the effectiveness of AE-GCN on cancer tissue, we applied AE-GCN to the human pancreatic ductal adenocarcinoma (PDAC) ST dataset (n = 428 spots). The histopathological image and annotations were taken as references (Figures 4A, B). We assessed these

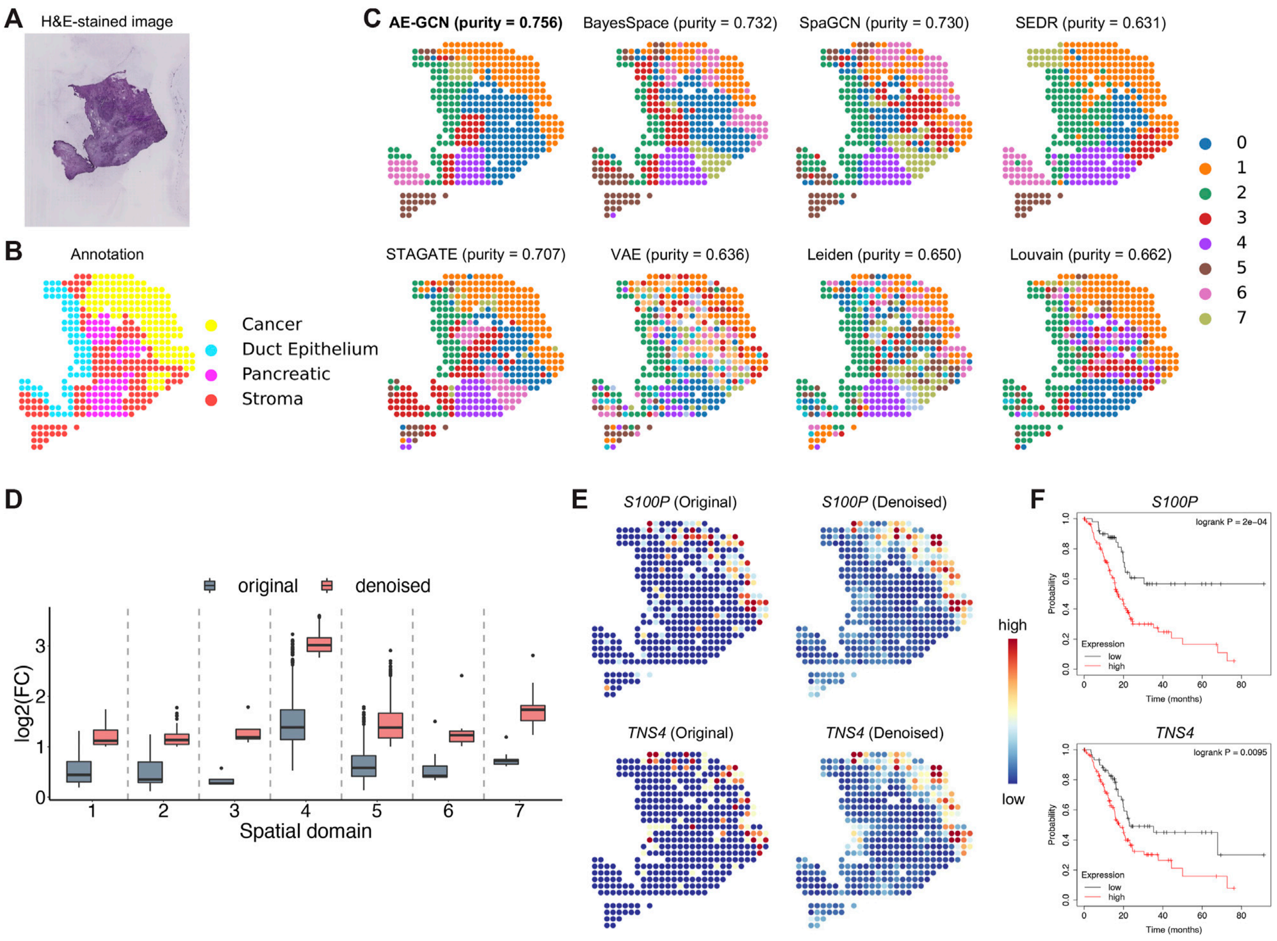


FIGURE 4
AE-GCN identifies tumor regions on human PDAC ST data. The H&E-stained image **(A)** and the corresponding manual annotation **(B)** are shown as references. **(C)** The identified spatial domains using all the compared methods are distinguished by different colors without strict correspondence. Cluster purity is used to compare the similarity between identified domains and the reference annotation. **(D)** The change in gene differential expression in each domain before and after data denoising. $\log_2$(FC): the logarithmic value of the gene expression fold change with base 2. **(E)** Spatial expression visualization of selected DEGs (i.e., *S100P* and *TNS4*) before and after data denoising. **(F)** Kaplan–Meier survival curves show the clinical relevance of the identified DEGs (i.e., *S100P* and *TNS4*).

spatial domain identification methods using cluster purity (see Methods) as the quantitative measure on cancer datasets with rough annotation information. AE-GCN achieved the highest cluster purity (purity = 0.756) and detected more spatially enriched functional regions in tumor tissue than other compared methods (Figure 4C).

Next, we examined whether AE-GCN could provide more insights into the underlying tumor heterogeneity, as data sparsity could hinder other downstream analytical tasks, for example, the identification of differentially expressed genes (DEGs). In this manner, we used the AE-GCN-reconstructed data to denoise the low-quality measurements and evaluated the effectiveness in recovering gene spatial expression patterns. Based on the denoised data, we selected the top 50 DEGs of each domain from the reconstructed data X' and compared the log fold change (LFC) of these DEGs before and after denoising (Figure 4D). Overall, the comparison highlights the significant improvement of biological specificity brought by AE-GCN denoising across the identified domains (Wilcoxon signed-rank test $P < 10^{-14}$, Figure 4D). In particular, we found that some DEG expression (e.g., *S100P* and *TNS4*) appeared more spatially smoothed on spots *in situ* (Figure 4E). These two DEGs were validated to be the potential prognostic risk factors for PDAC (Figure 4F). For example, *S100P* is ever reported to be involved in the aggressive properties of cancer cells and associated with poor prognosis (Wang et al., 2012) (Figure 4F); *TNS4* is associated with cancer cell motility and migration, whose high expression can indicate poor prognosis (Sakashita et al., 2008). These results indicate that AE-GCN has the potential to provide the in-depth biological insights into the underlying tumor heterogeneity from the perspectives of spatial domain detection and gene expression pattern recovery.

AE-GCN reveals more intratumor heterogeneity on invasive ductal carcinoma data

To illustrate the generalization ability of AE-GCN on cancer tissues, we next tested AE-GCN using the invasive ductal carcinoma

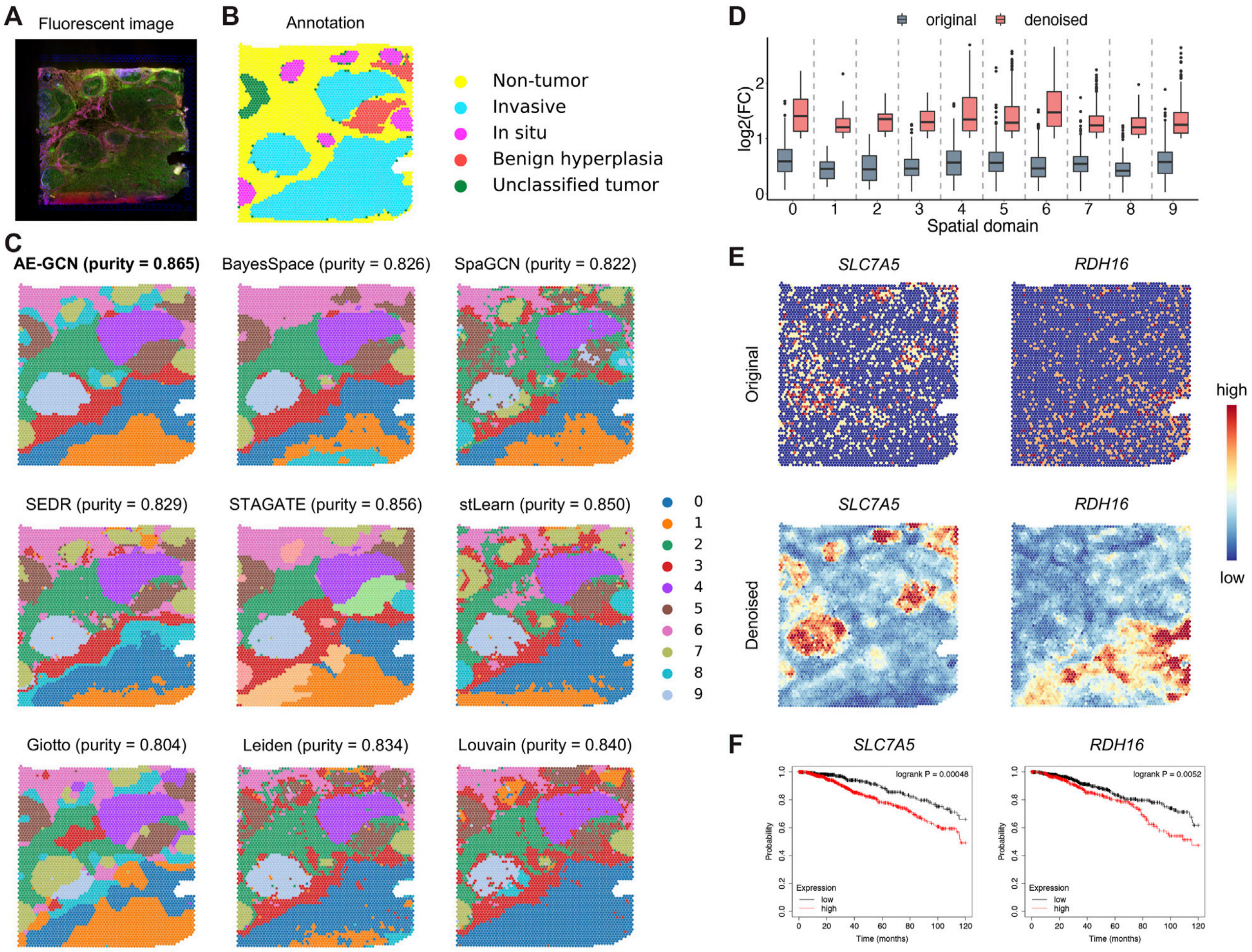


FIGURE 5
AE-GCN provides more biological insights into intratumor heterogeneity on the IDC 10x Visium dataset. The fluorescent image **(A)** and the corresponding manual annotation **(B)** are shown as references. Each spot is colored due to the annotation label in **(B)**. **(C)** The spatial domains obtained by all involved methods are distinguished using different colors without strict correspondence. Cluster purity is used to compare the similarities between identified outcomes and reference annotation. **(D)** The change gene FC before and after data denoising. **(E)** Spatial expression visualization of the selected domain-specific genes (i.e., *SLC7A5* and *RDH16*) before and after data denoising. **(F)** Kaplan–Meier survival curves show the clinical relevance of the newly identified DEGs (i.e., *SLC7A5* and *RDH16*).

(IDC) Visium dataset (n = 4,727 spots). The histopathological annotations from the original paper (Zhao et al., 2021) were taken as the reference (Figures 5A, B). We found that the identified domains of AE-GCN were highly consistent with the manual annotations (purity = 0.865, Figure 5C). Compared with the domains captured by other methods, the clustering partitions from AE-GCN showed clear spatial separations with few scatter points and high regional continuity.

Then, for functional gene identification, we identified the top 50 DEGs of each cluster from the denoised data X'. Similarly, based on the comparison before and after denoising, we found that AE-GCN significantly improves the LFCs of gene expression, revealing more biological specificity across domains, which may suggest the detection of new disease-associated genes (Figure 5D). For example, *SLC7A5* and *RDH16* are two newly found DEGs after denoising, whose spatial expression patterns are greatly enhanced after denoising (Figure 5E). Moreover, the two novel DEGs were shown to be the potential prognostic risk genes for breast cancer via survival analysis of independent clinical data (Figure 5F). Their biological functions in tumors indicate the prognostic relevance from previous studies. For example, *SLC7A5* is reported to involve in tumor cell metabolism and promotes cell proliferation (El Ansari et al., 2018). *RDH16* affects retinol metabolism to participate indirectly in breast cancer occurrence and progression (Gao et al., 2020). The application, along with the PDAC case, demonstrates that AE-GCN can unveil cancer heterogeneity from SRT data, enabling the discovery of novel spatial patterns of both samples and genes.

Discussion

Spatially resolved transcriptomics technologies measure gene expression on each spot while preserving spatial context, which can support computational methods to identify functional regions of tissue and further resolve organizational heterogeneity. The combined modeling of gene expression and spatial information enables the improved identification accuracy of spatial domains, especially for complex spatial architecture, e.g., tumor microenvironments. In this paper, AE-GCN combines the autoencoder and graph convolutional neural network to achieve effective latent representations from expression data itself and spot neighboring structure. The superiority of AE-GCN is shown not only on the accurate and fine-grained identification of spatial domains for multiple SRT platforms but also on the recovery or identification of gene spatial expression patterns. In particular, the application on cancer slices (i.e., IDC and PDAC) demonstrates that AE-GCN reveals more functional regions and novel cancer prognostic genes for interpreting cancer heterogeneity, suggesting that AE-GCN has great capability of unveiling tissue heterogeneity from SRT data.

The effectively combined modeling is key to the superiority of AE-GCN in the SRT study. Generally, AE models learn the representations from expression data itself, while GCN models learn the structured representations from the sample graph structure by providing an approximate second-order graph regularization, which may suffer from over-smoothing issues. AE-GCN combines the characteristics of these two deep learning methods and integrates them to learn effective representations so that AE is used to weaken the problem of overfitting while simultaneously learning the structured representations in GCN. Additionally, the proposed clustering-aware contrastive module in AE-GCN further promotes the combined model from processes independent of clustering targets to the model that achieves effective spatial clustering. Thus, AE-GCN can not only effectively use the information of the expression data itself but also reasonably regularize the learned information from expression data by spatial structure between spots, which has better advantages than the spatial domain detection methods based on a single-model design in SRT studies.

Currently, AE-GCN only models gene expression and spatial information from SRT data and cannot utilize histological images which are also provided by several SRT technologies, e.g., 10x Visium. Although some methods have used histological images in spatial domain detection, histological images are mainly used to enhance the quality of expression data and lack of modeling image data separately, e.g., stLearn (Pham et al., 2020). Compared with expression data and spatial information, histological image data are one type of modalities more suitable for deep learning modeling. The future work to extend AE-GCN is to integrate deep learning models for each multi-modal data characteristic (i.e., gene expression, histological images, and spatial information) to improve the performance of current methods in SRT research.

Data availability statement

The datasets presented in this study can be found in online repositories. The available web resources include human DLPFC datasets (available in spatialLIBD package), mouse hippocampus Slide-seqV2 dataset (https://singlecell.broadinstitute.org), human IDC 10x Visium dataset (https://www.10xgenomics.com/resources/datasets) and human PDAC ST dataset (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672). Python source code of AE-GCN is available at https://github.com/zccqq/AE-GCN.

Author contributions

QS conceived and designed the framework and the experiments. XL and WH performed the experiments. WH and XX developed the Python package and documentation website of the framework. QS analyzed the data and wrote the paper. QS and HZ revised the manuscript. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 61802141).

Acknowledgments

The authors thank the reviewers for useful suggestions.

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.

References

Atta, L., and Fan, J. (2021). Computational challenges and opportunities in spatially resolved transcriptomic data analysis. *Nat. Commun.* 12 (1), 5283. doi:10.1038/s41467-021-25557-9

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36 (5), 411–420. doi:10.1038/nbt.4096

De Beaurepaire, R. (2018). A review of the potential mechanisms of action of baclofen in alcohol use disorder. *Front. Psychiatry* 9, 506. doi:10.3389/fpsyt.2018.00506

Dong, K., and Zhang, S. (2022). Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* 13 (1), 1739–1812. doi:10.1038/s41467-022-29439-6

Dries, R., Zhu, Q., Dong, R., Eng, C.-H. .L., Li, H., Liu, K., et al. (2021). Giotto: A toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* 22 (1), 78–31. doi:10.1186/s13059-021-02286-2

El Ansari, R., Craze, M. .L., Miligy, I., Diez-Rodriguez, M., Nolan, C. .C., Ellis, I. .O., et al. (2018). The amino acid transporter SLC7A5 confers a poor prognosis in the highly proliferative breast cancer subtypes and is a key therapeutic target in luminal B tumours. *Breast Cancer Res.* 20, 21–17. doi:10.1186/s13058-018-0946-6

Fu, H., Hang, X. .U., and Chen, J. (2021). *Unsupervised spatial embedded deep representation of spatial transcriptomics*. bioRxiv.

Gao, C., Zhuang, J., Li, H., Liu, C., Zhou, C., Liu, L., et al. (2020). Development of a risk scoring system for evaluating the prognosis of patients with Her2-positive breast cancer. *Cancer Cell Int.* 20 (1), 121–212. doi:10.1186/s12935-020-01175-1

Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. .J., et al. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. methods* 18 (11), 1342–1351. doi:10.1038/s41592-021-01255-8

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., et al. (2018). Saver: Gene expression recovery for single-cell RNA sequencing. *Nat. methods* 15 (7), 539–542. doi:10.1038/s41592-018-0033-z

Hubert, L., and Arabie, P. (1985). Comparing partitions. *J. Classif.* 2 (1), 193–218. doi:10.1007/bf01908075

Kingma, D. .P., and Welling, M. (2019). An introduction to variational autoencoders. *Found. Trends® Mach. Learn.* 12 (4), 307–392. doi:10.1561/2200000056

Koike, S., Keino-Masu, K., Ohto, T., and Masu, M. (2006). The N-terminal hydrophobic sequence of autotaxin (ENPP2) functions as a signal peptide. *Genes cells* 11 (2), 133–142. doi:10.1111/j.1365-2443.2006.00924.x

Maniatis, S., Petrescu, J., and Phatnani, H. (2021). Spatially resolved transcriptomics and its applications in cancer. *Curr. Opin. Genet. Dev.* 66, 70–77. doi:10.1016/j.gde.2020.12.002

Marx, V. (2021). Method of the year: Spatially resolved transcriptomics. *Nat. methods* 18 (1), 9–14. doi:10.1038/s41592-020-01033-y

Nasab, R. .Z., Ghamsari, M. .R. .E., Argha, A., Macphillamy, C., Beheshti, A., Alizadehsani, R., et al. (2022). *Deep learning in spatially resolved transcriptomics: A comprehensive technical view*. arXiv preprint arXiv:221004453 2022.

Pardo, B., Spangler, A., Weber, L. .M., Page, S. .C., Hicks, S. .C., Jaffe, A. .E., et al. (2022). spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. *BMC genomics* 23 (1), 434. doi:10.1186/s12864-022-08601-w

Pham, D., Tan, X., Xu, J., Grice, L. .F., Lam, P. .Y., Raghubar, A., et al. (2020). stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv*.

Sakashita, K., Mimori, K., Tanaka, F., Kamohara, Y., Inoue, H., Sawada, T., et al. (2008). Prognostic relevance of Tensin4 expression in human gastric cancer. *Ann. Surg. Oncol.* 15, 2606–2613. doi:10.1245/s10434-008-9989-8

San Antonio, A., Liban, K., Ikrar, T., Tsyganovskiy, E., and Xu, X. (2014). Distinct physiological and developmental properties of hippocampal CA2 subfield revealed by using anti-Purkinje cell protein 4 (PCP4) immunostaining. *J. Comp. Neurology* 522 (6), 1333–1354. doi:10.1002/cne.23486

Ståhl, P. .L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. .F., Magnusson, J., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353 (6294), 78–82. doi:10.1126/science.aaf2403

Stickels, R. .R., Murray, E., Kumar, P., Li, J., Marshall, J. .L., Di Bella, D. .J., et al. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* 39 (3), 313–319. doi:10.1038/s41587-020-0739-1

Sunkin, S. .M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T. .L., Thompson, C. .L., et al. (2012). Allen Brain Atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic acids Res.* 41 (D1), D996-D1008–D1008. doi:10.1093/nar/gks1042

Traag, V. .A., Waltman, L., and Van Eck, N. .J. (2019). From Louvain to leiden: Guaranteeing well-connected communities. *Sci. Rep.* 9 (1), 5233–5312. doi:10.1038/s41598-019-41695-z

Wang, Q., Zhang, Y.-N., Lin, G.-L., Qiu, H.-Z., Wu, B., Wu, H.-Y., et al. (2012). S100P, a potential novel prognostic marker in colorectal cancer. *Oncol. Rep.* 28 (1), 303–310. doi:10.3892/or.2012.1794

Wolf, F. .A., Angerer, P., and Theis, F. .J. (2018). Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biol.* 19 (1), 15–5. doi:10.1186/s13059-017-1382-0

Xu, C., Jin, X., Wei, S., Wang, P., Luo, M., Xu, Z., et al. (2022). DeepST: Identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Res.* 50 (22), e131–e. doi:10.1093/nar/gkac901

Zeng, Z., Li, Y., Li, Y., and Luo, Y. (2022). Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biol.* 23 (1), 83–23. doi:10.1186/s13059-022-02653-7

Zhao, E., Stone, M. .R., Ren, X., Guenthoer, J., Smythe, K. .S., Pulliam, T., et al. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* 39 (11), 1375–1384. doi:10.1038/s41587-021-00935-2

Zwyea, S., Naji, L., and Almansouri, S. (2021). "Kaplan-Meier plotter data analysis model in early prognosis of pancreatic cancer," in Journal of Physics: Conference Series (IOP Publishing).

TYPE Methods
PUBLISHED 30 May 2023
DOI 10.3389/fgene.2023.1201934

Prediction of small molecule drug-miRNA associations based on GNNs and CNNs

Zheyu Niu, Xin Gao, Zhaozhi Xia, Shuchao Zhao, Hongrui Sun, Heng Wang, Meng Liu, Xiaohan Kong, Chaoqun Ma, Huaqiang Zhu, Hengjun Gao, Qinggong Liu, Faji Yang, Xie Song, Jun Lu and Xu Zhou*

Department of Hepatobiliary Surgery, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, China

Check for updates

OPEN ACCESS

EDITED BY
Junlin Xu,
Hunan University, China

REVIEWED BY
Wei Liu,
Xiangtan University, China
Yajie Meng,
Wuhan Textile University, China

*CORRESPONDENCE
Xu Zhou,
✉ zhouxu2008@sina.com

RECEIVED 07 April 2023
ACCEPTED 17 May 2023
PUBLISHED 30 May 2023

CITATION
Niu Z, Gao X, Xia Z, Zhao S, Sun H, Wang H, Liu M, Kong X, Ma C, Zhu H, Gao H, Liu Q, Yang F, Song X, Lu J and Zhou X (2023), Prediction of small molecule drug-miRNA associations based on GNNs and CNNs.
Front. Genet. 14:1201934.
doi: 10.3389/fgene.2023.1201934

MicroRNAs (miRNAs) play a crucial role in various biological processes and human diseases, and are considered as therapeutic targets for small molecules (SMs). Due to the time-consuming and expensive biological experiments required to validate SM-miRNA associations, there is an urgent need to develop new computational models to predict novel SM-miRNA associations. The rapid development of end-to-end deep learning models and the introduction of ensemble learning ideas provide us with new solutions. Based on the idea of ensemble learning, we integrate graph neural networks (GNNs) and convolutional neural networks (CNNs) to propose a miRNA and small molecule association prediction model (GCNNMMA). Firstly, we use GNNs to effectively learn the molecular structure graph data of small molecule drugs, while using CNNs to learn the sequence data of miRNAs. Secondly, since the black-box effect of deep learning models makes them difficult to analyze and interpret, we introduce attention mechanisms to address this issue. Finally, the neural attention mechanism allows the CNNs model to learn the sequence data of miRNAs to determine the weight of sub-sequences in miRNAs, and then predict the association between miRNAs and small molecule drugs. To evaluate the effectiveness of GCNNMMA, we implement two different cross-validation (CV) methods based on two different datasets. Experimental results show that the cross-validation results of GCNNMMA on both datasets are better than those of other comparison models. In a case study, Fluorouracil was found to be associated with five different miRNAs in the top 10 predicted associations, and published experimental literature confirmed that Fluorouracil is a metabolic inhibitor used to treat liver cancer, breast cancer, and other tumors. Therefore, GCNNMMA is an effective tool for mining the relationship between small molecule drugs and miRNAs relevant to diseases.

KEYWORDS
small molecule drug, miRNAs, graph neural networks, convolutional neural networks, CNN, liver cancer

Introduction

With the development of sequencing technology, the biomedical field has accumulated a large amount of medical data, which provides more convenience for researchers to study the relationship between diseases and drugs using these data. The prediction of the relationship between small molecule (SM) drugs and microRNAs (miRNAs) has become an important and rapidly developing area in pharmacology and pharmacogenomics research (Bartel, 2004; Beermann et al., 2016; Kozomara et al., 2019; Liu et al., 2022). miRNAs are small non-coding RNA molecules that regulate gene expression and play a key role in various biological processes, including the development of diseases (Cai et al., 2021; Peng et al., 2023). On the other hand, small molecule drugs have been widely used to treat diseases, but their impact on miRNA expression is not clear. However, there are still blind issues in using traditional biological experiments to identify small molecule drug-related miRNAs, which require a lot of experimental time and cost. With the increasing availability of large datasets, it is possible to predict the relationship between small molecule drugs and miRNAs and use this information to improve the efficacy and safety of drugs (Wang et al., 2019; Chen et al., 2020). This field has tremendous potential in discovering new therapeutic targets and developing personalized drugs (Chen et al., 2021; Liu et al., 2023; Xu et al., 2023).

Computational methods have played a crucial role in predicting the association between small molecule drugs and miRNAs (Xu et al., 2020; Zhang et al., 2023). As the available data on drugs and miRNAs continues to increase, various computational methods have been proposed to identify and predict their interactions. Lv et al. (2015) constructed a complete network by combining small molecule similarity networks, miRNA similarity networks, and known small molecule-miRNA association networks. They calculated the similarity of small molecules and miRNAs using a weighted combination strategy, and then used the RWR (Random Walk With Restart) algorithm to predict the potential associations between small molecule drugs and miRNAs. BNNRSMMA first defined a new matrix to represent the small molecule-miRNA heterogenous network using miRNA-miRNA similarity, small molecule-small molecule similarity, and known small molecule-miRNA associations. They then completed this matrix by minimizing its kernel parameter count and used alternating direction multiplication to further minimize the kernel parameter count and obtain prediction scores. They introduced a regularization term to tolerate noise in the integrated similarity. Wang et al. (2022a) proposed a novel dual-network collaborative matrix factorization (DCMF) method for predicting potential SM-miRNA associations. They first preprocessed the missing values in the SM-miRNA association matrix using the WKNKN method, and then constructed a matrix factorization model for the dual network to obtain feature matrices containing potential features of small molecules and miRNAs, respectively. Finally, the predicted SM-miRNA association score matrix was obtained by calculating the inner product of the two feature matrices. Li et al. (2016) proposed a network-based inference model for small molecule-miRNA networks (SMiR-NBI), which relies solely on known SM-miRNA associations. For a given SM, the initial resources are evenly allocated to its associated miRNAs. Then, the resources of each miRNA are allocated to all its associated SMs, and the resources are then redistributed from SMs to their associated miRNAs. The final resources obtained by the miRNAs reflect the likelihood of associations between the given SM and miRNAs. Guan et al. (2018) developed a new graphlet interaction-based inference model for predicting small molecule-miRNA associations (GISMMA). The complex relationships among SMs or miRNAs are described by graphlet interactions, which consist of 28 isomers. The association score for an SM-miRNA pair is calculated by counting the number of graphlet interactions. However, if neither the SM nor the miRNA has a known association, the model cannot predict the SM-miRNA association. Wang et al. (2022b) proposed an ensemble method for predicting small molecule-miRNA associations based on kernel ridge regression (EKRRSMMA). This method combines feature dimension reduction and ensemble learning to reveal potential SM-miRNA associations. Firstly, the authors constructed different feature subsets for SMs and miRNAs. Then, homogeneous base learners were trained on different feature subsets, and the average scores obtained from these base learners were used as the association scores for SM-miRNA pairs. Peng et al. (2022) proposed a new computational method based on deep autoencoder and scalable tree boosting model (DAESTB) to predict the associations between small molecules and miRNAs. Firstly, a high-dimensional feature matrix was constructed by integrating small molecule-small molecule similarity, miRNA-miRNA similarity, and known small molecule-miRNA associations. Secondly, the feature dimension of the integrated matrix was reduced using a deep autoencoder to obtain potential feature representations for each small molecule-miRNA pair. Finally, a scalable tree boosting model was used to predict potential associations between small molecules and miRNAs. Although these models have achieved promising results and played important roles in the development of computational methods for small molecule-miRNA association identification, they have certain issues or limitations: the experimental validation of small molecule-miRNA associations is very limited, and there are many negative associations. When performed on this noisy and sparse small molecule-miRNA association network, the predictors often detect many false negative associations.

Therefore, we propose a miRNA-molecule association prediction model (GCNNMMA) by integrating graph convolutional networks (GCNs) (Scarselli et al., 2008) and convolutional neural networks (CNNs) (Chen, 2015) (Figure 1). Firstly, GCNs are used to effectively learn the molecular structural graph data of small molecule drugs, and CNNs are used to learn the sequence data of miRNAs. Due to the black-box nature of deep learning models, it is difficult to analyze and interpret them. Therefore, GCNNMMA introduces a neural attention mechanism (Bahdanau et al., 2014) to address this issue. The neural attention mechanism enables CNNs to learn the weights of sub-sequences in miRNAs, thus predicting the associations between miRNAs and small molecule drugs.

Materials and methods

Datasets

For dataset 1, we obtained a total of 664 known small molecule-miRNA associations from SM2miR database (version 1.0) (Liu et al., 2013). Then a total of 831 small molecules were extracted and

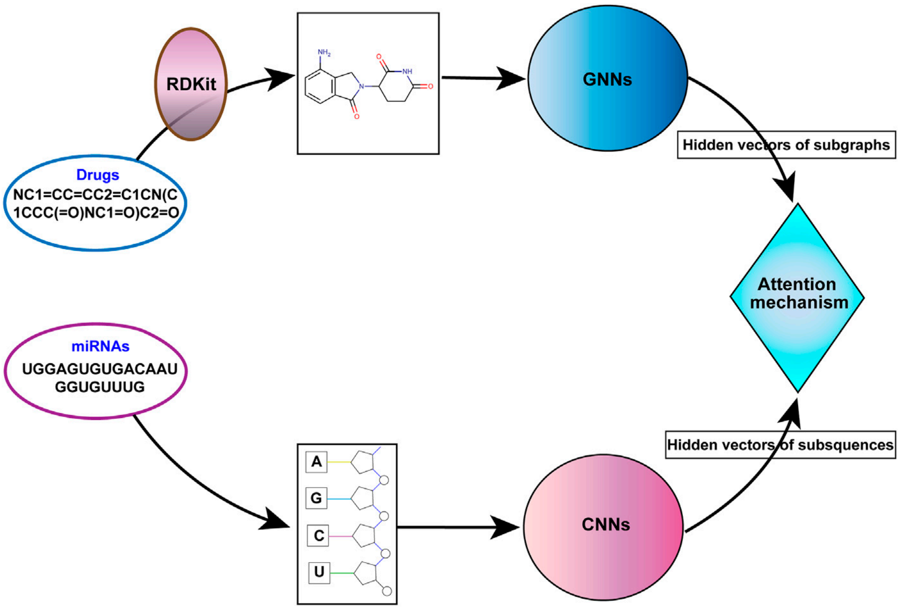

RDKit
Drugs
NC1=CC=CC2=C1CN(C1CCC(=O)NC1=O)C2=O
GNNs
Hidden vectors of subgraphs
Attention mechanism
miRNAs
UGGAGUGUGACAAUGGUGUUUG
A
G
C
U
CNNs
Hidden vectors of subsquences

FIGURE 1
The overall workflow of GCNNMMA.

TABLE 1 Statistics of datasets used in this study.

Dataset	No. of miRNAs	No. of molecules	No. of associations
Dataset 1	541	831	664
Dataset 2	2,460	680	60,212

integrated from SM2miR, DrugBank (Wishart et al., 2018), and PubChem (Kim et al., 2019). 541 miRNAs were collected from SM2miR, HMDD, miR2Disease, and PhenomiR (Ruepp et al., 2010). To evaluate our model performance more comprehensively, we constructed dataset 2, which contains 680 small molecules, 2,460 miRNAs, and 60,212 known small molecule-miRNA associations. Additionally, we downloaded corresponding small molecule drug SMILES data from DrugBank. The SMILES format data was used to describe the spatial structural information of small molecule drugs. Furthermore, we obtained corresponding miRNA sequence data from the miRbase database (Table 1).

Prediction model based on the integration of CNNs and GNNs

GNNs process small molecule drug data

End-to-end learning model GNNs has been shown to achieve good performance in many scenarios. Therefore, we first use two functions [the transformation function $tran(x)$ and the output function $f(x)$] in GNNs to map the molecular structure graph $G(V,E)$ of small molecule drugs to a low-dimensional vector $y\epsilon\mathbb{R}^{d}$. The transformation function $tran(x)$ updates the feature information of each node in the molecular graph $G(V,E)$ using information from neighboring nodes (atoms in the molecular structure graph) and neighboring edges (chemical bonds in the molecular structure graph). The output function $f(x)$ converts the updated node information in the molecular graph after the transformation function into a low-dimensional vector. In GNNs, both the transformation function and the output function are implemented as differentiable neural networks, and the parameters in the functions are automatically learned through the backpropagation process (Figure 2). The specific steps are as follows:

Subgraph embedding with radius r: Here, we use $G(V,E)$ to represent a molecular graph, where V is a set of nodes and E is a set of edges. In the molecular structure graph, $v_i\epsilon V$ represents the i-th atom and $e_{ij}\epsilon E$ represents the chemical bond between atom i and atom j. Because there are only a few types of nodes (hydrogen and carbon) and edges (double and single bonds) in the molecular graph, representative learning models cannot obtain effective learning results. To solve this problem, GCNNMMA introduces the concept of r-radius subgraphs. An r-radius subgraph describes the set of atoms and chemical bonds within a radius of r with a certain atom as the center. Here, we use $\Gamma(i,r)$ to represent the set of indices of all adjacent nodes in the subgraph with node i as the center and a radius of r. $\Gamma(i,0)$

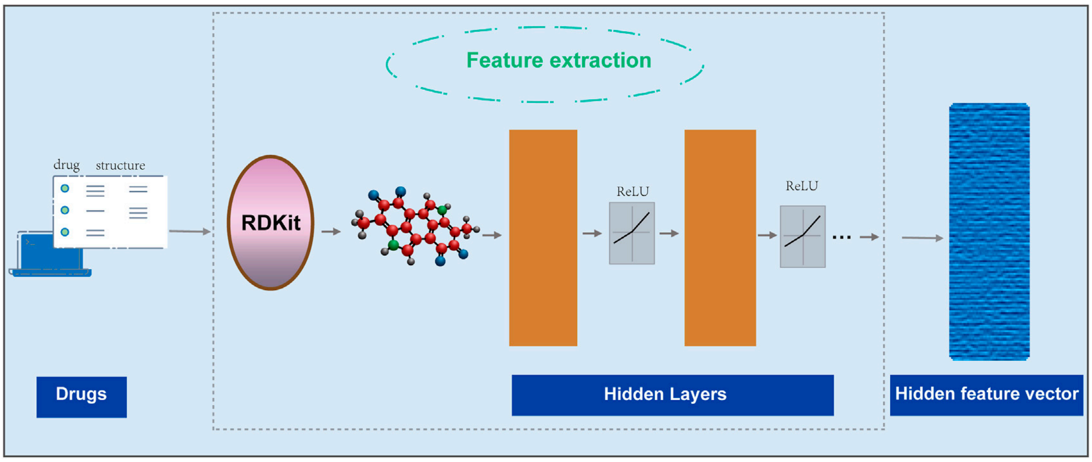


FIGURE 2
Using GNNs to extract features of small molecule drugs.

is the node i itself. We use the following definition to describe the subgraph with node v_i and a radius of r:

$$v_i^r = \left(V_i^r, E_i^r\right) \tag{1}$$

Where, $V_i^r = \left\{v_j | j\epsilon\Gamma(i,r)\right\}, E_i^r = \{e_{mn}\epsilon E| (m,n)\epsilon(\Gamma(i,r) \times \Gamma(i,r-1))\}$ Similarly, the subgraph with a radius of r can be defined for the edge e_{ij}: $e_{ij}^r = (V_i^{r-1} \cup V_j^{r-1}, E_i^r \cap E_j^r)$.

Vertex transformation function: In the molecular structure graph G, subgraph embedding can start from any vertex. $v_i^{(t)}\epsilon R^d$ is used to describe the vertex i at the t-th step of subgraph embedding information update. The update process is described as follows:

$$v_i^{(t)} = \sigma\left(v_i^{(t-1)} + \sum_{j\epsilon\Gamma(i)} h_{ij}^{(t)}\right) \tag{2}$$

Where $\sigma(x) = \frac{1}{(1+e^x)}$, $\Gamma(i)$ represents the set of neighbor node indices for vertex i. $h_{ij}^{(t)}$ is a hidden vector describing the information of neighbor node j and the edge e_{ij} between the two nodes for vertex i. It can be calculated using the following formula:

$$h_{ij}^{(t)} = max\left(0, W_{neighbor}{}^{*}\begin{bmatrix} v_j^{(t)} \\ e_{ij}^{(t)} \end{bmatrix} + b_{neighbor}\right) \tag{3}$$

Were, $W_{neighbor}\epsilon\mathbb{R}^{d\times 2d}$ is a weight matrix and $b_{neighbor}\epsilon\mathbb{R}^d$ is a bias matrix. $e_{ij}^{(t)}$ represents the t-th subgraph embedding information update between vertex i and vertex j. By summing the hidden vectors of adjacent nodes and iteratively updating, vertex embedding can gradually learn the global information of the molecular structure graph.

The edge transformation function: The process of updating edge embeddings are similar to the process of updating vertex embeddings. Here, $e_{ij}^{(t)}$ is used to represent the embedding of the edge between vertex i and vertex j. At the same time, the embeddings of adjacent vertices to the edge, $v_i^{(t)}$ and $v_j^{(t)}$, are used to update the edge embedding information. The update process is described as follows:

$$e_{ij}^{(t)} = \sigma\left(e_{ij}^{(t-1)} + g_{ij}^{(t-1)}\right) \tag{4}$$

The formula describes $g_{ij}^{(t-1)}$ as follows: $g_{ij}^{(t)} = max(0, W_{side}{}^{*}[v_i^{(t)} + v_j^{(t)}] + b_{side})$. $W_{side}\epsilon\mathbb{R}^{d\times 2d}$ is a weight matrix, and $b_{side}\epsilon\mathbb{R}^d$ is a bias vector.

Small molecule output function: To obtain the final output $y_{sm}\epsilon\mathbb{R}^d$, the model sums up the embeddings of each vertex in the molecular graph $V = \left\{v_1^{(t)}(t), v_2^{(t)}(t), \cdot\cdot\cdot, v_{|V|}^{(t)}(t)\right\}$. The process is described as follows:

$$y_{sm} = \frac{1}{|V|}\sum_{i=1}^{|V|} v_i^{(t)} \tag{5}$$

$|V|$ represents the number of vertices in the molecular graph.

Using CNNs to process miRNA sequence data

First, CNNs use filter functions to compute a hidden vector $y \in R^d$ based on the sub-sequences of the input sequence C and a weight matrix (learned parameters). The filter functions are implemented by neural networks. In CNNs, the overall function = $f(C)$ is differentiable and all parameters in $f(x)$ are learned through backpropagation (Figure 3). The specific steps are shown as follows:

Sequence input function

To apply CNNs to miRNA sequence data, First, miRNA sequences are defined as “words” consisting of n-length bases (Dong et al., 2006; Costa and De Grave, 2010), where n refers to the number of bases. Then, the miRNA sequence is divided into overlapping n-mers. In this study, to maintain a manageable and informative word vocabulary and to avoid using low-frequency sequence fragmentation in learning representations, a relatively small value of $n = 3$ was set for the number of bases. The miRNA sequence $S = x_1, x_2, \cdot\cdot\cdot, x_{|s|}$, where x_i is the i-th base pair

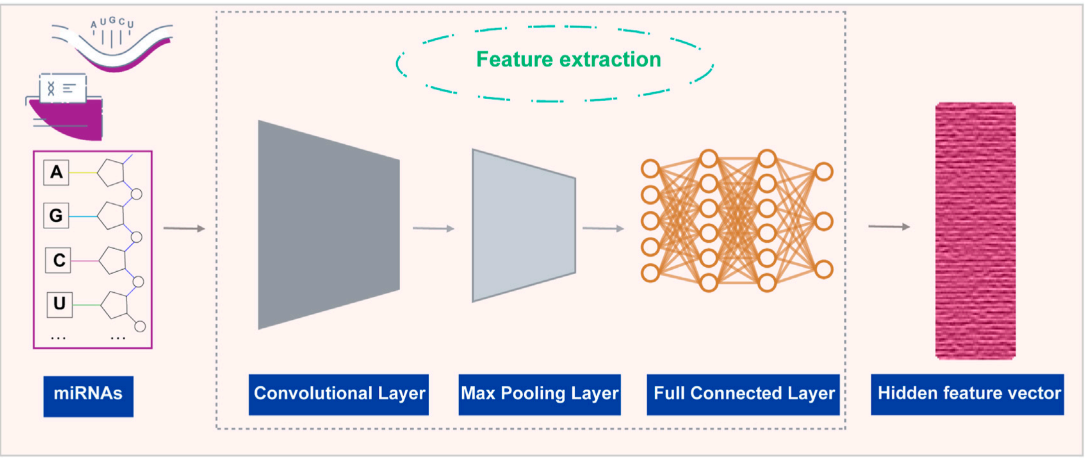


FIGURE 3
Using CNNs to extract features of miRNAs.

and $|s|$ is the length of the sequence, is then split into overlapping n-base pair segments. All words are then translated into randomly initialized embeddings, referred to as "word embeddings." The word embeddings are ordered as $X_1, X_2, \cdots, X_{|s|-1}X_{|s|}$, where $X_i \in \mathbb{R}^d$ is a d-dimensional embedding for the i-th word. Alternatively, we can consider a sequence whose elements consist of concatenated word embeddings. For example, a sequence composed of three consecutive embeddings would be $[X_1; X_2; X_3], [X_2; X_3; X_4] \cdots [X_{|S|-2}; X_{|S|-1}; X_{|s|}]$, where $[X_{i+1}; X_{i+2}; X_{i+3}]\epsilon\mathbb{R}^{3d}$ is the concatenation of X_{i+1}, X_{i+2}, and X_{i+3}. Here, $X_{i:\, i+w-1}$ refers to $[X_i; \cdots; X_{i+w-1}]$, where w is the window size. This processed sequence can be used as input for CNNs.

Filter function

Using $X_{i:\, i+w-1} = [X_i; X_{i+w-1}] = c_i^{(0)}\epsilon\mathbb{R}^{dw}$ as the input to the filter function $f(x)$, the output of the filter function is a hidden vector $c_i^{(1)}\epsilon\mathbb{R}^d$. The description of the hidden vector is as follows:

$$c_i^{(1)} = f\left(W_{conv}{}^{*}c_i^{(0)} + b_{conv}\right) \tag{6}$$

Where $f(x)$ is a non-linear activation function, $W_{conv}\epsilon\mathbb{R}^{d\times dw}$ is the weight matrix, and b_{conv} is the bias vector. By using the filter function repeatedly, multiple hidden vectors can be obtained:

$$c_i^{(t)} = f\left(W_{conv}{}^{*}c_i^{(t-1)} + b_{conv}\right) \tag{7}$$

Multiple hidden vectors form a hidden vector set $C = \left\{c_1^{(t)}, c_2^{(t)}, c_3^{(t)}, \ldots\ldots c_{|c|}^{(t)}\right\}$.

miRNA sequence output function. In order to obtain the final output $y_{miRNA}\epsilon\mathbb{R}^d$ from $C = \left\{c_1^{(t)}, c_2^{(t)}, c_3^{(t)}, \ldots\ldots c_{|c|}^{(t)}\right\}$, the average of C is taken. The process is described as follows:

$$y_{miRNA} = \frac{1}{|C|}\sum_{i=1}^{|C|} c_i^{(t)} \tag{8}$$

$|C|$ denotes the number of elements in set C.

Neural attention mechanism for predicting potential associations between miRNAs and small molecule drugs

GCNNMMA employs a neural attention mechanism to infer interactions between small molecules and subsequences in miRNA sequences. In the collection of hidden vector sequences $C = \left\{c_1^{(t)}, c_2^{(t)}, c_3^{(t)}, \ldots\ldots c_{|c|}^{(t)}\right\}$ for miRNA sub-sequences, each hidden vector sequence represents its corresponding miRNA sub-sequence. Different miRNA sub-sequences have different binding abilities and probabilities with small molecules. A neural attention mechanism is used to assign corresponding weights to each sub-sequence in the miRNA hidden vector sequence collection, which represents the importance of its association with small molecules. The weight calculation process is described as follows:

$$h_{sm} = f\left(W_{inter}{}^{*}y_{sm} + b_{inter}\right) \tag{9}$$

$$h_i = f\left(W_{inter}{}^{*}c_i + b_{inter}\right) \tag{10}$$

$$\alpha_i = \sigma\left(h_{sm}^T{}^{*}h_i\right) \tag{11}$$

Where W_{inter} is the weight matrix and b_{inter}_inter is the bias vector. α_i represents the strength of interaction between small molecules and miRNA sub-sequences. Based on the calculated attention weights, the final weighted sum can be obtained, as shown below:

$$y_{miRNA} = \sum_{i=1}^{|C|} \alpha_{i*}h_i \tag{12}$$

Finally, the model obtains the final classification output vector $Z\epsilon R^2$ by jointly considering y_{miRNA} and y_{sm}:

$$Z = W_{output}{}^{*}\left[y_{miRNA}; y_{sm}\right] + b_{output} \tag{13}$$

Where $W_{output} \in R^{2\times 2d}$ is the weight matrix and $b_{output} \in R^2$ is the bias vector. Finally, the output vector $Z = [y0, y1]]$ is passed through the softmax function to compute the associated probabilities:

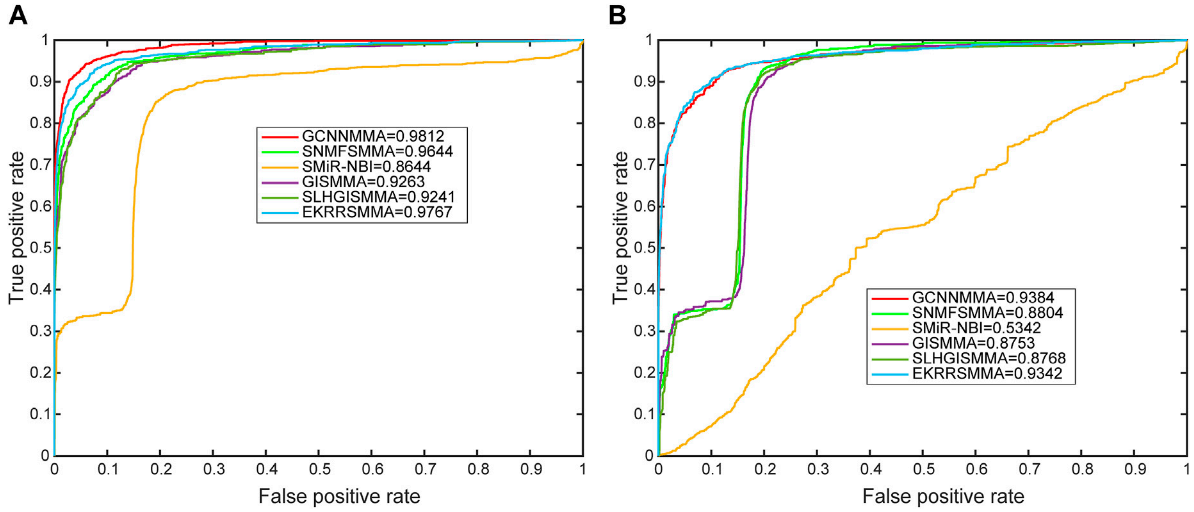


FIGURE 4
The ROC curves for GCNNMMA and benchmark algorithms for 5-fold CV on the **(A)** dataset 1 and **(B)** dataset 2.

$$P_t = \frac{exp(y_t)}{\sum_i y_i} \tag{14}$$

Results

Performance of GCNNMMA in the cross-validation

In this work, we compared the performance of the latest five models [SMiR-NBI (Li et al., 2016), GISMMA (Guan et al., 2018), SLHGISMMA (Yin et al., 2019), SNMFSMMA (Zhao et al., 2020), EKRRSMMA (Wang et al., 2022b)] with GCNNMMA, and conducted 5-fold cross-validation (CV) on both dataset 1 and dataset 2 to evaluate the predictive performance of GCNNMMA. All predicted small molecule miRNA pairs were ranked according to the obtained scores. Based on the rankings, we used receiver operating characteristic (ROC) curves to illustrate the performance of our models in the cross-validation runs. As shown in Figure 4, we found that GCNNMMA achieved the best predictive performance on both dataset 1 (AUC = 0.9812) and dataset 2 (AUC = 0.9384). This suggests that GCNNMMA performed the best in predicting the correlation between small molecule drugs and miRNAs.

GCNNMMA is superior to other popular methods in predicting miRNAs associated with new small molecule drugs

It is important to examine the performance of the above method in predicting new miRNAs related to small molecule drugs, in addition to testing the performance of global prediction of small molecule drug-miRNA relationships. A leave-one-out experiment is used to evaluate the ability of the algorithm to predict miRNAs related to new small molecule drugs. To compare the fairness of the test, we still use ROC as the indicator of predictive performance. The local LOOCV experiment was carried on the dataset 1 and dataset 2 (see Figure 5). GCNNMMA showed a higher performance over other approaches in terms of AUC on the dataset 2. Specifically, GCNNMMA obtained AUC value of 0.9367, outperforming that of SMiR-NBI (AUC = 0.6754), GISMMA (AUC = 0.8473), SLHGISMMA (AUC = 0.8532), SNMFSMMA (AUC = 0.9254), EKRRSMMA (AUC = 0.8751). In addition, we can find that the performance of GCNNMMA is also second only to SNMFSMMA on the dataset 1. This also sufficient GCNNMMA is also the best way to predict m miRNAs related to new small molecule drugs.

Case studies: identifying the relationship between small molecule drugs and miRNAs associated with liver cancer

To further verify the reliability capability of GCNNMMA, we take all known miRNAs-small molecule drug associations in the SM2miR dataset 1 as the training set, and regard the missing miRNAs-small molecule drug associations as candidate sets. After GCNNMMA predicted the interaction probabilities of all candidate miRNAs-small molecule drug associations, we then ranked them according to the predicted probabilities so that the top-ranked associations were most likely to interact. We also validated these top 30 associations by searching for corresponding PubMed literature, as shown in Table 2. Among the top 10, 20, and 30 predicted associations, we were able to validate 6, 12, and 20 associations, respectively through literature search. In the top 10 predicted associations, we found that 5 different miRNAs were associated with Fluorouracil (CID: 3385), a small molecule drug that belongs to the class of pyrimidine analogs and is an anti-metabolic drug used to treat tumors. It interferes with DNA synthesis by

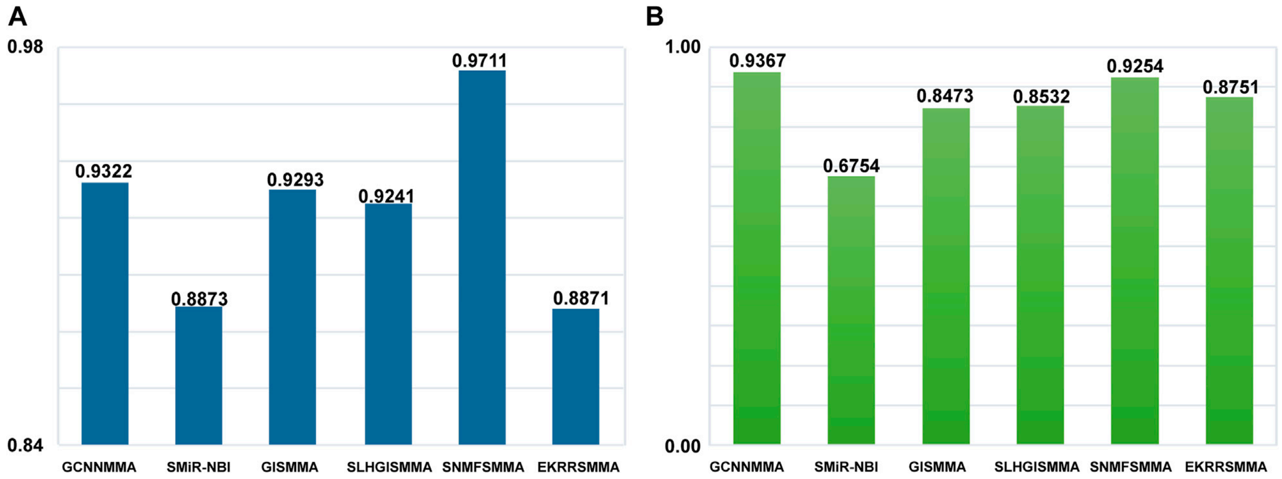


FIGURE 5
The ROC curves for GCNNMMA and benchmark algorithms for local LOOCV on the **(A)** dataset 1 and **(B)** dataset 2.

TABLE 2 Predicting the top 30 small molecule drugs associated with miRNAs.

Rank	CID	miRNA	Evidence (PubMed)	Rank	CID	miRNA	Evidence (PubMed)
1	3,229	hsa-mir-212	28,131,841	16	5,757	hsa-mir-542	17,765,232
2	3,385	hsa-mir-149	27,415,661	17	5,757	hsa-mir-663a	32,215,262
3	3,385	hsa-mir-1915	22,121,083	18	6,013	hsa-mir-135a-1	32,735,753
4	3,385	hsa-mir-203a	25,526,515	19	6,013	hsa-mir-29a	26,296,572
5	3,385	hsa-mir-320a	unconfirmed	20	10,635	hsa-mir-32	20,945,501
6	3,385	hsa-mir-483	unconfirmed	21	10,635	hsa-mir-630	20,945,501
7	3,385	hsa-mir-519c	26,386,386	22	31,401	hsa-mir-603	20,689,055
8	3,385	hsa-mir-617	21,743,970	23	36,462	hsa-mir-26b	31,985,026
9	5,311	hsa-mir-126	unconfirmed	24	36,462	hsa-mir-663a	31,639,426
10	5,311	hsa-mir-409	unconfirmed	25	60,750	hsa-mir-139	33,300,085
11	5,311	hsa-mir-574	unconfirmed	26	60,750	hsa-mir-211	25,789,319
12	5,311	hsa-mir-595	unconfirmed	27	60,750	hsa-mir-299	28,131,841
13	5,311	hsa-mir-744	unconfirmed	28	60,750	hsa-mir-326	unconfirmed
14	5,311	hsa-mir-760	unconfirmed	29	60,953	hsa-mir-137	22,740,910
15	5,757	hsa-mir-17	24,283,290	30	216,239	hsa-mir-664a	unconfirmed

blocking the conversion of deoxyuridine monophosphate to thymidine monophosphate (Ellison, 1961). Currently, Fluorouracil is used to treat diseases such as actinic keratosis, breast cancer, colon cancer, pancreatic cancer, gastric cancer, liver cancer, and superficial basal cell carcinoma (Lecluse and Spuls, 2015; Guo et al., 2020). Among the top 20 predicted associations, we discovered novel small molecule drugs associated with miRNAs and Estradiol (CID:5757), Testosterone (CID: 6013), and Dihydrotestosterone (CID: 10635). These three hormones have high bioavailability and can enhance cellular metabolism. These three hormones have high bioavailability and can enhance cellular metabolism (Pentikäinen et al., 2000). Among the top 30 predicted associations, we found that the small molecule drugs Etoposide (CID: 36462) (Wang et al., 2003) and Gemcitabine (CID: 60750) are used for cancer treatment. Etoposide is a semi-synthetic derivative with anti-tumor activity. It inhibits DNA synthesis by forming a complex with topoisomerase II and DNA, inducing double-stranded DNA breaks and preventing repair by blocking the binding of topoisomerase II. Accumulation of DNA breaks prevents cells from entering mitosis, leading to cell death (Uesaka et al., 2007).

Gemcitabine (CID: 60750) is a nucleoside analog used in chemotherapy that, like fluorouracil and other pyrimidine analogs, replaces a structural group of nucleic acids in DNA replication to form cytidine in this case. The formation of cytidine stops tumor growth as new nucleosides cannot attach to the "defective" nucleosides, leading to cell apoptosis (cell "suicide") (Hastak et al., 2010; Vogl et al., 2010). Currently, Gemcitabine is used to treat cancers such as non-small cell lung cancer, pancreatic cancer, bladder cancer, and breast cancer.

Discussion

The development of deep learning provides new approaches for predicting the association between small molecule drugs and miRNAs. We developed a prediction model called GCNNMMA based on graph neural networks (GNNs) and convolutional neural networks (CNNs), and validated its performance on two datasets. Experimental results show that GCNNMMA exhibited the best performance in the datasets. Compared with previous similarity-based models, our model extracts the characteristic information of small molecule drugs and miRNAs through GNN and CNN networks, avoiding the dependence on known association information. Furthermore, when predicting the top 30 associations in the dataset, GCNNMMA identified Gemcitabine (CID: 60750) related to hsa-mir-139 and Fluorouracil (CID: 3385) related to hsa-mir-149, both of which are used in cancer treatment by targeting the relevant miRNAs to inhibit cell division and induce cancer cell death. While GCNNMMA achieved good performance, there is still room for improvement, such as integrating multi-source data which remains a challenging problem. In the future, incorporating more data sources, such as miRNA spatial structure data and miRNA precursor data, could improve GCNNMMA. In addition, three-dimensional structural information can better reflect spatial information. One of the future research directions is to utilize the three-dimensional structural information of miRNAs and small molecule drugs to improve prediction accuracy.

Data availability statement

The program and data used in this study are publicly available at: https://github.com/niuzheyu123/GCNNMMA.git.

Author contributions

XZ conceived the study. ZN, XG, ZX, SZ, and HS performed experiments and data analysis. HW, ML, KX, and CM interpreted the data analysis. ZN, HZ HG, QL FY, XS, JL, and XZ drafted the manuscript and critically revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study is supported by Natural Science Foundation of Shandong Province, China (Grant Nos ZR2020MH259, ZR2016HB40), Natural Science Foundation of China Youth Project (Grant No. 81802379).

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.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). "Neural machine translation by jointly learning to align and translate.", arXiv preprint arXiv:1409.0473.

Bartel, D. P. (2004). MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* 1162, 281–297. doi:10.1016/s0092-8674(04)00045-5

Beermann, J., Piccoli, M. T., Viereck, J., and Thum, T. (2016). Non-coding RNAs in development and disease: Background, mechanisms, and therapeutic approaches. *Physiol. Rev.* 96, 1297–1325. doi:10.1152/physrev.00041.2015

Cai, L., Lu, C., Xu, J., Meng, Y., Wang, P., Fu, X., et al. (2021). Drug repositioning based on the heterogeneous information fusion graph convolutional network. *Briefings Bioinforma.* 22, bbab319. doi:10.1093/bib/bbab319

Chen, X., Zhou, C., Wang, C. C., and Zhao, Y. (2021). Predicting potential small molecule–miRNA associations based on bounded nuclear norm regularization. *Briefings Bioinforma.* 22, bbab328. doi:10.1093/bib/bbab328

Chen, X., Guan, N. N., Sun, Y. Z., Li, J. Q., and Qu, J. (2020). MicroRNA-small molecule association identification: From experimental results to computational models. *Briefings Bioinforma.* 21, 47–61. doi:10.1093/bib/bby098

Chen, Y. (2015). "Convolutional neural network for sentence classification,". MS thesis (University of Waterloo) Computer Science.

Costa, F., and De Grave, K. (2010). "Fast neighborhood subgraph pairwise distance kernel," in Proceedings of the 26th International Conference on Machine Learning (Madison, WI, United States: Omnipress), 255–262.

Dong, Q.-W., Wang, X., and Lin, L. (2006). Application of latent semantic analysis to protein remote homology detection. *Bioinformatics* 22.3, 285–290. doi:10.1093/bioinformatics/bti801

Ellison, R. R. (1961). Clinical applications of the fluorinated pyrimidines. *Med. Clin. N. Am.* 45.3, 677–688. doi:10.1016/s0025-7125(16)33880-9

Guan, N.-N., Sun, Y. Z., Ming, Z., Li, J. Q., and Chen, X. (2018). Prediction of potential small molecule-associated microRNAs using graphlet interaction. *Front. Pharmacol.* 9, 1152. doi:10.3389/fphar.2018.01152

Guo, P., Pi, C., Zhao, S., Fu, S., Yang, H., Zheng, X., et al. (2020). Oral co-delivery nanoemulsion of 5-fluorouracil and curcumin for synergistic effects against liver cancer. *Expert Opin. Drug Deliv.* 17.10, 1473–1484. doi:10.1080/17425247.2020.1796629

Hastak, K., Alli, E., and JamesFord, M. (2010). Synergistic chemosensitivity of triple-negative breast cancer cell lines to poly(ADP-Ribose) polymerase inhibition, gemcitabine, and cisplatin. *Cancer Res.* 70.20, 7970–7980. doi:10.1158/0008-5472.CAN-09-4521

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic acids Res.* 47, D1102–D1109. doi:10.1093/nar/gky1033

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). "miRBase: from microRNA sequences to function." *Nucleic acids Res.* 47. D155–D162. doi:10.1093/nar/gky1141

Lecluse, L. L. A., and Spuls, P. I. (2015). Photodynamic therapy versus topical imiquimod versus topical fluorouracil for treatment of superficial basal-cell carcinoma: A single blind, non-inferiority, randomised controlled trial: A critical appraisal. *Br. J. Dermatology* 172.1, 8–10. doi:10.1111/bjd.13460

Li, J., Lei, K., Wu, Z., Li, W., Liu, G., Liu, J., et al. (2016). Network-based identification of microRNAs as potential pharmacogenomic biomarkers for anticancer drugs. *Oncotarget* 7.29, 45584–45596. doi:10.18632/oncotarget.10052

Liu, W., Hui, L., and Li, H. (2022). Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder. *Briefings Bioinforma.* 23, 3. doi:10.1093/bib/bbac104

Liu, W., Yang, Y., Lu, X., Fu, X., Sun, R., Yang, L., et al. (2023). Nsrgrn: A network structure refinement method for gene regulatory network inference. *Briefings Bioinforma.*, bbad129. doi:10.1093/bib/bbad129

Liu, X., Wang, S., Meng, F., Wang, J., Zhang, Y., Dai, E., et al. (2013). SM2miR: A database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics* 29.3, 409–411. doi:10.1093/bioinformatics/bts698

Lv, Y., Wang, S., Meng, F., Yang, L., Wang, Z., Wang, J., et al. (2015). Identifying novel associations between small molecules and miRNAs based on integrated molecular networks. *Bioinformatics* 31.22, 3638–3644. doi:10.1093/bioinformatics/btv417

Peng, L., Cheng, Y., Yifan, C., and Wei, L. (2023). Predicting CircRNA-Disease associations via feature convolution learning with heterogeneous graph attention network. *IEEE J. Biomed. Health Inf.*, 1–11. doi:10.1109/JBHI.2023.3260863

Peng, L., Tu, Y., Huang, L., Li, Y., Fu, X., and Chen, X. (2022). Daestb: Inferring associations of small molecule–miRNA via a scalable tree boosting model based on deep autoencoder. *Briefings Bioinforma.* 23.6, bbac478. doi:10.1093/bib/bbac478

Pentikäinen, V., Erkkilä, K., Suomalainen, L., Parvinen, M., and Dunkel, L. (2000). Estradiol acts as a germ cell survival factor in the human testis *in vitro*. *J. Clin. Endocrinol. Metabolism* 85.5, 2057–2067. doi:10.1210/jcem.85.5.6600

Ruepp, A., Kowarsch, A., Schmidl, D., Buggenthin, F., Brauner, B., Dunger, I., et al. (2010). PhenomiR: A knowledgebase for microRNA expression in diseases and biological processes. *Genome Biol.* 11, R6–R11. doi:10.1186/gb-2010-11-1-r6

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Trans. neural Netw.* 20.1, 61–80. doi:10.1109/TNN.2008.2005605

Uesaka, T., Shono, T., Kuga, D., Suzuki, S. O., Niiro, H., Miyamoto, K., et al. (2007). Enhanced expression of DNA topoisomerase II genes in human medulloblastoma and its possible association with etoposide sensitivity. *J. neuro-oncology* 84, 119–129. doi:10.1007/s11060-007-9360-0

Vogl, T. J., Naguib, N. N. N., Nour-Eldin, N. E. A., Eichler, K., Zangos, S., and Gruber-Rouh, T. (2010). Transarterial chemoembolization (TACE) with mitomycin C and gemcitabine for liver metastases in breast cancer. *Eur. Radiol.* 20, 173–180. doi:10.1007/s00330-009-1525-0

Wang, C.-C., Chen, X., Qu, J., Sun, Y. Z., and Li, J. Q. (2019). Rfsmma: A new computational model to identify and prioritize potential small molecule–mirna associations. *J. Chem. Inf. Model.* 59, 1668–1679. doi:10.1021/acs.jcim.9b00129

Wang, C.-C., Zhu, C.-C., and Chen, X. (2022). Ensemble of kernel ridge regression-based small molecule–miRNA association prediction in human disease. *Briefings Bioinforma.* 23, bbab431. doi:10.1093/bib/bbab431

Wang, S.-H., Wang, C. C., Huang, L., Miao, L. Y., and Chen, X. (2022). Dual-network collaborative matrix factorization for predicting small molecule-miRNA associations. *Briefings Bioinforma.* 23, bbab500. bbab500. doi:10.1093/bib/bbab500

Wang, X., Furukawa, T., Nitanda, T., Okamoto, M., Sugimoto, Y., Akiyama, S. I., et al. (2003). Breast cancer resistance protein (BCRP/ABCG2) induces cellular resistance to HIV-1 nucleoside reverse transcriptase inhibitors. *Mol. Pharmacol.* 63.1, 65–72. doi:10.1124/mol.63.1.65

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic acids Res.* 46.D1, D1074–D1082. doi:10.1093/nar/gkx1037

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-impute: An accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 36.10, 3139–3147. doi:10.1093/bioinformatics/btaa109

Xu, J., Meng, Y., Lu, C., Cai, L., Zeng, X., et al. (2023). Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. *Cell Rep. Methods* 3, 100382. doi:10.1016/j.crmeth.2022.100382

Yin, J., Chen, X., Wang, C. C., Zhao, Y., and Sun, Y. Z. (2019). Prediction of small molecule–microRNA associations by sparse learning and heterogeneous graph inference. *Mol. Pharm.* 16.7, 3157–3166. doi:10.1021/acs.molpharmaceut.9b00384

Zhang, Z., Xu, J., Wu, Y., Liu, N., Wang, Y., and Liang, Y. (2023). CapsNet-LDA: Predicting lncRNA-disease associations using attention mechanism and capsule network based on multi-view data. *Briefings Bioinforma.* 24, bbac531. doi:10.1093/bib/bbac531

Zhao, Y., Chen, X., Yin, J., and Qu, J. (2020). Snmfsmma: Using symmetric nonnegative matrix factorization and kronecker regularized least squares to predict potential small molecule-microRNA association. *RNA Biol.* 17.2, 281–291. doi:10.1080/15476286.2019.1694732

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 12 June 2023
DOI 10.3389/fgene.2023.1170940

OPEN ACCESS

EDITED BY
Junlin Xu,
Hunan University, China

REVIEWED BY
Qiaoming Liu,
Harbin Institute of Technology, China
Sabine Colnot,
Ligue Nationale Contre le Cancer, France

*CORRESPONDENCE
Katja Eloranta,
✉ katja.eloranta@helsinki.fi

†These authors have contributed equally to this work and share first authorship

RECEIVED 21 February 2023
ACCEPTED 29 May 2023
PUBLISHED 12 June 2023

CITATION
Nousiainen R, Eloranta K, Isoaho N, Cairo S, Wilson DB, Heikinheimo M and Pihlajoki M (2023), *UBE2C* expression is elevated in hepatoblastoma and correlates with inferior patient survival. *Front. Genet.* 14:1170940. doi: 10.3389/fgene.2023.1170940

UBE2C expression is elevated in hepatoblastoma and correlates with inferior patient survival

Ruth Nousiainen[1†], Katja Eloranta[1*†], Noora Isoaho[2], Stefano Cairo[3,4,5], David B. Wilson[6,7], Markku Heikinheimo[1,7,8] and Marjut Pihlajoki[1]

[1]Pediatric Research Center, Children's Hospital, Helsinki University Hospital, University of Helsinki, Helsinki, Finland, [2]Division of Micro and Nanosystems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden, [3]Champions Oncology, Hackensack, NJ, United States, [4]Istituto di Ricerca Pediatrica, Padova, Italy, [5]XenTech, Evry, France, [6]Department of Developmental Biology, Washington University School of Medicine, St. Louis, MO, United States, [7]Department of Pediatrics, Washington University School of Medicine, St. Louis Children's Hospital, St. Louis, MO, United States, [8]Faculty of Medicine and Health Technology, Center for Child, Adolescent and Maternal Health Research, Tampere University, Tampere, Finland

Hepatoblastoma (HB) is the most common malignant liver tumor among children. To gain insight into the pathobiology of HB, we performed RNA sequence analysis on 5 patient-derived xenograft lines (HB-243, HB-279, HB-282, HB-284, HB-295) and 1 immortalized cell line (HUH6). Using cultured hepatocytes as a control, we found 2,868 genes that were differentially expressed in all of the HB lines on mRNA level. The most upregulated genes were *ODAM*, *TRIM71*, and *IGDCC3*, and the most downregulated were *SAA1*, *SAA2*, and *NNMT*. Protein-protein interaction analysis identified ubiquitination as a key pathway dysregulated in HB. *UBE2C*, encoding an E2 ubiquitin ligase often overexpressed in cancer cells, was markedly upregulated in 5 of the 6 HB cell lines. Validation studies confirmed UBE2C immunostaining in 20 of 25 HB tumor specimens *versus* 1 of 6 normal liver samples. The silencing of *UBE2C* in two HB cell models resulted in decreased cell viability. RNA sequencing analysis showed alterations in cell cycle regulation after *UBE2C* knockdown. *UBE2C* expression in HB correlated with inferior patient survival. We conclude that *UBE2C* may hold prognostic utility in HB and that the ubiquitin pathway is a potential therapeutic target in this tumor.

KEYWORDS

hepatoblastoma, liver tumor, pediatric cancer, ubiquitin, *UBE2C*

1 Introduction

Hepatoblastoma (HB) is the most common malignant pediatric liver neoplasm with an annual incidence of 1.9/1,000,000 (Aronson and Meyers, 2016; Feng et al., 2019) that has been increasing over the past decades (Linabery and Ross, 2008). The etiology of most cases of HB remains unknown, but preterm birth, birthweight less than 2500 g, and certain genetic conditions such as Familial Adenomatous Polyposis and Beckwith-Wiedemann syndrome are associated with increased risk of HB (Spector and Birch, 2012; Paquette et al., 2019). Wnt/β-catenin signaling has been identified as one of the pathways altered in the majority of HB tumors, but other molecular pathways involved in HB pathogenesis are not yet fully understood (Udatsu et al., 2001; Armengol et al., 2011). Current treatment of HB includes complete surgical resection combined with doxorubicin and cisplatin- or carboplatin-based neoadjuvant and adjuvant chemotherapy (Zsíros et al., 2010). These forms of chemotherapy are effective but often lead to serious long-term side effects including cardiotoxicity, ototoxicity, and nephrotoxicity (Spector and Birch, 2012; Volkova and Russell, 2012). Although the prognosis of HB has improved over the years, 20%–30% of HB patients still respond poorly to current treatment modalities (Sivaprakasam et al., 2011), so new therapeutic targets are needed.

Metabolic reprogramming is one of the hallmarks of cancer (Faubert et al., 2020). Highly proliferating tumor cells need to adapt to conditions such as hypoxia and lack of nutrients and thus require metabolic reprogramming to enhance their survival. Many of the oncogenes and tumor suppressors are participating in dysregulation of metabolic pathways in cancer (Nong et al., 2023). Also, genes coding for metabolic enzymes have been described to be mutated or aberrantly expressed in several tumor types (Sreedhar and Zhao, 2018). Major changes include alterations in glucose metabolism, known as Warburg effect, as well as in amino acid and lipid metabolism (Counihan et al., 2018). Alterations in genes regulating ubiquitination and deubiquitination and their role as modulators of the metabolic changes of tumor cells are also known to be essential in cancer progression (Sun et al., 2020).

Some changes in metabolic genes have already been demonstrated to be present in HB. In HB tumors, activating mutations in the Wnt/β-catenin pathway genes lead to altered glucose metabolism mediated by upregulation of *GLUT3* (Crippa et al., 2017). Immortalized HepG2 cells, originally derived from a HB, exhibit deranged bile acid metabolism (Kullak-Ublick et al., 1996) due in part to and downregulated *SLC10A1* (Wang et al., 2020). In the same study, the authors also found that downregulation of *SLC10A1* resulted in upregulated adenosine metabolism. Retinol metabolism and cytochrome P450 pathway have both been demonstrated to be downregulated in HB (Sekiguchi et al., 2020). Despite these findings, the gene expression behind the metabolic alterations taking place in HB still remain poorly understood.

The objective of this study was to characterize the landscape of metabolic genes in HB using RNA sequencing data and bioinformatics analyses. Our overarching goal was to identify potential treatment targets and novel biomarkers in HB.

2 Materials and methods

2.1 RNA sequencing and microarray datasets

Raw RNA sequencing datasets from previously published studies were obtained from Gene Expression Omnibus (GEO) database of National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/geo/) and European Genome-phenome Archive (EGA) (https://ega-archive.org/). Accession numbers were as follows: EGAS00001004827/EGAD00001006621 (HB-282, HB-295, HB-279, HB-284, HB-243), GSE140520 (PHH-D3, 1–3), GSE83518 (1HUH6HB, 2HUH6HB). Raw microarray data of gene expression in 53 HB tissue samples and 14 noncancerous liver tissue samples collected from the HB patients at the time of surgery were acquired from GEO, accession number GSE131329.

2.2 Identification of differentially expressed genes and differentially expressed metabolic genes

RNA sequencing dataset files were analyzed with Chipster software (https://chipster.rahtiapp.fi/) (Kallio et al., 2011). Reads were preprocessed using Trimmomatic and aligned to human reference genome (GRCh38) using HISAT2 (Kim et al., 2015). Reads per genes were counted with HTseq (Anders et al., 2015). Differential expression analysis was conducted with the edgeR package (Robinson et al., 2009). Differentially expressed genes (DEGs) were then filtered using cut-off criteria adjusted to *p*-value <0.05 and $|\log_2 FC| \geq 1.0$.

Microarray data were analyzed with Chipster plus the normalization tool for Affymetrix gene arrays (Li, 2001; Irizarry et al., 2003). Statistical tests were conducted using the “Two group tests” tool (empirical Bayes as test and BH as *p*-value adjustment method) (Smyth, 2004).

Human metabolic genes were obtained from The Virtual Metabolic Human database (VMH) (Supplementary Table S1) (Noronha et al., 2019). A list of DEGs in each cell line was compared to a list of human metabolic genes to further filter the results. Duplicates were removed from the list leaving 3,285 unique genes.

2.3 Protein-protein interaction (PPI) network construction

PPI networks of differentially expressed metabolic genes in each cell line were constructed with the online tool of STRING database (Szklarczyk et al., 2019) using 0.7 as minimum required interaction score. Results were visualized with Cytoscape (Shannon et al., 2003). Highly interconnected areas (clusters) of these interaction maps were identified and scored using the Cytoscape plugin Molecular Complex Detection (MCODE) (Bader and Hogue, 2003) using the following criteria: degree cut-off: 2; haircut: yes; fluff: no; node score cut-off: 0.2; K-core: 2; max. Depth: 100. Based on this score, highest-ranking cluster of each cell line was chosen to be investigated more thoroughly. PPI network of UBE2C was constructed using STRING with 0.4 as minimum required interaction score.

2.4 Kyoto encyclopedia of genes and genomes (KEGG) and gene ontology (GO) pathway enrichment analyses

Genes in each cell line’s highest-scoring cluster were uploaded to Enrichr (Chen et al., 2013; Kuleshov et al., 2016; Xie et al., 2021). Results of KEGG pathway and GO Biological Processes term enrichment analyses were imported from the website. Results with adjusted *p*-value of <0.05 were considered significant. The top 5 terms with lowest adjusted *p*-values from both KEGG and GO were chosen for each HB cell line. Python programming language (Python Software Foundation) with Matplotlib (Hunter, 2007), NumPy (Harris et al., 2020), and pandas (McKinney, 2010) libraries were used for handling and plotting this data.

2.5 Statistical analysis of clinical variables

Student's t-test, Mann-Whitney U test, and receiver operating characteristic (ROC) curves were used to analyze microarray gene expression data and clinical variables provided in GSE131329 dataset. Statistical significance was set to *p*-value <0.05. Analyses were conducted with R software (v. 4.0.3) or GraphPad Prism (v. 8.4.2; GraphPad, San Diego, CA, United States).

2.6 Gene co-expression analysis

An online tool for gene co-expression analysis, GeneFriends v 5.0 (Raina et al., 2022), was used to analyze the gene co-expression of *UBE2C*. Following input parameters were used: Species; *Homo Sapiens*, Data Source; SRA, Tissue: All tissue types, Object type; Gene, Seed gene: *UBE2C*, Pearson correlation threshold: 0.75.

2.7 Patient samples

Formalin-fixed paraffin-embedded (FFPE) HB tumor samples [n = 24, median age in years 3.37 (range 0.23–11.75)] and normal liver control samples (NL, n = 6, median age in years 12.5 (range 3–26)) were obtained from the Helsinki Biobank at Helsinki University Hospital. HB samples were originally collected at the time of surgical treatment from patients treated in Children's Hospital, Helsinki University Hospital between 1 January 1991, and 31 December 2017. Prior to resection, majority of patients had received preoperative chemotherapy. NL samples were collected from liver transplantation donors in Helsinki University Hospital. This study was approved by Helsinki University Hospital institutional ethical committee (HUS/3319/2018) and conducted in accordance with Finnish bylaws. Informed consent was obtained when samples were deposited to the Helsinki Biobank.

2.8 Immunohistochemistry

Samples were heated for 30 min at 60°C oven and deparaffinized with NeoClear (Merck-Millipore, Darmstadt, Germany). Target retrieval was performed using pH 9.0 target retrieval solution for 30 min at + 98°C (Dako, Glostrup, Denmark). Novolink Polymer Detection System Kit (Leica, Newcastle, United Kingdom) was used to block endogenous peroxidase activity and nonspecific binding. Samples were incubated with UBE2C polyclonal antibody at +4°C overnight (dilution 1:1,500; #PA5-102791; Invitrogen, Thermo Fisher Scientific). A polymerized reporter enzyme staining system (Novolink Polymer Detection System Kit) was used to visualize the bound antibody. UBE2C immunoreactivity was scored as strong nuclear staining (positive) or negative by two separate observers. Images were generated using 3DHISTECH Pannoramic 250 FLASH II digital slide scanner at Genome Biology Unit supported by HiLIFE and the Faculty of Medicine, University of Helsinki, and Biocenter Finland.

2.9 HB *in vitro* models

HB cell line HUH6 was obtained from Japanese Collection of Research Bioresources Cell Bank (Osaka, Japan). Cells were maintained in Dulbecco's modified Eagle's medium (DMEM)-glutaMAX (Gibco) supplemented with 10% FBS (Gibco), 100 U/mL penicillin (Gibco), and 100 μg/mL streptomycin sulfate (Gibco, Waltham, MA, United States). HB cell line HB-243 from patient-derived xenograft (PDX) was provided by XenTech (Evry, France) (Kats et al., 2019). HB-243 cells were cultured in Advanced DMEM/F12 (Gibco, Waltham, MA, United States) supplemented with 8% fetal bovine serum (FBS) (Gibco), 2 mM glutaMAX (Gibco), 100 U/mL penicillin (Gibco), 100 μg/mL streptomycin sulfate (Gibco) and 20 μM rock kinase inhibitor Y-27632 (S1049; SelleckChem, Houston, TX, United States). Absence of *mycoplasma* was regularly confirmed with PCR-based method (PromoCell, Heidelberg, Germany).

2.10 *UBE2C* silencing

UBE2C expression in HUH6 and HB-243 cell models was silenced by small interfering RNA (siRNA) transfection. The cells were exposed to 25 nM *UBE2C* ON-TARGETplus SMARTpool siRNA (cat# L-004693-00-0005) or ON-TARGETplus non-targeting (NT) control siRNA (cat# D-001810-10-05; both purchased from Horizon Discovery, Cambridge, United Kingdom). Dharmafect 4 (Horizon Discovery) was used to deliver the siRNAs into the HUH6 cells using the protocol provided. Knockdown efficacy was evaluated at mRNA and at protein level 48 h after transfection.

2.11 RNA and protein extraction

Total RNA and protein were extracted from cultured HUH6 and HB-243 cell models using NucleoSpin RNA/Protein extraction kit (Macherey-Nagel, Düren, Germany). Instructions provided by the manufacturer were followed.

2.12 Quantitative real-time polymerase chain reaction

Reverse transcription was carried out using iScript cDNA Synthesis Kit (Bio-Rad, Hercules, CA, United States). Quantitative polymerase chain reaction (qPCR) was performed using PowerUp SYBR Green Master Mix (Thermo Fisher Scientific, Fremont, CA, United States). The geometric mean of *B2M* and *PPIG* served as a reference. Primer sequenced were designed as follows: *B2M* 5′- GAT GAG TAT GCC TGC CGT GT—3′ (forward), 5′- CTG CTT ACA TGT CTT GAT CCC A- 3′ (reverse); *PPIG* 5′ -CAA TGG CCA ACA GAG GGA AG—3′(forward), 5′—CCA AAA ACA TGA TGC CCA—3′ (reverse); *UBE2C* 5′- CCG CCC GTA AAG G—3′ (forward), 5′- CTC AGG TCT TCA TAT ACT GTT CCA G -3′ (reverse).

2.13 Western blotting

Equal amounts of protein (10 μg) were loaded into Mini-Protein TGX stain-free gels (Bio-Rad) and separated using gel electrophoresis. Proteins were transferred to polyvinyl-fluoride membrane and 5% skimmed milk in Tris-buffered saline-Tween^{20} was utilized to block unspecific binding. Primary antibody incubation was performed at room temperature for overnight (UBE2C at dilution 1:500; #14234S; Cell Signaling Technology Inc., Danvers, MA, United States). Secondary antibody incubation was carried out at room temperature for 1 h (goat anti-rabbit IgG at dilution 1:10,000; #111-035-144, Jackson ImmunoResearch, West Grove, PA, United States). Protein bands were visualized using Enhanced Chemiluminescence detection kit (Amersham ECL reagent; GE Healthcare, Barrington, IL, United States). Protein quantification was performed with Image Lab software (version 6.0, Bio-Rad) by normalizing UBE2C band intensities to amount of total protein in corresponding lane utilizing stain-free technology (Gürtler et al., 2013).

2.14 RNA sequencing of *UBE2C* silenced HUH6 cells

HUH6 cells were cultured and treated with either *UBE2C* or non-targeting siRNAs. RNA and protein extraction were conducted as described above. Prior to sequencing, RNA concentration, quality, and integrity were assessed using Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, United States) and the TapeStation system (Agilent, Glostrup, Denmark). After quality assessment, RNA libraries were constructed applying polyA selection, and Illumina compatible cDNA libraries were prepared by GENEWIZ (Leipzig, Germany). Samples were then sequenced on Illumina NovaSeq 6,000 yielding 2 × 150 bp paired end reads (GENEWIZ). Processing of RNA sequencing data was done using Chipster software as described above. The cut-off criteria were set to adjusted p-value <0.1. No cut-off criteria were used for logFC. Enrichr was used to identify enriched pathways and ontologies.

2.15 Cell viability assays

Cell viability after *UBE2C* knockdown was evaluated with cell proliferation agent WST-1 and clonogenic survival assay. The WST-1 assay (Roche Diagnostics GmbH, Basel, Switzerland) was performed according to the manufacturer's instructions at timepoint of 48 h. For the clonogenic assay, HUH6 cells transfected with *UBE2C* or NT siRNAs were disaggregated into single-cell suspension and seeded at low density into 12-well plates. After culturing for 72 h, the cells were washed with phosphate-buffered saline (PBS), fixed with 4% paraformaldehyde, permeabilized with 100% methanol, stained with crystal violet and rinsed with dH2O. The area occupied by cell colonies in each well was calculated using Cell Profiler (McQuin et al., 2018).

2.16 Migration assay

Cell migration was assessed using transwell migration inserts (8 um pore size; Merck Millipore, Darmstadt, Germany). The bottom of each insert was pre-coated with collagen I (0.1 mg/mL; Sigma Aldrich, St. Louis, MO, United States) and placed into 24-well plates containing cell culture medium (10% serum). *UBE2C* or NT siRNA transfected cells were seeded to upper side of membrane in starvation medium (serum-free) (seeding density 5 × 10^5 cells per well). After culturing for 42 h, cells were fixed with 4% paraformaldehyde, permeabilized with 100% methanol, and stained with crystal violet. Non-migrated cells were removed from upper side of membrane with a cotton swab. In each insert, images were captured from five randomly chosen fields with Eclipse TS100 microscope supplemented with DS-Fi1 digital imaging system (magnification ×10; Nikon, Tokyo, Japan). The number of migrated cells was calculated with ImageJ software.

3 Results

3.1 Genes differentially expressed in HB cell lines vs. primary hepatocytes

The workflow is outlined in Figure 1. Five of the cell lines used in this study were established from aggressive HB tumors; the sixth was the immortalized human HB cell line, HUH6. Details of these cell lines are shown in Table 1 (Doi, 1976; Kats et al., 2019). RNA-sequencing analysis of these six cell lines identified approximately 9,000 differentially expressed genes (DEGs) in each cell line compared to primary hepatocytes (Figure 2A; Supplementary Table S2). Of these, approximately half were upregulated and half downregulated. Venn analysis showed that 2,868 of DEGs were shared among all 6 HB cell lines (Figure 2B). The top 20 most upregulated and most downregulated genes are shown in Figure 2C. The most upregulated genes were *ODAM, TRIM71,* and *IGDCC3*, while the most downregulated genes were *SAA1, SAA2,* and *NNMT*. Among the most upregulated genes were *GPC3*, *DLK1*, and *SP8*, previously connected to aggressive HB, underscoring the robustness of the analysis pipeline (Cairo et al., 2008; Zynger et al., 2008; Wagner et al., 2020).

3.2 Differentially expressed metabolic genes

A list of known human metabolic genes was obtained from Virtual Human Metabolomics (Supplementary Table S1). Venn analysis showed that 490 of these metabolic genes were shared among all 6 HB cell lines (Figure 3A; Supplementary Table S4). Approximately 1,400 DEGs in each HB cell line were classified as metabolic genes (Figure 3B; Supplementary Table S3). The 20 most upregulated and most downregulated DEGs overlapping with metabolic genes are listed in Figure 3C. The most upregulated metabolism-associated genes were *RBP2, DPEP1,* and *PCYT1B*, and the most downregulated genes were *NNMT, VNN1,* and *CYP2C18*. Out of these six genes, DPEP1, responsible for hydrolysis of several dipeptides, and nicotinamide N-methyltransferase NNMT have been demonstrated to play a

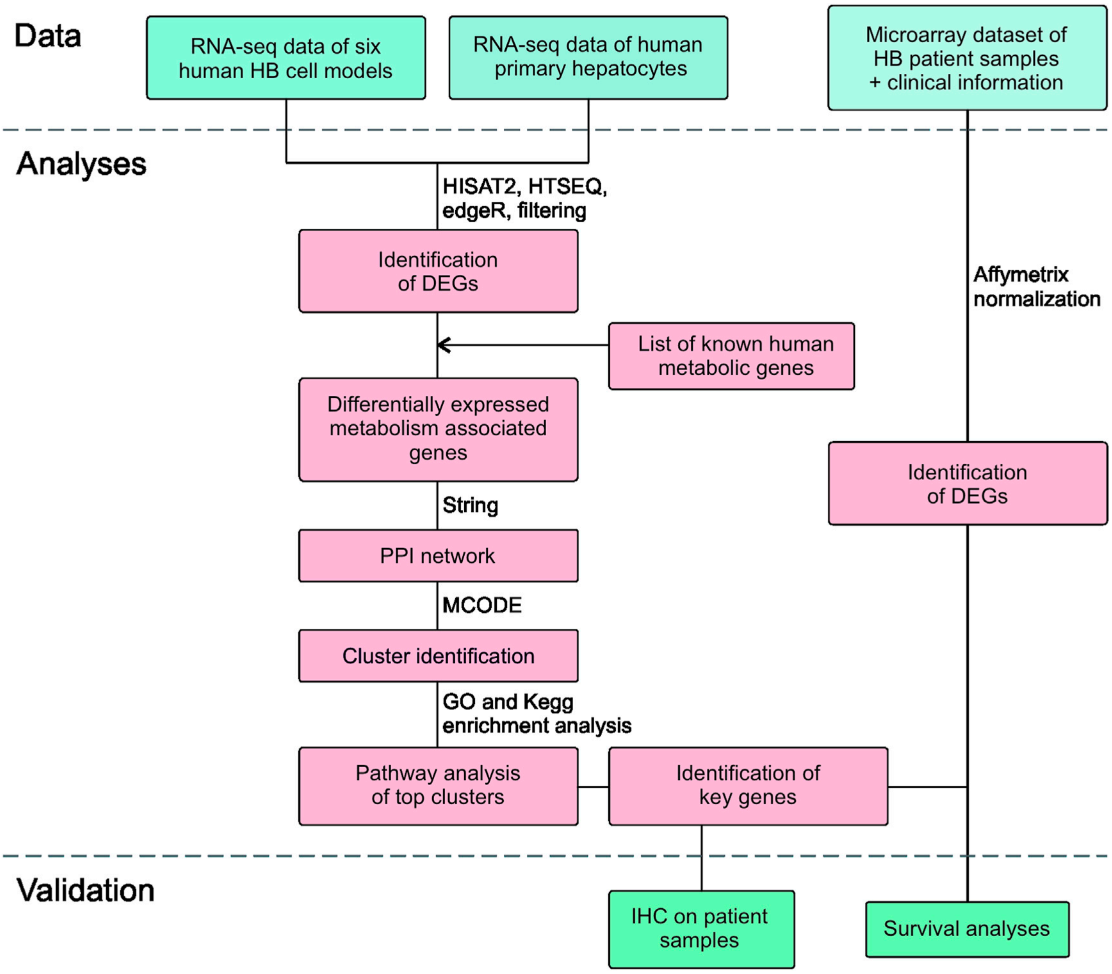

Data
RNA-seq data of six human HB cell models
RNA-seq data of human primary hepatocytes
Microarray dataset of HB patient samples + clinical information
Analyses
HISAT2, HTSEQ, edgeR, filtering
Identification of DEGs
Affymetrix normalization
List of known human metabolic genes
Differentially expressed metabolism associated genes
Identification of DEGs
String
PPI network
MCODE
Cluster identification
GO and Kegg enrichment analysis
Pathway analysis of top clusters
Identification of key genes
Validation
IHC on patient samples
Survival analyses

FIGURE 1
Flowchart of the study design.

TABLE 1 HB cell line characteristics.

Cell line	Age at sampling	Sex	Histology	Origin	References
HB-243	52 months	Male	embryonal	intrahepatic relapse	Kats et al., 2019
HB-279	79 months	Male	embryonal and macrotrabecular	primary tumor	Kats et al., 2019
HB-282	12 months	Male	embryonal	primary tumor	Kats et al., 2019
HB-284	83 months	Male	embryonal	peritoneal metastasis	Kats et al., 2019
HB-295	26 months	female	fetal	primary tumor	Kats et al., 2019
HUH6	12 months	Male	mixed, predominantly embryonal	primary tumor	Doi, (1976)

role in HB pathogenesis in previous studies (Cui et al., 2019; Rivas et al., 2020). Other genes were not previously reported in HB. These genes are involved in regulating oxidative phosphorylation (*RBP2*, *VNN1*), phosphatidylcholine biosynthesis (*PCYT1B*), and xenobiotic and retinoid metabolism (*CYP2C18*, *RBP2*) (Chen and Goldstein, 2012; Grinde et al., 2014; Váraljai et al., 2015; Giessner et al., 2018; Blaner et al., 2020).

3.3 PPI-network construction and clustering

Protein-to-protein interaction (PPI) networks describe the physical contact of proteins within cells. PPI networks were constructed to better understand the changes in cell physiology represented by transcriptome analysis in HB. Lists of differentially expressed metabolic genes in each cell line

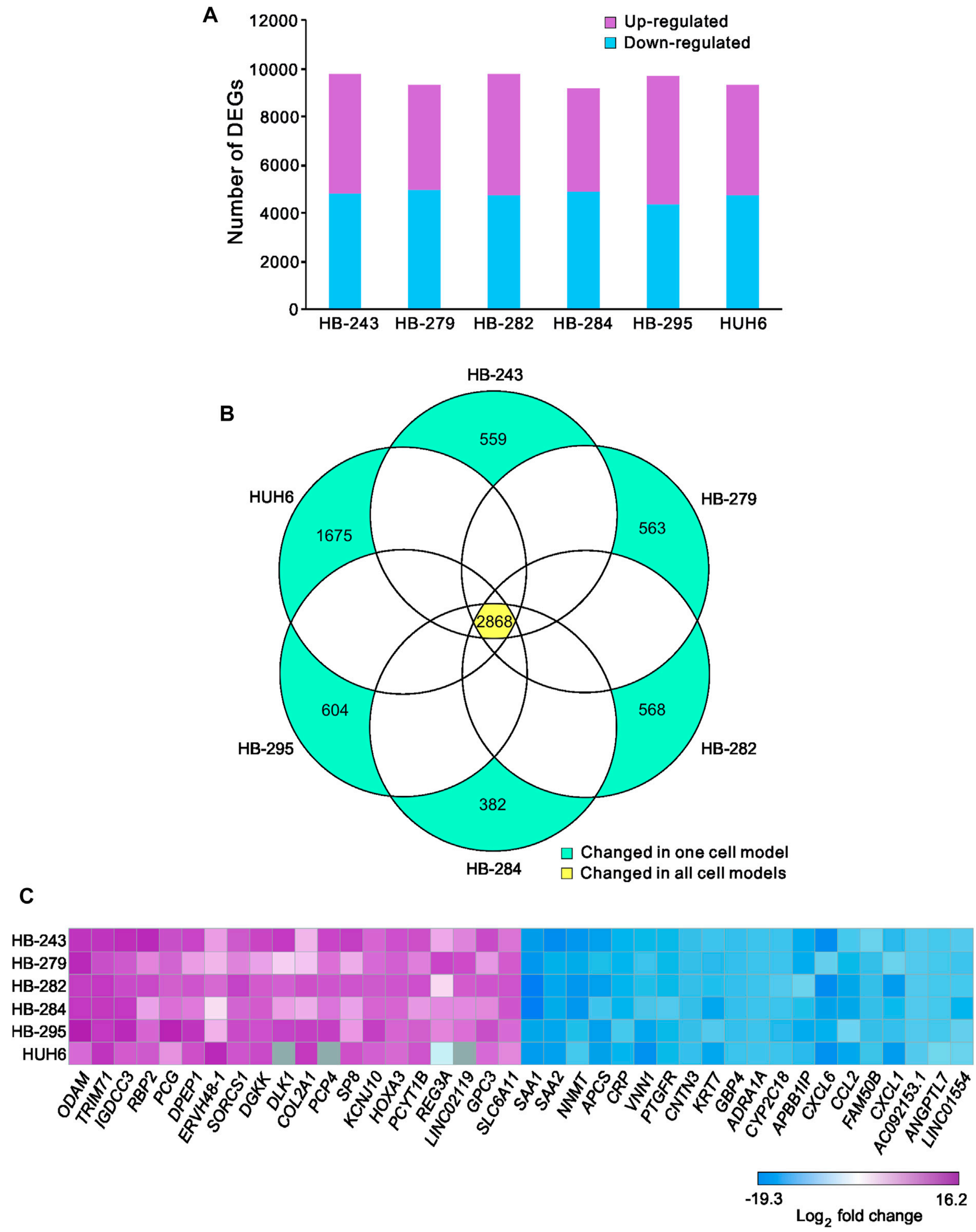


FIGURE 2
Differentially expressed genes in HB cell lines compared to primary hepatocytes Upregulated (pink) and downregulated (blue) genes in each HB cell line **(A)**. Venn diagram showing the number of significant differentially expressed genes compared to primary hepatocytes that were shared among all six HB cell models **(B)**. Heatmap of the 20 most downregulated and the 20 most upregulated differentially expressed genes sorted by $\log_2$FC **(C)**. Genes in gray = change not statistically significant. Differential expression analysis was conducted using the edgeR package.

(Supplementary Table S3) were used to construct the PPI networks using the STRING database. Highly interconnected areas (clusters) in each PPI-network were identified, scored, and ranked on size and density. The highest-scoring cluster of each cell line is shown in Figures 4A–F. The scores of the top clusters were 34 (HB-243), 24 (HB-279), 32 (HB-282), 29 (HB-

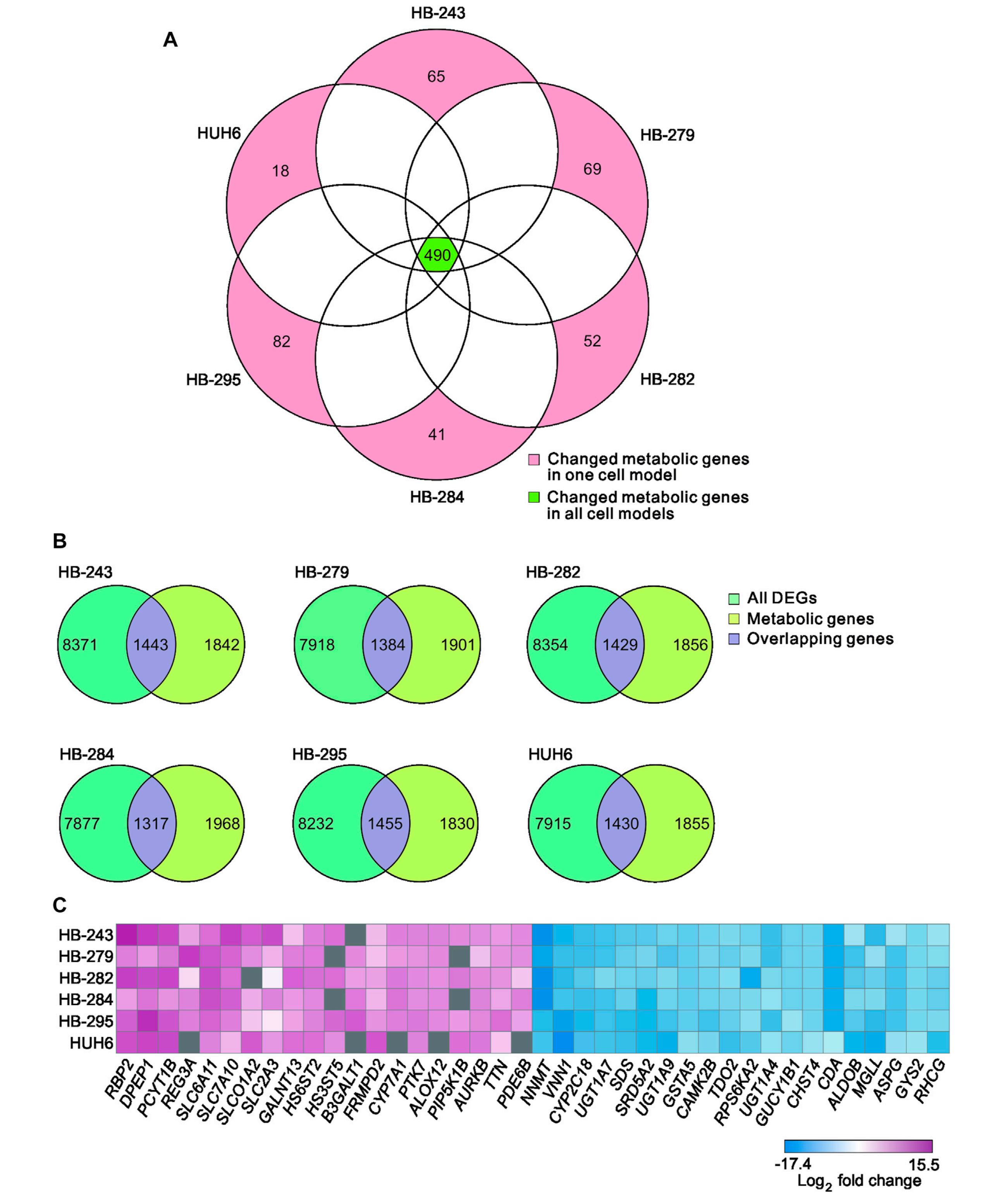


FIGURE 3
Differentially expressed metabolic genes in HB cell models Venn diagram showing the number of shared metabolic genes among the 6 HB cell lines **(A)**. Overlap of differentially expressed genes in each cell line and the list of all metabolic genes **(B)**. Heatmap of the 20 most up- and the 20 most downregulated metabolic genes in HB cell models **(C)**. Genes in gray = change not statistically significant.

284), 33 (HB-295) and 28 (HUH6). Upregulated genes found in one or more of these clusters included *UBE2C, UBE2D1, UBE2N, UBE2O, UBE2Q1, UBE2Q2, UBE2R2, UBE2S, DZIP3, HACE1, MGRN1, MIB2, MYLIP, RBBP6, RNF126, RNF138, RNF182, SIAH2, SMURF2, WWP1, ZNRF1,* and *ZNRF2.* Downregulated genes included *CBLB, HECTD2, HECTD3,*

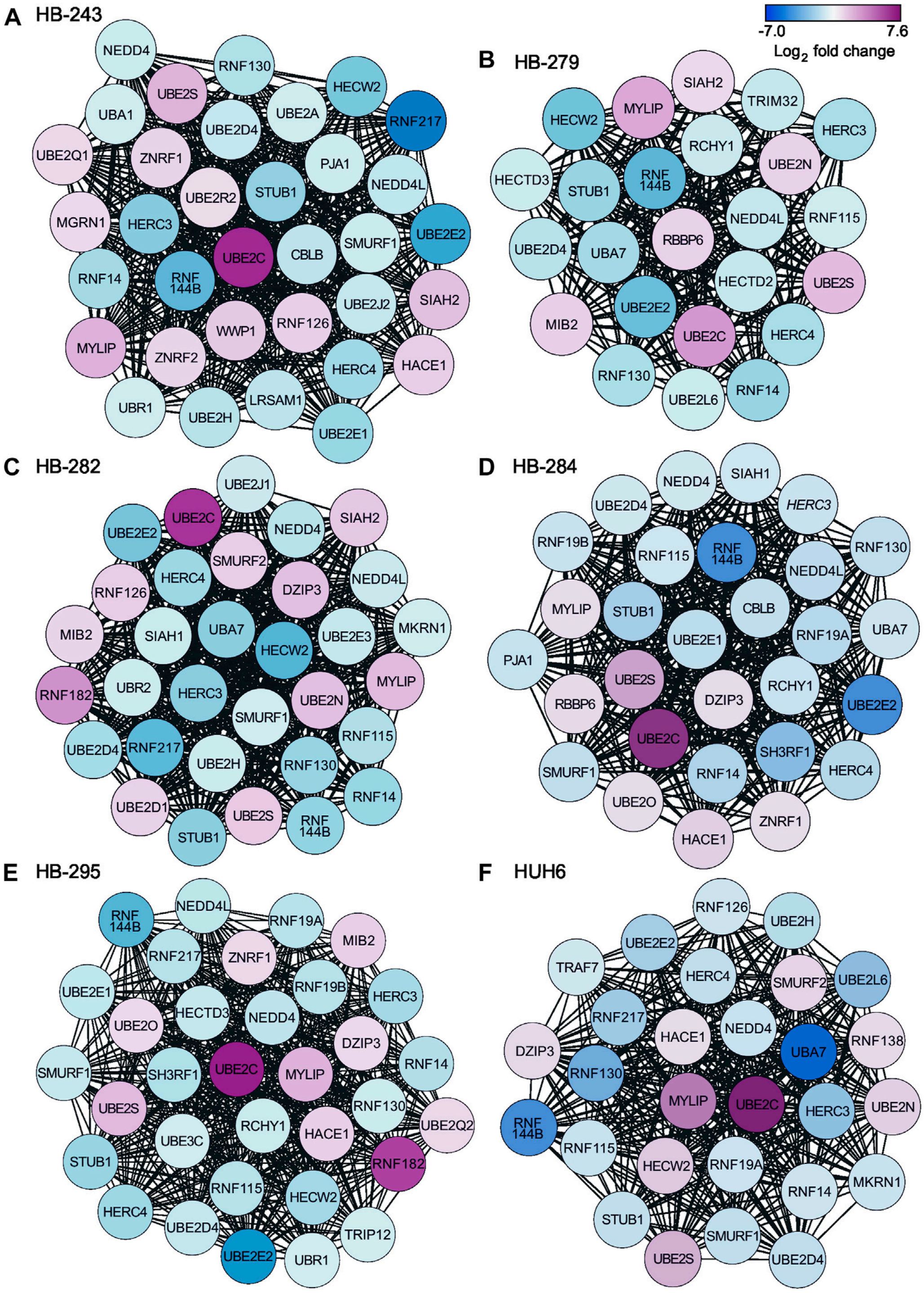


FIGURE 4
Highest-scoring clusters found in PPI networks of differentially expressed metabolic genes in each cell line. Clusters were ranked based on their score using MCODE plugin of Cytoscape. HB-243 **(A)**, HB-279 **(B)**, HB-282 **(C)**, HB-284 **(D)**, HB-295 **(E)**, and HUH6 **(F)**.

HECW2, HERC3, HERC4, LRSAM1, MKRN1, NEDD4, NEDD4L, PJA1, RCHY1, RNF115, RNF130, RNF14, RNF144B, RNF19A, RNF19B, RNF217, SH3RF1, SIAH1, SMURF1, STUB1, TRAF7, TRIM32, TRIP12, UBA1, UBA7, UBE2A, UBE2D4, UBE2E1, UBE2E2, UBE2E3, UBE2H, UBE2J1, UBE2J2, UBE2L6 UBE3C, UBR1, and *UBR2.*

3.4 KEGG and GO analyses of the highest-ranking clusters

Next, to identify enriched pathways and gene sets, the list of protein-coding genes in each cell line's highest-scoring PPI cluster were uploaded to the online tool Enrichr and the results ranked by *p*-value. The top 5 most

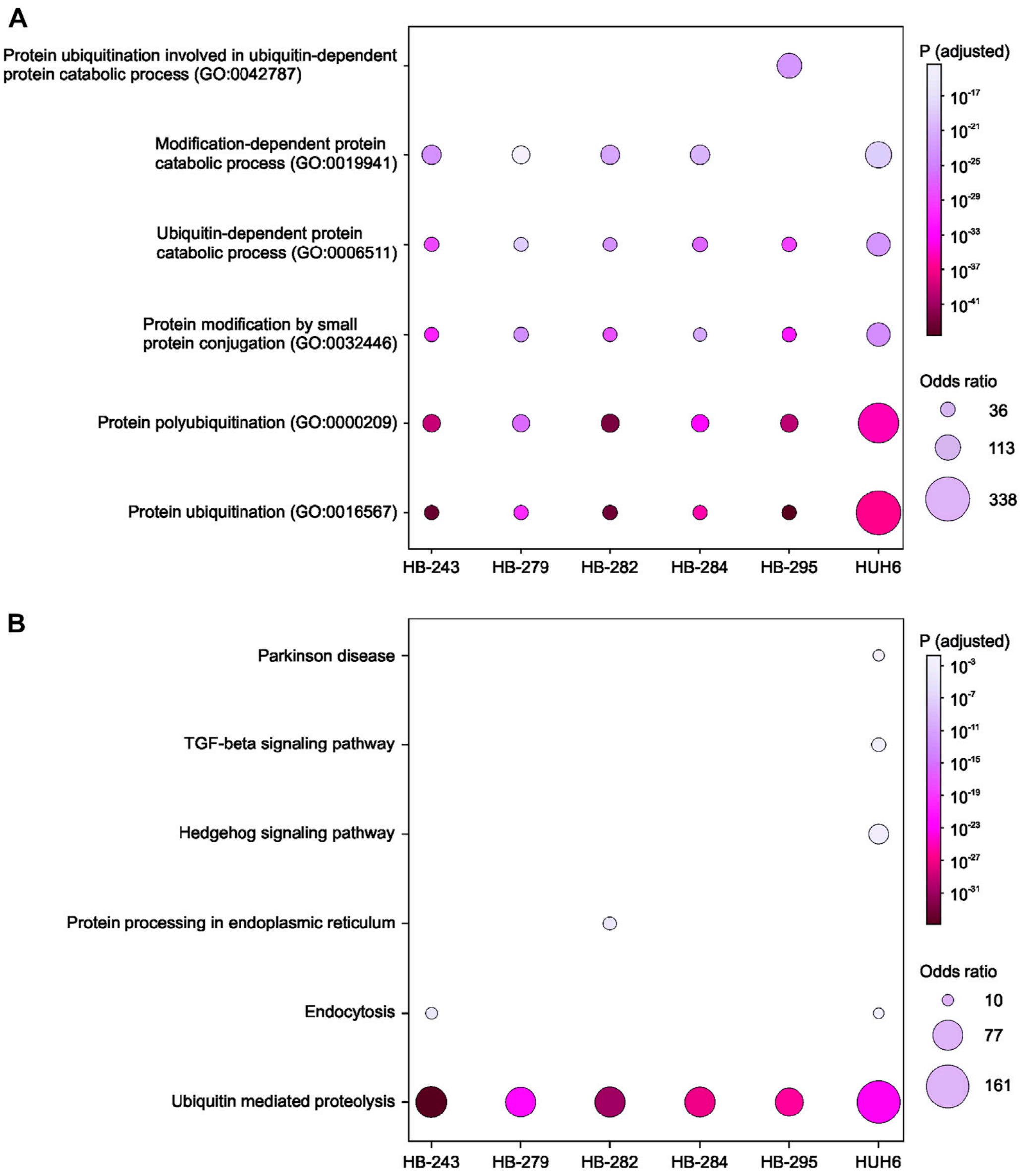


FIGURE 5
KEGG and GO analyses for enriched pathways. Genes in each cell line's top-ranking PPI clusters were analyzed to find common enriched pathways. Statistically significant ($p < 0.05$) GO biological processes **(A)** and KEGG pathways **(B)** are shown for each cell line.

statistically significant GO-terms in the highest-ranking clusters were protein ubiquitination, protein polyubiquitination, protein modification by small protein conjugation, modification-dependent protein catabolic processes, and protein ubiquitination involved in ubiquitin-dependent protein catabolic processes (Figure 5A). The corresponding KEGG-terms were ubiquitin mediated proteolysis, endocytosis, protein processing in endoplasmic reticulum, Hedgehog signaling pathway, TGF-β signaling pathway, and Parkinson's disease (Figure 5B).

3.5 Validation of findings with HB patient microarray dataset

Genes present in all six of the highest-scoring clusters in the PPI-network were *RNF130, UBE2E2, UBE2C, RNF14, HERC3, HERC4, STUB1, UBE2S, RNF144B, MYLIP,* and *UBE2D4*. Of these, four genes—*RNF130, UBE2C, HERC3,* and *RNF144B*—were found to be significantly altered in the GSE131329 microarray dataset. This dataset includes 53 HB tissue samples and 14 noncancerous liver (NCL) tissue samples collected at the time of surgery from HB patients. *RNF144B, RNF130,* and *HERC3* mRNA expression was downregulated compared to the normal liver samples ($\log_2$FC −0.84, adj. p-value $1.10*10^{-5}$; $\log_2$FC −0.6, adj. p-value $3.0*10^{-06}$, and $\log_2$FC −0.62, adj. p-value 0.00186, respectively), whereas *UBE2C* mRNA expression was upregulated in HB samples ($\log_2$FC 1.0; adj. p-value 0.0001). RNF144B, RNF130, and HERC3 act as a ubiquitin ligases and UBE2C is an ubiquitin conjugating enzyme. Expression of these 4 genes in both HB and NCL groups is shown in Figures 6A–D.

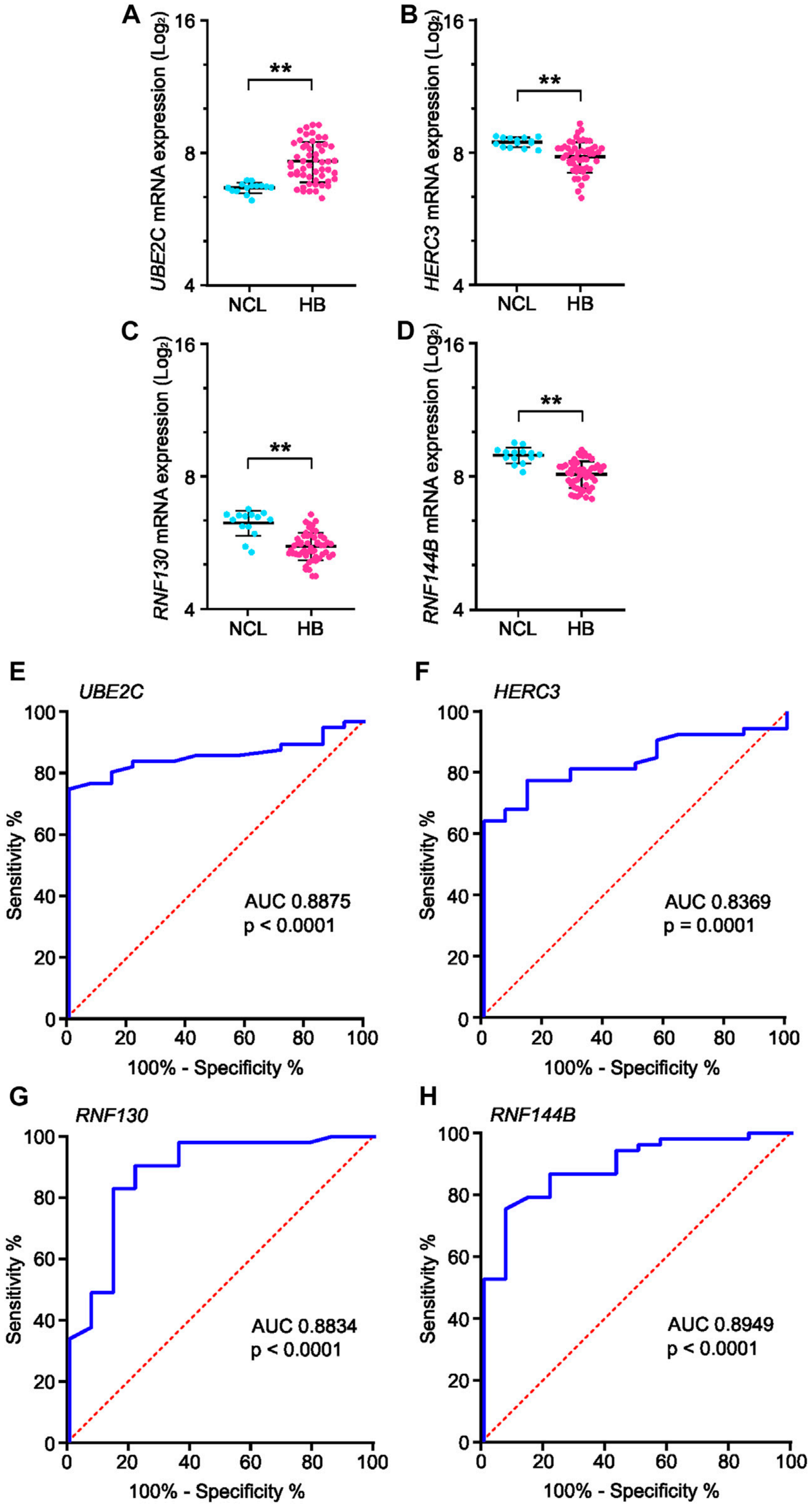


FIGURE 6
mRNA expression and ROC curve analysis of the four key genes in GSE131329 dataset. Expression of *UBE2C* **(A)**, *HERC3* **(B)**, *RNF130* **(C)** and *RNF144B* **(D)** in HB samples compared to noncancerous liver samples on mRNA level. ROC curve analysis of *UBE2C* **(E)**, *HERC3* **(F)**, *RNF130* **(G)** and *RNF144B* **(H)** assessing the suitability of each gene for discrimination of noncancerous liver (NCL) and HB. ** = $p < 0.01$. Grey dots represent gene expression of independent patients, the whiskers represent the first and third quartile, and the thick solid line is median **(A–D)**.

TABLE 2 Association of UBE2C, HERC3, RNF130, and RNF144B mRNA expression (log2) with clinical course of the disease assessed with Mann-Whitney U test.

	UBE2C	*HERC3*	*RNF130*	*RNF144B*
Distant metastasis NO (*n* = *39*) Median mRNA exp	7.330	7.940	9.400	8.240
Distant metastasis YES (*n* = *14*) Median mRNA exp	8.295	7.440	9.470	7.765
p-value	0.0046	0.0682	0.5592	0.0874
Event-free YES (*n* = *32*) Median mRNA exp	7.325	8.060	9.460	8.310
Event-free NO (*n* = *21*) Median mRNA exp	7.930	7.580	9.330	7.860
p-value	0.0303	0.0247	0.6619	0.0130
Overall survival ALIVE (*n* = *38*) Median mRNA exp	7.325	8.005	9.425	8.280
Overall survival DEAD (*n* = *15*) Median mRNA exp	8.240	7.560	9.510	7.840
p-value	0.0095	0.0581	0.4966	0.0269

TABLE 3 Clinical information of GSE131329 dataset. (HB = hepatoblastoma; NCL = non-cancerous liver).

	HB (*n* = 53)	NCL (*n* = 17)
Median age, months	22 (0–109)	17 (5–98)
Sex FEMALE, n (%)	47.2	57.1
Sex MALE, n (%)	52.8	42.9
PRETEXT = 1, n	9	—
PRETEXT = 2, n	15	—
PRETEXT = 3, n	18	—
PRETEXT = 4, n	11	—
Distant metastasis NO, n	39	—
Distant metastasis YES, n	14	—
Event-free NO, n	21	—
Event-free YES, n	32	—
Overall survival, alive (%)	71.7	—

3.6 Clinical analyses of potential key genes *UBE2C, RNF130, HERC3,* and *RNF144B*

To analyze the suitability of each selected gene (*UBE2C*, *RNF130*, *HERC3*, and *RNF144B*) to discriminate normal control liver samples (NCL) and HB, we performed ROC curve analysis. The area under curve (AUC) was 0.8875 for *UBE2C* (Figure 6E, 95% CI 0.81–0.96), 0.8369 for *HERC3* (Figure 6F, 95% CI 0.74–0.93), 0.8834 for *RNF130* (Figure 6G, 95% CI 0.77–0.99), and 0.8949 for *RNF144B* (Figure 6H, 95% CI 0.81–0.98). Next, we assessed the association of each of these genes with distant metastasis status, occurrence of events, and overall survival. High *UBE2C* mRNA expression was linked with distant metastasis (p-value <0.01), events (p-value <0.05), and death (p-value <0.01) (Table 2). Downregulation of *HERC3* and *RNF144B* was associated with occurrence of events (p-values <0.05) (Table 2). *RNF130* expression did not show a statistically significant association with any of the studied variables (Table 2). Clinical information adapted from GSE131329 dataset is summarized in Table 3.

3.7 UBE2C associated protein interactions and gene co-expression analysis

To analyze the functional enrichment of UBE2C specific protein interactions in general, STRING network analysis was carried out. UBE2C was used as the input protein. Thirty proteins were interacting with UBE2C when the cut-off was set to 0.4 (Figure 7A). Next, we assessed the level of differential RNA expression of these UBE2C interacting proteins in our RNA sequencing data of the six HB cell models. Of these 30 proteins, fourteen were found to be significantly differentially expressed on RNA level in all 6 HB models (Figure 7B).

Gene co-expression analysis was conducted for *UBE2C* using the online tool GeneFriends to further explore the relationship between *UBE2C* and related genes. Using *UBE2C* as the seed gene, top 10 co-expressed genes with the highest Pearson correlation values were *CCNB2, TOP2A, NUSAP1, CKS2, NUF2, CDK1, NEK2, PTTG1, CKS1B,* and lncRNA *RP11-102C16.3* (Supplementary Figure S1A–B). When comparing these protein interactions and gene co-expression results of UBE2C, three genes/proteins, CKS1B, CCNB2 and CDK1, were present in both networks.

3.8 Immunohistochemical staining of UBE2C in HB patient samples

To validate UBE2C protein expression in HB patient samples, immunohistochemical staining was performed on 6 NL (Table 4) and 25 HB (Table 5) samples. One of the 6 NL samples showed positive *UBE2C* staining on the cell membranes (Figures 8A, B). Of the HB samples, 5 were considered UBE2C-negatives (Figures 8C, D) and 20 UBE2C positive (Figures 8E, F). Compared to NL samples, staining for UBE2C in HB cells appeared stronger and localized to nuclei rather than the cell membrane.

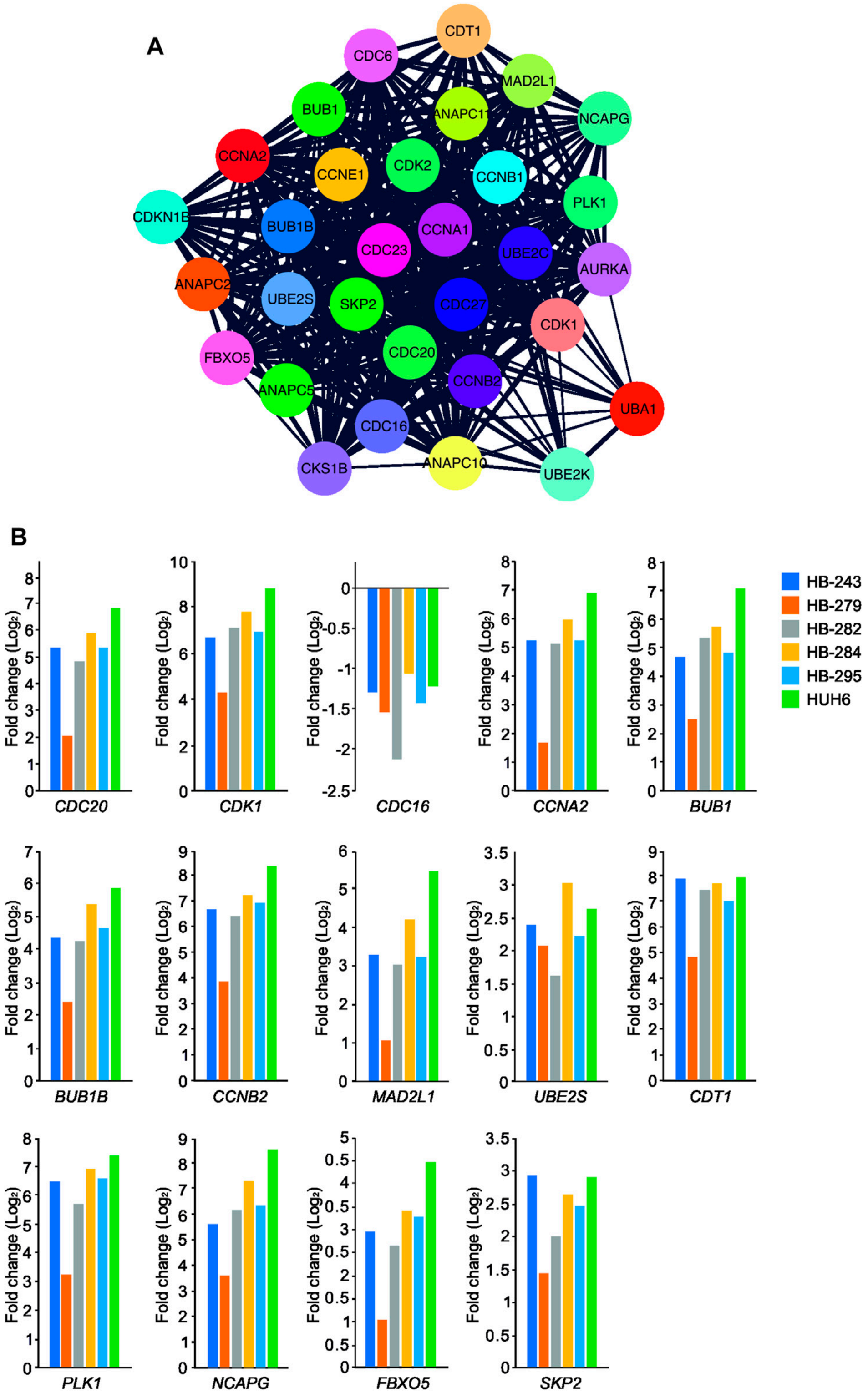


FIGURE 7
UBE2C protein-protein interaction network. UBE2C associated protein-protein interactions **(A)**. The fold change of UBE2C interacting DEGs (in relation to primary hepatocyte) emerging from our RNA sequencing analysis **(B)**.

TABLE 4 Liver samples from donors used in IHC.

Sample	UBE2C staining	Age at death (years)	Sex	Cause of death
NL1	–	3	M	anoxia, heart-related
NL2	–	8	M	traumatic brain injury
NL3	+	11	F	anoxia, trauma-related
NL4	–	26	F	traumatic brain injury
NL5	–	14	M	traumatic brain injury
NL6	–	16	F	spontaneous subdural hemorrhage

TABLE 5 Clinical information of HB patient tissue samples used in IHC.

Sample	UBE2C staining	Age at tx/res (years, age group)	Risk	Histology	Sex
HB1	+	3–7	high	NA	F
HB2	+	>7	high	epithelial, macrotrabecular	M
HB3	+	1–3	high	fetal epithelial, well differentiated	F
HB4	+	1–3	high	embryonal and fetal epithelial	F
HB5	+	>7	high	fetal epithelial	M
HB6	+	1–3	standard	mixed epithelial and mesenchymal	M
HB7	+	1–3	standard	fetal epithelial	M
HB8	–	3–7	high	fetal epithelial	F
HB9	+	3–7	high	epithelial	F
HB10	+	1–3	standard	fetal epithelial	M
HB11	+	>7	high	fetal epithelial	F
HB12	+	>7	high	fetal epithelial	M
HB13	+	<1	standard	fetal, teratoid features	M
HB14	+	1–3	standard	fetal epithelial	M
HB15	–	1–3	high	mixed epithelial and mesenchymal with teratoid features	F
HB16	+	<1	high	mixed epithelial and mesenchymal	M
HB17	+	1–3	high	embryonal and fetal epithelial	M
HB18	+	>7	high	embryonal and fetal epithelial	F
HB19	+	3–7	high	fetal epithelial	M
HB20	-	>7	high	embryonal and fetal epithelial	F
HB21	+	3–7	high	embryonal and fetal epithelial	F
HB22	–	3–7	high	fetal epithelial	M
HB23	–	3–7	high	embryonal epithelial	F
HB24	+	3–7	high	mixed	M
HB25	+	3–7	high	fetal epithelial, well differentiated	M

3.9 *UBE2C* silencing decreases cell viability and migration in HUH6 and HB-243 HB cell models

To explore UBE2C function *in vitro*, *UBE2C* was silenced in HUH6 and HB-243 cell lines using siRNA transfection. Non-targeting (NT) siRNAs were used as a control. After siRNA transfections, *UBE2C* expression was reduced 95% at mRNA and 80% at protein level in HUH6 (Figures 9A, B) and 98% at mRNA and 80% at protein level in HB-243. (Figures 9D, E). The effect of *UBE2C* knockdown on cell viability was evaluated using WST-1 assay. Relative cell viability

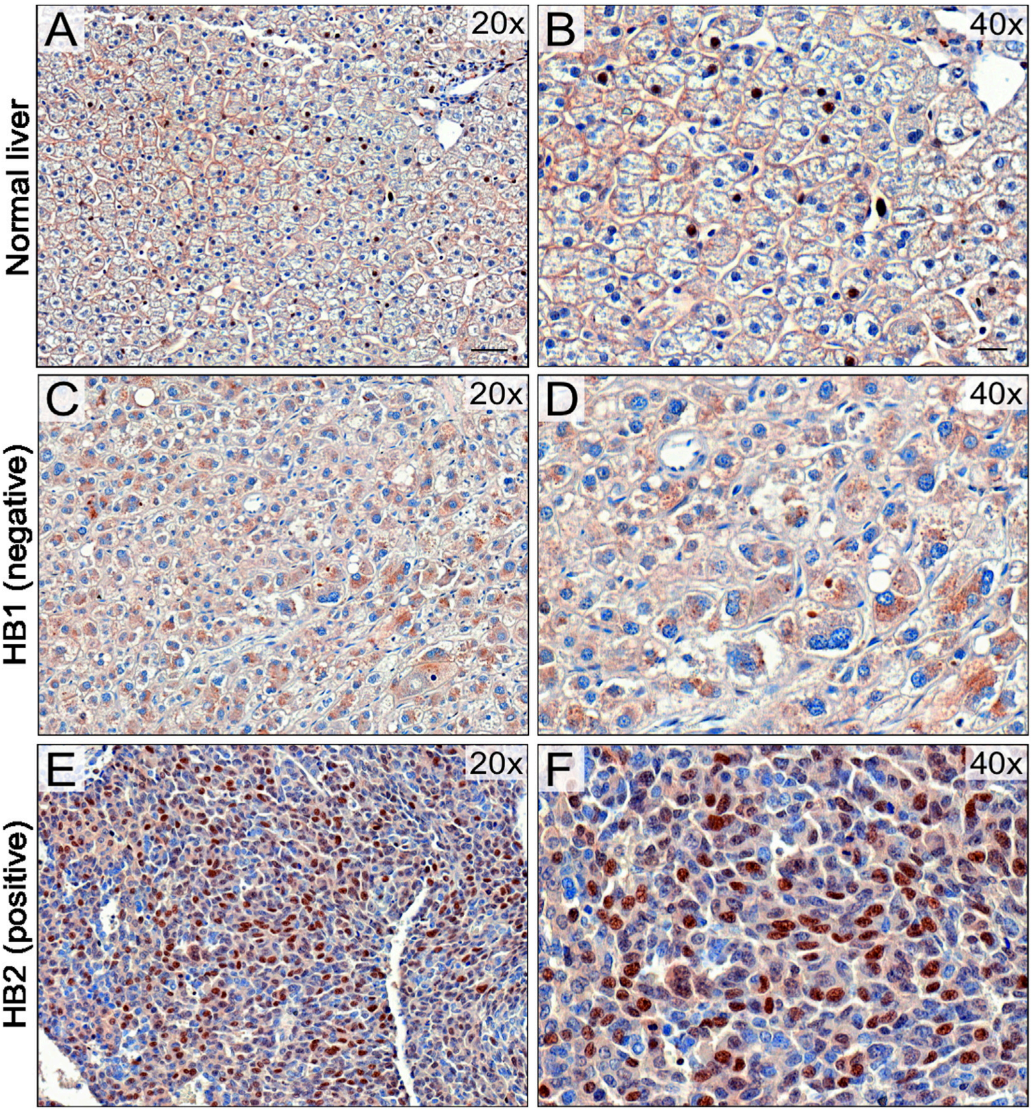


FIGURE 8
UBE2C protein expression in HB tissues. Immunohistochemical staining of UBE2C in 6 normal liver (NL) samples and 25 hepatoblastoma (HB) samples was done. Representative image of normal liver stained with UBE2C with ×20 **(A)** and ×40 **(B)** magnification. Representative image of HB liver staining negative for UBE2C with ×20 **(C)** and ×40 **(D)** magnification. Representative sample of HB liver staining positive for UBE2C ×20 **(E)** and ×40 **(F)** magnification. Scale bars = 50 μm **(B,D)** and 20 μm **(C,E)**.

decreased 24% in HUH6 (Figure 9C) and 44% in HB-243 (Figure 9F).

RNA sequencing of *UBE2C* silenced HUH6 cells showed 111 differentially expressed genes (Supplementary Table S5). Top 5 pathways and ontologies matching these genes included cell cycle related processes such as p53 regulation, DNA damage response, and G1/S checkpoint (Figures 10A–F).

Effects of *UBE2C* knockdown in HUH6 cells were further evaluated using a clonogenic assay, which showed a statistically significant fall in cell number in *UBE2C* silenced cells with the well area covered by cells decreasing 35% (Supplementary Figure S2A–C). The effect of *UBE2C* silencing on HUH6 cell migration was assessed with transwell assay, which demonstrated a statistically significant 65% decrease in the number of migrated cells compared to NT siRNA treated cells (Supplementary Figure S2D–F).

4 Discussion

Metabolic derangements have been associated with enhanced tumorigenesis and cancer progression in several tumor entities (Faubert et al., 2020), and these tumor-specific changes can be exploited to develop targeted therapies (Sullivan et al., 2016). In our RNA sequencing analysis of HB cell models, ubiquitination emerged as the most significantly altered metabolic pathway. The expression of three ubiquitin ligases (*HERC3*, *RNF130*, and *RNF144A*) and one ubiquitin conjugating enzyme (*UBE2C*) was significantly dysregulated in all studied HB models. These four genes were assessed more thoroughly in a HB patient dataset to validate their significance in the clinical setting. We noticed a remarkable association between high *UBE2C* expression and aggressive disease in the particular HB patient cohort.

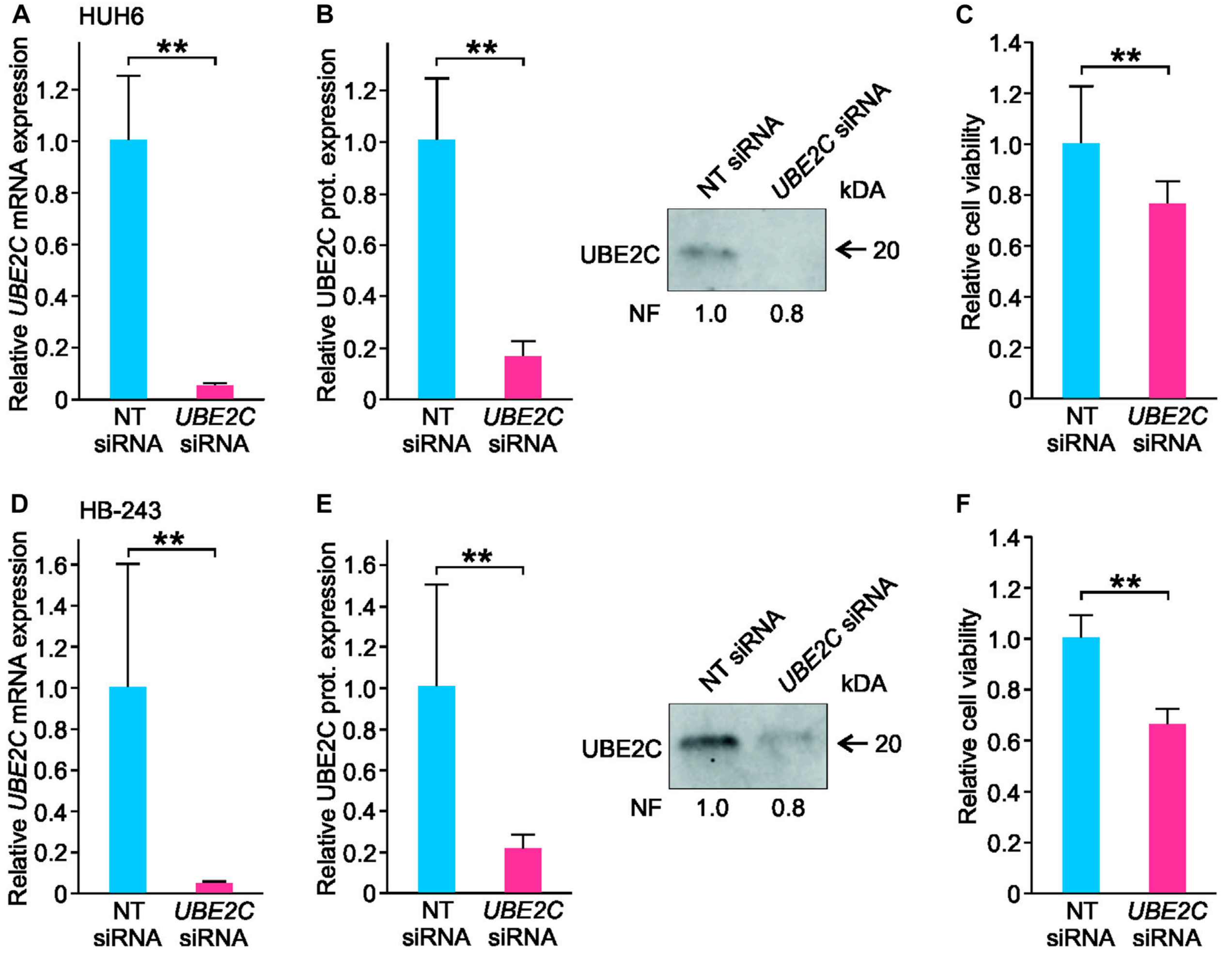


FIGURE 9
Knockdown of *UBE2C* and its effect on HUH6 and HB-243 cell viability. Following the siRNA transfection, *UBE2C* mRNA expression in HUH6 was reduced by 95% **(A)** and protein expression by 80% **(B)** compared to cells transfected with non-targeting (NT) siRNA. In HB-243, *UBE2C* mRNA expression was reduced by 98% **(D)** and protein expression by 80% **(E)**. WST-1 assay showed a 24% decrease in HUH6 **(C)** and 44% in HB-243 **(F)** in cell viability after *UBE2C* silencing. Bar plots are presented as relative values of mean of three independent experiments ± RSD. ***p*-value <0.01, NT = non-targeting. Normalization factor (NF) describing the amount of total protein in lane in relation to other lanes is given beneath the bands **(B, E)**.

Ubiquitination is a crucial mechanism for the degradation of short-lived proteins like those involved in cell cycle regulation (Guo et al., 2019; Zhang et al., 2021). In addition to protein degradation, ubiquitination regulates DNA repair, translation, and inflammation (Miranda and Sorkin, 2007). The three main steps in ubiquitination are activation (performed by ubiquitin-activating enzymes, E1s), conjugation (ubiquitin-conjugating enzymes, E2s), and ligation (ubiquitin-ligating enzymes, E3s) (Komander and Rape, 2012). Ubiquitination and deubiquitination are known to be modulated during cancer progression (Sun et al., 2020), and high *UBE2C* expression portends poor survival in various cancers including node-positive breast cancer (Loussouarn et al., 2009) and ovarian carcinomas (Berlingieri et al., 2007). The magnitude of *UBE2C* mRNA overexpression in HB cell lines was striking (up to 128-fold higher than primary hepatocytes), and *UBE2C* expression in HB clinical specimens was associated with increased risk of distant metastasis, events, and death. Our findings echo a recent study showing that *UBE2C* expression may be used as a diagnostic biomarker in hepatocellular carcinoma, the most frequent liver cancer in adults (Gao et al., 2021).

In HB tissue samples, we observed predominantly nuclear localization of UBE2C protein. In other cancers, both nuclear and cytoplasmic UBE2C immunoreactivity have been observed, but the functional significance of this is unclear (Shen et al., 2013; Ma et al., 2016; Palumbo et al., 2016). Kraft *et al.* showed that strong nuclear expression of UBE2C was linked with higher mitosis rate in melanoma suggesting that UBE2C localization in nuclei may be at least partially related to its role in the regulation of cell cycle associated proteins (Kraft et al., 2017).

It has previously been shown that in cancer cells UBE2C plays an important role in facilitation of protein degradation and dysregulation of the cell cycle (Sun et al., 2020). UBE2C overexpressing cells have the ability to override mitotic spindle checkpoints, which may lead to loss of genomic stability, a characteristic of cancer (Reddy et al., 2007). UBE2C is also suggested to be a potential oncogene enhancing migration and invasion in hepatocellular carcinoma (Xiong et al., 2019). Consistently,

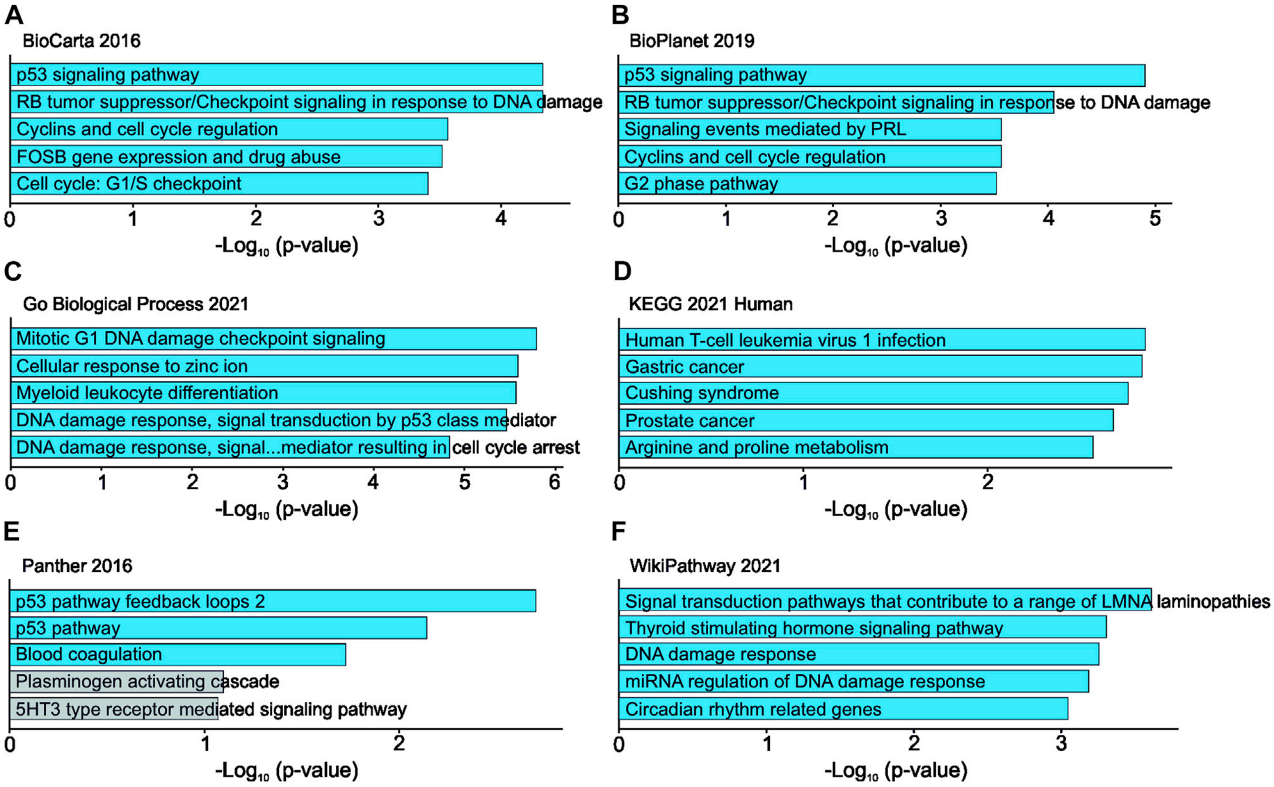


FIGURE 10
Effects of *UBE2C* knockdown in HUH6 cells on RNA level. *UBE2C* knockdown is linked with alterations in RNA expression of genes connected to cell cycle regulation and p53 signaling pathway. Top 5 pathways and ontologies ranked by *p*-value **(A–F)**.

we demonstrated that knockdown of *UBE2C* resulted in a decrease in HB viability, and preliminary results suggest that it could also have a negative effect on HB cell migration. Our RNA sequencing results supported the hypothesis that UBE2C participates in cell cycle regulation in HB. After *UBE2C* knockdown, we observed alterations in mRNA expression of *CDKN1A*, *CDK, PIK3C2B, PIDD1,* and *E2F2* genes which are known to participate in cell cycle regulation and the p53 signaling pathway. Changes in mRNA expression, however, were rather subtle. Effects of *UBE2C* knockdown on cell cycle were, however, not assessed with *in vitro* experiments in this article. Given the role of UBE2C as a post-translational factor rather than a direct regulator of gene expression, proteomics analysis could be conducted in future experiments. UBE2C overexpression has also previously been linked to increased ubiquitination and subsequent degradation of the tumor suppressor p53 in endometrial cancer (Liu et al., 2020). A novel therapy aimed at enhancing p53 activity has been suggested to be a potential treatment alternative for HB (Woodfield et al., 2021). If UBE2C participates in post-translational regulation of p53 expression in HB, its inhibition could lead to reactivation of p53.

Aurora Kinase A (AURKA), a serine/threonine kinase, has a critical role in regulating cell cycle and mitosis (Du et al., 2021). Expression of *AURKA* has been shown to be significantly higher in HB than in normal liver (Tian et al., 2021). In our study, *AURKA* was significantly upregulated in all studied HB cell lines. Treatment modalities targeting AURKA, such as alisertib, have shown promising results in preclinical studies of HB (Tan et al., 2020). Interestingly, increased AURKA expression has been demonstrated to correlate with upregulated *UBE2C* in cancer cells (http://gepia.cancer-pku.cn) (Naso et al., 2021). Furthermore, inhibition of *UBE2C* expression was shown to reduce the level of phosphorylation of AURKA and impair cell viability in gastric adenocarcinoma cells (Wang et al., 2017).

UBE2C links to another key HB gene, cyclin-dependent kinase 1 (*CDK1*) (Aghajanzadeh et al., 2020; Sun et al., 2021; Tian et al., 2021). CDK1 functions as a serine/threonine kinase and, like AURKA, plays an important role in cell cycle regulation. CDK1 has been reported to be upregulated in various cancers including hepatocellular carcinoma (Zhou et al., 2019). Consistent with previous studies, our RNA-sequencing results showed that *CDK1* was highly upregulated in HB cell models. *CDK1* siRNA knockdown was shown to inhibit the growth and invasiveness of HUH6 HB cells (Tian et al., 2021). A study of ovarian cancer cells showed that high UBE2C expression correlated with expression of CDK1. Knockdown of *UBE2C* induced G2/M arrest in the cells, which led to decreased CDK1 expression (Li et al., 2020). In our study, knockdown of *UBE2C* in HUH6 cells increased cyclin-dependent kinase inhibitor 1 (*CDKN1A*) expression and decreased cyclin-dependent kinase 2 (*CDK2*) expression at mRNA level, both of these having a role in G1/S transition. *CDK1* expression was not significantly altered.

There is some evidence that *UBE2C* overexpression impacts chemoresistance. Downregulation of *UBE2C* reversed resistance to cisplatin in ovarian cancer cell models (Li et al., 2020). UBE2C inhibition has been shown to increase doxorubicin sensitivity in breast cancer cells *in vitro* (Rawat et al., 2013). Cisplatin and doxorubicin are both widely used in HB management. In addition to targeted treatment, UBE2C expression status could thus be utilized in the evaluation of treatment resistance to conventional chemotherapy in HB. The proteasome inhibitor bortezomib slows HB progression *in vitro* and *in vivo* (Hooks et al., 2018). In colorectal carcinoma, bortezomib treatment has been demonstrated to downregulate UBE2C expression leading to decreased cell viability via stabilizing mitotic cyclins and inhibiting cell cycle progression (Bavi et al., 2011). Thus, high UBE2C expression could identify HB patients who may benefit from bortezomib treatment.

There are some limitations to this study. The reader should note that in this article, both mRNA- and protein expression of the genes in question are being used. While changes in mRNA-expression often correlate with changes in protein expression, this is not always the case given the several factors affecting the translation process and the final amount of protein in the tissue. This should be kept in mind while interpretating the results. In this study we have used immunohistochemistry and Western blotting to determine the protein expression levels of UBE2C in HB tissues. More extensive proteomics would, however, be required to further elucidate the actual protein expression levels of all the genes related to *UBE2C*. Liver matures throughout childhood, and the use of primary hepatocytes from adult donor as control cells in RNA sequencing analyses of the PDX models may have impacted our results. The noted effects of *UBE2C* silencing on mRNA expression level of cell cycle regulating genes should be further validated with *in vitro* and *in vivo* experiments in order to properly assess the effects on cell cycle. The number of HB patient samples available for this study limited clinical analyses and conclusions. This is unfortunately the case with most studies concerning HB, since the prevalence of HB is low and the availability of samples therefore limited.

Given the promising role of ubiquitin system as a target of new cancer treatments, the role and function of UBE2C in HB progression should be investigated further. The possible role of UBE2C and other ubiquitination-mediating enzymes in drug resistance is also intriguing. One possible technique that could be utilized is single-cell RNA sequencing (scRNAseq). Previously, Bondoc et al. have characterized HB tumor cell populations and identified driver tumor cell clusters using scRNAseq (Bondoc et al., 2021). Given the advantages of the technique, analysis of scRNAseq could provide new insight.

Taken together, we found that metabolic alterations taking place in HB tumors are diverse and that ubiquitination-related factors may have a significant role in HB progression. Notably, UBE2C expression was highly upregulated in all six HB cell lines as well as in patient samples at both mRNA and protein level. *In vitro* knockdown of *UBE2C* resulted in decreased cell division and motility. Moreover, high UBE2C expression was associated with inferior patient survival. These findings may be brought to the clinic to identify the high-risk HB patients for earlier treatment interventions.

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 at https://www.ncbi.nlm.nih.gov/geo/ (GSE140520, GSE83518, GSE131329) and https://ega-archive.org/ (EGAS00001004827/EGAD00001006621).

Ethics statement

The studies involving human participants were reviewed and approved by the Helsinki University Hospital institutional ethics committee. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

Conceptualization, KE and MP; formal analysis, RN and KE; funding acquisition, MH and MP; investigation, RN and KE; methodology, RN, KE, NI, and MP; project administration, KE and MP; resources, SC and MH; software, NI; supervision, KE, DW, MH, and MP; validation, RN; visualization, NI and MP; writing—original draft, RN; writing—review and editing, RN, KE, NI, SC, DW, MH, and MP. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by Biomedicum Helsinki Foundation Grants, Finska Läkaresällskapet, Helsinki University Central Hospital Research Grants, Päivikki and Sakari Sohlberg Foundation, Sigrid Jusélius Foundation, and Väre Foundation for Pediatric Cancer Research.

Conflict of interest

Author SC is employed by Champions Oncology and was previously employed by XenTech.

The remaining 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/fgene.2023.1170940/full#supplementary-material

References

Aghajanzadeh, T., Tebbi, K., and Talkhabi, M. (2020). Identification of potential key genes and miRNAs involved in Hepatoblastoma pathogenesis and prognosis. *J. Cell. Commun. Signal* 151 (15), 131–142. doi:10.1007/S12079-020-00584-1

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi:10.1093/bioinformatics/btu638

Armengol, C., Cairo, S., Fabre, M., and Buendia, M. A. (2011). Wnt signaling and hepatocarcinogenesis: The hepatoblastoma model. *Int. J. Biochem. Cell. Biol.* 43, 265–270. doi:10.1016/j.biocel.2009.07.012

Aronson, D. C., and Meyers, R. L. (2016). Malignant tumors of the liver in children. *Semin. Pediatr. Surg.* 25, 265–275. doi:10.1053/j.sempedsurg.2016.09.002

Bavi, P., Uddin, S., Ahmed, M., Jehan, Z., Bu, R., Abubaker, J., et al. (2011). Bortezomib stabilizes mitotic cyclins and prevents cell cycle progression via inhibition of UBE2C in colorectal carcinoma. *Am. J. Pathol.* 178, 2109–2120. doi:10.1016/j.ajpath.2011.01.034

Bader, G. D., and Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinforma.* 4, 2. doi:10.1186/1471-2105-4-2

Berlingieri, M. T., Pallante, P., Guida, M., Nappi, C., Masciullo, V., Scambia, G., et al. (2007). UbcH10 expression may be a useful tool in the prognosis of ovarian carcinomas. *Oncogene* 26 (14), 2136–2140. doi:10.1038/sj.onc.1210010

Blaner, W. S., Brun, P. J., Calderon, R. M., and Golczak, M. (2020). Retinol-binding protein 2 (RBP2): Biology and pathobiology. *Crit. Rev. Biochem. Mol. Biol.* 55 (2), 197–218. doi:10.1080/10409238.2020.1768207

Bondoc, A., Glaser, K., Jin, K., Lake, C., Cairo, S., Geller, J., et al. (2021). Identification of distinct tumor cell populations and key genetic mechanisms through single cell sequencing in hepatoblastoma. *Commun. Biol.* 4, 1049. doi:10.1038/S42003-021-02562-8

Cairo, S., Armengol, C., Reyniès, A. D., Wei, Y., Thomas, E., Renard, C. A., et al. (2008). Hepatic stem-like phenotype and interplay of Wnt/beta-catenin and Myc signaling in aggressive childhood liver cancer. *Cancer Cell.* 14, 471–484. doi:10.1016/J.CCR.2008.11.002

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinforma.* 14, 128. doi:10.1186/1471-2105-14-128

Chen, Y., and Goldstein, J. (2012). The transcriptional regulation of the human CYP2C genes. *Curr. Drug Metab.* 10, 567–578. doi:10.2174/138920009789375397

Counihan, J. L., Grossman, E. A., and Nomura, D. K. (2018). Cancer metabolism: Current understanding and therapies. *Chem. Rev.* 118, 6893–6923. doi:10.1021/ACS.CHEMREV.7B00775

Crippa, S., Ancey, P., Vazquez, J., Angelino, P., Rougemont, A. L., Guettier, C., et al. (2017). Mutant *CTNNB 1* and histological heterogeneity define metabolic subtypes of hepatoblastoma. *EMBO Mol. Med.* 9, 1589–1604. doi:10.15252/emmm.201707814

Cui, X., Liu, X., Han, Q., Zhu, J., Li, J., Ren, Z., et al. (2019). DPEP1 is a direct target of miR-193a-5p and promotes hepatoblastoma progression by PI3K/Akt/mTOR pathway. *Cell. Death Dis.* 10, 701–716. doi:10.1038/s41419-019-1943-0

Doi, I. (1976). Establishment of a cell line and its clonal sublines from a patient with hepatoblastoma. *Gan* 67, 1–10.

Du, R., Huang, C., Liu, K., Li, X., and Dong, Z. (2021). Targeting AURKA in cancer: Molecular mechanisms and opportunities for cancer therapy. *Mol. Cancer* 20, 15. doi:10.1186/s12943-020-01305-3

Faubert, B., Solmonson, A., and DeBerardinis, R. J. (2020). Metabolic reprogramming and cancer progression. *Science* 80, eaaw5473. doi:10.1126/science.aaw5473

Feng, J., Polychronidis, G., Heger, U., Frongia, G., Mehrabi, A., and Hoffmann, K. (2019). Incidence trends and survival prediction of hepatoblastoma in children: A population-based study. *Cancer Commun.* 39, 62. doi:10.1186/s40880-019-0411-7

Gao, S., Gang, J., Yu, M., Xin, G., and Tan, H. (2021). Computational analysis for identification of early diagnostic biomarkers and prognostic biomarkers of liver cancer based on GEO and TCGA databases and studies on pathways and biological functions affecting the survival time of liver cancer. *BMC Cancer* 21 (1), 791. doi:10.1186/S12885-021-08520-1

Giessner, C., Millet, V., Mostert, K. J., Gensollen, T., Vu Manh, T. P., Garibal, M., et al. (2018). Vnn1 pantetheinase limits the Warburg effect and sarcoma growth by rescuing mitochondrial activity. *Life Sci. Alliance* 1, e201800073. doi:10.26508/LSA.201800073

Grinde, M. T., Skrbo, N., Moestue, S. A., Rødland, E. A., Borgan, E., Kristian, A., et al. (2014). Interplay of choline metabolites and genes in patient-derived breast cancer xenografts. *Breast Cancer Res.* 16, R5. doi:10.1186/BCR3597

Guo, J., Wang, M., Wang, J-P., and Wu, C-X. (2019). Ubiquitin-conjugating enzyme E2T knockdown suppresses hepatocellular tumorigenesis via inducing cell cycle arrest and apoptosis. *World J. Gastroenterol.* 25, 6386–6403. doi:10.3748/WJG.V25.I43.6386

Gürtler, A., Kunz, N., Gomolka, M., Hornhardt, S., Friedl, A. A., McDonald, K., et al. (2013). Stain-Free technology as a normalization tool in Western blot analysis. *Anal. Biochem.* 433, 105–111. doi:10.1016/j.ab.2012.10.010

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nat* 585, 357–362. doi:10.1038/s41586-020-2649-2

Hooks, K. B., Audoux, J., Fazli, H., Lesjean, S., Ernault, T., Dugot-Senant, N., et al. (2018). New insights into diagnosis and therapeutic options for proliferative hepatoblastoma. *Hepatology* 68, 89–102. doi:10.1002/hep.29672

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* 9, 90–95. doi:10.1109/MCSE.2007.55

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi:10.1093/biostatistics/4.2.249

Kallio, M. A., Tuimala, J. T., Hupponen, T., Klemelä, P., Gentile, M., Scheinin, I., et al. (2011). Chipster: User-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* 12, 507. doi:10.1186/1471-2164-12-507

Kats, D., Ricker, C. A., Berlow, N. E., Noblet, B., Nicolle, D., Mevel, K., et al. (2019). Volasertib preclinical activity in high-risk hepatoblastoma. *Oncotarget* 10, 6403–6417. doi:10.18632/oncotarget.27237

Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat: A fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi:10.1038/nmeth.3317

Komander, D., and Rape, M. (2012). The ubiquitin code. *Ubiquitin Code* 81, 203–229. doi:10.1146/annurev-biochem-060310-170328

Kraft, S., Moore, J. B., Muzikansky, A., Scott, K. L., and Duncan, L. M. (2017). Differential UBE2C and HOXA1 expression in melanocytic nevi and melanoma. *J. Cutan. Pathol.* 44, 843–850. doi:10.1111/CUP.12997

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi:10.1093/nar/gkw377

Kullak-Ublick, G. A., Beuers, U., and Paumgartner, G. (1996). Molecular and functional characterization of bile acid transport in human hepatoblastoma HepG2 cells. *Hepatology* 23, 1053–1060. doi:10.1002/hep.510230518

Li, C., and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* 98, 31–36. doi:10.1073/pnas.011404098

Li, J., Zhi, X., Shen, X., Chen, C., Yuan, L., Dong, X., et al. (2020). Depletion of UBE2C reduces ovarian cancer malignancy and reverses cisplatin resistance via downregulating CDK1. *Biochem. Biophys. Res. Commun.* 523, 434–440. doi:10.1016/j.bbrc.2019.12.058

Linabery, A. M., and Ross, J. A. (2008). Trends in childhood cancer incidence in the U.S. *Cancer* 112, 416–432. doi:10.1002/cncr.23169

Liu, Y., Zhao, R., Chi, S., Zhang, W., Xiao, C., Zhou, X., et al. (2020). UBE2C is upregulated by estrogen and promotes epithelial–mesenchymal transition via p53 in endometrial cancer. *Mol. Cancer Res.* 18, 204–215. doi:10.1158/1541-7786.MCR-19-0561

Loussouarn, D., Campion, L., Leclair, F., Campone, M., Charbonnel, C., Ricolleau, G., et al. (2009). Validation of UBE2C protein as a prognostic marker in node-positive breast cancer. *Br. J. Cancer* 101, 166–173. doi:10.1038/sj.bjc.6605122

Ma, R., Kang, X., Zhang, G., Fang, F., Du, Y., and Lv, H. (2016). High expression of UBE2C is associated with the aggressive progression and poor outcome of malignant glioma. *Oncol. Lett.* 11, 2300–2304. doi:10.3892/OL.2016.4171

McKinney, W. (2010). “Data structures for statistical computing in Python,” in Proceedings of the 9th Python in Science Conference, Austin, TX, June 28- July 3, 2010, 56–61.

McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B. A., Karhohs, K. W., et al. (2018). CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* 16, e2005970. doi:10.1371/JOURNAL.PBIO.2005970

Miranda, M., and Sorkin, A. (2007). Regulation of receptors and transporters by ubiquitination: New insights into surprisingly similar mechanisms. *Mol. Interv.* 7, 157–167. doi:10.1124/MI.7.3.7

Naso, F. D., Boi, D., Ascanelli, C., Pamfil, G., Lindon, C., Paiardini, A., et al. (2021). Nuclear localisation of aurora-A: Its regulation and significance for aurora-A functions in cancer. *Oncogene* 40, 3917–3928. doi:10.1038/s41388-021-01766-w

Nong, S., Han, X., Xiang, Y., Qian, Y., Wei, Y., Zhang, T., et al. (2023). Metabolic reprogramming in cancer: Mechanisms and therapeutics. *MedComm* 4, e218. doi:10.1002/MCO2.218

Noronha, A., Modamio, J., Jarosz, Y., Guerard, E., Sompairac, N., Preciat, G., et al. (2019). The Virtual Metabolic Human database: Integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Res.* 47, D614–D624. doi:10.1093/nar/gky992

Palumbo, A., Jr, Costa, N. M. D., De Martino, M., Sepe, R., Pellecchia, S., de Sousa, V. P. L., et al. (2016). UBE2C is overexpressed in ESCC tissues and its abrogation attenuates the malignant phenotype of ESCC cell lines. *Oncotarget* 7, 65876–65887. doi:10.18632/ONCOTARGET.11674

Paquette, K., Coltin, H., Boivin, A., Amre, D., Nuyt, A. M., and Luu, T. M. (2019). Cancer risk in children and young adults born preterm: A systematic review and meta-analysis. *PLoS One* 14, e0210366. doi:10.1371/journal.pone.0210366

Raina, P., Guinea, R., Chatsirisupachai, K., Lopes, I., Farooq, Z., Guinea, C., et al. (2022). GeneFriends: Gene co-expression databases and tools for humans and model organisms. *Nucleic Acids Res.* 51, D145–D158. doi:10.1093/nar/gkac1031

Rawat, A., Gopal, G., Selvaluxmy, G., and Rajkumar, T. (2013). Inhibition of ubiquitin conjugating enzyme UBE2C reduces proliferation and sensitizes breast cancer cells to radiation, doxorubicin, tamoxifen and letrozole. *Cell. Oncol.* 36, 459–467. doi:10.1007/s13402-013-0150-8

Reddy, S. K., Rape, M., Margansky, W. A., and Kirschner, M. W. (2007). Ubiquitination by the anaphase-promoting complex drives spindle checkpoint inactivation. *Nature* 446, 921–925. doi:10.1038/nature05734

Rivas, M. P., Aguiar, T. F. M., Maschietto, M., Lemes, R. B., Caires-Júnior, L. C., Goulart, E., et al. (2020). Hepatoblastomas exhibit marked NNMT downregulation driven by promoter DNA hypermethylation. *Tumor Biol.* 42, 1010428320977124. doi:10.1177/1010428320977124

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Sekiguchi, M., Seki, M., Kawai, T., Yoshida, K., Yoshida, M., Isobe, T., et al. (2020). Integrated multiomics analysis of hepatoblastoma unravels its heterogeneity and provides novel druggable targets. *npj Precis. Oncol.* 4, 20–12. doi:10.1038/s41698-020-0125-y

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Shen, Z., Jiang, X., Zeng, C., Zheng, S., Luo, B., Zeng, Y., et al. (2013). High expression of ubiquitin-conjugating enzyme 2C (UBE2C) correlates with nasopharyngeal carcinoma progression. *BMC Cancer* 13, 192. doi:10.1186/1471-2407-13-192

Sivaprakasam, P., Gupta, A. A., Greenberg, M. L., Capra, M., and Nathan, P. C. (2011). Survival and long-term outcomes in children with hepatoblastoma treated with continuous infusion of cisplatin and doxorubicin. *J. Pediatr. Hematol. Oncol.* 33, e226–e230. doi:10.1097/MPH.0b013e31821f0eaf

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, 1027. doi:10.2202/1544-6115.1027

Spector, L. G., and Birch, J. (2012). The epidemiology of hepatoblastoma. *Pediatr. Blood Cancer* 59, 776–779. doi:10.1002/pbc.24215

Sreedhar, A., and Zhao, Y. (2018). Dysregulated metabolic enzymes and metabolic reprogramming in cancer cells. *Biomed. Rep.* 8, 3–10. doi:10.3892/br.2017.1022

Sullivan, L. B., Gui, D. Y., and Heiden, M. G. Vander (2016). Altered metabolite levels in cancer: Implications for tumour biology and cancer therapy. *Nat. Rev. Cancer* 16 (11), 680–693. doi:10.1038/nrc.2016.85

Sun, R., Li, S., Zhao, K., Diao, M., and Li, L. (2021). Identification of ten core hub genes as potential biomarkers and treatment target for hepatoblastoma. *Front. Oncol.* 11, 591507. doi:10.3389/FONC.2021.591507

Sun, T., Liu, Z., and Yang, Q. (2020). The role of ubiquitination and deubiquitination in cancer metabolism. *Mol. Cancer* 19 (1), 146. doi:10.1186/s12943-020-01262-x

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi:10.1093/nar/gky1131

Tan, J., Xu, W., Lei, L., Liu, H., Wang, H., Cao, X., et al. (2020). Inhibition of aurora kinase a by alisertib reduces cell proliferation and induces apoptosis and autophagy in HuH-6 human hepatoblastoma cells. *Onco Targets Ther.* 13, 3953–3963. doi:10.2147/OTT.S228656

Tian, L., Chen, T., Lu, J., Yan, J., Zhang, Y., Qin, P., et al. (2021). Integrated protein-protein interaction and weighted gene Co-expression network analysis uncover three key genes in hepatoblastoma. *Front. Cell. Dev. Biol.* 9, 631982. doi:10.3389/fcell.2021.631982

Udatsu, Y., Kusafuka, T., Kuroda, S., Miao, J., and Okada, A. (2001). High frequency of β-catenin mutations in hepatoblastoma. *Pediatr. Surg. Int.* 17, 508–512. doi:10.1007/s003830000576

Váraljai, R., Islam, A. B. M. M. K., Beshiri, M. L., Rehman, J., Lopez-Bigas, N., and Benevolenskaya, E. V. (2015). Increased mitochondrial function downstream from KDM5A histone demethylase rescues differentiation in pRB-deficient cells. *Genes. Dev.* 29, 1817–1834. doi:10.1101/GAD.264036.115

Volkova, M., and Russell, R. (2012). Anthracycline cardiotoxicity: Prevalence, pathogenesis and treatment. *Curr. Cardiol. Rev.* 7, 214–220. doi:10.2174/157340311799960645

Wagner, A. E., Schwarzmayr, T., Häberle, B., Vokuhl, C., Schmid, I., von Schweinitz, D., et al. (2020). SP8 promotes an aggressive phenotype in hepatoblastoma via FGF8 activation. *Cancers* 12, 2294. doi:10.3390/CANCERS12082294

Wang, J., Tian, R., Shan, Y., Li, J., Gao, H., Xie, C., et al. (2020). Metabolomics study of the metabolic changes in hepatoblastoma cells in response to NTCP/SLC10A1 overexpression. *Int. J. Biochem. Cell. Biol.* 125, 105773. doi:10.1016/j.biocel.2020.105773

Wang, R., Song, Y., Liu, X., Wang, Q., Wang, Y., Li, L., et al. (2017). UBE2C induces EMT through Wnt/β-catenin and PI3K/Akt signaling pathways by regulating phosphorylation levels of Aurora-A. *Int. J. Oncol.* 50, 1116–1126. doi:10.3892/ijo.2017.3880

Woodfield, S. E., Shi, Y., Patel, R. H., Chen, Z., Shah, A. P., Srivastava, R. K., et al. (2021). MDM4 inhibition: A novel therapeutic strategy to reactivate p53 in hepatoblastoma. *Sci. Rep.* 11 (1), 2967. doi:10.1038/s41598-021-82542-4

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., et al. (2021). Gene set knowledge discovery with Enrichr. *Curr. Protoc.* 1, e90. doi:10.1002/cpz1.90

Xiong, Y., Lu, J., Fang, Q., Lu, Y., Xie, C., Wu, H., et al. (2019). UBE2C functions as a potential oncogene by enhancing cell proliferation, migration, invasion, and drug resistance in hepatocellular carcinoma cells. *Biosci. Rep.* 39. doi:10.1042/BSR20182384

Zhang, R-Y., Liu, Z-K., Wei, D., Yong, Y. L., Lin, P., Li, H., et al. (2021). UBE2S interacting with TRIM28 in the nucleus accelerates cell cycle by ubiquitination of p27 to promote hepatocellular carcinoma development. *Signal Transduct. Target Ther.* 61 (6), 64–12. doi:10.1038/s41392-020-00432-z

Zhou, Z., Li, Y., Hao, H., Wang, Y., Zhou, Z., Wang, Z., et al. (2019). Screening hub genes as prognostic biomarkers of hepatocellular carcinoma by bioinformatics analysis. *Cell. Transpl.* 28, 76S–86S. doi:10.1177/0963689719893950

Zsíros, J., Maibach, R., Shafford, E., Brugieres, L., Brock, P., Czauderna, P., et al. (2010). Successful treatment of childhood high-risk hepatoblastoma with dose-intensive multiagent chemotherapy and surgery: Final results of the SIOPEL-3HR study. *J. Clin. Oncol.* 28, 2584–2590. doi:10.1200/JCO.2009.22.4857

Zynger, D. L., Gupta, A., Luan, C., Chou, P. M., Yang, G. Y., and Yang, X. J. (2008). Expression of glypican 3 in hepatoblastoma: An immunohistochemical study of 65 cases. *Hum. Pathol.* 39, 224–230. doi:10.1016/j.humpath.2007.06.006

Glossary

AURKA	aurora kinase A
CDK1	cyclin-dependent kinase 1
CDKN1A	cyclin-dependent kinase inhibitor 1A
CYP2C18	cytochrome P450 Family 2 subfamily C member 18
DEG	differentially expressed gene
DPEP1	dipeptidase 1
E2F2	E2F transcription factor 2
FFPE	formalin-fixed paraffin-embedded
GLUT3	glucose transporter 3
GO	gene ontology
HB	hepatoblastoma
HERC3	HECT and RLD domain containing E3 ubiquitin protein ligase 3
IGDCC3	immunoglobulin superfamily DCC subclass member 3
KEGG	kyoto encyclopedia of genes and genomes
NCL	non-cancerous liver
NL	normal liver
NNMT	nicotinamide N-methyltransferase
NT	non-targeting
ODAM	odontogenic ameloblast-associated protein
PBS	phosphate-buffered saline
PCYT1B	phosphate cytidylyltransferase 1B
PIDD1	P53-onduced death domain protein 1
PDX	patient-derived xenografts
PIK3C2B	phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 beta
PPI	protein-protein interaction
p53	tumor protein p53
RBP2	retinol binding protein 2
RNF130	ring finger protein 130
RNF144B	ring finger protein 144B
ROC	receiver operating characteristic
SAA1	serum amyloid A1
SAA2	serum amyloid A2
siRNA	small interfering RNA
SLC10A1	solute carrier family 10 member 1
TRIM71	tripartite motif containing 71
UBE2C	ubiquitin conjugating enzyme E2 C
VMH	virtual metabolic human
VNN1	vanin1

TYPE Original Research
PUBLISHED 29 June 2023
DOI 10.3389/fgene.2023.1193944

OPEN ACCESS

EDITED BY
William C. Cho,
QEH, Hong Kong SAR, China

REVIEWED BY
Gulden Olgun,
National Cancer Institute (NIH),
United States
Anca Maria Cimpean,
Victor Babes University of Medicine and
Pharmacy, Romania

*CORRESPONDENCE
Debo Chen,
✉ deboqz@163.com

†These authors have contributed equally to this work

RECEIVED 26 March 2023
ACCEPTED 21 June 2023
PUBLISHED 29 June 2023

CITATION
Wang Z, Du X, Lian W, Chen J, Hong C, Li L and Chen D (2023), A novel disulfidptosis-associated expression pattern in breast cancer based on machine learning.
Front. Genet. 14:1193944.
doi: 10.3389/fgene.2023.1193944

A novel disulfidptosis-associated expression pattern in breast cancer based on machine learning

Zhitang Wang†, Xianqiang Du†, Weibin Lian, Jialin Chen, Chengye Hong, Liangqiang Li and Debo Chen*

Department of Breast, The First Hospital of Quanzhou Affiliated to Fujian Medical University, Quanzhou, China

Background: Breast cancer (BC), the leading cause of cancer-related deaths among women, remains a serious threat to human health worldwide. The biological function and prognostic value of disulfidptosis as a novel strategy for BC treatment via induction of cell death remain unknown.

Methods: Gene mutations and copy number variations (CNVs) in 10 disulfidptosis genes were evaluated. Differential expression, prognostic, and univariate Cox analyses were then performed for 10 genes, and BC-specific disulfidptosis-related genes (DRGs) were screened. Unsupervised consensus clustering was used to identify different expression clusters. In addition, we screened the differentially expressed genes (DEGs) among different expression clusters and identified hub genes. Moreover, the expression level of DEGs was detected by RT-qPCR in cellular level. Finally, we used the least absolute shrinkage and selection operator (LASSO) regression algorithm to establish a prognostic feature based on DEGs, and verified the accuracy and sensitivity of its prediction through prognostic analysis and subject operating characteristic curve analysis. The correlation of the signature with the tumor immune microenvironment and tumor stemness was analyzed.

Results: Disulfidptosis genes showed significant CNVs. Two clusters were identified based on three DRGs (DNUFS1, LRPPRC, SLC7A11). Cluster A was found to be associated with better survival outcomes($p < 0.05$) and higher levels of immune cell infiltration($p < 0.05$). A prognostic signature of four disulfidptosis-related DEGs (KIF21A, APOD, ALOX15B, ELOVL2) was developed by LASSO regression analysis. The signature showed a good prediction ability. In addition, the prognostic signature in this study were strongly related to the tumor microenvironment (TME), tumor immune cell infiltration, tumor mutation burden (TMB), tumor stemness, and drug sensitivity.

Conclusion: The prognostic signature we constructed based on disulfidptosis-DEGs is a good predictor of prognosis in patients with BC. This prognostic signature is closely related to TME, and its potential correlation provides clues for further studies.

KEYWORDS

breast cancer, disulfidptosis, prognostic signature, tumor microenvironment, expression pattern

1 Introduction

Breast cancer (BC), the most frequently diagnosed malignancy in women, is a highly heterogeneous disease that accounted for 30% of female malignancies in 2020. This malignancy poses a great threat to women's health, due to its extremely high recurrence and mortality rates (Siegel et al., 2019; Sung et al., 2021). At present, treatment strategies for BC mainly include surgery, radiotherapy, chemotherapy, hormone therapy, targeted therapy, and immunotherapy. Despite these, however, the mortality rate for BC remains very high (Wang et al., 2021). Therefore, it is imperative to explore new therapeutic targets and reliable prognostic models in order to achieve optimal BC clinical outcomes.

Disulfidptosis is a new type of programmed cell death that has been found to be independent of apoptosis, iron death, necrotic apoptosis, and copper death (Vanden Berghe et al., 2014; Liu et al., 2023). Disulfidptosis is a rapid cell death mechanism caused by disulfide stress resulting from the accumulation of excess cysteine in cells, which usually occurs during glucose starvation (Liu et al., 2023). In glucose-deficient cancer cells expressing high levels of SLC7A11, a large accumulation of disulfide molecules leads to abnormal disulfide formation in the actin cytoskeleton, interfering with the organization of tissues and ultimately leading to the breakdown of the actine network and eventual cell death (Liu et al., 2020). We identified several genes involved in disulfidptosis that may provide novel strategies for predicting outcomes in patients with BC.

This study systematically studied the genomic characteristics of BC-specific disulfidptosis-related genes (DRGs). Based on DRGs, two disulfidptosis expression patterns were determined by unsupervised consensus clustering. The differences in prognosis, clinicopathological factors, and immune features between the two clusters were elucidated. In addition, the prognostic signature based on differentially expressed genes (DEGs) between the two disulfidptosis subtypes has been established to quantify disulfidptosis-related characteristics, high risk score predicted poor prognosis and higher TMB in BC patients. We then analyzed tumor microenvironment (TME) evaluation scores, tumor mutation burden (TMB) associations, RNA based stemness scores (RNAss) associations, and differences in chemotherapy sensitivity in the high-low risk group. These results suggest that disulfidptosis related genes play an important role in BC, which helps us to evaluate the prognosis of patients with BC and their response to chemotherapy and immunotherapy, and these genes may be potential synergistic targets to improve the therapeutic efficacy of BC.

2 Methods

2.1 Public data acquisition and preprocessing

Disulfidptosis-related gene lists were acquired from recently published literature (Liu et al., 2023). The gene expression data, corresponding survival information, copy number variations (CNVs), and somatic mutation data of patients with BC were obtained from The Cancer Genome Atlas (TCGA) database. Bulk RNA expression matrices were calibrated to the TPM format for subsequent analysis, and the GSE86166 and TCGA-BRCA bulk RNA expression matrices were integrated to form a complete queue. The data were then randomly divided at a ratio of 1:1, into training and test cohorts for subsequent analyses.

The "maftools" R package (version 4.2.2) was used to characterize DRGs and tumor mutation burden (TMB). The "ggpubr" R package was used to analyze the correlation between risk score and TMB, and the boxplot and correlation graph were used to visualize the results. Based on the CNV data, we analyzed the frequency of CNVs in DRGs and used the "RCircos" R package to locate CNVs on the 22 somatic human chromosomes, as well as the X/Y sex chromosomes.

2.2 Screening of BC-specific disulfidptosis-related genes

We investigated the differences in the expression levels of DRGs between tumor and normal samples. Statistical significance was considered to be $p < 0.05$. Univariate Cox regression and Kaplan–Meier (KM) analyses were used to screen for BC-specific DRGs. The "limma" and "reshape2" R packages were used to screen DRGs. The KM survival analysis and univariate Cox analysis based on above genes were performed using the R packages "survival" and "survminer." Venn diagrams were constructed using the R packages "ggplot2" and "VennDiagram."

2.3 Unsupervised clutering for disulfidptosis-related genes

A consensus clustering algorithm based on the R package "ConsensuClusterPlus" with 1000 permutations was used to calculate the number of disulfidptosis clusters in the overall cohorts. Principal component analysis (PCA) was conducted to verify the expression patterns using the R packages "limma" and "ggplot2."

2.4 Functional enrichment analysis

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed for patients between high- and low-risk groups using the "clusterProfiler" R package. Statistical significance was set at $p < 0.05$ for GO and KEGG pathways.

2.5 Analysis of correlation with immune infiltration

Based on the LM22 gene set on the CIBERSORT website, the CIBERSORT algorithm was used to estimate the total immune infiltration of high- and low-risk groups, as well as DRGs.

2.6 Screening of hub disulfidptosis-related DEGs

Gene expression between clusters was compared by "limma" R package, and the differentially expressed genes were obtained according to | FC | > 1, $p < 0.05$. These genes were included in univariate Cox analysis to obtain genes with important value. Least absolute shrinkage and selection operator (LASSO) Cox regression was used for 10-fold cross-validation of overall survival (OS), and genes related to disulfidptosis were screened. The "glmnet" R package was used to identify genetic signatures containing biomarkers that were the most helpful for prognosis, and risk scores were calculated for each sample in all datasets based on these signatures. The risk score was calculated using the following formula:

$$Risk\ score = (KIF21A^{*}0.218) + (APOD^{*} - 0.058) + (\mathrm{ALOX15B}^{*} - 0.071) + (\mathrm{ELOVL2}^{*} - 0.087)$$

To assess the predictive ability of disulfidptosis-related differentially expressed genes (DEGs), time-dependent receiver operating characteristic (ROC) at 3 years, 5 years, and 10 years of survival were analyzed in training and test data sets using the "timeROC" R package. For survival analysis, the optimal cut-off value of risk score was analyzed using the "Survival" R package, and the samples were divided into a high-risk and low-risk group. Kaplan–Meier analysis was used to investigate the prognostic significance of disulfidptosis-related DEGs. In addition, a prognostic nomogram was established based on the TCGA-BC dataset. Time-dependent calibration curves were plotted to predict the accuracy of the nomogram.

2.7 Cell culturing

The cell lines used in the study included human normal breast cell line MCF-10A and human breast cancer cell line MDA-MB-231were purchased from Procell (Wuhan, China). Cells were cultured in DMEM medium supplemented with 10% FBS (Gibco, United States) and antibiotics (Penicillin 100 U/mL, Streptomycin 100 mg/mL) (Gibco, United States). Cells were cultured at 37°C with 5% CO2.

2.8 RNA extraction and quantitative real-time PCR (qRT-PCR)

RNA was isolated using TRIzol reagent (Invitrogen, Thermo Fisher Scientific, Waltham, MA, United States), and reverse transcription was performed using the PrimeScriptTM RT Reagent Kit (Takara; Takara Bio, Shiga, Japan). SYBR Green PCR Master Mix (Takara) was used for qRT-PCR on a StepOnePlus System (Applied Biosystems, Thermo Fisher Scientific). Fold-changes in gene expression were determined using the $2^{-\Delta\Delta CT}$ method, using GAPDH for normalization. The primers used in this study are listed in Supplementary Table S1.

2.9 Statistical analysis

The Wilcoxon rank-sum test was used to compare differences between the two groups. The K–W test was performed to compare three or more groups. Kaplan–Meier analysis was used to evaluate survival differences between the low- and high-risk- groups. All statistical analyses were done using R version 4.2.2 with $p < 0.05$ indicating statistical significance.

3 Results

3.1 Genetic alterations analysis and screening of disulfidptosis-related genes in BC

We identified 10 genes (*NCKAP1*, *LRPPRC*, *NDUFS1*, *GYS1*, *SLC3A2*, *RPN1*, *SLC7A11*, *OXSM*, *NDUFA11*, and *NUBPL*) that were closely related to disulfidptosis. We first determined the somatic mutation levels, CNVs, gene expression levels, and prognostic values of DRGs in BC samples.

Somatic mutations were not widespread in these genes (Figure 1A). Somatic mutations in the DRGs were present in 47 of the 987 samples, a frequency of 4.76%. Among these, the mutation frequencies of *NCKAP1*, *LRPPRC*, *NDUFS1*, and *GYS1* were the highest. By investigating the frequency of the CNVs, we noticed that DRGs had widespread alterations in CNVs and that most genes had a gain status that was higher than the loss status. The primary genes showing CNV amplification were *SLC3A2* and *NUBPL*. By contrast, *NDUFA11* had the highest number of CNV deletions (Figure 1B). The positions of these 10 genes on the chromosome are shown in Figure 1C. We then analyzed the expression levels of these 10 genes in cancers and their adjacent normal tissues. *NDUFA11*, *LRPPRC*, *SLC7A11*, *SLC3A2*, *OXSM*, and *RPN1* showed higher expression levels in cancer tissues, whereas *NDUFS1* and *NUBPL* were expressed at lower levels ($p < 0.01$). The expression of *NCKAP1* and *GYS1* was not significantly different between cancer and adjacent normal tissues (Figure 1D). OS analysis showed that the group with high expression of *NDUFA11* and the group with low expression of *NDUFS1*, *SLC7A11*, *OXSM*, *NCKAP1*, and *LRPPRC* had better prognoses ($p < 0.05$; Figure 1E). There were no significant differences in OS between the *NUBPL*, *RPN1*, and *SLC3A2* expression groups.

3.2 Identification of BC-specific DRGs and distinct expression patterns

Univariate Cox regression analysis identified three primary genetic risk factors: *LRPPRC*, *NDUFS1*and *SLC7A11* ($p < 0.01$; Figure 2A). Three BC-specific DRGs were identified by intersections of eight DRGs, six prognostic DRGs, and three risk factors from the univariate cox regression analysis. These were the genes *NDUFS1*, *LRPPRC* and *SLC7A11* (Figure 2B). Based on these genes, unsupervised consensus clustering of the overall cohort was performed and patients with BC in the overall cohort were categorized into clusters A and B (Figures 2C, D). PCA showed that BC samples could be distinguished according to distinct expression patterns, and our KM survival curve showed that the median OS of cluster A was better than that of cluster B (Figure 2E).

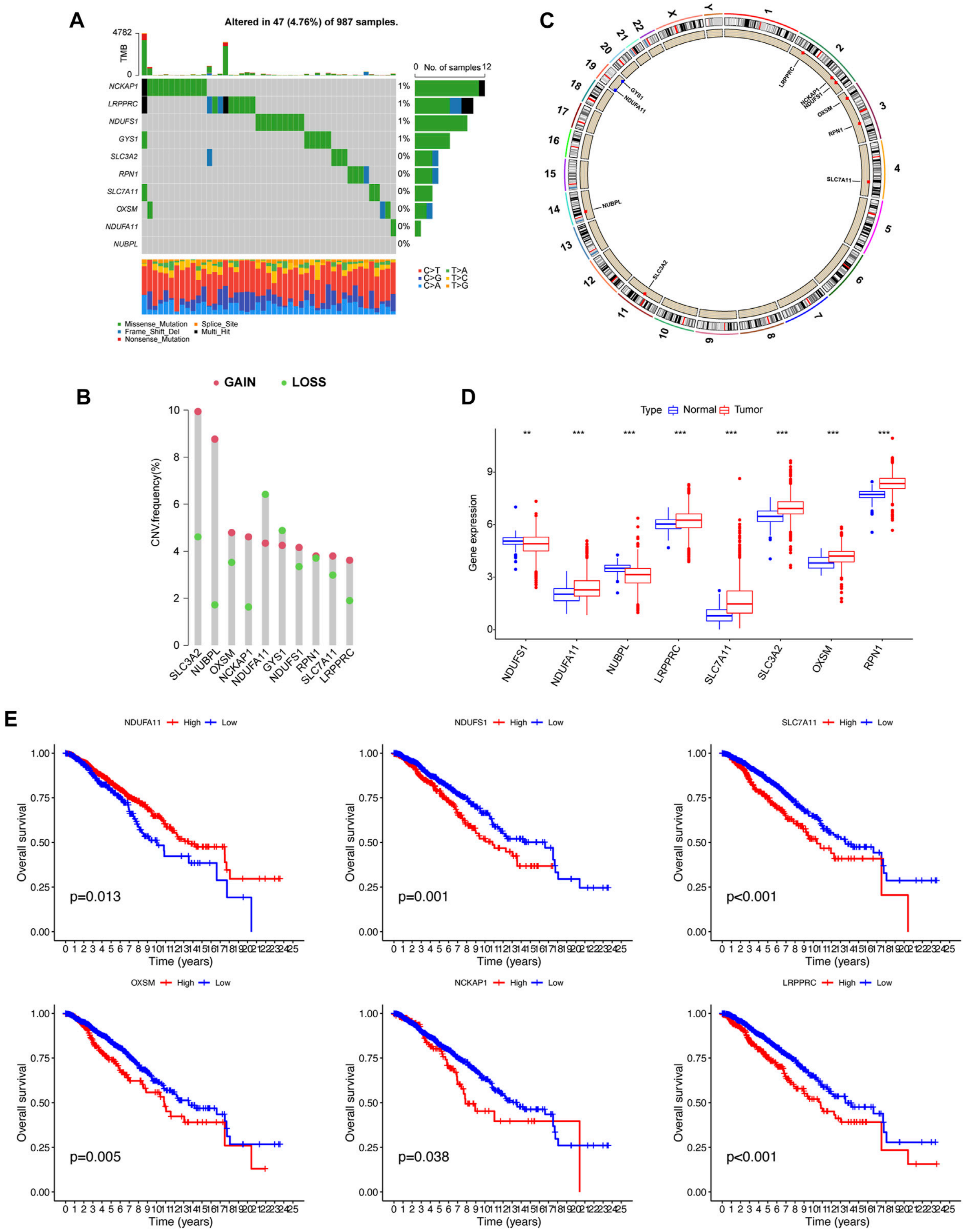


FIGURE 1
Gene mutational, copy number variations (CNV), differentially expressed, and survival analysis of disulfidptosis-related genes. **(A)** Waterfall plot showing the gene mutational frequency and types of genetic mutations. **(B, C)** Bar chart and circus show the CNV frequency and the position of the disulfidptosis-related genes on the chromosomes. **(D)** Gene expression analysis between normal and breast cancer samples. **(E)** K–M survival analysis between high and low expression of genes. $^{**}p < 0.01$, $^{***}p < 0.001$.

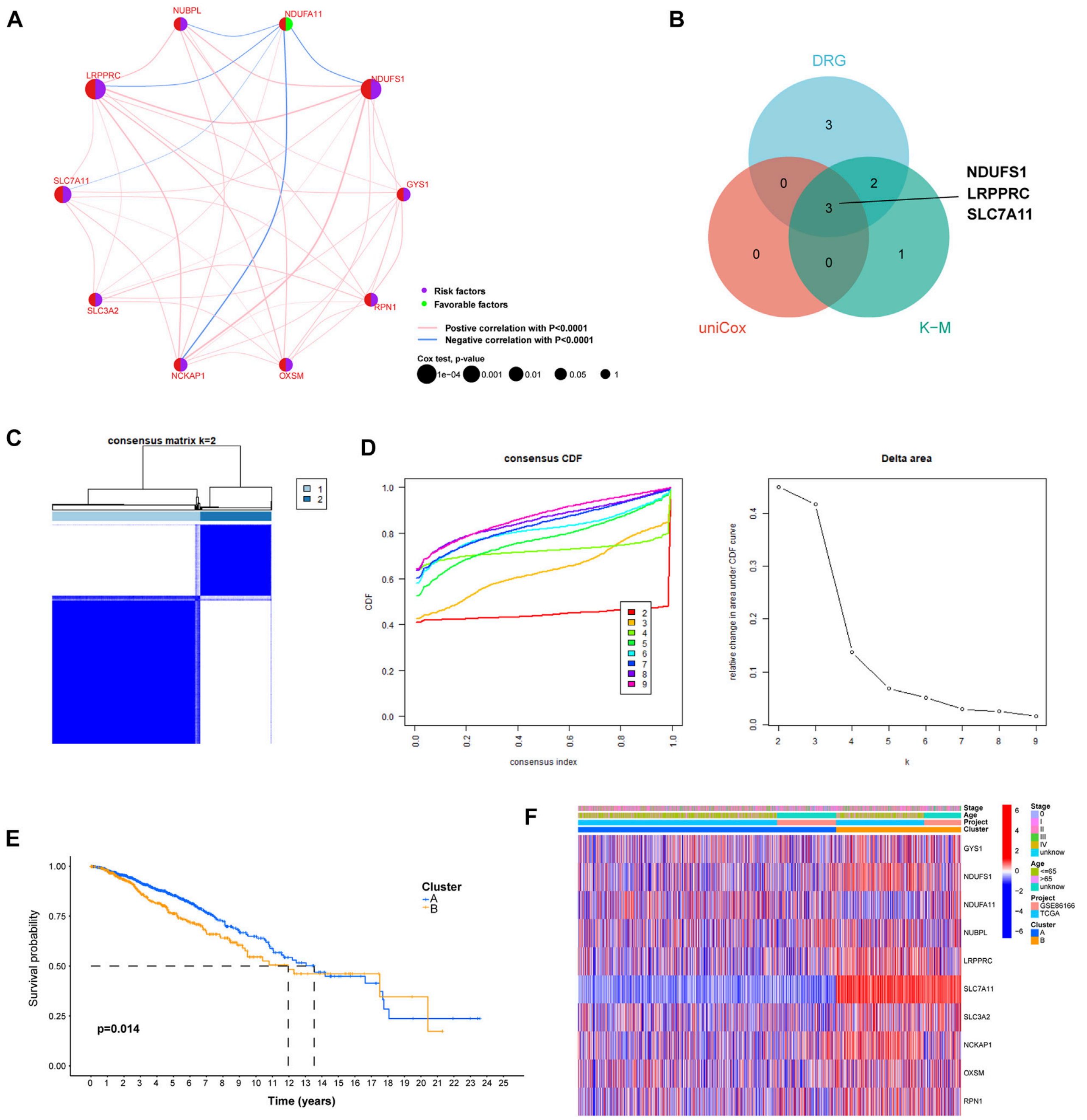


FIGURE 2
The construction of distinct disulfidptosis-related expression patterns. **(A)** Univariate Cox regression and correlation analysis between disulfidptosis-related genes. **(B)** Venn plot showing the shared genes according to the results of differentially expressed analysis, univariate Cox regression analysis, and K−M survival analysis. **(C)** The consensus clustering matrix (k = 2) was used to stratify Breast cancer (BC) patients into two clusters. **(D)** Consensus clustering model with cumulative distribution function (CDF) by k from 2-9. **(E)** K-M survival analysis between cluster A and **(B)** **(F)** The heat map shows differences in clinicopathological factors in each distinct cluster.

3.3 Correlation between expression patterns and BC molecular subtype

We constructed a heat map that showed the differences in the clinical factors between clusters A and B (Figure 2F). In order to further explore the relationship between breast cancer molecular subtypes and the expression pattern we identified, we drew the Sankey diagram and KM survival curve. The results showed that in cluster A, patients with luminal A, luminal B, HER2 and Basel subtypes account for 60.4%, 17.4%, 8.8% and 13.4%, respectively. In cluster B, luminal A, luminal B, HER2 and Basel subtypes accounted for 39.8%, 26.0%, 5.8% and 28.4%, respectively (Supplementary Figure S1A). The results indicated that the proportion of patients with Luminal A subtype in cluster A is significantly higher than that

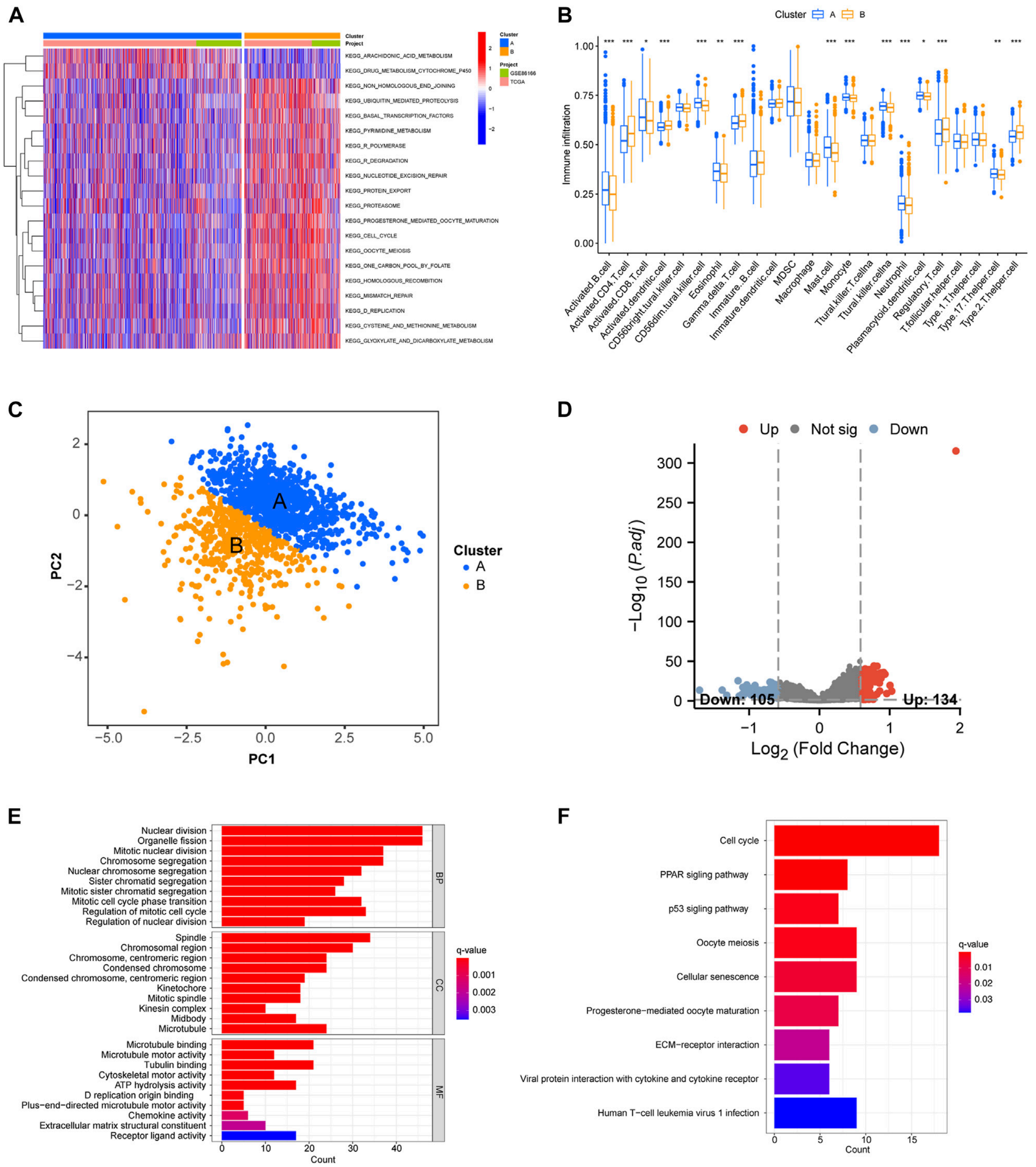


FIGURE 3
ssGSEA and immune infiltration analysis in distinct cluster and functional enrichment analysis of disulfidptosis. **(A)** Heat map plot showing our ssGSEA analysis of clusters **(A, B)**. **(B)** Box plot showing the differences between clusters **(A, B)**. **(C)** Principal Component Analysis (PCA) based on the two clusters. **(D)** The differentially expressed genes between cluster **(A, B)**. **(E, F)** GO and KEGG analysis of molecular subtype-related DEGs. **p* < 0.05, ***p* < 0.01, ****p* < 0.001.

in cluster B, and the proportion of patients with Basel subtype in cluster B is significantly higher than that in cluster A. The results of KM survival analysis showed that there was a significant difference in the prognosis of patients in the cluster A and B of luminal subtype, but no difference was found in HER2 and Basel subtypes. (Supplementary Figure S1B).

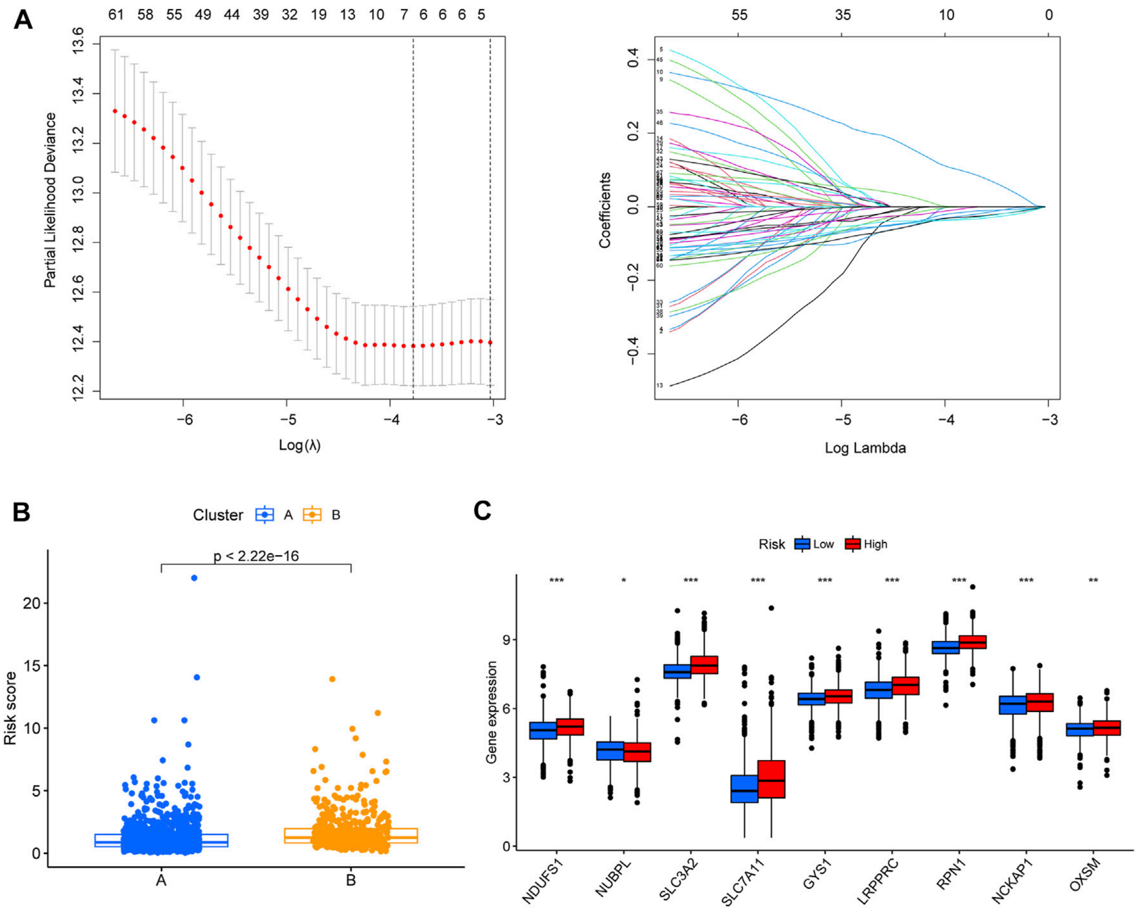


FIGURE 4
Establishment of disulfidptosis-related prognostic signature. **(A)** Lasso regression was used to establish the four-gene prognostic signature. **(B)** Box plot showing the differences in the risk score of patients between clusters **(A, B)**. **(C)** The differential gene expression analyses that were performed between low- and high-risk group. *p < 0.05, **p < 0.01, ***p < 0.001.

3.4 Analysis immune infiltrate level analysis and functional enrichment analysis between two clusters

GSVA functional enrichment analysis indicated that cluster A was mainly enriched in mutations pertaining to arachidonic acid and drug metabolism pathways. Cluster B was mainly enriched in tumor-related pathways (e.g., DNA replication and cell cycle) and metabolic pathways (e.g., primary bile acid biosynthesis, pyrimidine metabolism, cysteine and methionine metabolism, and glyoxylate and dicarboxylate metabolism; Figure 3A). As shown in Figure 3B, the extent of immune cell infiltration differed distinctly between clusters A and B. CD56^{bright} natural killer cells, immature B cells, immature dendritic cells, MDSC, macrophages, natural killer T cells, follicular helper T cells and Type 1 helper T cells were observed. The infiltration of immune cells was higher in cluster A than in cluster B. PCA analysis showed that cluster A and B could better distinguish patients into different group. Therefore, we further explored the difference between the two clusters (Figure 3C). 239 disulfidptosis-DEGs were identified between cluster A and B (Figure 3D). GO and KEGG enrichment analyses of disulfidptosis-DEGs showed that these genes were mainly enriched in cell division-related pathways (e.g., nuclear division, mitotic nuclear division, and chromosome segregation; Figure 3E). The results of the KEGG analysis showed that disulfidptosis-DEGs were significantly enriched in cancer-related pathways (e.g., cell cycle, p53 signaling pathway, and ECM-receptor interaction; Figure 3F).

3.5 Construction of prognostic signature

A total of 239 DEGs between clusters A and B, including 71 prognostic-associated disulfidptosis-DEGs were selected for univariate Cox regression analysis. A prognostic signature of four disulfidptosis-DEGs was then developed by LASSO regression analysis based on the training cohort (Figure 4A). We then verified the expression levels of four disulfidptosis-DEGs at the cellular level. KIF21A and ALOX15B were low expressed in cancer cells, APOD was high expressed in cancer cells, and ELOVL2 expression was not significantly different between cancer cells and normal cells (Supplementary Figure S2). Cluster B had a higher risk score than Cluster A (Figure 4B). Except for *NDUFA11*, the expression of nine of the DRGs differed between the high- and low-risk groups. Of these nine, *NUBPL*

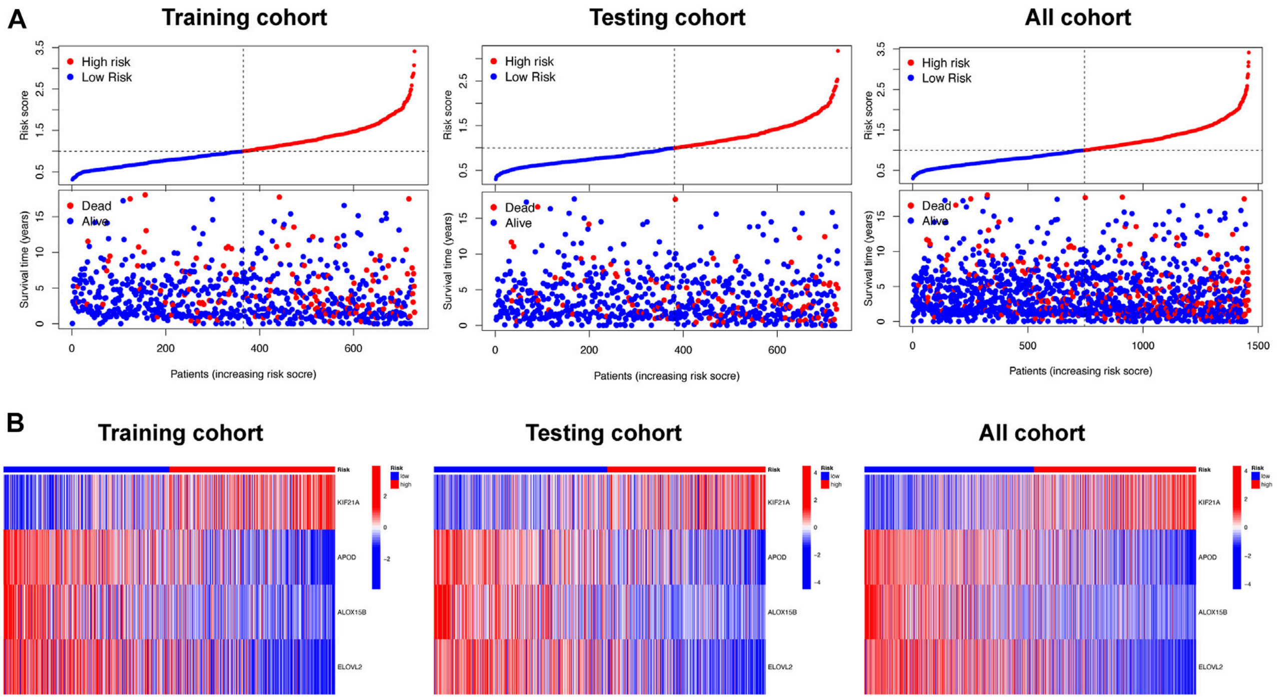


FIGURE 5
The relationship between survival status and risk score, and differential expression analysis of signature related genes, in the different risk groups. **(A)** Scatterplots showing the changes in survival statuses of BC patients as a function of increasing risk scores. **(B)** Heat map plots showing the differences between the low- and high-risk group in four signature-related genes.

was expressed at low levels in the high-risk group, whereas the remaining eight genes were highly expressed in the high-risk group (Figure 4C).

3.6 Validation of the disulfidptosis-related prognostic signature

The risk score of the high-risk group was higher than that of the low-risk group, and the number of deaths increased with risk score in the training and testing cohorts and all cohort (Figure 5A). A heatmap showed the differential expression of disulfidptosis-DEGs between the high- and low-risk groups (Figure 5B). Among these genes, *KIF21A* was highly expressed in the high-risk group, whereas *APOD*, *ALOX15B*, and *ELOVL2* were highly expressed in the low-risk group.

3.7 Evaluating the independent role of the prognostic signature and building a predictive nomogram for prognosis prediction

We also confirmed that the overall survival (OS) of the low-risk group was significantly longer than that of the high-risk group ($p < 0.05$; Figure 6A). We also explored the consistency of prognostic value of prognostic models across different molecular subtypes of BC. We found that in Luminal and Her2 subtypes, the PFS and DSS of high-risk group were worse than those of low-risk group. There was no difference in the prognosis of the high- and low-groups in Basel subtype, which may due to the small number of patients in the low-risk group (the number of patients with Basel subtype in the low- and the high-risk group was 14 and 175, respectively). However, we found that the 7-year PFS and DSS of the low-risk group was also significantly better than that of the high-risk group in the K-M survival curve. In general, the prognostic models had good prognostic value for different molecular subtypes of BC (Supplementary Figure S3). The AUCs of the prognostic signature suggested that the model had good predictive accuracy (Figure 6B). Nomograms are another quantitative model for predicting clinical outcomes in patients with BC. Therefore, a nomogram was developed based on the risk score and other clinical characteristics (e.g., age, disease stage and molecular subtype), so that the probability of survival at 1, 3, and 5 years for each patient with BC could be calculated (Figure 6C). The calibration charts used for internal validation of the line charts showed good agreement between the predicted OS results and actual observations (Figure 6D).

3.8 Analysis of immune cell infiltration, TMB, RNAss, and drug sensitivity

We used the CIBERSORT algorithm to calculate the correlation between the level of infiltration of 22 immune cells and the disulfidptosis-DEGs we identified. Among these, *APOD*

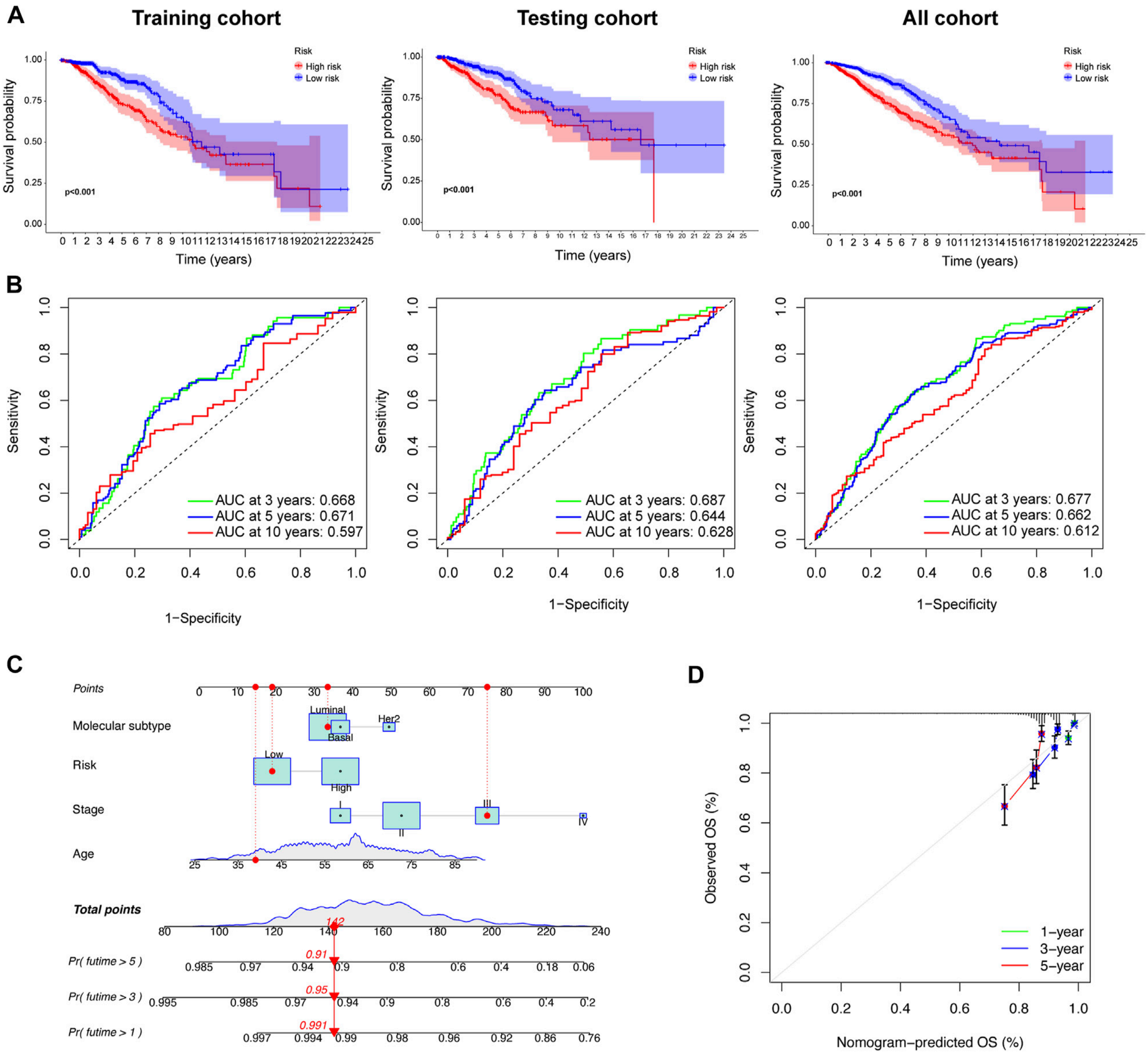


FIGURE 6
Prognostic value and reliability analyses of the prognostic signature for the training, testing, and all cohorts during development of the nomogram. **(A)** K–M survival analysis between low- and high-risk group in the three cohorts. **(B)** Receiver operating characteristic (ROC) curves were constructed, and the area under the ROC curve (AUC) were determined. **(C)** A nomogram was built based on prognostic signature and clinicopathological factors (age and disease stage). **(D)** The calibration curve showing the predictive accuracy of nomogram.

and naïve B cells, as well as *ELOVL2* and resting mast cells, showed significant positive correlations. *APOD* and M0 macrophages, as well as *ELOVL2* and CD4 resting memory T cells were negatively correlated (Figure 7A). We then analyzed the correlation between the content of stromal cells immune cells in the tumor microenvironment (TME), and the risk score. The low-risk group showed higher stromal and estimated scores (Figure 7B). Next, we analyzed whether there were differences in the TMB between the high- and low-risk groups. The results showed that the TMB frequency in the high-risk group was greater than in the low-risk group (Figure 7D). There was a positive correlation between TMB and risk score (Figure 7E). In BC, the TMBs of 20 genes with high mutation frequencies differed significantly between the high- and low-risk groups. For example, the mutation frequencies for *PIK3CA* were 23% and 46% in the high-and low-risk groups, respectively. *TP53* was mutated in 46% of the high-risk group and 18% of the low-risk group (Figure 7C). A positive correlation between RNAss and risk score was observed in tumor stemness analysis (Figure 7F). The results of drug sensitivity analysis showed that the sensitivity of low-risk group to cisplatin, cyclophosphamide, docetaxel, lapatinib, paclitaxel, and tamoxifen was higher than that of high-risk group, while the drug sensitivity of high-risk group to Ribociclib was higher than that of low-risk group, which could help to guide the selection of clinical treatment (Supplementary Figure S4).

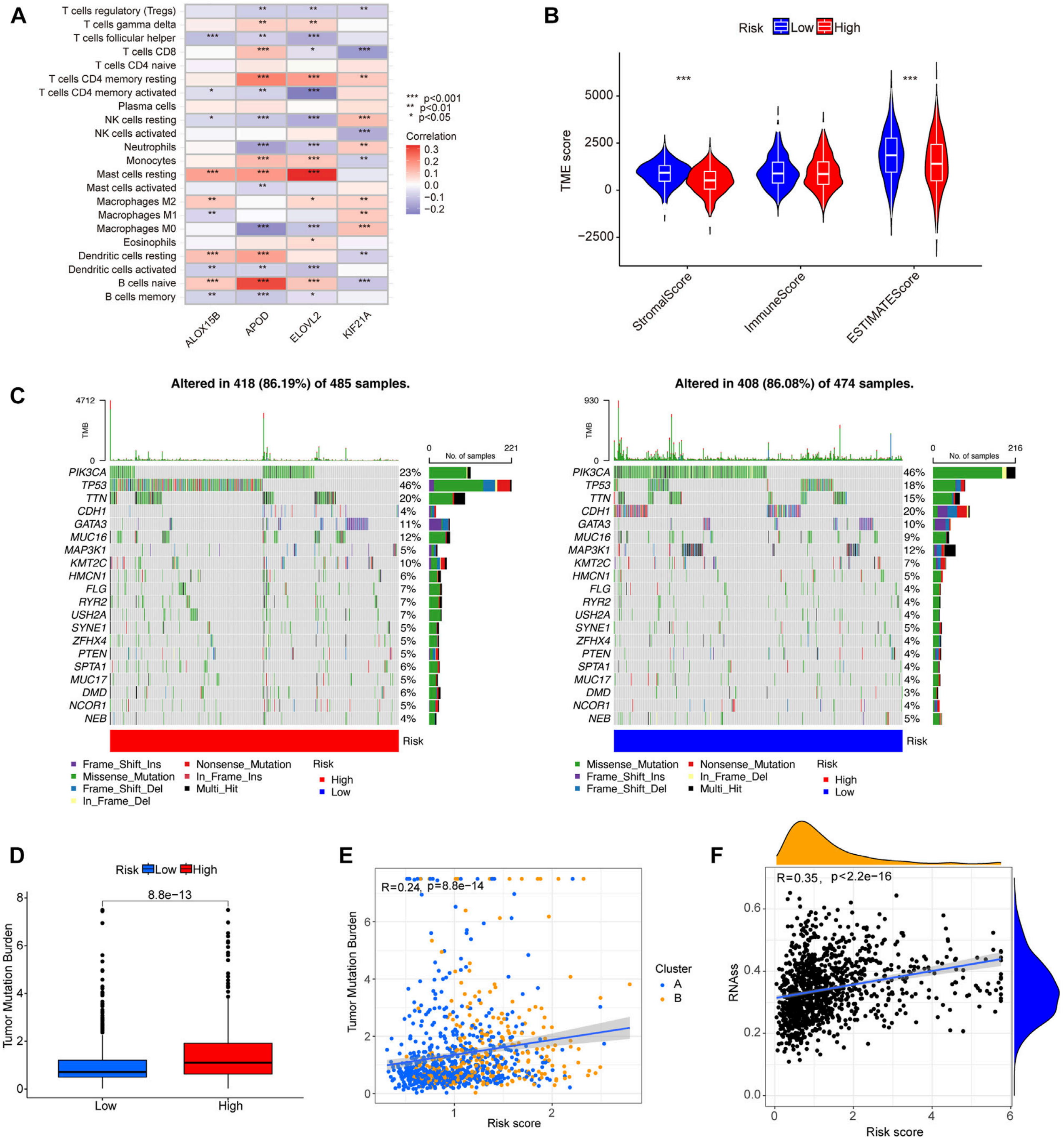


FIGURE 7
The correlation of tumor immune cell infiltration, gene mutational frequency, TMB, and RNAss with prognostic signature. **(A)** The heat map shows the correlation between four signature-related genes and level of tumor immune cell infiltration. **(B)** A violin plot showing the differences in stromal score, immune score, and estimate score between the different risk groups. **(C)** Waterfall plots showing the top 20 genes with highest gene mutational frequencies, and the types of gene mutations. **(D)** A box plot showing the difference in TMB between the low- and high-risk group. **(E)** Correlation analysis of TMB and molecular subtypes with risk score. **(F)** Correlation scatterplot showing the relationship between RNAss and risk score. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

4 Discussion

Cell energy metabolism is a necessary condition for maintaining biological development and internal environmental balance (Vander Heiden et al., 2009). Studies have shown that disulfide is closely related to energy metabolism in cancer cells. Cancer cells typically exhibit increased glucose uptake and, in the context of high SLC7A11 expression, limit NADPH production by glucose starvation or GLUT inhibition, resulting in massive accumulation of disulfide, defective oxidation-reduction reactions, and cell death

(Stockwell et al., 2017; Liu et al., 2020).Disulfidptosis has recently been identified as a new type of cell death (Liu et al., 2020; Liu et al., 2023), and a new therapeutic approach for targeting and killing cancer cells. Targeting and killing of cancer cells is a new therapeutic approach. Aberrant expression of the cystine transporter solute carrier family 7 member 11 (SLC7A11; also known as xCT), the 11th member of the seventh family of solute transporters, is a cystine/glutamate anti-transporter involved in amino acid transport across the plasma membrane (Conrad et al., 2018). In 2020, Gamboi et al. found that for cells to maintain cystine at non-toxic levels, cancer cells with high expression of SLC7A11 reduced cystine to more soluble cysteine, leading to the rapid depletion of NADPH pools and abnormal accumulation of disulfides, with resultant toxic effects that led to cell death (Liu et al., 2020). We found that the expression of SLC7A11 in BC tissues was much higher than that in adjacent normal tissues. Therefore, targeting the disulfidase pathway is a promising new strategy for BC therapy.

BC is a molecular heterogeneous disease. The classical molecular subtypes of BC classify patients into Luminal, HER2 and Basel subtypes and the prognostic characteristics and drug sensitivity are different among these molecular subtypes (Holm et al., 2017). In this study, we investigated the relationship between expression patterns we built and classical molecular subtypes of BC. We found that the proportion of patients with Luminal A subtype in cluster A is significantly higher than that in cluster B, and the proportion of patients with Basel subtype in cluster B is significantly higher than that in cluster A. The epidemiological study of breast cancer reported that the prognosis of luminal A is the best among four molecular subtypes, on the contrary, the basel subtype had the worst prognosis. This is also consistent with the results in survival analysis between cluster B and cluster A in our research. Besides, the results of subgroup analysis based on the three BRCA subtypes (Luminal, Her2, Basel) indicated that the expression pattern we identified can combined with BRCA molecular subtype for better predicting and improving the prognosis of patients with luminal subtype.

At present, there are few studies on constructing prognostic models based on disulfidptosis-related gene. Recent studies have found that disulfidptosis-related gene signature has an excellent ability to identify the immune landscape of patients with bladder cancer and predict their prognosis (Zhao et al., 2023).However, little research has been conducted on DRGs in BC. Therefore, in this study, we first integrated TCGA data and the GSE86166 dataset to screen three DRGs (*NDUFS1*, *LRPPRC*, and *SLC7A11*) with differential expression and prognostic value. According to the expression pattern of DRGs, BC patients were divided into two clusters, with significant differences in OS rate and immune cell infiltration level. Indicating that these DRGs participate in TME. Subcomponent PCA was used to evaluate the prognostic value of the two groups (clusters A and B). Subsequently, four disulfidptosis-DEGs with prognostic value were identified using LASSO Cox regression analysis, and a prognostic model was constructed. In the training and validation cohort, the OS difference between the high-risk group and the low-risk group indicates that the risk score can be used as an indicator to distinguish the BC survival rate. Multivariate Cox analysis showed that risk score, age and tumor stage were considered to be independent prognostic indicators of BC. In order to better quantify 1-year, 3-year, and 5-year OS in BC patients, a nomogram combining these independent prognostic factors was developed. The results of ROC and calibration curve showed that the nomogram had significant prognostic performance. This quantitative result can be used as a complementary tool to improve prognosis assessment and personalized treatment.

The tumor microenvironment includes a variety of complex cellular components, such as immune cells, stromal cells and tumor cells (Shi et al., 2022; Srinivasan et al., 2022). Their difference in composition and expression is one of the main causes of tumor heterogeneity. Elucidating tumor immune heterogeneity will help to identify effective synergistic targets to enhance the efficacy of BC therapy. The prognosis of cluster A was better than cluster B. Cluster A showed abundant infiltration of activated B, CD8$^+$ T, dendritic, natural killer cells and neutrophil. These immune cells kill tumor cells and promote immune responses and immunotherapy. In the constructed signature based on disulfidptosis-DEGs, the stromal and estimated scores of the low-risk group were higher than those of the high-risk group, and the immune score was also higher in the low-risk group than in the high-risk group, although the difference was not statistically significant. These findings suggest that disulfidptosis is associated with TME, and can be used to guide targeted immunotherapy.

Disulfidptosis is a novel type of cell death, and this study established a prognostic model based on disulfidptosis-DEGs for the first time. Our study adds to the understanding of the molecular biology of DRGs in BC. TCGA and GEO data were integrated to expand the sample size and improve the accuracy of the results. However, our study also had several limitations. First, this study mainly used the TCGA and GEO databases for analysis, and thus lacked real-world research, which urgently needs to be used for full verification of our results in the future. Second, the regulatory mechanism of DRGs in BC immune infiltration remains unclear, and further functional verification at tissue, cell and animal level is needed in the future. Finally, further research is needed to determine whether the model can be used to predict resistance to therapeutic agents in clinical practice.

5 Conclusion

We used consensus clustering to identify two disulfidptosis-molecular subtypes in breast cancer with different OS. We further constructed a prognostic signature based on disulfidptosis-DEGs that better predicted patient survival outcomes and tentatively identified the relationship between our risk model and the immune landscape. The results of our study provide useful insights into predicting the prognoses of patients with BC, and may even aid their treatment in clinical practice.

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

DC, ZW, and XD studied and designed all bioinformatics analysis. ZW and XD conducted the experiment. XD, WL, JC, CH, and LL critically reviewed the manuscript. JC, CH, and LL worked together on the manuscript. ZW, XD, and DC contributed equally to our research. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by grants from the Science and Technology Project of Quanzhou (No. 2020N030s), the Startup Fund for Scientific Research of Fujian Medical University (No. 2019QH1253), the Youth Project of Health Commission of Fujian Province (No. 2020QNB042), the Natural Science Foundation of Fujian Province, PRC (No. 2020J011275), and the Clinical research Special Fund of Wu Jieping Medical Foundation (320.6750.2021-10-60).

Acknowledgments

The authors want to thank the TCGA, and GEO for free use.

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/fgene.2023.1193944/full#supplementary-material

References

Conrad, M., Kagan, V. E., Bayir, H., Pagnussat, G. C., Head, B., Traber, M. G., et al. (2018). Regulation of lipid peroxidation and ferroptosis in diverse species. *Genes. Dev.* 32, 602–619. doi:10.1101/gad.314674.118

Holm, J., Eriksson, L., Ploner, A., Eriksson, M., Rantalainen, M., Li, J., et al. (2017). Assessment of breast cancer risk factors reveals subtype heterogeneity. *Cancer Res.* 77, 3708–3717. doi:10.1158/0008-5472.CAN-16-2574

Liu, X., Olszewski, K., Zhang, Y., Lim, E. W., Shi, J., Zhang, X., et al. (2020). Cystine transporter regulation of pentose phosphate pathway dependency and disulfide stress exposes a targetable metabolic vulnerability in cancer. *Nat. Cell. Biol.* 22, 476–486. doi:10.1038/s41556-020-0496-x

Liu, X., Nie, L., Zhang, Y., Yan, Y., Wang, C., Colic, M., et al. (2023). Actin cytoskeleton vulnerability to disulfide stress mediates disulfidptosis. *Nat. Cell. Biol.* 25, 404–414. doi:10.1038/s41556-023-01091-2

Shi, D. D., Guo, J. A., Hoffman, H. I., Su, J., Mino-Kenudson, M., Barth, J. L., et al. (2022). Therapeutic avenues for cancer neuroscience: Translational frontiers and clinical opportunities. *Lancet. Oncol.* 23, e62–e74. doi:10.1016/S1470-2045(21)00596-9

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA a cancer J. Clin.* 69, 7–34. doi:10.3322/caac.21551

Srinivasan, S., Kryza, T., Batra, J., and Clements, J. (2022). Remodelling of the tumour microenvironment by the kallikrein-related peptidases. *Nat. Rev. Cancer* 22, 223–238. doi:10.1038/s41568-021-00436-z

Stockwell, B. R., Friedmann Angeli, J. P., Bayir, H., Bush, A. I., Conrad, M., Dixon, S. J., et al. (2017). Ferroptosis: A regulated cell death nexus linking metabolism, redox biology, and disease. *Cell.* 171, 273–285. doi:10.1016/j.cell.2017.09.021

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA a cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Vanden Berghe, T., Linkermann, A., Jouan-Lanhouet, S., Walczak, H., and Vandenabeele, P. (2014). Regulated necrosis: The expanding network of non-apoptotic cell death pathways. *Nat. Rev. Mol. Cell. Biol.* 15, 135–147. doi:10.1038/nrm3737

Vander Heiden, M. G., Cantley, L. C., and Thompson, C. B. (2009). Understanding the warburg effect: The metabolic requirements of cell proliferation. *Sci. New York, N.Y.* 324, 1029–1033. doi:10.1126/science.1160809

Wang, D., Wei, G., Ma, J., Cheng, S., Jia, L., Song, X., et al. (2021). Identification of the prognostic value of ferroptosis-related gene signature in breast cancer patients. *BMC cancer* 21, 645. doi:10.1186/s12885-021-08341-2

Zhao, S., Wang, L., Ding, W., Ye, B., Cheng, C., Shao, J., et al. (2023). Crosstalk of disulfidptosis-related subtypes, establishment of a prognostic signature and immune infiltration characteristics in bladder cancer based on a machine learning survival framework. *Front. Endocrinol.* 14, 1180404. doi:10.3389/fendo.2023.1180404

TYPE Review
PUBLISHED 10 August 2023
DOI 10.3389/fgene.2023.1172108

OPEN ACCESS

EDITED BY
Junlin Xu,
Hunan University, China

REVIEWED BY
Yushuang Li,
Yanshan University, China
Jujuan Zhuang,
Dalian Maritime University, China

*CORRESPONDENCE
Jun Zhou,
✉ 15304690053@163.com
Pingping Bing,
✉ bpping@163.com

†The authors contributed equally to this study

RECEIVED 23 February 2023
ACCEPTED 20 April 2023
PUBLISHED 10 August 2023

CITATION
Zhu L, Xu R, Yang L, Shi W, Zhang Y, Liu J, Li X, Zhou J and Bing P (2023), Minimal residual disease (MRD) detection in solid tumors using circulating tumor DNA: a systematic review.
Front. Genet. 14:1172108.
doi: 10.3389/fgene.2023.1172108

Minimal residual disease (MRD) detection in solid tumors using circulating tumor DNA: a systematic review

Lemei Zhu[1,2,3†], Ran Xu[4†], Leilei Yang[4], Wei Shi[4], Yuan Zhang[1,2,3], Juan Liu[1,2,3], Xi Li[5], Jun Zhou[1,2*] and Pingping Bing[1,2*]

[1]Hunan Key Laboratory of the Research and Development of Novel Pharmaceutical Preparations, Changsha, China, [2]Academician Workstation, Changsha Medical University, Changsha, China, [3]School of Public Health, Changsha Medical University, Changsha, China, [4]Geneis Beijing Co., Ltd., Beijing, China, [5]Department of Orthopedics, Xiangya Hospital Central South University, Changsha, China

Minimal residual disease (MRD) refers to a very small number of residual tumor cells in the body during or after treatment, representing the persistence of the tumor and the possibility of clinical progress. Circulating tumor DNA (ctDNA) is a DNA fragment actively secreted by tumor cells or released into the circulatory system during the process of apoptosis or necrosis of tumor cells, which emerging as a non-invasive biomarker to dynamically monitor the therapeutic effect and prediction of recurrence. The feasibility of ctDNA as MRD detection and the revolution in ctDNA-based liquid biopsies provides a potential method for cancer monitoring. In this review, we summarized the main methods of ctDNA detection (PCR-based Sequencing and Next-Generation Sequencing) and their advantages and disadvantages. Additionally, we reviewed the significance of ctDNA analysis to guide the adjuvant therapy and predict the relapse of lung, breast and colon cancer et al. Finally, there are still many challenges of MRD detection, such as lack of standardization, false-negatives or false-positives results make misleading, and the requirement of validation using large independent cohorts to improve clinical outcomes.

KEYWORDS

MRD, tumor, CtDNA, biomarker, NGS

1 Introduction

Liquid biopsy, which has many advantages such as non-invasiveness, acceptability, repeatability and prediction of tumor burden and treatment response, has played an increasingly important role in the diagnosis and treatment of cancer. Cancer biomarkers can be extracted and analyzed from the blood, urine, pleural effusion, seroperitoneum, cerebrospinal fluid or saliva of cancer patients with this novel detection method. Circulating tumor cells (CTCs), cell free nucleic acids, exosomes and other biological components secreted into body fluids by cancer cells are all analytes of liquid biopsies, providing biomarkers such as somatic point mutations, amplifications, deletions, gene fusions, DNA methylation markers, miRNAs, proteins, and metabolites.

Cell-free DNA (cfDNA) consists of double-stranded DNA with a length of 150–200 base pairs that circulate mainly in the blood, released through apoptosis, necrosis, and phagocytosis (Corcoran and Chabner, 2018). The origin of cfDNA is hemopoietic cells such as erythrocytes, leukocytes and endothelial cells in healthy individuals, and normal

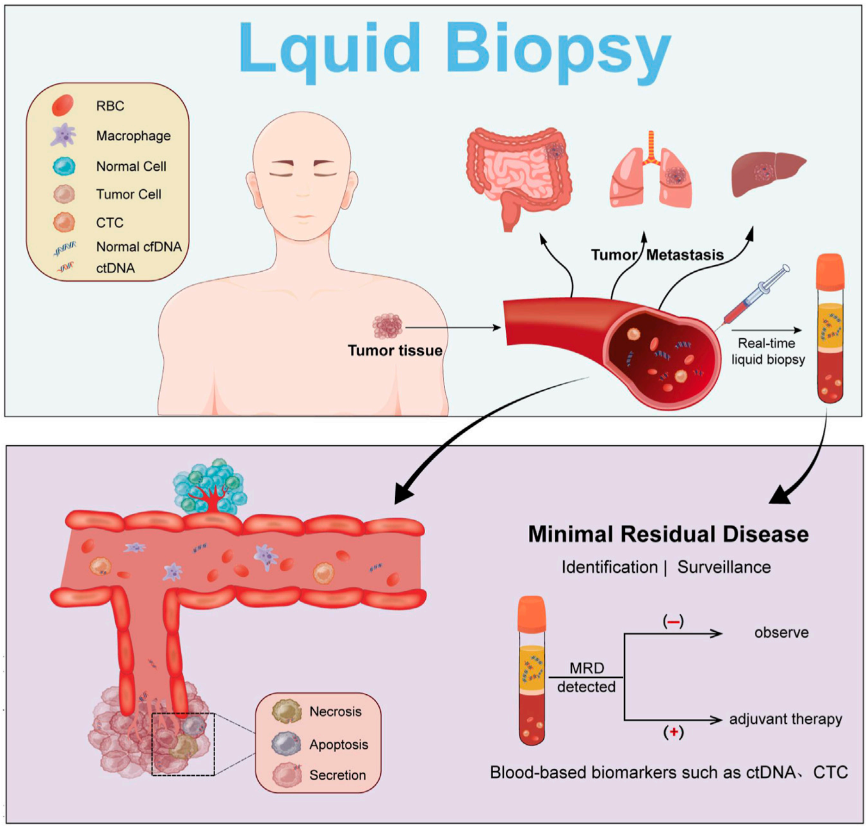


FIGURE 1
ctDNA during cancer progression. Detection of ctDNA is achieved by liquid biopsy, which allows monitoring and adjunctive treatment of MRD.

tissues damaged by ischemia, trauma, infection or inflammation can also contribute cfDNA (Schwarzenbach et al., 2011; Snyder et al., 2016). Circulating tumor DNA (ctDNA) is a rather minor fraction of cfDNA released by malignant tumors into the bloodstream or other bodily fluids (Diehl et al., 2008). ctDNA is shorter compared to cfDNA derived from non-cancer cells (Mouliere et al., 2011; Zheng et al., 2012; Underhill et al., 2016). ctDNA is generally more fragmented than non-mutant cfDNA, with a maximum enrichment between 90 and 150 bp compared with 250–320 bp (Bao et al., 2016; Mouliere et al., 2018; He et al., 2020a). ctDNA levels correlate with clinical and pathological features of cancer, including stage, tumor burden, localization, vascularization, and response to therapy (Diehl et al., 2008; Bettegowda et al., 2014; Heitzer et al., 2019). In addition, ctDNA levels vary according to tumor type, shedding rate, and other biological factors (Bettegowda et al., 2014; Siravegna et al., 2019).

MRD (Minimal residual disease) is defined as a small number of cancer cells that remain in the body after cancer treatment (those that do not respond to treatment or are resistant to drugs), which may ultimately lead to disease relapse. ctDNA tests can benefit patients with solid tumor for its capacity to confirm the existence of MRD during the postoperative period. On the other hand, MRD tests can monitor and assess the biomarkers that indicate the effectiveness of adjuvant chemotherapy as well as drug resistance (Guibert et al., 2020). (Figure 1)

2 ctDNA detection methods

The ctDNA detection methods are mainly divided into two main categories: the PCR-based detection methods such as Droplet Digital PCR (ddPCR), and the Next-Generation Sequenceing (NGS). These methods have significant differences in detection sensitivity, specificity and coverage.

2.1 Droplet digital PCR

Droplet Digital PCR (ddPCR) is to distribute the DNA sample into millions of water-oil emulsion droplets before the traditional PCR amplification, which means, each of the droplets either contains no gene under test or contains one gene. After PCR amplification, each microdroplet was detected one by one. The initial copy number or concentration of the DNA to be tested could be obtained according to the Poisson distribution principle and the number and proportion of positive microdroplet (Zonta et al., 2016; He et al., 2021). The ddPCR provides an absolute quantification that improves sensitivity at a low cost, and it can also achieve high specificity by designing the primers and probes individually.

The detection limit of ddPCR turns out to be about 0.1% (Dong et al., 2018; Corless et al., 2019; He et al., 2020b; Vessies et al., 2020). Reported LODs vary due to differences in the amount of ctDNA in plasma, sample quality, and analysis approaches. Although ddPCR is very effective for detecting small numbers of mutations identified from sequencing of tumor tissue or hot-spot mutations with a high prevalence, this detection approach has inferior clinical sensitivity for MRD than highly parallel NGS methods monitoring multiple mutations (Garcia-Murillas et al., 2015; Pietrasz et al., 2017; Schøler et al., 2017; Christensen et al., 2018).

The ddPCR method has the advantages of low cost and short detection time, but the disadvantage of detecting only known variants and analyzing only a limited number of variants localize its use in clinical practice as a supplement or conditional substitute for tissue biopsy for genotyping (Elazezy and Joosse, 2018; Wei et al., 2018; Franczak et al., 2019; Kerachian et al., 2019; Peng et al., 2022; Wang et al., 2022). Therefore, in most circumstances, ddPCR is not the preferred approach for solid tumor MRD detection.

2.2 PCR amplicon-based NGS

Next-generation sequencing (NGS) is a high-throughput technique that enables detection of billions of DNA molecules from a biological sample. Compared with ddPCR method, NGS can search for previously unidentified variations (Wan et al., 2017). With the emergence of more and more therapeutic related molecular targets, NGS has become increasingly important in cancer research. Although whole genome sequencing (WGS) or whole exome sequencing (WES) can provide more detailed genomic information, ctDNA NGS techniques for clinical application utilize amplicon-based NGS or hybridization capture-based NGS to provide clinically relevant information with higher sequencing depth at lower cost.

Amplicon-based NGS is one of the popular detection methods to identify specific ctDNA molecules. Gene-specific PCR amplicons are used to amplify the specific genomic regions originated from tumor-derived mutations before NGS is performed. Unique molecular identifiers (UMI) can help increasing sensitivity and specificity of NGS detection (Phallen et al., 2017; Goldberg et al., 2018). Forshew et al. first described Tagged-Amplicon Sequencing (TAm-Seq), demonstrating that cancer mutations with allele frequencies as low as 2% and sensitivity and specificity over 97% could be detected, and this technique was successfully applied to cancer mutation surveillance in patients with advanced ovarian cancer (Forshew et al., 2012). Later on, TAm-seq was used to apply NGS to a target panel which has detection limits as low as 0.01% (Gale et al., 2018). Although amplicon-based targeted NGS methods are highly sensitive and specific, amplification may potentially bias the observed mutant allele, and this technique is limited to the queried amplicon space while mutation detection performed (Chaudhuri et al., 2015; Wan et al., 2017; Abbosh et al., 2018).

2.3 Hybridization capture-based NGS

Hybrid capture-based NGS, which hybridize relevant DNA sequences to biotinylated probes before NGS is performed, is developed to improve the detection of multiple mutations of tumor with high sensitivity and without significant prior knowledge (Wan et al., 2017).

Newman et al. utilized an highly sensitive ctDNA detection technique named Cancer Personalized Profiling by Deep Sequencing (CAPP-Seq) (Newman et al., 2014a). Cancer WGS and WES data from databases such as The Cancer Genome Atlas (TCGA) and the Catalog of Somatic Mutations in Cancer (COSMIC) was used for bioinformatics analysis to identify recurrently mutated genomic regions in the population of a given cancer type. Biotinylated probes, designed according to these results, were applied to cfDNA capture before NGS of certain cancer patients in order to quantitate the ctDNA with a detection limit of 0.02%. UMI was used to reduce the effect of PCR errors and a bioinformatic error correction step called polishing was used to reduce the effect of stereotypical background artifacts (Newman et al., 2016). Recently ctDNA detection limit was improved to 1 part per million by utilizing multiple somatic mutations within individual DNA fragments to reduce the effect of both technical and biological errors (Kurtz et al., 2021).

The capability of CAPP-Seq includes detection of SNV, insertions/deletion (indel), and genomic rearrangements without individuation. Compared to amplicon-based NGS, Capp-seq shows more reliable detection of copy number changes and allows detection of fusion proteins (Gagan and Van Allen, 2015; Xiao et al., 2022). Otherwise, results from sequencing can reveal the mechanisms of carcinogenesis and drug resistance (Chabon et al., 2016; Khan et al., 2018).

2.4 Whole genome sequencing (WGS) and whole exome sequencing (WES)

NGS approaches have become prevalent for tumor sequencing and have also been applied to ctDNA detection. WGS applied to cfDNA achieves a sequencing depth of 0.1× and WES achieves a sequencing depth of 100× (Farris and Trimarchi, 2013; Heitzer et al., 2013; Murtaza et al., 2013; Cohen et al., 2017). Although some studies suggest that WGS is feasible for clinical application to certain patients, it is prohibitive for routine clinical implementation of WGS because of its cost and time required to perform WGS and the associated bioinformatic analysis (Welch et al., 2011; Chan et al., 2013). Therefore, WES turns out to be feasible to improve detection sensitivity and reduce cost while maintaining comprehensive coverage of likely mutated genomic regions. The exons are enriched for most of the pathogenic somatic mutations while they represent only 1.5% of the whole genome (Choi et al., 2009). Above all, there is an inverse correlation between sequencing breadth and detection cost, and sequencing depth *versus* detection limit of detection. Due to the low level of ctDNA in body fluids, targeted approaches, including hybridization capture-based NGS, PCR amplicon-based NGS, are superior to more extensive sequencing approaches such as whole exome or whole genome sequencing.

3 Application of ctDNA for MRD detection

ctDNA detection has shown promising clinical potential as a method to detect MRD in solid tumors after radical therapy and before clinical or radiographic disease recurrence (Chaudhuri et al., 2017). MRD status is closely associated with future radiological relapse and the detection of ctDNA after clinical treatment may improve the decision of the next therapeutic regimen. Most treatments are still based on strict chemotherapy regimens, although the probability of serious adverse effects is lower than in the past. Therefore, it is important to avoid unnecessary adjuvant chemotherapy when it can be established that patients may not benefit. ctDNA analysis showed that MRD was associated with poor prognosis in patients with malignant tumors. In this review, we focus on the significance of ctDNA analysis to guide the adjuvant therapy and predict the relapse of lung, breast and colon cancer.

3.1 Lung cancer

Among patients with non-metastatic lung cancer, some patients can be cured by primary surgical resection, radiotherapy and comprehensive treatment including chemotherapy (Kalemkerian et al., 2013; Ettinger et al., 2022). In fact, by the time recurrent or progressive lesions were detected by imaging tests after treatment, the patient's systemic tumor burden was significantly increased. Therefore, there is a great interest in whether MRD detection after radical resection of NSCLC can identify patients at risk of recurrence and provide personalized adjuvant therapy before the tumor burden increases (Kalemkerian et al., 2013; Chaudhuri et al., 2017; Chen et al., 2017; Chen et al., 2019a; Zhao et al., 2019; Peng et al., 2020; Ettinger et al., 2022).

The TRACERx study showed that MRD was predictive of recurrence before routine imaging and that more than 99% of MRD-negative patients did not relapse after treatment (Abbosh et al., 2017). The time interval between the increase in ctDNA levels after surgery and the clinical diagnosis of cancer recurrence provides an opportunity for clinical intervention.

Dynamic study prospectively revealed the dynamic changes of ctDNA in patients with primary lung cancer after surgery (Chen et al., 2019b). After tumor resection, ctDNA level decreased rapidly in patients with surgical lung cancer. The half-life of ctDNA after radical resection of lung cancer is only 35 min. They proposed that 3 days after R0 resection can be used as a baseline for postoperative monitoring of lung cancer.

Chaudhuri et al. (Chaudhuri et al., 2017) introduced their research utilizing CAPP-seq to detect ctDNA. After 36 months of MRD detection, 100% of the patients with detectable ctDNA had disease progression, while 93% of those without detectable ctDNA had no progression of cancer (HR = 43.4, $p < 0.001$). The long-term survival rate of patients without ctDNA detected in MRD was significantly higher than that of patients with ctDNA detected ($p < 0.001$). They suggested that both node-positive and node-negative patients with stage I to III NSCLC may benefit from personalized adjuvant therapy. Patients without tissue material may benefit from tyrosine kinase inhibitors (TKI) or immune checkpoint inhibitors (ICI) with assessing actiable mutations and mutational burden in ctDNA.

In the study of Kuang's, they detected tumor tissue-specific mutated ctDNA in preoperative plasma samples from 19 (50%) patients (Kuang et al., 2020), and preoperative ctDNA in plasma was consistent with that in tissue. Compared with patients with undetectable ctDNA after chemotherapy, the RFS of ctDNA-positive patients after chemotherapy was worse (HR = 8.68, p = 0.022). ctDNA-negative patients after chemotherapy had better long-term efficacy than patients with positive ctDNA after chemotherapy (HR = 4.76, p = 0.047).

Gale et al. reported their study using patient-specific assays with up to 48 amplicons targeting tumour-specific variants unique to each patient to monitor postoperative MRD (Gale et al., 2022). Of the 48 patients whose samples were collected 1–3 days after surgery, ctDNA was detected in 12 samples (25%), with a median eVAF of 0.0026%. Therefore, in the case of complete excision of the disease, ctDNA may be present transiently in the blood at low concentrations. ctDNA was detectable in 18/28 (64.3%) patients with clinical recurrence of primary tumors. ctDNA detection had clinical specificity >98.5% and preceded clinical detection of relapse of the primary tumour by a median of 212.5 days. They suggested that MRD detection may be best delayed beyond the first few days as well, because ctDNA was detectable during 1–3 days after surgery in 25% patients, but half of them did not have clinical relapse.

A recent study identified a potentially cured population of localized NSCLC by longitudinal MRD detection (Zhang et al., 2022). From 261 patients with stage I to III NSCLC who underwent definitive surgery, 913 peripheral blood samples were successfully detected by MRD assay. In the surveillance population, only 6 patients (3.2%) with longitudinally undetectable MRD relapsed, with a negative predictive value of 96.8%. The authors identified these patients with longitudinally undetectable MRD as potentially cured patients. The peak risk for detectable MRD was approximately 18 months after the landmark detected. The positive predictive value of longitudinal detectable MRD was 89.1%, and the median lead time was 3.4 months. MRD detection is not ideal for the monitoring of patients with only brain recurrence (n = 1/5, 20%). Further subgroup analysis showed that patients with undetectable MRD may not benefit from adjuvant therapy. In addition, the authors suggest that the risk of developing detectable MRD decreased progressively 18 months after the biomarker discovery.

In conclusion, MRD detection can identify patients at risk of recurrence earlier and is a practical prognostic factor after radical NSCLC surgery. Positive ctDNA after treatment may indicate the presence of MRD, which may be a signal suggesting a change in treatment regimen. After treatment, ctDNA can change from positive to negative, which means that surgery or adjuvant therapy can remove MRD, thereby changing disease progression and survival.

3.2 Breast cancer

Although tumor biopsy has long been the standard method for tumor detection, its limitations have made minimally invasive and relatively inexpensive liquid biopsy an alternative. For patients with early-stage breast cancer, ctDNA testing can monitor tumor burden and treatment response, so as to guide therapeutic regimen selection.

Riva et al. described their study that massively parallel sequencing (MPS) was performed on patients with nonmetastatic triple-negative breast cancer (TNBC) and droplet digital PCR (ddPCR) was used to monitor TP53 mutations expressed in tumor tissues (Riva et al., 2017). Patients were treated with neoadjuvant chemotherapy prior to surgery, ctDNA levels decreased rapidly during NCT and no MRD was detected postoperatively. The slow decline in ctDNA levels during NCT is closely associated with shorter survival.

Another study used whole exome sequencing to detect mutations in tumor tissue (Parsons et al., 2020). They then performed an individualized MRD assay to detect mutations in cfDNA. This approach was 100-fold more sensitive than ddPCR when tracking individual mutation. MRD detection at 1 year was strongly associated with distant recurrence (HR = 20.8; 95% confidence interval, 7.3–58.9). The median lead time from first detectable ctDNA to clinical recurrence was 18.9 months.

In the neoadjuvant I-SPY 2 TRIAL, cfDNA was isolated from 291 plasma samples of 84 high-risk early breast cancer patients (Magbanua et al., 2021). 16 patient-specific mutations were identified by whole exome sequencing of pretreated tumors, and then ultra-deep sequencing of cfDNA from patients was performed with this personalized ctDNA detection panel. Patients with positive ctDNA after 3 weeks of neoadjuvant chemotherapy had a significantly lower probability of pathological complete response (pCR) after treatment than patients with negative ctDNA (odds ratio 4.33, $p = 0.012$). All patients who achieved pCR were ctDNA negative after neoadjuvant chemotherapy (n = 17, 100%). While ctDNA-positive patients (14%) who failed to achieve pCR (n = 43) showed a significantly high risk of metastatic relapse [HR 10.4; 95% CI 2.3–46.6]. 86% of those who did not achieve pCR and had negative ctDNA had a favorable prognosis. The author suggested that even in patients who did not achieve pCR, insufficient ctDNA clearance was an important predictor of poor treatment response and metastatic tumor recurrence, and clearance was associated with improved survival.

In advanced or metastatic tumors, ctDNA has high clinical value and development prospects because of its relatively high detection rate (Diehl et al., 2008; Tie et al., 2015; Jiang et al., 2022). Recently, Liu et al. introduced their research of metastatic breast cancer (Liu et al., 2022a). They established a novel ctDNA-level Response Evaluation Criterion in Solid Tumors (ctle-RECIST) to assess treatment response and predict progression-free survival (PFS) based on ctDNA alteration levels and variant allele frequency (VAF). By monitoring and analyzing the ctDNA of 223 patients with metastatic breast cancer at different time points before and after treatment, the results showed that the median PFS of patients without ctDNA changes was significantly longer than that of patients with ctDNA changes (6.63 vs 4.9–5.7 months, $p < 0.05$). In addition, they found that ctDNA detection may be a good complement to radiological assessment, due to the median PFS of double DCR group tended to be longer than that of single DCR group [HR 1.41 (0.93–2.13), $p = 0.107$].

In the treatment of breast cancer patients, PARP inhibitors are synthetically lethal to TNBC tumors carrying BRCA1/2 aberrations by impairing DNA repair mechanisms (Helleday et al., 2005). Genomic alterations detected by longitudinal plasma sampling can identify genes that are resistant to PARP inhibitors such as olaparib and velipariib. Mutations in the TP53 and PIK3CA gene in ctDNA have been sensitive and specific circulating blood biomarkers (Dawson et al., 2013). In addition, ESR1-mutated ctDNA has also been identified as a predictive marker of response to aromatase inhibitor therapy (Guttery et al., 2015; Schiavon et al., 2015). These studies suggest that ctDNA detection can be used to track molecular alterations in patients before and after treatment to develop personalized targeted therapies.

3.3 Colorectal cancer

Compared with the lack of sufficient tumor tissue in the specimen and the need for a long test cycle in the tissue biopsy, the utilization of liquid biopsy to detect ctDNA is expected to become an effective tool to promote precision medicine.

For stage II CRC patients, most of them did not receive postoperative chemotherapy. MRD detection is needed to identify 10%–15% of those patients who still have residual lesions after surgery (Osterman and Glimelius, 2018). Postoperative chemotherapy may help reducing the risk of relapse for those who have positive ctDNA. For stage III CRC patients, 30% of them had clinical recurrence after receiving postoperative chemotherapy (Osterman and Glimelius, 2018). At the same time, most patients with stage III colorectal cancer receive postoperative chemotherapy, although more than 50% of patients are cured by surgery (Böckelman et al., 2015; Påhlman et al., 2016; Babaei et al., 2018). Therefore, MRD detection is one potential approach to address the problem of how to better identify patients who could benefit from postoperative adjuvant therapy.

In a previous study, 40% of patients with stage II colorectal cancer who received 6 months of conventional adjuvant chemotherapy had an absolute risk reduction of only 3%–5%, despite the risk associated with potentially serious adverse events and without means to monitor the efficacy of adjuvant therapy (Wirtzfeld et al., 2009). In another study of patients with stage III colorectal cancer, at least one somatic mutation was identified in tumor tissue from all 96 evaluable patients (Tie et al., 2019). ctDNA was detectable in 20 of 96 (21%) postoperative samples and was associated with poor recurrence-free survival (HR, 3.8; 95% CI, 2.4–21.0; $p < 0.001$). For patients received chemotherapy, 15 of 88 (17%) samples were ctDNA positive, with a 30% estimated 3-year RFI. While for those ctDNA undetectable, the 3-year RFI was 77% (HR, 6.8; 95% CI, 11.0–157.0; $p < 0.001$). The author found out that postoperative ctDNA status was independently associated with RFI and significantly outperformed standard clinicopathologic characteristics as a prognostic marker. They later utilized meta-analysis to summarize their previous studies and concluded that the 5-year recurrence-free rate and overall survival rate of patients with non-metastatic CRC who had detectable ctDNA after surgery were poorer (Tie et al., 2016; Tie et al., 2019; Tie et al., 2019; Tie et al., 2021). In this meta-analysis, they combined individual patient data from three independent cohort studies of non-metastatic colorectal cancer (CRC). A massively parallel sequencing platform SafeSeqS was used to analyze ctDNA from 485 CRC patients. ctDNA was detected in 59 (12%) patients postoperatively and the risk of recurrence increases exponentially with increasing ctDNA mutation allele frequency (MAF) (HR, 1.2, 2.5 and 5.8 for MAF of 0.1%, 0.5% and 1%). ctDNA was detected in 3 of 20 patients (15%) with local regional recurrence and 27 of 60 patients (45%) with distant recurrence ($p = 0.018$). This also implies that ctDNA is a better predictor of distant recurrence than local regional recurrence.

An observational GALAXY study recently analyzed MRD in patients with stage I-IV colorectal cancer (Taniguchi et al., 2021; Kotaka et al., 2022). Within the 188 MRD-positive patients, 95 received postoperative adjuvant chemotherapy. ctDNA levels decreased at a significantly faster rate in patients who received adjuvant chemotherapy than in those who did not receive adjuvant chemotherapy (68% vs. 7%; HR: 17.1; $p < 0.001$). Furthermore, patients received adjuvant chemotherapy had significantly longer 6-month DFS than those who did not (84% vs. 34%; HR: 0.15; $p < 0.001$).

In a recent study, Liu et al. used a technique that allows multiple tests of one single cfDNA sample using different methods (Liu et al., 2022b). They detected MRD using 3 approaches for each sample: personalized detection targeting tumour-informed mutations, universal panel for genes frequently mutated in colorectal cancer (CRC), and low depth sequencing for copy number alterations (CNAs). MRD positivity on personalized detection after neoadjuvant therapy was significantly associated with an increased risk of recurrence (HR = 27.38; $p < 0.0001$). Post-nat universal Panel was good at predicting recurrence in patients with high clinical risk, but not in patients with low clinical risk. CNAs analysis also showed a compromised performance in predicting recurrence.

Current methods for monitoring disease status in patients with metastatic colorectal cancer include radiographic imaging techniques and detection of serum CEA levels. However, serum CEA levels may be increased in only 70%–80% of patients (Goldstein and Mitchell, 2005).

In a study of patients with metastatic colorectal cancer (Garlan et al., 2017), ≥80% ctDNA clearance after first-line or second-line chemotherapy was associated with significantly improved objective response rates (47.1% vs. 0%; $p = 0.003$) and longer median PFS (8.5 months vs. 2.4 months; HR 0.19, 95% CI 0.09–0.40; $p < 0.0001$) and OS (27.1 months vs. 11.2 months; HR 0.25, 95% CI 0.11–0.57; $p < 0.001$). In another study, the authors used amplicon based deep sequencing to detect ctDNA in mCRC patients (Osumi et al., 2019). Patients with lower ctDNA levels (≤50%) showed significantly longer PFS and OS than patients with higher ctDNA levels (>50%) 8 weeks after initiation of chemotherapy.

In patients with stage II and III CRC, based on current studies, it has been demonstrated that ctDNA may be a useful prognostic marker after surgery to guide initial adjuvant therapy and monitor postoperative recurrence. ctDNA analysis can potentially transform the postoperative management of CRC by enabling risk stratification, chemotherapy monitoring, and early recurrence detection.

3.4 Other tumors

In recent studies, ctDNA has emerged as a potential biomarker for minimal residual disease (MRD) after treatment of many solid tumors (Lou et al., 2018; Xiong et al., 2019; He et al., 2020c; Chen et al., 2020; Xu et al., 2020). In a study of patients with locally advanced unresectable or metastatic gastric cancer, patients with low ctDNA levels significantly prolonged DFS after the first cycle of chemotherapy (3 months) compared with patients with high ctDNA levels (COX regression $p = 0.0228$) (Normando et al., 2018). In another study, advanced gastric cancer patients with higher ctDNA levels were more likely to have peritoneal recurrence and significantly lower 5-year overall survival rate than patients with lower ctDNA levels (39.2% vs 45.8%, $p = 0.039$) (Fang et al., 2016). Carrying ctDNA mutations was associated with poor prognosis among patients with late stage gastric cancer. In a study of gastrointestinal malignancies, ctDNA levels were higher in the gastrointestinal tumor group than in the carcinoma *in situ* group and healthy controls ($p = 0.019$) (Lan et al., 2017). For recurrent gastric cancer, persistent high levels of ctDNA and an increasing trend were observed after surgery (Wen et al., 2015). In addition, ctDNA levels tended to be more sensitive than CEA levels in predicting recurrence during postoperative monitoring.

In a study of metastatic gastroesophageal cancer, ctDNA was detectable in plasma before treatment in 75% of 72 patients and correlated well with mutations on metastatic biopsy (86% agreement) (van Velzen et al., 2022). The detection of multiple mutations in baseline plasma ctDNA was associated with poorer overall survival (OS, HR 2.16, 95% CI 1.10–4.28; $p = 0.027$) and PFS (PFS, HR 2.71, 95% CI 1.28–5.73; $p = 0.009$), and the VAF was associated with baseline tumor volume (Pearson's R 0.53, $p < 0.0001$). In addition, patients with residual ctDNA detected after 9 weeks of treatment had worse OS and PFS (OS: HR 4.95, 95% CI 1.53–16.04; $p = 0.008$; PFS: HR 4.08, 95% CI 1.31–12.75; $p = 0.016$).

A large proportion of the patients with early and intermediate stage liver cancer after surgery will have recurrence. In a recent study, peripheral blood samples were collected from all patients after surgery and analyzed by next-generation sequencing based on hybrid capture (Ye et al., 2022). The recurrence rates of ctDNA positive group and ctDNA negative group were 60.9% and 27.8%, respectively. Multivariate Cox regression analysis showed that postoperative ctDNA was an independent prognostic factor for DFS (HR: 6.074, 95% Cl: 2.648–13.929, $p < 0.001$) and OS (HR: 4.829, 95% CI: 1.508–15.466, $p = 0.008$). The prognosis of patients with negative ctDNA was better than that of patients with positive ctDNA regardless of tumor stage. In addition, the authors suggested that the combination of ctDNA and AFP detection could improve the prediction performance.

The value of ctDNA in predicting early postoperative tumor recurrence and monitoring tumor burden in patients with hepatocellular carcinoma (HCC) was investigated in another prospective study (Zhu et al., 2022). They utilized NGS to analyze the ctDNA sequences before and after surgery, and whole exome sequencing was used to detect the DNA of HCC and adjacent tissues. During a median follow-up of 17.7 months, 9 patients (22%) experienced cancer relapse. The positive rate of ctDNA in the non-recurrence group was significantly lower than that in the recurrence group, and ctDNA positivity was associated with significantly shorter recurrence-free survival (RFS). The author suggested that median VAF of baseline ctDNA was an independent predictor of RFS in HCC patients.

Pancreatic cancer is an aggressive solid tumor with a poor prognosis. Currently used biomarkers that are often used to identify advanced pancreatic cancer also do not indicate prognosis. A recent study used hybrid capture-based NGS to sequence ctDNA in patients with metastatic pancreatic cancer (Guan et al., 2022). In 40 tumor tissue samples, mutations in KRAS (87.5%, N = 35) and TP53 (77.5%, N = 31) were more common, and ≥3 mutations in driver genes were strongly associated with overall survival (OS). Univariate analysis showed a significant association between CDKN2A or SMAD4 mutation in ctDNA and PFS in 35 blood samples. Cox hazard proportion model showed that CDKN2A mutation in ctDNA (HR = 16.1, 95% CI = 4.4–59.1, $p <$ 0.001) were significantly associated with OS. Patients' CDKN2A mutation in ctDNA (HR = 6.8, 95% CI = 2.3–19.9, $p = 0.001$) and SMAD4 mutation (HR = 3.0, 95% CI = 1.1–7.9, $p = 0.031$) were significantly associated with PFS. Disease progression detected by ctDNA was 0.9 months earlier than radiological imaging (mean PFS: 4.6 m vs. 5.5m, $p = 0.004$).

In another research of patients with borderline resectable pancreatic cancer, no significant decrease in median RFS or OS was observed in ctDNA-positive patients before treatment or after NAC (Kitahata et al., 2022). The median OS of patients (723 days) with positive ctDNA was significantly shorter than that of patients with negative ctDNA (not reached; $p = 0.0148$). The hazard ratio for adjusted survival risk increased from 4.13 times to 17.71 times for patients with a risk factor (detectable ctDNA or CA19-9>37 U/ml) compared with patients without risk factors (both $p = 0.0055$).

Perioperative systemic chemotherapy can improve the prognosis of upper tract urothelial carcinoma (UTUC). A recent study utilized NGS to analyze perioperative ctDNA to identify patients with poor prognosis who require perioperative chemotherapy (Nakano et al., 2022). They performed targeted ultra-deep sequencing of plasma free DNA (cfDNA) and albugemma DNA, as well as whole-exome sequencing of cancer tissue, thereby eliminating possible false positives. ctDNA was positive in 23 of 50 untreated UTUC patients (46%) and in 17 of 43 localized UTUC patients (40%). Among preoperative risk factors, only preoperative ctDNA score >2% was a significant and independent risk factor associated with poor recurrence-free survival (RFS). In addition, the presence of ctDNA early after surgery was significantly associated with poor RFS, suggesting the presence of MRD.

In another study of urothelial carcinoma, the authors improved the performance of the prognostic model by combining ctDNA sequence aggregate VAF (aVAF) values with clinical factors, including age, sex, and liver metastases (KyrillusShohdy et al., 2022). In consecutive ctDNA samples, an increase in ctDNA aVAF of ≥1 predicted disease progression within 6 months in 90% of patients. The majority of patients with aVAFs≤0.7 in three consecutive ctDNA samples achieved a durable clinical response (≥6 months).

4 Challenges of MRD detection

When we perform MRD detection, the number of specific variants we focused on was very small because the total number of gene copies in the plasma samples was limited. As we all know, MRD detection often requires a high sequencing depth, the sensitivity of ctDNA analysis is limited, and when VAF lowers close to LOD, the number of specific variants in the sample may be demanding. In addition, the tumor fraction of cfDNA varies between cancer entities and even between patients with the same cancer entity (Bachet et al., 2018; Normanno et al., 2018; Jiang and Yan, 2021; Huo et al., 2022). In some ctDNA-based studies, it has been found that tumor micrometastases represent a higher tumor burden than residual local disease, and therefore can shed higher ctDNA levels (Azad et al., 2020; Tie et al., 2021). Therefore, some false-negative results cannot be prevented due to biological factors such as low DNA shedding in some tumors or the location of the metastasis itself. The sensitivity of different types of mutations is also different. The ability of different techniques to detect single nucleotide mutations differs from that of structural variants (e.g., fusion) or copy number variants (e.g., copy number amplification). ctDNA analysis is less sensitive to detect structural variants or copy number variants. Therefore, the interpretation of ctDNA results needs to take into account that the amount of ctDNA may not be sufficient to detect specific types of variation. Pascual J, Attard G, Bidard FC, Curigliano G, De Mattos-Arruda L, Diehn M, Italiano A, Lindberg J, Merker JD, Montagut C, Normanno N, Pantel K, Pentheroudakis G, Popat S, Reis-Filho JS, Tie J, Seoane J, Tarazona N, Yoshino T, Turner NC. ESMO recommendations on the use of circulating tumour DNA assays for patients with cancer: a report from the ESMO Precision Medicine Working Group. Ann Oncol. 2022 August; 33(8):750–768. doi: 10.1016/j.annonc.2022.05.520. Epub 2022 July 6. PMID: 35809752.

DNA fragments from the clonal hematopoiesis of indeterminate potential (CHIP) or non-neoplastic hematopoietic stem cells can lead to false-positive ctDNA results, which can be reduced by advanced bioinformatics analysis or comparison of ctDNA sequencing with leukocytes and/or matched tumor tissue (Steensma et al., 2015; Snyder et al., 2016). These mutations represent a confounding factor when analyzing actual tumor variants in the absence of white blood cell (WBC) control samples (Razavi et al., 2019). Therefore, additional NGS analysis of leukocytes is recommended to rule out CHIP-related variants, especially in the case of MRD or early cancer detection.

Agreement between ctDNA and tissue-based NGS results is typically defined as the presence or absence of identical genomic alterations in a single gene on both molecular platforms. The main reasons for inconsistent blood and tissue detection are biopsy location and time, different DNA shedding, tumor heterogeneity, and epigenetic modifications. Lack of standardization between ctDNA tests is another barrier, which limits the understanding of the available results. Inconsistent ctDNA results may be the result of several variables, including the time of sample collection, sample collection process, sample storage method, library construction process, utilization of unique molecular identifiers and bioinformatic analysis.

Accurate risk assessment and adjuvant therapy are very important for cancer patients. ctDNA testing can accurately identify the MRD after primary tumor resection, and thus identify the patient population that needs further adjuvant chemotherapy, so as to avoid unnecessary additional treatment. In addition, determining the duration of adjuvant therapy based on ctDNA clearance can help reduce adverse reactions. However, many researchers also suggested that adjuvant therapy based on negative ctDNA testing should not be excluded due to the low standardization of ctDNA detection procedures and the limitations of ctDNA testing techniques.

Although preliminary data on the clinical application of ctDNA in MRD detection is promising, most of the studies that provide evidence to support it are small, limited in scope and require validation using large independent cohorts (Corcoran and Chabner, 2018; Heitzer et al., 2019). It is only through these further studies that we can solve the next important question of whether acting on positive ctDNA MRD results can improve clinical outcomes or whether ctDNA MRD can be used to more precisely guide adjuvant therapy.

5 Conclusion

Overall, MRD aids in the management of cancer at all stages, including screening, guiding adjuvant treatment, predicting relapse early, initiating systemic treatment and monitoring response, and genotyping resistance. Liquid biopsy, espesially ctDNA, can be used as an alternative to tumor tissue detection, especially when tissue biopsy is not feasible or time does not permit. New technologies are being developed, such as methylation pattern-based sequencing which have the potential to optimize ctDNA detection for use in a wide range of scenarios. In the future, we need to carry out more intervention studies to provide stronger evidence support for the application of MRD detection methods, so as to achieve the purpose of integration with clinical routine. Through the monitoring of ctDNA, the therapeutic regimen can be adjusted in time, and the treatment effect can be improved to maximize the survival time of patients.

Author contributions

LZ, WS, and PB contributed to conception and design of the study. LZ organized the database. YZ performed the statistical analysis. RX wrote the first draft of the manuscript. XL, JZ, JZ, and PB wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

The work was supported by the Hunan Provincial Education Commission Foundation (19B068,20A056,20C0202, 22C0669); The Hunan Provincial Health Commission Foundation (No.202112041226, D202302088596); the Innovation and Entrepreneurship Education Base of Public Health and Preventive Medicine (Hunan Education Bureau Notice 2019 No.333-93); and the Funding by young backbone teachers of Hunan province training program foundation of Changsha Medical University (Hunan Education Bureau Notice 2021 No.29-26).

Conflict of interest

Authors LZ, YZ, JL were employed by Geneis Beijing Co.

The remaining 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.

References

Abbosh, C., Birkbak, N. J., and Swanton, C. (2018). Early stage NSCLC: challenges to implementing ctDNA-based screening and MRD detection. *Nat. Rev. Clin. Oncol.* 15 (9), 577–586. doi:10.1038/s41571-018-0058-3

Abbosh, C., Birkbak, N. J., Wilson, G. A., Jamal-Hanjani, M., Constantin, T., Salari, R., et al. (2017). Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* 545 (7655), 446–451. doi:10.1038/nature22364

Azad, T. D., Chaudhuri, A. A., Fang, P., Qiao, Y., Esfahani, M. S., Chabon, J. J., et al. (2020). Circulating tumor DNA analysis for detection of minimal residual disease after chemoradiotherapy for localized esophageal cancer. *Gastroenterology* 158, 494–505. doi:10.1053/j.gastro.2019.10.039

Babaei, B., Jansen, L., Erningvan, E., Vaes, S., Glimelius, U., Jansen, L., van Erning, F. N., et al. (2018). Administration of adjuvant chemotherapy for stage II-III colon cancer patients: an European population-based study. *Int. J. Cancer* 142, 1480–1489. doi:10.1002/ijc.31168

Bachet, J., Bouché, O., Laurent-Puig, P., Dubreuil, O., Garcia, M. L., Meurisse, A., et al. (2018). RAS mutation analysis in circulating tumor DNA from patients with metastatic colorectal cancer: the AGEO RASANC prospective multicenter study. *Ann. Oncol.* 29, 1211–1219. doi:10.1093/annonc/mdy061

Bao, M. H., Luo, H. Q., Chen, L. H., Tang, L., Ma, K. F., Xiang, J., et al. (2016). Impact of high fat diet on long non-coding RNAs and messenger RNAs expression in the aortas of ApoE(-/-) mice. *Sci. Rep.* 6, 34161. doi:10.1038/srep34161

Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., et al. (2014). Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* 6, 224ra24. doi:10.1126/scitranslmed.3007094

Böckelman, C., Engelmann, B. E., Kaprio, T., Kaprio, T., and Glimelius, B. (2015). Risk of recurrence in patients with colon cancer stage II and III: a systematic review and meta-analysis of recent literature. *Acta Oncol.* 54, 5–16. doi:10.3109/0284186X.2014.975839

Chabon, J. J., Simmons, A. D., Lovejoy, A. F., Esfahani, M. S., Newman, A. M., Haringsma, H. J., et al. (2016). Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. *Nat. Commun.* 7, 11815–11914. doi:10.1038/ncomms11815

Chan, K. C. A., Jiang, P., Zheng, Y. W. L., Liao, G. J. W., Sun, H., Wong, J., et al. (2013). Cancer genome scanning in plasma: detection of tumorassociated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin. Chem.* 59, 211–224. doi:10.1373/clinchem.2012.196014

Chaudhuri, A. A., Binkley, M. S., Osmundson, E. C., Alizadeh, A. A., and Diehn, M. (2015). Predicting radiotherapy responses and treatment outcomes through analysis of circulating tumor DNA. *Semin. Radiat. Oncol.* 25, 305–312. doi:10.1016/j.semradonc.2015.05.001

Chaudhuri, A. A., Chabon, J. J., Lovejoy, A. F., Newman, A. M., Stehr, H., Azad, T. D., et al. (2017). Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. *Cancer Discov.* 7, 1394–1403. doi:10.1158/2159-8290.CD-17-0716

Chen, K., Zhao, H., Shi, Y., Yang, F., Wang, L. T., Kang, G., et al. (2019). Perioperative dynamic changes in circulating tumor DNA in patients with lung cancer (DYNAMIC). *Clin. Cancer Res.* 25 (23), 7058–7067. doi:10.1158/1078-0432.CCR-19-1213

Chen, S. J., Li, Y. Z., Zhi, S., Ding, Z., Huang, Y., Wang, W., et al. (2020). lncRNA xist regulates osteoblast differentiation by sponging miR-19a-3p in aging-induced osteoporosis. *Aging Dis.* 11 (5), 1058–1068. doi:10.14336/AD.2019.0724

Chen, S. Y., Zhao, Y., Shen, F., Long, D., Yu, T., and Lin, X. (2019). Introduction of exogenous wild-type p53 mediates the regulation of oncoprotein 18/stathmin signaling via nuclear factor-B in non-small cell lung cancer NCI-H1299 cells. *Onncol Rep.* 41 (3), 2051–2059. doi:10.3892/or.2019.6964

Chen, X., Liao, Y., Long, D., Yu, T., Shen, F., and Lin, X. (2017). The Cdc2/Cdk1 inhibitor, purvalanol A, enhances the cytotoxic effects of taxol through Op18/stathmin in non-small cell lung cancer cells *in vitro*. *Int. J. Mol. Med.* 40 (1), 235–242. doi:10.3892/ijmm.2017.2989

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19096–19101. doi:10.1073/pnas.0910672106

Christensen, E., Nordentoft, I., Vang, S., Birkenkamp-Demtröder, K., Jensen, J. B., Agerbæk, M., et al. (2018). Optimized targeted sequencing of cell-free plasma DNA from bladder cancer patients. *Sci. Rep.* 8, 1917. doi:10.1038/s41598-018-20282-8

Cohen, J. D., Javed, A. A., Thoburn, C., Wong, F., Tie, J., Gibbs, P., et al. (2017). Combined circulating tumor DNA and protein biomarkerbased liquid biopsy for the earlier detection of pancreatic cancers. *Proc. Natl. Acad. Sci. U. S. A.* 114, 10202–10207. doi:10.1073/pnas.1704961114

Corcoran, R. B., and Chabner, B. A. (2018). Application of cell-free DNA analysis to cancer treatment. *N. Engl. J. Med.* 379, 1754–1765. doi:10.1056/NEJMra1706174

Corless, B. C., Chang, G. A., Cooper, S., Syeda, M. M., Shao, Y., Osman, I., et al. (2019). Development of novel mutation-specific droplet digital PCR assays detecting TERT promoter mutations in tumor and plasma samples. *J. Mol. Diagn* 21, 274–285. doi:10.1016/j.jmoldx.2018.09.003

Dawson, S. J., Tsui, D. W., Murtaza, M., Biggs, H., Rueda, O. M., Chin, S. F., et al. (2013). Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* 368, 1199–1209. doi:10.1056/NEJMoa1213261

Diehl, F., Schmidt, K., Choti, M. A., Romans, K., Goodman, S., Li, M., et al. (2008). Circulating mutant DNA to assess tumor dynamics. *Nat. Med.* 14, 985–990. doi:10.1038/nm.1789

Dong, L., Wang, S., Fu, B., and Wang, J. (2018). Evaluation of droplet digital PCR and next generation sequencing for characterizing DNA reference material for KRAS mutation detection. *Sci. Rep.* 8, 9650. Nature Publishing Group. doi:10.1038/s41598-018-27368-3

Elazezy, M., and Joosse, S. A. (2018). Techniques of using circulating tumor DNA as a liquid biopsy component in cancer management. *Comput. Struct. Biotechnol. J.* 16, 370–378. doi:10.1016/j.csbj.2018.10.002

Ettinger, D. S., Wood, D. E., Aisner, D. L., Akerley, W., Bauman, J. R., Bharat, A., et al. (2022). Non-small cell lung cancer, version 3.2022, NCCN clinical practice guidelines in oncology. *J. Natl. Compr. Canc Netw.* 20 (5), 497–530. doi:10.6004/jnccn.2022.0025

Fang, W. L., Lan, Y. T., Huang, K. H., Liu, C. A., Hung, Y. P., et al. (2016). Clinical significance of circulating plasma DNA in gastric cancer. *Int. J. Cancer* 138, 2974–2983. doi:10.1002/ijc.30018

Farris, C., and Trimarchi, J. M. (2013). Plasma-seq: a novel strategy for metastatic prostate cancer analysis. *Genome Med.* 5, 35. doi:10.1186/gm439

Forshew, T., Murtaza, M., Parkinson, C., Gale, D., Tsui, D. W. Y., Kaper, F., et al. (2012). Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* 4, 136ra68. doi:10.1126/scitranslmed.3003726

Franczak, C., Filhine-Tresarrieu, P., Gilson, P., Merlin, J. L., Au, L., and Harle, A. (2019). Technical considerations for circulating tumor DNA detection in oncology. *Expert Rev. Mol. Diagn* 19 (2), 121–135. doi:10.1080/14737159.2019.1568873

Gagan, J., and Van Allen, E. M. (2015). Next-generation sequencing to guide cancer therapy. *Genome Med.* 7, 80–10. doi:10.1186/s13073-015-0203-x

Gale, D., Heider, K., Ruiz-Valdepenas, A., Hackinger, S., Perry, M., Marsico, G., et al. (2022). Residual ctDNA after treatment predicts early relapse in patients with early-stage non-small cell lung cancer. *Ann. Oncol.* 33 (5), 500–510. doi:10.1016/j.annonc.2022.02.007

Gale, D., Lawson, ARJ., Howarth, K., Madi, M., Durham, B., Smalley, S., et al. (2018). Development of a highly sensitive liquid biopsy platform to detect clinically-relevant cancer mutations at low allele fractions in cell-free DNA. *PLoS ONE* 13, e0194630. doi:10.1371/journal.pone.0194630

Gao, Y. X., Liu, Y. J., Liu, Y. H., Peng, Y., Yuan, B., Fu, Y., et al. (2021). UHRF1 promotes androgen receptor-regulated CDC6 transcription and anti-androgen receptor drug resistance in prostate cancer through KDM4C-Mediated chromatin modifications. *Cancer Lett.* 520, 172–183. doi:10.1016/j.canlet.2021.07.012

Garcia-Murillas, I., Schiavon, G., Weigelt, B., Ng, C., Hrebien, S., Cutts, R. J., et al. (2015). Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci. Transl. Med.* 7, 302ra133. doi:10.1126/scitranslmed.aab0021

Garlan, F., Laurent-Puig, P., Sefrioui, D., Siauve, N., Didelot, A., Sarafan-Vasseur, N., et al. (2017). Early evaluation of circulating tumor DNA as marker of therapeutic efficacy in metastatic colorectal cancer patients (PLACOL study). *Clin. Cancer Res.* 23, 5416–5425. doi:10.1158/1078-0432.CCR-16-3155

Goldberg, S. B., Narayan, A., Kole, A.J., Decker, R. H., Teysir, J., Lee, A., et al. (2018). Early assessment of lung cancer immunotherapy response via circulating tumor DNA. *Clin. Cancer Res.* 24, 1872–1880. doi:10.1158/1078-0432.CCR-17-1341

Goldstein, M. J., and Mitchell, E. P. (2005). Carcinoembryonic antigen in the staging and follow-up of patients with colorectal cancer. *Cancer Investig.* 23, 338–351. doi:10.1081/cnv-58878

Guan, S., Deng, G., Sun, J., Han, Q., Lv, Y., Xue, T., et al. (2022). Evaluation of circulating tumor DNA as a prognostic biomarker for metastatic pancreatic adenocarcinoma. *Front. Oncol.* 12, 926260. doi:10.3389/fonc.2022.926260

Guibert, N., Pradines, A., Favre, G., and Mazieres, J. (2020). Current and future applications of liquid biopsy in nonsmall cell lung cancer from early to advanced stages. *Eur. Respir. Rev.* 29 (155), 190052. doi:10.1183/16000617.0052-2019

Guttery, D. S., Page, K., Hills, A., Woodley, L., Marchese, S. D., Rghebi, B., et al. (2015). Noninvasive detection of activating estrogen receptor 1 (ESR1) mutations in estrogen receptor-positive metastatic breast cancer. *Clin. Chem.* 61, 974–982. doi:10.1373/clinchem.2015.238717

He, B. S., Dai, C., Lang, J. D., Bing, P., Tian, G., Wang, B., et al. (2020). A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochim. Biophys. Acta Mol. Basis Dis.* 1866 (11), 165916. doi:10.1016/j.bbadis.2020.165916

He, B. S., Hou, F. X., Ren, C. J., Bing, P., and Xiao, X. (2021). A review of current in silico methods for repositioning drugs and chemical compounds. *Front. Oncol.* 11, 711225. doi:10.3389/fonc.2021.711225

He, B. S., Lang, J. D., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020). TOOme: a novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. *Front. Bioeng. Biotechnol.* 8, 394. doi:10.3389/fbioe.2020.00394

He, B. S., Zhang, Y. X., Zhou, Z., Wang, B., Liang, Y., Lang, J., et al. (2020). A neural network framework for predicting the tissue-of-origin of 15 common cancer types based on RNA-seq data. *Front. Bioeng. Biotechnol.* 8, 737. doi:10.3389/fbioe.2020.00737

Heitzer, E., Ulz, P., Belic, J., Gutschi, S., Quehenberger, F., Fischereder, K., et al. (2013). Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome Med.* 5, 30–16. doi:10.1186/gm434

Heitzer, E., Haque, I. S., Roberts, CES., and Roberts, C. E. S. (2019). Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat. Rev. Genet.* 20, 71–88. doi:10.1038/s41576-018-0071-5

Helleday, T., Bryant, H. E., and Schultz, N. (2005). Poly (ADP-ribose) polymerase (PARP-1) in homologous recombination and as a target for cancer therapy. *Cell Cycle* 4, 1176–1178. doi:10.4161/cc.4.9.2031

Huo, R., Liu, Y., Xu, H. M., Li, J., Xin, R., Xing, Z., et al. (2022). Associations between carotid atherosclerotic plaque characteristics determined by magnetic resonance imaging and improvement of cognition in patients undergoing carotid endarterectomy. *Quant. Imaging Med. Surg.* 12 (5), 2891–2903. doi:10.21037/qims-21-981

Jiang, X. W., and Yan, M. (2021). Comparing the impact on the prognosis of acute myocardial infarction critical patients of using midazolam, propofol, and dexmedetomidine for sedation. *BMC Cardiovasc Disord.* 21 (1), 584–2021. doi:10.1186/s12872-021-02385-9

Jiang, Z. R., Yang, L. H., Jin, L. Z., Yi, L. M., Bing, P. P., Zhou, J., et al. (2022). Identification of novel cuproptosis-related lncRNA signatures to predict the prognosis and immune microenvironment of breast cancer patients. *Front. Oncol.* 20 (12), 988680. doi:10.3389/fonc.2022.988680

Kalemkerian, G. P., Akerley, W., Bogner, P., Borghaei, H., Chow, L. Q., Downey, R. J., et al. (2013). Small cell lung cancer. *J. Natl. Compr. Canc Netw.* 11 (1), 78–98. doi:10.6004/jnccn.2013.0011

Kerachian, M. A., Poudineh, A., and Thiery, J. P. (2019). Cell free circulating tumor nucleic acids, a revolution in personalized cancer medicine. *Crit. Rev. Oncol. Hematol.* 144, 102827. doi:10.1016/j.critrevonc.2019.102827

Khan, K. H., Cunningham, D., Werner, B., Vlachogiannis, G., Spiteri, I., Heide, T., et al. (2018). Longitudinal liquid biopsy and mathematical modeling of clonal evolution forecast time to treatment failure in the PROSPECT-C phase II colorectal cancer clinical trial. *Cancer Discov.* 8 (10), 1270–1285. doi:10.1158/2159-8290.CD-17-0891

Kitahata, Y., Kawai, M., Hirono, S., Okada, K. I., Miyazawa, M., Motobayashi, H., et al. (2022). Circulating tumor DNA as a potential prognostic marker in patients with borderline-resectable pancreatic cancer undergoing neoadjuvant chemotherapy followed by pancreatectomy. *Ann. Surg. Oncol.* 29 (3), 1596–1605. doi:10.1245/s10434-021-10985-0

Kotaka, M., Shirasu, H., Watanabe, J., Yamazaki, K., Hirata, K., Akazawa, N., et al. (2022). Association of circulating tumor DNA dynamics with clinical outcomes in the adjuvant setting for patients with colorectal cancer from an observational GALAXY study in CIRCULATE-Japan. *J. Clin. Oncol.* 40, 9. doi:10.1200/jco.2022.40.4_suppl.009

Kuang, P. P., Li, N., Liu, Z., Sun, T. Y., Wang, S. Q., Hu, J., et al. (2020). Circulating tumor DNA analyses as a potential marker of recurrence and effectiveness of adjuvant chemotherapy for resected non-small-cell lung cancer. *Front. Oncol.* 10, 595650. doi:10.3389/fonc.2020.595650

Kurtz, D. M., Soo, J., Co Ting Keh, L., Alig, S., Chabon, J. J., Sworder, B. J., et al. (2021). Enhanced detection of minimal residual disease by targeted sequencing of phased variants in circulating tumor DNA. *Nat. Biotechnol.* 39, 1537–1547. doi:10.1038/s41587-021-00981-w

KyrillusShohdy, S., Villamar, D. M., Cao, Y., Trieu, J., Price, K. S., Nagy, R., et al. (2022). Serial ctDNA analysis predicts clinical progression in patients with advanced urothelial carcinoma. *Br. J. Cancer* 126 (3), 430–439. doi:10.1038/s41416-021-01648-8

Lan, Y-T., Chen, M-H., Fang, W-L., Hsieh, C. C., Jhang, F. Y., et al. (2017). Clinical relevance of cell-free DNA in gastrointestinal tract malignancy. *Oncotarget* 8 (2), 3009–3017. doi:10.18632/oncotarget.13821

Liu, B., Hu, Z., Ran, J., Xie, N., Tian, C., Tang, Y., et al. (2022). The circulating tumor DNA (ctDNA) alteration level predicts therapeutic response in metastatic breast cancer: novel prognostic indexes based on ctDNA. *Breast* 65, 116–123. doi:10.1016/j.breast.2022.07.010

Liu, W., Li, Y., Tang, Y., Song, Q., Wang, J., et al. (2022). Response prediction and risk stratification of patients with rectal cancer after neoadjuvant therapy through an analysis of circulating tumour DNA. *EBioMedicine* 78, 103945. doi:10.1016/j.ebiom.2022.103945

Lou, Z., GongYQZhou, X., and Hu, G. H. (2018). Low expression of miR-199 in hepatocellular carcinoma contributes to tumor cell hyper-proliferation by negatively suppressing XBP1. *Oncol. Lett.* 16 (5), 6531–6539. doi:10.3892/ol.2018.9476

Magbanua, M. J. M., Swigart, L. B., Wu, H. T., Hirst, G. L., Yau, C., Wolf, D. M., et al. (2021). Circulating tumor DNA in neoadjuvant-treated breast cancer reflects response and survival. *Ann. Oncol.* 32 (2), 229–239. doi:10.1016/j.annonc.2020.11.007

Mouliere, F., Chandrananda, D., Piskorz, A. M., Morris, J., Moore, E. K., Ahlborn, L. B., et al. (2018). Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* 10, eaat4921. doi:10.1126/scitranslmed.aat4921

Mouliere, F., Robert, B., Peyrotte, E. A, Rio, M. D, Ychou, M., Molina, F., et al. (2011). High fragmentation characterizes tumour-derived circulating DNA. *PLoS ONE* 6, e23418. doi:10.1371/journal.pone.0023418

Murtaza, M., Dawson, S. J., Tsui, D. W. Y., Gale, D., Forshew, T., Piskorz, A. M., et al. (2013). Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* 497, 108–112. doi:10.1038/nature12065

Nakano, K., Koh, Y., Yamamichi, G., Yumiba, S., Tomiyama, E., Matsushita, M., et al. (2022). Perioperative circulating tumor DNA enables the identification of patients with poor prognosis in upper tract urothelial carcinoma. *Cancer Sci.* 113 (5), 1830–1842. doi:10.1111/cas.15334

Newman, A. M., Bratman, S. V., Stehr, H., Lee, L. J., Liu, C. L., Diehn, M., et al. (2014). Factera: a practical method for the discovery of genomic rearrangements at breakpoint resolution. *Bioinformatics* 30, 3390–3393. doi:10.1093/bioinformatics/btu549

Newman, A. M., Bratman, S. V., To, J., Wynne, J. F., Eclov, N. C. W., Modlin, L. A., et al. (2014). An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* 20, 548–554. doi:10.1038/nm.3519

Newman, A. M., Lovejoy, A. F., Klass, D. M., Kurtz, D. M., Chabon, J. J., Scherer, F., et al. (2016). Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* 34, 547–555. doi:10.1038/nbt.3520

Normando, S. R. C., Delgado, P. O., Rodrigues, A., David Filho, W. J., Fonseca, F. L. A., Cruz, F. J. S. M., et al. (2018). Circulating free plasma tumor DNA in patients with advanced gastric cancer receiving systemic chemotherapy. *BMC Clin. Pathol.* 18, 12. doi:10.1186/s12907-018-0079-y

Normanno, N., Esposito Abate, R., Lambiase, M., Forgione, L., Cardone, C., Iannaccone, A., et al. (2018). RAS testing of liquid biopsy correlates with the outcome of metastatic colorectal cancer patients treated with first-line FOLFIRI plus cetuximab in the CAPRI-GOIM trial. *Ann. Oncol.* 29, 112–118. doi:10.1093/annonc/mdx417

Osterman, E., and Glimelius, B. (2018). Recurrence risk after up-to-date colon cancer staging, surgery, and pathology: analysis of the entire Swedish population. *Dis. Colon. Rectum* 61, 1016–1025. doi:10.1097/DCR.0000000000001158

Osumi, H., Shinozaki, E., Yamaguchi, K., and Zembutsu, H. (2019). Early change in circulating tumor DNA as a potential predictor of response to chemotherapy in patients with metastatic colorectal cancer. *Sci. Rep.* 9, 17358. doi:10.1038/s41598-019-53711-3

Påhlman, L. A., Hohenberger, W. M, Matzel, K., Sugihara, K., and Quirke, P. (2016). Should the benefit of adjuvant chemotherapy in colon cancer Be Re-evaluated? *J. Clin. Oncol.* 34, 1297–1299. doi:10.1200/JCO.2015.65.3048

Parsons, H. A., Rhoades, J., Reed, S. C., Gydush, G., Ram, P., Exman, P., et al. (2020). Sensitive detection of minimal residual disease in patients treated for early-stage breast cancer. *Clin. Cancer Res.* 26 (11), 2556–2564. doi:10.1158/1078-0432.CCR-19-3005

Peng, M., Huang, Q., Yin, W., Tan, S., Chen, C., Liu, W., et al. (2020). Circulating tumor DNA as a prognostic biomarker in localized non-small cell lung cancer. *Front. Oncol.* 10, 561598. doi:10.3389/fonc.2020.561598

Peng, P. X., Luan, Y. S., Sun, P., Wang, L., Zeng, X., Wang, Y., et al. (2022). Prognostic factors in stage IV colorectal cancer patients with resection of liver and/or pulmonary metastases: a population-based cohort study. *Front. Oncol.* 12, 850937. doi:10.3389/fonc.2022.850937

Phallen, J., Mark, S., Leal, A., Adleff, V., Hruban, C., Hruban, C., et al. (2017). Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.* 9, eaan2415. doi:10.1126/scitranslmed.aan2415

Pietrasz, D., Pécuchet, N., Garlan, F., Didelot, A., Dubreuil, O., Doat, S., et al. (2017). Plasma circulating tumor DNA in pancreatic cancer patients is a prognostic marker. *Clin. Cancer Res.* 23, 116–123. doi:10.1158/1078-0432.CCR-16-0806

Razavi, P., Li, B. T., Brown, D. N., Jung, B., Hubbell, E., Shen, R., et al. (2019). High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat. Med.* 25, 1928–1937. doi:10.1038/s41591-019-0652-7

Riva, F., Bidard, F. C., Houy, A., Saliou, A., Madic, J., Rampanou, A., et al. (2017). Patient-specific circulating tumor DNA detection during neoadjuvant chemotherapy in triple-negative breast cancer. *Clin. Chem.* 63, 691–699. doi:10.1373/clinchem.2016.262337

Schiavon, G., Hrebien, S., Garcia-Murillas, I., Cutts, R. J., Pearson, A., Tarazona, N., et al. (2015). Analysis of ESR1 mutation in circulating tumor DNA demonstrates evolution during therapy for metastatic breast cancer. *Sci. Transl. Med.* 7, 313ra182. doi:10.1126/scitranslmed.aac7551

Schøler, L. V., Reinert, T., Ørntoft, M-B. W., Kassentoft, C. G., Árnadóttir, S. S., Vang, S., et al. (2017). Clinical implications of monitoring circulating tumor DNA in patients with colorectal cancer. *Clin. Cancer Res.* 23, 5437–5445. doi:10.1158/1078-0432.CCR-17-0510

Schwarzenbach, H., Hoon, DSB., and Pantel, K. (2011). Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* 11, 426–437. doi:10.1038/nrc3066

Siravegna, G., Mussolin, B., Venesio, T., Marsoni, S., Seoane, J., et al. (2019). How liquid biopsies can change clinical practice in oncology. *Ann. Oncol.* 30, 1580–1590. doi:10.1093/annonc/mdz227

Snyder, M. W., Kircher, M., Hill, A. J, Daza, R. M, and Shendure, J. (2016). Cell-free DNA comprises an *in vivo* nucleosome footprint that informs its tissues-of-origin. *Cell* 164, 57–68. doi:10.1016/j.cell.2015.11.050

Steensma, D., Bejar, R., Ebert, B., Lindsley, R. C., Sekeres, M. A., Hasserjian, R. P., et al. (2015). Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* 126, 9–16. doi:10.1182/blood-2015-03-631747

Taniguchi, H., Nakamura, Y., Kotani, D., Yukami, H., Mishima, S., Sawada, K., et al. (2021). CIRCULATE-Japan: circulating tumor DNA-guided adaptive platform trials to refine adjuvant therapy for colorectal cancer. *Cancer Sci.* 112 (7), 2915–2920. doi:10.1111/cas.14926

Tie, J., Cohen, J. D., Lo, S. N., Wang, Y., Li, L., Christie, M., et al. (2021). Prognostic significance of postsurgery circulating tumor DNA in nonmetastatic colorectal cancer: individual patient pooled analysis of three cohort studies. *Int. J. Cancer.* 148, 1014–1026. doi:10.1002/ijc.33312

Tie, J., Cohen, J. D., Wang, Y., Li, L., Christie, M., Simons, K., et al. (2019). Serial circulating tumour DNA analysis during multimodality treatment of locally advanced rectal cancer: a prospective biomarker study. *Gut* 68, 663–671. doi:10.1136/gutjnl-2017-315852

Tie, J., Kinde, I., Wang, Y., Wong, H. L., Roebert, J., Christie, M., et al. (2015). Circulating tumor DNA as an early marker of therapeutic response in patients with metastatic colorectal cancer. *Ann. Oncol.* 26, 1715–1722. doi:10.1093/annonc/mdv177

Tie, J., Wang, Y., Tomasetti, C., Li, L., Springer, S., Kinde, I., et al. (2016). Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Sci. Transl. Med.* 8, 346ra92. 346ra392. doi:10.1126/scitranslmed.aaf6219

Tie, J., Cohen, J. D., Wang, Y., Simons, K., Simons, K., et al. (2019). Circulating tumor DNA analyses as markers of recurrence risk and benefit of adjuvant therapy for stage III colon cancer. *JAMA Oncol.* 5, 1710–1717. doi:10.1001/jamaoncol.2019.3616

Underhill, H. R., Kitzman, J. O., Hellwig, S., Welker, N. C., Daza, R., Baker, D. N., et al. (2016). Fragment length of circulating tumor DNA. *PLoS Genet.* 12, e1006162. doi:10.1371/journal.pgen.1006162

van Velzen, M. J. M., Creemers, A., van den Ende, T., Schokker, S., Krausz, S., Reinten, R. J., et al. (2022). Circulating tumor DNA predicts outcome in metastatic gastroesophageal cancer. *Gastric Cancer* 25 (5), 906–915. doi:10.1007/s10120-022-01313-w

Vessies, D. C. L., Greuter, M. J. E., van Rooijen, K. L., Linders, T. C., Lanfermeijer, M., Ramkisoensing, K. L., et al. (2020). Performance of four platforms for KRAS mutation detection in plasma cell-free DNA: ddPCR, idylla, COBAS z480 and BEAMing. *Sci. Rep.* 10, 8122. Nature Publishing Group. doi:10.1038/s41598-020-64822-7

Wang, N. N., Chen, J., Chen, W. J., Shi, Z., Yang, H., Liu, P., et al. (2022). The effectiveness of case management for cancer patients: an umbrella review. *BMC Health Serv. Res.* 22 (1), 1247. doi:10.1186/s12913-022-08610-1

Wan, J. C. M., Massie, C., Garcia-Corbacho, J., Brenton, J. D., et al. (2017). Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* 17, 223–238. doi:10.1038/nrc.2017.7

Wei, S. S., Sun, T. L., Du, J., Zhang, B., Xiang, D., and Li, W. (2018). Xanthohumol, a prenylated flavonoid from Hops, exerts anticancer effects against gastric cancer *in vitro*. *Oncol. Rep.* 40 (6), 3213–3222. doi:10.3892/or.2018.6723

Welch, J. S., Larson, D. E., Wallis, J., Klco, J. M., Kulkarni, S., et al. (2011). Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA* 305, 1577–1584. doi:10.1001/jama.2011.497

Wen, L., Cheng, F. Z., Zhou, Y. Y., and Yin, C. (2015). MiR-26a enhances the sensitivity of gastric cancer cells to cisplatin by targeting NRAS and E2F2. *Saudi J. Gastroenterol.* 21 (5), 313–319. doi:10.4103/1319-3767.166206

Wirtzfeld, D. A., Mikula, L., Gryfe, R., Dicks, E. L., Parfrey, P., et al. (2009). Concordance with clinical practice guidelines for adjuvant chemotherapy in patients with stage I–III colon cancer: experience in 2 Canadian provinces. *Can. J. Surg.* 52, 92–97.

Xiao, Y. B., Liu, M. Q., Ning, Q., Xiang, Z., Zheng, X., et al. (2022). MiR-26a-5p regulates proliferation, apoptosis, migration and invasion via inhibiting hydroxysteroid dehydrogenase like-2 in cervical cancer cell. *BMC Cancer* 22 (1), 876. doi:10.1186/s12885-022-09970-x

Xiong, T., Li, Z. H., Huang, X. L., Lu, K., Xie, W., Zhou, Z., et al. (2019). TO901317 inhibits the development of hepatocellular carcinoma by LXRα/

Glut1 decreasing glycometabolism. *Am. J. Physiol. Gastrointest. Liver Physiol.* 316 (5), G598–G607. doi:10.1152/ajpgi.00061.2018

Xu, H. F., Wang, H., Zhao, W., Fu, S., Li, Y., Ni, W., et al. (2020). SUMO1 modification of methyltransferase-like 3 promotes tumor progression via regulating Snail mRNA homeostasis in hepatocellular carcinoma. *Theranostics* 10 (13), 5671–5686. doi:10.7150/thno.42539

Ye, K., Fan, Q., Yuan, M., Wang, D., Xiao, L., Long, G., et al. (2022). Prognostic value of postoperative circulating tumor DNA in patients with early- and intermediate-stage hepatocellular carcinoma. *Front. Oncol.* 12, 834992. doi:10.3389/fonc.2022.834992

Zhang, J-T., Liu, S-Y., Wu, Y-L., Liu, S. Y. M., Yan, H. H., Ji, L., et al. (2022). Longitudinal undetectable molecular residual disease defines potentially cured population in localized non–small cell lung cancer. *Cancer Discov.* 12 (7), 1690–1701. doi:10.1158/2159-8290.CD-21-1486

Zhao, Y., Chen, S. Y., Shen, F., Long, D., Yu, T., Wu, M., et al. (2019). *In vitro* neutralization of autocrine IL-10 affects Op18/stathmin signaling in non-small cell lung cancer cells. *Oncol. Rep.* 41 (1), 501–511. doi:10.3892/or.2018.6795

Zheng, Y. W., Chan, K. C., SunJiangSuChenLunHungLeeWong, H. P. X. E. Z. F. M. E. C. V. J., Su, X., et al. (2012). Nonhematopoietically derived DNA is shorter than hematopoietically derived DNA in plasma: a transplantation model. *Clin. Chem.* 58, 549–558. doi:10.1373/clinchem.2011.169318

Zhu, G-Q., Liu, W-R., Tang, Z., Qu, W. F., Fang, Y., Jiang, X. F., et al. (2022). Serial circulating tumor DNA to predict early recurrence in patients with hepatocellular carcinoma: a prospective study. *Mol. Oncol.* 16 (2), 549–561. doi:10.1002/1878-0261.13105

Zonta, E., Garlan, F., Pécuchet, N., Perez-Toralla, K., Caen, O., Milbury, C., et al. (2016). Multiplex detection of rare mutations by picoliter droplet based digital PCR: sensitivity and specificity considerations. *PLoS ONE* 11, e0159094. doi:10.1371/journal.pone.0159094

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 16 August 2023
DOI 10.3389/fgene.2023.1238095

Check for updates

OPEN ACCESS

EDITED BY
Junlin Xu,
Hunan University, China

REVIEWED BY
XianFang Tang,
Wuhan Textile University, China
Wenyan Wang,
Anhui University of Technology, China

*CORRESPONDENCE
Zejun Li,
✉ lzjfox@hnit.edu.cn
Lian Duan,
✉ duanlian301@163.com

[†]These authors have contributed equally to this work and share first authorship

RECEIVED 10 June 2023
ACCEPTED 19 July 2023
PUBLISHED 16 August 2023

CITATION
Su Z, Lu H, Wu Y, Li Z and Duan L (2023), Predicting potential lncRNA biomarkers for lung cancer and neuroblastoma based on an ensemble of a deep neural network and LightGBM.
Front. Genet. 14:1238095.
doi: 10.3389/fgene.2023.1238095

Predicting potential lncRNA biomarkers for lung cancer and neuroblastoma based on an ensemble of a deep neural network and LightGBM

Zhenguo Su[1†], Huihui Lu[2†], Yan Wu[3], Zejun Li[4*] and Lian Duan[5,6,7,8*]

[1]Clinical Lab, Yantai Affiliated Hospital of Binzhou Medical University, Yantai, China, [2]Department of Thoracic Cardiovascular Surgery, Hunan Province Directly Affiliated TCM Hospital, Zhuzhou, China, [3]Geneis (Beijing) Co., Ltd., Beijing, China, [4]School of Computer Science, Hunan Institute of Technology, Hengyang, China, [5]Faculty of Pediatrics, The Chinese PLA General Hospital, Beijing, China, [6]Department of Pediatric Surgery, The Seventh Medical Center of PLA General Hospital, Beijing, China, [7]National Engineering Laboratory for Birth Defects Prevention and Control of Key Technology, Beijing, China, [8]Beijing Key Laboratory of Pediatric Organ Failure, Beijing, China

Introduction: Lung cancer is one of the most frequent neoplasms worldwide with approximately 2.2 million new cases and 1.8 million deaths each year. The expression levels of programmed death ligand-1 (PDL1) demonstrate a complex association with lung cancer. Neuroblastoma is a high-risk malignant tumor and is mainly involved in childhood patients. Identification of new biomarkers for these two diseases can significantly promote their diagnosis and therapy. However, *in vivo* experiments to discover potential biomarkers are costly and laborious. Consequently, artificial intelligence technologies, especially machine learning methods, provide a powerful avenue to find new biomarkers for various diseases.

Methods: We developed a machine learning-based method named LDAenDL to detect potential long noncoding RNA (lncRNA) biomarkers for lung cancer and neuroblastoma using an ensemble of a deep neural network and LightGBM. LDAenDL first computes the Gaussian kernel similarity and functional similarity of lncRNAs and the Gaussian kernel similarity and semantic similarity of diseases to obtain their similar networks. Next, LDAenDL combines a graph convolutional network, graph attention network, and convolutional neural network to learn the biological features of the lncRNAs and diseases based on their similarity networks. Third, these features are concatenated and fed to an ensemble model composed of a deep neural network and LightGBM to find new lncRNA–disease associations (LDAs). Finally, the proposed LDAenDL method is applied to identify possible lncRNA biomarkers associated with lung cancer and neuroblastoma.

Results: The experimental results show that LDAenDL computed the best AUCs of 0.8701, 107 0.8953, and 0.9110 under cross-validation on lncRNAs, diseases, and lncRNA-disease pairs on Dataset 1, respectively, and 0.9490, 0.9157, and 0.9708 on Dataset 2, respectively. Furthermore, AUPRs of 0.8903, 0.9061, and 0.9166 under three cross-validations were obtained on Dataset 1, and 0.9582, 0.9122, and 0.9743 on Dataset 2. The results demonstrate that LDAenDL significantly outperformed the other four classical LDA prediction methods (i.e., SDLDA, LDNFSGB, IPCAF, and LDASR). Case studies demonstrate that CCDC26 and IFNG-AS1 may be new biomarkers of lung cancer, SNHG3 may associate with PDL1 for lung cancer, and HOTAIR and BDNF-AS may be potential biomarkers of neuroblastoma.

Conclusion: We hope that the proposed LDAenDL method can help the development of targeted therapies for these two diseases.

KEYWORDS

lncRNA, biomarker, lung cancer, neuroblastoma, deep neural network, LightGBM

1 Introduction

Long non-coding RNAs (lncRNAs) are non-coding RNAs with more than 200 nucleotides (Bertone et al., 2004; Peng et al., 2022a; Peng et al., 2022b). LncRNAs play an important role in the development and progression of various diseases (Lanjanian et al., 2021; Meng et al., 2021; Yang and Li 2021; Peng et al., 2022c). LncRNAs have dense associations with many diseases, for example, lung cancer, colorectal cancer, prostate cancer, and Alzheimer's disease (Klattenhoff et al., 2013; Tan et al., 2013; Chakravarty et al., 2014; He et al., 2014; Zhang et al., 2014). LncRNA H19 is associated with the under-regulation of renal carcinoma cells (Wang et al., 2015). The expression of EGOT in breast cancer is much lower than one in adjacent noncancerous tissues (Broadbent et al., 2008). NEAT1 is overexpressed in prostate cancer cells (Pasmant et al., 2011). The identification of lncRNA-disease associations (LDAs) helps us to further understand the biological processes and the molecular mechanisms of various complex diseases. However, the number of known and experimentally validated LDAs is very small. Thus, it is important to identify potential LDAs. Determining LDAs through *in vivo* experiments is costly and time-consuming, therefore, it is necessary to design efficient computational approaches for identifying potential LDAs (Meng et al., 2021; Peng et al., 2022d). Computational LDA prediction methods are categorized as biological network-based methods and machine learning-based methods.

Biological network-based methods use network algorithms for association prediction (Liu et al., 2023a). This type of method first constructs heterogeneous networks of lncRNAs and diseases and then identifies LDAs via matrix decomposition, random walk, and so on. To predict potential LDAs, LRWRHLDA combined Laplace normalized random walk with restart (Wang et al., 2022), LDGRNMF used graph regularized nonnegative matrix factorization (Wang et al., 2021), DSCMF developed a dual sparse collaborative matrix factorization approach (Liu et al., 2021a), RWSF-BLP added random walk-based multi-similarity fusion to bidirectional label propagation (Xie et al., 2021), HBRWRLDA utilized bi-random walk on hypergraphs (Xie et al., 2022), and MHRWRLDA exploited a random walk model with restart through multiplex and heterogeneous networks (Yao et al., 2021).

With the fast advance of RNA sequencing technologies, artificial intelligence has obtained wide applications in biomedical data analysis (Peng et al., 2023a; Peng et al., 2023b; Xu et al., 2023). Notably, artificial intelligence technologies, especially machine learning methods, have been widely applied to predict miRNA-disease associations (Liu et al., 2022) and circRNA-disease associations (Liu et al., 2023b). To find new LDAs, HGATLDA developed a novel heterogeneous graph attention network model (Zhao et al., 2022), DeepMNE extracted multi-omics data and designed a deep multi-network embedding model (Ma, 2022), iLncDA-LTR is a rank-based method (Wu et al., 2022), MAGCNSE utilized a graph convolutional network (Liang et al., 2022), LDAformer extracted topological features and used a transformer encoder for LDA classification (Zhou et al., 2022), BiGAN explored a bidirectional generative adversarial network (Yang et al., 2021), and SVDNVLDA extracted linear and non-linear features and used an XGBoost for LDA prediction (Li et al., 2021).

Computational methods have found many potential LDAs, however, network-based methods were more likely to favor well-investigated lncRNAs or diseases and can not predict LDAs for new lncRNAs or new diseases. Machine learning-based methods failed to effectively integrate different kernels from multiple data sources. Thus, in this study, we developed a machine learning-based method named LDAenDL to detect potential lncRNA biomarkers for lung cancer and neuroblastoma based on an ensemble of a deep neural network and LightGBM.

2 Materials and methods

As shown in Figure 1, LDAenDL first computes the Gaussian kernel similarity and functional similarity of lncRNAs and the Gaussian kernel similarity and semantic similarity of diseases to obtain their similar networks. Next, LDAenDL combines a graph convolutional network (GCN) (Kipf and Welling, 2016), graph attention network (GAT) (Velickovic et al., 2017), and convolutional neural network (Gu et al., 2018) to learn the biological features of lncRNAs and diseases based on their similarity networks. Third, these features are concatenated and fed to an ensemble model composed of a deep neural network (DNN) and LightGBM to find new LDAs. Finally, LDAenDL was applied to identify possible lncRNA biomarkers associated with lung cancer and neuroblastoma.

2.1 Data preparation

We used two human LDA datasets that were provided by Chen et al. (2012) and Cui et al. (2018). Dataset 1 contains 605 LDAs between 157 diseases and 82 lncRNAs. Dataset 2 contains 1,529 LDAs between 190 diseases and 89 lncRNAs. An LDA network can be denoted as $Y \in \mathfrak{R}^{n \times m}$ where $y_{ij} = 1$ if lncRNA l_i interacts with disease d_j, otherwise, it equals 0.

2.2 Similarity computation

Inspired by the LDA-DLPU method (Peng et al., 2022a), we computed the Gaussian kernel similarity and functional similarity of

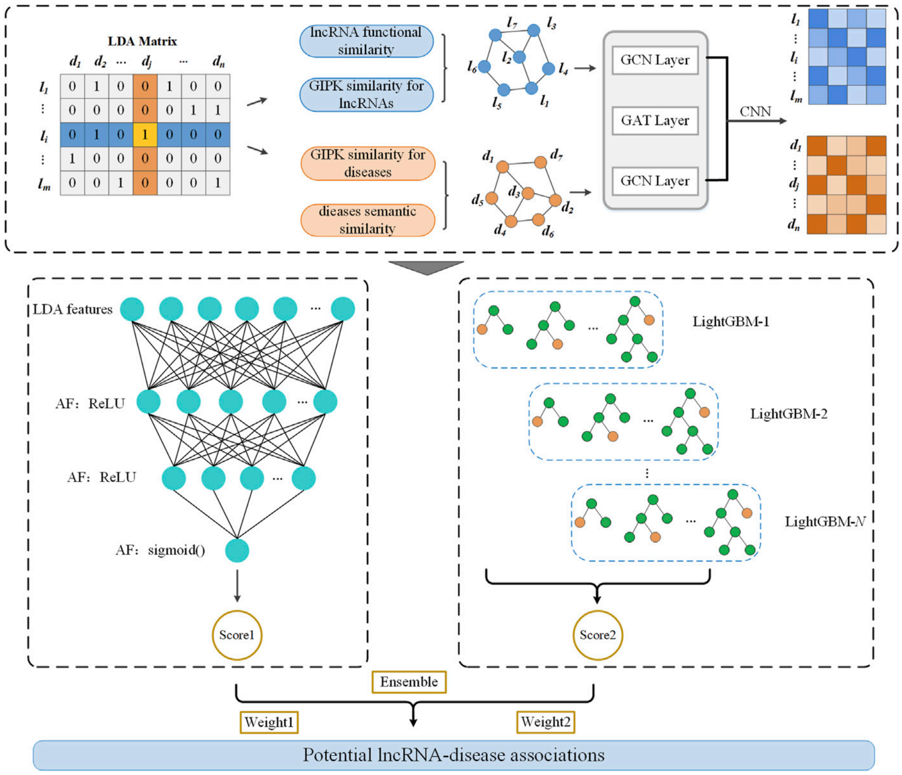

LDA Matrix
d_1 d_2 ... d_j ... d_n
l_1
l_i
l_m
lncRNA functional similarity
GIPK similarity for lncRNAs
GIPK similarity for diseases
dieases semantic similarity
GCN Layer
GAT Layer
GCN Layer
CNN
LDA features
AF: ReLU
AF: ReLU
AF: sigmoid()
Score1
LightGBM-1
LightGBM-2
LightGBM-N
Score2
Ensemble
Weight1
Weight2
Potential lncRNA-disease associations

FIGURE 1
The pipeline of LDAenDL.

lncRNAs and the Gaussian kernel similarity and semantic similarity of diseases. Based on the computed lncRNA similarity and disease similarity matrices, we learned the features of lncRNAs and diseases by combining a GCN, GAT, and CNN.

2.3 Feature learning

Dai et al. (2022) designed a hybrid graph representation learning model (GraphCDA) to represent the features of circRNAs and diseases and obtained better circRNA-disease association prediction performance. Inspired by GraphCDA proposed by Dai et al. (2022), we exploit a GraphCDA-based LDA feature learning model.

2.3.1 Graph convolutional network

A GCN was applied to obtain the feature representations of lncRNAs and diseases based on their similarity networks. For a GCN G, it is denoted as an adjacency matrix $S \in R^{N\times N}$ with N nodes where each node can be described as an F-dimensional vector. And GCN outputs node representation matrix H^{new} in Eqs 1, 2:

$$H^{new} = GCN(S, H) \tag{1}$$

$$\mathrm{GCN}(S, H) = \sigma\left(A^{-\frac{1}{2}} S' A^{-\frac{1}{2}} H Q\right) \tag{2}$$

where $S' = I + S$, $A = \sum_j S'_{i,j}$ and $Q \in R^{F\times F}$ denote degree matrix and trainable weight matrix, and $\sigma(\cdot)$ denotes a ReLU activation function.

2.3.2 Graph attention network

A GAT (Veličković et al., 2017) uses multi-head attention to set weights for all adjacent nodes based on their importance. LDAenDL introduces a GAT layer between two GCN layers to help the GCN to extract high-level features of lncRNAs and diseases.

For the GCN G, a GAT layer outputs node representations H^{new} in Eq. 3:

$$H^{new} = GAT(S, H) \tag{3}$$

For K attention mechanisms in multi-head attention and its weight matrix W_k, let $\overrightarrow{H_i}$ denote the input feature vector of the i-th lncRNA, its feature representation $\vec{H}_i^{new}$ in H^{new} can be denoted as Eq. 4:

$$\vec{H}_i^{new} = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\neq i}^{n}\phi_{ij}^{k} W_k \vec{H}_i\right) \tag{4}$$

where ϕ_{it}^{k} denotes the k-th attention coefficients between two lncRNA nodes i and t:

$$\phi_{ij}^{k} = \frac{\exp\left(\mathrm{f}\left(a_k^T\left[W_k\vec{H}_i \parallel W_k\vec{H}_j \parallel B_k S_{ij}\right]\right)\right)}{\sum_{t\neq i}\exp\left(\mathrm{f}\left(a_k^T\left[W_k\vec{H}_i \parallel W_k\vec{H}_t \parallel B_k S_{it}\right]\right)\right)} \tag{5}$$

where $\|$ denotes a concatenation operation, f denotes the LeaklyReLU activation function, $a_k \in R^{2F+1}$ denotes a weight vector related to the k-th attention mechanism, and B_k denotes the weight of an edge S_{ij}.

2.3.3 Feature representation of lncRNAs and diseases

For a lncRNA similarity network G_c, its adjacency matrix C, and node feature matrix $H_c^{(0)} \in R^{N_c \times F_c}$, we alternately use GCN and GAT layers to obtain the graph feature representation of lncRNAs at different levels in Eq. 6:

$$\begin{cases} H_c^{(1)} = GCN\left(C, H_c^{(0)}\right) \\ H_c^{(2)} = GAT\left(C, H_c^{(1)}\right) \\ H_c^{(3)} = GCN\left(C, H_c^{(2)}\right) \end{cases} \tag{6}$$

Thus, a 1D CNN is used to produce the lncRNA feature representation matrix X_c by combining the output features $H_c^{(1)}$ and $H_c^{(3)}$ in the different GCN layers.

Similarly, the graph feature representations of diseases at different levels are denoted by Eq. 7:

$$\begin{cases} \mathrm{H}_d^{(1)} = \mathrm{GCN}\left(D, H_d^{(0)}\right) \\ \mathrm{H}_d^{(2)} = \mathrm{GAT}\left(D, H_d^{(1)}\right) \\ \mathrm{H}_d^{(3)} = \mathrm{GCN}\left(D, H_d^{(2)}\right) \end{cases} \tag{7}$$

A 1D CNN is used to produce the disease feature representation matrix X_c by combining the output features $H_{\mathrm{d}}^{(1)}$ and $H_{\mathrm{d}}^{(3)}$ in the different GCN layers.

2.3.4 Preference matrix construction

The preference matrix U that describes all lncRNA-disease pairs can be represented as Eq. 8 based on X_c and X_d:

$$\mathrm{U} = {X_c}^{\mathrm{T}} X_d \tag{8}$$

We used binary cross-entropy as the activation function to evaluate the difference between the preference matrix U and the known adjacency matrix R. By minimizing the loss function on two LDA datasets, the feature representation matrices X_c and X_d of lncRNAs and diseases are learned.

2.4 LDA prediction

2.4.1 DNN

We built a DNN to predict new LDAs based on known LDAs and the learned LDA features. The DNN contains an input layer, an output layer, and multiple hidden layers. In the input layer, there are F neurons that are the same as the number of LDA features.

Given an LDA sample x, the input layer with k inputs is represented by Eq. 9:

$$x = [x_1, x_2, \ldots x_k] \tag{9}$$

where x_i denotes the i-th feature in a sample x.

The hidden layer is represented by Eq. 10:

$$h_j = \sum\nolimits_{i=1}^{k} w_i x_i + b_j \tag{10}$$

where w_i and b_j denote the weight of x_i and the bias in the j-th hidden layer, respectively.

The output in the j-th hidden layer is denoted by Eq. 11:

$$h = f\left(h_j\right) \tag{11}$$

where f denotes a ReLU activation function. Finally, the output layer with the sigmoid function outputs the LDA prediction results in Eq. 12:

$$\sigma(h) = \frac{1}{1 + e^{-h}} \tag{12}$$

2.4.2 LightGBM

In this section, we built a LightGBM (Ke et al., 2017) to identify new LDAs. For a training set $X = \{(x_i, y_i)\}_{i=1}^{n}$ with n lncRNA-disease pair, LightGBM intends to build an approximation of $\hat{f}$ to a certain function $f(x)$ by minimizing the expected value of loss function $L(y, f(x))$ by Eq. 13:

$$\hat{f} = \arg\min_{f} \; E_{x,y}[L(y, f(x))] \tag{13}$$

LightGBM integrates T regression trees $\sum_{t=1}^{T} f_t(X)$ to approximate the final model by Eq. 14:

$$f_T(X) = \sum\nolimits_{t=1}^{T} f_t(X) \tag{14}$$

The regression trees are expressed as $w_{q(x)}, q \in \{1, 2, \ldots, J\}$, where J, q, and w denote the number of leaves, the decision rules of the tree, and the sample weight of leaf nodes, respectively.

At step t, LightGBM is trained in an additive form:

$$\Gamma_t = \sum_{i=1}^{n} L(y_i, F_{t-1}(x_i) + f_t(x_i)) \tag{15}$$

The objective function (15) is rapidly approximated with Newton's method (Sun et al., 2020).

To solve the objective function of LightGBM, we removed the constant term for simplicity, and model (15) can be represented as Eq. 16:

$$\Gamma_t \cong \sum_{i=1}^{n}\left(g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right) \tag{16}$$

where g_i and h_i are the first-order and second-order gradients related to the loss function. Given the sample set I_j related to leaf j, Eq. 16 is transformed to Eq. 17:

$$\Gamma_t = \sum_{j=1}^{J}\left(\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right) \tag{17}$$

Given a certain tree structure $q(x)$, for each leaf node w_j^*, its optimal leaf weight and the extreme value of Γ_k could be computed by Eq. 18:

$$\begin{gathered} w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \\ \Gamma_T^* = -\frac{1}{2}\sum_{j=1}^{J} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} \end{gathered} \tag{18}$$

where Γ_T^* is a scoring function used to evaluate the quality of a tree structure q. Finally, Model (15) can be denoted as:

$$G = \frac{1}{2}\left(\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda}\right) \tag{19}$$

TABLE 1 Comparison of LDAenDL with the other four methods under CV1.

		SDLDA	LDNFSGB	IPCARF	LDASR	LDAenDL
Precision	Dataset 1	0.8514 ± 0.0509	0.7004 ± 0.0639	0.4878 ± 0.1309	0.6726 ± 0.1200	**0.8764 ± 0.0493**
	Dataset 2	**0.9399 ± 0.0154**	0.8552 ± 0.0393	0.6615 ± 0.0966	0.8405 ± 0.0300	0.9391 ± 0.0290
Recall	Dataset 1	0.6521 ± 0.0732	0.6092 ± 0.0790	0.5721 ± 0.1580	0.5129 ± 0.0946	**0.7019 ± 0.0639**
	Dataset 2	0.8239 ± 0.0437	0.8021 ± 0.0498	0.6434 ± 0.1545	0.7358 ± 0.0562	**0.8304 ± 0.0523**
Accuracy	Dataset 1	0.7799 ± 0.0341	0.6769 ± 0.0423	0.4906 ± 0.0951	0.6417 ± 0.0597	**0.7996 ± 0.0312**
	Dataset 2	0.8857 ± 0.0283	0.8323 ± 0.0230	0.6526 ± 0.0775	0.7972 ± 0.0268	**0.8879 ± 0.0289**
F1-score	Dataset 1	0.7365 ± 0.0563	0.6462 ± 0.0451	0.5125 ± 0.1100	0.5668 ± 0.0536	**0.7768 ± 0.0399**
	Dataset 2	0.8775 ± 0.0278	0.8260 ± 0.0230	0.6401 ± 0.1017	0.7827 ± 0.0260	**0.8804 ± 0.0334**
AUC	Dataset 1	0.8023 ± 0.0477	0.7346 ± 0.0465	0.5096 ± 0.1432	0.7057 ± 0.0420	**0.8701 ± 0.0339**
	Dataset 2	0.9366 ± 0.0195	0.8839 ± 0.0270	0.7104 ± 0.0997	0.8641 ± 0.0256	**0.9490 ± 0.0220**
AUPR	Dataset 1	0.8461 ± 0.0553	0.7239 ± 0.0626	0.5336 ± 0.1423	0.6775 ± 0.0971	**0.8903 ± 0.0273**
	Dataset 2	0.9533 ± 0.0129	0.8832 ± 0.0307	0.7128 ± 0.1012	0.8671 ± 0.0252	**0.9582 ± 0.0167**

The bold value denotes the best performance.

where I_L and I_R denote the example sets in the left and right subtrees of q, respectively.

2.4.3 Ensemble learning

Through the solution of models (12) and (15), we can identify potential LDAs based on a DNN and LightGBM. Ensemble learning has better prediction accuracy than a single model. To further improve LDA prediction accuracy, we combined a DNN and LightGBM and developed an ensemble model for LDA identification through soft voting in Eq. 16:

$$\text{Score} = \alpha C_{\text{DNN}} + \beta C_{Light\text{GBM}} \tag{20}$$

where C_{DNN} and $C_{Light\text{GBM}}$ denote LDA prediction results from the DNN and LightGBM, respectively. α and β are their weights with values of 0.4 and 0.6, respectively. In particular, a lncRNA–disease pair is taken as an LDA if its association probability is greater than 0.5; otherwise, the pair is taken as a negative LDA.

3 Results

3.1 Evaluation metrics

In this article, we compared our proposed LDAenDL method with four LDA prediction methods, SDLDA, LDNFSGB, IPCAF, and LDASR. Precision, recall, accuracy, F1-score, AUC, and AUPR were used to compare the

TABLE 2 Comparison of LDAenDL with the other four methods under CV2.

		SDLDA	LDNFSGB	IPCARF	LDASR	LDAenDL
Precision	Dataset 1	0.8854 ± 0.0377	0.7548 ± 0.0639	0.5583 ± 0.0910	0.7462 ± 0.0613	**0.9135 ± 0.0317**
	Dataset 2	0.9232 ± 0.0331	0.8005 ± 0.0625	0.5557 ± 0.1473	0.7625 ± 0.0749	**0.9528 ± 0.0225**
Recall	Dataset 1	**0.7182 ± 0.0694**	0.7309 ± 0.0646	0.7538 ± 0.1067	0.6431 ± 0.0757	0.6649 ± 0.0814
	Dataset 2	**0.8579 ± 0.0655**	0.6936 ± 0.0794	0.5279 ± 0.1969	0.5758 ± 0.0894	0.4616 ± 0.1702
Accuracy	Dataset 1	**0.8187 ± 0.0282**	0.7552 ± 0.0291	0.5766 ± 0.0740	0.7165 ± 0.0339	0.8005 ± 0.0381
	Dataset 2	**0.9043 ± 0.0174**	0.7670 ± 0.0432	0.5593 ± 0.1159	0.7010 ± 0.0463	0.7196 ± 0.0821
F1-score	Dataset 1	**0.7917 ± 0.0519**	0.7407 ± 0.0526	0.6339 ± 0.0715	0.6873 ± 0.0512	0.7664 ± 0.0593
	Dataset 2	**0.8886 ± 0.0475**	0.7402 ± 0.0577	0.5190 ± 0.1434	0.6485 ± 0.0555	0.6032 ± 0.1612
AUC	Dataset 1	0.8788 ± 0.0274	0.8329 ± 0.0273	0.6402 ± 0.1004	0.7951 ± 0.0317	**0.8953 ± 0.0284**
	Dataset 2	**0.9559 ± 0.0160**	0.8603 ± 0.0363	0.5992 ± 0.1601	0.8045 ± 0.0362	0.9157 ± 0.0420
AUPR	Dataset 1	0.8934 ± 0.0387	0.8163 ± 0.0537	0.6355 ± 0.1217	0.7914 ± 0.0542	**0.9061 ± 0.0254**
	Dataset 2	**0.9561 ± 0.0354**	0.8292 ± 0.0680	0.6040 ± 0.1476	0.7630 ± 0.0717	0.9122 ± 0.0436

The bold value denotes the best performance.

TABLE 3 Comparison of LDAenDL with the other four methods under CV3.

		SDLDA	LDNFSGB	IPCARF	LDASR	LDAenDL
Precision	Dataset 1	**0.8782 ± 0.0306**	0.7782 ± 0.0270	0.7069 ± 0.0478	0.7695 ± 0.0393	0.8637 ± 0.0312
	Dataset 2	0.9178 ± 0.0154	0.8548 ± 0.0156	0.7693 ± 0.0850	0.8553 ± 0.0189	**0.9351 ± 0.0157**
Recall	Dataset 1	0.7256 ± 0.0376	0.8169 ± 0.0408	0.6155 ± 0.0652	0.6836 ± 0.0342	**0.8234 ± 0.0314**
	Dataset 2	0.8824 ± 0.0198	0.8818 ± 0.0204	0.5034 ± 0.1469	0.8204 ± 0.0238	**0.8999 ± 0.0179**
Accuracy	Dataset 1	0.8120 ± 0.0216	0.7916 ± 0.0256	0.6793 ± 0.0403	0.7385 ± 0.0283	**0.8462 ± 0.0229**
	Dataset 2	0.9015 ± 0.0114	0.8658 ± 0.0127	0.6793 ± 0.0753	0.8405 ± 0.0129	**0.9186 ± 0.0126**
F1-score	Dataset 1	0.7939 ± 0.0260	0.7965 ± 0.0262	0.6563 ± 0.0492	0.7233 ± 0.0289	**0.8426 ± 0.0232**
	Dataset 2	0.8996 ± 0.0119	0.8679 ± 0.0129	0.5995 ± 0.1312	0.8371 ± 0.0137	**0.9171 ± 0.0130**
AUC	Dataset 1	0.8774 ± 0.0200	0.8578 ± 0.0234	0.7384 ± 0.0466	0.8133 ± 0.0218	**0.9110 ± 0.0197**
	Dataset 2	0.9560 ± 0.0081	0.9346 ± 0.0074	0.7680 ± 0.0882	0.9143 ± 0.0112	**0.9708 ± 0.0062**
AUPR	Dataset 1	0.8952 ± 0.0177	0.8489 ± 0.0289	0.7409 ± 0.0515	0.8131 ± 0.0277	**0.9166 ± 0.0203**
	Dataset 2	0.9639 ± 0.0063	0.9273 ± 0.0098	0.7689 ± 0.0924	0.9100 ± 0.0136	**0.9743 ± 0.0058**

The bold value denotes the best performance.

performance of LDAenDL with the four methods. The six metrics have been defined by Peng et al. (2022b) (Shen et al., 2022).

TABLE 4 Comparison of LDAenDL with individual models under CV1.

		DNN	LightGBM	LDAenDL
Precision	Dataset 1	**0.8772 ± 0.0461**	0.8569 ± 0.0511	0.8764 ± 0.0493
	Dataset 2	0.9149 ± 0.0375	0.9386 ± 0.0278	**0.9391 ± 0.0290**
Recall	Dataset 1	0.6851 ± 0.0694	0.7106 ± 0.0714	**0.7019 ± 0.0639**
	Dataset 2	**0.8337 ± 0.0510**	0.8278 ± 0.0533	0.8304 ± 0.0523
Accuracy	Dataset 1	0.7930 ± 0.0317	0.7939 ± 0.0340	**0.7996 ± 0.0312**
	Dataset 2	0.8772 ± 0.0288	0.8865 ± 0.0295	**0.8879 ± 0.0289**
F1-score	Dataset 1	0.7664 ± 0.0429	0.7737 ± 0.0446	**0.7768 ± 0.0399**
	Dataset 2	0.8711 ± 0.0321	0.8786 ± 0.0344	**0.8804 ± 0.0334**
AUC	Dataset 1	**0.8712 ± 0.0373**	0.8622 ± 0.0340	0.8701 ± 0.0339
	Dataset 2	0.9308 ± 0.0209	**0.9497 ± 0.0227**	0.9490 ± 0.0220
AUPR	Dataset 1	0.8842 ± 0.0327	0.8822 ± 0.0284	**0.8903 ± 0.0273**
	Dataset 2	0.9449 ± 0.0190	**0.9586 ± 0.0171**	0.9582 ± 0.0167

The bold value denotes the best performance.

3.2 Comparison of LDAenDL with the other four methods

To implement the performance evaluation, inspired by the three cross-validations proposed by Zhou et al. (2021), we conducted cross-validations on lncRNAs (CV1), diseases (CV2), and lncRNA-disease pairs (CV3). Tables 1–3 give the precision, recall, accuracy, F1-score, AUC, and AUPR under CV1, CV2, and CV3 on two LDA datasets. In Tables 1–6, the bold font in each row denotes the best performance.

Under CV1, LDAenDL randomly took 80% of lncRNAs as training samples, and the rest were taken as test samples to investigate the LDA prediction ability for new lncRNAs. The results from Table 1 show that our proposed LDAenDL approach obtained the best precision, recall, accuracy, F1-score, AUC, and AUPR on two datasets under CV1 except that it computed slightly lower precision on Dataset 2 (0.9391 vs. 0.9399). It computed the highest AUPRs of 0.8903 and 0.9582, and far exceeded the AUPR values computed by SDLDA (i.e., 0.8461 and 0.9533).

Figure 2 shows the AUC and AUPR values computed by LDAenDL and the other four methods on two datasets under CV1. The results demonstrated that LDAenDL can discover possible diseases associated with a new lncRNA.

Under CV2, LDAenDL randomly took 80% of diseases as training samples, and the rest were taken as test samples to investigate the LDA prediction ability for new diseases. The results from Table 2 show that our proposed LDAenDL approach obtained better precision, AUC, and AUPR on two datasets under CV2. However, SDLDA computed higher recall, accuracy, and F1-score than LDAenDL, which may be caused by smaller disease samples.

Figure 3 shows the AUC and AUPR values computed by LDAenDL and the other four methods on two datasets under CV2. The results show that LDAenDL can be applied to screen possible lncRNAs associated with a new disease.

Under CV3, LDAenDL randomly took 80% of lncRNA-disease pairs as training samples, and the rest were taken as test samples to investigate the LDA prediction ability. The results from Table 3 show that our proposed LDAenDL approach obtained the best precision, recall, accuracy, F1-score, AUC, and AUPR on two datasets under CV3. It computed the highest AUCs of 0.9110 and 0.9708 and far exceeded

TABLE 5 Comparison of LDAenDL with individual models under CV2.

		DNN	LightGBM	LDAenDL
Precision	Dataset 1	0.9049 ± 0.0383	0.8927 ± 0.0309	**0.9135 ± 0.0317**
	Dataset 2	0.9274 ± 0.0412	0.9439 ± 0.0283	**0.9528 ± 0.0225**
Recall	Dataset 1	0.6182 ± 0.1006	**0.6873 ± 0.0734**	0.6649 ± 0.0814
	Dataset 2	0.3426 ± 0.1457	**0.5370 ± 0.1739**	0.4616 ± 0.1702
Accuracy	Dataset 1	0.7759 ± 0.0453	**0.8017 ± 0.0336**	0.8005 ± 0.0381
	Dataset 2	0.6580 ± 0.0689	**0.7533 ± 0.0842**	0.7196 ± 0.0821
F1-score	Dataset 1	0.7289 ± 0.0794	**0.7740 ± 0.0493**	0.7664 ± 0.0593
	Dataset 2	0.4835 ± 0.1531	**0.6678 ± 0.1537**	0.6032 ± 0.1612
AUC	Dataset 1	0.8853 ± 0.0374	0.8869 ± 0.0281	**0.8953 ± 0.0284**
	Dataset 2	0.8412 ± 0.0512	**0.9164 ± 0.0441**	0.9157 ± 0.0420
AUPR	Dataset 1	0.8882 ± 0.0368	0.8981 ± 0.0257	**0.9061 ± 0.0254**
	Dataset 2	0.8416 ± 0.0530	**0.9150 ± 0.0466**	0.9122 ± 0.0436

The bold value denotes the best performance.

TABLE 6 Comparison of LDAenDL with individual models under CV3.

		DNN	LightGBM	LDAenDL
Precision	Dataset 1	0.8561 ± 0.0357	0.8477 ± 0.0320	**0.8637 ± 0.0312**
	Dataset 2	0.9214 ± 0.0171	0.9322 ± 0.0157	**0.9351 ± 0.0157**
Recall	Dataset 1	**0.8241 ± 0.0373**	0.8110 ± 0.0381	0.8234 ± 0.0314
	Dataset 2	0.8983 ± 0.0204	0.8936 ± 0.0176	**0.8999 ± 0.0179**
Accuracy	Dataset 1	0.8419 ± 0.0244	0.8322 ± 0.0265	**0.8462 ± 0.0229**
	Dataset 2	0.9106 ± 0.0130	0.9142 ± 0.0122	**0.9186 ± 0.0126**
F1-score	Dataset 1	0.8389 ± 0.0247	0.8284 ± 0.0277	**0.8426 ± 0.0232**
	Dataset 2	0.9095 ± 0.0134	0.9124 ± 0.0126	**0.9171 ± 0.0130**
AUC	Dataset 1	0.9076 ± 0.0225	0.9015 ± 0.0204	**0.9110 ± 0.0197**
	Dataset 2	0.9562 ± 0.0107	0.9692 ± 0.0064	**0.9708 ± 0.0062**
AUPR	Dataset 1	0.9067 ± 0.0244	0.9082 ± 0.0215	**0.9166 ± 0.0203**
	Dataset 2	0.9611 ± 0.0102	0.9728 ± 0.0061	**0.9743 ± 0.0058**

The bold value denotes the best performance.

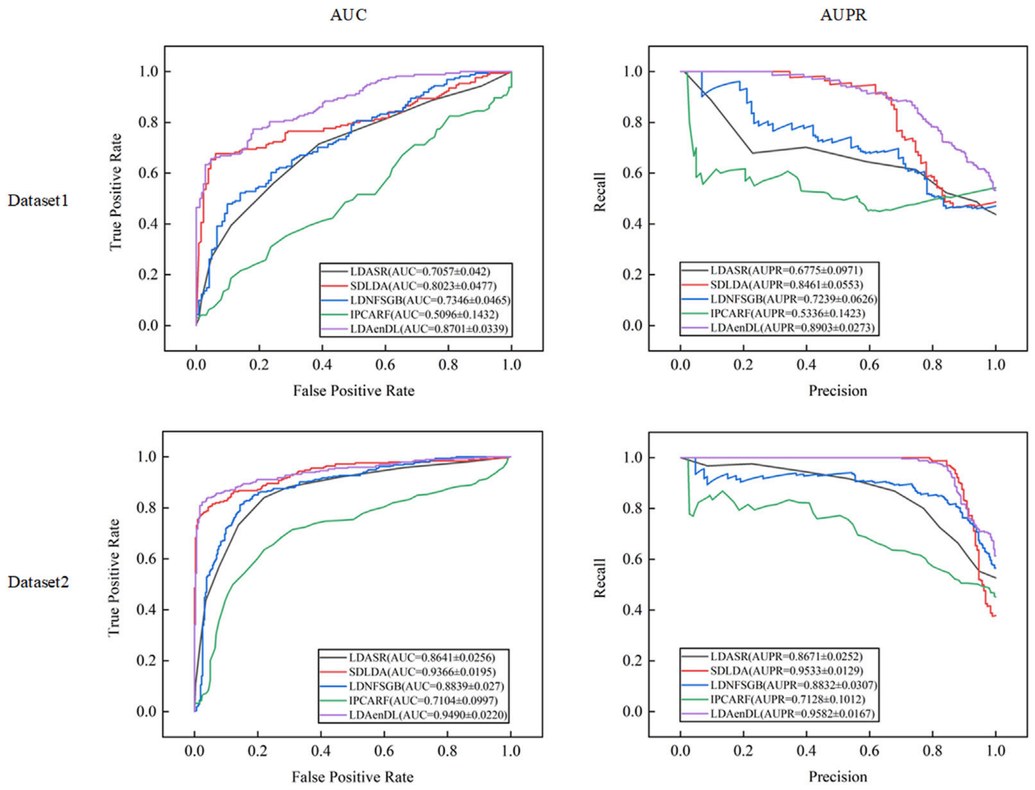


FIGURE 2
The AUC and AUPR values of five LDA prediction methods under CV1.

those computed by SDLDA (i.e., 0.8774 and 0.9560). Furthermore, our LDAenDL approach computed the highest AUPRs of 0.9166 and 0.9743 and far exceeded those computed by SDLDA (i.e., 0.8952, and 0.9639).

Figure 4 shows the AUC and AUPR values computed by LDAenDL and the other four methods on two datasets under CV3. The results demonstrated that LDAenDL could find potential LDAs based on known LDAs.

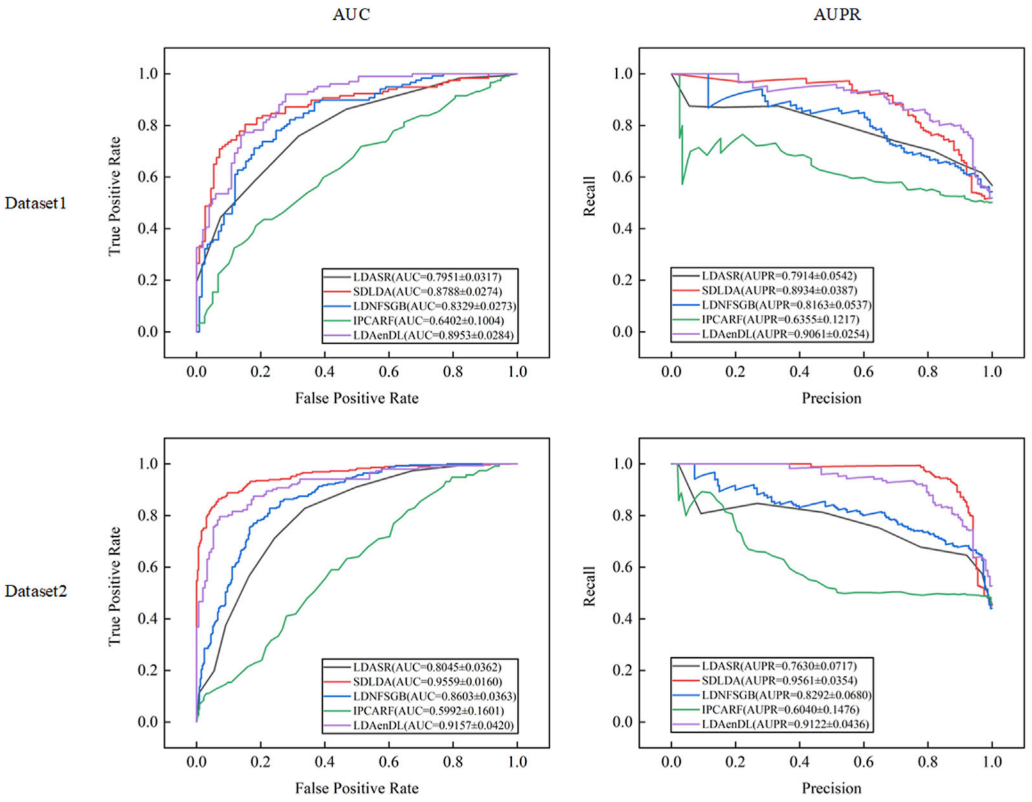


FIGURE 3
The AUC and AUPR values of five LDA prediction methods under CV2.

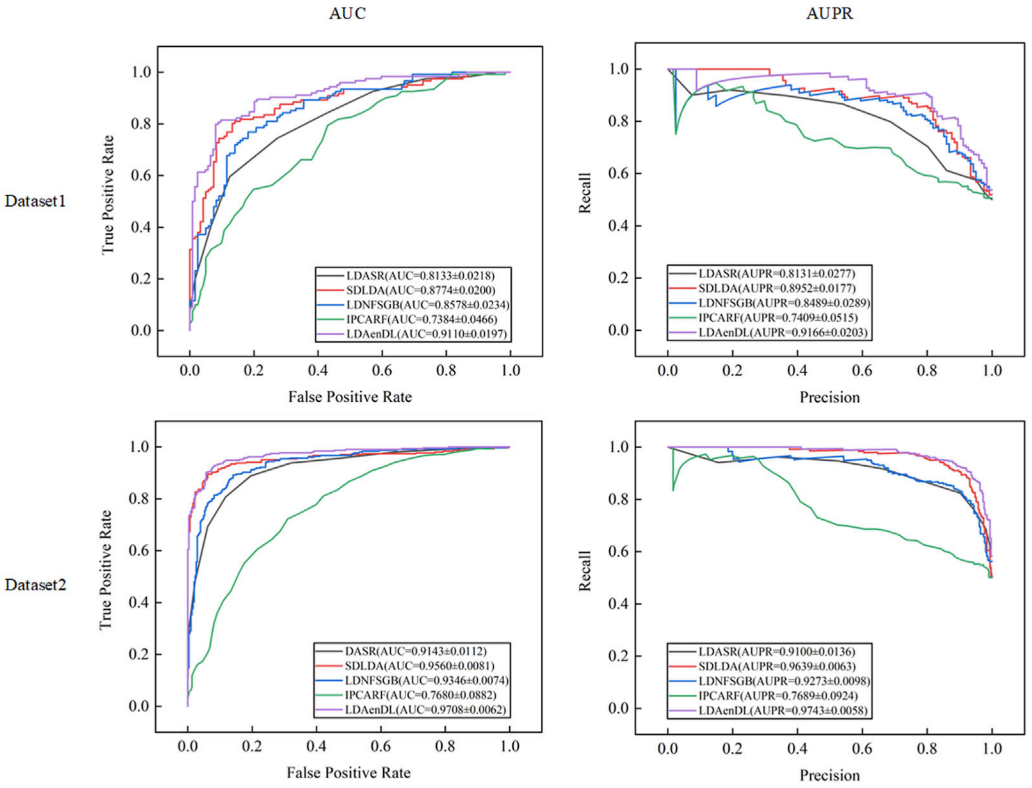


FIGURE 4
The AUC and AUPR values of five LDA prediction methods under CV3.

TABLE 7 The predicted top 20 lncRNA biomarkers for lung cancer in each of the two datasets.

Dataset 1			Dataset 2		
Rank	lncRNA	Evidence	Rank	lncRNA	Evidence
1	TUG1	27485439, 31532756	1	TUG1	27485439, 31532756
2	CRNDE	28550688, 30982057	2	DLEU2	31721438
3	DANCR	30535487, 32196604	3	WT1-AS	32349718
4	MIAT	29795987	4	CRNDE	28550688, 30982057
5	NPTN-IT1	27896272, 29416684	5	DANCR	30535487, 32196604
6	HNF1A-AS1	25863539	6	SNHG11	32239719
7	LINC00032	Unconfirmed	7	IFNG-AS1	Unconfirmed
8	WT1-AS	32349718	8	HULC	30575912
9	CBR3-AS1	32945466	9	XIST	29812958
10	HULC	30575912	10	PCA3	Unconfirmed
11	CCDC26	Unconfirmed	11	SRA1	Unconfirmed
12	SNHG3	31602642	12	HAR1A	Unconfirmed
13	PVT1	27904703	13	DSCAM-AS1	32280246
14	BCAR4	28537678	14	NPTN-IT1	27896272, 29416684
15	PTENP1	32698750	15	TCL6	Unconfirmed
16	RMST	Unconfirmed	16	PTENP1	32698750
17	LSINCT5	20214974	17	PANDAR	28121347
18	MIR155HG	32432745	18	TDRG1	31742752
19	BOK-AS1	Unconfirmed	19	KCNQ1OT1	31486494
20	KCNQ1OT1	31486494	20	IGF2-AS	28471495

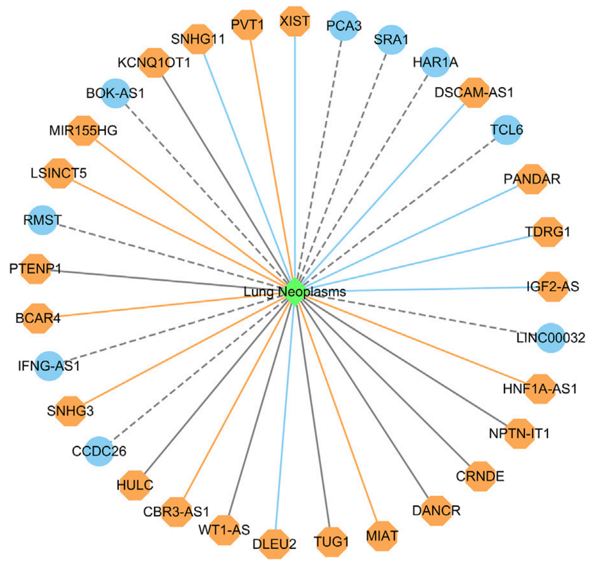


FIGURE 5
The top 20 predicted lncRNA biomarkers for lung cancer in each of the two datasets (The repeated lncRNAs in the two datasets have been removed). This figure was drawn using Cytoscape (Shannon et al., 2003).

3.3 Comparison of LDAenDL with individual models

To measure the effect of the ensemble algorithm on LDA prediction performance, we compared LDAenDL with two individual models, DNN, and LightGBM. Tables 4–6 show the precision, recall, accuracy, F1-score, AUC, and AUPR of the DNN, LightGBM, and LDAenDL under CV1, CV2, and CV3, respectively.

Under CV1, as shown in Table 4, LDAenDL outperformed the DNN and LightGBM on two LDA datasets for the majority of conditions. LDAenDL computed the best accuracy and F1-score on the two datasets. Although LDAenDL computed slightly lower AUC value than the DNN on dataset 1, and still slightly lower AUC than LightGBM on dataset 2, their differences were very small. For example, the DNN computed an AUC of 0.8712 while LDAenDL computed 0.8701 on dataset 1, and the DNN calculated an AUC of 0.9497 while LDAenDL calculated 0.9490 on dataset 2. LDAenDL obtained the best AUPR on dataset 1, and LightGBM obtained an AUPR of 0.9586 while LDAenDL obtained an AUPR of 0.9582.

Under CV2, as shown in Table 5, LDAenDL outperformed the DNN under all conditions on two LDA datasets. Recall, accuracy,

TABLE 8 The top 20 predicted lncRNA biomarkers for neuroblastoma in each of the two datasets.

Dataset 1			Dataset 2		
Rank	lncRNA	Evidence	Rank	lncRNA	Evidence
1	HOTAIR	Unconfirmed	1	BDNF-AS	Unconfirmed
2	HNF1A-AS1	Unconfirmed	2	SNHG4	32614236
3	CDKN2B-AS1	Unconfirmed	3	BANCR	Unconfirmed
4	GAS5	28035057	4	HAR1A	Unconfirmed
5	CCAT1	Unconfirmed	5	HCP5	33189302
6	TUG1	Unconfirmed	6	TUG1	Unconfirmed
7	UCA1	Unconfirmed	7	HOTAIR	Unconfirmed
8	CRNDE	Unconfirmed	8	SRA1	Unconfirmed
9	WT1-AS	Unconfirmed	9	TERC	Unconfirmed
10	BANCR	Unconfirmed	10	SPRY4-IT1	Unconfirmed
11	WRAP53	Unconfirmed	11	KCNQ1OT1	31433907
12	SPRY4-IT1	Unconfirmed	12	IGF2-AS	30914706
13	CCAT2	33475889	13	PTENP1	Unconfirmed
14	CCDC26	Unconfirmed	14	CCAT1	Unconfirmed
15	PVT1	Unconfirmed	15	PCAT1	Unconfirmed
16	HULC	Unconfirmed	16	NPTN-IT1	Unconfirmed
17	CASC2	Unconfirmed	17	DGCR5	Unconfirmed
18	DANCR	34050113	18	HULC	Unconfirmed
19	KCNQ1OT1	31433907	19	BOK-AS1	Unconfirmed
20	7SK	Unconfirmed	20	BCYRN1	Unconfirmed

and F1-score computed by LightGBM were slightly better than LDAenDL on the two datasets. But it calculated the best AUC and AUPR on dataset 1.

Under CV3, as shown in Table 6, LDAenDL computed the highest precision, recall, accuracy, F1-score, AUC, and AUPR on the two LDA datasets except that it computed a slightly lower recall on dataset 1. The results demonstrate that LDAenDL is appropriate to predict possible LDAs from unknown lncRNA-disease pairs.

3.4 Case study

3.4.1 Identifying possible lncRNA biomarkers for lung cancer

Lung cancer is one of the most prevalent causes of mortality globally. It mainly contains small cell lung cancer and non-small cell lung cancer. Targeted drug therapy is its one therapeutic option (Lahiri et al., 2023). We used the proposed LDAenDL method to predict possible lncRNA biomarkers for lung cancer. Table 7 shows the predicted top 20 lncRNA biomarkers for lung cancer. The 20 lncRNA biomarkers associated with lung cancer have no known association information with lung cancer in the two datasets.

In dataset 1, LDAenDL predicted that CCDC26 could be associated with lung cancer. CCDC26 can enhance thyroid cancer malignant progression (Ma et al., 2021). It promotes imatinib resistance in human gastrointestinal stromal tumors (Yan et al., 2019). Its inhibition could increase the sensitivity of doxorubicin in MDR-CML cells (Liu et al., 2021b). In this study, we predicted that CCDC26 could be associated with lung cancer in dataset 1.

In dataset 2, LDAenDL predicted that IFNG-AS1 could be associated with lung cancer. IFNG-AS1 has been reported in long-lasting memory T cells (Castellucci et al., 2021). It can boost interferon gamma generation in human natural killer cells (Stein et al., 2019). We identified that IFNG-AS1 could be associated with lung cancer in Dataset 2.

Figure 5 shows the top 20 predicted lncRNAs associated with lung cancer in each of the two datasets. Yellow solid lines and blue solid lines denote lncRNA-lung cancer associations confirmed by the literatures among the predicted top 20 associations on datasets 1 and 2, respectively. Grey solid lines denote the predicted and co-occurring lncRNA-lung cancer associations that can be confirmed by the literatures in the two datasets, and grey dashed lines denote the predicted and unconfirmed lncRNA-lung cancer associations in the two datasets. The repeated lncRNAs in the two datasets have been removed.

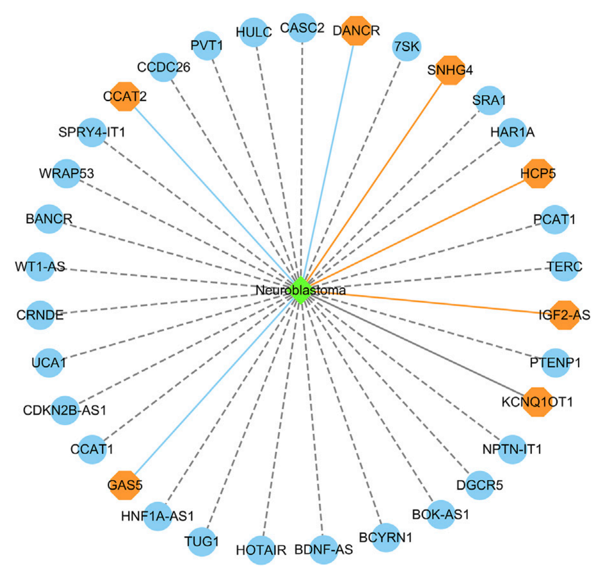


FIGURE 6
The top 20 predicted lncRNA biomarkers for neuroblastoma in each of the two datasets. (The repeated lncRNAs in the two datasets have been removed). This figure was drawn using Cytoscape (Shannon et al., 2003).

3.4.2 Identifying possible lncRNAs associated with PDL1 for lung cancer

Recent advances in lung cancer treatment have demonstrated significant responses in patients when they were treated with programmed death-1/programmed death-ligand 1 (PD-1/PD-L1) checkpoint blockade immunotherapies (Lahiri et al., 2023). To find possible lncRNAs associated with PDL1 for lung cancer, inspired by LPI-DLDN proposed by Peng et al. (2022a), we first downloaded the sequence of PDL1 from the UniProt database. Next, we extracted the biological features of PDL1 and depicted PDL1 as a 10,029-dimensional vector using BioTriangle. Finally, we used cosine similarity to compute the similarities between PDL1 and the other proteins in a lncRNA-protein interaction dataset (Li et al., 2015) and found the top 3 proteins with the highest interaction probabilities with PDL1. The results show that SNHG3 has a higher interaction probability with PDL1 and has been reported to be associated with lung cancer.

3.4.3 Identifying possible lncRNA biomarkers for neuroblastoma

Neuroblastoma is the most frequent pediatric solid tumor and accounts for approximately 15% of childhood cancer-related mortality (Zafar et al., 2021). We used the proposed LDAenDL method to identify possible lncRNA biomarkers for neuroblastoma. Table 8 shows the top 20 predicted lncRNA biomarkers for neuroblastoma in each of the two datasets. The repeated lncRNAs in the two datasets have been removed.

In dataset 1, we predicted that HOTAIR could be associated with neuroblastoma with the highest probability. HOTAIR is a novel oncogenic biomarker in human cancer (Rajagopal et al., 2020). Its knockdown can promote radiosensitivity in colorectal cancer (Liu et al., 2020). It also can enhance the carcinogenesis of gastric (Zhang et al., 2020). We identified that HOTAIR may be one biomarker of neuroblastoma in dataset 1.

In dataset 2, we predicted that BDNF-AS could be associated with neuroblastoma with the highest probability. PABPC1-induced stabilization of BDNF-AS helps the inhibition of malignant progression in glioblastoma cells (Su et al., 2020). It can regulate the miR-9-5p/BACE1 pathway that affects neurotoxicity in Alzheimer's disease (Ding et al., 2022). We identified that BDNF-AS is a possible biomarker of neuroblastoma in dataset 2.

Figure 6 shows the top 20 predicted lncRNAs associated with neuroblastoma in each of the two datasets. Yellow solid lines and blue solid lines denote lncRNA-neuroblastoma associations confirmed by the literatures among the predicted top 20 associations on datasets 1 and 2, respectively. Grey solid lines denote the predicted and co-occuring lncRNA-neuroblastoma associations that can be confirmed by the literatures in the two datasets, and grey dashed lines denote the predicted and unconfirmed lncRNA-neuroblastoma associations in the two datasets. The repeated lncRNAs in the two datasets have been removed.

4 Conclusion

Lung cancer and neuroblastoma are two human diseases that severely affect the human body. Detecting new biomarkers for them contributes to their diagnosis and therapy. Experimental biomarker identification methods are costly and laborious. Thus, we developed a machine learning-based method named LDAenDL to predict possible lncRNA biomarkers for the two diseases based on an ensemble of a deep neural network and LightGBM. LDAenDL first computed lncRNA similarity and disease similarity and then combined a GCN, GAT, and CNN to learn the biological features of lncRNAs and diseases. Finally, these features were fed to a DNN and LightGBM to find new LDAs.

LDAenDL was compared with the other four classical LDA prediction methods (i.e., SDLDA, LDNFSGB, IPCAF, and LDASR). The results showed that LDAenDL computed the best AUCs and AUPRs under three cross-validations on two LDA datasets, demonstrating the optimal LDA prediction performance of LDAenDL. We further identified possible lncRNA biomarkers for lung cancer and neuroblastoma. The results demonstrated that CCDC26 and IFNG-AS1 may be new biomarkers for lung cancer, SNHG3 may be associated with PDL1 for lung cancer, and HOTAIR and BDNF-AS may be potential biomarkers for neuroblastoma.

In the future, we will combine data from multiple sources, for example, miRNA, circRNA, and drugs, to improve LDA identification performance. We will also design a new deep-learning model to efficiently extract the biological features of lncRNAs and diseases for LDA prediction. We hope that the proposed LDAenDL can help the development of targeted therapies for these two diseases.

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

Conceptualization: ZS, HL, ZL, and LD; Investigation: ZS and HL; Methodology: ZS, HL, ZL, and LD; Project administration: YW and LD; Software: ZS and ZL; Writing-original draft: ZS and HL; Writing-review and editing: ZS, HL, ZL, and LD. All authors contributed to the article and approved the submitted version.

Conflict of interest

Author YW was employed by Geneis (Beijing) Co., Ltd.

The remaining 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.

References

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science* 306 (5705), 2242–2246. doi:10.1126/science.1103388

Broadbent, H. M., Peden, J. F., Lorkowski, S., Goel, A., Ongen, H., Green, F., et al. (2008). Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. *Hum. Mol. Genet.* 17 (6), 806–814. doi:10.1093/hmg/ddm352

Castellucci, L. C., Almeida, L., Cherlin, S., Fakiola, M., Francis, R. W., Carvalho, E. M., et al. (2021). A genome-wide association study identifies SERPINB10, CRLF3, STX7, LAMP3, IFNG-AS1, and KRT80 as risk loci contributing to cutaneous leishmaniasis in Brazil. *Clin. Infect. Dis.* 72 (10), e515–e525. doi:10.1093/cid/ciaa1230

Chakravarty, D., Sboner, A., Nair, S. S., Giannopoulou, E., Li, R., Hennig, S., et al. (2014). The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. *Nat. Commun.* 5 (1), 5383. doi:10.1038/ncomms6383

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2012). LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic acids Res.* 41 (D1), D983–D986. doi:10.1093/nar/gks1099

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., et al. (2018). MNDR v2. 0: an updated resource of ncRNA–disease associations in mammals. *Nucleic acids Res.* 46 (D1), D371–D374. doi:10.1093/nar/gkx1025

Dai, Q., Liu, Z., Wang, Z., Duan, X., and Guo, M. (2022). GraphCDA: a hybrid graph representation learning framework based on GCN and GAT for predicting disease associated circRNAs. *Briefings in Bioinformatics* 23 (5), bbac379. doi:10.1093/bib/bbac379

Ding, Y., Luan, W., Wang, Z., and Cao, Y. (2022). LncRNA BDNF-AS as ceRNA regulates the miR-9-5p/BACE1 pathway affecting neurotoxicity in Alzheimer's disease. *Archives Gerontology Geriatrics* 99, 104614. doi:10.1016/j.archger.2021.104614

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognit.* 77, 354–377. doi:10.1016/j.patcog.2017.10.013

He, X., Tan, X., Wang, X., Jin, H., Liu, L., Ma, L., et al. (2014). C-Myc-activated long noncoding RNA CCAT1 promotes colon cancer cell proliferation and invasion. *Tumor Biol.* 35, 12181–12188. doi:10.1007/s13277-014-2526-4

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: a highly efficient gradient boosting decision tree. *Adv. neural Inf. Process. Syst.* 30. doi:10.5555/3294996.3295074

Kipf, T. N., and Welling, M. (2016). *Semi-supervised classification with graph convolutional networks. arXiv preprint* arXiv:1609.02907.

Klattenhoff, C. A., Scheuermann, J. C., Surface, L. E., Bradley, R. K., Fields, P. A., Steinhauser, M. L., et al. (2013). Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell.* 152 (3), 570–583. doi:10.1016/j.cell.2013.01.003

Lahiri, A., Maji, A., Potdar, P. D., Singh, N., Parikh, P., Bisht, B., et al. (2023). Lung cancer immunotherapy: progress, pitfalls, and promises. *Mol. Cancer* 22 (1), 40–37. doi:10.1186/s12943-023-01740-y

Lanjanian, H., Nematzadeh, S., Hosseini, S., Torkamanian-Afshar, M., Kiani, F., Moazzam-Jazi, M., et al. (2021). High-throughput analysis of the interactions between viral proteins and host cell RNAs. *Comput. Biol. Med.* 135, 104611. doi:10.1016/j.compbiomed.2021.104611

Li, A., Ge, M., Zhang, Y., Peng, C., and Wang, M. (2015). Predicting long noncoding RNA and protein interactions using heterogeneous network model. *BioMed Res. Int.* 2015, 671950. doi:10.1155/2015/671950

Li, J., Li, J., Kong, M., Wang, D., Fu, K., and Shi, J. (2021). Svdnvlda: predicting lncRNA-disease associations by singular value decomposition and node2vec. *BMC Bioinforma.* 22, 538. doi:10.1186/s12859-021-04457-1

Liang, Y., Zhang, Z. Q., Liu, N. N., Wu, Y. N., Gu, C. L., and Wang, Y. L. (2022). Magcnse: predicting lncRNA-disease associations using multi-view attention graph convolutional network and stacking ensemble model. *BMC Bioinforma.* 23 (1), 189. doi:10.1186/s12859-022-04715-w

Liu, Y., Chen, X., Chen, X., Liu, J., Gu, H., Fan, R., et al. (2020). Long non-coding RNA HOTAIR knockdown enhances radiosensitivity through regulating microRNA-93/ATG12 axis in colorectal cancer. *Cell. Death Dis.* 11 (3), 175. doi:10.1038/s41419-020-2268-8

Liu, J. X., Gao, M. M., Cui, Z., Gao, Y. L., and Li, F. (2021a). Dscmf: prediction of LncRNA-disease associations based on dual sparse collaborative matrix factorization. *BMC Bioinforma.* 22 (3), 241. doi:10.1186/s12859-020-03868-w

Liu, Z., Wang, Y., Xu, Z., Yuan, S., Ou, Y., Luo, Z., et al. (2021b). Analysis of ceRNA networks and identification of potential drug targets for drug-resistant leukemia cell K562/ADR. *PeerJ* 9, e11429. doi:10.7717/peerj.11429

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., et al. (2022). Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder. *Briefings Bioinforma.* 23 (3), bbac104. doi:10.1093/bib/bbac104

Liu, W., Yang, Y., Lu, X., Fu, X., Sun, R., Yang, L., et al. (2023a). Nsrgrn: a network structure refinement method for gene regulatory network inference. *Briefings Bioinforma.* 24 (3), bbad129. doi:10.1093/bib/bbad129

Liu, W., Tang, T., Lu, X., Fu, X., Yang, Y., and Peng, L. (2023b). Mpclcda: predicting circRNA–disease associations by using automatically selected meta-path and contrastive learning. *Briefings Bioinforma.* 24, bbad227. doi:10.1093/bib/bbad227

Ma, X., Li, Y., Song, Y., and Xu, G. (2021). Long noncoding RNA CCDC26 promotes thyroid cancer malignant progression via miR-422a/EZH2/Sirt6 axis. *OncoTargets Ther.* 14, 3083–3094. doi:10.2147/OTT.S282011

Ma, Y. (2022). Deepmne: deep multi-network embedding for lncRNA-disease association prediction. *IEEE J. Biomed. Health Inf.* 26 (7), 3539–3549. doi:10.1109/JBHI.2022.3152619

Meng, J., Kang, Q., Chang, Z., and Luan, Y. (2021). PlncRNA-HDeep: plant long noncoding RNA prediction using hybrid deep learning based on two encoding styles. *BMC Bioinforma.* 22 (3), 242. doi:10.1186/s12859-020-03870-2

Pasmant, E., Sabbagh, A., Vidaud, M., and Bièche, I. (2011). ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* 25 (2), 444–448. doi:10.1096/fj.10-172452

Peng, L., Huang, L., Lu, Y., Liu, G., Chen, M., and Han, G. (2022a). "Identifying possible lncRNA-disease associations based on deep learning and positive-unlabeled learning," in *2022 IEEE international conference on bioinformatics and biomedicine (BIBM)* (IEEE), 168–173.

Peng, L., Tan, J., Tian, X., and Zhou, L. (2022b). EnANNDeep: an ensemble-based lncRNA–protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models. *Interdiscip. Sci. Comput. Life Sci.* 14 (1), 209–232. doi:10.1007/s12539-021-00483-y

Peng, L., Wang, C., Tian, X., Zhou, L., and Li, K. (2022c). Finding lncrna-protein interactions based on deep learning with dual-net neural architecture. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 19 (6), 3456–3468. doi:10.1109/TCBB.2021.3116232

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022d). Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. *Briefings Bioinforma.* 23 (4), bbac234. doi:10.1093/bib/bbac234

Peng, L., Tan, J., Xiong, W., Zhang, L., Wang, Z., Yuan, R., et al. (2023a). Deciphering ligand–receptor-mediated intercellular communication based on ensemble deep learning and the joint scoring strategy from single-cell transcriptomic data. *Comput. Biol. Med.* 16 (2023), 107137. doi:10.1016/j.compbiomed.2023.107137

Peng, L., Yuan, R., Han, C., Han, G., Tan, J., Wang, Z., et al. (2023b). CellEnBoost: a boosting-based ligand-receptor interaction identification model for cell-to-cell communication inference. *IEEE Trans. NanoBioscience*, 1–11. doi:10.1109/TNB.2023.3278685

Rajagopal, T., Talluri, S., Akshaya, R. L., and Dunna, N. R. (2020). HOTAIR LncRNA: a novel oncogenic propellant in human cancer. *Clin. Chim. acta* 503, 1–18. doi:10.1016/j.cca.2019.12.028

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (2022). VDA-RWLRLS: an anti-SARS-CoV-2 drug prioritizing framework combining an unbalanced bi-random walk and Laplacian regularized least squares. *Comput. Biol. Med.* 140, 105119. doi:10.1016/j.compbiomed.2021.105119

Stein, N., Berhani, O., Schmiedel, D., Duev-Cohen, A., Seidel, E., Kol, I., et al. (2019). IFNG-AS1 enhances interferon gamma production in human natural killer cells. *Iscience* 11, 466–473. doi:10.1016/j.isci.2018.12.034

Su, R., Ma, J., Zheng, J., Liu, X., Liu, Y., Ruan, X., et al. (2020). PABPC1-induced stabilization of BDNF-AS inhibits malignant progression of glioblastoma cells through STAU1-mediated decay. *Cell. Death Dis.* 11 (2), 81. doi:10.1038/s41419-020-2267-9

Sun, X., Liu, M., and Sima, Z. (2020). A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Res. Lett.* 32, 101084. doi:10.1016/j.frl.2018.12.032

Tan, L., Yu, J. T., Hu, N., and Tan, L. (2013). Non-coding RNAs in Alzheimer's disease. *Mol. Neurobiol.* 47, 382–393. doi:10.1007/s12035-012-8359-5

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). *Graph attention networks*. arXiv preprint arXiv:1710.10903.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *stat* 1050 (20), 10–48550. doi:10.48550/arXiv.1710.10903

Wang, L., Cai, Y., Zhao, X., Jia, X., Zhang, J., Liu, J., et al. (2015). Down-regulated long non-coding RNA H19 inhibits carcinogenesis of renal cell carcinoma. *Neoplasma* 62 (3), 412–418. doi:10.4149/neo_2015_049

Wang, M. N., You, Z. H., Wang, L., Li, L. P., and Zheng, K. (2021). Ldgrnmf: lncRNA-disease associations prediction based on graph regularized non-negative matrix factorization. *Neurocomputing* 424, 236–245. doi:10.1016/j.neucom.2020.02.062

Wang, L., Shang, M., Dai, Q., and He, P. A. (2022). Prediction of lncRNA-disease association based on a Laplace normalized random walk with restart algorithm on heterogeneous networks. *BMC Bioinforma.* 23 (1), 5–20. doi:10.1186/s12859-021-04538-1

Wu, H., Liang, Q., Zhang, W., Zou, Q., Hesham, A. E. L., and Liu, B. (2022). iLncDA-LTR: identification of lncRNA-disease associations by learning to rank. *Comput. Biol. Med.* 146, 105605. doi:10.1016/j.compbiomed.2022.105605

Xie, G., Huang, B., Sun, Y., Wu, C., and Han, Y. (2021). RWSF-BLP: a novel lncRNA-disease association prediction model using random walk-based multi-similarity fusion and bidirectional label propagation. *Mol. Genet. Genomics* 296, 473–483. doi:10.1007/s00438-021-01764-3

Xie, G., Zhu, Y., Lin, Z., Sun, Y., Gu, G., Li, J., et al. (2022). Hbrwrlda: predicting potential lncRNA–disease associations based on hypergraph bi-random walk with restart. *Mol. Genet. Genomics* 297 (5), 1215–1228. doi:10.1007/s00438-022-01909-y

Xu, J., Xu, J., Meng, Y., Lu, C., Cai, L., Zeng, X., et al. (2023). Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. *Cell. Rep. Methods* 3, 100382. doi:10.1016/j.crmeth.2022.100382

Yan, J., Chen, D., Chen, X., Sun, X., Dong, Q., Hu, C., et al. (2019). Downregulation of lncRNA CCDC26 contributes to imatinib resistance in human gastrointestinal stromal tumors through IGF-1R upregulation. *Braz. J. Med. Biol. Res.* 52, e8399. doi:10.1590/1414-431x20198399

Yang, Q., and Li, X. (2021). BiGAN: lncRNA-disease association prediction based on bidirectional generative adversarial network. *BMC Bioinforma.* 22, 357. doi:10.1186/s12859-021-04273-7

Yang, M., Zhao, L., Hu, X., Feng, H., and Kang, X. (2021). Identification of key mRNAs and lncRNAs associated with the effects of anti-TWEAK on osteosarcoma. *Curr. Bioinforma.* 16 (1), 154–161. doi:10.2174/1574893615999200626191405

Yao, Y., Ji, B., Lv, Y., Li, L., Xiang, J., Liao, B., et al. (2021). Predicting LncRNA–disease association by a random walk with restart on multiplex and heterogeneous networks. *Front. Genet.* 12, 712170. doi:10.3389/fgene.2021.712170

Zafar, A., Wang, W., Liu, G., Wang, X., Xian, W., McKeon, F., et al. (2021). Molecular targeting therapies for neuroblastoma: progress and challenges. *Med. Res. Rev.* 41 (2), 961–1021. doi:10.1002/med.21750

Zhang, E. B., Yin, D. D., Sun, M., Kong, R., Liu, X. H., You, L. H., et al. (2014). P53-regulated long non-coding RNA TUG1 affects cell proliferation in human non-small cell lung cancer, partly through epigenetically regulating HOXB7 expression. *Cell. death Dis.* 5 (5), e1243. doi:10.1038/cddis.2014.201

Zhang, J., Qiu, W. Q., Zhu, H., Liu, H., Sun, J. H., Chen, Y., et al. (2020). HOTAIR contributes to the carcinogenesis of gastric cancer via modulating cellular and exosomal miRNAs level. *Cell. death Dis.* 11 (9), 780. doi:10.1038/s41419-020-02946-4

Zhao, X., Zhao, X., and Yin, M. (2022). Heterogeneous graph attention network based on meta-paths for lncrna–disease association prediction. *Briefings Bioinforma.* 23 (1), bbab407. doi:10.1093/bib/bbab407

Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021). LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncRNA–protein interaction identification. *BMC Bioinforma.* 22 (1), 479. doi:10.1186/s12859-021-04399-8

Zhou, Y., Wang, X., Yao, L., and Zhu, M. (2022). LDAformer: predicting lncRNA-disease associations based on topological feature extraction and transformer encoder. *Briefings Bioinforma.* 23 (6), bbac370. doi:10.1093/bib/bbac370

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 10 November 2023
DOI 10.3389/fgene.2023.1237167

Check for updates

OPEN ACCESS

EDITED BY
Taoyang Wu,
University of East Anglia, United Kingdom

REVIEWED BY
Wei Shen,
Chongqing Medical University, China
Haiying Zou,
Shantou University, China
Saeid Ghorbian,
Islamic Azad University of Ahar, Iran

*CORRESPONDENCE
Haopeng Yu,
✉ yuhaopeng@wchscu.cn
Yong Yuan,
✉ yongyuan@scu.edu.cn

RECEIVED 08 June 2023
ACCEPTED 20 October 2023
PUBLISHED 10 November 2023

CITATION
Chen Y, Kuang Y, Luan S, Yang Y, Ying Z, Li C, Gao J, Yuan Y and Yu H (2023), DASES: a database of alternative splicing for esophageal squamous cell carcinoma. *Front. Genet.* 14:1237167. doi: 10.3389/fgene.2023.1237167

DASES: a database of alternative splicing for esophageal squamous cell carcinoma

Yilong Chen[1,2], Yalan Kuang[1,2], Siyuan Luan[1], Yongsan Yang[1,2], Zhiye Ying[1,2], Chunyang Li[1,2], Jinhang Gao[3,4], Yong Yuan[1]* and Haopeng Yu[1,2]*

[1]Department of Thoracic Surgery and West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China, [2]Med-X Center for Informatics, Sichuan University, Chengdu, China, [3]Department of Gastroenterology, West China Hospital, Sichuan University, Chengdu, China, [4]Laboratory of Gastroenterology and Hepatology, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, China

Esophageal carcinoma ranks as the sixth leading cause of cancer-related mortality globally, with esophageal squamous cell carcinoma (ESCC) being particularly prevalent among Asian populations. Alternative splicing (AS) plays a pivotal role in ESCC development and progression by generating diverse transcript isoforms. However, the current landscape lacks a specialized database focusing on alternative splicing events (ASEs) derived from a large number of ESCC cases. Additionally, most existing AS databases overlook the contribution of long non-coding RNAs (lncRNAs) in ESCC molecular mechanisms, predominantly focusing on mRNA-based ASE identification. To address these limitations, we deployed DASES (http://www.hxdsjzx.cn/DASES). Employing a combination of publicly available and in-house ESCC RNA-seq datasets, our extensive analysis of 346 samples, with 93% being paired tumor and adjacent non-tumor tissues, led to the identification of 257 novel lncRNAs in esophageal squamous cell carcinoma. Leveraging a paired comparison of tumor and adjacent normal tissues, DASES identified 59,094 ASEs that may be associated with ESCC. DASES fills a critical gap by providing comprehensive insights into ASEs in ESCC, encompassing lncRNAs and mRNA, thus facilitating a deeper understanding of ESCC molecular mechanisms and serving as a valuable resource for ESCC research communities.

KEYWORDS

esophageal squamous cell carcinoma, alternative splicing, database, novel lncRNA, isoform

1 Introduction

Esophageal carcinoma (EC), a type of malignant tumor affecting the esophagus, is a major global health concern with an estimated annual incidence of over 600,000 and mortality of over 500,000, making it the seventh most common malignant tumor and the sixth leading cause of cancer-related death globally (Sung et al., 2021). There are significant regional differences in the incidence of EC, which can be divided into esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC), according to the pathological type, with nearly 79% of ESCC occurring in Asian countries (Morgan et al., 2022). Although the incidence of ESCC has shown a decreasing trend in certain countries (Liang et al., 2017), ESCC continues to be a pressing public health issue on account of its increased fatality rate (Abnet et al., 2018). Majority of ESCC patients present at an advanced stage during medical consultation, and conventional surgical interventions often exhibit suboptimal effectiveness or even fail to achieve a radical resection in some cases (He et al., 2022; Pape et al., 2022), with a 5-year survival rate of less than 30% (Allemani et al., 2018). As tumor molecular biology and immune escape mechanisms are more thoroughly studied, a growing number of targeted and immune drugs are being investigated as potential treatments to prolong the survival time of patients with ESCC (Kojima et al., 2020; Costoya and Arce, 2023).

Alternative splicing (AS) is a post-transcriptional regulatory process that generates various RNA isoforms by employing diverse splicing patterns, thereby playing a pivotal role in regulating protein production, especially during developmental and differentiation processes (Yang et al., 2016; Bonnal et al., 2020). When AS is not properly regulated, it can result in the production of oncogenic isoforms, which can contribute to the growth and progression of tumors (Zhang et al., 2021). ESCC patients exhibit a high frequency of alternative splicing events (ASEs), which are associated with tumor initiation, progression, invasion, and immune evasion (Dlamini et al., 2021; Wu et al., 2021). Meanwhile, AS has potential importance in the treatment of ESCC, and several studies suggest that intervention in AS can enhance the sensitivity of ESCC cells to chemotherapy drugs (Siegfried and Karni, 2018; Sciarrillo et al., 2020). Additionally, AS has been shown to impact the efficacy of immunotherapy for ESCC by influencing the expression and presentation of tumor antigens, ultimately affecting the recognition and attack of tumor cells by immune cells (Duan et al., 2021; Wu et al., 2021). Thus, AS has important implications and value for a deeper understanding of the molecular mechanisms of ESCC and the development of therapeutic and immunotherapeutic strategies.

Currently, several databases are available that encompass ASEs, including some that cover ESCC, such as TCGASpliceSeq (Deng et al., 2021) and OncoSplicing (Zhang et al., 2022), developed based on data from The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015). However, despite the inclusion of multiple cancer types, the number of ESCC cases in these databases is limited, with only 96 cases available (Yang et al., 2023). Furthermore, most of these databases primarily rely on oligo dT and poly A sequencing techniques, focusing on AS identification in protein-coding genes, with limited attention given to AS events involving long non-coding RNAs (lncRNAs). In contrast, although ESCC-specific databases, such as ESCCdb (Yang et al., 2023) and CCGD-ESCC (Peng et al., 2018), encompass a larger number of cases, they lack the specific annotation of ASEs. Considering the significant relationship between lncRNA expression and ESCC development and progression (Li et al., 2019b; Razavi and Ghorbian, 2019; Sadeghpour and Ghorbian, 2019; Aalijahan and Ghorbian, 2020; Liu et al., 2020; Ghasemzadeh and Ghorbian, 2023) and the absence of specialized AS-related databases for ESCC, we developed the Database of Alternative Splicing for Esophageal Squamous cell carcinoma (DASES) (http://www.hxdsjzx.cn/DASES), which utilizes two main sets of data. The first set consists of our in-house total transcriptome sequencing data, derived from ESCC patients at the West China Hospital of Sichuan University. The second set is total transcriptome sequencing data from 11 published projects related to ESCC. Through the integration of known transcripts, the identification of novel lncRNAs, and the paired comparison of isoforms between tumor and adjacent normal tissues, DASES provides a comprehensive and precise catalog of ASEs in ESCC, filling a critical gap in the field and offering a valuable resource for ESCC research communities.

2 Materials and methods

2.1 Data collection

DASES contains raw data from two sources. The first source includes total RNA sequencing data on both tumor and adjacent normal tissues from 63 ESCC patients in West China Hospital of Sichuan University. The second source includes publicly available total RNA sequencing data on ESCC patients from the European Bioinformatics Institute (EBI). To ensure high-quality data, we employed strict search criteria to select suitable samples from EBI (Figure 1): 1) the samples were obtained from human ESCC tissues; 2) the data included RNA sequencing; and 3) the data had sufficient information available. We excluded the cell line RNA-seq data, RNA-seq data from esophageal adenocarcinoma or other parts of the esophagus, and any data without sufficient information. It is essential to emphasize that the included datasets were not specifically targeted or enriched for circular RNA (circRNA) or small RNA during the sequencing and library preparation processes.

2.2 Data quality control and lncRNA identification

In this study, we used a series of bioinformatics tools to identify potential lncRNAs and mRNAs associated with ESCC (Figure 2). First, the raw reads obtained from RNA-seq were subjected to quality control using Trim Galore software (version 0.6.4; https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) to obtain clean reads. Then, we used STAR (version 2.7.3a) (Dobin et al., 2013) and HISAT2 software programs (version 2.2.1) (Kim et al., 2019) for sequence alignment of the clean data, resulting in the generation of BAM and SAM files for each sample, respectively. Next, we used Cufflinks (version 2.2.0) (Trapnell et al., 2010) and StringTie software programs (version 2.1.4) (Pertea et al., 2015) to assemble the BAM and SAM files, respectively, generating GTF files for each sample. We then used StringTie software to merge all the assembled GTF files, obtaining a preliminary merged GTF file. Subsequently, we utilized GffCompare software (version 0.12.2) (Pertea and Pertea, 2020) and reference transcripts to identify potential lncRNAs and mRNAs. We selected transcripts with class code "i" or "u" as potential lncRNA candidates and those with class code "=," "c," or "j" as mRNA candidates. Finally, we predicted the lncRNA candidates using CPAT (version 3.0.2) (Wang et al., 2013) and PLEK software programs (version 1.2) (Li et al., 2014), and selected those predicted as non-coding RNAs by both tools as lncRNA candidates. We merged the lncRNA and mRNA candidates to generate a comprehensive GTF file containing all potential lncRNA and mRNA candidates associated with ESCC. All the analyses were conducted using the human genome hg38 (release 84) reference provided by Ensembl (https://ensembl.org/Homo_sapiens/Info/Index).

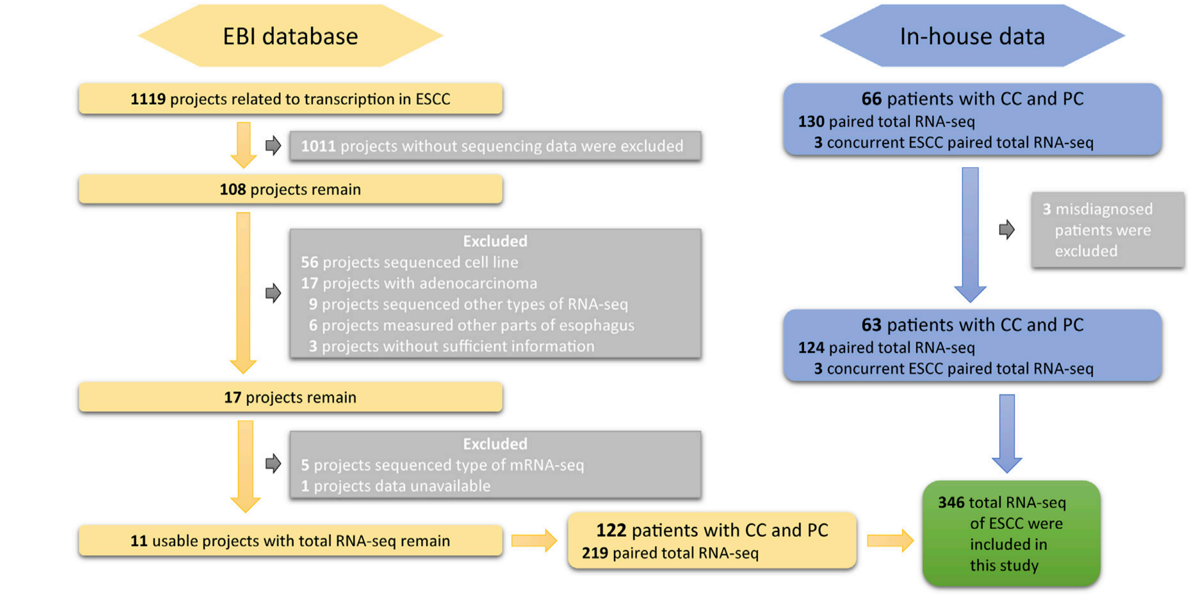

EBI database
1119 projects related to transcription in ESCC
1011 projects without sequencing data were excluded
108 projects remain
Excluded
56 projects sequenced cell line
17 projects with adenocarcinoma
9 projects sequenced other types of RNA-seq
6 projects measured other parts of esophagus
3 projects without sufficient information
17 projects remain
Excluded
5 projects sequenced type of mRNA-seq
1 projects data unavailable
11 usable projects with total RNA-seq remain
122 patients with CC and PC
219 paired total RNA-seq
In-house data
66 patients with CC and PC
130 paired total RNA-seq
3 concurrent ESCC paired total RNA-seq
3 misdiagnosed patients were excluded
63 patients with CC and PC
124 paired total RNA-seq
3 concurrent ESCC paired total RNA-seq
346 total RNA-seq of ESCC were included in this study

FIGURE 1
Overview of the screening and inclusion criteria for ESCC patient transcriptome data in the study.

2.3 Alternative splicing event identification

ASEs can manifest in different ways, including skipping an exon (SE), including or excluding a mutually exclusive exon (MXE), using alternative 5′ or 3′ splice sites (A5SS or A3SS), or retaining an intron (RI). To determine the occurrence of ASEs, we compared two different transcript isoforms derived from the same gene. Specifically, we performed paired comparisons between tumor and adjacent normal groups. In these comparisons, we assigned the term "included isoform" to the isoform containing exons when comparing two transcripts. Conversely, the isoform lacking exons was referred to as the "excluded isoform." The designation of the "included isoform" was based on having a shorter intron length, whereas the "excluded isoform" had a longer intron length (Figure 3). By comparing the splice junctions and exon–intron boundaries between these two isoforms, we identified and quantified the specific ASEs present in the transcriptome.

To comprehensively identify ASEs associated with ESCC, we employed rMATS software (version 3.1.0) (Shen et al., 2014) with a stringent splicing difference cutoff of 0.0001. Given the publicly available data literature reports, which indicated that all the whole-transcriptome data utilized dUTP-based library construction techniques, we considered the fr-firststrand library type during the analysis of aligned reads in BAM format. By comparing exon inclusion levels between tumor and adjacent normal groups, we detected differential ESCC-related ASEs. We only retained ASEs with a percent spliced in (PSI) (Katz et al., 2010) value greater than 0 and that were present in at least two samples. The results of splicing with only reads that span splicing junction based on GTF files were selected as the ESCC-related ASEs. Furthermore, to establish the coordinates of ASEs, we considered that each ASE comprises two transcript isoforms, with each isoform potentially containing 0–2 introns. In order to define the boundaries of ASE, we determined the minimum coordinate of the intron within the event as the starting coordinate and the maximum coordinate of the intron as the ending coordinate.

2.4 Expression quantitative analysis

For quantitative analysis of the RNA-seq data, we employed the merged GTF file as the reference annotation. The BAM files, generated from the alignment step, were subjected to subsequent analysis using Cuffnorm (version 2.2.0), which is a part of Cufflinks software. This tool allowed us to estimate the expression levels of individual isoforms, providing fragments per kilobase of transcript per million mapped reads (FPKM) values.

2.5 Potential affected the protein domain by alternative splicing

To assess the potential overlap between ASEs and protein domains, we adopted a conservative approach. We focused only on the known protein domains that intersected with ASEs, disregarding the predictions from various tools. Initially, we retrieved protein domain information from the Ensembl database, specifically focusing on the hg38 version of protein domain annotations (release 109), which includes InterPro coordinates, associated transcripts, and corresponding genes. We then mapped the InterPro coordinates onto genomic coordinates using appropriate alignment algorithms as follows:

$$Start_{genomic} = Start_{CDS} + 3 \times Start_{InterPro} - 3,$$
$$End_{genomic} = Start_{CDS} + 3 \times End_{InterPro} - 1,$$

where $Start_{genomic}$ and $End_{genomic}$ represent the start and end sites of the protein domain on the genome, respectively, $Start_{CDS}$ refers to

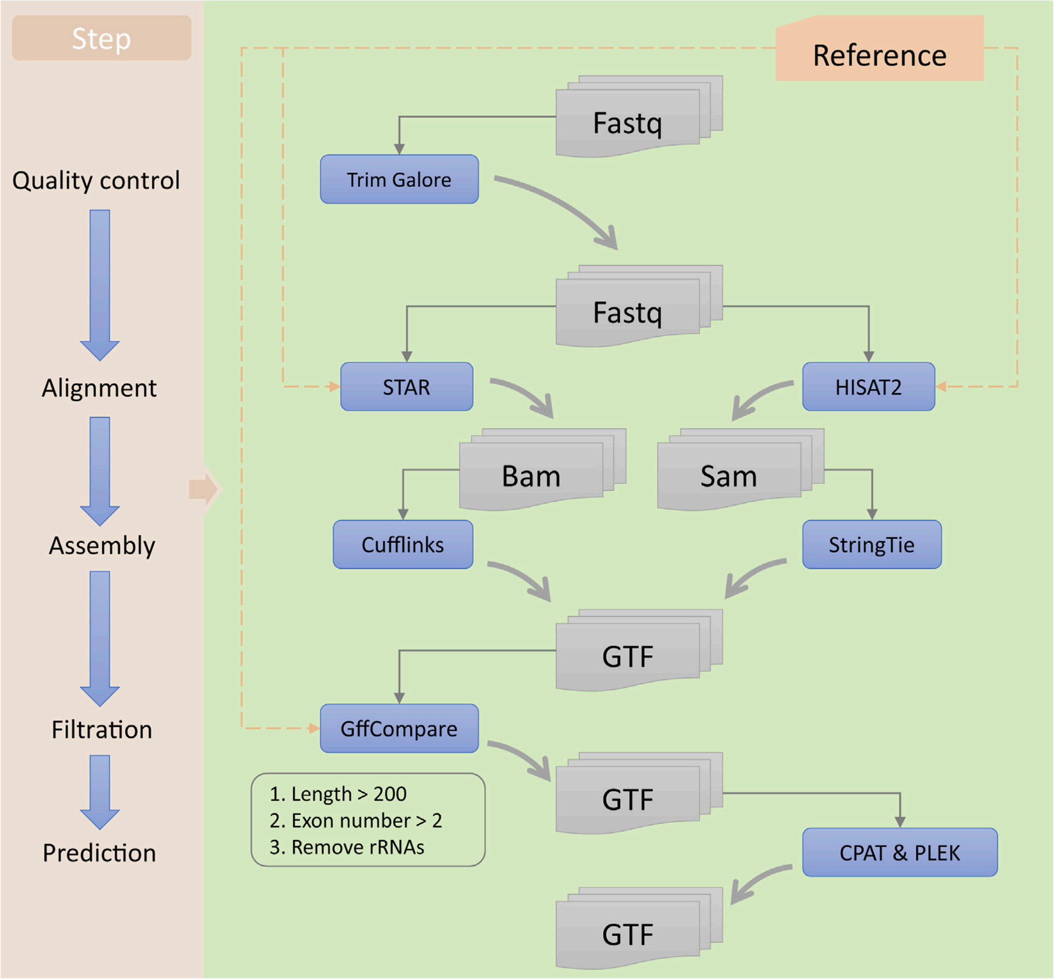


FIGURE 2
Workflow for the identification of lncRNAs. The left panel represents each step of the identification process, while the right panel includes the tools used and the types of input and output files for each step.

the first CDS start site of the corresponding transcript, and $Start_{InterPro}$ and $End_{InterPro}$ indicate the start and end sites of the protein domain on InterPro coordinates, respectively. Subsequently, we scrutinized whether there was any intersection or overlap between the genomic coordinates of protein domains and genomic coordinates of ASEs. In the cases where a protein domain exhibited any intersection or overlap with an ASE, we deemed it as having a significant overlap with the respective ASE.

2.6 Deployment of DASES

DASES is readily accessible through its website at http://www.hxdsjzx.cn/DASES, and no registration or login is required for usage. The current version of DASES was deployed utilizing MySQL (version 8.0.18) (http://www.mysql.com) and operates on a Linux-based Aliyun web server. Server-side scripting was implemented using Tomcat (version 8.0) (http://tomcat.apache.org/) and JAVA (version 1.8) (https://www.oracle.com/technetwork/java/index.html), providing the necessary functionality. The user-friendly web interface of DASES was created using Bootstrap (version 3.3.7) (https://v3.bootcss.com) and jQuery (version 2.1.1) (http://jquery.com) for seamless interaction and enhanced user experience. Genomic visualization capabilities were achieved using JBrowse (http://jbrowse.org) and IGV (https://igv.org), while additional visualizations were facilitated by ECharts (https://echarts.apache.org/zh/index.html). The web interface of DASES comprises various modules, including Home, Search, Browse, Genome Browser, Download, and About, ensuring comprehensive and intuitive access to the platform's features and information.

3 Results

3.1 Data and database overview

Following the application of quality control measures, a total of 14 patients, corresponding to 28 samples, were eliminated from the dataset. Currently, DASES encompasses data from the in-house

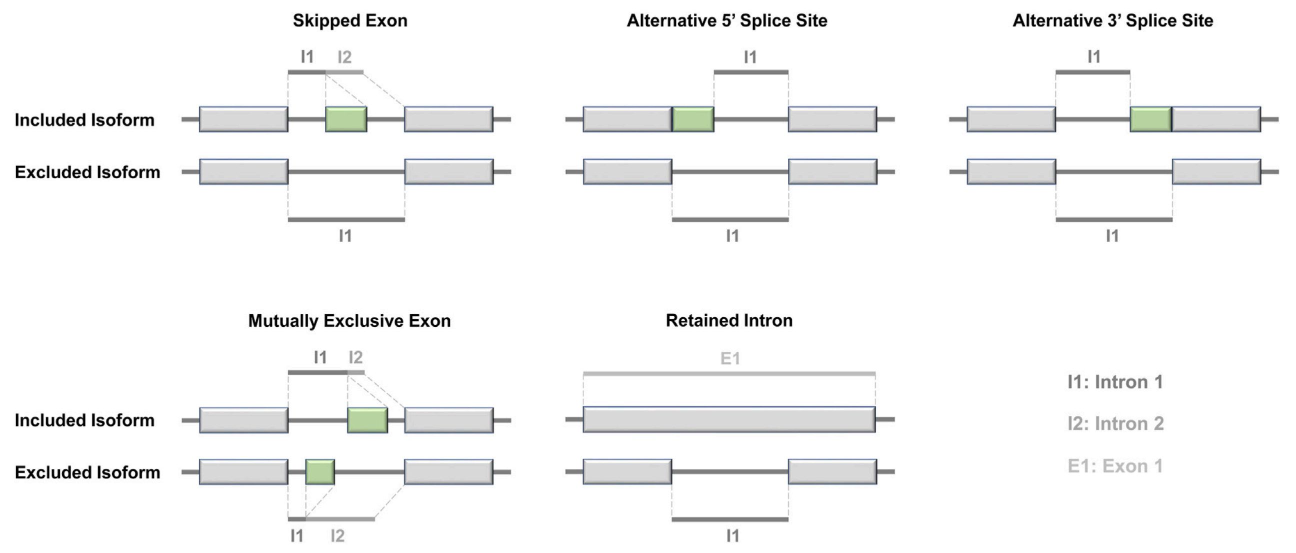


FIGURE 3
Classification of alternative splicing events.

TABLE 1 Information on whole-transcriptome data in ESCC patients from one in-house study and 11 publicly available studies.

Study accession	Number of samples (tumor: adjacent)[a]	Geographic position	Layout	Sequencing library	Data accession
PRJCA017448	64:63	China	Paired	dUTP	https://ngdc.cncb.ac.cn/search/?dbId=hra&q=PRJCA017448
PRJNA793370	3:3	China	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/793370
PRJNA843947	6:6	China	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/843947
PRJNA784605	4:4	China	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/784605
PRJNA665149	18:18	China	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/665149
PRJNA689307	8:8	China	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/689307
PRJNA629358	10:10	China	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/629358
PRJNA594797	3:3	China	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/594797
PRJNA608223	0:25	Kazakhstan	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/608223
PRJNA533799	23:23	Korea	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/533799
PRJNA435587	7:7	China	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/435587
PRJNA298963	15:15	China	Paired	dUTP	https://www.ncbi.nlm.nih.gov/bioproject/298963

[a]The number of samples on tumor tissues versus the number of samples on adjacent normal tissues for ESCC patients in each study.

study and 11 publicly available studies (Table 1), comprising a total of 346 samples, with 185 distinct ESCC patients represented. We identified 257 novel lncRNAs (Figure 4A) and a total of 59,094 ASEs by using a tumor versus adjacent normal strategy, with 31,777 belonging to SE, 7,546 utilizing A5SS, 9,279 utilizing A3SS, 2,653 involving MXE, and 7,839 RI (Figure 4B).

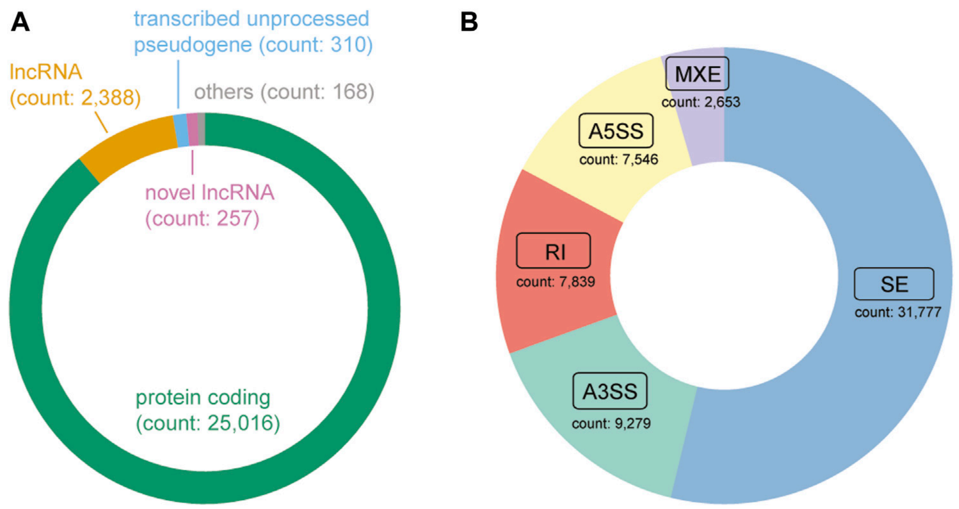


FIGURE 4
Composition of the gene type and alternative splicing event type in DASES. **(A)** Distribution of gene biotypes, where "Others" comprise a combination of transcribed unitary pseudogene, transcribed processed pseudogene, miRNA, TEC, unprocessed pseudogene, and IG C gene. **(B)** Distribution of alternative splicing event types.

To facilitate easy access and utilization of the database, we designed a user-friendly web interface featuring various modules. The Home page provides users with a concise overview of DASES, accompanied by illustrative diagrams showcasing the five major types of ASEs (Figure 5A). The Search page offers four different search options, namely, gene, transcript, ASE ID, and genomic region, facilitating easy and efficient data retrieval (Figure 5B). The Browse page provides a comprehensive list of all ASE IDs, allowing users to narrow down their queries by applying filters based on the ASE type or study name (Figure 5C). The Genome Browser page enables users to visualize the genomic regions associated with ASEs (Figure 5D). The Download page offers convenient access to essential files, including processed GTF file and ASE-related data, which can be downloaded for further analysis (Figure 5E). Lastly, the About page serves as a valuable resource, providing a detailed pipeline overview of the entire database, along with comprehensive explanations of important interface features, including headers and abbreviations for primary tables, enabling users to fully comprehend and navigate the database with ease.

3.2 Diversified search strategies

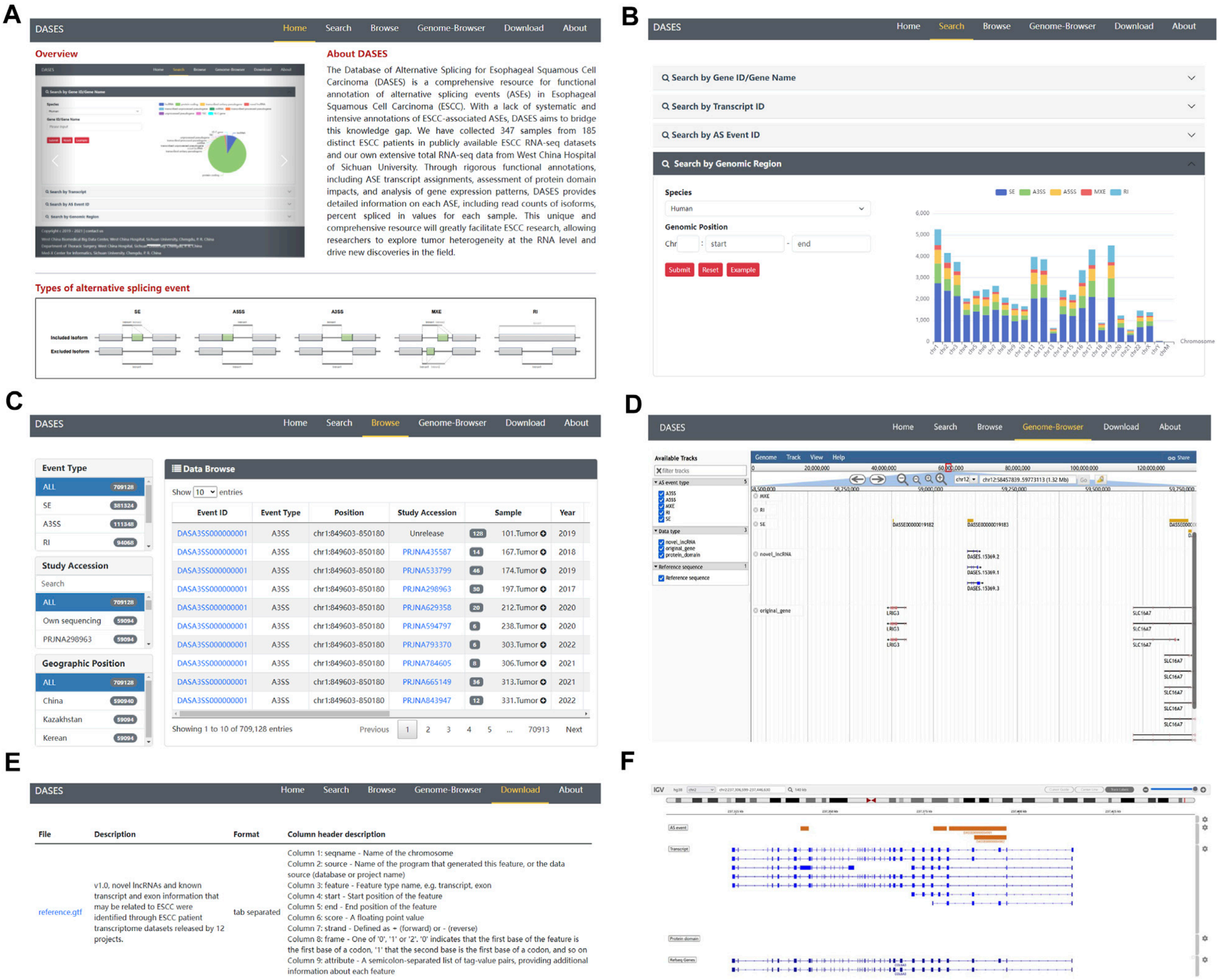
In DASES, we present a comprehensive search system comprising four dimensions (Figure 5B). The first dimension allows users to conduct searches based on the gene ID or gene name, thereby retrieving pertinent gene information alongside details concerning gene-associated ASEs. By employing the second dimension, users can search using the transcript ID, obtaining transcript-specific information, expression levels across samples, and insights into transcript-associated ASEs. The third dimension facilitates searches based on the ASE ID, yielding ASE-related details, including the exon junction count (EJC), intron junction count (IJC), and PSI values. Finally, the fourth dimension empowers users to search by genomic coordinates, resulting in the retrieval of ASE information specific to designated genomic loci.

3.3 Genome Browser visualization

The Genome Browser in DASES comprises two distinct sections. The first section facilitates the visualization of all ASEs (Figure 5D). Within the Genome Browser page, users can utilize diverse tracks to filter ASEs based on specific criteria, including ASE types and chromosome numbers. They also have the option to display or conceal tracks associated with ASEs, transcripts, genes, and protein domains. The second section is accessible via the Search page (Figure 5F). When users conduct searches for genes, transcripts, or specific ASEs, the pertinent information is presented visually on the Genome Browser. This seamless integration of search results with the Genome Browser offers users a contextual perspective on the genomic location of these elements.

3.4 Significant association between ESCC TNM staging and ASE frequency

ASEs have been closely linked to tumorigenesis and cancer progression. To investigate whether the frequency of ASEs exhibits an association with TNM staging in ESCC, we conducted a comprehensive analysis using data from DASES. As shown in Supplementary Figures S1A, D, both the frequency of genes undergoing alternative splicing (AS-gene frequency) and the frequency of ASEs in genes exhibiting AS (ASE frequency) exhibited a substantial increase within ESCC tissues when compared to adjacent normal tissues. Furthermore, our analysis unveiled a significant trend in the correlation between ESCC TNM staging and AS-gene frequency (Supplementary Figures S1B, C), as well as ASE frequency (Supplementary Figures S1E, F). These findings



About DASES

The Database of Alternative Splicing for Esophageal Squamous Cell Carcinoma (DASES) is a comprehensive resource for functional annotation of alternative splicing events (ASEs) in Esophageal Squamous Cell Carcinoma (ESCC). With a lack of systematic and intensive annotations of ESCC-associated ASEs, DASES aims to bridge this knowledge gap. We have collected 347 samples from 185 distinct ESCC patients in publicly available ESCC RNA-seq datasets and our own extensive total RNA-seq data from West China Hospital of Sichuan University. Through rigorous functional annotations, including ASE transcript assignments, assessment of protein domain impacts, and analysis of gene expression patterns, DASES provides detailed information on each ASE, including read counts of isoforms, percent spliced in values for each sample. This unique and comprehensive resource will greatly facilitate ESCC research, allowing researchers to explore tumor heterogeneity at the RNA level and drive new discoveries in the field.

Data Browse

Show 10 entries

Event ID	Event Type	Position	Study Accession	Sample		Year
DASA3SS000000001	A3SS	chr1:849603-850180	Unrelease	128	101.Tumor	2019
DASA3SS000000001	A3SS	chr1:849603-850180	PRJNA435587	14	167.Tumor	2018
DASA3SS000000001	A3SS	chr1:849603-850180	PRJNA533799	46	174.Tumor	2019
DASA3SS000000001	A3SS	chr1:849603-850180	PRJNA298963	30	197.Tumor	2017
DASA3SS000000001	A3SS	chr1:849603-850180	PRJNA629358	20	212.Tumor	2020
DASA3SS000000001	A3SS	chr1:849603-850180	PRJNA594797	6	238.Tumor	2020
DASA3SS000000001	A3SS	chr1:849603-850180	PRJNA793370	6	303.Tumor	2022
DASA3SS000000001	A3SS	chr1:849603-850180	PRJNA784605	8	306.Tumor	2021
DASA3SS000000001	A3SS	chr1:849603-850180	PRJNA665149	36	313.Tumor	2021
DASA3SS000000001	A3SS	chr1:849603-850180	PRJNA843947	12	331.Tumor	2022

Showing 1 to 10 of 709,128 entries Previous 1 2 3 4 5 ... 70913 Next

File	Description	Format	Column header description
reference.gtf	v1.0, novel lncRNAs and known transcript and exon information that may be related to ESCC were identified through ESCC patient transcriptome datasets released by 12 projects.	tab separated	Column 1: seqname - Name of the chromosome Column 2: source - Name of the program that generated this feature, or the data source (database or project name) Column 3: feature - Feature type name, e.g. transcript, exon Column 4: start - Start position of the feature Column 5: end - End position of the feature Column 6: score - A floating point value Column 7: strand - Defined as + (forward) or - (reverse) Column 8: frame - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on Column 9: attribute - A semicolon-separated list of tag-value pairs, providing additional information about each feature

FIGURE 5
Overview of DASES module interfaces. **(A)** Home interface, **(B)** Search interface, **(C)** Browse interface, **(D)** Genome Browser, **(E)** Download interface, and **(F)** Genome Browser after performing a search.

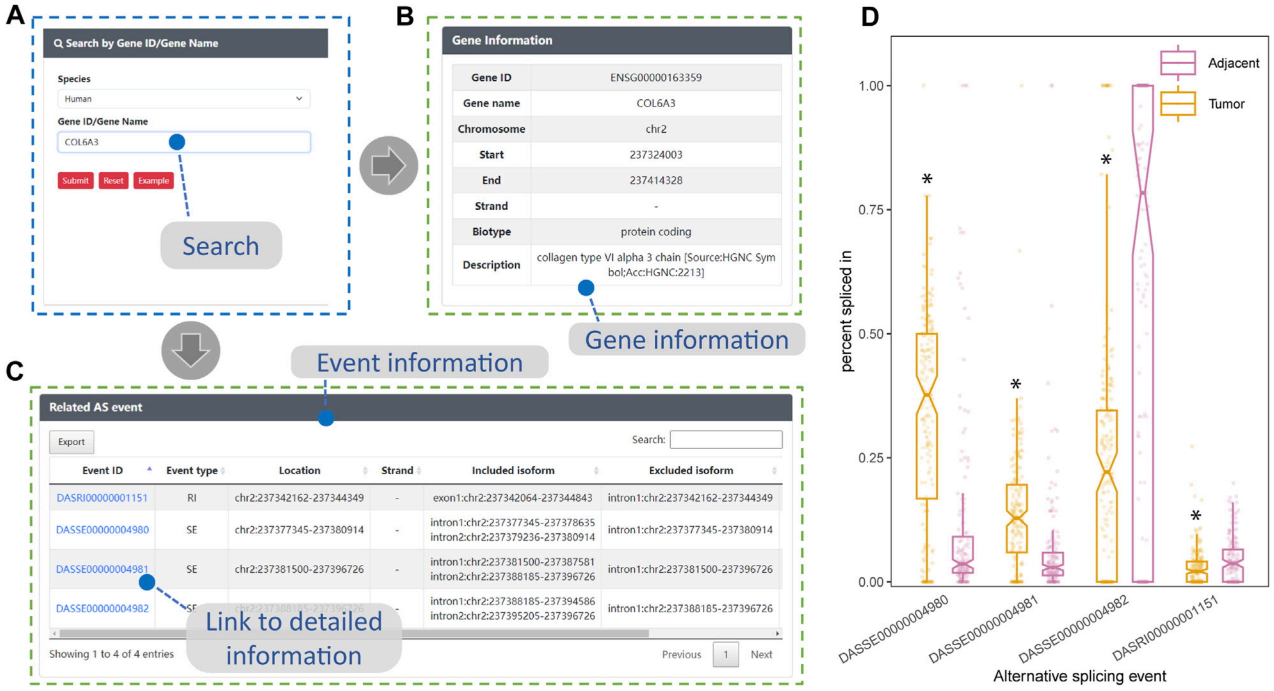
underscore a compelling association between ESCC TNM staging and the frequency of ASEs, suggesting their potential relevance in the context of ESCC progression. The source of DASES facilitates the exploration of these intricate relationships, providing a valuable platform for future research in this field.

3.5 Consistency with literature findings for *COL6A3* in DASES

The expression of *COL6A3* in both bulk esophageal tissue and single esophageal tissue samples exhibited a relatively high level, as evidenced by data obtained from the GTEx website (https://gtexportal.org/home/). Utilizing the search interface of DASES, we specifically queried *COL6A3* (Figures 6A, B), leading to the identification of four ASEs, i.e., three SE events and one RI event (Figure 6C). Notably, our findings closely align with the observations reported by Ding, who identified three SE-type ASEs of *COL6A3* from 11 samples in their study of ESCC tissues (Ding et al., 2021). Intriguingly, in addition to the ASEs reported by Ding, we discovered an additional RI-type ASE, “DASRI00000001151” (chr2: 237342162–237344349), which was not addressed by Ding. This discrepancy could be attributed to our larger sample size, which enabled us to identify more *COL6A3*-related ASEs. Importantly, we observed statistically significant differences in the PSI values of these four ASEs between the tumor and adjacent normal tissue groups, further highlighting their potential significance in ESCC (Figure 6D). This robust consistency between our findings and those of Ding provides substantial evidence for the reliability of ESCC-related ASEs documented within DASES, thereby reinforcing their validity through corroboration with findings from other literature reports.

4 Discussion

In this study, we successfully constructed DASES. By integrating publicly available RNA-seq from ESCC patient tissues, DASES



Search by Gene ID/Gene Name

Species: Human

Gene ID/Gene Name: COL6A3

Submit Reset Example

Gene Information

Gene ID	ENSG00000163359
Gene name	COL6A3
Chromosome	chr2
Start	237324003
End	237414328
Strand	-
Biotype	protein coding
Description	collagen type VI alpha 3 chain [Source:HGNC Symbol;Acc:HGNC:2213]

Related AS event

Export Search:

Event ID	Event type	Location	Strand	Included isoform	Excluded isoform
DASRI00000001151	RI	chr2:237342162-237344349	-	exon1:chr2:237342064-237344843	intron1:chr2:237342162-237344349
DASSE00000004980	SE	chr2:237377345-237380914	-	intron1:chr2:237377345-237378635 intron2:chr2:237379236-237380914	intron1:chr2:237377345-237380914
DASSE00000004981	SE	chr2:237381500-237396726	-	intron1:chr2:237381500-237387581 intron2:chr2:237388185-237396726	intron1:chr2:237381500-237396726
DASSE00000004982	SE	chr2:237388185-237396726	-	intron1:chr2:237388185-237394586 intron2:chr2:237395205-237396726	intron1:chr2:237388185-237396726

Showing 1 to 4 of 4 entries Previous 1 Next

FIGURE 6
Consistency between DASES results and literature findings for *COL6A3*. This figure showcases the validation process of DASES by employing an example search for the highly expressed *COL6A3* in esophageal tissue. **(A)** Process of searching for *COL6A3* using the search interface. The search results, including relevant descriptions of *COL6A3* in **(B)** and detailed information on the identified alternative splicing events that are related to *COL6A3* in **(C)**. Discovery of associated alternative splicing events that are consistent with literature reports (Ding et al., 2021). **(D)** Disparity in percent spliced in values for these four alternative splicing events between the tumor and adjacent normal tissue groups, as determined by the Mann–Whitney test. **p*-value < 0.05.

provides a comprehensive resource for the identification and exploration of ASEs potentially associated with ESCC. Moreover, DASES stands out as the first specialized database dedicated to ESCC-associated ASEs, addressing the existing gap in ESCC-specific databases in the field of AS.

DASES employed a tumor versus adjacent normal strategy to identify ESCC-associated ASEs, presenting several notable advantages. First, by comparing samples within the same patient, DASES effectively highlights splicing events that are highly likely to be functionally relevant to the development and progression of ESCC, which could reduce the confounding effects of individual genetic variations or splicing differences that are unrelated to ESCC. This strategy has been demonstrated to be effective in previous studies (Xiong et al., 2015; Kahles et al., 2018). Moreover, in line with other research conclusions, the approach enables the identification of ESCC-specific ASEs that may serve as potential biomarkers or therapeutic targets as they reflect the unique molecular characteristics of ESCC (Kalsotra and Cooper, 2011; Sebestyén et al., 2016). Furthermore, by comparing splicing patterns within the same patient, DASES minimizes inter-individual variations and provides a more robust assessment of the splicing changes specifically related to ESCC, enhancing the reliability of the identified ASEs in DASES. In a word, this approach allows for a more effective, accurate, and reliable characterization of ESCC-related AS.

DASES serves as a comprehensive resource that includes whole-transcriptome data to investigate both known ASEs linked to ESCC and novel lncRNAs, along with their associated ASEs that could potentially be implicated in ESCC. The identification of novel lncRNAs and their associated ASEs in ESCC holds great promise for advancing our understanding of the disease. Several studies have highlighted the importance of lncRNAs in cancer development and progression, including ESCC (Chen et al., 2018). These long non-coding RNAs regulate gene expression, modulate signaling pathways, and contribute to the hallmarks of cancer (Huarte, 2015). Therefore, the incorporation of lncRNA-associated ASEs in DASES provides valuable insights into the regulatory complexity underlying ESCC. Moreover, we offer a comprehensive GTF file that incorporates both the known transcriptome information and the newly discovered lncRNAs in DASES. This resource enables the in-depth exploration of the expression patterns, functional implications, and potential interactions of these newly identified lncRNAs in the context of ESCC.

We recognize that ESCC is a multifactorial disease with various pathogenic genes, including but not limited to *TP53*, *NOTCH1*, *CDKN2A*, and *COL6A3* (Gao et al., 2014; Li et al., 2019b; Liu et al., 2022; Ko et al., 2023). Our Gene Ontology (GO) enrichment analysis of differentially expressed genes highlighted "cell adhesion" as one of the top-ranked GO terms closely linked to cancer, with *COL6A3* being among the genes significantly associated with this GO term. Given these factors, we selected *COL6A3* as a representative gene for demonstrating the utility of DASES. Our analysis of ASEs within *COL6A3* revealed intriguing findings. Although consistent with the study conducted by Ding et al. (2021) on paired ESCC tissues for the most part, our dataset uncovered an additional RI-type ASE within *COL6A3* not reported by Ding et al. (2021). This discrepancy could be attributed to our larger sample size, enabling us to capture more ASEs. This novel finding highlights the value of our database in complementing existing knowledge and uncovering potentially clinically relevant splicing events. It also underscores the significance of leveraging a comprehensive resource like DASES to complement existing studies and expand our knowledge of this complex disease.

It is important to acknowledge that DASES also has certain limitations and areas that can be further improved. First, DASES currently focuses exclusively on ESCC patient tissue data and lacks representation from other species. However, human patient tissue data remain valuable, and future versions will include data from diverse organizations to broaden its scope. Second, DASES primarily relies on whole-transcriptome data, neglecting other sequencing data types. Integrating multiple omics data types can enhance our understanding of ESCC mechanisms. Moreover, in evaluating the impact of ASEs on proteins, we only considered instances where ASEs occur directly within protein domains. However, there are other ways in which proteins can be affected, such as a frameshift occurring before a protein domain or ASEs occurring in scaffold regions, which can influence their three-dimensional structure.

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.

Ethics statement

The studies involving humans were approved by the Ethics Committee of West China Hospital of Sichuan University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

YC was responsible for data analysis, visualization, and drafting the manuscript. YK designed the web interface. YYa deployed the database. SL performed data collection and data screening. ZY prepared the software program for analysis and database deployment. HY and YYu supervised the project and revised the draft of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Youth Science Foundation of China (Grant No. 32100927).

Acknowledgments

The authors thank the team members involved in the West China Biomedical Big Data Center, West China Hospital of Sichuan University, for their support.

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/fgene.2023.1237167/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Relationship between the frequency of ASEs and ESCC TNM staging. AS-gene frequency denotes the frequency of genes undergoing AS, while the ASE frequency indicates the frequency of ASEs in genes exhibiting AS. **(A,D)** show the statistically significant difference in the AS-gene frequency and ASE frequency between tumor and adjacent normal tissues, as determined by the Mann–Whitney test (**p*-value < 0.05). **(B,C)** demonstrate the statistically significant trends in the AS-gene frequency concerning ESCC T staging and N staging, and **(E,F)** reveal the statistically significant trends in the ASE frequency with respect to ESCC T staging and N staging, as determined by the Cochran–Armitage trend test ($^{\#}p$-value < 0.05).

References

Aalijahan, H., and Ghorbian, S. (2020). Clinical application of long non-coding RNA-UCA1 as a candidate gene in progression of esophageal cancer. *Pathol. Oncol. Res.* 26, 1441–1446. doi:10.1007/s12253-019-00711-3

Abnet, C. C., Arnold, M., and Wei, W. Q. (2018). Epidemiology of esophageal squamous cell carcinoma. *Gastroenterology* 154, 360–373. doi:10.1053/j.gastro.2017.08.023

Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Nikšić, M., et al. (2018). Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet* 391, 1023–1075. doi:10.1016/S0140-6736(17)33326-3

Bonnal, S. C., López-Oreja, I., and Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer - implications for care. *Nat. Rev. Clin. Oncol.* 17, 457–474. doi:10.1038/s41571-020-0350-x

Chen, T., Chen, X., Zhang, S., Zhu, J., Tang, B., Wang, A., et al. (2021). The genome sequence archive family: toward explosive data growth and diverse data types. *Genomics Proteomics Bioinforma.* 19, 578–583. doi:10.1016/j.gpb.2021.08.001

Chen, X., Chen, Z., Yu, S., Nie, F., Yan, S., Ma, P., et al. (2018). Long noncoding RNA LINC01234 functions as a competing endogenous RNA to regulate CBFB expression by sponging miR-204-5p in gastric cancer. *Clin. Cancer Res.* 24, 2002–2014. doi:10.1158/1078-0432.CCR-17-2376

Costoya, J. A., and Arce, V. M. (2023). Cancer cells escape the immune system by increasing stemness through epigenetic reprogramming. *Cell Mol. Immunol.* 20, 6–7. doi:10.1038/s41423-022-00953-3

Deng, Y., Luo, H., Yang, Z., and Liu, L. (2021). LncAS2Cancer: a comprehensive database for alternative splicing of lncRNAs across human cancers. *Brief. Bioinform* 22, bbaa179. doi:10.1093/bib/bbaa179

Ding, J., Li, C., Cheng, Y., Du, Z., Wang, Q., Tang, Z., et al. (2021). Alterations of RNA splicing patterns in esophagus squamous cell carcinoma. *Cell Biosci.* 11, 36. doi:10.1186/s13578-021-00546-z

Dlamini, Z., Hull, R., Mbatha, S. Z., Alaouna, M., Qiao, Y. L., Yu, H., et al. (2021). Prognostic alternative splicing signatures in esophageal carcinoma. *Cancer Manag. Res.* 13, 4509–4527. doi:10.2147/CMAR.S305464

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635

Duan, Y., Jia, Y., Wang, J., Liu, T., Cheng, Z., Sang, M., et al. (2021). Long noncoding RNA DGCR5 involves in tumorigenesis of esophageal squamous cell carcinoma via SRSF1-mediated alternative splicing of Mcl-1. *Cell Death Dis.* 12, 587. doi:10.1038/s41419-021-03858-7

Gao, Y. B., Chen, Z. L., Li, J. G., Hu, X. D., Shi, X. J., Sun, Z. M., et al. (2014). Genetic landscape of esophageal squamous cell carcinoma. *Nat. Genet.* 46, 1097–1102. doi:10.1038/ng.3076

Ghasemzadeh, S., and Ghorbian, S. (2023). Investigation of clinical significant utility of LncRNA-linc02389 in patients with esophageal squamous cell carcinoma. *J. Kermanshah Univ. Med. Sci.* 27, e136290. doi:10.5812/jkums-136290

He, W., Wang, C., Wu, L., Wan, G., Li, B., Han, Y., et al. (2022). Tislelizumab plus chemotherapy sequential neoadjuvant therapy for non-cCR patients after neoadjuvant chemoradiotherapy in locally advanced esophageal squamous cell carcinoma (ETNT): an exploratory study. *Front. Immunol.* 13, 853922. doi:10.3389/fimmu.2022.853922

Huarte, M. (2015). The emerging role of lncRNAs in cancer. *Nat. Med.* 21, 1253–1261. doi:10.1038/nm.3981

Kahles, A., Lehmann, K. V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., et al. (2018). Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell* 34, 211–224.e6. doi:10.1016/j.ccell.2018.07.001

Kalsotra, A., and Cooper, T. A. (2011). Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.* 12, 715–729. doi:10.1038/nrg3052

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* 7, 1009–1015. doi:10.1038/nmeth.1528

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi:10.1038/s41587-019-0201-4

Ko, K. P., Huang, Y., Zhang, S., Zou, G., Kim, B., Zhang, J., et al. (2023). Key genetic determinants driving esophageal squamous cell carcinoma initiation and immune evasion. *Gastroenterology* 165, 613–628.e20. doi:10.1053/j.gastro.2023.05.030

Kojima, T., Shah, M. A., Muro, K., Francois, E., Adenis, A., Hsu, C. H., et al. (2020). Randomized phase III KEYNOTE-181 study of pembrolizumab versus chemotherapy in advanced esophageal cancer. *J. Clin. Oncol.* 38, 4138–4148. doi:10.1200/JCO.20.01888

Li, A., Zhang, J., and Zhou, Z. (2014). PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinforma.* 15, 311. doi:10.1186/1471-2105-15-311

Li, W., Zhang, L., Guo, B., Deng, J., Wu, S., Li, F., et al. (2019a). Exosomal FMR1-AS1 facilitates maintaining cancer stem-like cell dynamic equilibrium via TLR7/NFκB/c-Myc signaling in female esophageal carcinoma. *Mol. Cancer* 18, 22. doi:10.1186/s12943-019-0949-7

Li, Y., Sun, Y., Yang, Q., Wu, J., Xiong, Z., Li, S., et al. (2019b). Variants in COL6A3 gene influence susceptibility to esophageal cancer in the Chinese population. *Cancer Genet.* 238, 23–30. doi:10.1016/j.cancergen.2019.07.003

Liang, H., Fan, J. H., and Qiao, Y. L. (2017). Epidemiology, etiology, and prevention of esophageal squamous cell carcinoma in China. *Cancer Biol. Med.* 14, 33–41. doi:10.20892/j.issn.2095-3941.2016.0093

Liu, J., Liu, Z. X., Wu, Q. N., Lu, Y. X., Wong, C. W., Miao, L., et al. (2020). Long noncoding RNA AGPG regulates PFKFB3-mediated tumor glycolytic reprogramming. *Nat. Commun.* 11, 1507. doi:10.1038/s41467-020-15112-3

Liu, T., Zhao, X., Lin, Y., Luo, Q., Zhang, S., Xi, Y., et al. (2022). Computational identification of preneoplastic cells displaying high stemness and risk of cancer progression. *Cancer Res.* 82, 2520–2537. doi:10.1158/0008-5472.CAN-22-0668

Morgan, E., Soerjomataram, I., Rumgay, H., Coleman, H. G., Thrift, A. P., Vignat, J., et al. (2022). The global landscape of esophageal squamous cell carcinoma and esophageal adenocarcinoma incidence and mortality in 2020 and projections to 2040: new estimates from GLOBOCAN 2020. *Gastroenterology* 163, 649–658.e2. doi:10.1053/j.gastro.2022.05.054

Pape, M., Vissers, P., de Vos-Geelen, J., Hulshof, M., Gisbertz, S. S., Jeene, P. M., et al. (2022). Treatment patterns and survival in advanced unresectable esophageal squamous cell cancer: a population-based study. *Cancer Sci.* 113, 1038–1046. doi:10.1111/cas.15262

Partners, C. M. a. (2022). Database resources of the national genomics data center, China national center for bioinformation in 2022. *Nucleic Acids Res.* 50, D27–D38. doi:10.1093/nar/gkab951

Peng, L., Cheng, S., Lin, Y., Cui, Q., Luo, Y., Chu, J., et al. (2018). CCGD-ESCC: a comprehensive database for genetic variants associated with esophageal squamous cell carcinoma in Chinese population. *Genomics Proteomics Bioinforma.* 16, 262–268. doi:10.1016/j.gpb.2018.03.005

Pertea, G., and Pertea, M. (2020). GFF utilities: GffRead and GffCompare. *F1000Res* 9, ISCB Comm J-304. doi:10.12688/f1000research.23297.2

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi:10.1038/nbt.3122

Razavi, M., and Ghorbian, S. (2019). Up-regulation of long non-coding RNA-PCAT-1 promotes invasion and metastasis in esophageal squamous cell carcinoma. *EXCLI J.* 18, 422–428. doi:10.17179/excli2018-1847

Sadeghpour, S., and Ghorbian, S. (2019). Evaluation of the potential clinical prognostic value of lncRNA-BANCR gene in esophageal squamous cell carcinoma. *Mol. Biol. Rep.* 46, 991–995. doi:10.1007/s11033-018-4556-2

Sciarrillo, R., Wojtuszkiewicz, A., Assaraf, Y. G., Jansen, G., Kaspers, G., Giovannetti, E., et al. (2020). The role of alternative splicing in cancer: from oncogenesis to drug resistance. *Drug Resist Updat* 53, 100728. doi:10.1016/j.drup.2020.100728

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M. A., et al. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res.* 26, 732–744. doi:10.1101/gr.199935.115

Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., et al. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* 111, E5593–E5601. doi:10.1073/pnas.1419161111

Siegfried, Z., and Karni, R. (2018). The role of alternative splicing in cancer drug resistance. *Curr. Opin. Genet. Dev.* 48, 16–21. doi:10.1016/j.gde.2017.10.001

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. Pozn.* 19, A68–A77. doi:10.5114/wo.2014.47136

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi:10.1038/nbt.1621

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., and Li, W. (2013). CPAT: coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41, e74. doi:10.1093/nar/gkt006

Wu, Q., Zhang, Y., An, H., Sun, W., Wang, R., Liu, M., et al. (2021). The landscape and biological relevance of aberrant alternative splicing events in esophageal squamous cell carcinoma. *Oncogene* 40, 4184–4197. doi:10.1038/s41388-021-01849-8

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347, 1254806. doi:10.1126/science.1254806

Yang, J., Bi, L., Wang, C., Wang, G., Gou, Y., Dong, L., et al. (2023). ESCCdb: a comprehensive database and key regulator exploring platform based on cross dataset comparisons for esophageal squamous cell carcinoma. *Comput. Struct. Biotechnol. J.* 21, 2119–2128. doi:10.1016/j.csbj.2023.03.026

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., et al. (2016). Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164, 805–817. doi:10.1016/j.cell.2016.01.029

Zhang, Y., Qian, J., Gu, C., and Yang, Y. (2021). Alternative splicing and cancer: a systematic review. *Signal Transduct. Target Ther.* 6, 78. doi:10.1038/s41392-021-00486-7

Zhang, Y., Yao, X., Zhou, H., Wu, X., Tian, J., Zeng, J., et al. (2022). OncoSplicing: an updated database for clinically relevant alternative splicing in 33 human cancers. *Nucleic Acids Res.* 50, D1340–D1347. doi:10.1093/nar/gkab851

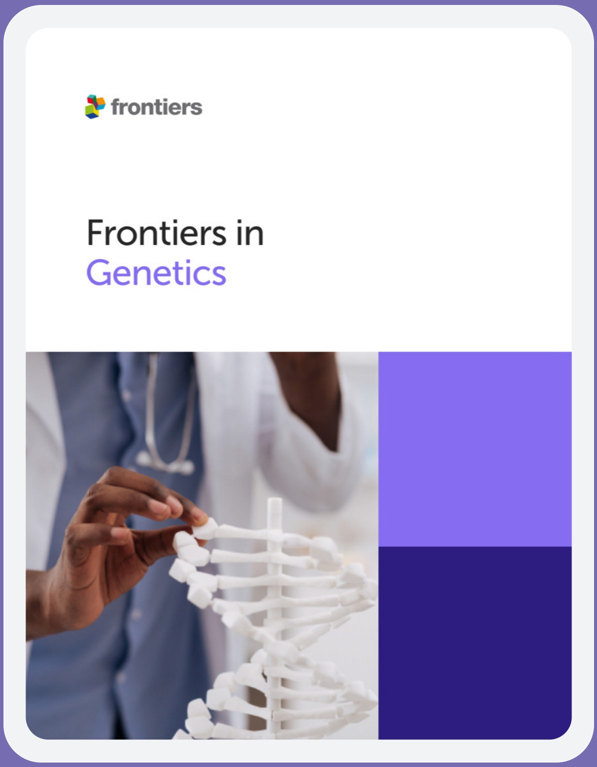
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

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