

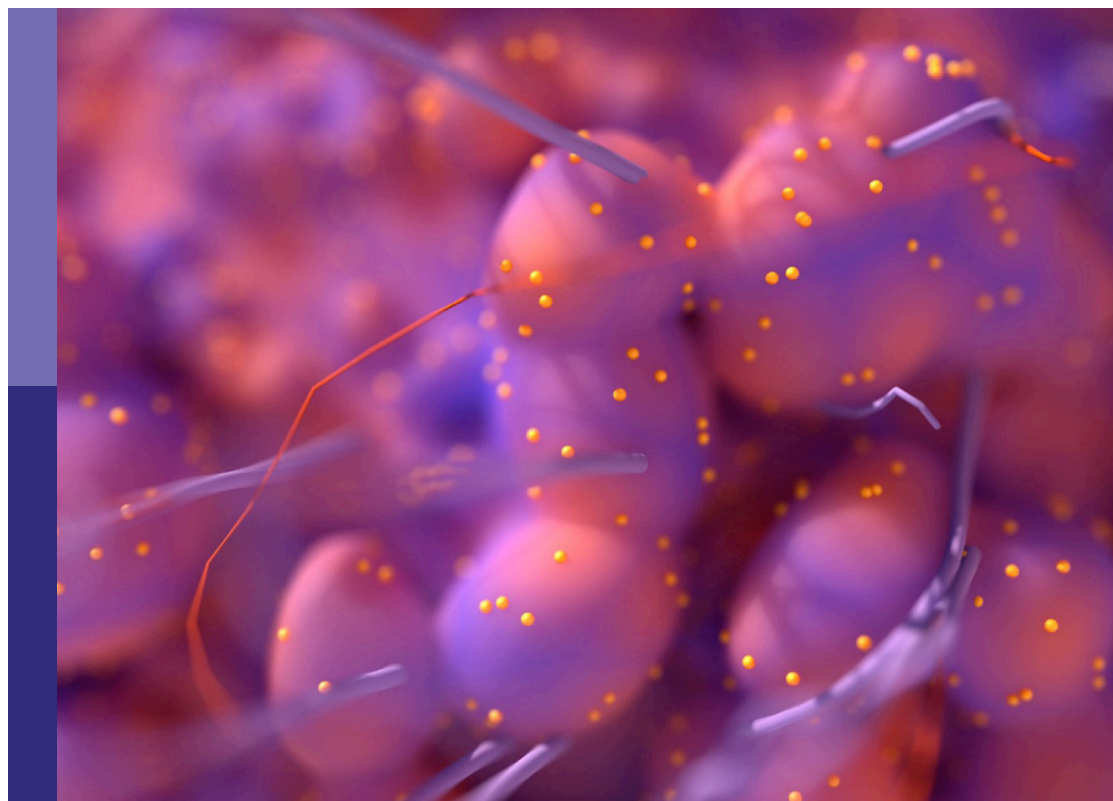
Identification of immune-related biomarkers for cancer diagnosis based on multi-omics data

Edited by

Liang Cheng, Xin Zhang, Chuan-Xing Li, Rui Guo and Tianyi Zhao

Published in

Frontiers in Oncology



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-83251-314-9
DOI 10.3389/978-2-83251-314-9

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

Identification of immune-related biomarkers for cancer diagnosis based on multi-omics data

Topic editors

Liang Cheng — Harbin Medical University, China

Xin Zhang — Jiangmen Central Hospital, China

Chuan-Xing Li — Respiratory Medicine Unit, Department of Medicine & Centre for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden

Rui Guo — Brigham and Women's Hospital, Harvard Medical School, United States

Tianyi Zhao — Harbin Institute of Technology, China

Citation

Cheng, L., Zhang, X., Li, C.-X., Guo, R., Zhao, T., eds. (2023). *Identification of immune-related biomarkers for cancer diagnosis based on multi-omics data*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-314-9

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.

Table of contents

- 06 **Editorial: Identification of immune-related biomarkers for cancer diagnosis based on multi-omics data**
Liang Cheng, Xin Zhang, Chuan-Xin Li, Rui Guo and Tianyi Zhao
- 09 **Meta-Analysis of Efficacy From CTLA-4 and PD-1/PD-L1 Inhibitors in Cancer Patients**
Li Xu, Xin Yan and Weiyue Ding
- 18 **ZKSCAN5 Activates *VEGFC* Expression by Recruiting SETD7 to Promote the Lymphangiogenesis, Tumour Growth, and Metastasis of Breast Cancer**
Jingtong Li, Zhifeng Yan, Jianli Ma, Zhong Chu, Huizi Li, Jingjing Guo, Qingyuan Zhang, Hui Zhao, Ying Li and Tao Wang
- 30 **XGBG: A Novel Method for Identifying Ovarian Carcinoma Susceptible Genes Based on Deep Learning**
Ke Feng Sun, Li Min Sun, Dong Zhou, Ying Ying Chen, Xi Wen Hao, Hong Ruo Liu, Xin Liu and Jing Jing Chen
- 37 **A Systematic Framework for Identifying Prognostic Genes in the Tumor Microenvironment of Colon Cancer**
Jinyang Liu, Yu Lan, Geng Tian and Jialiang Yang
- 47 **Inferring Gene Regulatory Networks From Single-Cell Transcriptomic Data Using Bidirectional RNN**
Yanglan Gan, Xin Hu, Guobing Zou, Cairong Yan and Guangwei Xu
- 57 **NKX2-8/PTHrP Axis-Mediated Osteoclastogenesis and Bone Metastasis in Breast Cancer**
Ainiwaerjiang Abudourousuli, Suwen Chen, Yameng Hu, Wanying Qian, Xinyi Liao, Yingru Xu, Libing Song, Shuxia Zhang and Jun Li
- 68 **A Novel Secreted Protein-Related Gene Signature Predicts Overall Survival and Is Associated With Tumor Immunity in Patients With Lung Adenocarcinoma**
Shuaijun Chen, Jun Zhang, Qian Li, Lingyan Xiao, Xiao Feng, Qian Niu, Liqin Zhao, Wanli Ma and Hong Ye
- 84 **Prediction of Transcription Factor Binding Sites Using a Combined Deep Learning Approach**
Linan Cao, Pei Liu, Jialong Chen and Lei Deng
- 94 **Heterogeneity Analysis of Bladder Cancer Based on DNA Methylation Molecular Profiling**
Shuyu Wang, Dali Xu, Bo Gao, Shuhan Yan, Yiwei Sun, Xinxing Tang, Yanjia Jiao, Shan Huang and Shumei Zhang
- 104 **Prediction of Gastric Cancer-Related Genes Based on the Graph Transformer Network**
Yan Chen, Xuan Sun and Jiaxing Yang

- 111 **Multi-Omics Integration-Based Prioritisation of Competing Endogenous RNA Regulation Networks in Small Cell Lung Cancer: Molecular Characteristics and Drug Candidates**
Xiao-Jun Wang, Jing Gao, Qin Yu, Min Zhang and Wei-Dong Hu
- 133 **Exome Sequencing Reveals Genetic Variability and Identifies Chronic Prognostic Loci in Chinese Sarcoidosis Patients**
Qian Zhang, Hui Huang, Meijun Zhang, Chuling Fang, Na Wang, Xiaoyan Jing, Jian Guo, Wei Sun, Xiaoyu Yang and Zuojun Xu
- 146 **Evaluating the Microsatellite Instability of Colorectal Cancer Based on Multimodal Deep Learning Integrating Histopathological and Molecular Data**
Wenjing Qiu, Jiasheng Yang, Bing Wang, Min Yang, Geng Tian, Peizhen Wang and Jialiang Yang
- 154 **Computational Characterizing Necroptosis Reveals Implications for Immune Infiltration and Immunotherapy of Hepatocellular Carcinoma**
Jun Zhu, Tenghui Han, Shoujie Zhao, Yejing Zhu, Shouzheng Ma, Fenghua Xu, Tingting Bai, Yuxin Tang, Yungang Xu and Lei Liu
- 170 **Pathogens and Pathogenesis in Wheezing Diseases in Children Under 6**
Yongjun Tang, Yaxiong Yang, Ruohui He, Rong Huang, Xiangrong Zheng and Chentao Liu
- 178 **A novel cuproptosis-related prognostic lncRNA signature and lncRNA MIR31HG/miR-193a-3p/TNFRSF21 regulatory axis in lung adenocarcinoma**
Xiaocong Mo, Di Hu, Pingshan Yang, Yin Li, Shoaib Bashir, Aitao Nai, Feng Ma, Guoxia Jia and Meng Xu
- 195 **Deep-LC: A Novel Deep Learning Method of Identifying Non-Small Cell Lung Cancer-Related Genes**
Mo Li, Guang xian Meng, Xiao wei Liu, Tian Ma, Ge Sun and HongMei He
- 201 **Comprehensive analyses unveil novel genomic and immunological characteristics of micropapillary pattern in lung adenocarcinoma**
Yansong Huo, Leina Sun, Jie Yuan, Hua Zhang, Zhenfa Zhang, Lianmin Zhang, Wuhao Huang, Xiaoyan Sun, Zhe Tang, Yingnan Feng, Huilan Mo, Zuoquan Yang, Chao Zhang, Zicheng Yu, Dongsheng Yue, Bin Zhang and Changli Wang
- 213 **Identification of methylation signatures associated with CAR T cell in B-cell acute lymphoblastic leukemia and non-hodgkin's lymphoma**
Jiwei Song, FeiMing Huang, Lei Chen, KaiYan Feng, Fangfang Jian, Tao Huang and Yu-Dong Cai
- 226 **A computational method for large-scale identification of esophageal cancer-related genes**
Xin He, Wei-Song Li, Zhen-Gang Qiu, Lei Zhang, He-Ming Long, Gui-Sheng Zhang, Yang-Wen Huang, Yun-mei Zhan and Fan Meng

- 234 **Predicting recurrence and metastasis risk of endometrial carcinoma *via* prognostic signatures identified from multi-omics data**
Ling Li, Wenjing Qiu, Liang Lin, Jinyang Liu, Xiaoli Shi and Yi Shi
- 243 **Identification of novel cuproptosis-related lncRNA signatures to predict the prognosis and immune microenvironment of breast cancer patients**
Zi-Rong Jiang, Lin-Hui Yang, Liang-Zi Jin, Li-Mu Yi, Ping-Ping Bing, Jun Zhou and Jia-Sheng Yang
- 256 **Predicting non-small cell lung cancer-related genes by a new network-based machine learning method**
Yong Cai, Qiongya Wu, Yun Chen, Yu Liu and Jiying Wang
- 263 **SVMMDR: Prediction of miRNAs-drug resistance using support vector machines based on heterogeneous network**
Tao Duan, Zhufang Kuang and Lei Deng
- 277 **B cell deficiency promotes the initiation and progression of lung cancer**
Han Wu, Chen Chen, Lixing Gu, Jiapeng Li, Yunqiang Yue, Mengqing Lyu, Yeting Cui, Xiaoyu Zhang, Yu Liu, Haichuan Zhu, Xinghua Liao, Tongcun Zhang, Fan Sun and Weidong Hu
- 289 **Identifying anal and cervical tumorigenesis-associated methylation signaling with machine learning methods**
Fangfang Jian, FeiMing Huang, Yu-Hang Zhang, Tao Huang and Yu-Dong Cai
- 301 **Functional and embedding feature analysis for pan-cancer classification**
Jian Lu, JiaRui Li, Jingxin Ren, Shijian Ding, Zhenbing Zeng, Tao Huang and Yu-Dong Cai
- 316 **Proteomics and phosphoproteomics of chordoma biopsies reveal alterations in multiple pathways and aberrant kinases activities**
Jing Hang, Hanqiang Ouyang, Feng Wei, Qihang Zhong, Wanqiong Yuan, Liang Jiang and Zhongjun Liu
- 331 **Nomogram of intra-abdominal infection after surgery in patients with gastric cancer: A retrospective study**
Yue Zhang, Zhengfei Wang, Zarrin Basharat, Mengjun Hu, Wandong Hong and Xiangjian Chen
- 343 **Microscopic polyangiitis presenting with persistent cough and hemoptysis in pediatrics: A case report and review of the literature**
Yantong Zhu and Xiangrong Zheng



OPEN ACCESS

EDITED AND REVIEWED BY
Heather Cunliffe,
University of Otago, New Zealand

*CORRESPONDENCE

Liang Cheng
✉ liangcheng@hrbmu.edu.cn
Xin Zhang
✉ zhangx45@mail3.sysu.edu.cn
Chuan-Xin Li
✉ chuan-xing.li@ki.se
Rui Guo
✉ rguo2@bwh.harvard.edu
Tianyi Zhao
✉ zty2009@hit.edu.cn

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 09 December 2022

ACCEPTED 13 December 2022

PUBLISHED 26 December 2022

CITATION

Cheng L, Zhang X, Li C-X,
Guo R and Zhao T (2022) Editorial:
Identification of immune-related
biomarkers for cancer diagnosis
based on multi-omics data.
Front. Oncol. 12:1119622.
doi: 10.3389/fonc.2022.1119622

COPYRIGHT

© 2022 Cheng, Zhang, Li, Guo and
Zhao. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Editorial: Identification of immune-related biomarkers for cancer diagnosis based on multi-omics data

Liang Cheng^{1*}, Xin Zhang^{2*}, Chuan-Xin Li^{3*}, Rui Guo^{4*}
and Tianyi Zhao^{5*}

¹Jiangmen Key Laboratory of Clinical Biobanks and Translational Research, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, ²Jiangmen Central Hospital, Jiangmen, China, ³Respiratory Medicine Unit, Department of Medicine & Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden, ⁴Brigham and Women's Hospital, Harvard Medical School, Stockholm, United States, ⁵School of Medicine and Health, Harbin Institute of Technology, Harbin, China

KEYWORDS

cancer, biomarkers (BMs), multi-omics data, machine learning, cancer diagnosis

Editorial on the Research Topic

Identification of immune-related biomarkers for cancer diagnosis based on multi-omics data

Cancer has become one of the leading causes of mortality around the globe. Activating the innate immune signal pathway and inducing the anti-tumor immune response plays a key role in the efficacy of tumor treatment, especially in preventing the recurrence of residual tumor cells. With the development of high-throughput sequencing technology, multi-omics data for cancer has become accessible. These data have given researchers increasing opportunities to explore genetic risk, regulatory mechanisms, and protein function of the immune microenvironment in cancers. However, it is still a big challenge to utilize these data effectively and to mine knowledge from them. Artificial intelligence algorithms and statistical methods have shown great potential to take advantage of omics data and reveal mechanisms of immune function in cancer.

Here, we organized a Research Topic on “*Identification of Immune-Related Biomarkers for Cancer Diagnosis Based on Multi-Omics Data*.” In total, about 30 outstanding works were presented in this thematic issue, ten of which have been highlighted as follows.

- [Xu et al.](#) performed a meta-analysis by downloading data from PubMed, Google Scholar, and Embase databases on randomized clinical trials compared ipilimumab, nivolumab, pembrolizumab, or atezolizumab with non-immunotherapy controls. Median overall survival (OS) and median progression-free survival (PFS) were selected to evaluate the efficacy of cytotoxic T-lymphocyte-associated protein 4 (CTLA-4), programmed cell

death 1 (PD-1), and programmed death ligand 1 (PD-L1) inhibitors. Utilizing the random-effect model, hazard ratios (HRs) with 95 confidence intervals (CIs) were calculated by R software. The meta-analysis suggested that ICIs were associated with obvious improvements in PFS and OS compared with non-ICI therapies.

- Sun et al. introduced a novel disease-susceptible gene prediction method, XGBG, to study ovarian carcinomas (OCs) in more depth. Firstly, they employed the graph convolutional network (GCN) to reconstruct the gene features based on both gene features and network topological structure. Then, a boosting method was utilized to predict OC susceptible genes. The final XGBG model achieved a high AUC of 0.7541 and an AUPR of 0.8051. This method is helpful in further understanding the etiology and pathology of OC, and may be used as strong theoretical evidence for drug design.
- Chen et al. developed a novel method named “DBN-GTN” to identify gastric cancer-related genes on a large scale. This method built a heterogeneous network using a disease similarity network and a gene interaction network. Meanwhile, the deep belief network (DBN) was applied to reduce the dimension of features. This method used multiple features of genes and gastric cancer to identify the patterns of gastric cancer-related genes, which can be used to find more gastric cancer-related genes, and it performed best among four traditional methods and five similar methods. This paper provides support to further explain the genetic risk, susceptibility, and drug screening of gastric cancer.
- Liu et al. investigated prognostic genes in the tumor microenvironment of colon cancer using gene expression profiles and clinical information from colon adenocarcinoma (COAD) and rectal adenocarcinoma (READ). Meanwhile, they utilized the nine key prognostic genes obtained to build the independent prognostic model. They calculated stromal and immune scores for each sample and identified nine key prognostic genes including *HOXC8*, *SRPX*, *CCL22*, *CD72*, *IGLON5*, *SERPING1*, *PCOLCE2*, *FABP4*, and *ARL4C* by LASSO Cox regression analysis. This work may help in clinical decisions and improve the prognosis for colon cancer.
- Cao et al. developed an accurate and interpretable attention-based hybrid approach, DeepARC, which combined a convolutional neural network (CNN) and recurrent neural network (RNN) to predict transcription factor binding sites (TFBS). Taking advantage of the attention mechanism, DeepARC can gain greater access to valuable information about the motif and bring interpretability to the work of searching for motifs through the attention weight graph. Moreover, DeepARC achieved an average area under the receiver operating characteristic curve (AUC) score of 0.908 on five cell lines in the benchmark dataset. This method predicts better than existing state-of-the-art methods and has good interpretability.
- Qiu et al. developed a novel deep-learning framework to study the association between MSI status and several molecules including mRNA, miRNA, lncRNA, DNA methylation, and copy number variation (CNV) using colorectal cancer data from The Cancer Genome Atlas (TCGA). The fusion models integrating the H&E image with a single type of molecule has higher prediction accuracies than that using the H&E image alone, with the highest AUC of 0.952 achieved when combining the H&E image with DNA methylation data. This study may have clinical significance in practice and provide a reference for future studies.
- Duan et al. proposed an SVM-based method, SVMMDR, to predict the relationship between miRNAs and drug resistance based on the miRNAs-drug resistance association data from the ncDR database. The SVMMDR integrated miRNAs-drug resistance association, miRNAs sequence similarity, drug chemical structure similarity, and other similarities, extracted path-based Heterosim features, and obtained inclined diffusion features through restart random walk. By combining the multiple features, the prediction score between miRNAs and drug resistance was obtained based on the SVM. The final the average AUC of the SVMMDR method was 0.978 in 10-fold cross-validation. This work shows that SVMMDR has a significant performance advantage compared with existing methods.
- He et al. proposed a method for large-scale identification of esophageal cancer-related genes by computational methods, GCNLMF, to improve the efficiency of esophageal cancer genetic susceptibility research. This method fused graph convolutional networks and logical matrix factorization to effectively identify esophageal cancer-related genes through the association between genes. The GCNLMF achieved an AUC of 0.927 and AUPR of 0.86 in 10-fold cross-validation. In the final comparison with the other five methods, GCNLMF performed best. This study provides a new algorithm for finding signature genes in esophageal cancer and offers new insights into the future development of esophageal cancer research.
- Li et al. first downloaded the mRNA, microRNA (miRNA), long non-coding RNA (lncRNA), copy number variation (CNV) data, and clinical information of patients with endometrial cancer (EC) from The Cancer Genome Atlas (TCGA). Then, differential

expression analyses were performed to screen potential prognostic markers and establish prediction models using three classifiers. Finally, the prediction model achieved an area under the curve of 0.763, and an accuracy of 0.819 under 10-fold cross-validation. This work develops a computational model using omics information, which can predict the recurrence and metastasis risk of EC accurately, thereby avoiding improper treatment, and improving the prognosis of patients.

- [Li et al.](#) proposed a novel deep-learning method named Deep-LC for predicting NSCLC-related genes. Firstly, they built a gene interaction network and used graph convolutional networks to extract features of genes and interactions between gene pairs. Then, a simple convolutional neural network module was used as the decoder to decide whether the gene was related to the disease. Deep-LC is an end-to-end method, and from the evaluation results, they can conclude that Deep-LC performs well in mining potential Non-Small Cell Lung Cancer-related genes and performs better than existing state-of-the-art methods. This work provides new insights for future research in non-small cell lung cancer.

Author contributions

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

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.



Meta-Analysis of Efficacy From CTLA-4 and PD-1/PD-L1 Inhibitors in Cancer Patients

Li Xu^{1,2}, Xin Yan¹ and Weiyue Ding^{1,3*}

¹ College of Computer Science and Technology, Harbin Engineering University, Harbin, China, ² Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China, ³ School of Mathematics, Harbin Institute of Technology, Harbin, China

OPEN ACCESS

Edited by:

Liang Cheng,
Harbin Medical University, China

Reviewed by:

Xiangtao Li,
Jilin University, China
Yang Yu,
Shenyang Normal University, China

*Correspondence:

Weiyue Ding
wyding0501@hotmail.com

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 15 February 2022

Accepted: 17 March 2022

Published: 28 April 2022

Citation:

Xu L, Yan X and Ding W (2022) Meta-Analysis of Efficacy From CTLA-4 and PD-1/PD-L1 Inhibitors in Cancer Patients. *Front. Oncol.* 12:876098. doi: 10.3389/fonc.2022.876098

Introduction: Immune checkpoint inhibitors (ICIs) have been approved to prolong overall survival (OS), compared to other treatments. However, the recent studies reported consistent and inconsistent results. Hence, we conducted this meta-analysis to evaluate the efficacy of ICIs.

Materials and Methods: The articles were identified by searching PubMed, Embase, and Google Scholar published up to December 2021. A total of 12,126 participants (6,450 cases and 5,676 controls) were involved in the meta-analysis. Median OS and median progression-free survival (PFS) were selected to evaluate the efficacy of cytotoxic T-lymphocyte-associated protein 4 (CTLA-4), programmed cell death 1 (PD-1), and programmed death ligand 1 (PD-L1) inhibitors (ipilimumab, nivolumab or pembrolizumab, and atezolizumab, respectively). Utilizing the random-effect model, hazard ratios (HRs) with 95 confidence intervals (CIs) were calculated by R software.

Results: We observed a significant association between cancer patients and ICIs in OS (HR = 0.79, CI = 0.74–0.84) and PFS (HR = 0.80, CI = 0.75–0.86).

Conclusions: The meta-analysis suggested that ICIs were associated with obvious improvements in PFS and OS compared with non-ICI therapies.

Keywords: immune checkpoint inhibitors, meta-analysis, cytotoxic T-lymphocyte-associated protein 4, overall survival, programmed cell death 1, progression-free survival, programmed death ligand 1

INTRODUCTION

Cancer, an enormous burden on society, is one of the main reasons of death in both developed and developing countries. According to the global cancer statistics, there were about 19.3 million new cancer cases and nearly 10.0 million cancer deaths in 2020 worldwide (1). The immune system can recognize and prepare to eliminate cancer but is controlled by inhibitory receptors and ligands (2).

Immune checkpoints are regulatory pathways in the immunome that inhibit a part of an active immune response against a specific target or a group of targets (3). These immune checkpoint pathways are often able to keep self-tolerance and limit incidental tissue damage during the antimicrobial immune response; thus, immune destruction can be averted by cancers. There is no doubt that tumors co-opt certain immune checkpoint pathways as a main mechanism of immune resistance, especially against T cells that are specific for tumor antigens (4).

Immune checkpoint inhibitors (ICIs), which regain the efficacy of tumor-specific T cells in the tumor microenvironment, enhancing the immune system's ability to recognize and eradicate tumors, are breakthroughs in the treatment of cancer and have made significant advances in both hematological and solid tumor oncology (5). They have been approved for use in melanoma, bladder cancer, non-small cell lung cancer (NSCLC), stomach cancer, renal cell carcinoma (RCC) and head and neck squamous cell carcinoma and will be approved for other types in the foreseeable future tumors (6, 7).

The US Food and Drug Administration (FDA) approved ipilimumab as the first CTLA-4 inhibitor of advanced melanoma. Nivolumab and pembrolizumab were the first of two PD-1 inhibitors approved for advanced melanoma, and atezolizumab was the first programmed death ligand 1 (PD-L1) inhibitor approved by the FDA (8–11).

Recent studies showed that ICIs could prolong the overall survival (OS) of cancer patients, compared with placebo, dacarbazine, everolimus, paclitaxel, chemotherapy, and other therapy methods or drugs (12–24). However, the studies reported inconsistent results. In 2013, Reck et al. randomly assigned 130 SCLC patients to receive paclitaxel with placebo (control) or ipilimumab 10 mg/kg in two alternative regimens, concurrent ipilimumab or phased ipilimumab, and declared that ipilimumab did not prolong the overall survival (OS) of SCLC patients (25). In 2014, Kwon et al. did a double-blind, multicenter, randomized, phase 3 trial with 799 metastatic castration-resistant prostate cancer (399 to ipilimumab and 400 to placebo) patients and reported that no obvious difference in overall survival was found between the ipilimumab group and the placebo group (26). In 2016, Beer et al. randomly assigned 400 and 202 metastatic castration-resistant prostate cancer patients to ipilimumab and to placebo, respectively, and discovered that ipilimumab did not increase the overall survival (OS) of patients with metastatic castration-resistant prostate cancer (27). Reck et al. randomly assigned 478 small-cell lung cancer (SCLC) patients to the chemotherapy plus ipilimumab group and 476 SCLC patients to the chemotherapy plus placebo group and got a conclusion that ipilimumab plus chemotherapy did not prolong OS compared with chemotherapy alone in SCLC patients (28). Larkin et al. randomly assigned 272 melanoma patients to the nivolumab group and 133 melanoma patients to chemotherapy and found that nivolumab did not prolong OS compared with chemotherapy alone in SCLC (29). Owonikoko et al. randomly assigned 278 SCLC patients to the ipilimumab group and 278

SCLC patients to the chemotherapy group and found that ipilimumab did not prolong OS compared with chemotherapy alone in SCLC patients (30). Spigel et al. randomly assigned 284 SCLC patients to the nivolumab group and 285 SCLC patients to the chemotherapy group and got a conclusion that nivolumab did not prolong OS compared with chemotherapy alone in SCLC patients (31). Hence, to get a more convincing result, we performed a meta-analysis to study the efficacy of ipilimumab, nivolumab, pembrolizumab, and atezolizumab, compared to other therapies or other drugs.

MATERIALS AND METHODS

Search Strategy

We identified all randomized clinical trials that compared ipilimumab, nivolumab, pembrolizumab, or atezolizumab with the non-immunotherapy control arms from January 1, 2007, to December 31, 2021. The articles we collected were searched by using the keywords “overall survival” or “OS,” “progression-free survival” or “PFS,” “immune checkpoint inhibitors” or “immune checkpoint blockade” or “ICIs” or “ipilimumab” or “nivolumab” or “pembrolizumab” or “atezolizumab” in the PubMed, Google Scholar and Embase databases. The articles we selected were written in English.

Study Selection Criteria

Trials were eligible for inclusion if they met the following criteria: (1) trials that involved patients must receive cancer treatment; (2) trials that had adequate data available including OS and PFS; (3) trials were phase 2 or phase 3 randomized clinical trials (RCTs); and (4) the articles published must be written in English.

Data Extraction

We extracted the following information from each study and selected the items including first author's last name, year of publication, phase of RCTs, the name of the ICIs (ipilimumab, nivolumab, pembrolizumab or atezolizumab) and control arms, number of patients ICIs and control groups, and the hazard ratios (HRs) of OS and PFS. All the duplicated studies were excluded.

Statistics Analysis

To calculate the overall incidence and HR of OS and PFS, we combined estimates by exploiting the fixed-effect model with the Mantel and Haenszel method and by employing the random-effect models with the DerSimonian and Laird method. The statistical analysis was performed with the R software package named Meta. The HR with 95% confidence interval (CI) was calculated to access the association between overall survival and ICIs.

Two quantities, Cochran's Q and I^2 , were used to access the heterogeneity in different types of ICIs groups and subgroups. Statistical heterogeneity was assessed using Cochran's Q statistic, and the p value ranging from 0% to 100%, to measure the significance level of inconsistency. If the value of I^2 is less than 50%, or the p value of heterogeneity is greater than 0.10, the fixed-effect model is applied, otherwise the random effect model is employed. After the heterogeneity test, we exploited the R meta package to conduct the meta-analysis with the random-effect model.

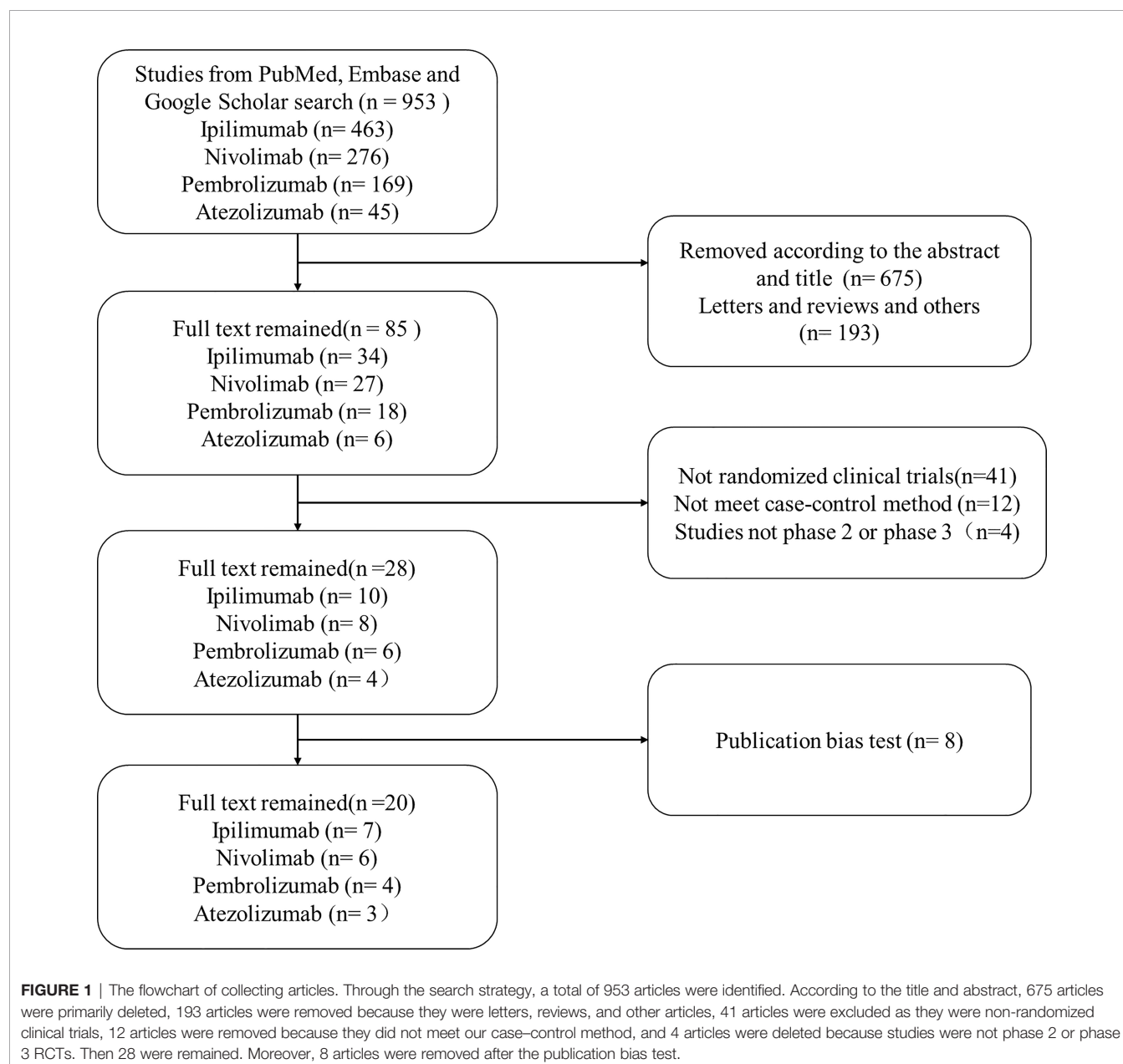
Egger's test (32) and Begg's test (33) were selected to access the publication bias for OS and PFS. When a two-tailed p value was less than 0.05, the publication bias was extremely significant.

Moreover, the potential publication bias was assessed by Begg's funnel plots to check the relative symmetry of the overall estimated individual study estimates.

RESULTS

Literature Search

A flowchart for the article selection is shown in **Figure 1**. Through the search strategy, a total of 953 articles were identified. According to the title and abstract, 675 articles were primarily deleted, 193 articles were removed because they



were letters, reviews, and other articles, 41 articles were excluded as they were not randomized clinical trials (RCTs), 12 articles were removed because they did not meet our case-control method, and 4 articles were deleted because studies were not phase 2 or phase 3 RCTs. Then 28 were remained, and 8 articles were removed because of the publication bias test. Finally, 20 articles were left, including 7 ipilimumab articles, 6 nivolumab articles, 4 pembrolizumab articles, and 3 atezolizumab articles, respectively (shown in **Table 1**). Moreover, the data we extracted from the articles were accessed in clinical trial databases with specific identifiers.

Heterogeneity Test

The summary result of heterogeneity is directly shown in **Table 2**. In the OS group, we found little heterogeneity in total with $I^2 = 41\%$, $p = 0.02$, and we chose to select the random-effect model according to the method we used. In the tumor subgroup heterogeneity test, we did not find obvious significant heterogeneity in the melanoma ($I^2 = 38\%$, $p = 0.16$), SCLC ($I^2 = 0\%$, $p = 0.83$), NSCLC ($I^2 = 0\%$, $p = 0.48$), and urothelial cancer ($I^2 = 0\%$, $p = 0.73$) subgroups, but a significant heterogeneity was found in the prostate cancer ($I^2 = 77\%$, $p = 0.04$) subgroup. In the PFS group, we also found

TABLE 1 | The primary characteristics of the 23 articles.

Study	Year	Treatment	Arm	Phase	Tumor	No.		OS		PFS	
						ICIs	Control	HR	95%CI	HR	95%CI
Hodi et al. (12)	2010	Ipi+Gp100	Gp100	3	Melanoma	403	136	0.68	0.55–0.65	0.81	0.66–1
Hodi et al. (12)	2010	Ipilimumab	Gp100	3	Melanoma	137	136	0.66	0.51–0.87	0.64	0.5–0.83
Robert et al. (13)	2011	Ipi+ DTIC	Dacarbazine	3	Melanoma	250	252	0.716	0.558–0.872	NA	NA
Reck et al. (25)	2013	CP+Con Ipi	Chemotherapy	2	SCLC	43	55	0.947	0.585–1.583	0.93	0.588–1.481
Reck et al. (25)	2013	CP+Seq Ipi	Chemotherapy	2	SCLC	42	55	0.753	0.461–1.232	0.927	0.59–1.45
Kwon et al. (26)	2014	Ipilimumab	Placebo	3	Prostate	399	400	0.83	0.71–0.96	0.7	0.61–0.82
Weber et al. (17)	2015	Nivolumab	Chemotherapy	3	Melanoma	272	133	0.95	0.73–1.24	1.03	0.78–1.36
Brahmer et al. (15)	2015	Nivolumab	Docetaxel	3	NSCLC	135	137	0.59	0.43–0.81	0.62	0.47–0.81
Borghaei et al. (14)	2015	Nivolumab	Docetaxel	3	NSCLC	292	290	0.73	0.59–0.89	0.92	0.77–1.11
Ribas et al. (16)	2015	Pembrolizumab	Chemotherapy	2	Melanoma	180	179	0.87	0.67–1.12	0.58	0.46–0.73
Beer et al. (27)	2016	Ipilimumab	Placebo	3	Prostate	400	202	1.11	0.88–1.39	0.67	0.55–0.8
Reck et al. (28)	2016	Ipilimumab	VP16+Plt	3	SCLC	478	476	0.936	0.807–1.085	0.85	0.75–0.97
Herbst et al. (19)	2016	Pembrolizumab 2mg	Chemotherapy	3	NSCLC	344	343	0.71	0.58–0.88	0.88	0.73–1.04
Herbst et al. (19)	2016	Pembrolizumab 10mg	Chemotherapy	3	NSCLC	346	343	0.61	0.49–0.75	0.79	0.66–0.94
Fehrenbacher et al. (18)	2016	Atezolizumab	Docetaxel	3	NSCLC	144	133	0.69	0.52–0.92	0.92	0.71–1.2
Rittmeyer et al. (21)	2016	Atezolizumab	Docetaxel	3	NSCLC	425	425	0.73	0.62–0.81	0.95	0.82–1.1
Bellmunt et al. (20)	2017	Pembrolizumab	Chemotherapy	3	Urothelial	270	272	0.73	0.59–0.91	0.98	0.81–1.19
Larkin et al. (29)	2018	Nivolumab	Chemotherapy	3	Melanoma	272	133	0.95	0.70–1.29	1	0.78–1.44
Paz-Ares et al. (22)	2019	Nivolumab + chemo	Chemotherapy	3	NSCLC	377	388	0.81	0.67–0.97	0.62	0.52–0.73
Owonikoko et al. (30)	2019	Ipilimumab	Placebo	3	SCLC	278	278	0.84	0.69–1.02	0.67	0.56–0.81
Rudin et al. (24)	2020	Pembrolizumab + etoposide	Placebo+ etoposide	3	SCLC	228	225	0.8	0.64–0.98	0.75	0.61–0.91
Galsky et al. (23)	2020	Atezolizumab + chemotherapy	Placebo+ chemotherapy	3	Urothelial	451	400	0.8	0.70–0.96	0.83	0.69–1.0
Spigel et al. (31)	2021	Nivolumab	Chemotherapy	3	SCLC	284	285	0.86	0.72–1.04	1.41	1.18–1.69

As shown in **Table 1**, a total of 12,126 participants (6,450 cases and 5,676 controls) from 20 articles were included in the meta-analysis. The name of the first author, the publication year, the tumor type of the study, the phase of the RCTs, the name of the ICIs (ipilimumab, nivolumab, pembrolizumab, or atezolizumab) in the experimental groups and non-ICI therapies in the control groups, the number of patients in the ICIs and control groups, and the HR of OS and PFS.

TABLE 2 | The summary of OS and PFS heterogeneity test.

Subgroup	OS		PFS	
	I ²	p	I ²	p
Melanoma	38.00%	0.16	73.00%	<0.01
SCLC	0.00%	0.83	28.00%	0.24
NSCLC	0.00%	0.48	45.00%	0.11
Prostate	77%	0.04	0.00%	0.72
Urothelial	0.00%	0.38	61.00%	0.11
Total	0.41	0.02	0.58	<0.01

heterogeneity in total with $I^2 = 58\%$, $p < 0.01$, and we chose to select the random-effect model according to the method we used. In the tumor subgroup heterogeneity test, we did not find obvious significant heterogeneity in the SCLC ($I^2 = 28\%$, $p = 0.24$), NSCLC ($I^2 = 45\%$, $p = 0.11$) and prostate cancer ($I^2 = 0\%$, $p = 0.72$) subgroups, but a significant heterogeneity was found in the melanoma ($I^2 = 73\%$, $p < 0.01$) and urothelial cancer ($I^2 = 61\%$, $p = 0.11$) subgroups.

Publication Bias Analysis and Sensitivity Analysis

The p -values of Begg's test and Egger's test were applied for OS and PFS. We did not find publication bias in OS by Begg's test ($p = 0.5436$) and Egger's test ($p = 0.6849$), and in PFS by Begg's test ($p = 0.9483$) and Egger's test ($p = 0.9774$). The result of the OS and PFS publication bias analysis is directly reflected in **Figure 2** by using Begg's funnel plot.

Association of ICIs With Overall Survival

The OS analysis was included in 23 studies, and the PFS analysis was included in 20 studies (shown in **Table 1**). **Figure 3** shows the results of OS, and **Figure 4** shows the results of PFS. **Table 3** shows the summary of the melanoma, SCLC, NSCLC, prostate cancer, and urothelial cancer ($I^2 = 0\%$, $p = 0.73$) subgroups, but a significant heterogeneity was found in the prostate cancer subgroup meta-analysis and overall meta-analysis.

In the OS analysis, the ICIs were associated with substantially ameliorated OS (HR = 0.79, CI = 0.74–0.84), compared with non-ICI therapies. In the subgroup analyses, melanoma, SCLC, NSCLC, and urothelial cancer patients treated with ICIs were associated more with OS compared with non-ICI therapies (HR = 0.78, CI = 0.69–0.89; HR = 0.87, CI = 0.80–0.95; HR = 0.71, CI = 0.66–0.77; HR = 0.79, CI = 0.68–0.91),

respectively. However, prostate cancer was not significantly associated with improved OS (HR = 0.95, CI = 0.71–1.26).

Association of ICIs With Progression-Free Survival

In the PFS analysis, the ICIs were associated with significantly improved PFS (HR = 0.80, CI = 0.75–0.86), compared with non-ICI therapies. In subgroup analyses, melanoma, SCLC, NSCLC, and prostate cancer patients treated with ICIs were associated more with PFS compared with non-ICI therapies (HR = 0.78, CI = 0.63–0.98; HR = 0.78, CI = 0.69–0.89; HR = 0.86, CI = 0.77–0.95; HR = 0.69, CI = 0.61–0.77), respectively. However, urothelial cancer was not significantly associated with improved PFS (HR = 0.88, CI = 0.72–1.05).

DISCUSSION

In our meta-analysis, a total of 12,126 participants (6,450 cases and 5,676 controls), treated with ICIs and non-ICI arms, from 20 articles were included.

In total, among 12,126 patients in our meta-analysis, 2,423 patients (1,514 cases and 969 controls) were included into the melanoma subgroup, 2,727 patients (1,353 cases and 1,374 controls) were included into the SCLC subgroup, 4,122 patients (2,063 cases and 2,059 controls) were included into the NSCLC subgroup, 1,401 patients (799 cases and 602 controls) were included into the prostate subgroup, and 1,393 patients (721 cases and 672 controls) were included into the urothelial cancer subgroup.

To our knowledge, this is the comprehensive meta-analysis to assess the efficacy of ICIs (ipilimumab, pembrolizumab, nivolumab, and atezolizumab) in different types of tumors, including melanoma, SCLC, NSCLC, prostate cancer, and urothelial cancer. Results of trials on ICIs have been published, while the clinical value of ICIs is still controversial. To further investigate the efficacy of

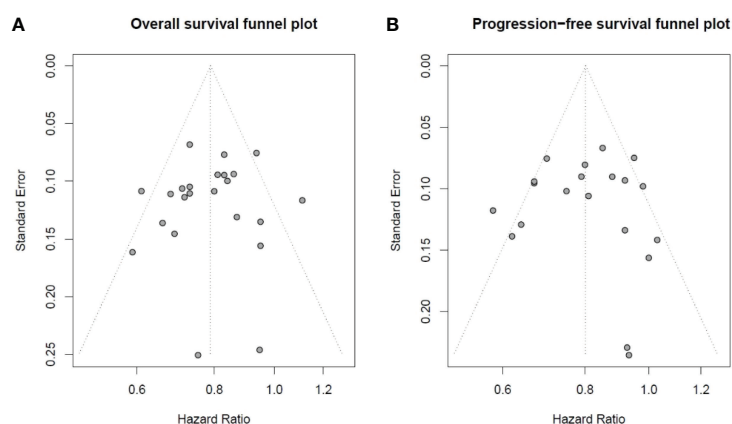


FIGURE 2 | Begg's funnel plot of overall survival and progression-free survival studies: **(A)** Begg's funnel plot of overall survival studies to evaluate publication bias. **(B)** Begg's funnel plot of progression-free survival studies to evaluate publication bias.

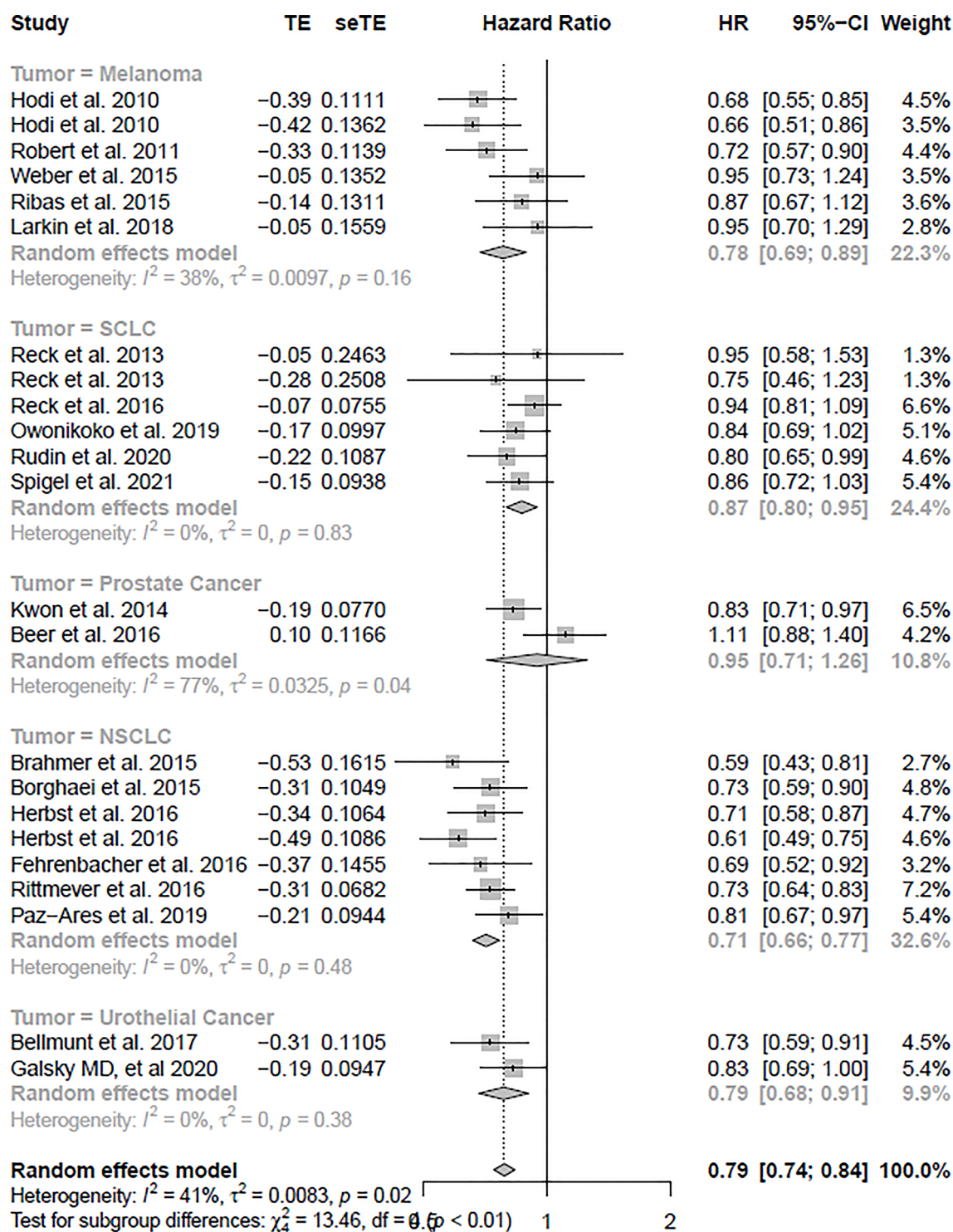


FIGURE 3 | The forest plot of OS in the random-effect model.

ICIs, we made five subgroups of melanoma, SCLC, NSCLC, prostate cancer, and urothelial cancer with OS and PFS.

The pooled analyses indicated that ICIs were associated with obviously ameliorated PFS and OS compared with non-ICI

arms. In OS subgroup analyses, melanoma, SCLC, NSCLC, and urothelial cancer patients treated with ICIs were associated more with OS compared with non-ICI therapies. However, prostate cancer was not significantly associated with improved

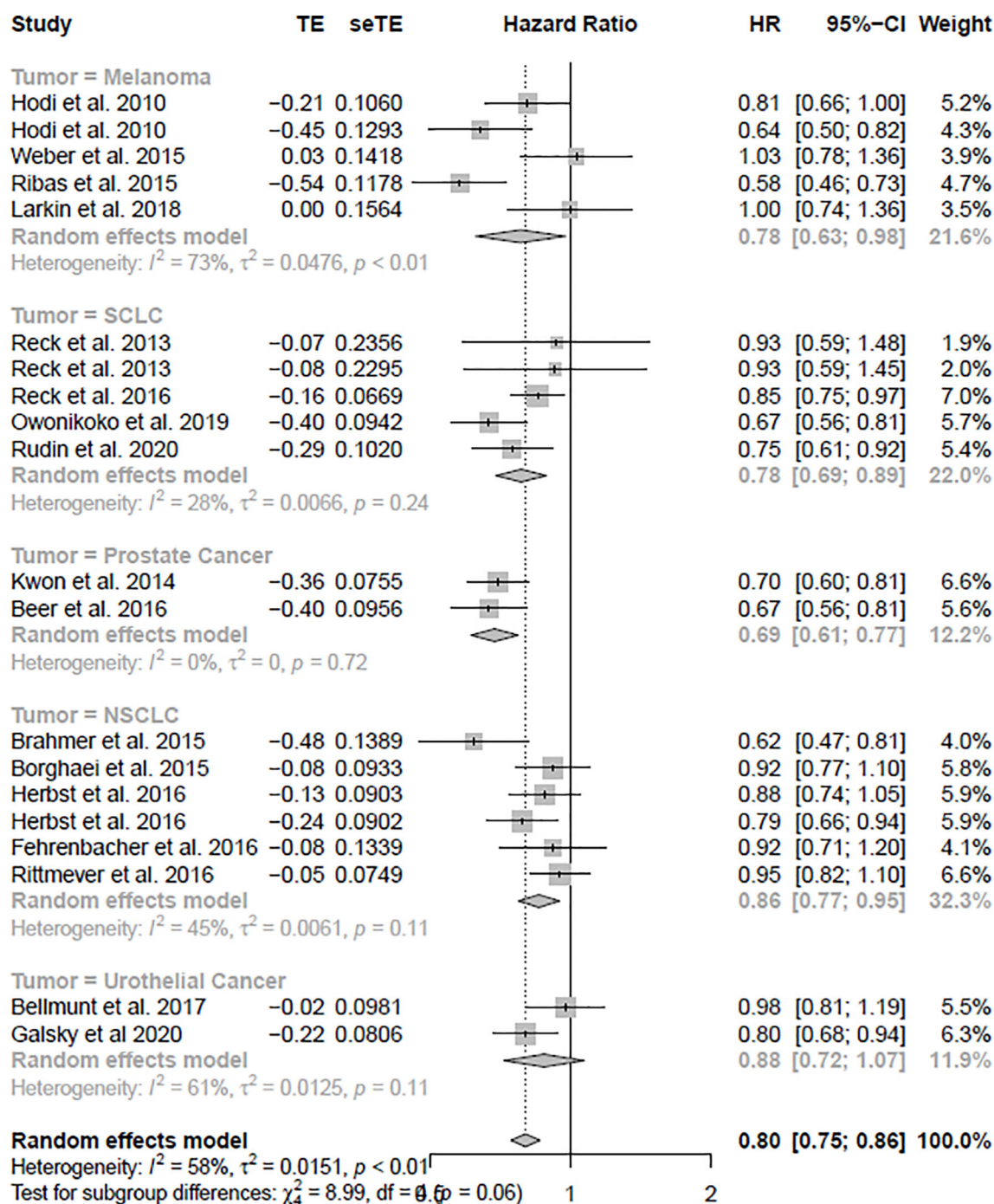


FIGURE 4 | The forest plot of PFS in the random effect model.

OS. In PFS subgroup analyses, melanoma, SCLC, NSCLC, and prostate cancer patients treated with ICIs were associated more with PFS compared with non-ICI therapies. However, urothelial cancer was not significantly associated with improved PFS.

However, the meta-analysis had some limitations. To begin with, the number of participants in our meta-

analysis was 12,126, and more studies should be added to this meta-analysis. Second, some heterogeneity existed in this meta-analysis, especially in the PFS group. It should be solved in the further study. Besides, more studies should be added into the prostate cancer patients and urothelial cancer subgroups.

TABLE 3 | The summary of the meta-analysis with OS and PFS.

Subgroup	OS		PFS	
	HR	95% CI	HR	95% CI
Melanoma	0.78	0.69–0.89	0.78	0.63–0.98
SCLC	0.87	0.80–0.95	0.78	0.69–0.89
NSCLC	0.71	0.66–0.77	0.86	0.77–0.95
Prostate	0.95	0.71–1.26	0.69	0.61–0.77
Urothelial	0.79	0.68–0.91	0.88	0.72–1.07
Total	0.79	0.74–0.84	0.80	0.75–0.86

CONCLUSIONS

This meta-analysis got a conclusion that immune checkpoint inhibitors were associated with obviously ameliorated PFS and OS compared with non-ICI therapies.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

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

REFERENCES

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660
- Pauczek RD, Baltimore D, Li G. The Cellular Immunotherapy Revolution: Arming the Immune System for Precision Therapy. *Trends Immunol* (2019) 40(4):292–309. doi: 10.1016/j.it.2019.02.002
- Li L, Jiang M, Qi L, Wu Y, Song D, Gan J, et al. Pyroptosis, a New Bridge to Tumor Immunity. *Cancer Sci* (2021) 112(10):3979–94. doi: 10.1111/cas.15059
- Schlotter CM, Tietze L, Vogt U, Heinsen CV, Hahn A. Ki67 and Lymphocytes in the Pretherapeutic Core Biopsy of Primary Invasive Breast Cancer: Positive Markers of Therapy Response Prediction and Superior Survival. *Hormone Mol Biol Clin Invest* (2017) 32(2):20170022. doi: 10.1515/hmbci-2017-0022
- Zhou Z, Li M. Evaluation of BRCA1 and BRCA2 as Indicators of Response to Immune Checkpoint Inhibitors. *JAMA Netw Open* (2021) 4(5):e217728. doi: 10.1001/jamanetworkopen.2021.7728
- Marrone KA, Ying W, Naidoo J. Immune-Related Adverse Events From Immune Checkpoint Inhibitors. *Clin Pharmacol Ther* (2016) 100(3):242–51. doi: 10.1002/cpt.394
- De Velasco G, Je Y, Bossé D, Awad MM, Ott PA, Moreira RB, et al. Comprehensive Meta-Analysis of Key Immune-Related Adverse Events From CTLA-4 and PD-1/PD-L1 Inhibitors in Cancer Patients. *Cancer Immunol Res* (2017) 5(4):312–8. doi: 10.1158/2326-6066.CIR-16-0237

AUTHOR CONTRIBUTIONS

WD and LX wrote the manuscript. WD and XY collected the data and conducted the experiment. LX performed the project. WD interpreted the results. WD and LX developed the analytical tools. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded in part by the National Natural Science Foundation of China, grant number 62172122, and the Fundamental Research Funds for the Central Universities, Jilin University, grant number 93K172021K04.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.876098/full#supplementary-material>

- Ledford H. Melanoma Drug Wins US Approval. *Nature* (2011) 471(7340):561. doi: 10.1038/471561a
- Kazandjian D, Suzman DL, Blumenthal G, Mushti S, He K, Libeg M, et al. FDA Approval Summary: Nivolumab for the Treatment of Metastatic Nonsmall Cell Lung Cancer With Progression on or After Platinum-Based Chemotherapy. *Oncologist* (2016) 21(5):634–42. doi: 10.1634/theoncologist.2015-0507
- Ma W, Gilligan BM, Yuan J, Li T. Current Status and Perspectives in Translational Biomarker Research for PD-1/PD-L1 Immune Checkpoint Blockade Therapy. *J Hematol Oncol* (2016) 9:47. doi: 10.1186/s13045-016-0277-y
- Hazarika M, Chuk MK, Theoret MR, Mushti S, He K, Weis SL, et al. US FDA Approval Summary: Nivolumab for Treatment of Unresectable or Metastatic Melanoma Following Progression on Ipilimumab. *Clin Cancer Res* (2017) 23(14):3484–8. doi: 10.1158/1078-0432.CCR-16-0712
- Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, et al. Improved Survival With Ipilimumab in Patients With Metastatic Melanoma. *N Engl J Med* (2010) 363:711–23. doi: 10.1056/NEJMoa1003466
- Robert C, Thomas L, Bondarenko I, O'Day S, Weber J, Garbe C, et al. Ipilimumab Plus Dacarbazine for Previously Untreated Metastatic Melanoma. *N Engl J Med* (2011) 364(26):2517–26. doi: 10.1056/NEJMoa1104621
- Borghaei H, Paz-Ares L, Horn L, Spigel DR, Steins M, Ready NE, et al. Nivolumab Versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer. *N Engl J Med* (2015) 373(17):1627–39. doi: 10.1056/NEJMoa1507643

15. Brahmer J, Reckamp KL, Baas P, Crinò L, Eberhardt WEE, Poddubskaya E, et al. Nivolumab Versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. *N Engl J Med* (2015) 373(2):123–35. doi: 10.1056/NEJMoa1504627
16. Ribas A, Puzanov I, Dummer R, Schadendorf D, Hamid O, Robert C, et al. Pembrolizumab Versus Investigator-Choice Chemotherapy for Ipilimumab-Refractory Melanoma (KEYNOTE-002): A Randomised, Controlled, Phase 2 Trial. *Lancet Oncol* (2015) 16(8):908–18. doi: 10.1016/S1470-2045(15)00083-2
17. Weber JS, D'Angelo SP, Minor D, Hodi FS, Gutzmer R, Neyns B, et al. Nivolumab Versus Chemotherapy in Patients With Advanced Melanoma Who Progressed After Anti-CTLA-4 Treatment (CheckMate 037): A Randomised, Controlled, Open-Label, Phase 3 Trial. *Lancet Oncol* (2015) 16(4):375–84. doi: 10.1016/S1470-2045(15)70076-8
18. Fehrenbacher L, Spira A, Ballinger M, Kowanzet M, Vansteenkiste J, Mazieres J, et al. Atezolizumab Versus Docetaxel for Patients With Previously Treated Non-Small-Cell Lung Cancer (POPLAR): A Multicentre, Open-Label, Phase 2 Randomised Controlled Trial. *Lancet* (2016) 387(10030):1837–46. doi: 10.1016/S0140-6736(16)00587-0
19. Herbst RS, Baas P, Kim DW, Felip E, Pérez-Gracia JL, Han JY, et al. Pembrolizumab Versus Docetaxel for Previously Treated, PD-L1-Positive, Advanced Non-Small-Cell Lung Cancer (KEYNOTE-010): A Randomised Controlled Trial. *Lancet* (2016) 387(10027):1540–50. doi: 10.1016/S0140-6736(15)01281-7
20. Bellmunt J, De Wit R, Vaughn DJ, Fradet Y, Lee JL, Fong L, et al. Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma. *N Engl J Med* (2017) 376(11):1015–26. doi: 10.1056/NEJMoa1613683
21. Rittmeyer A, Barlesi F, Waterkamp D, Park K, Ciardiello F, Von Pawel J, et al. Atezolizumab Versus Docetaxel in Patients With Previously Treated Nonsmall-Cell Lung Cancer (OAK): A Phase 3, Open-Label, Multicentre Randomised Controlled Trial. *Lancet* (2017) 389(10066):255–65. doi: 10.1016/S0140-6736(16)32517-X
22. Paz-Ares L, Ciuleanu TE, Yu X, Salman PA. LBA3 Nivolumab (NIVO)+ Platinum-Doublet Chemotherapy (Chemo) vs Chemo as First-Line (1L) Treatment (Tx) for Advanced Non-Small Cell Lung Cancer (aNSCLC): CheckMate 227-Part 2 Final Analysis. *Ann Oncol* (2019) 30:xi67–8. doi: 10.1093/annonc/mdz453.004
23. Galsky MD, Ariba JÁV, Bamias A, Davis ID, De Santis M, Kikuchi E, et al. Atezolizumab With or Without Chemotherapy in Metastatic Urothelial Cancer (IMvigor130): A Multicentre, Randomised, Placebo-Controlled Phase 3 Trial. *Lancet* (2020) 10236:1547–57. doi: 10.1016/S0140-6736(20)30230-0
24. Rudin CM, Awad MM, Navarro A, Gottfried M, Peters S, Csösz T, et al. Pembrolizumab or Placebo Plus Etoposide and Platinum as First-Line Therapy for Extensive-Stage Small-Cell Lung Cancer: Randomized, Double-Blind, Phase III KEYNOTE-604 Study. *J Clin Oncol* (2020) 38(21):2369. doi: 10.1200/JCO.20.00793
25. Reck M, Bondarenko I, Luft A, Serwatowski P, Barlesi F, Chacko R, et al. Ipilimumab Incombination With Paclitaxel and Carboplatin as First-Line Therapy in Extensive-Disease-Small-Cell Lung Cancer: Results From a Randomized, Double-Blind, Multicenter Phase 2 Trial. *Ann Oncol* (2013) 24(1):75–83. doi: 10.1093/annonc/mds213
26. Kwon ED, Drake CG, Scher HI, Fizazi K, Bossi A, Van den Eertwegh AJM, et al. Ipilimumab Versus Placebo After Radiotherapy in Patients With Metastatic Castration-Resistant Prostate Cancer That had Progressed After Docetaxel Chemotherapy (CA184-043): A Multicentre, Randomised, Double-Blind, Phase 3 Trial. *Lancet Oncol* (2014) 15(7):700–12. doi: 10.1016/S1470-2045(14)70189-5
27. Beer TM, Kwon ED, Drake CG, Fizazi K, Logothetis C, Gravis G, et al. Randomized, Double-Blind, Phase III Trial of Ipilimumab Versus Placebo in Asymptomatic or Minimally Symptomatic Patients With Metastatic Chemotherapy-Naive Castration-Resistant Prostate Cancer. *J Clin Oncol* (2016) 35(1):40–7. doi: 10.1200/JCO.2016.69.1584
28. Reck M, Luft A, Szczesna A, Havel L, Kim SW, Akerley W, et al. Phase III Randomized Trial of Ipilimumab Plus Etoposide and Platinum Versus Placebo Plus Etoposide and Platinum in Extensive-Stage Small-Cell Lung Cancer. *J Clin Oncol* (2016) 34(31):3740–8. doi: 10.1200/JCO.2016.67.6601
29. Larkin J, Minor D, D'Angelo S, Neyns B, Smylie M, Miller WH Jr., et al. Overall Survival in Patients With Advanced Melanoma Who Received Nivolumab Versus Investigator's Choice Chemotherapy in CheckMate 037: A Randomized, Controlled, Open-Label Phase III Trial. *J Clin Oncol* (2018) 36(4):383. doi: 10.1200/JCO.2016.71.8023
30. Owonikoko TK, Kim HR, Govindan R, Ready N, Reck M, Peters S, et al. Nivolumab (Nivo) Plus Ipilimumab (Ipi), Nivo, or Placebo (Pbo) as Maintenance Therapy in Patients (Pts) With Extensive Disease Small Cell Lung Cancer (ED-SCLC) After First-Line (1L) Platinum-Based Chemotherapy (Chemo): Results From the Double-Blind, Randomized Phase III CheckMate 451 Study. *Ann Oncol* (2019) 30:ii77. doi: 10.1093/annonc/mdz094
31. Spigel DR, Vicente D, Ciuleanu TE, Gettinger S, Peters S, Horn L, et al. Second-Line Nivolumab in Relapsed Small-Cell Lung Cancer: CheckMate 331. *Ann Oncol* (2021) 32:5:631–41. doi: 10.1016/j.annonc.2021.01.071
32. Egger M, Davey Smith G, Schneider M, Minder C. Bias in Meta-Analysis Detected by a Simple Graphical Test. *BMJ* (1997) 315(7109):629–34. doi: 10.1136/bmj.315.7109.629
33. Begg CB, Mazumdar M. Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics* (1994) 50(4):1088–101. doi: 10.2307/2533446

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.

The reviewer XL declared a shared affiliation with the author XL to the handling editor at the time of review.

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.

Copyright © 2022 Xu, Yan and Ding. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



ZKSCAN5 Activates VEGFC Expression by Recruiting SETD7 to Promote the Lymphangiogenesis, Tumour Growth, and Metastasis of Breast Cancer

Jingtong Li^{1†}, Zhifeng Yan^{2†}, Jianli Ma^{3†}, Zhong Chu¹, Huizi Li⁴, Jingjing Guo⁵, Qingyuan Zhang^{1*}, Hui Zhao^{5*}, Ying Li^{6*} and Tao Wang^{6*}

OPEN ACCESS

Edited by:

Xin Zhang,
Jiangmen Central Hospital, China

Reviewed by:

Guangchao Sui,
Northeast Forestry University, China
Jian Zhang,
Fourth Military Medical University,
China

*Correspondence:

Qingyuan Zhang
0566@hrbmu.edu.cn
Hui Zhao
Lyuyww1996@sina.com
Ying Li
2321211929@qq.com
Tao Wang
13910928773@163.com

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

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 13 February 2022

Accepted: 15 March 2022

Published: 05 May 2022

Citation:

Li J, Yan Z, Ma J, Chu Z, Li H, Guo J,
Zhang Q, Zhao H, Li Y and Wang T
(2022) ZKSCAN5 Activates
VEGFC Expression by Recruiting
SETD7 to Promote the
Lymphangiogenesis, Tumour Growth,
and Metastasis of Breast Cancer.
Front. Oncol. 12:875033.
doi: 10.3389/fonc.2022.875033

¹ Department of Medical Oncology, Harbin Medical University Cancer Hospital, Harbin, China, ² Department of Obstetrics and Gynecology, Seventh Medical Center of Chinese People's Liberation Army (PLA) General Hospital, Beijing, China, ³ Department of Radiation Oncology, Harbin Medical University Cancer Hospital, Harbin, China, ⁴ Department of Nutrition, People's Liberation Army (PLA) Rocket Force Characteristic Medical Center, Beijing, China, ⁵ Department of Oncology, Fourth Medical Center of Chinese People's Liberation Army (PLA) General Hospital, Beijing, China, ⁶ Department of Oncology, Fifth Medical Center of Chinese People's Liberation Army (PLA) General Hospital, Beijing, China

The growth of lymphatic vessels (lymphangiogenesis) plays a pivotal role in breast cancer progression and metastasis and the immune response. Vascular endothelial growth factor C (VEGFC) has been demonstrated to accelerate cancer metastasis and modulate the immune system by enhancing lymphangiogenesis. However, it remains largely unclear how transcription factors physically regulate VEGFC expression by interacting with histone-modifying enzymes. Like many histone-modifying enzymes, SETD7 plays a key role in cell proliferation and inhibits tumour cell differentiation. In this study, we identified the role of the transcription factor zinc finger with KRAB and SCAN domains 5 (ZKSCAN5) in interacting with histone methyltransferase SETD7 and mediating VEGFC transcription and tumour lymphangiogenesis. ZKSCAN5 interacts with and recruits SETD7 to the VEGFC promoter. By regulating breast cancer-secreted VEGFC, ZKSCAN5 could induce the tube formation of lymph endothelial cells, which promotes tumour proliferation, migration, and metastasis. Clinically, the expression of ZKSCAN5 was frequently upregulated in patients with breast cancer and positively correlated with the expression of VEGFC and the number of lymphatic microvessels. ZKSCAN5 is a poor prognostic factor for patients with breast cancer. Our results characterise the role of ZKSCAN5 in regulating VEGFC transcription and predict ZKSCAN5 as a breast cancer therapeutic target.

Keywords: VEGFC, lymphangiogenesis, breast cancer, proliferation and metastasis, SETD7, ZKSCAN5

INTRODUCTION

Breast cancer is the leading cause of death among women worldwide (1, 2). Lymphangiogenesis, a pivotal component for tumour metastasis, immune escape, and growth, has frequently been shown to occur in human breast cancer (3). Vascular endothelial growth factor C (VEGFC), a member of the VEGF family, is an important regulator of lymphangiogenesis (4–6). VEGFC stimulates the

formation of new lymph vessels and provides a route for detached cancer cells to metastasise to distant sites (7). Numerous experiments have demonstrated that tumour-secreted VEGFC is a key cytokine involved in tumour development and the immune response (8, 9). A high VEGFC level is associated with significantly decreased overall and disease-free survival in many solid tumours (10, 11). Thus, VEGFC appears to be an attractive therapeutic target for cancers. Thus, discovering novel factors that regulate VEGFC expression is of great significance.

The transcriptional regulation of VEGFC is one of the most significant ways to control VEGFC expression (12). A small number of transcription factors have been reported to regulate VEGFC expression at the mRNA level, including Six1 (13) and forkhead box k1 (FOKK1) (14). However, novel VEGFC transcriptional factors controlling the VEGFC mRNA level remain largely unknown.

Zinc finger with KRAB and SCAN domains 5 (ZKSCAN5) is a transcription factor that belongs to one of the Krüppel-like zinc finger family members. ZKSCAN5 is pivotal in the process of spermatogenesis (15). It has been validated that ZKSCAN5 is closely linked with oesophageal squamous cell carcinoma tumorigenesis (16). However, the biological functions of ZKSCAN5 remain largely unknown. SETD7 interacts with and methylates a large number of transcription factors, such as BRG1 (17), E2F1 (18), and SMAD3 (19). SETD7-mediated methylation could facilitate the recruitment of transcription factors to chromatin (20, 21).

In this study, we found that ZKSCAN5 interacts with SETD7 and increases VEGFC transcription by facilitating the recruitment of the ZKSCAN5/SETD7 complex to the VEGFC promoter. In addition, ZKSCAN5 has been recognised as a novel critical regulator for the expression of VEGFC and contributes to tumour lymphangiogenesis. ZKSCAN5 promotes the proliferation, migration, and tube formation of human lymphocyte endothelial cells (HLECs). Furthermore, ZKSCAN5 is positively correlated with the expression of VEGFC and could be a valuable prognostic marker for poor survival of breast cancer.

MATERIALS AND METHODS

Plasmids, Antibodies, siRNAs, and Reagents

PCR-amplified fragments were inserted into pGEX-KG (Amersham Pharmacia Biotech, Amersham, UK) or pET-28a (Novagen) to produce plasmids expressing fusion proteins of GST or His. The FLAG-tagged ZKSCAN5 and SETD7 as well as the MYC-tagged ZKSCAN5 and SETD7 eukaryotic expression vectors were constructed by cloning PCR-amplified sequences into pcDNA3 (Invitrogen, Carlsbad, CA, USA). The luciferase reporters of the VEGFC promoter were constructed by cloning promoter DNA fragments obtained from genomic DNA into the pGL4-Basic vector (Promega, Madison, WI, USA).

Anti-Flag (A8592), anti-GAPDH (G9295), anti-Flag M2 agarose (A2220), anti-ZKSCAN5 (SAB4501021), and anti-SETD7 (SAB1306218) antibodies were obtained from Sigma-Aldrich (St. Louis, MO, USA); anti-Myc (sc-40HRP) antibody was obtained from Santa Cruz Biotechnology (Dallas, TX, USA); anti-H3K4me2 (17–677) and anti-H3K4me3 (17–678) antibodies were obtained from Millipore (Burlington, MA, USA); anti-H3K4me (ab8895), anti-SETD7 (ab14820), anti-VEGFC (ab83905), and anti-LYVE1 (ab10278) antibodies were obtained from Abcam (Cambridge, MA, USA); anti-SET1 (A300-289A) and anti-mixed lineage leukaemia protein 1 (MLL1; A300-374A) antibodies were obtained from Bethyl (Montgomery, TX, USA); and anti-His (27471001) and anti-GST (RPN1236) antibodies were obtained from GE Healthcare Life Sciences (Chicago, IL, USA).

The sequences of ZKSCAN5 and SETD7, both short hairpin RNAs (shRNAs) and siRNAs, are provided in **Supplementary Table S1**. A lentiviral pSIH-H1-Puro vector was used to express shRNAs, and stable cell lines were generated using lentiviral transduction (System Biosciences, Palo Alto, CA, USA). siRNAs were chemically synthesised (GenePharma, Shanghai). (R)-PFI-2 (HY-18627A) was obtained from MedChemExpress (Princeton, NJ, USA). GSK-LSD1 2HCL (S7574) and CPI-455 HCL (S8287) were obtained from Selleck (Houston, TX, USA).

Cell Culture, Transfection, and Luciferase Reporter Assay

Human embryonic kidney 293T cells, breast cancer cells ZR75-1 (ER+), and MDA-MB-231 (ER-) were purchased from ATCC and cultured in DMEM (Invitrogen) with 10% FBS (HyClone, Logan, UT, USA). Lipofectamine 2000 Reagent (Invitrogen) was used for transfection. Integration of lentiviruses was achieved by co-transfecting recombinant lentivirus vectors and pPACK Packaging Plasmid Mix (System Biosciences) into 293T cells using the MegaTran Reagent (OriGene, Rockville, MD, USA). Stable cell lines were kept for approximately 2 months in 1 µg/ml puromycin. The Dual Luciferase Reporter Assay System from Promega was used to perform luciferase reporter assays.

Screening for Transcription Factors Regulating the VEGFC Promoter

High-throughput screening assays were performed according to the manufacturer's instructions (OriGene). In brief, screening assay reagents were added to each 384-well plate containing VEGFC-Luc reporter vector (100 ng), galactosidase reporter (100 ng), and distinct cDNA plasmids (60 ng). The mixture was kept at room temperature for 20 min until complex formation, and ZR75-1 cells were added at a density of 7,500 cells/well. After 48 h of incubation, the cells were collected, and subsequently, luciferase activities were analysed.

Real-Time Reverse Transcription-PCR

Cellular RNA was isolated by using the TRIzol reagent (Invitrogen). Using the Quantscript RT Kit (Promega), reverse transcription of the extracted RNA into cDNA was performed. The relative expression of VEGFC was normalised to β -actin

expression. The primers used for quantitative real-time reverse transcription (qRT-PCR) were as follows: VEGFC-forward: 5'-CTCGGATGCTGGAGATGAC-3', VEGFC-reverse: 5'-GGCTGGGAAGAGTTTGTT-3'.

Wound Healing Assays

A micropipette tip was used to scrape the cells in a six-well plate. The cells were cultured in ZKSCAN5-related conditioned medium. Cell migration was monitored and imaged with a microscope at the indicated times. The cell migratory abilities were recorded and analysed with ImageJ software.

Tube Formation Assay

We placed the thawed extracellular matrix (ECM) gel solution into 96 prechilled sterile well plates, and then they were incubated for 1 h at 37°C to allow the matrix solution to solidify. Cell suspensions of $1.5\text{--}3 \times 10^4$ cells/well were added to the cured ECM gel. The cells were incubated at 37°C for 6–18 h. An inverted microscope was then used to observe and photograph the tube formation.

GST Pull-Down and Coimmunoprecipitation Assays

Purified His or GST fusion proteins bound to GST beads supplemented with protease inhibitors were co-incubated at 4°C for 4 h. After washing, the precipitated components were subjected to Western blot analysis. Cells were harvested and lysed using sonication to perform a coimmunoprecipitation assay. The supernatant of the cell lysates was incubated with antibodies at 4°C overnight, followed by incubation with Protein A Agarose (Santa Cruz) at 4°C overnight. The beads were dissolved in 2× SDS loading buffer after washing thrice with lysis buffer washing. Western blot was performed using specific antibodies as indicated.

Chromatin Immunoprecipitation and Re-ChIP

A Magna ChIP Test Kit (Millipore, Burlington, MA, USA) was used for chromatin immunoprecipitation (ChIP) determination according to the manufacturer's instructions. Briefly, 1×10^7 ZR75-1 cells were cross-linked with 1% formaldehyde (Sigma) at room temperature, and then 0.25 M glycine was added after 10 min. Chromatin was sonicated to a size range of 200–1,000-bp fragments for ChIP analysis. The primary immunoprecipitation complexes were washed, eluted with 10 mM DTT at 37°C for 30 min, and diluted to 1:50 in re-ChIP buffer followed by re-ChIP with the secondary antibodies. Real-time PCR was conducted to detect the relative mRNA expression. **Supplementary Table S2** summarises the primers used for quantitative real-time PCR analysis.

In Vivo Tumour Growth and Metastasis Analysis

The animal study was approved and monitored by the Ethics Committee of Harbin Medical University Cancer Hospital (the ID of animal experiment ethical approval: SYDW2021-056). For *in vivo* tumour estimation, nude mice were inoculated

subcutaneously with 1×10^7 ZR75-1 cells with different constructs on the right side. The tumour size was calculated, and the mice were euthanised at the indicated time. The resected tumour was preserved in liquid nitrogen.

BALB/c mice were injected with 1×10^6 MDA-MB-231 cells labelled with luciferase carrying the indicated constructs into the lateral tail vein. All mice were euthanised after 50 days. All lungs were excised for metastatic foci analysis.

Immunohistochemistry

Primary breast cancer tissues and adjacent normal tissues were obtained from 116 patients at the Harbin Medical University Cancer Hospital (the ID of clinical experiment ethical approval: SYLC2021-063). Informed consent was obtained from the patients, and all study protocols were approved by the Institutional Review or Committees of Harbin Medical University Cancer Hospital. Anti-ZKSCAN5 (SAB4501021), anti-VEGFC (ab83905), and anti-LYVE1 (ab10278) primary antibodies were used at 1:100, 1:100, and 1:50 dilutions, respectively. The H-score of ZKSCAN5 or VEGFC was calculated by multiplying the percentage of positive cells and staining intensity.

Statistical Analyses

Statistical significance was assessed by using the two-tailed Student's *t*-tests. The correlation expression and clinicopathologic characteristics were determined using the Pearson's χ^2 tests. The Kaplan–Meier method was used to estimate the overall and disease-free survival. All calculations were conducted with the SPSS 20.0 software. $p < 0.05$ was considered to indicate statistical significance.

RESULTS

ZKSCAN5 Mediates the Transcription of VEGFC in Breast Cancer Cells

To determine the possible transcription factors regulating VEGFC transcription, we selected a transcription factor from the full-length cDNA transfection array of zr75-1 breast cancer cells from −1,058 to +1 bp by using the VEGFC promoter-luciferase (VEGFC-Luc) reporter. Besides the previously reported transcription factor Six1, we identified a novel transcriptional factor, i.e., ZKSCAN5. With an increase in ZKSCAN5 expression vector transfection doses, the VEGFC-Luc reporter activity gradually increased in both ZR75-1 and MDA-MB-231 cells (**Figure 1A**). By contrast, the knockdown of ZKSCAN5 decreased VEGFC-Luc reporter activity (**Figure 1B**). In accordance with the results of luciferase reporter analysis, knockdown of ZKSCAN5 decreased the VEGFC mRNA level (**Figure 1C**). Since the subcellular localisation of ZKSCAN5 has not been reported, we investigated the subcellular localisation of ZKSCAN5 by performing cytosolic–nuclear separation and immunofluorescence assay. The results showed that ZKSCAN5 was mainly located in the nucleus, which provided the cellular basis of ZKSCAN5 to transcriptionally regulate VEGFC expression (**Figures 1D, E**).

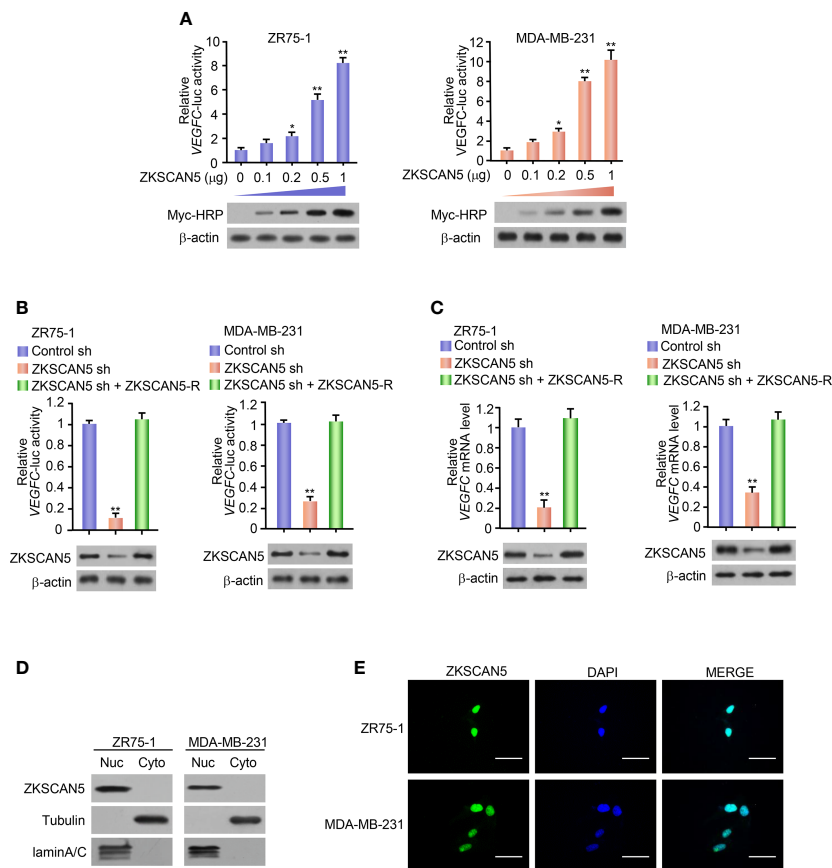


FIGURE 1 | ZKSCAN5 regulates the expression of *VEGFC* in breast cancer cells. **(A)** Luciferase reporter genes were determined in ZR75-1 and MDA-MB-231 breast cancer cells co-transfected with different concentrations of *VEGFC* reporter and myc-ZKSCAN5. A representative immunoblot showed the expression of myc-HRP. β-Actin was used as a control for loading. All values shown are expressed as the average value ± SD obtained from three independent experiments. * $p < 0.05$, ** $p < 0.01$, and empty vector. **(B)** Luciferase reporter gene detection in ZR75-1 and MDA-MB-231 breast cancer cells co-transfected with *VEGFC*-Luc and ZKSCAN5 shRNA, *VEGFC*-Luc and control shRNA, or *VEGFC*-Luc and ZKSCAN5 shRNA plus shRNA-resistant ZKSCAN5 (ZKSCAN5-R). The representative Western blot shows ZKSCAN5 expression. Among them, β-actin was used as the loading control. **(C)** Real-time RT-PCR was used to analyse the *VEGFC* expression in ZR75-1 and MDA-MB-231 cells, which were transfected with ZKSCAN5 shRNA, control shRNA, or ZKSCAN5 shRNA plus shRNA-resistant ZKSCAN5 (ZKSCAN5-R). The representative Western blot further showed the expression of ZKSCAN5. Data shown are the mean ± SD of triplicate measurements from experiments that have been repeated three times with similar results **(B, C)**. ** $p < 0.01$ versus control shRNA. **(D)** Cytoplasmic and nuclear ZKSCAN5 protein levels in two types of breast cancer cell lines, ZR75-1 and MDA-MB-231. Tubulin was used as the cytoplasmic control, and lamin A/C was used as the nuclear protein-loading control. **(E)** Immunofluorescence images of ZKSCAN5 cellular localisation in green, and nuclei stained in blue (DAPI).

ZKSCAN5-Regulated VEGFC Promotes the Proliferation, Migration, and Tube Formation of HLECs

Cancer cell-secreted VEGFC markedly enhanced the proliferation and migration of lymphocyte endothelial cells. Because ZKSCAN5 improved the secretion of VEGFC by breast cancer cells, the effects of the conditioned medium on HLEC proliferation and migration were investigated in ZKSCAN5 knockdown stable cell lines. The ZKSCAN5 knockdown ZR75-1 or MDA-MB-231 cell-conditioned medium decreased HLEC proliferation. The conditioned medium from these cells re-expressing ZKSCAN5 could rescue these effects (Figures 2A, B). A similar tendency was also detected in HLEC migration analysis (Figures 2C, D).

The evolution of capillary lymph ducts by lymphatic endothelial cells is the key aspect of lymphangiogenesis.

Therefore, we examined whether the expression of ZKSCAN5-mediated VEGFC could affect HLEC tube formation *in vitro*. The conditioned medium of ZKSCAN5 knockdown breast cancer cells constrained tube formation, which could be rescued by ZKSCAN5 re-expression in the ZKSCAN5 knockdown cells (Figures 2E, F). Collectively, these results illustrate that ZKSCAN5 enhances the expression of VEGFC and promotes HLEC tube formation and lymphangiogenesis.

ZKSCAN5 Regulates Breast Cancer Tumour Growth and Lung Metastasis *In Vivo*

To determine the phenotype of ZKSCAN5 *in vivo*, we examined the effect of ZKSCAN5 on breast cancer growth by injecting breast cancer cells containing this structure into the back of

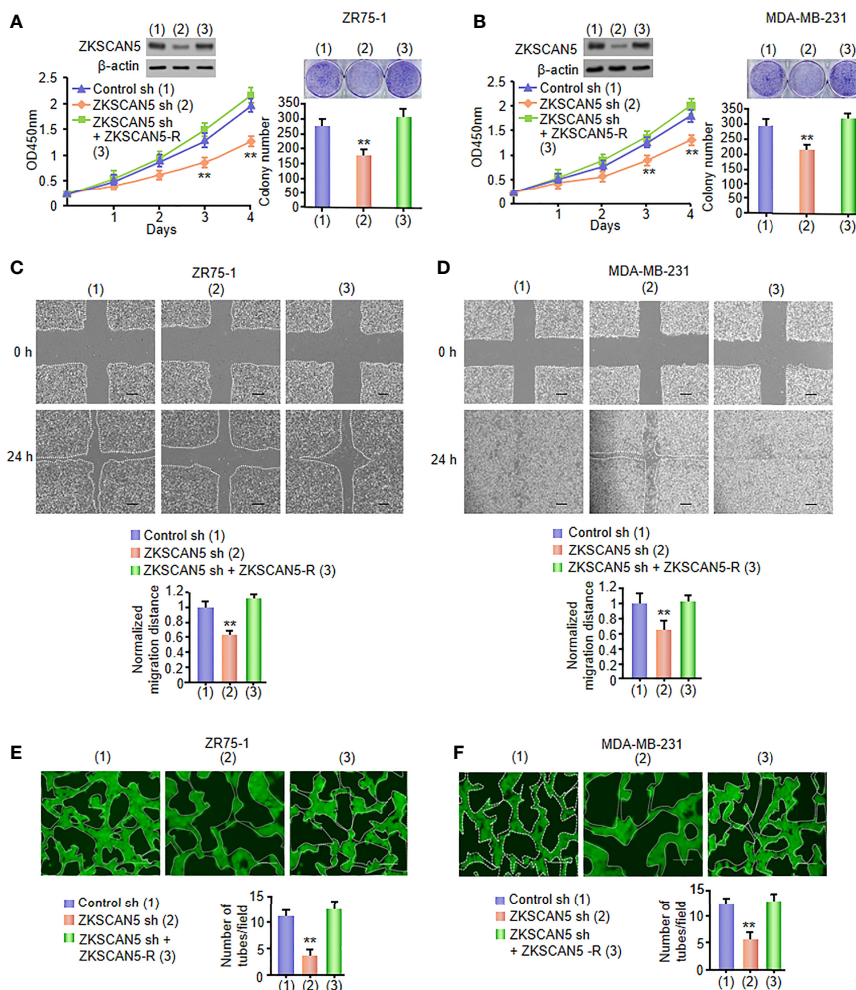


FIGURE 2 | VEGFC secreted by cancer cells, under the influence of ZKSCAN5, regulates HLEC proliferation, migration, and tube formation. **(A, B)** Cell proliferation and colony formation assays in HLECs cultured in conditioned medium come from ZR75-1 or MDA-MB-231 cells stably infected with lentivirus carrying ZKSCAN5 shRNA or ZKSCAN5 shRNA plus ZKSCAN5-R. The representative Western blot displays the expression of ZKSCAN5. ****p < 0.01** versus the control shRNA group **(A, B)**. **(C, D)** Wound healing assays for HLECs cultured in conditioned medium from ZR75-1 or MDA-MB-231 cells, which were stably infected as in **(A)**. The image shown is one of the representative results **(C, D)**. Scale bar: 100 μ m. **(E, F)** Tube formation assays for HLECs cultured in the conditioned medium from ZR75-1 or MDA-MB-231 cells, which were stably infected as in **(A)**. All values shown are the mean \pm SD of triplicate measurements and were repeated three times with analogous results **(C, D)**. ***p < 0.05** versus control shRNA. ****p < 0.01** versus control shRNA.

BALB/C nude mice. As expected, ZKSCAN5 knockdown significantly inhibited the growth of breast cancer tumours. This could be rescued by ZKSCAN5 re-expression in the ZKSCAN5 knockdown cells (**Figures 3A–C**).

Since metastases occur in about 10% of patients with breast cancer, and nearly half of distant metastases occur in the lungs, we investigated the effect of this pathway on breast cancer tumour metastasis. Compared with that in the control group, diffuse pulmonary nodules were significantly reduced in the ZKSCAN5 knockout group. Importantly, ZKSCAN5 re-expression in the ZKSCAN5 knockdown cells dramatically rescued lung metastasis (**Figures 3D, E**). A histological examination of the lungs confirmed the presence of metastases. In conclusion, ZKSCAN5 regulates breast cancer tumour growth and lung metastasis *in vivo*.

ZKSCAN5 Recruits the Histone Methyltransferase SETD7 to the VEGFC Promoter

To further investigate the transcription mechanisms of ZKSCAN5 on regulating VEGFC expression in breast cancer cells, we confirmed the binding site of ZKSCAN5 on the VEGFC promoter. We used JASPAR to predict conserved binding sequences of ZKSCAN5 and its binding sites to the VEGFC promoter (**Figure 4A**). A luciferase assay demonstrated that nucleotides from –658 to –608 bp on the VEGFC promoter contained a possible ZKSCAN5-binding site (**Figure 4B**). ChIP assay revealed that ZKSCAN5 was specifically recruited into the –658- to –608-bp region of the VEGFC promoter, and not the –608- to –558-bp region or 2 kb upstream of the VEGFC promoter (**Figure 4C**).

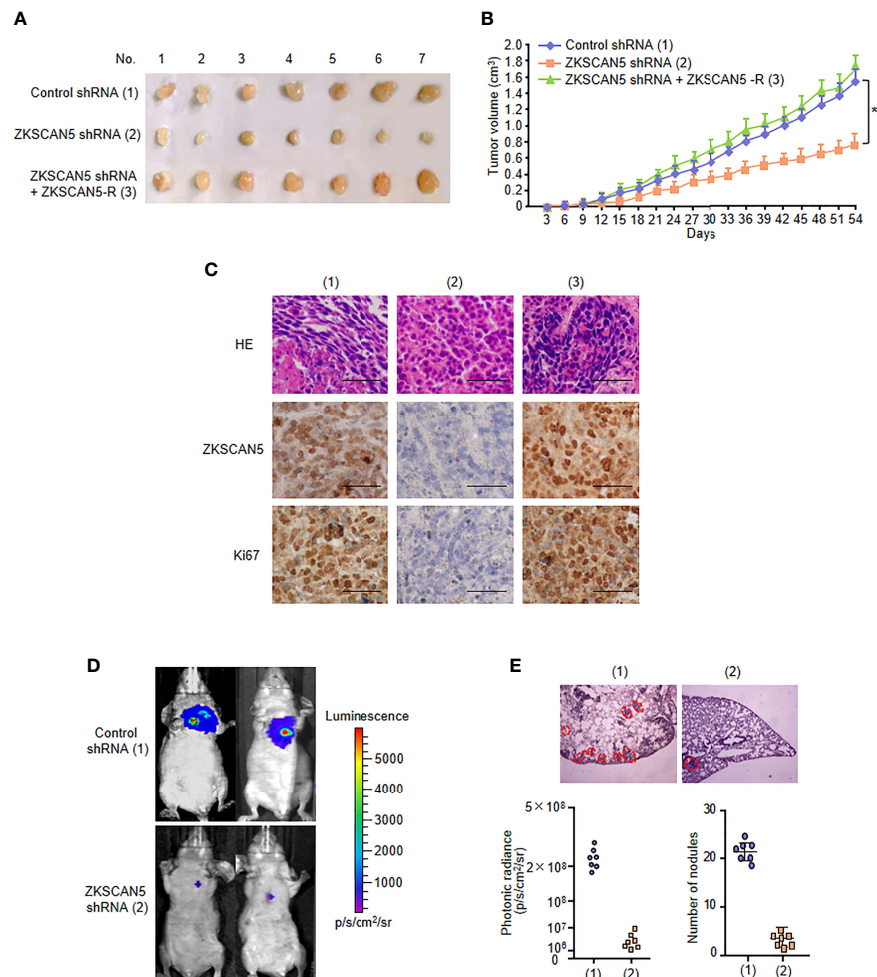


FIGURE 3 | ZKSCAN5 regulates the growth of breast cancer tumours and lung metastasis *in vivo*. **(A, B)** ZR75-1 cells stably infected with the lentivirus carrying the indicated constructs were injected subcutaneously into the nude mice ($n = 7$ per group). The tumour volume was measured every 3 days, and the growth curve was plotted **(B)**. **(C)** Representative IHC staining of ZKSCAN5 and Ki67 and H&E staining images of tumours resected from nude mice. Scale bar, 50 μ m. **(D)** MDA-MB-231 cells stably expressing the constructs were injected through the tail vein to construct a breast cancer cell metastasis model in nude mice ($n = 7$ per group). **(E)** Anatomical and histological analyses of representative lung metastases were carried out. The number of tumour tubercles was determined under an anatomical microscope. Symbols represent individual mice. ** $p < 0.01$ versus the corresponding control.

Transcriptional activation or repression can be led by histone methylation. Methylation of histone H3 at lysine 4 (H3K4) is supposed to be a transcriptional activating mark. Given that ZKSCAN5 benefits VEGFC transcription, we then investigated whether H3K4 methylation enriched the ZKSCAN5-binding region. The specificity of the H3K4 methyl antibodies was validated before the ChIP assay. As expected, GSK-LSD1 (200 μ M, 12 h), an LSD1 inhibitor, specifically increases the levels of H3K4me2 and H3K4me3 but does not affect H3K4me1. CPI-455 (10 μ M, 5 days), the inhibitor of KDM5 demethylases, only increases the level of H3K4me3 but does not affect H3K4me2. These findings demonstrate that the H3K4methyl antibodies used in our experiments are specific without cross-reactivity (**Supplementary Figure S1**). H3K4 dimethylation (H3K4me2) and trimethylation (H3K4me3), but not H3K4 monomethylation

(H3K4me), were enriched at the -658- to -608-bp region, despite the positive control H3K4me being enriched at the GAPDH promoters (**Figure 4D**).

Next, we investigated which histone methyltransferase precisely regulates the dimethylation or trimethylation of H3K4 on the ZKSCAN5-binding region (-658 to -608 bp). Like ZKSCAN5, SETD7 was also recruited to the ZKSCAN5-binding site on the VEGFC promoter (**Figure 4E**). As previously reported (22, 23), although MLL1 and SET1A were recruited to the promoters of homeobox-containing 7 and plasma membrane ATPase 1 separately, they were not recruited to the binding site of the VEGFC promoter (**Figure 4E**). Re-ChIP experiments were performed to determine whether ZKSCAN5 was associated with SETD7 on the -658- to -608-bp region of the VEGFC promoter (**Figure 5A**). Importantly, ZKSCAN5 knockdown reduced the

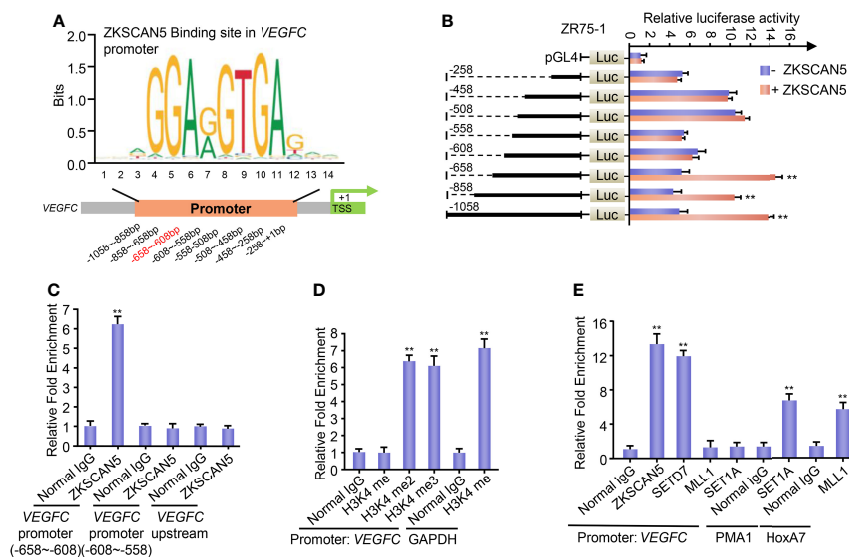


FIGURE 4 | (A) Conserved binding sequences of the transcription factor ZKSCAN5 (JASPAR: <http://jaspar.genereg.net/>) and its binding sites to the VEGFC promoter. **(B)** Luciferase activity of various VEGFC promoter constructs in ZR75-1 cells transfected with ZKSCAN5 or empty vector. Data shown are the mean \pm SD of triplicate measurements and were repeated three times with similar results. ** $p < 0.01$ versus empty vector with corresponding promoter reporter. **(C)** ChIP analysis of the occupancy of ZKSCAN5 on the putative ZKSCAN5-binding sites of the VEGFC promoter in ZR75-1 cells. **(D)** ChIP analysis of the occupancy of H3K4me, H3K4me2, and H3K4me3 on the VEGFC promoter in ZR75-1 cells. The GAPDH promoter has the function of H3K4me2- and H3K4me3-positive control. **(E)** ChIP analysis of the occupancy of ZKSCAN5 and different histone methyltransferases on the VEGFC promoter in ZR75-1 cells. Positive controls of SET1 and MLL1 were promoters of PMA1 and HoxA7, respectively.

recruitment of SETD7, H3K4me2, and H3K4me3, to the -658- to -608-bp region of the VEGFC promoter (**Figure 5B**). Knockout of SETD7 reduced the recruitment of H3K4me2 and H3K4me3 to the -658- to -608-bp region of the VEGFC promoter (**Figure 5B**). The same trend was observed by using (R)-PFI-2 (1 μ M, 2 h), an inhibitor of SETD7 (**Figure 5C**).

Based on the fact that ZKSCAN5 could mediate the enrichment of SETD7, we investigated whether ZKSCAN5 could substantially interact with SETD7. Endogenous ZKSCAN5 pointedly coimmunoprecipitated with endogenous SETD7 using ZR75-1 cells (**Figure 5D**). Since the His-labelled ZKSCAN5 protein interacts with the purified GST-SETD7, but not GST alone, the functional interaction between ZKSCAN5 and SETD7 is explicit (**Figure 5E**). ZKSCAN5 (215–366) contains the SCAN domain related to Set7 but does not contain other ZKSCAN5 deletion mutants (**Figure 5F**). SET7 (215–366) contains the SET fragment (SET), which interacted with ZKSCAN5, whereas SET7 (108–214) containing the middle-region fragment (MF) and SET7 N-terminal region (1–107) containing the NF domain did not (**Figure 5G**). These results show that ZKSCAN5 and SETD7 may construct complexes in the -658- to -608-bp region of the VEGFC promoter.

ZKSCAN5 Positively Correlates With VEGFC Expression and Plays a Prognostic Role in Breast Cancer

We first performed immunohistochemistry (IHC) on 116 human breast cancer samples to demonstrate the clinical significance of ZKSCAN5. Before this test, the specificity of the antibodies for

ZKSCAN5 in IHC was determined by immunoblotting lysates from MDA-MB-231 and ZR75-1 breast cancer cells transfected with ZKSCAN5 siRNAs (**Supplementary Figure S2**). Interestingly, ZKSCAN5 expression increased in cancer tissues compared to that in the adjacent paracancerous tissues ($p = 2.33 \times 10^{-6}$; **Figure 6A**). The associations between ZKSCAN5 expression and lymph vessel number stained by the specific marker LYVE-1 were investigated. ZKSCAN5 was positively related to VEGFC expression in breast cancer tissues ($p = 9.0 \times 10^{-6}$). Tumours with high ZKSCAN5 expression had more lymph vessels compared with low ZKSCAN5 expression (**Figures 6B, C**). Moreover, we observed that higher ZKSCAN5 expression indicated reduced disease-free ($p = 1.842 \times 10^{-4}$) and overall survival ($p = 0.006$; **Figure 6D**). In conclusion, these findings imply the importance of ZKSCAN5 in lymphangiogenesis and the prognosis of breast cancer.

DISCUSSION

Lymphatic vasculature is considered a crucial factor in the modulation of normal homeostasis and many diseases (24). VEGFC is one of the most important regulators of tumour lymphangiogenesis. Emerging evidence shows that various aspects of tumour development can be promoted through the autocrine regulation of VEGFC. It is reported that VEGFC can also regulate the immune system, making it easier for tumour cells to escape immune surveillance. The proliferation and migration of lymphatic

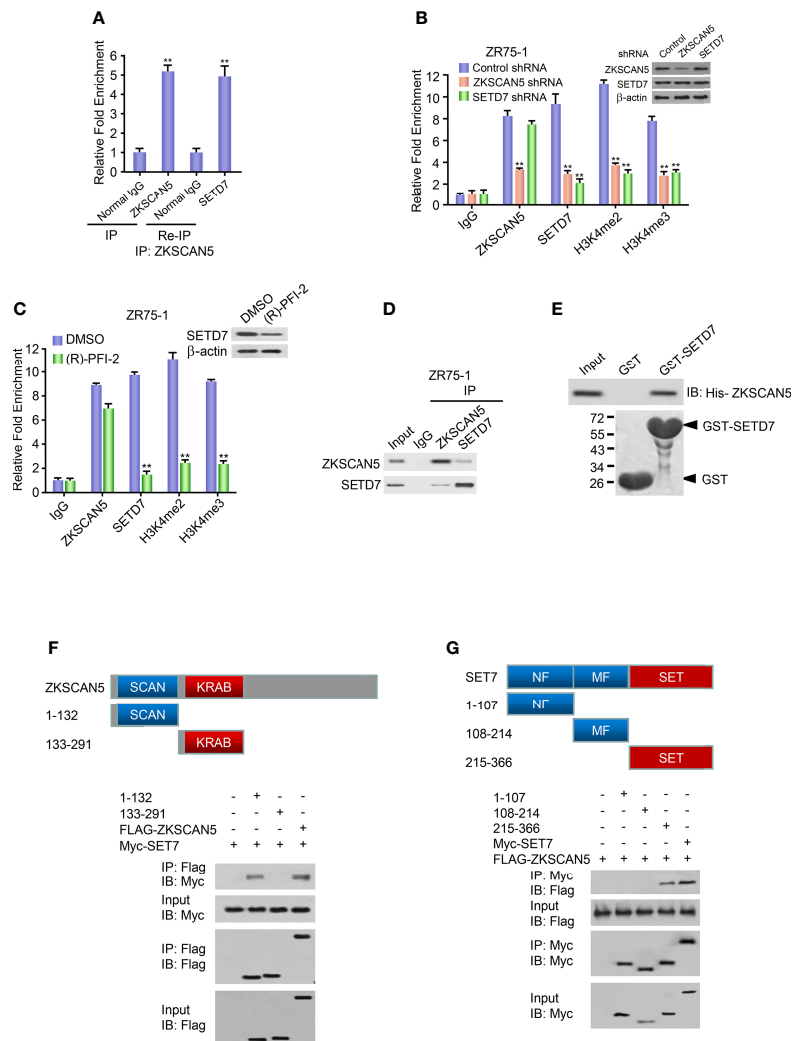


FIGURE 5 | ZKSCAN5 and SETD7 constructed a complex on the -658 to -608-bp region of the VEGFC promoter. **(A)** Re-ChIP analysis of the occupancy of ZKSCAN5 and SETD7 on the VEGFC promoter (-658 to -608 bp) in ZR75-1 cells. **(B)** ChIP analysis of ZR75-1 cells stably infected with lentivirus carrying ZKSCAN5 shRNA or SETD7 shRNA on VEGFC promoter (-658 to -608 bp) with the indicated antibodies. Western blot revealed the knockdown effects of ZKSCAN5 and SETD7. ** $p < 0.01$ versus corresponding control shRNA. **(C)** ChIP analysis using the SETD7 antibody in ZR75-1 cells treated with DMSO or (R)-PFI-2 on the VEGFC promoter (-658 to -608 bp). **(D)** Reciprocal coimmunoprecipitation analysis of endogenous interactions among ZKSCAN5 and SETD7. **(E)** GST pull-down analysis of direct interactions between ZKSCAN5 and SETD7. Purified His-tagged ZKSCAN5 and GST-SETD7 or GST was used. **(F)** Mapping of the interaction region of SET7 in ZKSCAN5. HEK293T cells were co-transfected with MYC-tagged SET7 and FLAG-tagged ZKSCAN5 or its deletion mutants. Anti-FLAG immunoprecipitation was used to precipitate cell lysates, followed by immunoblotting with the specified antibody. The schematic diagram shows ZKSCAN5 and its deletion mutants. **(G)** The mapping highlights the interaction region of ZKSCAN5 in SET7. HEK293T cells were co-transfected with MYC-tagged ZKSCAN5 and FLAG-tagged SET7 or its deletion mutants. Immunoprecipitation of the cell lysate was analysed in (A). The schematic diagram shows SET7 and its deletion mutants; MF, middle region fragment; SET, SET domain-containing fragment. All values shown are the mean \pm SD of triplicate measurements from experiments that have been repeated three times with similar results.

endothelial cells are prerequisites for lymphangiogenesis (25). The expression of VEGFD in breast tumours was significantly higher than that in the non-adjacent control (26, 27). The expression of VEGFC was significantly higher than that of VEGFD in patients with breast cancer, as revealed by investigating TCGA database (Supplementary Figure S3, $p < 0.0001$). A study demonstrated that primary breast tumours induce sentinel lymph node lymphangiogenesis and that tumour-derived VEGFC plays an important role in their lymphangiogenesis in breast cancer, but not VEGFD (28). VEGFD seemed to exert proliferative activity in

invasive breast carcinomas. VEGFC was found to be an independent indicator of a patient's poor prognosis (29). Thus, elucidating the molecular mechanisms underlying VEGFC expression modulation in cancer cells is of great significance.

The significant upregulation and downregulation of VEGFC expression in tumours were mainly caused by transcriptional regulation (30). Transcription factors, such as Six1 (13) and FOXK1 (14), enhanced VEGFC transcription among cancer cells. However, other transcriptional factors that regulate VEGFC expression remain largely unknown. Here, we

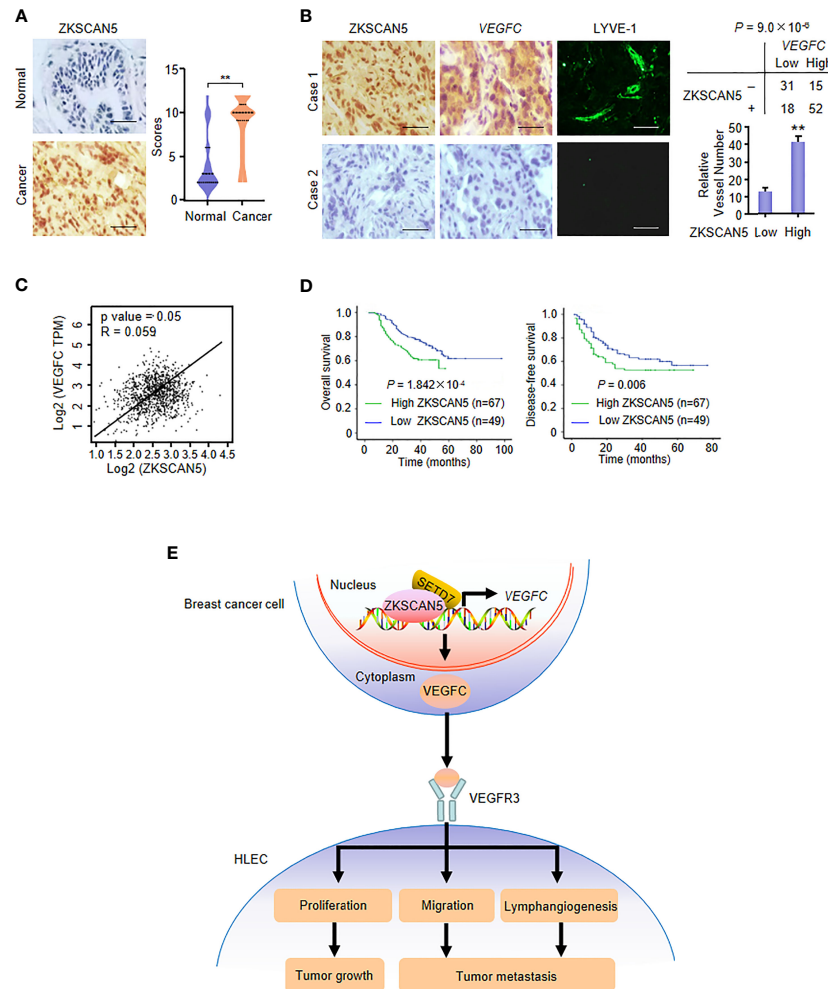


FIGURE 6 | ZKSCAN5 is a prognostic marker of breast cancer and is positively correlated with VEGFC expression. **(A)** Representative immunohistochemical staining of ZKSCAN5 in human cancerous breast tissues and adjacent normal breast tissues. Scale bar: 25 μ m. ZKSCAN5 expression scores were plotted and compared (Mann-Whitney *U* test). **(B)** Representative immunohistochemical staining of ZKSCAN5 in human breast cancer samples. Scale bar: 25 μ m. To quantify lymphatic microvessel density, images were obtained from eight regions of each tissue to calculate the number of vessels more accurately. The correlation of ZKSCAN5 with VEGFC expression or lymph microvessel number (positive LYVE-1 staining) is shown. The *p* value was generated using Pearson's χ^2 test (ZKSCAN5 and VEGFC) and the Wilcoxon rank-sum test (LYVE-1). **(C)** Database analysis showed that ZKSCAN5 expression was positively correlated with VEGFC expression in breast cancer patients. **(D)** The total survival time (above) and disease-free survival time (below) of patients with breast cancer were estimated by the Kaplan-Meier method. The review samples are represented by the markers on the chart lines. **(E)** Proposed model for ZKSCAN5 modulation of VEGFC expression as well as its tumour-promoting function. ZKSCAN5 recruited SETD7 into the VEGFC promoter, which promoted the increase of VEGFC transcription and secretion in breast cancer cells. VEGFC is secreted by cancer cells, binding to VEGFR3, promoting proliferation, migration and lymphangiogenesis of HLEC. Ultimately, they can lead to tumour growth and metastasis.

identified ZKSCAN5 as a novel transcriptional factor for VEGFC expression regulation. We chose ZR75-1 and MDA-MB-231 breast cancer cell lines to exclude the influence of the ER status. ZKSCAN5 can not only activate the activity of the VEGFC-Luc reporter but also increase the expression of VEGFC mRNA. ZKSCAN5 was localised predominantly in the nucleus, which provided the cellular basis of ZKSCAN5 to regulate VEGFC expression transcriptionally. ZKSCAN5 binds to the promoter section (−2,911 to −2,859 bp) of VEGFC in breast cancer cells. Cancer cell-secreted VEGFC regulated by ZKSCAN5 controls HLEC proliferation, migration, and tube formation (**Figure 6E**). As a new clinical prognostic marker for

breast cancer, ZKSCAN5 has a positive correlation with VEGFC expression. Thus, targeting ZKSCAN5 will be an effective way to control lymphangiogenesis in breast cancer.

Transcriptional regulation consists of changes in transcription factor binding and a complex programme of epigenetic changes regulated by histone-modifying enzymes and DNA methyltransferase (31, 32). However, the characteristics of transcription factor binding-related histone modification enzymes remain unclear. Unlike many other methyltransferases, SETD7 only monomethylates H3K4, resulting in transcription activation in HeLa cells (33). SETD7 has initially been defined as H3K4me1. It remains controversial whether SETD7-mediated H3K4me1 is critical for the

transcriptional regulation of its target genes (34). Although SETD7 is one H3K4-specific methyltransferase, SETD7-mediated p53 methylation is not a major regulatory event and does not affect p53 activity markedly *in vivo* (35). Although Guo et al. established the physical interaction between ISL1 and SETD7, as a histone H3K4-specific methyltransferase (36), SETD7 activates its expression in gastric cancer cells by binding to the ZEB1 promoter (37). Despite SETD7 being generally considered a monomethyltransferase, it has also been shown to catalyze the dimethylation of specific substrates, depending on the sequence contexts of the methylation sites (38). For example, researchers discovered that SETD7-mediated H3K4me3 enrichment on the lncRNA DRAIC promoter regulated the growth and metastasis of gliomas (39). We confirmed a functional role for ZKSCAN5 in recruiting SETD7 to the specific target VEGFC. ZKSCAN5 directly interacts with SETD7 and forms a complex with SETD7 on the VEGFC promoter. ZKSCAN5 knockdown reduces the recruitment of SETD7, H3K4me2, and H3K4me3. SETD7 knockdown or inhibition decreases H3K4me2 and H3K4me3 expression on the VEGFC promoter. Notably, SETD7 is known to be a transcriptional coactivator for ZKSCAN5 in regulating VEGFC transcription. Our study showed that SETD7 could play an important role in the transcriptional regulation of ZKSCAN5 as a cofactor of H3K4me2 or H3K4me3. The different results from previous studies may be caused by the following factors: first, previous studies have focused on different cell lines; second, many factors can affect the target genes excluding DNA methylation; third, sequence contexts of the methylation sites could lead to different outcomes. Taken together, our study indicates the critical role of ZKSCAN5 in epigenetic regulation and suggests that methylation of H3K4me2 and H3K4me3 by SETD7 is required for ZKSCAN5-induced VEGFC transcription.

ZKSCAN5 is proposed to play an important role during spermatogenesis (15). In humans, alternatively spliced ZKSCAN5 transcripts with different 5'-untranslated regions have been confirmed (40). However, the biological function of ZKSCAN5 is currently largely unknown. To the best of our knowledge, here, for the first time, we uncovered the function of ZKSCAN5 in modulating VEGFC expression, lymphangiogenesis, and breast cancer cell growth. In addition, we found that ZKSCAN5 overexpression was positively correlated with a poor prognosis in patients with breast cancer. ZKSCAN5 is the first identified sequence-specific DNA-binding transcription factor that can bind to the VEGFC promoter (**Figure 4D**). Our results pertaining to ZKSCAN5 supplement previous findings of the biological functions of ZKSCAN5. However, there have been few relevant studies on ZKSCAN5 since its discovery in 1999 (41). ZKSCAN3, a transcription factor in the same family as ZKSCAN5, plays a role in many types of tumours (42–44). ZKSCAN3 is a zinc finger transcription factor with KRAB and SCAN domains. It upregulates the expression of genes related to the cell cycle, resulting in cell proliferation, migration, angiogenesis, and proteolysis. Therefore, ZKSCAN3 promotes the tumour progression, invasion, and migration and cell growth. Silencing its expression can significantly suppress the malignancy, tumorigenicity of xenotransplants, and growth and metastasis of tumour cells. Knocking out this key molecule in tumour cells can also lead to the enhancing of the antitumor effects of drugs. The wide expression

of ZKSCAN3 in tumour cells makes it an important potential target for tumour therapy. ZKSCAN5 and ZKSCAN3 belong to the zinc finger transcription factor family. We propose that their functions would also be similar and, along with ZKSCAN3, can be targeted for the development of tumour therapy.

It has been reported that SETD7 methylates ER α , which plays an important role in breast cancer development and progression (45). Recently, SETD7 has been found to potentially methylate β -catenin, which plays a key role in cytodifferentiation, cell proliferation, and tumorigenesis (46). Inactivation or elimination of SETD7 causes G1/S cell-cycle arrest in osteosarcoma and pulmonary carcinoma cells after DNA damage (47, 48). It remains to be explored whether SETD7 has tissue-specific effects on regulating the growth of cancer cells. The present study provides some evidence that SETD7 is an oncogene in breast cancer. Inhibiting SETD7 expression may be a good strategy for breast cancer treatment. It is very interesting to study these inhibitors that may restrain tumour cell growth and lymphangiogenesis.

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 Ethics Committee of Harbin Medical University Cancer Hospital. The patients/participants provided their written informed consent to participate in this study. The animal study was reviewed and approved by the Ethics Committee of Harbin Medical University Cancer Hospital. 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

QZ, HZ, YL and TW conceived the idea of the study; JL, ZY and JM analysed the data; ZC, HL, JG interpreted the results; JL, ZY and JM wrote the paper; all authors discussed the results and revised the manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China [grant number 81730074].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.875033/full#supplementary-material>

REFERENCES

- Ginsburg O, Bray F, Coleman MP, Vanderpuye V, Eniu A, Kotha SR, et al. The Global Burden of Women's Cancers: A Grand Challenge in Global Health. *Lancet* (2017) 389(10071):847–60. doi: 10.1016/S0140-6736(16)31392-7
- Runowicz CD, Leach CR, Henry NL, Henry KS, Mackey HT, Cowens-Alvarado RL, et al. American Cancer Society/American Society of Clinical Oncology Breast Cancer Survivorship Care Guideline. *J Clin Oncol* (2016) 34(6):611–35. doi: 10.1200/JCO.2015.64.3809
- Yamauchi H, Cristofanilli M, Nakamura S, Hortobagyi GN, Ueno NT. Molecular Targets for Treatment of Inflammatory Breast Cancer. *Nat Rev Clin Oncol* (2009) 6(7):387–94. doi: 10.1038/nrclinonc.2009.73
- Morfoisse F, Renaud E, Hantelys F, Prats AC, Garmy-Susini B. Role of Hypoxia and Vascular Endothelial Growth Factors in Lymphangiogenesis. *Mol Cell Oncol* (2015) 2(4):e1024821. doi: 10.1080/23723556.2015.1024821
- Coso S, Bovay E, Petrova TV. Pressing the Right Buttons: Signaling in Lymphangiogenesis. *Blood* (2014) 123(17):2614–24. doi: 10.1182/blood-2013-12-297317
- Jussila L, Alitalo K. Vascular Growth Factors and Lymphangiogenesis. *Physiol Rev* (2002) 82(3):673–700. doi: 10.1152/physrev.00005.2002
- Ellis LM, Hicklin DJ. VEGF-Targeted Therapy: Mechanisms of Anti-Tumour Activity. *Nat Rev Cancer* (2008) 8(8):579–91. doi: 10.1038/nrc2403
- Price DJ, Miralem T, Jiang S, Steinberg R, Avraham H. Role of Vascular Endothelial Growth Factor in the Stimulation of Cellular Invasion and Signaling of Breast Cancer Cells. *Cell Growth Differ* (2001) 12(3):129–35. doi: 10.1007/BF02829520
- Ferrara N, Gerber HP, LeCouter J. The Biology of VEGF and its Receptors. *Nat Med* (2003) 9(6):669–76. doi: 10.1038/nm0603-669
- Carmeliet P. VEGF as a Key Mediator of Angiogenesis in Cancer. *Oncology* (2005) 69(Suppl 3):4–10. doi: 10.1159/000088478
- Gasparini G, Toi M, Gion M, Verderio P, Dittadi R, Hanatani M, et al. Prognostic Significance of Vascular Endothelial Growth Factor Protein in Node-Negative Breast Carcinoma. *J Natl Cancer Inst* (1997) 89(2):139–47. doi: 10.1093/jnci/89.2.139
- Stacker SA, Achen MG, Jussila L, Baldwin ME, Alitalo K. Lymphangiogenesis and Cancer Metastasis. *Nat Rev Cancer* (2002) 2(8):573–83. doi: 10.1038/nrc863
- Wang C-A, Jedlicka P, Patrick AN, Micalizzi DS, Lemmer KC, Deutsch E, et al. SIX1 Induces Lymphangiogenesis and Metastasis via Upregulation of VEGF-C in Mouse Models of Breast Cancer. *J Clin Invest* (2012) 122(5):1895–906. doi: 10.1172/JCI59858
- Zheng S, Yang L, Zou Y, Liang JY, Liu P, Gao G, et al. Long non-Coding RNA HUMT Hypomethylation Promotes Lymphangiogenesis and Metastasis via Activating FOXK1 Transcription in Triple-Negative Breast Cancer. *J Hematol Oncol* (2020) 13(1):17. doi: 10.1186/s13045-020-00852-y
- Dreyer SD, Zheng Q, Zabel B, Winterpacht A, Lee B. Isolation, Characterization, and Mapping of a Zinc Finger Gene, ZFP95, Containing Both a SCAN Box and an Alternatively Spliced KRAB A Domain. *Genomics* (1999) 62(1):119–22. doi: 10.1006/geno.1999.5981
- Chen N, Zhang G, Fu J, Wu Q. Identification of Key Modules and Hub Genes Involved in Esophageal Squamous Cell Carcinoma Tumorigenesis using WCGNA. *Cancer Control* (2020) 27(1):1073274820978817. doi: 10.1177/1073274820978817
- Okabe J, Orlowski C, Balcerzyk A, Tikellis C, Thomas MC, Cooper ME, et al. Distinguishing Hyperglycemic Changes by SET7 in Vascular Endothelial Cells. *Circ Res* (2012) 110(8):1067–76. doi: 10.1161/CIRCRESAHA.112.266171
- Kontaki H, Talianidis I. Lysine Methylation Regulates E2F1-Induced Cell Death. *Mol Cell* (2010) 39(1):152–60. doi: 10.1016/j.molcel.2010.06.006
- Shuttleworth VG, Gaughan L, Nawafa L, Mooney CA, Cobb SL, Sheerin NS, et al. The Methyltransferase SET9 Regulates TGFβ1 Activation of Renal Fibroblasts via Interaction With SMAD3. *J Cell Sci* (2018) 131(1):jcs207761. doi: 10.1242/jcs.207761
- Kassner I, Barandun M, Fey M, Rosenthal F, Hottiger MO. Crosstalk Between SET7/9-Dependent Methylation and ARTD1-Mediated ADP-Ribosylation of Histone H1.4. *Epigenet Chromatin* (2013) 6(1):1. doi: 10.1186/1756-8935-6-1
- Tuano NK, Okabe J, Ziemann M, Cooper ME, El-Osta A. SET7 Mediated Interactions Regulate Transcriptional Networks in Embryonic Stem Cells. *Nucleic Acids Res* (2016) 44(19):9206–17. doi: 10.1093/nar/gkw621
- Mishra BP, Ansari KI, Mandal SS. Dynamic Association of MLL1, H3K4 Trimethylation With Chromatin and HOX Gene Expression During the Cell Cycle. *FEBS J* (2009) 276(6):1629–40. doi: 10.1111/j.1742-4658.2009.06895.x
- Schneider J, Wood A, Lee JS, Schuster R, Dueker J, Maguire C, et al. Molecular Regulation of Histone H3 Trimethylation by Compass and the Regulation of Gene Expression. *Mol Cell* (2005) 19(6):849–56. doi: 10.1016/j.molcel.2005.07.024
- Stachura J, Wachowska M, Kilarski WW, Güç E, Golab J, Muchowicz A. The Dual Role of Tumor Lymphatic Vessels in Dissemination of Metastases and Immune Response Development. *Oncoimmunology* (2016) 5(7):e1182278. doi: 10.1080/2162402X.2016.1182278
- Betterman KL, Harvey NL. The Lymphatic Vasculature: Development and Role in Shaping Immunity. *Immunol Rev* (2016) 271(1):276–92. doi: 10.1111/imr.12413
- Gu Y, Qi X, Guo S. Lymphangiogenesis Induced by VEGF-C and VEGF-D Promotes Metastasis and a Poor Outcome in Breast Carcinoma: A Retrospective Study of 61 Cases. *Clin Exp Metastasis* (2008) 25(7):717–25. doi: 10.1007/s10585-008-9180-4
- Hunter S, Nault B, Ugwuagbo KC, Maiti S, Majumder M. MIR526B and MIR655 Promote Tumour Associated Angiogenesis and Lymphangiogenesis in Breast Cancer. *Cancers (Basel)* (2019) 11(7):938. doi: 10.3390/cancers11070938
- Zhao YC, Ni XJ, Wang MH, Zha XM, Zhao Y, Wang S. Tumor-Derived VEGF-C, But Not VEGF-D, Promotes Sentinel Lymph Node Lymphangiogenesis Prior to Metastasis in Breast Cancer Patients. *Med Oncol* (2012) 29(4):2594–600. doi: 10.1007/s12032-012-0205-0
- Mylona E, Alexandrou P, Mpakali A, Giannopoulou I, Liapis G, Markaki S, et al. Clinicopathological and Prognostic Significance of Vascular Endothelial Growth Factors (VEGF)-C and -D and VEGF Receptor 3 in Invasive Breast Carcinoma. *Eur J Surg Oncol* (2007) 33(3):294–300. doi: 10.1016/j.ejso.2006.10.015
- Pağès G, Pouyssegur J. Transcriptional Regulation of the Vascular Endothelial Growth Factor Gene—a Concert of Activating Factors. *Cardiovasc Res* (2005) 65(3):564–73. doi: 10.1016/j.cardiores.2004.09.032
- Vaissière T, Sawan C, Herceg Z. Epigenetic Interplay Between Histone Modifications and DNA Methylation in Gene Silencing. *Mutat Res* (2008) 659(1–2):40–8. doi: 10.1016/j.mrrev.2008.02.004
- Chi P, Allis CD, Wang GG. Covalent Histone Modifications—Miswritten, Misinterpreted and Mis-Erased in Human Cancers. *Nat Rev Cancer* (2010) 10(7):457–69. doi: 10.1038/nrc2876
- Lenstra DC, Damen E, Leenders RGG, Blaauw RH, Rutjes FPJT, Wegert A, et al. Structure–activity Relationship Studies on (R)-Pfi-2 Analogues as Inhibitors of Histone Lysine Methyltransferase SETD7. *ChemMedChem* (2018) 13(14):1405–13. doi: 10.1002/cmdc.201800242
- Lee J, Shao NY, Paik DT, Wu H, Guo H, Termglinchan V, et al. SETD7 Drives Cardiac Lineage Commitment Through Stage-Specific Transcriptional Activation. *Cell Stem Cell* (2018) 22(3):428–44.e5. doi: 10.1016/j.stem.2018.02.005
- Lehnertz B, Rogalski JC, Schulze FM, Yi L, Lin S, Kast J, et al. P53-Dependent Transcription and Tumor Suppression are Not Affected in SET7/9-Deficient Mice. *Mol Cell* (2011) 43(4):673–80. doi: 10.1016/j.molcel.2011.08.006
- Nishioka K, Chuikov S, Sarma K, Erdjument-Bromage H, Allis CD, Tempst P, et al. SET9, a Novel Histone H3 Methyltransferase That Facilitates Transcription by Precluding Histone Tail Modifications Required for Heterochromatin Formation. *Genes Dev* (2002) 16(4):479–89. doi: 10.1101/gad.967202
- Guo T, Wen X-Z, Li Z-Y, Han H-B, Zhang C-G, Bai Y-H, et al. ISL1 Predicts Poor Outcomes for Patients With Gastric Cancer and Drives Tumor Progression Through Binding to the ZEB1 Promoter Together With SETD7. *Cell Death Dis* (2019) 10(2):33. doi: 10.1038/s41419-018-1278-2
- Dhayalan A, Kudithipudi S, Rathert P, Jeltsch A. Specificity Analysis-Based Identification of New Methylation Targets of the SET7/9 Protein Lysine Methyltransferase. *Chem Biol* (2011) 18(1):111–20. doi: 10.1016/j.chembiol.2010.11.014
- Li C, Feng SY, Chen L. SET7/9 Promotes H3K4ME3 at lncRNA DRAIC Promoter to Modulate Growth and Metastasis of Glioma. *Eur Rev Med Pharmacol Sci* (2020) 24(23):12241–50. doi: 10.26355/eurrev_202012_24016
- Kim YH, Choe SH, Song BS, Park SJ, Kim MJ, Park YH, et al. Macaca Specific Exon Creation Event Generates a Novel ZKSCAN5 Transcript. *Gene* (2016) 577(2):236–43. doi: 10.1016/j.gene.2015.11.051

41. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the Human Tissue-Specific Expression by Genome-Wide Integration of Transcriptomics and Antibody-Based Proteomics. *Mol Cell Proteomics* (2014) 13(2):397–406. doi: 10.1074/mcp.M113.035600
42. Kawahara T, Inoue S, Ide H, Kashiwagi E, Ohtake S, Mizushima T, et al. ZKSCAN3 Promotes Bladder Cancer Cell Proliferation, Migration, and Invasion. *Oncotarget* (2016) 7(33):53599–610. doi: 10.18632/oncotarget.10679
43. Zhang X, Jing Y, Qin Y, Hunsucker S, Meng H, Sui J, et al. The Zinc Finger Transcription Factor ZKSCAN3 Promotes Prostate Cancer Cell Migration. *Int J Biochem Cell Biol* (2012) 44(7):1166–73. doi: 10.1016/j.biocel.2012.04.005
44. Chi Y, Xu H, Wang F, Chen X, Shan Z, Sun Y, et al. ZKSCAN3 Promotes Breast Cancer Cell Proliferation, Migration and Invasion. *Biochem Biophys Res Commun* (2018) 503(4):2583–9. doi: 10.1016/j.bbrc.2018.07.019
45. Subramanian K, Jia D, Kapoor-Vazirani P, Powell DR, Collins RE, Sharma D, et al. Regulation of Estrogen Receptor Alpha by the SET7 Lysine Methyltransferase. *Mol Cell* (2008) 30(3):336–47. doi: 10.1016/j.molcel.2008.03.022
46. Shen C, Wang D, Liu X, Gu B, Du Y, Wei FZ, et al. SET7/9 Regulates Cancer Cell Proliferation by Influencing B-Catenin Stability. *FASEB J* (2015) 29(10):4313–23. doi: 10.1096/fj.15-273540
47. Lezina L, Aksenova V, Ivanova T, Purmessur N, Antonov AV, Tentler D, et al. Kmtase SET7/9 Is a Critical Regulator of E2F1 Activity Upon Genotoxic Stress. *Cell Death Differ* (2014) 21(12):1889–99. doi: 10.1038/cdd.2014.108
48. Francis NJ, Rowlands M, Workman P, Jones K, Aherne W. Small-Molecule Inhibitors of the Protein Methyltransferase SET7/9 Identified in a High-Throughput Screen. *J Biomol Screen* (2012) 17(8):1102–9. doi: 10.1177/1087057112452137

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.

Copyright © 2022 Li, Yan, Ma, Chu, Li, Guo, Zhang, Zhao, Li and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



XGBG: A Novel Method for Identifying Ovarian Carcinoma Susceptible Genes Based on Deep Learning

Ke Feng Sun^{1†}, Li Min Sun^{2†}, Dong Zhou^{3†}, Ying Ying Chen⁴, Xi Wen Hao⁵, Hong Ruo Liu^{2*}, Xin Liu^{2*} and Jing Jing Chen^{6*}

OPEN ACCESS

Edited by:

Tianyi Zhao,
Harbin Institute of Technology, China

Reviewed by:

Jingyu Huang,
Wuhan University, China
Yuansong Zhao,
University of Texas Health Science
Center at Houston, United States

*Correspondence:

Hong Ruo Liu
liuhongruo@sina.com
Xin Liu
648087759@qq.com
Jing Jing Chen
chenjingjing0401@163.com

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 16 March 2022

Accepted: 08 April 2022

Published: 12 May 2022

Citation:

Sun KF, Sun LM, Zhou D,
Chen YY, Hao XW, Liu HR, Liu X
and Chen JJ (2022) XGBG:
A Novel Method for Identifying
Ovarian Carcinoma Susceptible
Genes Based on Deep Learning.
Front. Oncol. 12:897503.
doi: 10.3389/fonc.2022.897503

¹ Department of Obstetrics and Gynecology, First Affiliated Hospital, Heilongjiang University of Chinese Medicine, Harbin, China, ² Department of Oncology, The Second Affiliated Hospital of Dalian Medical University, Dalian, China, ³ Department of Oncology, Affiliated Zhongshan Hospital of Dalian University, Dalian, China, ⁴ Department of Nephrology, The First Affiliated Hospital of Heilongjiang University of Chinese Medicine, Harbin, China, ⁵ Heilongjiang University of Chinese Medicine, Harbin, China, ⁶ Department of Rheumatology and Immunology, The First Hospital Affiliated to Army Medical University, Chongqing, China

Ovarian carcinomas (OCs) represent a heterogeneous group of neoplasms consisting of several entities with pathogenesis, molecular profiles, multiple risk factors, and outcomes. OC has been regarded as the most lethal cancer among women all around the world. There are at least five main types of OCs classified by the fifth edition of the World Health Organization of tumors: high-/low-grade serous carcinoma, mucinous carcinoma, clear cell carcinoma, and endometrioid carcinoma. With the improved knowledge of genome-wide association study (GWAS) and expression quantitative trait locus (eQTL) analyses, the knowledge of genomic landscape of complex diseases has been uncovered in large measure. Moreover, pathway analyses also play an important role in exploring the underlying mechanism of complex diseases by providing curated pathway models and information about molecular dynamics and cellular processes. To investigate OCs deeper, we introduced a novel disease susceptible gene prediction method, XGBG, which could be used in identifying OC-related genes based on different omics data and deep learning methods. We first employed the graph convolutional network (GCN) to reconstruct the gene features based on both gene feature and network topological structure. Then, a boosting method is utilized to predict OC susceptible genes. As a result, our model achieved a high AUC of 0.7541 and an AUPR of 0.8051, which indicates the effectiveness of the XGBG. Based on the newly predicted OC susceptible genes, we gathered and researched related literatures to provide strong support to the results, which may help in understanding the pathogenesis and mechanisms of the disease.

Keywords: ovarian cancer, susceptible genes, XGBG, deep learning, pathway analyses

INTRODUCTION

Ovarian carcinomas (OCs) are one of the most fatal cancers in women; a scientific study of the disease is of vital priority due to its high death rate (1). A better understanding of the entities and molecules that contribute to the pathogenesis and progression of OC is essential to improve the diagnostics and treatment of the disease. Although the etiologic causes of OCs have not been recognized well, genetic factors that caused mutations in the disease have been examined profoundly with the help of many genetic approaches. However, there are still many disease susceptible genes not identified, and it is of vital importance to explore the mechanism and underlying pathogenic factors to better understand the disease and make a contribution in treating the disease.

A genome-wide association study (GWAS) is an approach utilized in genetics research to associate specific genetic variants [single-nucleotide polymorphisms (SNPs)] with a specific disease. It has identified hundreds of risk genetic variants (SNPs) that may result in ovarian cancers (2–6). However, these studies can only explain a small fraction of disease-related regions in a functional point of view (7–9). Since many risk alleles may locate in the non-protein-coding regions to regulate the expression of target genes (10), though GWAS provides strong support in revealing the associations between variants and traits, it is not comprehensive to discover the disease-related genes or gene regulators merely based on GWAS datasets.

Expression quantitative trait loci (eQTLs) are genomic loci that explain variation in expression levels of genes, which can be regarded as an additional evidence for identifying disease-related genes. eQTLs indicate the chromosomal loci that can explain variance in expression traits. These distinguishing characteristics from most expression quantitative traits are not the product of the expression of a single gene. With the help of eQTL analyses, a lot of causal genes for multiple types of cancers have been identified, such as kidney cancers, prostate cancers, breast cancers (9, 11, 12), and other complex diseases such as Alzheimer's disease and schizophrenia (13, 14). Therefore, it is more worthy to discover disease causal genes based on the integration of both GWAS and eQTL datasets.

In addition to the genetic information derived from GWAS and eQTL datasets to understand the mechanisms of complex diseases, investigation and identification of molecular pathways are also important in exploring the underlying mechanism of diseases. Pathway analysis is a typical efficient analysis to explore the biology of genes and proteins that are differentially expressed in biological processes. There are many widely accepted pathway databases such as KEGG and BioCarta that can provide illustrative information to study diseases from the view of pathway system (15, 16). According to the information of molecular dynamics and cellular processes, genes and gene products are annotated based on different functions and characteristics (17). Since complex diseases are not only caused by a single gene or a single biological process, it is important to understand the diseases and identify disease causal genes from the point of view of a pathway system.

In this article, we proposed a novel OC causal gene identification method, XGPG, integrating gene features from both genomic point and pathway annotation point. We first employed the graph convolutional network (GCN) to reconstruct the gene feature based on both gene feature and network topological structure, then utilized a boosting method, extreme gradient boosting (XGBoost), to predict OC-related susceptible genes as a binary classification problem. By applying this method, we built an efficient gene prediction model and prioritized more putative genes associated with OCs.

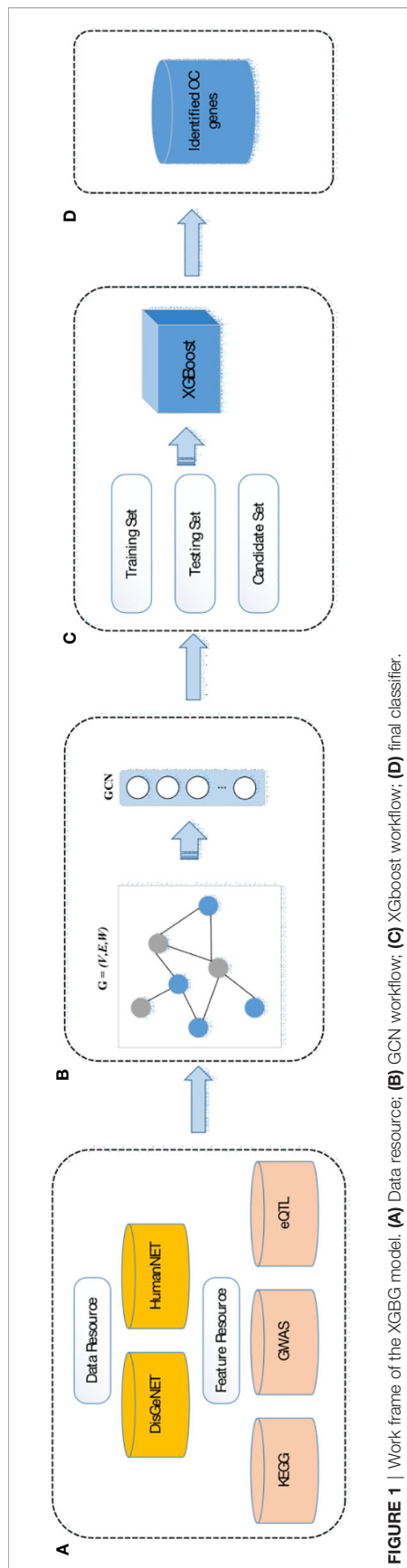
METHODS

Framework

Our method, XGPG, contains 4 main parts, data collection, feature extraction, gene feature reconstruction based on both gene feature and network topology structure, and OC causal gene prediction based on the constructed XGBoost model. In the first section (A), we manually collected different types of ovarian diseases including OC-related genes from the DisGeNET database (18) and then we obtained gene features from the GWAS Catalog, GTEx Portal, and KEGG database for different features (19, 20). Furthermore, we collected gene interaction information from the HumanNet database (21). (B) Thus, we extracted gene features from GWAS data, eQTL data, and pathway annotations, and then extracted gene network structure topological features based on the gene–gene interaction network. (C) After the feature extraction process, we utilized the GCN model to reconstruct the integrated gene features based on both gene feature and topological structure for a more precise representation of collected genes. (D) In the disease gene prediction part, a boosting model, XGBoost, is employed for constructing the prediction model and to prioritize OC-related genes. The work frame is shown in **Figure 1**.

Data Collection

We first downloaded published verified ovarian cancer-related genes from the DisGeNET database; after filtering, the dataset contains 3,181 genes to be regarded as a positive gene set. To construct a balanced training set, we randomly selected 3,171 genes that have interactions with positive genes but have no associations with ovarian diseases. These genes are used to construct the negative gene set. Then, we downloaded gene interaction information from the HumanNet database to build the gene–gene interaction network. For the prediction of OC causal genes, we also downloaded 721 ovarian disease-related genes as candidate genes to construct the prediction gene set. To extract gene features, we downloaded GWAS data from the GWAS Catalog and obtained 9,793,553 susceptible loci associated with OC, and we downloaded eQTL data from the GTEx v8 database including 25,325 susceptible loci detected in ovary tissue based on gene expression level. Moreover, we downloaded gene-pathway information from the KEGG database, including 343 annotated pathways.



Feature Extraction

We extracted gene features from three aspects, namely, GWAS data, eQTL, data and KEGG pathway information. We first obtained the detailed gene location information of the training and predictive gene data, including chromosome name, start position, and end position. Then, the genes are mapped to the SNPs provided by GWAS data. To construct the SNP feature, we sorted the gene-mapped SNPs by p -value and extracted the top 5 significant SNPs as the SNP feature of the gene. Thus, the SNP feature can be denoted as a 5-D vector:

$$F_{SNP} = [D_1, D_2, D_3, D_4, D_5] \quad (1)$$

For those genes that have less than 5 mapped SNPs, we set the value to 9×10^{-6} to avoid calculation error. For the expression feature, we mapped the genes to eQTL data based on gene location information and then extracted the top 5 significant eQTL p -values as expression feature. We also set the value to 9×10^{-6} for those genes mapped to less than 5 loci to avoid the calculation error. Thus, the expression feature can be denoted as a 5-D vector:

$$F_{exp} = [D_1, D_2, D_3, D_4, D_5] \quad (2)$$

We then downloaded the KGML files from the KEGG database, representing the details for computational analysis and pathway relations in KEGG pathways. According to the KGML files, we can obtain the genes that participate in each KEGG annotated pathway. In total, the KEGG database has annotated 343 pathways; thus, the pathway feature of each gene can be denoted as a 343-D vector; the value is set to 1 if the gene is in the pathway process or set to 0 *vice versa*:

$$F_{path} = [D_1, D_2, \dots, D_{342}, D_{343}] \quad (3)$$

$$D_i = \begin{cases} 0, & \text{if gene is in pathway } i \\ 1, & \text{if gene is not in pathway } i \end{cases} \quad (4)$$

Thus, the primary feature representation of each gene can be denoted as a 353-D vector including the SNP feature, the expression feature, and the pathway feature. Since the feature matrix could be very sparse and is not comprehensive, we further utilized the GCN model to reconstruct the feature representation with the information of the gene interaction network topological structure.

Feature Reconstruction by GCN

We first downloaded the gene–gene interaction information from the HumanNet database and constructed a gene–interaction network of the training set with dimensions of $6,352 \times 6,352$. Then, the adjacent matrix can be constructed based on the topological structure of the net. Next, the gene interaction network with gene features is input to the GCN model to reconstruct the gene features to obtain a more comprehensive feature representation. Consider the graph $G = (V, E, W)$, where V is the nodes, E is the edge, and W is the weight matrix encoding the associations between nodes. In the

GCN model, Rectified Linear Units (ReLU) is used as the activation function. We input the gene feature matrix X to the GCN model and then the gene feature can be extracted by the propagation rule of each layer:

$$H^{l+1} = \sigma(LH^l X W^l) \quad (5)$$

$$\text{ReLU}(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad (6)$$

where σ is the non-linearity activation function; here, we used ReLU.

Lastly, gene feature representation is reconstructed by GCN.

Gene Prediction Based on XGBoost Model

XGBoost is a state-of-the-art boosting method that has been widely employed in many kinds of data mining problems. It can also be used in classification and regression problems. Boosting is an ensemble learning algorithm that firstly train a weak model and then train an enhanced model to improve the errors by iteration. By iteration, the new model can fit the residuals of the previous model. Here, we utilized the “xgboost” package in R to perform the training and prediction process. In order to evaluate our prediction model, we performed a 10-fold cross-validation on the 6,352 training set. Since the training set is composed of 3,181 positive samples and 3,171 negative samples, we randomly divided them into 10 groups, and 9 of them is used to train the model and the last one is used to test the model based on the labels at each time. Grid searches were performed to evaluate the best performance of the parameters of the model.

RESULTS

Measurement of Model Performance

Since we have assessed the performance of our model based on 10 CVs with training sets, the ROC curve and PR curve are used to measure the performance of the model; the curves of 10 CVs are shown in **Figure 2**. The AUC and AUPR of 10 CVs are shown in **Table 1**. As a result, we obtained the average AUPR of 0.8051 and the average AUC of 0.7541. We chose the best performance model with an AUPR of 0.8301 and an AUC of 0.7770 to predict the OC causal genes.

Performance Comparison Between Models

Although we have proved the performance of XGPG by 10 CVs on the training set, there have been many other machine learning and deep learning methods used in classification problems, such as random forest (RF), Naïve Bayesian (NB), support vector machine (SVM), and deep neural network (DNN). To better illustrate the effectiveness and credibility of XGPG, we also compared it with SVM, RF, Naïve Bayes, and DNN. In order to ensure the consensus of the input to each model, all the gene features are reconstructed by GCN. The results are shown in **Figure 3**. As shown in the figure, SVM and RF perform better than NB and DNN, but they are far behind the XGBoost model.

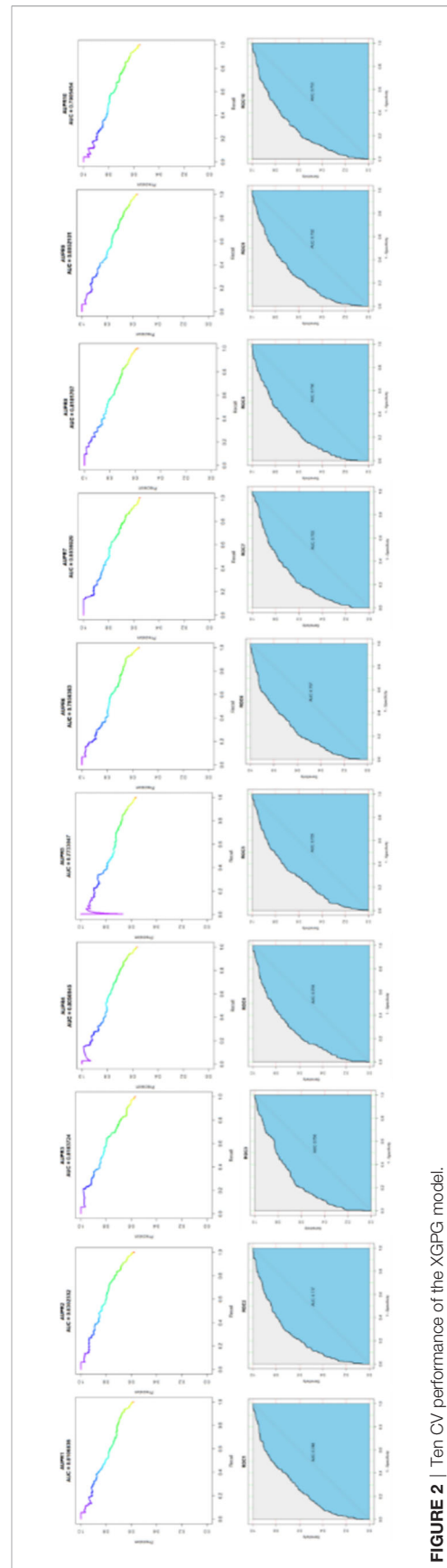
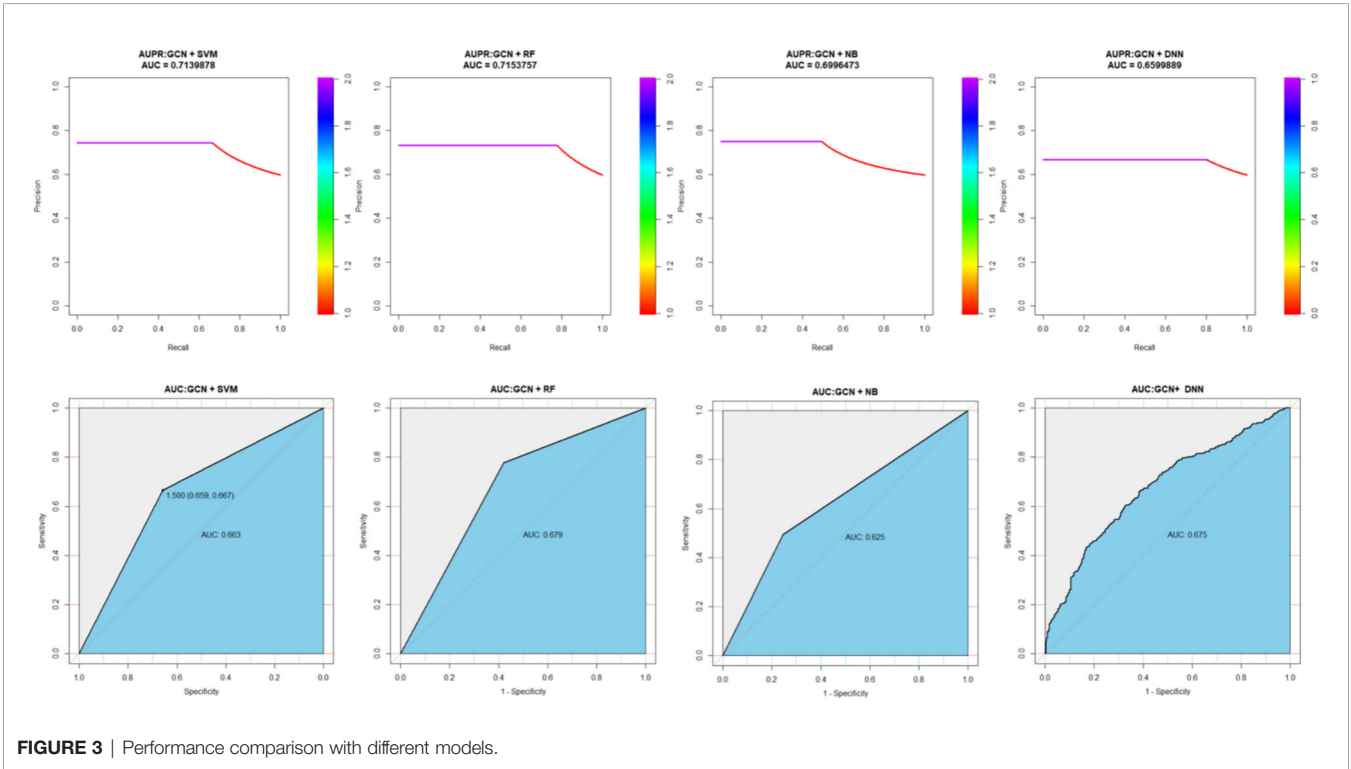


FIGURE 2 | Ten CV performance of the XGPG model.

TABLE 1 | AUPR and AUC of 10 CVs.

	1	2	3	4	5	6	7	8	9	10	Ave
AUPR	0.8102	0.8301	0.8187	0.8050	0.7731	0.7959	0.8034	0.8191	0.8050	0.7904	0.8051
AUC	0.7484	0.7770	0.7558	0.7592	0.7294	0.7568	0.7533	0.7564	0.7516	0.7532	0.7541



OC Gene Prediction Process

Since we have demonstrated the performance of our method and chose the best model to predict the OC genes, we then performed the gene prediction process with 721 verified ovary disease-related genes obtained from DisGeNET to further identify genes that are significantly associated with OCs. We also extracted the gene features as mentioned in the *Feature Extraction* section and built the gene interaction network to obtain the topological structure. After the gene prediction process by XGPG, we finally prioritize the candidate genes by the score resulting from the XGBoost model.

CASE STUDY

According to the results, our method predicted 148 (score threshold is 0.8) and 45 (score threshold is 0.9) OC causal genes from 721 candidate susceptible genes. We listed the top 20 genes in **Table 2**. As shown in **Table 2**, some of predicted genes have been reported to have direct or indirect associations with OC. Studies have indicated that KNG1 is highly related to the gonadotropin-releasing hormone (GnRH) (22), which is a hypothalamic neuropeptide that plays an important role in the reproductive system. Investigators have made a great effort to

TABLE 2 | Top 20 predicted OC causal genes.

Symbol	NCBI ID	Symbol	NCBI ID
3827	KNG1	2147	F2
2162	F13A1	5577	PRKAR2B
5921	RASA1	4086	SMAD1
657	BMPR1A	58	ACTA1
2688	GH1	2690	GHR
1489	CTF1	3489	IGFBP6
186	AGTR2	4879	NPPB
22806	IKZF3	10370	CITED2
8204	NRIP1	406954	MIR181A2
407021	MIR29A	407036	MIR32

develop GnRH agonists and antagonists for the treatment of tumors such as ovarian cancers (23). Coagulation factor II (F2) is found to be overexpressed in various epithelial neoplasms including ovarian cancer (24); F2 receptor, also known as PAR1, has been provided to be differentially expressed in ovarian cancer tissue (25). F13A, also known as coagulation factor XIII A, has been proven to have a significantly higher concentration in OC plasma, which may be a powerful tool for the clinical diagnosis and prognostic prediction of the disease (26). RASA1 is a member of the RAS-GAP family, which has been reported to play an important role in cell proliferation and

migration in several types of cancers, including OC, by inhibiting the malignant progression of OC cells in a high level (27). Furthermore, SMAD1 can regulate BMPs (such as BMP1A), resulting in aberrant BMP signaling in ovarian cancer pathology (28, 29). The IGF system has been implicated in OC since it has a key role in normal growth and development. In the Yang study, they proved that IGFBP-6 may have profound effects on the migration of two ovarian cancer cell lines, which may help in developing an IGFBP-6-based therapeutic for ovarian cancers (30). Since AGTR1 has been demonstrated to be the main effector of RAS and AGTR1 protein was detected in 86% of OC tissues, AGTR2 is the antagonist of AGTR1, which means that it also plays an important role in the pathology of OC (31). NPPB is a secreted protein that has been proven to maintain a high level in the blood of women with ovarian cancer, which indicates that NPPB may be a novel biomarker for the detection of EOC (32).

DISCUSSION

OCs are one of the most dangerous cancers for women. It is important and essential to understand the mechanisms of the disease. In this study, we proposed an OC causal gene prediction method, XGPG, based on the deep learning method and the boosting method. Since GWASs have identified lots of susceptible loci associated with OC, due to the theory of linkage disequilibrium (LD), SNPs can regulate the pathologies of traits on the expression level of target genes. Thus, we integrated both GWAS and eQTL data to integrate the gene feature from both genetic and expression levels. Moreover, since complex diseases are not only caused by a single gene or SNP, it is important to also take gene–gene interaction into consideration. We built the gene interaction network to extract

the gene network topological structure. Based on both gene feature and structure feature, we can reconstruct the gene feature representation by the GCN model and then perform the prediction process using the XGBoost model. We obtained a high AUPR 0.8051 of and an AUC of 0.7541 on the training set composed of 3,181 positive samples and 3,171 negative samples after 10-fold cross-validation. Compared with 4 other models, SVM, RF, NB and DNN, our model performed much better. Then, we performed the OC prediction process on the 721 candidate genes and derived a prioritized gene list. As a result, our method predicted 148 (score threshold is 0.8) and 45 (score threshold is 0.9) OC causal genes. From the results, prioritized genes such as F13A, RASA, SMAD, and AGTR2, and several other genes are published and proved to be associated with OC, which also proved the effectiveness of our method. In summary, our method is helpful in further understanding the etiology and pathology of OC, and may be used as a strong theoretical evidence for drug design.

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

KS, LS, and DZ designed the experiments, analyzed the data, and wrote the manuscript. YC and XH analyzed the bioinformatic data. HL and XL provided important ideas. This whole work is guided by JC. All authors contributed to the article and approved the submitted version.

REFERENCES

1. Siegel R, Ward E, Brawley O, Jemal A. Cancer Statistics, 2011: The Impact of Eliminating Socioeconomic and Racial Disparities on Premature Cancer Deaths. *CA: Cancer J Clin* (2011) 61(4):212–36. doi: 10.3322/caac.20121
2. Song H, Ramus SJ, Tyrer J, Bolton KL, Gentry-Maharaj A, Wozniak E, et al. A Genome-Wide Association Study Identifies a New Ovarian Cancer Susceptibility Locus on 9p22. 2. *Nat Genet* (2009) 41(9):996–1000. doi: 10.1038/ng.424
3. Permuth-Wey J, Lawrenson K, Shen HC, Velkova A, Tyrer JP, Chen Z, et al. Identification and Molecular Characterization of a New Ovarian Cancer Susceptibility Locus at 17q21. 31. *Nat Commun* (2013) 4(1):1–12. doi: 10.1038/ncomms2613
4. Pharoah PD, Tsai Y-Y, Ramus SJ, Phelan CM, Goode EL, Lawrenson K, et al. GWAS Meta-Analysis and Replication Identifies Three New Susceptibility Loci for Ovarian Cancer. *Nat Genet* (2013) 45(4):362–70. doi: 10.1038/ng.2564
5. Bolton KL, Tyrer J, Song H, Ramus SJ, Notaridou M, Jones C, et al. Common Variants at 19p13 are Associated With Susceptibility to Ovarian Cancer. *Nat Genet* (2010) 42(10):880–4. doi: 10.1038/ng.666
6. Shen H, Fridley BL, Song H, Lawrenson K, Cunningham JM, Ramus SJ, et al. Epigenetic Analysis Leads to Identification of HNF1B as a Subtype-Specific Susceptibility Gene for Ovarian Cancer. *Nat Commun* (2013) 4(1):1–10. doi: 10.1038/ncomms2629
7. Grisanzio C, Werner L, Takeda D, Awoyemi BC, Pomerantz MM, Yamada H, et al. Genetic and Functional Analyses Implicate the NUDT11, HNF1B, and SLC22A3 Genes in Prostate Cancer Pathogenesis. *Proc Natl Acad Sci* (2012) 109(28):11252–7. doi: 10.1073/pnas.1200853109
8. Pomerantz MM, Shrestha Y, Flavin RJ, Regan MM, Penney KL, Mucci LA, et al. Analysis of the 10q11 Cancer Risk Locus Implicates MSMB and NCOA4 in Human Prostate Tumorigenesis. *PLoS Genet* (2010) 6(11):e1001204. doi: 10.1371/journal.pgen.1001204
9. Bojesen SE, Pooley KA, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, et al. Multiple Independent Variants at the TERT Locus Are Associated With Telomere Length and Risks of Breast and Ovarian Cancer. *Nat Genet* (2013) 45(4):371–84. doi: 10.1038/ng.2566
10. Hazelett DJ, Rhie SK, Gaddis M, Yan C, Lakeland DL, Coetzee SG, et al. Comprehensive Functional Annotation of 77 Prostate Cancer Risk Loci. *PLoS Genet* (2014) 10(1):e1004102. doi: 10.1371/journal.pgen.1004102
11. Yang MQ, Li D, Yang W, Zhang Y, Liu J, Tong W. A Gene Module-Based eQTL Analysis Prioritizing Disease Genes and Pathways in Kidney Cancer. *Comput Struct Biotechnol J* (2017) 15:463–70. doi: 10.1016/j.csbj.2017.09.003
12. Loo LW, Lemire M, Le Marchand L. In Silico Pathway Analysis and Tissue Specific cis-eQTL for Colorectal Cancer GWAS Risk Variants. *BMC Genomics* (2017) 18(1):1–14. doi: 10.1186/s12864-017-3750-2
13. Patel D, Zhang X, Farrell JJ, Chung J, Stein TD, Lunetta KL, et al. Cell-Type-Specific Expression Quantitative Trait Loci Associated With Alzheimer

- Disease in Blood and Brain Tissue. *Trans Psychiatry* (2021) 11(1):1–17. doi: 10.1038/s41398-021-01373-z
14. Cai L, Huang T, Su J, Zhang X, Chen W, Zhang F, et al. Implications of Newly Identified Brain eQTL Genes and Their Interactors in Schizophrenia. *Mol Therapy-Nucleic Acids* (2018) 12:433–42. doi: 10.1016/j.omtn.2018.05.026
 15. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* (2000) 28(1):27–30. doi: 10.1093/nar/28.1.27
 16. Nishimura D. BioCarta. *Biotech Softw Internet Rep: Comput Softw J Sci* (2001) 2(3):117–20. doi: 10.1089/152791601750294344
 17. Ge H, Liu Z, Church GM, Vidal M. Correlation Between Transcriptome and Interactome Mapping Data From *Saccharomyces Cerevisiae*. *Nat Genet* (2001) 29(4):482–6. doi: 10.1038/ng776
 18. Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI. DisGeNET: A Cytoscape Plugin to Visualize, Integrate, Search and Analyze Gene–Disease Networks. *Bioinformatics* (2010) 26(22):2924–6. doi: 10.1093/bioinformatics/btq538
 19. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malagone C, et al. The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019. *Nucleic Acids Res* (2019) 47(D1):D1005–D12. doi: 10.1093/nar/gky1120
 20. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) Project. *Nat Genet* (2013) 45(6):580–5. doi: 10.1038/ng.2653
 21. Hwang S, Kim CY, Yang S, Kim E, Hart T, Marcotte EM, et al. HumanNet V2: Human Gene Networks for Disease Research. *Nucleic Acids Res* (2019) 47(D1):D573–D80. doi: 10.1093/nar/gky1126
 22. Tripathi PH, Akhtar J, Arora J, Saran RK, Mishra N, Polisetty RV, et al. Quantitative Proteomic Analysis of GnRH Agonist Treated GBM Cell Line LN229 Revealed Regulatory Proteins Inhibiting Cancer Cell Proliferation. *BMC Cancer* (2022) 22(1):1–12. doi: 10.1186/s12885-022-09218-8
 23. Ma S, Pradeep S, Villar-Prados A, Wen Y, Bayraktar E, Mangala LS, et al. GnRH-R–Targeted Lytic Peptide Sensitizes BRCA Wild-Type Ovarian Cancer to PARP Inhibition. *Mol Cancer Ther* (2019) 18(5):969–79. doi: 10.1158/1535-7163.MCT-18-0770
 24. Grisaru-Granovsky S, Salah Z, Maoz M, Pruss D, Beller U, Bar-Shavit R. Differential Expression of Protease Activated Receptor 1 (Par1) and Py397fak in Benign and Malignant Human Ovarian Tissue Samples. *Int J Cancer* (2005) 113(3):372–8. doi: 10.1002/ijc.20607
 25. Wang F-q, Fisher J, Fishman DA. MMP-1-PAR1 Axis Mediates LPA-Induced Epithelial Ovarian Cancer (EOC) Invasion. *Gynecolog Oncol* (2011) 120(2):247–55. doi: 10.1016/j.ygyno.2010.10.032
 26. Xu Y, Xu Y, Wang C, Xia B, Mu Q, Luan S, et al. Mining TCGA Database for Gene Expression in Ovarian Serous Cystadenocarcinoma Microenvironment. *PeerJ* (2021) 9:e11375. doi: 10.7717/peerj.11375
 27. Hu J, Wang L, Chen J, Gao H, Zhao W, Huang Y, et al. The Circular RNA Circ-ITCH Suppresses Ovarian Carcinoma Progression Through Targeting miR-145/RASA1 Signaling. *Biochem Biophys Res Commun* (2018) 505(1):222–8. doi: 10.1016/j.bbrc.2018.09.060
 28. Herrera B, van Dinther M, Ten Dijke P, Inman GJ. Autocrine Bone Morphogenetic Protein-9 Signals Through Activin Receptor-Like Kinase-2/Smad1/Smad4 to Promote Ovarian Cancer Cell Proliferation. *Cancer Res* (2009) 69(24):9254–62. doi: 10.1158/0008-5472.CAN-09-2912
 29. Edson MA, Nalam RL, Clementi C, Franco HL, DeMayo FJ, Lyons KM, et al. Granulosa Cell-Expressed BMPRI1A and BMPRI1B Have Unique Functions in Regulating Fertility But Act Redundantly to Suppress Ovarian Tumor Development. *Mol Endocrinol* (2010) 24(6):1251–66. doi: 10.1210/me.2009-0461
 30. Yang Z, Bach LA. Differential Effects of Insulin-Like Growth Factor Binding Protein-6 (IGFBP-6) on Migration of Two Ovarian Cancer Cell Lines. *Front Endocrinol* (2015) 5:231. doi: 10.3389/fendo.2014.00231
 31. Park Y-A, Choi CH, Do I-G, Song SY, Lee JK, Cho YJ, et al. Dual Targeting of Angiotensin Receptors (AGTR1 and AGTR2) in Epithelial Ovarian Carcinoma. *Gynecolog Oncol* (2014) 135(1):108–17. doi: 10.1016/j.ygyno.2014.06.031
 32. Lawrenson K, Grun B, Lee N, Mhawech-Fauceglia P, Kan J, Swenson S, et al. NPPB Is a Novel Candidate Biomarker Expressed by Cancer-Associated Fibroblasts in Epithelial Ovarian Cancer. *Int J Cancer* (2015) 136(6):1390–401. doi: 10.1002/ijc.29092

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.

Copyright © 2022 Sun, Sun, Zhou, Chen, Hao, Liu, Liu and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Systematic Framework for Identifying Prognostic Genes in the Tumor Microenvironment of Colon Cancer

Jinyang Liu^{1,2}, Yu Lan^{1,2}, Geng Tian^{1,2} and Jialiang Yang^{1,2,3*}

¹ Department of Sciences, Geneis Beijing Co., Ltd., Beijing, China, ² Department of Data Mining, Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, ³ PhD Workstation, Chifeng Municipal Hospital, Chifeng, China

OPEN ACCESS

Edited by:

Liang Cheng,
Harbin Medical University, China

Reviewed by:

Shuo Yang,
City University of Hong Kong, Hong
Kong SAR, China
Cangzhi Jia,
Dalian Maritime University, China

*Correspondence:

Jialiang Yang
yangjl@geneis.cn

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 18 March 2022

Accepted: 19 April 2022

Published: 19 May 2022

Citation:

Liu J, Lan Y, Tian G and Yang J (2022)
A Systematic Framework for Identifying
Prognostic Genes in the Tumor
Microenvironment of Colon Cancer.
Front. Oncol. 12:899156.
doi: 10.3389/fonc.2022.899156

As one of the most common cancers of the digestive system, colon cancer is a predominant cause of cancer-related deaths worldwide. To investigate prognostic genes in the tumor microenvironment of colon cancer, we collected 461 colon adenocarcinoma (COAD) and 172 rectal adenocarcinoma (READ) samples from The Cancer Genome Atlas (TCGA) database, and calculated the stromal and immune scores of each sample. We demonstrated that stromal and immune scores were significantly associated with colon cancer stages. By analyzing differentially expressed genes (DEGs) between two stromal and immune score groups, we identified 952 common DEGs. The significantly enriched Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms for these DEGs were associated with T-cell activation, immune receptor activity, and cytokine–cytokine receptor interaction. Through univariate Cox regression analysis, we identified 22 prognostic genes. Furthermore, nine key prognostic genes, namely, *HOXC8*, *SRPX*, *CCL22*, *CD72*, *IGLON5*, *SERPING1*, *PCOLCE2*, *FABP4*, and *ARL4C*, were identified using the LASSO Cox regression analysis. The risk score of each sample was calculated using the gene expression of the nine genes. Patients with high-risk scores had a poorer prognosis than those with low-risk scores. The prognostic model established with the nine-gene signature was able to effectively predict the outcome of colon cancer patients. Our findings may help in the clinical decisions and improve the prognosis for colon cancer.

Keywords: colon cancer, tumor microenvironment, DEGs, prognostic genes, risk score, prognostic model

INTRODUCTION

Colon cancer is a common malignant tumor, ranking second among cancers in causing cancer-related deaths in the United States. Statistics from 2016 and 2017 estimated that approximately 147,950 individuals would be diagnosed with colon cancer in 2020, with 53,200 of these individuals dying from the disease (1–3). In China, colon cancer has the fifth highest incidence and mortality

among all cancers (4). The cure and survival rates for colon cancer have increased because of early cancer screening and improvements in treatment (5, 6).

The tumor microenvironment (TME) is composed of tumor cells and surrounding immune cells, stromal cells, and extracellular matrices (ECMs) (7–11). Tumor cells can interact closely with their niche, with mesenchymal stromal cells playing a role in tumor cells escaping surveillance of the immune system (12, 13). Stromal cells promote tumor growth by overexpressing growth signals in cancer cells (14). There is growing evidence that the TME results in tumor progression by participating in multiple biological processes, including immune cell activation and recruitment, angiogenesis, and ECM remodeling (8, 15). Therapeutic strategies targeting the TME have emerged as a promising approach for cancer treatment in recent years (16, 17). Many studies have indicated that TME can affect a patient's clinical outcome and response to therapy (18, 19). Tumor-infiltrating immune cells have been proven to significantly influence tumor progression and the efficacy of anti-tumor therapy (20).

The function of multiple cell types in the TME of colon cancer has been well elucidated. In addition to acting as a physical scaffolding for tumor cells, ECM also contributes to colon cancer cells adhesion, immune evasion, and metastasis (21). Tumor-associated neutrophils enhance invasiveness by influencing angiogenesis and response to vascular endothelial growth factor (VEGF) inhibition in colon cancer (22). Higher numbers of CD4⁺ T cells can improve survival and patient benefits (23), whereas infiltrated inefficient T cells can drive tumor immune resistance (24). Malignant cells may avoid immune surveillance by suppressing dendritic cells, and colon cancer stem cells can evolve into malignant cells by accumulating genetic and epigenetic alterations and interacting with the TME as well (25). In summary, the TME of colon cancer promotes a pro-inflammatory milieu, and therefore, anti-inflammatory agents can be used to treat colon cancer (26).

In this study, we explored the relationship between stromal and immune scores of colon cancer and clinical variables. We then identified nine key prognostic genes in the TME of colon cancer. We established a novel prognostic model of the nine-gene signature that effectively predicted the outcome of colon cancer patients.

MATERIALS AND METHODS

Colon Cancer Data Collection From the TCGA Database and GEO Database

Gene expression data and corresponding clinical information of COAD and READ patients used in our study were downloaded from The Cancer Genome Atlas (TCGA) database (<https://portal.gdc.cancer.gov/>). Detailed clinical characterization of the patients was summarized in **Table 1**. The Gene Expression Omnibus (GEO) database [GSE39582 ($n = 585$)] was used to validate the relationship between the expression of nine key prognostic genes and the survival of colon cancer patients.

TABLE 1 | Clinical characterizations of patients.

Clinicopathologic variables	Category	Count (%) ($n = 633$)
Sex	Female	294 (46.4)
	Male	335 (52.9)
	Unknown	4 (0.6)
Age	≤65	253 (39.9)
	>65	376 (59.3)
	Unknown	4 (0.6)
Stage	Stage I	109 (17.2)
	Stage II	229 (36.1)
	Stage III	181 (28.5)
	Stage IV	90 (14.2)
	Unknown	24 (3.7)
Stage T	T1	20 (3.1)
	T2	109 (17.2)
	T3	428 (67.6)
	T4	70 (11.0)
	Unknown	6 (0.9)
Stage N	N0	357 (56.3)
	N1	151 (23.8)
	N2	118 (18.6)
	Unknown	7 (1.1)
Stage M	M0	467 (73.7)
	M1	89 (14.0)
	Unknown	77 (12.1)
Survival status	Alive	499 (78.8)
	Death	130 (20.5)
	Unknown	4(0.6)

Calculation of the Stromal and Immune Scores and Identification of DEGs

We calculated the immune and stromal scores in each tumor sample using the “estimate” R package, and the gene expression matrix of colon cancer patients from the TCGA database was used as input (27). Patients were subsequently separated into high-stromal and low-stromal score groups or high-immune and low-immune score groups based on the median scores, respectively. DEGs were identified using the “limma” R package, ($FDR < 0.05$ and $|\log_2(\text{fold change})| > 1$ as the cutoff values (28, 29). The “heatmap” R package was employed to display the expression level of the top 40 DEGs. The “VennDiagram” R package was used to display the overlapping genes (30).

Enrichment Analysis of Intersection DEGs

To explore the potential functions and pathways of these intersection DEGs, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed by using the “enrichplot” package and the “clusterProfiler” package (31), with the threshold set as $p\text{-value} < 0.05$.

Identification of Key Prognostic Genes Within Intersection DEGs

Univariate Cox regression analysis was used for identifying the relationship between gene expression and overall survival (OS), tumor samples of patients were divided into a high-expression group and a low-expression group according to the median gene expression level, $p\text{-value} < 0.05$ was considered as the threshold,

and 22 genes were identified as candidate prognostic genes. A least absolute shrinkage and selector operation (LASSO) algorithm was used to identify key prognostic genes with the “glmnet” R package (32). Lambda.min was the cutoff point at which the minimum mean cross-validated error occurs. Genes or indexes whose coefficient was not 0 at lambda.min were selected as key prognostic genes. The risk score of each sample was calculated using the following formula:

$$\text{risk score} = \sum_i^n \text{Expi} * \text{Coefi}$$

Coef indicated the coefficient of genes and Exp indicated the expression level of genes. All patients were grouped into the high-risk group and low-risk group based on the median risk score. The “SurvivalROC” of R package was used to display the performance of all prognostic factors to predict the survival of colon cancer patients.

Statistical Analysis

The correlation analysis was performed using Spearman’s correlation analysis. Survival curves were compared using the Kaplan–Meier method and the log-rank test. Cox regression analysis was used to calculate hazard ratios (HRs) and 95% confidence intervals (CIs). All tests were two-sided, and a $p < 0.05$ was considered to indicate significance.

RESULTS

Stromal and Immune Scores Were Markedly Related to Colon Cancer Stages

To investigate the relationship between stromal, immune scores, and clinical variables, we calculated the immune and stromal scores of each tumor sample. Patients with more malignant tumors exhibited lower immune scores than those with less malignant tumors (M1 vs. M0; N1 vs. N0; stage IV vs. stage I or stage II) (Figures 1A–C), whereas there were no differences in the distribution of immune scores among T1–4 patients (Figure 1D). We also observed no differences in the distribution of stromal scores among M0–1 patients or stage I–IV (Supplementary Figures 1A, B). The stromal scores for patients with more malignant tumors (N1 and T4) were higher compared to those with less malignant tumors (N0, T1, and T2) (Figures 1E, F). We did not observe significant associations between stromal scores or immune scores and age or sex (Supplementary Figures 1C–F).

Identification of Intersection DEGs

We identified 1,814 DEGs in high versus low immune and high versus low stromal score groups. The heatmap showed the gene expressions of the top 40 DEGs based on stromal scores and the top 40 DEGs based on immune scores, respectively (Figures 2A, B). We identified 948 common upregulated DEGs (Figure 2C) and four common downregulated DEGs (Figure 2D). GO enrichment analyses demonstrated that the main enriched

terms for these intersection DEGs were T-cell activation, positive regulation of cytokine production, and immune receptor activity (Figure 2E). The significantly enriched KEGG terms were chemokine signaling pathway and cytokine–cytokine receptor interaction (Figure 2F).

Identification of Key Prognostic Genes

Univariate Cox regression analysis was used for exploring the relationship between gene expression and OS (33). We identified 22 candidate prognostic genes, including 20 high-risk genes and two low-risk genes (Figure 3A). LASSO Cox regression analysis was used to identify key prognostic genes and build a model that can predict the prognosis of colon cancer patients (Figures 3B, C); we obtained nine key prognostic genes (Table 2). The OS between the low- and high-risk groups classified by our prognostic model was significantly different ($p = 8.202 \times 10^{-5}$, Figure 3D). Next, we constructed the prognostic risk model with the nine-gene signature to predict 3- and 5-year OS; the area under the curve (AUC) of ROC curves of 3 and 5 years were 0.666 and 0.711, respectively (Figure 3E). To explore the correlation between the nine-gene risk score and TME score, we performed the Spearman’s correlation test, and the results showed that the nine-gene risk score was significantly correlated with the stromal or immune scores (Supplementary Figures 2A, B).

Prognostic Genes Influenced the Proportion of Infiltrating Immune Cells

The relative abundances of 22 immune cells in the tumor tissue of colon cancer patients are shown in Figure 4A, with M0 macrophages (21.61%), CD4+ resting memory T cells (16.29%), and M2 macrophages (11.97%) being the primary contributors to immune cell infiltration. CD8+ T cells exhibited a positive correlation with CD4+ memory T cells and a negative correlation with M0 macrophages (Figure 4B). The infiltration proportion of naïve B cells, M1 macrophages, and M2 macrophages was higher in high-risk score groups versus low-risk score groups, whereas the low-risk group had higher regulatory T cells (Tregs) ($p = 0.009$) (Figure 4C).

Validation in the GEO Database

An external colon cancer dataset from the GEO database (GSE39582) was used to validate the correlation between the expression of the nine key prognostic genes and OS. Survival analysis was performed, and only two genes were matched in the dataset, including *CCL22* and *ARL4C* (Figures 5A–C).

DISCUSSION

TME was related to the development and progression of tumors and had the potential to influence responses to therapies. We obtained the immune and stromal scores that could reflect the degree of immune infiltration of corresponding cells in tumor tissue. We confirmed that stromal and immune scores were significantly related to colon cancer stages. Patients with more

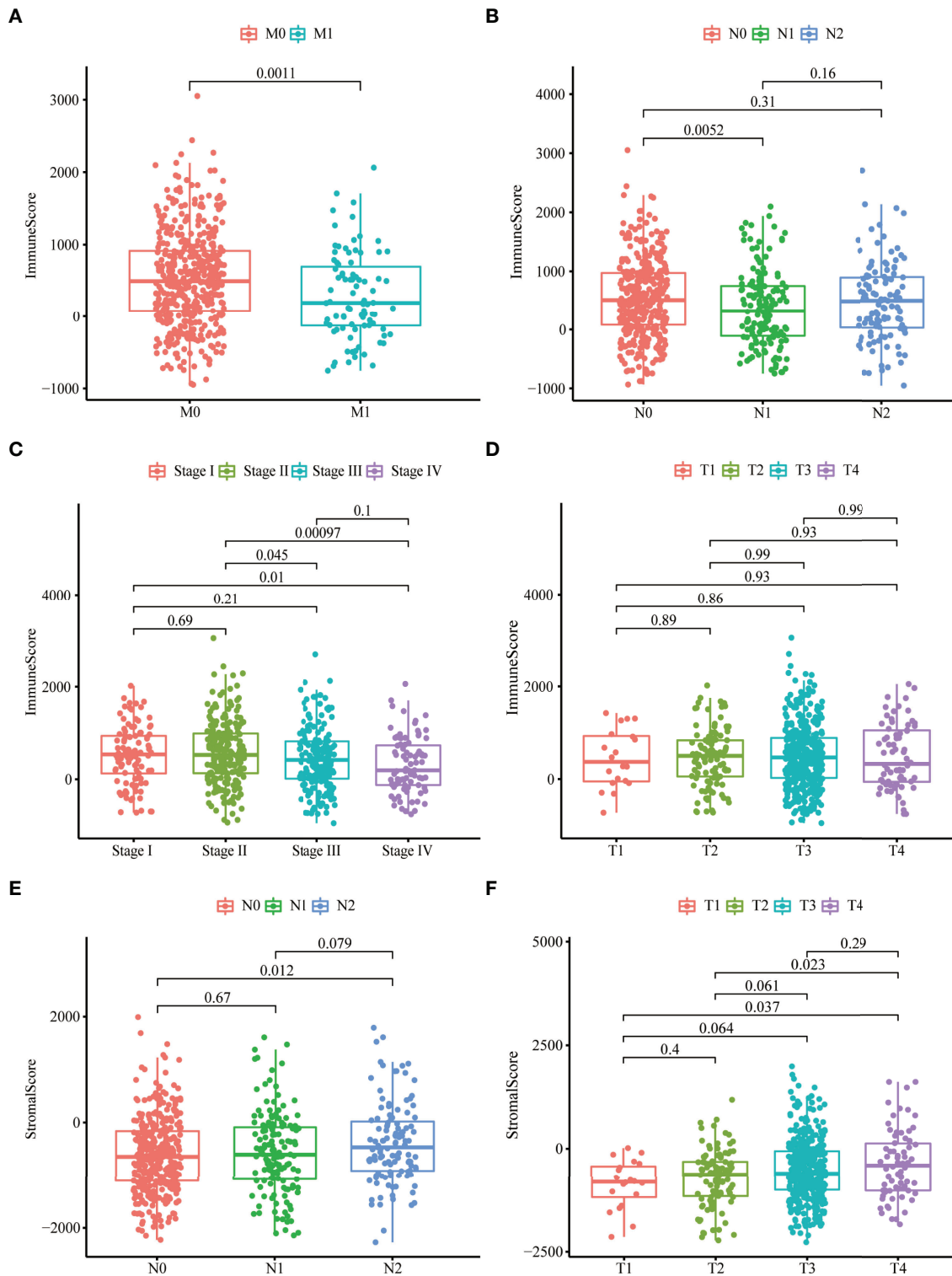


FIGURE 1 | Stromal and immune scores were markedly related to colon cancer stages. **(A–D)** Distribution of immune scores in nonmetastatic (M0) patients and distant metastases (M1) patients **(A)**, N0–2 patients **(B)**, stage I–IV patients **(C)**, T1–4 patients **(D)**. **(E, F)** Distribution of stromal scores in N0–2 patients **(E)** and T1–4 patients **(F)**.

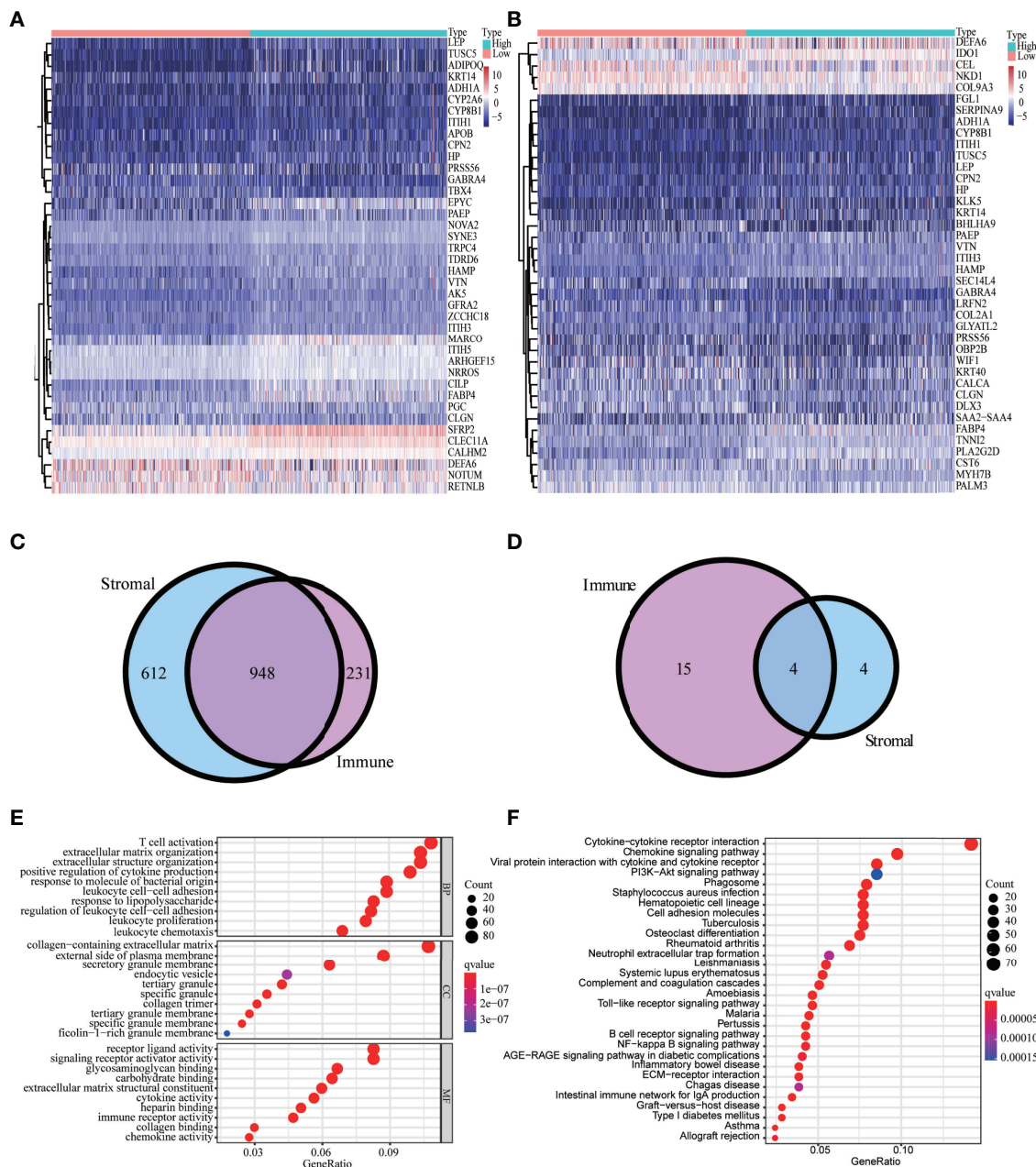


FIGURE 2 | Analysis of DEGs-based stromal and immune scores. The heatmap of the top 40 DEGs based on stromal scores (A) and immune scores (B). Venn diagrams displaying the number of upregulated DEGs (C) and downregulated DEGs (D) detected in both groups. Top 30 enriched ($p < 0.05$) GO terms (E) and KEGG terms (F).

malignant tumors (M1, N1, and stage IV) have lower immune scores than those with less malignant tumors (M0, N0, stage I, and stage II); in contrast, the stromal scores for late-stage (N2 and T4) patients were higher compared to early-stage (N0, T1, and T2) patients. In the early stage of tumorigenesis, the TME of colon cancer was remodeled, the number of infiltrating stromal cells was raised, and the number of immune cells

was decreased. Stromal cells helped tumor cells escape from being attacked by the immune system, and the lethality of some immune cells to tumors began to weaken. Disrupting the stability of TME thus induced tumor development.

We obtained DEGs between high versus low stromal and immune score groups and further identified 948 co-upregulated DEGs and four co-downregulated DEGs. The main enriched GO

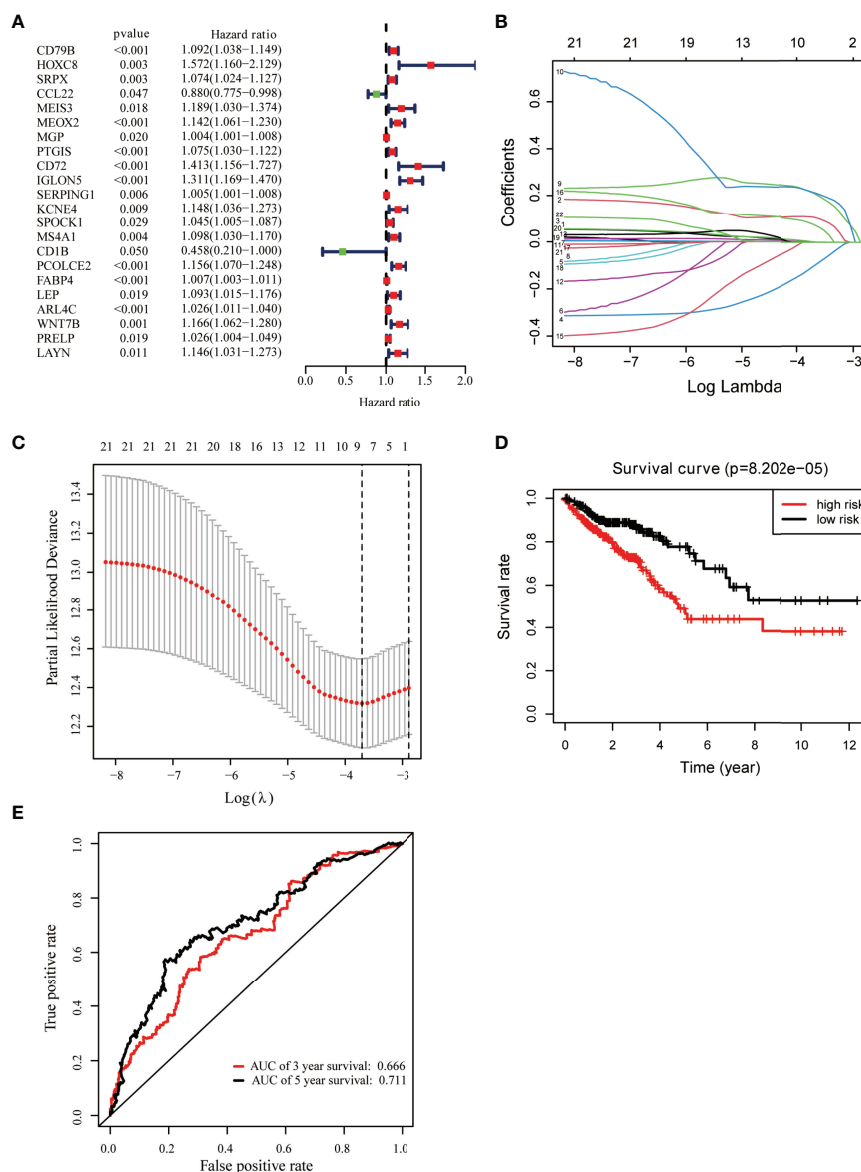


FIGURE 3 | Identification of key prognostic genes within intersection DEGs. **(A)** Forest plot of risk genes: Red represented high-risk genes (hazard ratios, HR > 1); green represented low-risk genes (HR < 1). **(B)** Constructing the LASSO coefficient prediction model. **(C)** Selecting variables in LASSO regression with minimum criteria by 1,000 times cross-validation. **(D)** Overall survival between high- and low-risk score groups. **(E)** ROC curves for predicting 3- and 5-year overall survival probability with the nine-gene score.

terms for the intersection DEGs were T-cell activation and ECM organization. Additionally, the significantly enriched KEGG terms were chemokine signaling pathway, and cytokine-cytokine receptor interaction.

Moreover, univariate Cox regression analysis was performed to determine the association between the expression of DEGs and survival, and we screened out 22 risk genes as candidate prognostic factors; using LASSO Cox regression analysis, we identified nine key prognostic genes. These genes have previously been reported to be associated with the

development and progression of tumors. Fatty acid-binding protein 4 (*FABP4*) released from adipocytes could promote invasion in prostate cancer (34). ADP-ribosylation factor (Arf)-like protein 4c (*Arl4c*) expression was upregulated upon activation of Wnt- β -catenin and growth factor-Ras signaling and contributed to tubulogenesis and tumorigenesis (35). Serine proteinase inhibitor family G1 (*SERPING1*) downregulation was associated with poor prognosis in prostate cancer (PCa) (36). *SRPX2* was involved in tumor suppression and progression (37). Homeobox C8 (*HOXC8*) was a transcription factor that

TABLE 2 | Nine key prognostic genes.

Gene	Coef
CCL22	-0.103078
FABP4	0.000222
ARL4C	0.000224
SERPING1	0.000982
SRPX	0.002037
PCOLCE2	0.077076
HOXC8	0.108125
CD72	0.209473
IGLN5	0.211910

had been reported, and high expression of *HOXC8* was associated with poor prognosis of cervical cancer (38). Several types of immune cells, such as dendritic cells and macrophages, secreted *CCL22* upon activation (39–41). *CCL22* could recruit T regulatory cells and controlled the growth of tumor cells in melanoma. However, the relationship between *CCL22* and colon cancer was unknown. The level of chemokine *CCL22* was

increased in COAD (42). Our study demonstrated that all these genes might be crucial biomarkers in the TME of colon cancer. We also found that *ARL4C* and *HOXC8* were upregulated, and *FABP4*, *PCOLCE2*, *SERPING1*, and *SRPX* were downregulated in tumor tissue compared to corresponding healthy tissue (**Supplementary Figure 3**). More work is needed to be done to investigate the association between the expression of these genes and colon cancer proliferation, metastasis, and invasion. We calculated the risk score of each sample using the gene expression of the nine genes, and we demonstrated that patients with high-risk scores have a poorer prognosis than those with a lower-risk score. Furthermore, we established an independent prognostic model that was able to effectively predict the outcome of colon cancer patients with the nine-gene signature.

Finally, analysis of immune cells' infiltration revealed the M0 macrophages, CD4+ resting memory T cells, and M2 macrophages with the highest proportion. Many studies had shown that M2 macrophages may promote tumor progression

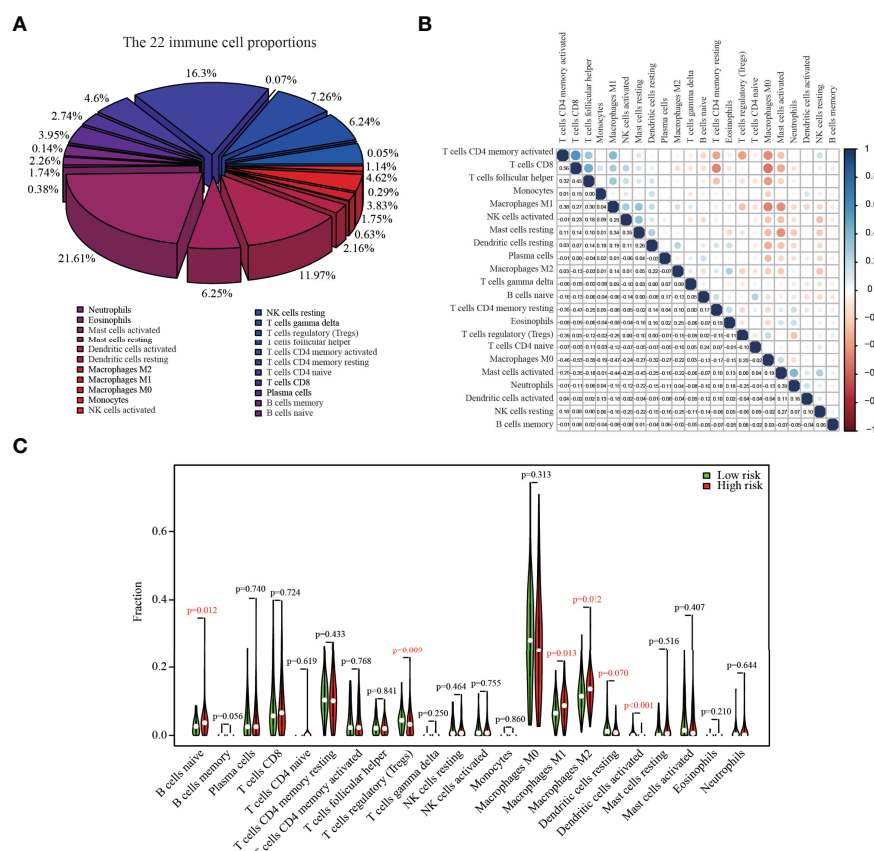


FIGURE 4 | The composition of 22 immune cells in colon cancer tumors from the TCGA dataset. **(A)** The relative abundances of 22 immune cells in the tumor tissue of colon cancer patients. **(B)** The correlation matrix between different cell types; the size of the circle represented the degree of correlation. **(C)** Fractions of infiltrating immune cells in high versus low risk score groups.

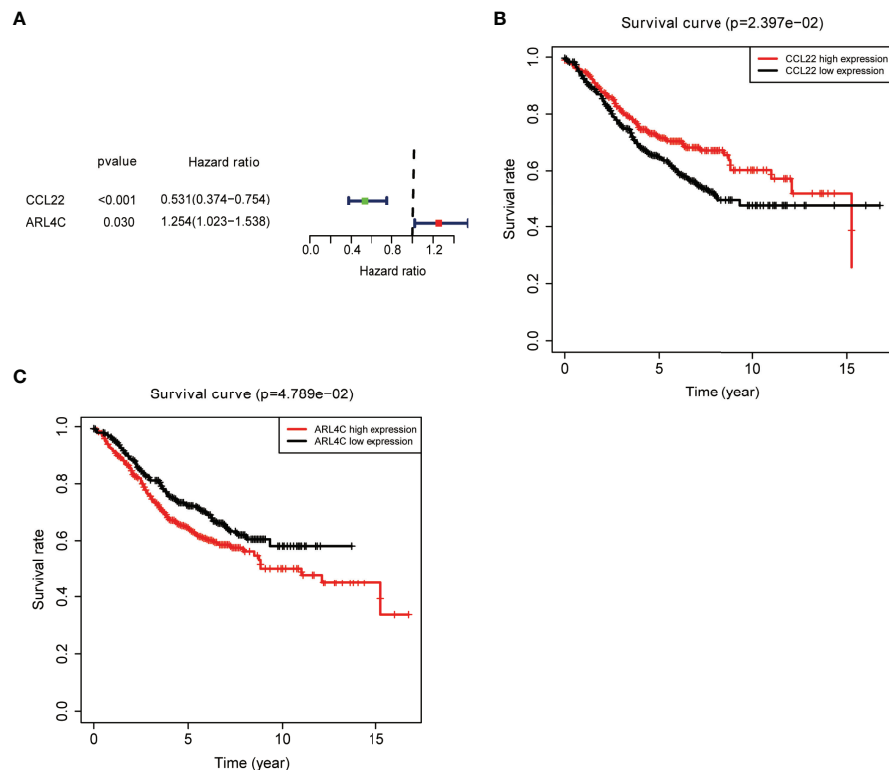


FIGURE 5 | External validation of key prognostic genes using the GEO database. **(A)** Forest plot of risk genes: Red represents high-risk genes (hazard ratios, HR > 1); green represents low-risk genes (HR < 1). **(B, C)** Overall survival between high and low *CCL22* **(B)** and *ARL4C* **(C)** expression groups.

(43), invasiveness (44), and angiogenesis (45). We also demonstrated that the infiltration proportion of M2 macrophages was higher in high-risk score groups versus low-risk score groups.

Because our study was a pure bioinformatics analysis based on the TCGA database, further biological experiments were needed to validate our results. Moreover, whether these nine key prognostic genes could improve the diagnostic accuracy and therapeutic response for colon cancer in actual clinical practice requires further verification.

CONCLUSION

In conclusion, we identified nine potential prognostic markers for colon cancer through a systematic bioinformatics analysis. A novel prognostic model established with the nine-gene signature effectively predicted the outcome of colon cancer patients. More work is needed to validate our findings.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: <https://github.com/benstevens/TME>.

AUTHOR CONTRIBUTIONS

JY contributed to the conception and design of the study. JL performed the statistical and bioinformatic analysis. JL and YL wrote the manuscript with the help of JY. GT provided suggestions for figure preparation. All authors contributed to revising the manuscript, and read and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.899156/full#supplementary-material>

Supplementary Figure 1 | Relationship between stromal and immune scores and colon cancer clinical variables. **(A, B)** Distribution of stromal scores in nonmetastatic (M0) patients and distant metastases (M1) patients **(A)**. Stage I–IV patients **(B)**. **(C–F)** Distribution of immune scores **(C)** and stromal scores **(D)** in different age groups, immune scores **(E)**, and stromal scores **(F)** in different sex groups.

Supplementary Figure 2 | The correlation between the nine gene score and tumor microenvironment scores. **(A)** Stromal score; **(B)** Immune score.

Supplementary Figure 3 | Comparison of the expressions of the nine key prognostic genes in tumor tissue and corresponding healthy tissue of colon cancer patients.

REFERENCES

- Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, et al. Colorectal Cancer Statistics, 2020. *CA Cancer J Clin* (2020) 70(3):145–64. doi: 10.3322/caac.21601
- Cheng L, Qi C, Yang H, Lu M, Cai Y, Fu T, et al. Gutmgene: A Comprehensive Database for Target Genes of Gut Microbes and Microbial Metabolites. *Nucleic Acids Res* (2021) 50(D1):D795–800. doi: 10.1093/nar/gkab786
- Zhao T, Hu Y, Zang T, Cheng L. MRTFB Regulates the Expression of NOMO1 in Colon. *Proc Natl Acad Sci USA* (2020) 117(14):7568–9. doi: 10.1073/pnas.2000499117
- Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer Statistics in China, 2015. *CA Cancer J Clin* (2016) 66(2):115–32. doi: 10.3322/caac.21338
- Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, et al. Cancer Treatment and Survivorship Statistics, 2019. *CA Cancer J Clin* (2019) 69(5):363–85. doi: 10.3322/caac.21565
- Hong J, Lin X, Hu X, Wu X, Fang W. A Five-Gene Signature for Predicting the Prognosis of Colorectal Cancer. *Curr Gene Ther* (2021) 21(4):280–9. doi: 10.2174/1566523220666201012151803
- Catalano V, Turdo A, Di Franco S, Dieli F, Todaro M, Stassi G. Tumor and its Microenvironment: A Synergistic Interplay. *Semin Cancer Biol* (2013) 23(6 Pt B):522–32. doi: 10.1016/j.semcancer.2013.08.007
- Roma-Rodriguez C, Mendes R, Baptista PV, Fernandes AR. Targeting Tumor Microenvironment for Cancer Therapy. *Int J Mol Sci* (2019) 20(4):840. doi: 10.3390/ijms20040840
- Mizuno R, Kawada K, Itatani Y, Ogawa R, Kiyasu Y, Sakai Y. The Role of Tumor-Associated Neutrophils in Colorectal Cancer. *Int J Mol Sci* (2019) 20(3):529. doi: 10.3390/ijms20030529
- Song Z, Chen X, Shi Y, Huang R, Wang W, Zhu K, et al. Evaluating the Potential of T Cell Receptor Repertoires in Predicting the Prognosis of Resectable Non-Small Cell Lung Cancers. *Mol Ther Methods Clin Dev* (2020) 18:73–83. doi: 10.1016/j.omtm.2020.05.020
- Cienfuegos-Jimenez O, Vazquez-Garza E, Rojas-Martinez A. CAR-NK Cells for Cancer Therapy: Molecular Redesign of the Innate Antineoplastic Response. *Curr Gene Ther* (2021) 22(4):303–18. doi: 10.2174/1566523222666211217091724
- Poggi A, Musso A, Dapino I, Zocchi MR. Mechanisms of Tumor Escape From Immune System: Role of Mesenchymal Stromal Cells. *Immunol Lett* (2014) 159(1–2):55–72. doi: 10.1016/j.imlet.2014.03.001
- Sindhu RK, Madaan P, Chandel P, Akter R, Adilakshmi G, Rahman MH. Therapeutic Approaches for the Management of Autoimmune Disorders via Gene Therapy: Prospects, Challenges, and Opportunities. *Curr Gene Ther* (2021) 22(3):245–61. doi: 10.2174/1566523221666210916113609
- Oya Y, Hayakawa Y, Koike K. Tumor Microenvironment in Gastric Cancers. *Cancer Sci* (2020) 111(8):2696–707. doi: 10.1111/cas.14521
- Huot JR, Novinger LJ, Pin F, Bonetto A. HCT116 Colorectal Liver Metastases Exacerbate Muscle Wasting in a Mouse Model for the Study of Colorectal Cancer Cachexia. *Dis Model Mech* (2020) 13(1):dmm043166. doi: 10.1242/dmm.043166
- Joyce JA. Therapeutic Targeting of the Tumor Microenvironment. *Cancer Cell* (2005) 7(6):513–20. doi: 10.1016/j.ccr.2005.05.024
- Yang H, Qi C, Li B, Cheng L. Non-Coding RNAs as Novel Biomarkers in Cancer Drug Resistance. *Curr Med Chem* (2021) 29(5):837–48. doi: 10.2174/0929867328666210804090644
- Galon J, Pagès F, Marincola FM, Thurin M, Trinchieri G, Fox BA, et al. The Immune Score as a New Possible Approach for the Classification of Cancer. *J Transl Med* (2012) 10:1. doi: 10.1186/1479-5876-10-1
- Pitt JM, Marabelle A, Eggermont A, Soria JC, Kroemer G, Zitvogel L. Targeting the Tumor Microenvironment: Removing Obstruction to Anticancer Immune Responses and Immunotherapy. *Ann Oncol* (2016) 27(8):1482–92. doi: 10.1093/annonc/mdw168
- Chen DS, Mellman I. Elements of Cancer Immunity and the Cancer-Immune Set Point. *Nature* (2017) 541(7637):321–30. doi: 10.1038/nature21349
- Tsagkaris C, Papakosta V, Miranda AV, Zacharopoulou L, Danilchenko V, Matiasheva L, et al. Gene Therapy for Angelman Syndrome: Contemporary Approaches and Future Endeavors. *Curr Gene Ther* (2020) 19(6):359–66. doi: 10.2174/1566523220666200107151025
- Itatani Y, Yamamoto T, Zhong C, Molinolo AA, Ruppel J, Hegde P, et al. Suppressing Neutrophil-Dependent Angiogenesis Abrogates Resistance to Anti-VEGF Antibody in a Genetic Model of Colorectal Cancer. *Proc Natl Acad Sci USA* (2020) 117(35):21598–608. doi: 10.1073/pnas.2008112117
- Kuwahara T, Hazama S, Suzuki N, Yoshida S, Tomochika S, Nakagami Y, et al. Intratumoural-Infiltrating CD4+ and Foxp3+ T Cells as Strong Positive Predictive Markers for the Prognosis of Resectable Colorectal Cancer. *Br J Cancer* (2019) 121(8):659–65. doi: 10.1038/s41416-019-0559-6
- D'Alterio C, Buoncervello M, Ieranò C, Napolitano M, Portella L, Rea G, et al. Targeting CXCR4 Potentiates Anti-PD-1 Efficacy Modifying the Tumor Microenvironment and Inhibiting Neoplastic PD-1. *J Exp Clin Cancer Res* (2019) 38(1):432. doi: 10.1186/s13046-019-1420-8
- Jahanafrooz Z, Mosafar J, Akbari M, Hashemzadeh M, Mokhtarzadeh A, Baradaran B. Colon Cancer Therapy by Focusing on Colon Cancer Stem Cells and Their Tumor Microenvironment. *J Cell Physiol* (2020) 235(5):4153–66. doi: 10.1002/jcp.29337
- Terzić J, Grivennikov S, Karin E, Karin M. Inflammation and Colon Cancer. *Gastroenterology* (2010) 138(6):2101–14.e2105. doi: 10.1053/j.gastro.2010.01.058
- Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-García W, et al. Inferring Tumour Purity and Stromal and Immune Cell Admixture From Expression Data. *Nat Commun* (2013) 4:2612. doi: 10.1038/ncomms3612
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res* (2015) 43(7):e47. doi: 10.1093/nar/gkv007
- Yang J, Huang T, Petralia F, Long Q, Zhang B, Argmann C, et al. Synchronized Age-Related Gene Expression Changes Across Multiple Tissues in Human and the Link to Complex Diseases. *Sci Rep* (2015) 5:15145. doi: 10.1038/srep15145
- Chen H, Boutros PC. VennDiagram: A Package for the Generation of Highly-Customizable Venn and Euler Diagrams in R. *BMC Bioinf* (2011) 12:35. doi: 10.1186/1471-2105-12-35
- Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *Omics* (2012) 16(5):284–7. doi: 10.1089/omi.2011.0118
- Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* (2010) 33(1):1–22. doi: 10.18637/jss.v033.i01
- Yang J, Ju J, Guo L, Ji B, Shi S, Yang Z, et al. Prediction of HER2-Positive Breast Cancer Recurrence and Metastasis Risk From Histopathological Images and Clinical Information via Multimodal Deep Learning. *Comput Struct Biotechnol J* (2022) 20:333–42. doi: 10.1016/j.csbj.2021.12.028
- Uehara H, Kobayashi T, Matsumoto M, Watanabe S, Yoneda A, Bando Y. Adipose Tissue: Critical Contributor to the Development of Prostate Cancer. *J Med Invest* (2018) 65(1.2):9–17. doi: 10.2152/jmi.65.9
- Matsumoto S, Fujii S, Kikuchi A. Arl4c is a Key Regulator of Tubulogenesis and Tumorigenesis as a Target Gene of Wnt-β-Catenin and Growth Factor-Ras Signalling. *J Biochem* (2017) 161(1):27–35. doi: 10.1093/jb/mvw069
- Peng S, Du T, Wu W, Chen X, Lai Y, Zhu D, et al. Decreased Expression of Serine Protease Inhibitor Family G1 (SERPING1) in Prostate Cancer can Help Distinguish High-Risk Prostate Cancer and Predicts Malignant Progression. *Urol Oncol* (2018) 36(8):366.e361–6.e369. doi: 10.1016/j.urolonc.2018.05.021
- Pawlowski K, Muszewska A, Lenart A, Szczepińska T, Godzik A, Grynberg M. A Widespread Peroxiredoxin-Like Domain Present in Tumor Suppression- and Progression-Implicated Proteins. *BMC Genomics* (2010) 11:590. doi: 10.1186/1471-2164-11-590
- Huang Y, Chen L, Guo A. Upregulated Expression of HOXC8 Is Associated With Poor Prognosis of Cervical Cancer. *Oncol Lett* (2018) 15(5):7291–6. doi: 10.3892/ol.2018.8200
- Vulcano M, Albanesi C, Stoppacciaro A, Bagnati R, D'Amico G, Struyf S, et al. Dendritic Cells as a Major Source of Macrophage-Derived Chemokine/CCL22 *In Vitro* and *In Vivo*. *Eur J Immunol* (2001) 31(3):812–22. doi: 10.1002/1521-4141(200103)31:3<812::AID-IMMU812>3.0.CO;2-L
- Mantovani A, Gray PA, Van Damme J, Sozzani S. Macrophage-Derived Chemokine (MDC). *J Leukoc Biol* (2000) 68(3):400–4. doi: 10.1189/jlb.68.3.400

41. Klarquist J, Tobin K, Farhangi Oskuei P, Henning SW, Fernandez MF, Dellacecca ER, et al. Ccl22 Diverts T Regulatory Cells and Controls the Growth of Melanoma. *Cancer Res* (2016) 76(21):6230–40. doi: 10.1158/0008-5472.CAN-16-0618
42. Langenes V, Svensson H, Börjesson L, Gustavsson B, Bemark M, Sjöling Å, et al. Expression of the Chemokine Decoy Receptor D6 Is Decreased in Colon Adenocarcinomas. *Cancer Immunol Immunother* (2013) 62(11):1687–95. doi: 10.1007/s00262-013-1472-0
43. Lee YS, Song SJ, Hong HK, Oh BY, Lee WY, Cho YB. The FBW7-MCL-1 Axis is Key in M1 and M2 Macrophage-Related Colon Cancer Cell Progression: Validating the Immunotherapeutic Value of Targeting PI3Kγ. *Exp Mol Med* (2020) 52(5):815–31. doi: 10.1038/s12276-020-0436-7
44. Illemann M, Bird N, Majeed A, Sehested M, Laerum OD, Lund LR, et al. MMP-9 Is Differentially Expressed in Primary Human Colorectal Adenocarcinomas and Their Metastases. *Mol Cancer Res* (2006) 4(5):293–302. doi: 10.1158/1541-7786.MCR-06-0003
45. Barbera-Guillem E, Nyhus JK, Wolford CC, Friece CR, Sampsel JW. Vascular Endothelial Growth Factor Secretion by Tumor-Infiltrating Macrophages

Essentially Supports Tumor Angiogenesis, and IgG Immune Complexes Potentiate the Process. *Cancer Res* (2002) 62(23):7042–9.

Conflict of Interest: All authors were employed by Geneis Beijing Co., Ltd.

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.

Copyright © 2022 Liu, Lan, Tian and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Inferring Gene Regulatory Networks From Single-Cell Transcriptomic Data Using Bidirectional RNN

Yanglan Gan¹, Xin Hu¹, Guobing Zou², Cairong Yan¹ and Guangwei Xu^{1*}

¹ School of Computer Science and Technology, Donghua University, Shanghai, China, ² School of Computer Engineering and Science, Shanghai University, Shanghai, China

OPEN ACCESS

Edited by:

Liang Cheng,
Harbin Medical University, China

Reviewed by:

Jianxing Zheng,
Shanxi University, China
Yuzhong Peng,
Nanning Normal University, China

*Correspondence:

Guangwei Xu
gwxu@dhu.edu.cn

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 19 March 2022

Accepted: 22 April 2022

Published: 26 May 2022

Citation:

Gan Y, Hu X, Zou G, Yan C
and Xu G (2022) Inferring Gene
Regulatory Networks From
Single-Cell Transcriptomic
Data Using Bidirectional RNN.
Front. Oncol. 12:899825.
doi: 10.3389/fonc.2022.899825

Accurate inference of gene regulatory rules is critical to understanding cellular processes. Existing computational methods usually decompose the inference of gene regulatory networks (GRNs) into multiple subproblems, rather than detecting potential causal relationships simultaneously, which limits the application to data with a small number of genes. Here, we propose BiRGRN, a novel computational algorithm for inferring GRNs from time-series single-cell RNA-seq (scRNA-seq) data. BiRGRN utilizes a bidirectional recurrent neural network to infer GRNs. The recurrent neural network is a complex deep neural network that can capture complex, non-linear, and dynamic relationships among variables. It maps neurons to genes, and maps the connections between neural network layers to the regulatory relationship between genes, providing an intuitive solution to model GRNs with biological closeness and mathematical flexibility. Based on the deep network, we transform the inference of GRNs into a regression problem, using the gene expression data at previous time points to predict the gene expression data at the later time point. Furthermore, we adopt two strategies to improve the accuracy and stability of the algorithm. Specifically, we utilize a bidirectional structure to integrate the forward and reverse inference results and exploit an incomplete set of prior knowledge to filter out some candidate inferences of low confidence. BiRGRN is applied to four simulated datasets and three real scRNA-seq datasets to verify the proposed method. We perform comprehensive comparisons between our proposed method with other state-of-the-art techniques. These experimental results indicate that BiRGRN is capable of inferring GRN simultaneously from time-series scRNA-seq data. Our method BiRGRN is implemented in Python using the TensorFlow machine-learning library, and it is freely available at <https://gitee.com/DHUEDLab/bi-rgrn>.

Keywords: gene regulatory network, recurrent neural network, gene expression, single-cell transcriptomic data, bidirectional structure

1 INTRODUCTION

Gene regulatory mechanisms are crucial to understanding diverse dynamic processes such as development, stress response and disease (1). Cell states and the dynamics of cell behavior are governed by complex gene interactions (2), which in turn define cellular morphology and functions. Such regulatory interactions can be modeled as a gene regulatory network (GRN), where nodes are

regulators and their target genes, and edges represent the regulatory relationships between genes (3). Unraveling GRNs is one of the major challenges in the field of computational biology, which allows us to pinpoint key factors that determine phenotype in health systems as well as in diseases (4, 5).

A plethora of computational or statistical approaches have been developed for inferring networks from observational gene expression data (6–8). The widely used algorithm GENIE3 decomposes the inference of gene regulatory networks into different regression subproblems. Using tree-based ensemble methods, the expression pattern of each target gene is predicted by the expression of all the other genes (9). ENNET also considers the inference problem as a regression task, which is solved by a decision tree optimizing the least-squares loss function (10). It builds the model additively using a boosting procedure. PPCOR reconstructs gene regulatory network by calculating partial correlation coefficient and semi-partial correlation coefficient between genes (11). PIDC exploits information theory to infer the regulatory relationship between genes (12). Biologically, it is assumed that changes in regulators should precede changes in their targets in time. However, such time information is not available in steady-state gene expression data, and thus GRNs constructed from these data have limited ability to capture dynamic regulatory relationships between genes. Several methods have been proposed to infer GRNs based on time-series gene expression data to address this issue. The algorithm LEAP reconstructs gene regulatory networks by calculating the Pearson correlation coefficient. With pseudo-time data information, the algorithm defines a fixed-size time window and assumes that the earlier expressed gene in this window can affect other genes (13). SCODE infers regulatory networks based on ordinary differential equations and linear regression (14). The method SINCERITIES adopts the Kolmogorov–Smirnov distance to quantify the distance between two cumulative distribution functions of gene expressions from subsequent time points, and recovers directed regulatory relationships among genes by employing regularized linear regression (15). BiXGBoost infers the regulatory network through both forward and reverse directions, separately considering the regulatory genes and target genes of specific genes, and uses the gradient boosting decision tree to integrate the final regulatory relationship (16). The algorithm GRGNN proposes an end-to-end gene regulation graph neural network approach to reconstruct GRNs from scratch utilizing gene expression data in both a supervised and a semi-supervised framework (17). DeepSEM is a neural network version of the structural equation model (SEM) to explicitly model the regulatory relationships among genes (18). These efforts mainly focus on intracellular interactions, inferring gene regulatory relationships within a specific cell. Recently developed methods for spatial transcriptomics are now providing high-throughput information about both the expression patterns of genes within a single cell and the spatial relationships between cells (19–21). The algorithm CNNC is a supervised framework for gene relationship inference, using convolutional neural networks to analyze summarized co-occurrence histograms from pairs of

genes in scRNA-seq data (22). GCNG transforms the problem of gene regulation network reconstruction into a classification problem. It uses a graph convolutional neural network to fit cell location information and gene expression data and infer the final result (23).

Although much progress has been made, inferring a network of regulatory interactions between genes is still challenging. On one hand, for time-series scRNA-seq data, methods for reconstructing GRNs on bulk data are not directly applicable. As the biological meaning of a sample changes from the average for several cells in bulk data to the value for a single cell, the form of the gene expression data is also changed. Meanwhile, as the approaches devised for single-cell transcriptomics typically require a large number of time points to infer GRNs, they are usually suitable for a small number of genes. Adding a few genes to a network inference analysis may require the inference algorithm to consider many additional regulatory interactions between them. As the number of genes grows, the number of edges and the demand for input data might explode.

Here, we present BiGRN, a novel method of inferring GRNs from time-series scRNA-seq data. BiGRN adopts a bidirectional recurrent neural network to infer GRNs. The recurrent neural network is a deep neural network that can capture complex, non-linear, and dynamic relationships among variables. It maps a neuron to a gene, and maps the connections between neural network layers to the regulatory relationship between genes, giving a good solution to model GRN with biological closeness and mathematical flexibility. Then we transform the reconstruction of GRNs into a regression problem, using the gene expression data of the previous time points to predict the gene expression data of the later time point. Meanwhile, we adopt a bidirectional structure and incorporate an incomplete set of prior knowledge to improve the accuracy and stability of the algorithm. To evaluate the performance of BiGRN, we apply it to four simulated datasets and three real single-cell transcriptomic datasets. We performed a comparison of our results with other state-of-the-art techniques, which shows the better performance of our proposed model.

2 MATERIALS

2.1 The BiGRN Method

In this work, we propose a new computational method BiGRN to reconstruct gene regulatory networks based on bidirectional recurrent neural network and multiple prior networks. The overview of the BiGRN is shown in **Figure 1**. The proposed algorithm consists of the following three main steps. Firstly, we train a deep neural network to infer preliminary gene regulatory networks, where neurons are mapped to genes, and the links between adjacent layers of the neural network are related to gene regulation relationships. Secondly, we incorporate incomplete prior knowledge to filter the candidate regulatory edges obtained in the first step. Finally, we adopt a voting strategy to integrate multiple candidate regulatory networks and utilize a bidirectional strategy to optimize the inferred GRN.

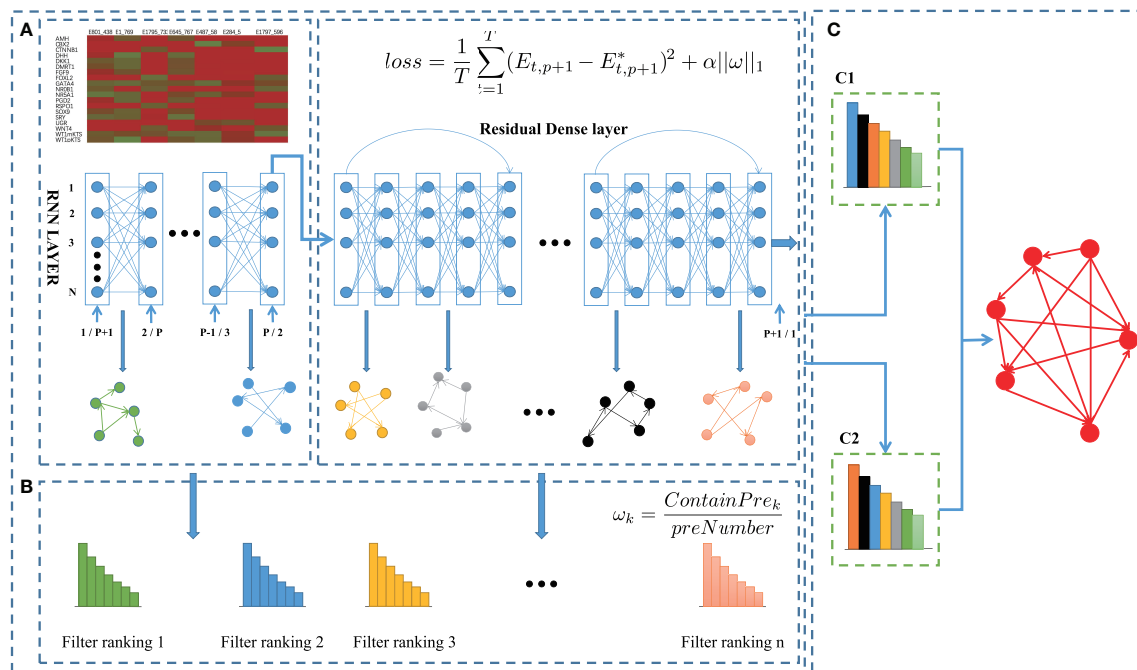


FIGURE 1 | BiGRNN reconstructs GRNs from time-series single cell transcriptome data using bidirection RNN. **(A)** Inferring initial gene regulatory network with RNN. **(B)** Incorporating incomplete prior knowledge to adjust candidate regulatory edges. **(C)** Adopting a voting strategy to integrate multiple candidate regulatory networks, and further utilizing bidirectional model to optimize the inferred GRN.

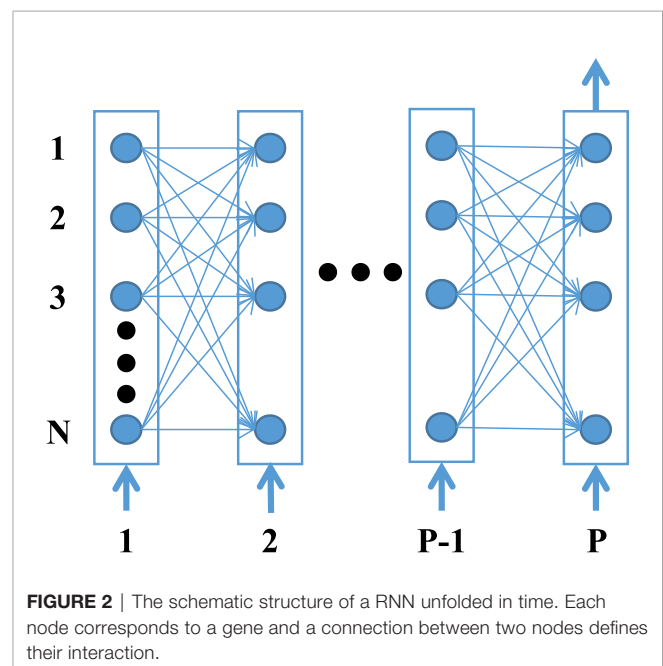
2.1.1 Step 1: Training RNN to Infer the Initial Gene Regulatory Networks

Inferring gene regulatory network from single-cell transcriptomic data is actually to construct a directed graph, where the nodes represent the genes, and the edges represent the regulatory relationships among genes. If we assume that the expression pattern of gene i at time point $p+1$ is the total regulatory effect of the expression values of all genes at the previous p time points, the regulation process can be described as the following function (16):

$$e_{p+1}^i = f^i(E_p) + \epsilon_i \quad (1)$$

where e_{p+1}^i represents the expression value of gene i at the time point $p+1$, E_p represents the expression value of all genes at the previous p time points, and ϵ_i represents the influence of external noise. Specifically, p is the time lag, which represents the maximum time delay of the interaction between genes.

Here, to model the regulation process of different genes in a parallel manner, we adopt RNN to formalize gene regulatory networks (24). A recurrent neural network is a type of artificial neural network that can capture complex, non-linear, and dynamic relationships among variables. It is mainly used for processing sequential data like time series and solving ordinal or temporal problems. As shown in the example RNN (**Figure 2**), each node represents a particular gene and the edges between the nodes represent the regulatory interactions among the genes.



Each layer of the neural network defines the gene expression level of the genes at a specific time point. The expression level of all genes at the time point $p+1$ depends upon the expression level of all the genes at the preceding p time points and the weights of the

corresponding connecting edges with that particular gene (25, 26). Then the regulation process can be formulated as:

$$E_{p+1} = F(E_p) + \in \quad (2)$$

where E_{p+1} represents the expression value of all genes at the time point $p+1$.

To improve the stability of the algorithm, BiGRN integrates multiple fully connected layers with the RNN to train gene expression data. Therefore, the proposed network structure consists of an RNN, multiple fully connected layers with ResNet residual connections (27), and an output layer. In detail, the proposed RNN contains p layers corresponding to p time points, with multiple inputs and one output. Subsequently, the output of the RNN is used as the input of these fully connected layers. To avoid the over-fitting problem usually caused by the deep neural network, BiGRN adds a ResNet residual connection for every five fully connected layers. In the experiment, we set the number of the connected layers ranging from 10 to 100. We find that too few fully connected layers will lead to a significant decrease in the stability of the algorithm, whereas too many fully connected layers can not improve the accuracy but increase the running time of the algorithm. Therefore, we use 50 fully connected layers and add a ResNet structure. To train the deep neural network, we take the gene expression data of the genes at the previous p time points as input, and the gene expression data at the $p+1$ time point as output. Then, the problem is transformed into a supervised regression problem, which overcomes the difficulty of obtaining training labels.

Here, we utilize mean square loss (MSE) as the regression loss function for deep neural network training. The RNN is a fully connected structure, whereas the regulatory network is usually sparsely connected. Thus, we add L1 regularization in the objective function, aiming to control the sparsity of the resulted weight matrix w . The loss function is defined as follows:

$$\text{loss} = \frac{1}{T} \sum_{t=1}^T (E_{t,p+1} - E_{t,p+1}^*)^2 + \alpha \|w\|_1 \quad (3)$$

where $E_{t,p+1}^*$ and $E_{t,p+1}$ respectively represent the predicted and the real expression value of all genes at the time point $t+p+1$. $\alpha\|w\|_1$ is the regularized term.

For the training process, when the objective function converges to the minimum, the algorithm extracts the multiple weight matrixes between the RNN layer and each fully connected layer. Then we normalize each basic weight matrix separately. According to the proposed network structure, the weight matrix corresponds to the regulatory relationships among genes, which can be used to reconstruct a candidate gene regulatory network. For each matrix, we take the top m (Usually 1.2 times the number of inferred regulation edge) connections as the candidate regulatory edges. As multiple weight matrixes are obtained after the training process, we can infer multiple candidate gene regulatory networks, which are used as the basic voters to determine the final regulatory edges in the following steps.

2.1.2 Step 2: Incorporating Prior Knowledge to Adjust Candidate Regulatory Edges

During the above training process, the final loss function of the model usually cannot be completely reduced to zero due to the influence of external noise. Meanwhile, in convex optimization problems, there are a large number of approximate solutions near the global optimal. In order to improve the accuracy of the GRN inference, some prior knowledge can be utilized to filter the candidate regulatory edges. The previous method, such as NetREX and MiPGRN, assumes that the prior network and the target GRN have some similarity, and then bias the optimization procedure toward networks that overlap with the prior (28, 29). Here, if the initial candidate GRN defined by the basic weight matrix has more overlap with the prior network, it is considered to be closer to the final inferred GRN. Correspondingly, this candidate GRN is assigned a higher voting weight in the following ensemble process. Specifically, the weight of the candidate GRN is calculated according to the following strategy:

$$\omega_k = \frac{\text{ContainPre}_k}{\text{preNumber}} \quad (4)$$

where ω_k represents the weight of the k_{th} initial GRN, ContainPre_k denotes the number of candidate edges in the k_{th} inferred GRN overlapping with the prior network, and preNumber represents the number of the prior edges.

As the usable prior knowledge usually does not exist for given datasets, here we adopt a general strategy to obtain an incomplete prior edge set. We utilize different computational algorithms to predict the putative GRNs, apply the method NETREX to optimize the predictions, and then integrate the top 10% of the resulted edges to obtain an incomplete prior edge set (29). Through evaluating different methods, here we select three methods, including GRNBOOST2, PPCOR, and PIDC. These three methods respectively adopt a different strategy to predict GRNs. NetREX is an algorithm based on Network Component Analysis (NCA) to optimize the predicted GRN (28).

2.1.3 Step 3: Utilizing a Bidirectional Model to Optimize the Inferred GRN

Based on the deep neural network, we obtain K candidate GRNs, and each candidate GRN possesses an adjusted weight matrix. Next, we integrate these K different initial gene regulatory networks. The voting strategy is the addition of weights, and finally a global regulatory edge ranking is obtained according to the weights. For the regulatory edge of gene i to gene j , the weight e_{ij} is calculated as:

$$e_{ij} = \sum_{k=1}^K \omega_k * e_{ij}^k \quad (5)$$

where ω_k represents the weight of the k_{th} candidate GRN, and e_{ij}^k represents the regulatory edge of gene i to gene j in the k_{th} candidate GRN.

Inspired by the bidirectional model of the algorithm BiXGBoost (16), we further utilize the bidirectional model to fully mine the regulatory genes and target genes. Different from BiXGBoost which proposes local_in and local_out models to deal

with forward and reverse inference, we use forward time-series expression data and reverse time-series expression data to respectively infer two regulatory networks. For the reverse time series data, the weight matrix obtained by the model represents the regulatory strength between gene. Next, considering the directionality of the regulatory relationship, we assume that genes expressed at earlier time points regulate genes expressed at later time points. Therefore, for the reverse inference, the input of the algorithm is the gene expression data at p time points of $p+1, p, p-1, \dots, 2$, and the output is the gene expression data of the first time point. After getting the trained model, the algorithm extracts the weight matrix ω^r , and the subsequent operations are consistent with the forward model. Then the algorithm will eventually get two regulatory networks, and also use voting strategies to integrate forward and reverse results to get the final inferred regulatory network:

$$e_{ij}^* = e_{ij}^f + e_{ij}^r \quad (6)$$

where e_{ij}^f represents the weight e_{ij} obtained from forward inferring, and e_{ij}^r represents the weight e_{ij} obtain the reverse inferred GRN. Based on the calculated new weights of these edges, we rank the regulatory edges and select the top m regulatory edges to form the inferred GRN.

2.2 Datasets

Real scRNA-seq data sets. In order to evaluate the performance of the proposed algorithm on real scRNA-seq datasets, we select three widely used scRNA-seq data sets as the previous method SCODE did (14). The first dataset is derived from primitive endoderm (PrE) cells differentiated from mouse ES cells (measured at 0, 12, 24, 48, and 72 hours, respectively) and contains 456 cells (30). The second dataset is derived from examining direct reprogramming from mouse embryonic fibroblast (MEF) cells to myocytes (measured on 0, 2, 5, and 22 days), and this data set contains 405 cells (31). The third dataset is the scRNA-seq data of definitive endoderm cells derived from human ES cell differentiation (measured at 0, 12, 24, 36, 72, and 96 hours, respectively), and this dataset contains 758 cells (32). In order to verify the inferred GRN on these scRNA-seq datasets, SCODE used the transcription factor regulation network database (<http://www.regulatorynetworks.org>), which was constructed from DNaseI footprints and TF-binding motifs (33, 34). They integrated the TF regulatory networks of human and mouse, and extracted 100*100 TF regulatory networks for each dataset. We use this regulatory network as the correct network for each data set, and calculate the AUC value of the inferred network.

Simulated data sets. For real single-cell gene expression datasets, it is usually difficult to obtain the real labels for the edges in the gene regulatory network. In order to verify the effectiveness of the proposed method and compare it with existing methods, four simulated datasets are also used to evaluate the inferred results (6). These four data sets are all generated by the Boolean model simulating real cell expression data (35). The advantage of using the Boolean model is that it can

be used as a real biological regulatory network to evaluate the performance of the reconstructed regulatory network. We utilize the four gene expression data sets of gonadal sex determination (GSD), hematopoietic stem cell differentiation (HSC), ventral spinal cord development (VSC), and mammalian cortical development (mCAD) to evaluate the performance of the algorithm. These four datasets all contain 10 simulation subsets composed of 2000 cells. The detailed information of the data sets is shown in **Table 1**.

2.3 Evaluation Metrics

To evaluate the performance of different methods in inferring GRNs, we utilize two widely-used metrics AUROC and AUPRC. Specifically, AUROC is the area under the ROC based on TPR and FPR. AUPRC is the area under the PRC based on the precision rate and the recall rate.

$$TPR = \frac{TP}{TP + FN} \quad (7)$$

$$FPR = \frac{FP}{FP + TN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = TPR \quad (10)$$

where TP and FP indicate the numbers of true and false positives, and TN and FN are true and false negatives. For the simulated datasets, we calculated the average of the AUROC and AUPRC to evaluate the accuracy of the inferred network on different subsets. Further, we calculated the overall score of $AUROC_{score}$ and $AUPRC_{score}$. The definition is as follows:

$$AUROC_{score} = \frac{1}{n} \sum_{i=1}^n AUROC_i \quad (11)$$

$$AUPRC_{score} = \frac{1}{n} \sum_{i=1}^n AUPRC_i \quad (12)$$

where n represents the number of subsets in each dataset (taking the dataset GSD as an example, n is 10). $AUROC_i$ and $AUPRC_i$

TABLE 1 | Details of time-seris gene expression datasets used in the experiment.

Dataset	Genes	Time points	Cells
GSD	19	734	2000
HSC	11	731	2000
VSC	8	492	2000
mCAD	5	492	2000
Real Dataset1	100	456	456
Real Dataset2	100	405	405
Real Dataset3	100	758	758

respectively denote the average AUROC and AUPRC of the algorithm on the i_{th} data set.

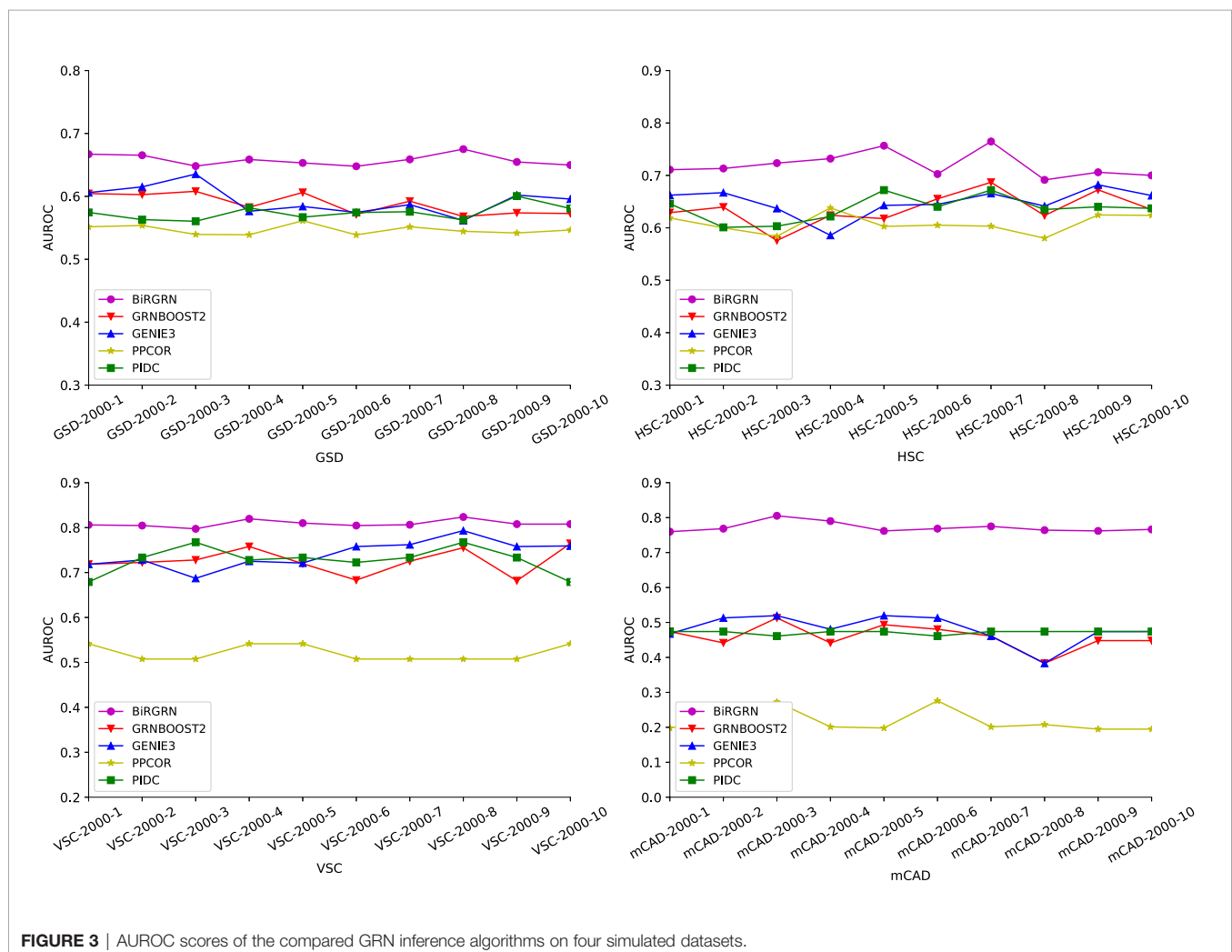
3 RESULTS

3.1 Performance on Simulated Data Sets

To evaluate the effectiveness of BiGRN, We apply the proposed GRN inference method to four simulated datasets, including datasets related to hematopoietic stem cell differentiation (HSC), gonadal sex determination (GSD), ventral spinal cord development (VSC), and mammalian cortical development (mCAD). In detail, each dataset is generated by the Boolean model in previous study (6), including 10 data subsets composed of 2000 cells and multiple time points. **Table 1** lists the detailed information of these datasets. We take each synthetic network as the ground truth and adopt two metrics to evaluate the inferred GRNs. We utilize both the area under the receiver operating characteristic curve and the area under the precision-recall curve (AUROC/AUPRC) as our evaluation metrics across the 10 different datasets. Further, we compare BiGRN with four

widely used methods, including three prior algorithms GRNBOOST2 (36), PPCOR, PIDC, and the classic algorithm GENIE3.

Figures 3, 4 respectively show the AUROC and AUPRC of these compared methods on the four datasets. As can be seen, BiGRN outperforms the compared methods on all four simulated datasets. We observe significant improvement over the three methods (GRNBOOST2, PPCOR, and PIDC) using the provided prior edge sets. Also, BiGRN performs better than the widely used method GENIE3. Compared with the second-ranked algorithm on GSD, BiGRN has a 6.2% increase in AUROC and a 33.3% increase in AUPRC. On the dataset HSC, BiGRN achieves an improvement of 11.3% in AUROC and 10.2% in AUPRC over the other methods. For the dataset VSC, BiGRN has a 3.8% higher AUROC and a 13.2% higher AUPRC than the second-ranked algorithm, whereas the performance of PPCOR is not as good as other methods. And as shown in the figures, the compared algorithms perform poorly on mCAD, and the AUROC values of the four algorithms are only around 0.5. In contrast, our proposed BiGRN reaches a mean AUROC of 0.8. Compared with the second-ranked algorithm, the AUROC of



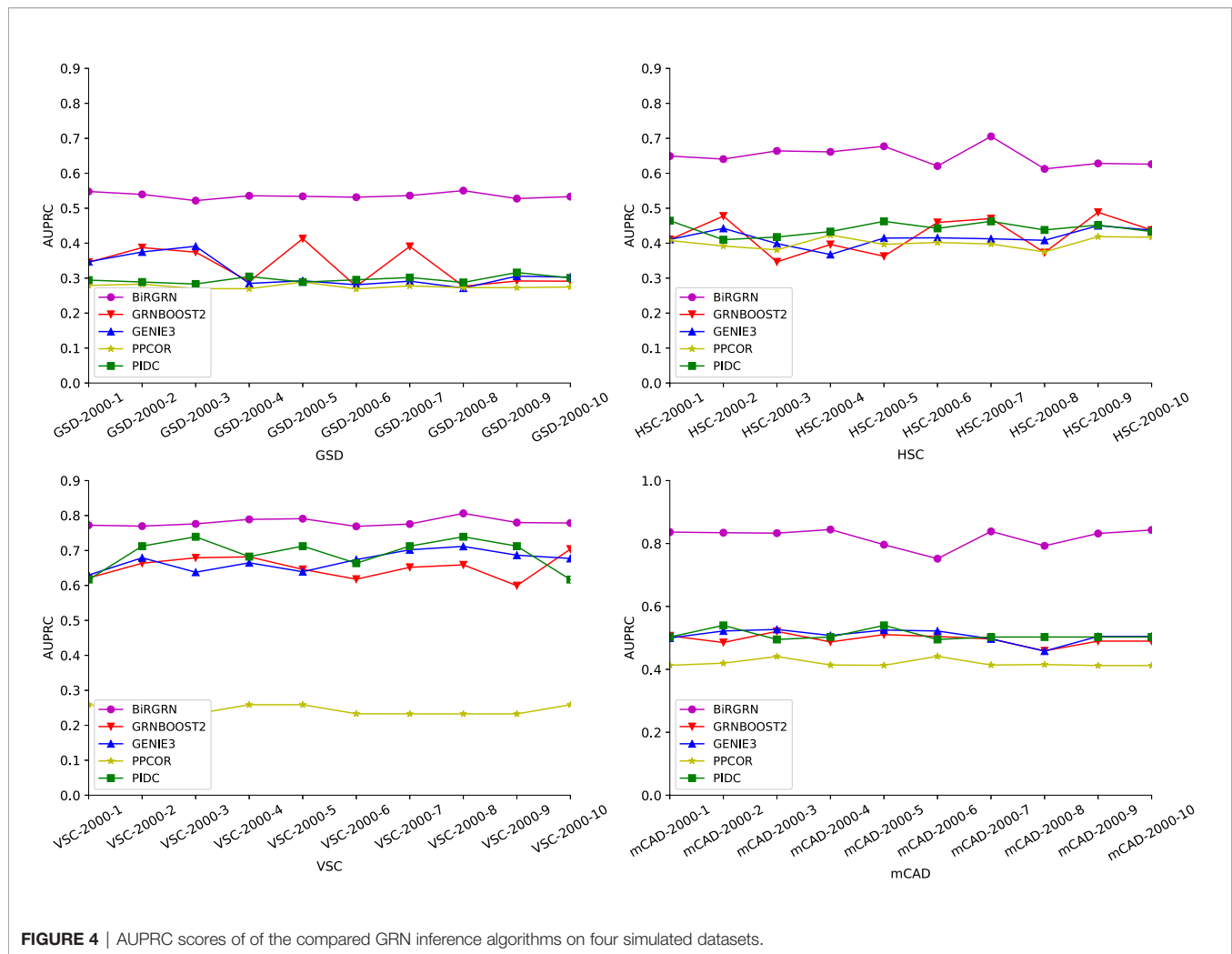


FIGURE 4 | AUPRC scores of the compared GRN inference algorithms on four simulated datasets.

BiGRN increases by 55%, AUPRC increases by 56.1%. Furthermore, **Figure 5** presents the overall score of these algorithms on the four datasets, the histogram of the overall score also intuitively shows that the algorithm in this paper has a better performance.

3.2 Performance on the Real scRNA-Seq Data Sets

We next measure the performance of BiGRN for inferring GRNs on real datasets. Here, BiGRN is applied to three real time-series scRNA-seq datasets. As previous studies did (14), the inferred GRN is validated by the TF regulatory network based on DNaseI footprints and TF-binding motifs. We calculate the AUROC values of BiGRN given 15% of the prior knowledge and compared them with four widely used methods, including GENIE3, LEAP, BiXGBoost, and SCODE. Specifically, GENIE3 is a classic random forest-based method for inferring GRNs. The algorithm BiXGBoost adopts local-in and local-out models to utilize time information in two directions and integrates

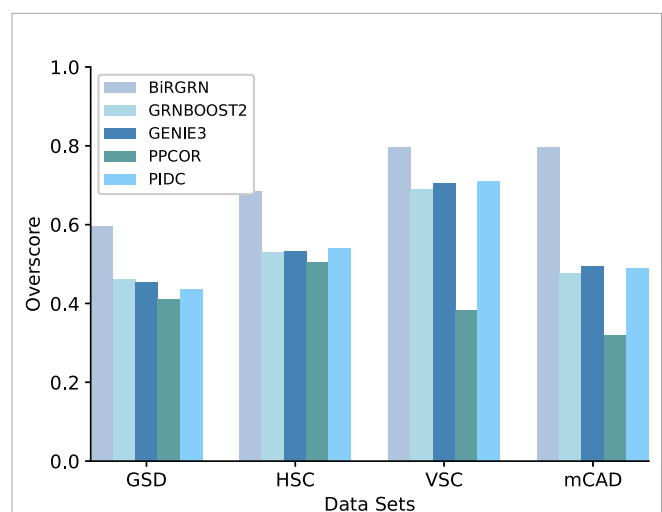


FIGURE 5 | The overall score of the algorithm on the four simulated datasets.

TABLE 2 | The AUROC value of the algorithm on three real scRNA-seq datasets.

Algorithm	Dataset 1	Dataset 2	Dataset 3
BiGRN	0.571	0.573	0.562
GENIE3	0.503	0.498	0.507
LEAP	0.487	0.5	0.494
SCODE	0.536	0.581	0.523
BiXGBoost	0.509	0.479	0.510

The value in bold represents the highest value in the column.

XGBoost to evaluate the feature importance. LEAP and SCODE are two advanced GRN inference methods for scRNA-seq data.

Table 2 presents the performance of these compared methods on the real scRNA-seq datasets. Compared to other network inference algorithms, our proposed algorithm BiGRN can infer TF regulatory networks with high performance. On Dataset 1 and Dataset 3, the AUROC values of BiGRN are obviously higher than those of the four previous algorithms. Compared with the second-ranked algorithm SCODE, the AUROC of BiGRN is increased by 6.5% on dataset1, and the AUROC of BiGRN is increased by 7.4% on dataset3. On Dataset 2, the performance of BiGRN is close to the best performance. These results indicate that the RNN structure utilized in BiGRN has a high capability of incorporating time point information, which is effective in network inference.

We also record the runtime of each method on three real data sets. As shown in **Table 3**, LEAP and GENIE3 have the highest efficiency. The runtime of BiGRN is at the median level among several methods. On Dataset 1 and Dataset 2, BiGRN runs for 1min and 58s, which is much faster than SCODE and BiXGBoost. These results show that BiGRN can efficiently use temporal information to rapidly reconstruct gene regulatory networks.

3.3 Ablation Study

As BiGRN is mainly composed of the bidirectional RNN integrating the forward and reverse training, and the voting model incorporating prior knowledge, we further investigate the

impact of the different components on the overall performance. Accordingly, we obtain three variants of BiGRN, including BiGRN-Prior(the model removing incorporated prior knowledge), BiGRN-Forward (the model removing forward training), and BiGRN-Reverse (the model removing reverse training). We respectively carry out the ablation study on the four simulated datasets. **Table 4** summarizes the performance comparison between BiGRN and these three variants.

We first evaluate the contribution of prior information for guiding the voting process in the model. The results show that the removal of the prior information results in a slight drop in performance. Without incorporating prior information, the network is able to reconstruct a relatively coarse segmentation. Without further guidance of prior information, it might be not able to refine it properly. To further inspect the effectiveness of the bidirectional model, we respectively compare the performance of the BiGRN without forwarding training and reverse training. From the table, we observe that the performance of two single directional training models is similar, and they are slightly lower than those of the bidirectional training model. This result of ablation Study indicates the forward training and the reverse training might be complementary to each other, and thus the bidirectional RNN structure is capable of capturing more regulation relationships among genes. On the whole, these results demonstrate that both the components are contributive to the performance of BiGRN.

4 CONCLUSION

Many cellular processes, either in development or disease progression are governed by complex gene regulatory mechanisms. GRN reverse engineering methods attempt to infer GRNs from large-scale transcriptomic data using computational or statistical models. A plethora of GRN inference methods has been proposed. However, with the development of single-cell sequencing technology, traditional GRN inference methods designed for bulk transcriptomic data

TABLE 3 | The runtime of each method for three real datasets.

Runtime ¹	BiGRN	SCODE	GENIE3	LEAP	BiXGBoost
Dataset 1	1min58s	7min3s	58s	6s	min49s
Dataset 2	1min58s	6min39s	52s	4s	3min21s
Dataset 3	2min22s	8min49s	1min6s	11s	3min58s

¹All algorithms except BiXGBoost are tested on Beeline(a benchmarking software for GRN inference algorithms). The computations were performed on a Lenovo Legion R7000 2020 equipped with a 3.0GHz AMD Ryzen 5 4600H processor a 4GB NVIDIA GeForce GTX 1650Ti and 16GB of 3200MHz DDR4 RAM.

TABLE 4 | The AUROC value of the algorithm and three variants on the simulated datasets.

Dataset	BiGRN	Prior network	Forward	Reverse
GSD	0.597	0.544	0.583	0.587
HSC	0.684	0.586	0.656	0.660
VSC	0.795	0.624	0.761	0.763
mCAD	0.796	0.678	0.796	0.792

The value in bold represents the highest value in the row.

might be unsuitable to process large quantities of scRNA-seq data. In this paper, we proposed a novel computational method BiGRN to infer GRNs from time-series scRNA-seq data. BiGRN utilizes a bidirectional recurrent neural network to infer GRNs. The recurrent neural network is a complex neural network, which can capture complex, non-linear, and dynamic relationships among variables. It maps a neuron to a gene, and maps the connections between neural network layers to the regulatory relationship between genes, giving a good solution to model GRN with biological closeness and mathematical flexibility. Then we transform the reconstruction of GRNs problem into a regression problem that uses the gene expression data of the previous time points to predict the gene expression data of the later time node. In order to improve the accuracy of the algorithm, the method can use an incomplete set of prior knowledge. The developed model has been tested on four simulated data and three real datasets. We performed a comparison of our results with other state-of-the-art techniques which shows the superiority of our proposed model. The experiments conducted on simulated datasets and real scRNA-seq datasets demonstrate that BiGRN can infer gene regulatory networks with high performance, which that the proposed bidirectional RNN structure is effective in GRN inference.

REFERENCES

- Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. *Cell* (2018) 172:650–65. doi: 10.1016/j.cell.2018.01.029
- Fiers MW, Minnoye L, Aibar S, Bravo González-Blas C, Kalender Atak Z, Aerts S. Mapping Gene Regulatory Networks From Single-Cell Omics Data. *Briefings Funct Genomics* (2018) 17:246–54. doi: 10.1093/bfpg/elx046
- Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-Specific Regulatory Circuits Reveal Variable Modular Perturbations Across Complex Diseases. *Nat Methods* (2016) 13:366–70. doi: 10.1038/nmeth.3799
- Iacono G, Massoni-Badosa R, Heyn H. Single-Cell Transcriptomics Unveils Gene Regulatory Network Plasticity. *Genome Biol* (2019) 20:1–20. doi: 10.1186/s13059-019-1713-4
- Fazlaty H, Rago L, Youssef KK, Ocaña OH, Garcia-Asencio F, Arcas A, et al. A Gene Regulatory Network to Control Emt Programs in Development and Disease. *Nat Commun* (2019) 10:1–16. doi: 10.1038/s41467-019-13091-8
- Pratapa A, Jaliha AP, Law JN, Bharadwaj A, Murali T. Benchmarking Algorithms for Gene Regulatory Network Inference From Single-Cell Transcriptomic Data. *Nat Methods* (2020) 17:147–54. doi: 10.1038/s41592-019-0690-6
- Delgado FM, Gómez-Vela F. Computational Methods for Gene Regulatory Networks Reconstruction and Analysis: A Review. *Artif Intell Med* (2019) 95:133–45. doi: 10.1016/j.artmed.2018.10.006
- Castro DM, De Veaux NR, Miraldi ER, Bonneau R. Multi-Study Inference of Regulatory Networks for More Accurate Models of Gene Regulation. *PLoS Comput Biol* (2019) 15:e1006591. doi: 10.1371/journal.pcbi.1006591
- Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring Regulatory Networks From Expression Data Using Tree-Based Methods. *PLoS One* (2010) 5:e12776. doi: 10.1371/journal.pone.0012776
- Ślawek J, Arodz T. Ennet: Inferring Large Gene Regulatory Networks From Expression Data Using Gradient Boosting. *BMC Syst Biol* (2013) 7:1–13. doi: 10.1186/1752-0509-7-106
- Kim S. Ppcor: An R Package for a Fast Calculation to Semi-Partial Correlation Coefficients. *Commun Stat Appl Methods* (2015) 22:665. doi: 10.5351/CSAM.2015.22.6.665

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The real dataset can be found in <https://github.com/hmatsu1226/SCODE>, the simulated datasets are all from Beeline and can be found in <https://github.com/Murali-group/Beeline>.

AUTHOR CONTRIBUTIONS

YG and XH are responsible for the main idea, as well as the completion of the manuscript. XH has developed the algorithm and performed data analysis. GZ, CY, and GX have coordinated data preprocessing and supervised the effort. All authors have read and approved the final manuscript.

FUNDING

This work was sponsored in part by the National Natural Science Foundation of China (62172088), National Key Research and Development Program of China (2016YFC0901704), and Shanghai Natural Science Foundation (21ZR1400400, 19ZR1402000).

- Chan TE, Stumpf MP, Babbie AC. Gene Regulatory Network Inference From Single-Cell Data Using Multivariate Information Measures. *Cell Syst* (2017) 5:251–67. doi: 10.1016/j.cels.2017.08.014
- Specht AT, Li J. Leap: Constructing Gene Co-Expression Networks for Single-Cell RNA-Sequencing Data Using Pseudotime Ordering. *Bioinformatics* (2017) 33:764–6. doi: 10.1093/bioinformatics/btx729
- Matsumoto H, Kiryu H, Furusawa C, Ko MS, Ko SB, Gouda N, et al. Scode: An Efficient Regulatory Network Inference Algorithm From Single-Cell RNA-Seq During Differentiation. *Bioinformatics* (2017) 33:2314–21. doi: 10.1093/bioinformatics/btx194
- Papili, Gao N, Ud-Dean SM, Gandrillon O, Gunawan R. Sincerities: Inferring Gene Regulatory Networks From Time-Stamped Single Cell Transcriptional Expression Profiles. *Bioinformatics* (2018) 34:258–66. doi: 10.1093/bioinformatics/btx575
- Zheng R, Li M, Chen X, Wu FX, Pan Y, Wang J. Bixgboost: A Scalable, Flexible Boosting-Based Method for Reconstructing Gene Regulatory Networks. *Bioinformatics* (2019) 35:1893–900. doi: 10.1093/bioinformatics/bty908
- Wang J, Ma A, Ma Q, Xu D, Joshi T. Inductive Inference of Gene Regulatory Network Using Supervised and Semi-Supervised Graph Neural Networks. *Comput Struct Biotechnol J* (2020) 18:3335–43. doi: 10.1016/j.csbj.2020.10.022
- Shu H, Zhou J, Lian Q, Li H, Zhao D, Zeng J, et al. Modeling Gene Regulatory Networks Using Neural Network Architectures. *Nat Comput Sci* (2021) 1:491–501. doi: 10.1038/s43588-021-00099-8
- Song Q, Su J. Dstg: Deconvoluting Spatial Transcriptomics Data Through Graph-Based Artificial Intelligence. *Briefings Bioinf* (2021) 22:1–13. doi: 10.1093/bib/bbaa414
- Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and Analysis of Cell-Cell Communication Using Cellchat. *Nat Commun* (2021) 12:1–20. doi: 10.1038/s41467-021-21246-9
- Kim J, Jakobsen S T, Natarajan KN, Won KJ. Tenet: Gene Network Reconstruction Using Transfer Entropy Reveals Key Regulatory Factors From Single Cell Transcriptomic Data. *Nucleic Acids Res* (2021) 49:e1–1. doi: 10.1093/nar/gkaa1014
- Yuan Y, Bar-Joseph Z. Deep Learning for Inferring Gene Relationships From Single-Cell Expression Data. *Proc Natl Acad Sci* (2019) 116:27151–8. doi: 10.1073/pnas.1911536116

23. Yuan Y, Bar-Joseph Z. Gcng: Graph Convolutional Networks for Inferring Gene Interaction From Spatial Transcriptomics Data. *Genome Biol* (2020) 21:1–16. doi: 10.1186/s13059-020-02214-w
24. Zaremba W, Sutskever I, Vinyals O. Recurrent Neural Network Regularization. *ArXiv Prepr ArXiv* (2014) 1409:2329. doi: 10.48550/arXiv.1409.2329
25. Cheng L, Hou ZG, Lin Y, Tan M, Zhang WC, Wu FX. Recurrent Neural Network for Non-Smooth Convex Optimization Problems With Application to the Identification of Genetic Regulatory Networks. *IEEE Trans Neural Networks* (2011) 22:714–26. doi: 10.1109/TNN.2011.2109735
26. Biswas S, Acharyya S. A Bi-Objective Rnn Model to Reconstruct Gene Regulatory Network: A Modified Multi-Objective Simulated Annealing Approach. *IEEE/ACM Trans Comput Biol Bioinf* (2018) 15:2053–9. doi: 10.1109/TCBB.2017.2771360
27. He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV: IEEE (2016). 770–8. p.
28. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network Component Analysis: Reconstruction of Regulatory Signals in Biological Systems. *Proc Natl Acad Sci* (2003) 100:15522–7. doi: 10.1073/pnas.2136632100
29. Gan Y, Xin Y, Hu X, Zou G. Inferring Gene Regulatory Network From Single-Cell Transcriptomic Data by Integrating Multiple Prior Networks. *Comput Biol Chem* (2021) 93:107512. doi: 10.1016/j.compbiolchem.2021.107512
30. Shimosato D, Shiki M, Niwa H. Extra-Embryonic Endoderm Cells Derived From Es Cells Induced by Gata Factors Acquire the Character of Xen Cells. *BMC Dev Biol* (2007) 7:1–12. doi: 10.1186/1471-213X-7-80
31. Treutlein B, Lee QY, Camp JG, Mall M, Koh W, Shariati SAM, et al. Dissecting Direct Reprogramming From Fibroblast to Neuron Using Single-Cell Rna-Seq. *Nature* (2016) 534:391–5. doi: 10.1038/nature18323
32. Chu LF, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, et al. Single-Cell Rna-Seq Reveals Novel Regulators of Human Embryonic Stem Cell Differentiation to Definitive Endoderm. *Genome Biol* (2016) 17:1–20. doi: 10.1186/s13059-016-1033-x
33. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. *Cell* (2012) 150:1274–86. doi: 10.1016/j.cell.2012.04.040
34. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, et al. Conservation of Trans-Acting Circuitry During Mammalian Regulatory Evolution. *Nature* (2014) 515:365–70. doi: 10.1038/nature13972
35. Giacomantonio CE, Goodhill GJ. A Boolean Model of the Gene Regulatory Network Underlying Mammalian Cortical Area Development. *PloS Comput Biol* (2010) 6:e1000936. doi: 10.1371/journal.pcbi.1000936
36. Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, et al. Grnboost2 and Arboreto: Efficient and Scalable Inference of Gene Regulatory Networks. *Bioinformatics* (2019) 35:2159–61. doi: 10.1093/bioinformatics/bty916

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.

Copyright © 2022 Gan, Hu, Zou, Yan and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



NKX2-8/PTHrP Axis-Mediated Osteoclastogenesis and Bone Metastasis in Breast Cancer

Ainiwaerjiang Abudourousuli^{1,2†}, Suwen Chen^{1,2†}, Yameng Hu^{1,2†}, Wanying Qian^{1,2}, Xinyi Liao^{1,2}, Yingru Xu^{1,2}, Libing Song³, Shuxia Zhang^{4*} and Jun Li^{1,2*}

¹ Key Laboratory of Liver Disease of Guangdong Province, The Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, China, ² Department of Biochemistry, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China, ³ State Key Laboratory of Oncology in South China, Cancer Center, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University, Guangzhou, China, ⁴ Department of Oncology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

OPEN ACCESS

Edited by:

Liang Cheng,
Harbin Medical University, China

Reviewed by:

Jincheng Zeng,
Guangdong Medical University, China
Jiarong Chen,
Jiangmen Central Hospital, China

*Correspondence:

Jun Li
lijun37@mail.sysu.edu.cn
Shuxia Zhang
zhangshx8@mail2.sysu.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 29 March 2022

Accepted: 29 April 2022

Published: 30 May 2022

Citation:

Abudourousuli A, Chen S, Hu Y, Qian W, Liao X, Xu Y, Song L, Zhang S and Li J (2022) NKX2-8/PTHrP Axis-Mediated Osteoclastogenesis and Bone Metastasis in Breast Cancer. *Front. Oncol.* 12:907000. doi: 10.3389/fonc.2022.907000

Bone metastasis is one of the most common distant metastasis of breast cancer, which could cause serious skeletal disease and increased cancer-related death. Therefore, identification of novel target(s) to develop therapeutics would improve patient outcomes. The role of NKX2-8 in modulation of bone remodeling was determined using osteoclastogenesis and micro-CT assays. The expression of NKX2-8 was examined via immunohistochemistry analysis in 344 breast cancer tissues. The mechanism underlying NKX2-8-mediated PTHrP downregulation was investigated using biotinylated deactivated Cas9 capture analysis, chromatin immunoprecipitation, co-immunoprecipitation assays. A bone-metastatic mouse model was used to examine the effect of NKX2-8 dysregulation on breast cancer bone metastasis and the impact of three PTHrP inhibitor on prevention of breast cancer bone metastasis. The downregulated expression of NKX2-8 was significantly correlated with breast cancer bone metastasis. *In vivo* bone-metastatic mouse model indicated that silencing NKX2-8 promoted, but overexpressing NKX2-8 inhibited, breast cancer osteolytic bone metastasis and osteoclastogenesis. Mechanistically, NKX2-8 directly interacted with HDAC1 on the PTHrP promoter, which resulted in a reduction of histone H3K27 acetylation, consequently transcriptionally downregulated PTHrP expression in breast cancer cells. Furthermore, targeting PTHrP effectively inhibited NKX2-8-downregulation-mediated breast cancer bone metastasis. Taken together, our results uncover a novel mechanism underlying NKX2-8 downregulation-mediated breast cancer bone metastasis and represent that the targeting PTHrP might be a tailored treatment for NKX2-8 silencing-induced breast cancer bone metastasis.

Keywords: NKX2-8, breast cancer, bone metastasis, osteoclastogenesis, PTHrP

INTRODUCTION

Recently, it has been recently reported that the incidence of breast cancer has become the highest and one of top leading cause of cancer death in the world (1), generally caused by the outgrowth of cancer cells in distant organs, such as bone, brain, liver and lungs (2). More than 70% advanced breast cancer develops bone metastasis that significantly reduced the quality of life and survival of patients (3, 4). However, current treatments for bone metastasis have limited efficacy. Despite the temporary effect of bisphosphonates and denosumab (an anti-receptor activator of NF- κ B ligand (RANKL) monoclonal antibody) on decreasing the risk of skeletal-related events (SREs), no significant effect has been observed in terms of overall survival. Thus, it is an urgent necessity to identify the vital molecule(s) that contribute to bone metastasis of breast cancer, which could be used as diagnostic marker and novel treatment target for bone-metastasis of breast cancer.

Breast cancer bone metastasis are mostly osteolytic metastases, caused by recruitment and activation of osteoclasts to the tumor-bone interface to form aberrant bone resorption (5). Evidences proves that cancer-secreted factors directly or indirectly activated osteoclasts to absorb the bone matrix, leading to formation of a “bone pre-metastatic niche” to support cancer bone metastasis, thus generating a “vicious cycle” between bone-metastatic tumor cells and the bone pre-metastatic niche (4, 5). Parathyroid-hormone related peptide (PTHrP) is the most important osteoclast-activating factors released by cancer cells, including breast cancer cells (3). It has been proven that breast cancer cells do not express RANKL, but produce PTHrP to elevate RANKL secretion from osteoblasts to activate osteoclasts (6). Although it has been widely demonstrated that PTHrP plays the central role in breast cancer bone metastasis, the molecular mechanism in regulation of PTHrP expression remain largely unclear.

Human Nk2 homeobox 8 (NKX2-8), a NK2-related transcription factor, was reported to be downregulated in multiple tumors, which contributed to initiation, progression and development of cancer (7–13). Our previous studies have demonstrated that downregulation of NKX2-8 significantly contributed to malignant progression and development of bladder cancer (10) and esophageal squamous carcinoma (11), and deletion of NKX2-8 conferred chemoresistance on epithelial ovarian cancer (13). Herein, we demonstrated that NKX2-8 transcriptionally downregulated the PTHrP expression through directly interacting with histone deacetylase 1 (HDAC1) on the PTHrP promoter, resulting in a reduction of histone H3K27 acetylation, which transcriptionally downregulated PTHrP level in breast cancer. Silencing NKX2-8 induced PTHrP expression to activate osteoclasts, which generated a bone-metastasis supported “bone pre-metastatic niche”. These results not only demonstrated the crucial role of NKX2-8 reduction in osteoclastogenesis-induced breast cancer bone metastasis but also represent a potential therapeutic strategy to treat bone metastasis of NKX2-8-downregulated breast cancer.

MATERIALS AND METHODS

Cell Lines and Cell Culture

The breast cancer cell lines, including MDA-MB-231 and SCP2, osteoclast precursors Raw 264.7 cells and osteoblast precursors MC3T3-E1 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) (Gibco, Grand Island, NY, USA) plus with 10% fetal bovine serum (FBS; Gibco). SCP2 cell line was kindly provided by Prof. Guohong Hu in the School of Medicine, Chinese Academy of Sciences and Shanghai Jiao Tong University. The abovementioned cell lines have been examined the contamination of mycoplasma and authenticated by short tandem repeat (STR) fingerprinting at the department of Forensic in Sun Yat-Sen University (China).

Plasmids, Retroviral Infection and Transfection

The retroviral vector pMSCV-neo was employed to construct the pMSCV-neo/NKX2-8 plasmid. Short hairpin RNAs (shRNAs) targeting NKX2-8, PTHrP, and HDAC1 were cloned into the pSuper Retro viral vector. The detailed information of shRNA oligonucleotides is presented in **Supplementary Table 1**. The region of human PTHrP promoter, which included nucleotides from –2000 to +500 around transcription start site, was subcloned into pGL3-Control luciferase reporter vector (Promega, Madison, WI, USA). Stable cell lines expressing NKX2-8, or NKX2-8 shRNA(s), PTHrP shRNA(s) and HDAC1 shRNA(s) were established by retroviral infection and 10 days selection with puromycin (0.5 μ g/mL).

RNA Extraction, Reverse Transcription and Real-Time PCR

The Trizol (Life Technologies, Carlsbad, CA, USA) reagent was used to extract total RNA from the indicated cells performed as the manufacturer's instructions. The extracted RNA was reverse transcribed to cDNA, which was used in the quantitative real-time PCR step of the quantitative real-time reverse transcription PCR (qRT-PCR) protocol. The primers and probes used for qRT PCR were designed using the Primer Express v 2.0 software (Applied BioSystems, Foster City, CA, USA). The expression level of indicated genes was normalized to the expression of the housekeeping gene GAPDH (encoding glyceraldehyde-3-phosphate dehydrogenase) and calculated as $2^{-[(C_t \text{ of gene}) - (C_t \text{ of GAPDH})]}$, which C_t indicated the cycle threshold for each gene.

All primers are listed in **Supplementary Table 1**.

Luciferase Assay

Total 1×10^3 indicated cells were cultured in 48-well for 24 h and then transfected with control or correspond luciferase reporter plasmids (100 ng) and pRL-TK renilla plasmid (5ng) (Promega, Madison, WI) using the Lipofectamine 3000 reagent (Invitrogen, Carlsbad, CA, USA). The Dual Luciferase Reporter Assay Kit (Promega, Madison, WI) was used to measure the luciferase and renilla signals after transfection at 48h using the protocol provided by the manufacturer.

Chromatin Immunoprecipitation (ChIP) Assay

The chromatin immunoprecipitation (ChIPs) assay kit (Cell Signaling Technology, Danvers, MA, USA) was used for ChIP assay according to protocol provided by the manufacturer's instructions. In brief, the indicated cells were cultured on 100-mm culture dish around 70%~80% confluence and were fixed to cross-link proteins with DNA using 1% formaldehyde. The cell lysates were sonicated to shear DNA into small uniform fragments. Equal amounts of supernatants using anti-NKX2-8 (Abcam), or anti-HDAC1 (Abcam), or anti-H3K27ac (Abcam) and anti-IgG antibodies (Millipore, Billerica, MA) with protein G magnetic beads were immunoprecipitated overnight at 4°C. The cross-linked protein/DNA complexes were collected by magnetic pull down, and then were eluted from beads by elution buffer. PCR analysis with the indicated primers was conducted using the free DNA reversed from cross-linked protein/DNA complexes. The indicated ChIP primers are showed in **Supplementary Table 1**.

Chemical Reagents

3 types of PTHrP inhibitors, the PTHrP neutralizing antibody (T-4512) were purchased from Bachem (Torrance, CA, USA). 6-thioguanine (6-TG) was purchased from Sigma-Aldrich (St. Louis, MO, USA), and PTHrP₇₋₃₄ was purchased from GL Biochem (Shanghai, China).

Enzyme-Linked Immunosorbent Assay (ELISA)

The PTHrP level in the culture medium from breast cancer cells was measured using a Human PTHrP (Parathyroid hormone-related protein) ELISA Kit (EH1058, FineTest, Wuhan Fine Biotech Co., Ltd., Wuhan, China), and analyzed according to the manufacturer's instructions. The levels of transforming growth factor beta (TGF- β), RANKL, and osteoprotegerin (OPG) in culture medium were measured using mouse TGF- β ELISA Kit (ab119557), human RANKL ELISA Kit (ab213841), mouse RANKL ELISA Kit (ab100749), human OPG ELISA Kit (ab100617) and mouse OPG ELISA Kit (ab203365), respectively. The SpectraMax i3x Multi-Mode Microplate Reader (Molecular Devices, San Jose, CA, USA) was used to read data at 450 nm.

Patient Information

A total of 20 tumor-adjacent normal breast tissues and 344 paraffin-embedded breast cancer samples were performed in this study. All samples that were histopathologically and clinically diagnosed at the third Affiliated Hospital, Sun Yat-sen University Cancer Center, and the First Affiliated Hospital from 2005 to 2019. The protocols used in this study were approved by the Institutional Research Ethics Committee of Sun Yat-sen University for the use of these clinical materials for research purposes. All Patients' samples were obtained according to the Declaration of Helsinki and each patient signed a written informed consent for all the procedures.

Immunohistochemistry (IHC) Assay

IHC assay was used to measure the NKX2-8 protein level *via* anti-NKX2-8 antibody (1:100; Sigma-Aldrich Cat# AV31856), and

PTHrP protein level *via* anti-NKX2-8 antibody (1:100; LSBio Cat# LS-C31524-100) in 20 normal breast tissue and 344 breast cancer specimens, according previous report (12). Axio Imager.Z2 system (Carl Zeiss Co. Ltd., Jena, Germany) was used to capture the immunohistochemistry images. The degree of immunostaining of formalin-fixed, paraffin-embedded sections were reviewed and scored separately by two independent pathologists blinded to the histopathological features and patient data of the samples. The scores were determined by combining the proportion of positively-stained tumor cells and the intensity of staining. The scores given by the two independent pathologists were combined into a mean score for further comparative evaluation. Tumor cell proportions were scored as follows: 0, no positive tumor cells; 1, <10% positive tumor cells; 2, 10–35% positive tumor cells; 3, 35–75% positive tumor cells; 4, >75% positive tumor cells. The staining intensity was graded according to the following standard: 1, no staining; 2, weak staining (light yellow); 3, moderate staining (yellow brown); 4, strong staining (brown). The staining index (SI) was calculated as the product of the staining intensity score and the proportion of positive tumor cells. Using this method of assessment, we evaluated protein expression in normal breast tissues, breast cancer tissues and bone metastasis tissues by determining the SI, with possible scores of 0, 2, 3, 4, 6, 8, 9, 12, and 16. Samples with an SI ≥ 8 were determined as high expression and samples with an SI < 8 were determined as low expression. Cutoff values were determined on the basis of a measure of heterogeneity using the log-rank test with respect to overall survival.

CAPTURE System

CAPTURE system using the biotinylated deactivated Cas9 (dCas9) was performed according to a previous report (14). In brief, the genomic locus-associated protein in the breast cancer cells transfected with the CAPTURE system, including a FB-dCas9 and a biotin ligase BirA (purchased from Addgene, Watertown, MA, USA; 100547 and 100548), and target-specific single guide RNAs (targeting the promoter of PTHrP, listed in **Supplementary Table 1**), were isolated using streptavidin purification and then for mass spectrometry analysis.

Cell Growth Assay

A 3-(4, 5-Dimethylthiazol-2-yl)-2, 5-diphenyltetrazolium bromide (MTT) assay was employed to examine the cell growth. In brief, the indicated cells (8×10^3) were cultured in 96-well plate at the indicated time point, and 0.5 mg/mL MTT was added to each well for 4 h. Then removing MTT, adding dimethyl sulfoxide (DMSO), and mixing vigorously. The SpectraMax i3x Multi-Mode Microplate Reader (Molecular Devices) was used to measure the absorbance at 490nm.

Osteoclastogenesis Assay

Osteoclast precursor cells (1×10^5) were cultured on 24-well clusters containing glass coverslips (Thermo Fisher Scientific, Waltham, MA, USA) and grown in the conditioned media (CM), alone or treated with anti-PTHrP antibody (10 μ g/mL), or 6-TG (10 μ M), or PTHrP₇₋₃₄ (0.5 μ M). Media were changed at every other day. Osteoclasts were counted on day 6. The osteoclasts

cultured on plastic dishes were fixed with 4% paraformaldehyde/phosphate-buffered saline (PBS) and tartrate-resistant acid phosphatase (TRAP) in the cells was stained using a commercial kit (387A-1KT; Sigma-Aldrich). The TRAP-positive multinucleated cells that contain > 3 nuclei defined as osteoclasts.

Xenografted Tumor Models

The animal study was reviewed and approved by the Sun Yat-sen University Animal Care Committee. Bone-metastasis assay was performed using the indicated luciferase-expressing breast cancer cells (1×10^5) that were intracardially injected into nu/nu nude mice (5 weeks old). Anti-PTHrP antibody (Abcam) or anti-rabbit IgG were administered intraperitoneally twice a week. The 6-TG (1.0 mg/kg in 100 μ L of PBS), or PTHrP₇₋₃₄ (200 μ g/kg in μ L of PBS) was injected subcutaneously daily. The SIEMENS micro-computed tomography (μ CT) system (SIEMENS, Munich, Germany) was used to detect the osteolytic lesions in tibia and femur of hind limb. The bones of mice were harvested for further analysis.

In Vivo Quantification of Osteoclasts

Hind limbs were fixed in paraformaldehyde solution (4%), decalcified in 14.3% EDTA for 4 days at 37°C with daily changes of EDTA, and then embedded in paraffin wax. Sections were stained with hematoxylin and eosin using Mayer's hematoxylin solution, stained with TRAP (using a TRAP kit, 387A-1KT; Sigma-Aldrich) according to the manufacturer's protocols. The numbers of TRAP⁺ osteoclasts were determined on a 3 mm length of endocortical surface and viewed under an optical microscope (Olympus, DP72, Tokyo, Japan).

Statistics

All the data presented in this study were showed as the mean \pm standard deviation (SD) and n represents the number of independent experiments performed on different mice, or different batches of cells, or different clinical tissues. Statistical analysis was performed either the Student's two-tailed t-test or one-way analysis of variance (ANOVA). Bivariate correlations were calculated between study variables using Spearman's rank correlation coefficients. Survival curves were plotted using the Kaplan-Meier method was used to plot survival curves that were compared using the log-rank test. Univariate and multivariate Cox regression analyses were used to analyze the significance of various variables for survival. The P-values that less than were considered statistically significant. The GraphPad Prism 7 (GraphPad Inc., La Jolla, CA, USA) and SPSS 19.0 (IBM Corp., Armonk, NY, USA) statistical software were used for statistical analysis and P-values were represented as * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and N.S. means not significant ($P > 0.05$).

RESULTS

Reduced NKX2-8 Is Associated With Progression of Bone-Metastasis in Breast Cancer

SCP2 cell line, which is the bone-tropism cell line derived from MDA-MB-231 cells, is used to study the bone-tropism of breast

cancer metastasis (15). To screen the key factors involved in breast cancer bone metastasis, MS-based proteomics was performed in SPC2 and MDA-MB-231-parental cells. Analysis of protein profiling showed that a total of 34 proteins, including 18 elevated proteins and 16 reduced proteins, were dysregulated in SPC2 cells compared with MDA-MB-231-parental cells (Figure 1A and Supplementary Table 2). Among them, NKX2-8 levels were found to be significantly decreased in bone-metastatic tissues compared to normal breast tissues, or non-metastatic breast cancer tissues, respectively (Figures 1B, C, and Supplementary Figure 1). Furthermore, statistical analysis showed that patients with NKX2-8 high-expressing breast cancer had much longer bone-metastasis-free survival than the patients with NKX2-8 low-expressing breast cancer ($P = 0.011$ and Figure 1D). Taken together, our results indicate that the reduced NKX2-8 is linked to the development of bone-metastasis in breast cancer.

NKX2-8 Suppresses Bone Metastasis of Breast Cancer Cells

The biological role of NKX2-8 in breast cancer organ-specific metastasis was then examined using NKX2-8-silenced SCP-2 and NKX2-8-overexpressing MDA-MB-231 breast cancer cell lines (Figure 2A), and then injected intracardially the corresponding cells into nude mice. The NKX2-8-silenced breast cancer cells-injected mice exhibited earlier bone metastases, as revealed by μ CT analysis, and histology examination (Figures 2B, C). Consistently, compared with that in the control mice, the NKX2-8-overexpressing breast cancer cells displayed delayed bone metastases, and reduced bone metastasis lesions/osteolytic areas (Figures 2B, C). Histological TRAP staining showed that NKX2-8-overexpressing breast cancer cells significantly suppressed activation of osteoclasts (Figure 2C). Collectively, these results demonstrated the NKX2-8 specifically inhibits bone metastasis of breast cancer cells by reducing osteoclastogenesis.

Downregulation of NKX2-8-Induced PTHrP Promotes Osteoclastogenesis

Interestingly, similar with treatment with the conditioned media (CM) from control cells, there was no effect on osteoclastogenesis when osteoclasts were treated directly with that of NKX2-8-silenced cells (Figure 3A). Nevertheless, CM from osteoblasts was found to significantly increase the TRAP-positive multinuclear osteoclasts number and enhance the TRAP enzymatic activity, when osteoblasts were pretreated with CM from NKX2-8-silenced cells (Figure 3A). These results suggested that CM from NKX2-8-silenced cells indirectly promotes osteoclastogenesis. Considering that the RANKL/OPG axis is vital for osteoclastogenesis, we next examined the RANKL/OPG ratio in osteoblasts under induction by CM from breast cancer cells. The relative RANKL/OPG ratio was dramatically elevated after treatment with CM from NKX2-8-silenced cells, but was decreased after treatment with CM from NKX2-8-overexpressing cells (Figure 3B). Consistently, the expression of differentiation marker and activation marker of osteoclasts, such as Acp5, Ctsk, Nfat-c1, C-fos, and Dc-stamp, were significantly upregulated in osteoclasts treated with CM from

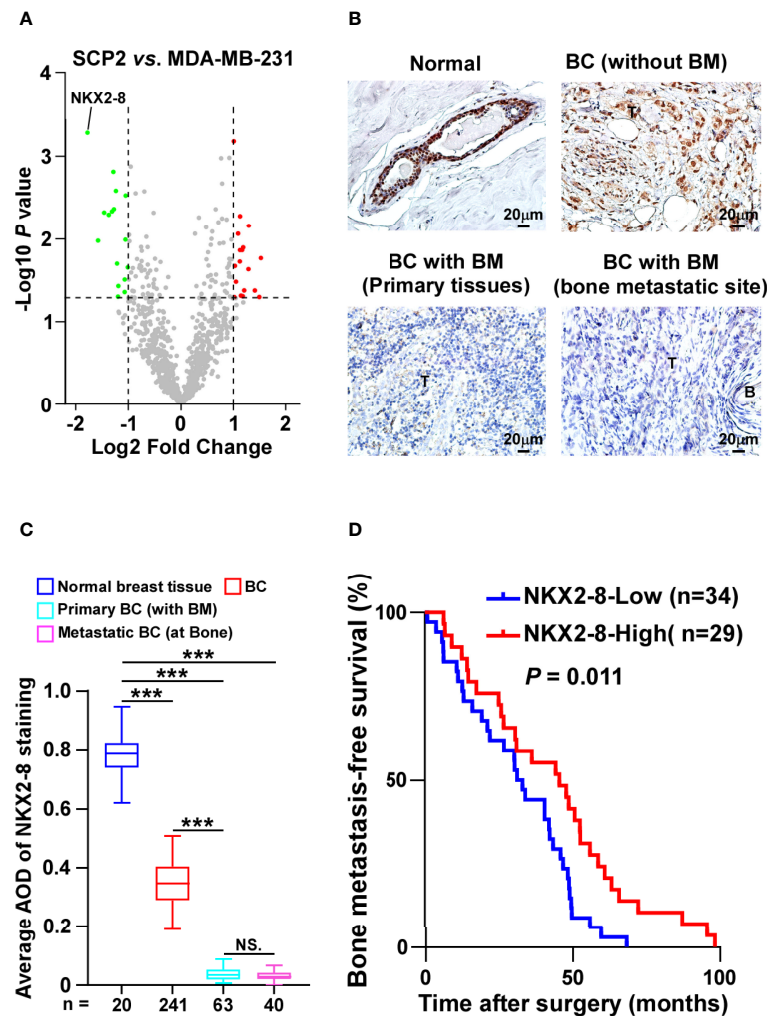


FIGURE 1 | NKX2-8 is associated with progression of bone-metastasis in breast cancer. **(A)** The upregulated and downregulated proteins in SCP2 cells compared with MDA-MB-231 cells were analyzed using Volcano plot. **(B, C)** Representative IHC images **(B)** and quantification **(C)** of NKX2-8 level in normal breast tissue (n = 20), primary non-bone metastatic breast cancer tissues (n = 241), primary bone-metastatic breast cancer tissues (n = 63), and bone-metastatic breast cancer tissues (n = 40). Scale bar, 50 μm . **(D)** Kaplan-Meier analysis of bone metastasis-free survival curves in patients with NKX2-8 high- vs. low-expressed breast cancer with bone metastasis (n = 63; P = 0.011, log-rank test). *** means P < 0.001, N.S. means not significant (P > 0.05).

osteoblasts that pretreated with CM from NKX2-8-silenced breast cancer cells, but decreased in response to overexpression of NKX2-8 (**Supplementary Figure 2**). These results indicated that the NKX2-8 silencing induced-secretome promotes osteoclastogenesis *via* osteoblasts-secreted RANKL.

Among the osteoclastogenesis regulators that have been reported previously (16–19), PTHrP was one of the most upregulated secreted proteins in NKX2-8-silenced cells, but was downregulated in NKX2-8-overexpressing cells (**Figure 3C**). Consistently, the secreted PTHrP protein levels were also drastically decreased in NKX2-8-overexpressed cells, but elevated in NKX2-8-silenced cells, (**Figure 3D**), suggesting the potential role of PTHrP in breast cancer bone-metastasis.

Silencing PTHrP abrogated NKX2-8-silenced induction of RANKL secretion, the TRAP-positive multinuclear osteoclasts

number, the enzymatic activity of TRAP (**Figures 3E, F**), which demonstrated that NKX2-8 silencing induced-PTHrP promotes osteoclastogenesis *via* osteoblasts-secreted RANKL. Moreover, the NKX2-8 silencing-promoted vicious cycle was significantly blocked by silencing PTHrP, as indicated by downregulation of bone matrix-released TGF- β and reduced growth rates of breast cancer cells (**Figure 3G**). Taken together, our results strongly indicated a critical role of the NKX2-8/PTHrP/RANKL axis in the regulation of osteoclastogenesis *in vivo*, which resulted in a vicious cycle between tumor cells and osteoclasts (**Figure 3H**).

NKX2-8 Transcriptionally Represses the Expression of PTHrP

Next, the role of NKX2-8 reduction in PTHrP expression were examined. As shown in **Figure 4A**, the NKX2-8 ChIP assay

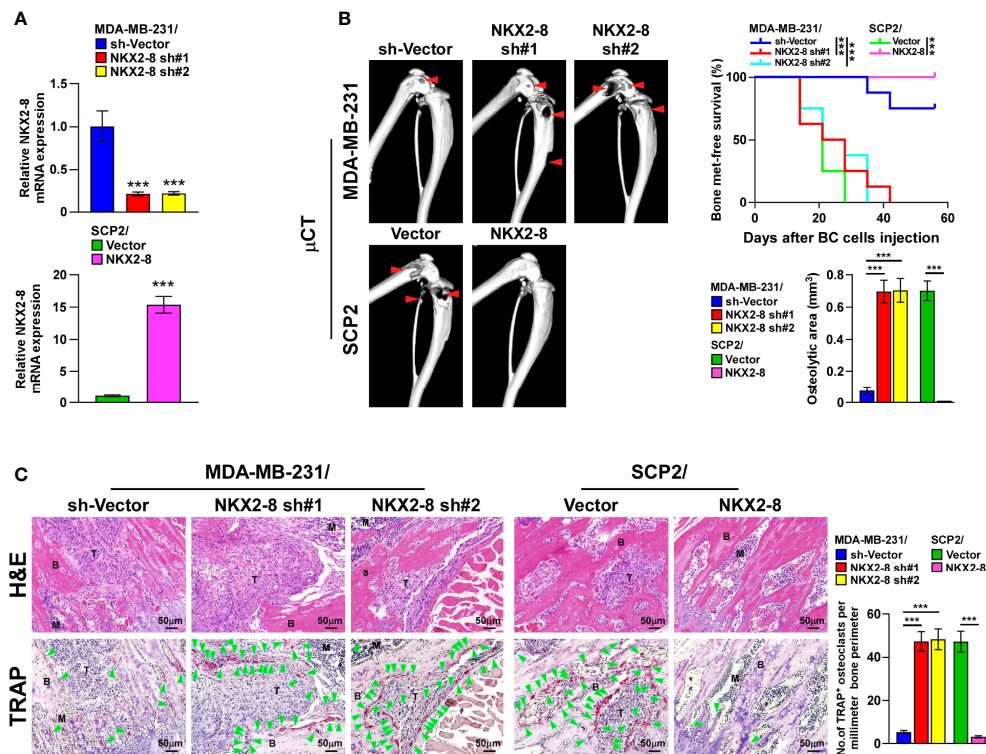


FIGURE 2 | NKX2-8 silencing promotes bone metastasis in breast cancer. **(A)** Real-time PCR analysis of NKX2-8 expression in sh-vector- and NKX2-8-shRNA(s)-transduced SCP2 cells and in vector- and NKX2-8-transduced MDA-MB-231 cells. GAPDH served as the loading control. **(B)** Left: μ CT images of bone lesions from representative mice. Right: Quantification of the μ CT osteolytic lesion area, and Kaplan-Meier bone metastasis-free survival curve of the indicated mice groups in experimental metastasis phase ($n = 8$ /group). **(C)** Histological H&E images (left-upper) and TRAP images (left-lower), and quantification (right) of the osteolytic area and TRAP-positive osteoclasts as shown in the indicated mice ($n = 8$ /group). Scale bar, 50 μ m. BC, Breast Cancer; BM, Bone Metastasis. * means $P < 0.05$, *** means $P < 0.001$.

results suggested that NKX2-8 was most significantly associated with the PTHrP promoter in MDA-MB-231 cells. Furthermore, the NKX2-8-silenced cells displayed increased, but NKX2-8-overexpressing cell exhibited decreased, the luciferase activities of genes with a NKX2-8-specific binding site (NBS) (Figure 4B). Whereas, we did not observe the effect on the luciferase activities of the NKX2-8 NBS-deleted promoter (Figure 4B). The reverse correlation adverse of NKX2-8 with PTHrP expression was also demonstrated using IHC analysis in breast cancer tissues, which NKX2-8 expression was adversely associated with the PTHrP level (Figure 4C). Importantly, no correlation of NKX2-8 and PTHrP was found in the breast cancer tissues with other organs metastasis (Supplementary Figure 3). Real-time PCR analysis revealed that the PTHrP expression was downregulated in the NKX2-8 high-expressing breast but upregulated in the NKX2-8 low-expressing (Figure 4D). Therefore, these results provided further evidence that NKX2-8 transcriptional downregulates PTHrP expression *via* directly targeting on its promoters.

The NKX2-8/HDAC1 Complex Is Involved in NKX2-8-Inhibited PTHrP Expression

In order to investigate the molecular mechanism in which NKX2-8-mediated transcriptional suppression of PTHrP,

we performed the biotinylated deactivated Cas9 (dCas9) capture analysis and found that both NKX2-8 and HDAC1 were identified to target on the PTHrP promoter in MDA-MB-231 cells (Figure 5A and Supplementary Table 3). Furthermore, ChIP assays revealed that downregulation of NKX2-8 significantly reduced, whereas upregulation of NKX2-8 enriched, the HDAC1 level on the PTHrP promoter in breast cancer cells (Figure 5B). Conversely, silencing NKX2-8 drastically increased, but overexpressing NKX2-8 decreased, the enrichment of H3K27 acetylation (H3K27ac) on the promoter of PTHrP gene in breast cancer cells (Figure 5C), which demonstrated that NKX2-8 inhibited PTHrP expression *via* HDAC1 to reduce H3K27ac on the promoter of PTHrP. Meanwhile, silencing HDAC1 significantly increased the PTHrP expression in NKX2-8 overexpression cells (Figure 5D). These results suggested that HDAC1 played an important role in NKX2-8-mediated PTHrP expression. In line with this hypothesis, we found that knockdown of HDAC1 has no impact on the enrichment of NKX2-8, but significantly increased the level of H3K27ac, on the PTHrP promoter (Figures 5E–G), which further supported the notion that NKX2-8 inhibits PTHrP expression *via* HDAC1 to reduce H3K27ac on the promoter of PTHrP.

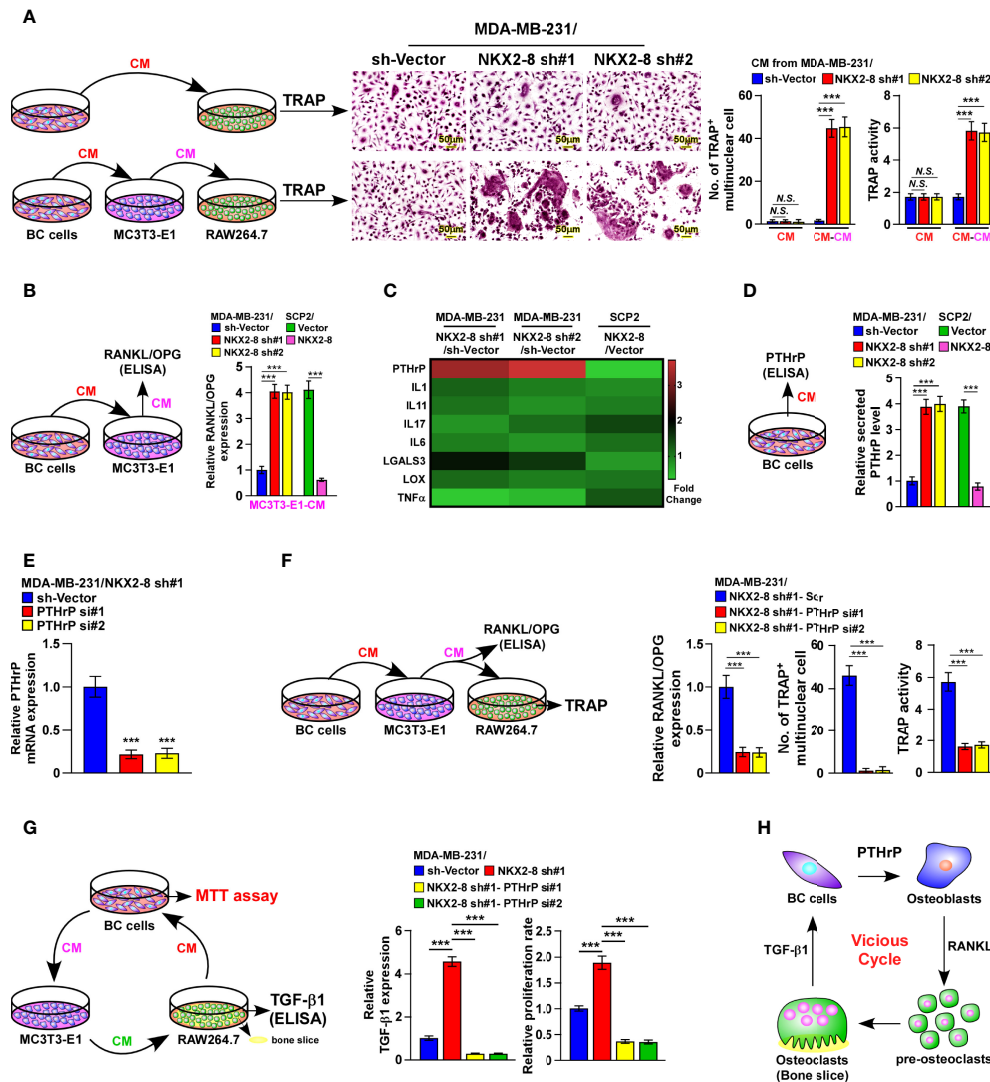


FIGURE 3 | NKX2-8 silencing-induced PTHrP promotes osteoclastogenesis *in vitro*. **(A)** Left: schematic illustration of osteoclastogenesis under treatment with CM from breast cancer cells or from osteoblasts pretreated with CM from breast cancer cells. Middle: osteoclast differentiation assays assessed using TRAP staining cultured CM obtained from the indicated cells. Right: Quantification of the number of TRAP-positive multinuclear osteoclasts and TRAP activity. **(B)** ELISA analysis of the RANKL/OPG ratio in CM from osteoblast precursor cells MC3T3-E1 cells cultured in the CM obtained from the indicated breast cancer cells. **(C)** Real-time PCR analysis indicating osteoclastogenesis regulator expression in the indicated cells. The pseudocolor represents the intensity scale of genes in the indicated cells generated by a log2 transformation. GAPDH serves as the loading control. **(D)** ELISA analysis of secreted PTHrP levels in CM from the indicated cells. **(E)** Real-time PCR analysis of PTHrP levels in sh-vector- and PTHrP-shRNA(s)-transduced MDA-MB-231/NKX2-8 sh#1 cells. GAPDH served as the loading control. **(F)** Left: schematic illustration of osteoclastogenesis under treatment with CM from osteoblasts pretreated with CM from breast cancer cells. Right: ELISA analysis of the RANKL/OPG ratio in CM from osteoblast precursor cells MC3T3-E1 cells in the presence of CM from the indicated breast cancer cells, and quantification of the TRAP-positive multinuclear osteoclasts number and TRAP activity cultured in the CM obtained from the indicated cells. **(G)** Left: schematic illustration of breast cancer CM-induced "vicious cycle" between the indicated cells. Middle: The TGF- β 1 levels analyzed using ELISA assay in CM from RAW 264.7 cells cultured onto the bone slice in CM obtained from the breast cancer cells. Right: MTT assay analysis of proliferation rate of indicated cells from experiment in left panel. **(H)** Schematic illustration of PTHrP-induced "vicious cycle" between the indicated cells. BC: Breast Cancer. N.S. means not significant ($P > 0.05$), *** means $P < 0.001$.

Consistently, silencing HDAC1 in the breast cancer cells significantly abolished the reductive effect of NKX2-8 on the formation of TRAP⁺-multinuclear osteoclasts and TRAP enzymatic activity (Figures 6A, B). Taken together, these results suggest that the NKX2-8/HDAC1 repressor complex is involved in NKX2-8-inhibited PTHrP expression in breast cancer cells.

Targeting PTHrP Inhibit NKX2-8 Silencing-Induced Osteoclastogenesis and Bone Metastasis

To determine the critical effect of tumor-derived PTHrP on NKX2-8 silencing-mediated osteoclastogenesis, the osteoclast precursor cells was treated with three types of PTHrP

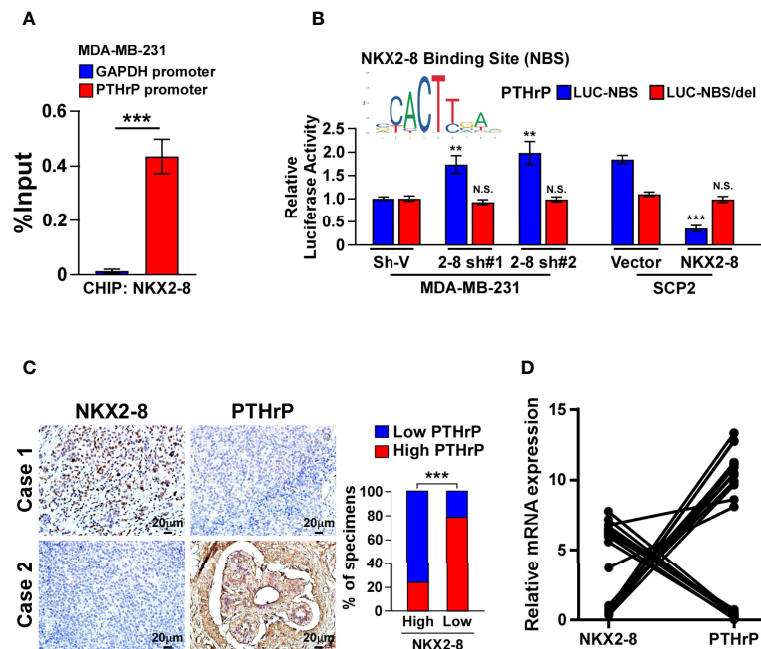


FIGURE 4 | NKX2-8 transcriptionally represses PTHrP. **(A)** ChIP analysis of the enrichment of NKX2-8 on promoter of PTHrP gene in MDA-MB-231 cells. **(B)** Relative luciferase activities of the PTHrP promoter in the indicated breast cancer cells. **(C)** NKX2-8 levels were negatively associated with PTHrP expression in breast cancer tissues (n = 304). The representative IHC staining images (left) and expression correlation (right) of NKX2-8 and PTHrP protein in breast cancer tissues. Scale bars, 20 μ m. **(D)** Relative expression of NKX2-8 and PTHrP in primary breast cancer tissues (n = 20) with bone metastasis (n = 10; $P < 0.001$). ** means $P < 0.01$, *** means $P < 0.001$ N.S. means not significant ($P > 0.05$).

inhibitors, including the peptide antagonist PTHrP₇₋₃₄, the neutralizing antibody and the chemical inhibitor 6-thioguanine (6-TG), under conditions of CM from NKX2-8-silenced cells. Osteoclastogenesis analyses indicated that all three type of inhibitors efficiently impaired the osteoclastogenesis effects of the CM from NKX2-8-silenced cells, as PTHrP inhibition completely abolishing the induced effects of NKX2-8 silencing on the TRAP-positive multinuclear osteoclasts and TRAP enzymatic activity (**Figures 7A–C**). These results further supported the notion that NKX2-8 inhibited osteoclastogenesis *via* PTHrP *in vitro*. Moreover, we found that treatment with all three PTHrP inhibitor exhibited dramatic blocked effect on vicious cycle induced by silencing NKX2-8, as indicated by decreased level of bone matrix-released TGF- β 1 and slower growth rates of breast cancer cells (**Supplementary Figure 4**). Therefore, our results indicate that targeting PTHrP contributes to inhibition of the osteoclastogenesis and vicious cycle induced by NKX2-8 silencing.

We injected NKX2-8 silenced MDA-MB-231 cells intracardially into nude mice and treated them with PTHrP₇₋₃₄ or anti-PTHrP antibody or 6-TG, which markedly suppressed the bone metastasis signals and bone destruction, as monitored by μ CT analysis (**Figure 7D**). Consistently, compared with that in the control mice, the mice treated with PTHrP₇₋₃₄ or anti-PTHrP antibody or 6-TG displayed no bone metastases and bone

metastasis lesions/osteolytic areas (**Figures 7E, F**). Collectively, our data demonstrate that targeting PTHrP inhibited NKX2-8 silencing-induced osteoclastogenesis and bone metastasis.

DISCUSSION

As the most prevalent form of metastasis, bone metastasis occurs over 70% breast cancer patients with advanced diseases (20). It can cause bone pain, osteoporosis, pathological fractures, and other skeletal-related events, which significantly reduce the patient's quality of life, and can even be lethal (4, 21). The outcome of breast cancer with bone metastasis is extremely poor, which a 1-year survival rate of breast cancer patient is only 40–59% (22). Although current therapeutic modalities, including chemoradiotherapy and anti-osteolytic drugs, have made significant improvement on reduce the incidence rate associated with bone metastasis, these treatments only provide minimal benefit to patients' survival (4). Therefore, developing novel effective therapeutic strategies to treat patients with bone metastasis of breast cancer is urgent. In this study, we observed that NKX2-8-silencing-induced PTHrP protein plays a vital role in bone metastasis of breast cancer by promoting osteoclastogenesis. Importantly, targeting PTHrP significantly decreased osteoclastogenesis and effectively suppressed the

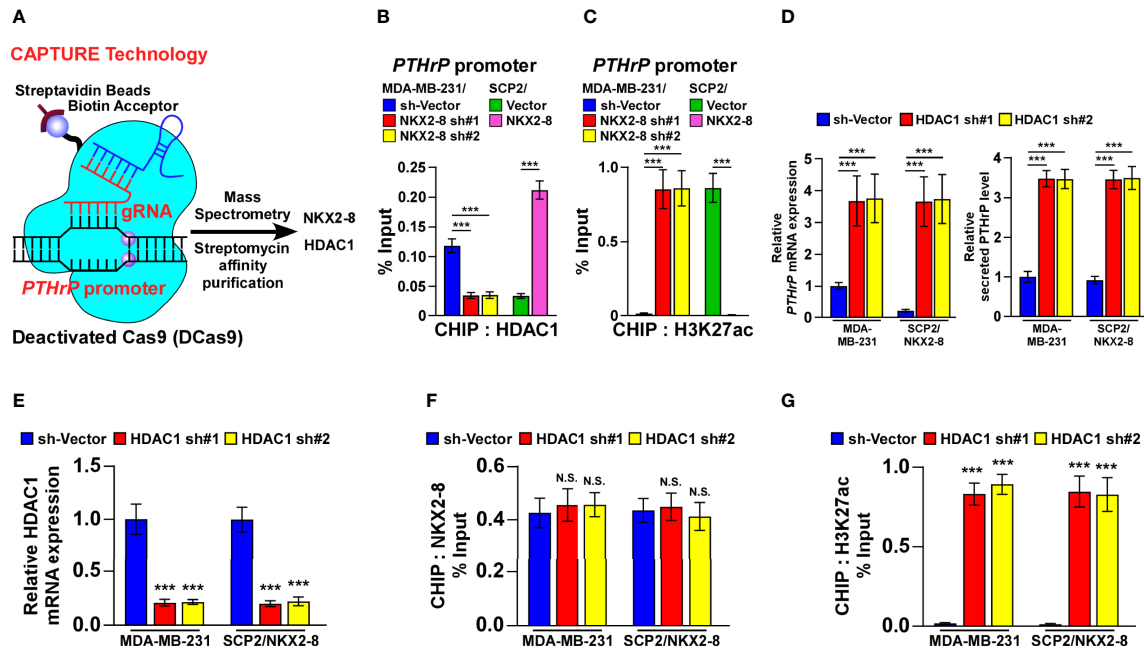


FIGURE 5 | NKX2-8 interacts with HDAC1 in NKX2-8-inhibited PTHrP transcription. **(A)** Schematic diagram of interaction between NKX2-8 and the HDAC1 in the PTHrP promoter using CAPTURE approach. **(B)** ChIP analysis of the enrichment of HDAC1 in the PTHrP promoter in the indicated breast cancer cells. **(C)** ChIP analysis of the enrichment of H3K27ac in the PTHrP promoter in the indicated breast cancer cells. **(D)** Real-time PCR analysis of PTHrP mRNA (left) and ELISA analysis of serum PTHrP (right) expression in the indicated cells. GAPDH was used as the loading control. **(E)** The relative expression of HDAC1 in the indicated cells using real-time PCR analysis. GAPDH served as the loading control. **(F)** ChIP analysis of the association of NKX2-8 with the promoter of PTHrP gene in the indicated breast cancer cells. **(G)** The relative enrichment of H3K27ac on the PTHrP promoter determined by ChIP analysis in the indicated breast cancer cells. N.S. means not significant ($P > 0.05$), *** means $P < 0.001$.

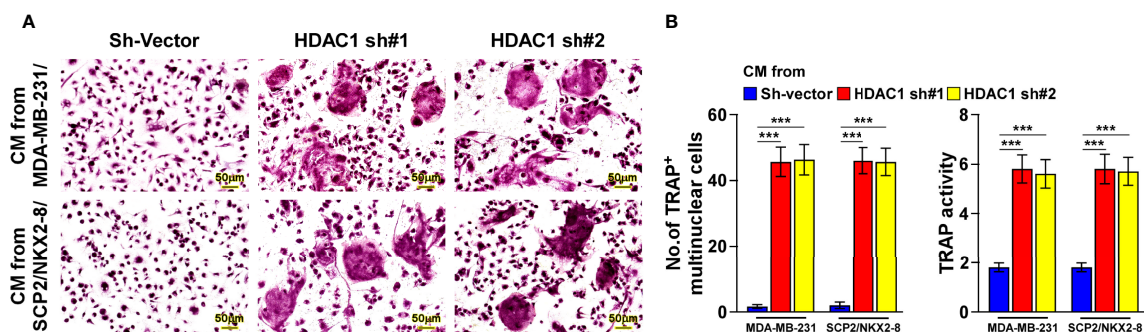


FIGURE 6 | NKX2-8 represses osteoclastogenesis via HDAC1. **(A)** TRAP staining images of the indicated CM-treated pre-osteoclasts. Scale bar, 20μm or 50μm. **(B)** The number quantification of TRAP-positive multinuclear cells and examination of TRAP activity. *** means $P < 0.001$.

progression of breast cancer bone metastasis. Therefore, these findings shed light on a potential mechanism underlying breast cancer bone metastasis and might represent a potential clinical strategy for treatment of breast cancer bone metastasis.

Breast cancer cells have the ability of “organotropic metastasis”, i.e., preferential metastasis to specific organs, which is regulated by subtypes of breast cancer, in which the

tumor-induced pre-metastatic niche plays a vital role in engrafting and surviving of metastatic cells (23). It has been reported that only a very small fraction of the breast cancer cells possessing the ability to form highly aggressive, osteolytic bone metastases. Thus, we used SCP2 cells and MDA-MB-231, a derivative bone-tropism cell line that was isolated by Yibin Kang et al. (15). We found that NKX2-8 levels were markedly

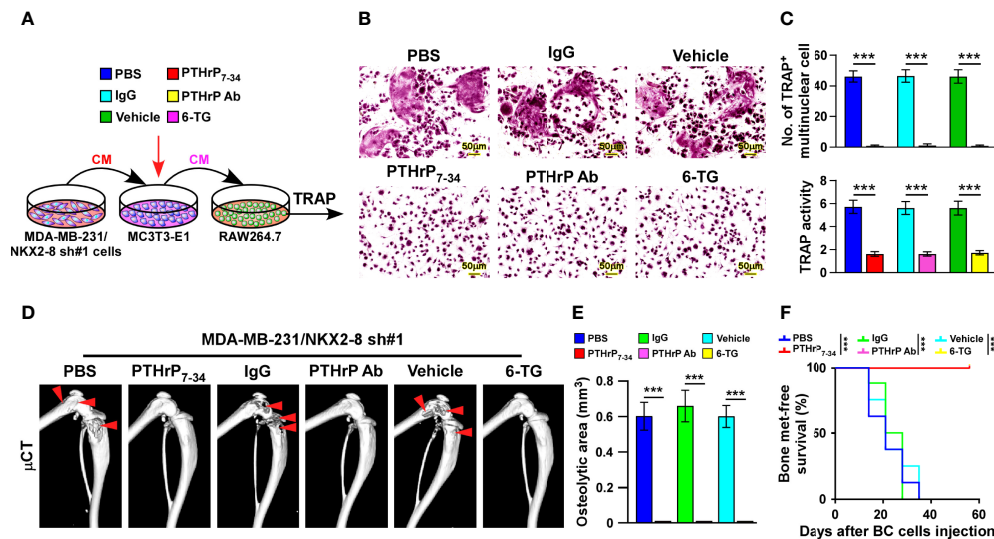


FIGURE 7 | Targeting PTHrP inhibits NKX2-8 silencing-induced bone metastasis. **(A)** Schematic illustration of osteoclastogenesis in the indicated condition. **(B)** Osteoclast differentiation assays analyzed as shown by TRAP staining of the indicated CM-treated pre-osteoclasts with addition of PBS or PTHrP₇₋₃₄, IgG or anti-PTHrP antibody, or vehicle or 6-TG. **(C)** Quantification of the number of TRAP-positive multinuclear cells (upper) and examination of TRAP activity (lower). **(D)** μ CT images of bone lesions from mice injected with MDA-MB-231/NKX2-8 sh#1, and treated with PBS or PTHrP₇₋₃₄ or IgG or anti-PTHrP antibody or vehicle or 6-TG. **(E, F)** Normalized BLI signals of bone metastases quantification of the μ CT osteolytic lesion area **(E)**, and Kaplan-Meier bone metastasis-free survival curve **(F)** of the indicated mice in the experimental metastasis phase ($n = 8/\text{group}$). *** means $P < 0.001$.

decreased in SCP2 cells compared with those in MDA-MB-231 parental cells. Subsequent experiments revealed an organ-specific correlation between NKX2-8 expression and bone metastasis. Importantly, our findings demonstrated that silencing NKX2-8 in breast cancer cells influenced their ability to affect the pre-metastatic niche. Silencing of NKX2-8 in breast cancer cells activated PTHrP transcription. Consequently, PTHrP derived from NKX2-8 silenced-breast cancer cells contribute to formation of pre-bone metastatic niche through alteration of RANKL/OPG ratio *via* acting on osteoblasts, resulting in instigating osteoclastogenesis that led to metastatic bone destruction.

NKX2-8 is a transcription factor that participates in the progression or chemoresistance in multiple tumors (7–13). Our results showed that NKX2-8 plays a critical role in breast cancer bone metastasis. Specifically, NKX2-8 interacts with HDAC1 to form a complex that suppresses PTHrP transcription in breast cancer cells with no bone metastasis, while silencing NKX2-8 could eliminate its inhibition of bone metastasis. Our results provided a new paradigm for NKX2-8 in cancer, especially in bone metastasis, in which it was involved in reshaping the bone microenvironment for and enhance metastasis.

Overall, the present data provided in current study suggest a potential therapeutic application of PTHrP inhibitors for treatment and prevention of downregulation of NKX2-8-induced breast cancer bone metastasis.

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.

ETHICS STATEMENT

The animal study was reviewed and approved by Sun Yat-sen University Animal Care Committee.

AUTHOR CONTRIBUTIONS

AA designed the experiments and analyzed data. YH performed *in vitro* cell studies. SC performed the xenograft tumor experiments. WQ performed staining, immunohistochemical and pathological analysis. XL performed the CHIP, real time PCR, immunoprecipitation and western blot. YX analyzed mass spectrometry data. LS, SZ, and JL supervised the whole study and wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Natural Science Foundation of China (No. 82030078, 81830082, 82072609, 81621004 and 82003128).

ACKNOWLEDGMENTS

We would like to thank Prof. Guohong Hu of the Chinese Academy of Sciences and Shanghai Jiao Tong University

School of Medicine for kindly providing the SCP2 cell lines. We would like to thank the native English speaking scientists of Elixigen Company (Huntington Beach, California) for editing our manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.907000/full#supplementary-material>

REFERENCES

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660
- Hess KR, Varadhachary GR, Taylor SH, Wei W, Raber MN, Lenzi R, et al. Metastatic Patterns in Adenocarcinoma. *Cancer* (2006) 106(7):1624–33. doi: 10.1002/cncr.21778
- Mundy GR. Metastasis to Bone: Causes, Consequences and Therapeutic Opportunities. *Nat Rev Cancer* (2002) 2(8):584–93. doi: 10.1038/nrc867
- Weilbaecher KN, Guise TA, McCauley LK. Cancer to Bone: A Fatal Attraction. *Nat Rev Cancer* (2011) 11(6):411–25. doi: 10.1038/nrc3055
- Ell B, Kang Y. Snapshot: Bone Metastasis. *Cell* (2012) 151(3):690–e1. doi: 10.1016/j.cell.2012.10.005
- Thomas RJ, Guise TA, Yin JJ, Elliott J, Horwood NJ, Martin TJ, et al. Breast Cancer Cells Interact With Osteoblasts to Support Osteoclast Formation. *Endocrinology* (1999) 140(10):4451–8. doi: 10.1210/endo.140.10.7037
- Kajiyama Y, Tian J, Locker J. Regulation of Alpha-Fetoprotein Expression by Nkx2.8. *Mol Cell Biol* (2002) 22(17):6122–30. doi: 10.1128/MCB.22.17.6122-6130.2002
- Hagihara A, Miyamoto K, Furuta J, Hiraoka N, Wakazono K, Seki S, et al. Identification of 27 5' CpG Islands Aberrantly Methylated and 13 Genes Silenced in Human Pancreatic Cancers. *Oncogene* (2004) 23(53):8705–10. doi: 10.1038/sj.onc.1207783
- Harris T, Pan Q, Sironi J, Lutz D, Tian J, Sapkar J, et al. Both Gene Amplification and Allelic Loss Occur at 14q13.3 in Lung Cancer. *Clin Cancer Res* (2011) 17(4):690–9. doi: 10.1158/1078-0432.CCR-10-1892
- Yu C, Zhang Z, Liao W, Zhao X, Liu L, Wu Y, et al. The Tumor-Suppressor Gene Nkx2.8 Suppresses Bladder Cancer Proliferation Through Upregulation of Foxo3a and Inhibition of the Mek/Erk Signaling Pathway. *Carcinogenesis* (2012) 33(3):678–86. doi: 10.1093/carcin/bgr321
- Lin C, Song L, Gong H, Liu A, Lin X, Wu J, et al. Nkx2-8 Downregulation Promotes Angiogenesis and Activates Nf-Kappab in Esophageal Cancer. *Cancer Res* (2013) 73(12):3638–48. doi: 10.1158/0008-5472.CAN-12-4028
- Qu L, Deng B, Zeng Y, Cao Z. Decreased Expression of the Nkx2.8 Gene Correlates With Tumor Progression and a Poor Prognosis in Hcc Cancer. *Cancer Cell Int* (2014) 14:28. doi: 10.1186/1475-2867-14-28
- Zhu J, Wu G, Song L, Cao L, Tan Z, Tang M, et al. Nkx2-8 Deletion-Induced Reprogramming of Fatty Acid Metabolism Confers Chemoresistance in Epithelial Ovarian Cancer. *EBioMedicine* (2019) 43:238–52. doi: 10.1016/j.ebiom.2019.04.041
- Liu X, Zhang Y, Chen Y, Li M, Zhou F, Li K, et al. In Situ Capture of Chromatin Interactions by Biotinylated Dcas9. *Cell* (2017) 170(5):1028–43.e19. doi: 10.1016/j.cell.2017.08.003
- Kang Y, Siegel PM, Shu W, Drobnjak M, Kakonen SM, Cordon-Cardo C, et al. A Multigenic Program Mediating Breast Cancer Metastasis to Bone. *Cancer Cell* (2003) 3(6):537–49. doi: 10.1016/s1535-6108(03)00132-6
- Jones DH, Kong YY, Penninger JM. Role of Rankl and Rank in Bone Loss and Arthritis. *Ann Rheum Dis* (2002) 61 (Suppl 2):ii32–9. doi: 10.1136/ard.61.suppl_2.ii32
- Zhang S, Xu Y, Xie C, Ren L, Wu G, Yang M, et al. Rnf219/Alpha-Catenin/Lgals3 Axis Promotes Hepatocellular Carcinoma Bone Metastasis and Associated Skeletal Complications. *Adv Sci (Weinh)* (2021) 8(4):2001961. doi: 10.1002/adv.202001961
- Nakajima K, Kho DH, Yanagawa T, Harazono Y, Hogan V, Chen W, et al. Galectin-3 Cleavage Alters Bone Remodeling: Different Outcomes in Breast and Prostate Cancer Skeletal Metastasis. *Cancer Res* (2016) 76(6):1391–402. doi: 10.1158/0008-5472.CAN-15-1793
- Cox TR, Rumney RMH, Schoof EM, Perryman L, Hoye AM, Agrawal A, et al. The Hypoxic Cancer Secretome Induces Pre-Metastatic Bone Lesions Through Lysyl Oxidase. *Nature* (2015) 522(7554):106–10. doi: 10.1038/nature14492
- Coleman RE. Metastatic Bone Disease: Clinical Features, Pathophysiology and Treatment Strategies. *Cancer Treat Rev* (2001) 27(3):165–76. doi: 10.1053/ctrv.2000.0210
- Coleman RE. Clinical Features of Metastatic Bone Disease and Risk of Skeletal Morbidity. *Clin Cancer Res* (2006) 12(20 Pt 2):6243s–9s. doi: 10.1158/1078-0432.CCR-06-0931
- Cetin K, Christiansen CF, Svaerke C, Jacobsen JB, Sorensen HT. Survival in Patients With Breast Cancer With Bone Metastasis: A Danish Population-Based Cohort Study on the Prognostic Impact of Initial Stage of Disease at Breast Cancer Diagnosis and Length of the Bone Metastasis-Free Interval. *BMJ Open* (2015) 5(4):e007702. doi: 10.1136/bmjopen-2015-007702
- Chen W, Hoffmann AD, Liu H, Liu X. Organotropism: New Insights Into Molecular Mechanisms of Breast Cancer Metastasis. *NPJ Precis Oncol* (2018) 2(1):4. doi: 10.1038/s41698-018-0047-0

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.

Copyright © 2022 Abudourousuli, Chen, Hu, Qian, Liao, Xu, Song, Zhang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Novel Secreted Protein-Related Gene Signature Predicts Overall Survival and Is Associated With Tumor Immunity in Patients With Lung Adenocarcinoma

Shuaijun Chen^{1†}, Jun Zhang^{2†}, Qian Li¹, Lingyan Xiao³, Xiao Feng¹, Qian Niu⁴, Liqin Zhao⁴, Wanli Ma^{4,5*} and Hong Ye^{1,5*}

OPEN ACCESS

Edited by:

Yuanrong Wang,
People's Liberation Army General
Hospital, China

Reviewed by:

Congkuan Song,
Wuhan University, China
Tianhao Li,
Second Affiliated Hospital of Hainan
Medical University, China

*Correspondence:

Wanli Ma
whmawl@aliyun.com
Hong Ye
dr_hong_ye@163.com

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

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 08 February 2022

Accepted: 09 May 2022

Published: 03 June 2022

Citation:

Chen S, Zhang J, Li Q, Xiao L, Feng X,
Niu Q, Zhao L, Ma W and Ye H (2022)
A Novel Secreted Protein-Related
Gene Signature Predicts Overall
Survival and Is Associated With
Tumor Immunity in Patients With
Lung Adenocarcinoma.
Front. Oncol. 12:870328.
doi: 10.3389/fonc.2022.870328

¹ Department of Pathophysiology, School of Basic Medicine, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, ² Department of Obstetrics and Gynecology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, ³ Department of Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, ⁴ Department of Respiratory and Critical Care Medicine, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, ⁵ Key Laboratory of Respiratory Diseases, National Health Commission of China, Wuhan, China

Secreted proteins are important proteins in the human proteome, accounting for approximately one-tenth of the proteome. However, the prognostic value of secreted protein-related genes has not been comprehensively explored in lung adenocarcinoma (LUAD). In this study, we screened 379 differentially expressed secretory protein genes (DESPRGs) by analyzing the expression profile in patients with LUAD from The Cancer Genome Atlas database. Following univariate Cox regression and least absolute shrinkage and selection operator method regression analysis, 9 prognostic SPRGs were selected to develop secreted protein-related risk score (SPRrisk), including CLEC3B, C1QTNF6, TCN1, F2, FETUB, IGFBP1, ANGPTL4, IFNE, and CCL20. The prediction accuracy of the prognostic models was determined by Kaplan–Meier survival curve analysis and receiver operating characteristic curve analysis. Moreover, a nomogram with improved accuracy for predicting overall survival was established based on independent prognostic factors (SPRrisk and clinical stage). The DESPRGs were validated by quantitative real-time PCR and enzyme-linked immunosorbent assay by using our clinical samples and datasets. Our results demonstrated that SPRrisk can accurately predict the prognosis of patients with LUAD. Patients with a higher risk had lower immune, stromal, and ESTIMATE scores and higher tumor purity. A higher SPRrisk was also negatively associated with the abundance of CD8⁺ T cells and M1 macrophages. In addition, several genes of the human leukocyte antigen family and immune checkpoints were expressed in low levels in the high-SPRrisk group. Our results provided some insights into assessing individual prognosis and choosing personalized treatment modalities.

Keywords: lung adenocarcinoma, secretome, secreted protein-related risk score, gene signature, immune landscape

INTRODUCTION

Lung cancer is the main cause of cancer-related deaths worldwide and accounts for approximately one-quarter of all cancer-related deaths, 82% of which are directly caused by cigarette smoking (1). Lung adenocarcinoma (LUAD) is the most common histological type, accounting for nearly 40% of all lung cancer cases (2). Despite intensive research and the development of several new targeted agents and immunotherapies, the survival rates for patients with LUAD remain dismal. The 5-year survival rate of lung adenocarcinoma is only 4–17% (3). Over 60% of patients with lung cancer are not diagnosed until the late stages of the disease (4). Therefore, early detection and personalized treatment may significantly improve patient survival.

Nowadays, with the rapid development of high-throughput sequencing technologies, many bioinformatics studies aim to identify biomarkers that can establish prognosis or predict drug response in patients with cancer (5, 6). Despite these advances, several critical limitations remain to be addressed. First, it is difficult to obtain tumor tissue samples. Second, performing transcriptome sequencing is expensive. These issues limit the use of combined gene signature models on a larger scale. Moreover, the available tumor tissues are usually from the intermediate or advanced stages of tumor progression, indicating missed opportunities for early detection and clinical intervention. Therefore, it is necessary to find an easy and attractive method to evaluate the prognosis of patients with LUAD.

In recent years, a large-scale, high-throughput protein expression, purification, and screening platform has been developed, establishing a secreted protein library (7). Over 2,000 human genes have been reported to encode known secreted proteins, including hormones, cytokines, proteases, antibodies, poison, and growth factors (8, 9). The proteins were classified into three major categories: (i) blood proteins, (ii) locally secreted proteins, and (iii) intracellular proteins (10). These proteins play important physiological roles in various biological processes, such as cell signal transduction, adhesion, migration, and immune defense (11–13). Meanwhile, several secreted proteins in diseases might serve as early-stage diagnostic and prognostic markers. These secreted proteins are also considered as new therapeutic agents or as targets for small molecule or antibody drug development (14)—for example, anterior gradient-2 (AGR2) is a secreted protein reported to be highly expressed in a variety of tumor types. Thus, AGR2 is related to the proliferation, metastasis, invasion, and drug resistance of tumor cells, making it an attractive target for early diagnosis and tumor therapy (15–17). Additionally, IL-6 plays a critical role in chronic inflammation, autoimmune diseases, infectious diseases, metabolic diseases, and cancer, and thus the IL-6 cytokine family has been used as a diagnostic or prognostic indicator of disease activity and response to therapy (18–21). Moreover, the IL-6 family of cytokines is now regarded as a major therapeutic target for clinical interventions (18, 20, 22, 23). Therefore, delving into the study of secreted proteins allows clinicians to evaluate the prognosis of patients with early-stage diseases and holds promise for individualized therapeutic interventions.

With regard to LUAD, a series of secreted proteins have been reported to be dysregulated and involved in LUAD progression. Widely used serum tumor markers, such as carcinoembryonic antigen, carbohydrate antigen 199, and neuron-specific enolase, have been used for the early diagnosis and classification of lung cancer (24). Many chemokines have been implicated in the modulation of the immune response, which has diverse functions in LUAD. It has been reported that CXCL17 expression in lung cancer cells could promote tumor progression (25). In addition, CCL20 was upregulated in patients with relapsed lung cancer and could accelerate cell proliferation through the ERK signaling pathway (26). Higher serum levels of IL-22 and HGF were observed in patients with non-small cell lung cancer (NSCLC) than in healthy subjects. Elevated serum IL-22 and loss of IL-34 expression have been associated with a poor prognosis in patients with NSCLC and LUAD, respectively (27). Pang *et al.* reported that RCC2 overexpression could induce JNK activation and upregulate MMPs (such as MMP-1, MMP-2, and MMP-9), which belong to a family of metastasis-related secretory proteins, in LUAD (28). To the best of our knowledge, a systematic investigation of secreted proteins in LUAD has not been reported.

Since the detection of secreted proteins is convenient, economical, and a minimally invasive intervention, developing a prognostic signature of secreted protein-related genes (SPRGs) is of great interest. In this study, we aimed to develop a useful tool to evaluate the prognostic role of secreted protein-related risk score (SPRrisk) based on large-scale RNA-seq data for LUAD from The Cancer Genome Atlas (TCGA) cohort and Gene Expression Omnibus (GEO) databases. We further used least absolute shrinkage and selection operator method (LASSO) regression and multivariate Cox regression analyses to investigate potential secreted protein-related prognostic genes and constructed SPRrisk to predict survival in patients with LUAD.

MATERIALS AND METHODS

Data Acquisition

The LUAD level 3 RNA-seq data (read counts) and corresponding clinical information of 535 tumor samples and 59 normal samples were downloaded from TCGA (<https://portal.gdc.cancer.gov/>) as a training cohort, and the ENSEMBL gene ID was converted into a gene name for the subsequent analysis. The LUAD microarray data GSE72094 ($n = 442$) and GSE31210 ($n = 246$) were downloaded with complete clinical data from GEO (<http://www.ncbi.nlm.nih.gov/geo>) to serve as the validation sets. GSE72094 was from the chip platform GPL15048 (Rosetta/Merck Human RSTA Custom Affymetrix 2.0 microarray) (29), and the CEL files were normalized against their median sample using the IRON algorithm (30). GSE31210 was from the chip platform GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array) (31), and the mRNA expression data were normalized by the MAS5 algorithm (32). All the genes detected with more than one probe were calculated by mean

expression, and the gene expression data were log2-transformed before the analyses. Despite the large number of secreted proteins, only those secreted into the plasma were selected for our study. Finally, a total of 730 secreted protein-related genes (SPRGs) ultimately remained. The SPRG list was retrieved from the HPA database (<https://www.proteinatlas.org/humanproteome/blood+protein/secreted+to+blood>) and is provided in **Supplementary Table S1**.

Construction and Validation of a Prognostic Secreted Protein-Related Gene Signature

Using the RNA-Seq data of TCGA LUAD dataset and the list of SPRGs obtained as detailed above, we finally got the SPRGs expression profiles of patients with LUAD. TCGA LUAD read count data of SPRGs were then processed with the “edgeR” R package (version 3.36.0) for normalization and differential expression analysis. The following criteria were applied to filter differentially expressed secreted protein-related genes (DESPRGs) between tumor tissues and normal tissues: false discovery rate (FDR) < 0.05 and |log2 fold change| > 1 (33). Following this, univariate Cox regression analysis was applied to identify SPRGs with prognostic values using the “survival” package of the R software. Next, we conducted LASSO regression to narrow the range of prognostic genes, removed overfitting between genes, and calculated risk scores according to LASSO regression coefficients with the “glmnet” R package (34, 35). Therefore, a final model with 9 variables was obtained at the end. The risk score of each patient was calculated by multiplying the gene expression by the regression coefficient. The formula was established as follows: Secreted protein-related signature (SPRS) = $\sum_i \beta_i * Exp_i$. The final risk model was as follows: SPRGrisk = (-0.0290 * CLEC3B expression) + (0.1830 * C1QTNF6 expression) + (0.0020 * TCN1 expression) + (0.0123 * F2 expression) + (0.0522 * FETUB expression) + (0.0381 * IGFBP1 expression) + (0.0185 * ANGPTL4 expression) + (0.0107 * IFNE expression) + (0.0051 * CCL20 expression). The expression level of each gene was log2-normalized.

Evaluation of the Predictive Efficacy of the Prognostic Model

All patients were classified as high or low risk based on their median risk score, and survival curves were used to assess the predictive power of the prognostic model between the high- and low-risk groups with “survminer” R package. “timeROC” package was applied to evaluate the prognostic model’s ability to predict outcomes in patients with LUAD (36), and the areas under the curve (AUC) at different time points of all the variables were compared. Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) were performed to explore the distribution of the high- and low-risk groups using the “Rtsne” R package, with the expression matrix of the 9 selected genes as the input. Given the presence of correlation between SPRrisk and other risk factors (stage, gender, and TP53 mutation), to make the results more convincing, we performed the additional multivariate Cox analyses after adjustments for other clinicopathologic

characteristics. The detailed method was as follows: after removing the clinicopathologic parameter that needs to be adjusted, multivariate Cox analysis was performed on the remaining parameters, including SPRrisk. To evaluate the proposed SPRG model in comparison with other models, different risk scores were calculated for each patient using different models, and the AUC at different time points of all the different risk scores was drawn on the same set of coordinate axes for comparison.

Construction and Evaluation of a Nomogram

To identify the best prognostic indicators of the survival outcome of LUAD patients, univariate and multivariate Cox regression analyses were performed. Finally, variables whose *p*-value was less than 0.05 (*p* < 0.05) were selected to build a nomogram. The nomogram was constructed and evaluated by employing the R packages “rms”, “regplot”, and “Hmisc”. Moreover, the concordance index (C-index) and calibration curve were adapted to appraise the availability of this nomogram in both the training set and the validation set. The ROC analysis was also performed to assess the accuracy of the nomogram for 1-, 3-, and 5-year overall survival of patients with LUAD.

Characterization of the Immune Cell Landscape and the Prediction of Therapeutic Sensitivity in Patients With LUAD

The ESTIMATE algorithm was adopted to estimate the immune scores and stromal scores of LUAD patients with the R package “estimate” (37). In order to determine the composition of immune cells in the tumor microenvironment of each sample, we performed deconvolution with support vector regression using the CIBERSORT algorithm. The LM22 immune cell signature matrix was downloaded from the CIBERSORT website (<https://cibersort.stanford.edu/>). CIBERSORT was run for 1,000 permutations, and quantile normalization was applied. The potential response of patients with LUAD to immunotherapy was evaluated by the tumor immune dysfunction and exclusion (TIDE) score and immunophenoscore (IPS). Data is available for download from TIDE (<http://tide.dfci.harvard.edu/>) or The Cancer Immunome Atlas (TCIA) (<https://tcia.at/home>) (38, 39).

Tissue Samples and Quantitative Real-Time Polymerase Chain Reaction

All tissue and blood samples were collected from the Thoracic Surgery Department of Wuhan Union Hospital, which was approved by the Medical Ethics Committee of the hospital. Written informed consent was obtained from each involved patient. A total of 25 lung adenocarcinoma tissue samples and 25 non-tumor lung tissues were obtained from the tumor and adjacent tissue of lung adenocarcinoma patients who underwent tumor resection between October 2019 and July 2021. All included patients were newly diagnosed and had not received any relevant treatment prior to surgery, and follow-up started at

the date of diagnosis and ended at death or on October 31, 2021. For gene expression studies, the total RNA from tissues was isolated with TRIZOL reagents (Takara, Otsu, Japan). RNA extraction was performed according to the manufacturer's protocols. RNA was reverse-transcribed into cDNA by RT-PCR using Hiscript[®] Q RT SuperMix (Vazyme, Nanjing, China) in a 20- μ l total sample volume. The parameters of reaction were as follows: 95°C for 30 s, followed by 40 cycles at 95°C for 5 s and 60°C for 1 min. Then, the gene expression levels were measured by quantitative PCR (qPCR). qPCR was performed in a CFX Connect Real-Time PCR Detection System (Bio-Rad, Hercules, CA, USA) using SYBR green supermix (Vazyme, Nanjing, China). The total amount of mRNA was normalized to endogenous GAPDH mRNA. The $2^{-\Delta\Delta C_t}$ method was used to calculate the related gene expression levels. The primer sequences are listed in **Supplementary Table S2**.

Enzyme-Linked Immunosorbent Assay

We used enzyme-linked immunosorbent assay (ELISA) kits to measure the plasma levels of FETUB [Human Fetuin B (FETUB) ELISA Kit; Reddot Biotech], IGFBP1 [Human Insulin Like Growth Factor Binding Protein 1 (IGFBP1) ELISA Kit; Reddot Biotech], TCN1 [Human Transcobalamin I (TCN1) ELISA Kit; Reddot Biotech], ANGPTL4 [Human Angiopoietin Like Protein 4 (ANGPTL4) ELISA Kit; Reddot Biotech], and CCL20 (Human CCL20/MIP-3 alpha ELISA Kit, Proteintech). Approximately 1 ml of blood was collected in EDTA-coated tubes on ice (BD Vacutainer), centrifuged at 4°C (2,000 \times g, 10 min), aliquoted, and stored at -20°C until the assay was performed using the ELISA kits. Subsequent steps were carried out following the manufacturer's protocol.

Statistical Analysis

All statistical analyses were conducted using R version 4.0.0 (2020-04-24). To test for differential expression across two groups (tumor and normal), the *p*-values were adjusted for multiple testing based on the FDR according to the Benjamini-Hochberg approach. The survival analysis was performed using the Kaplan-Meier (KM) method, and the subgrouping of the samples was stratified by medians of gene expression levels. Student's *t*-test or one-way analysis of variance was used to analyze differences between groups in variables with a normal distribution. Differences in proportions were compared by chi-square test. If not specified above, a *P*-value less than 0.05 was considered statistically significant, and all *P*-values were two-tailed.

RESULTS

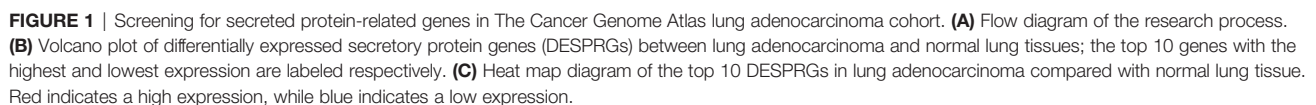
Identification of Differentially Expressed Secreted Protein-Related Genes in the Cancer Genome Atlas Training Cohort

A flow chart was developed to systematically describe our study (**Figure 1A**). We firstly obtained the gene expression profiles of the SPRGs from TCGA LUAD dataset. The gene list was

compiled from the literature and the HPA database, which included 730 genes encoding the proteins secreted into plasma. To screen for DESPRGs, differential expression analysis of the data was performed using the edgeR software. Finally, we obtained 379 DESPRGs, including 281 upregulated and 98 downregulated SPRGs (**Figure 1B**). The criteria to indicate a significant differential expression were as follows: $|\log_2\text{-fold change}| > 1$ and $\text{FDR} < 0.05$. The most obvious genes with an elevated expression were *CGA*, *ALB*, *FGB*, *FGF19*, *CALCA*, alpha fetoprotein (*AFP*), *GCG*, *INSL4*, *GC*, and *SERPINA4*. The most significantly downregulated genes included *CSF3*, *FCN3*, *ANGPT4*, *CD5L*, *CLEC3B*, *VEGFD*, *PI16*, *SCUBE1*, *DNASE1L3*, and *FOLR3* (**Figure 1C**). Evidently, a large number of SPRGs encoding proteins secreted into the plasma were differentially expressed in LUAD compared with those in adjacent normal samples. Among these genes, *AFP* and *VEGFD* are well-known tumor biomarkers that have been extensively applied in the early screening of tumors (40, 41). *FGF19* has been implicated in the pathogenesis of several cancers, including hepatocellular carcinoma in mice and potentially in humans (42). Some of these DESPRGs belong to the endocrine signaling pathway (including *CGA* and *GCG*), whereas some DESPRGs are involved in the regulation of the immune system process (including *CSF3* and *CD5L*). These results altogether suggested that these SPRGs had a potential prognostic value in patients with LUAD.

Establishment of Secreted Protein-Related Gene Signatures for Prognosis

Next, the prognostic role of SPRGs in patients with LUAD was examined. Using univariate Cox regression analysis on the stated DESPRGs, we identified 86 overall survival-associated genes in the samples of patients with LUAD in TCGA cohort ($P < 0.05$) (**Supplementary Figure S1**). Subsequently, we performed LASSO Cox regression analysis to identify the most robust marker genes for prognosis. Tenfold cross-validation was applied to prevent over-fitting, with a selected optimal λ value of 0.0602 (**Figures 2A, B**). Finally, an ensemble of 9 genes (*CLEC3B*, *C1QTNF6*, *TCN1*, *F2*, *FETUB*, *IGFBP1*, *ANGPTL4*, *IFNE*, and *CCL20*) was identified. The genes' individual nonzero LASSO coefficients and the distribution of LASSO coefficients of the gene signature are shown in **Figure 2C**. Meanwhile, KM survival analysis was performed for each gene separately, and the survival curves were plotted. The results indicated that patients with LUAD with a high expression of *C1QTNF6*, *TCN1*, *F2*, *FETUB*, *IGFBP1*, *ANGPTL4*, *IFNE*, and *CCL20* had a poor prognosis, whereas patients with a high expression of *CLEC3B* had a better prognosis (**Supplementary Figures S2A–I**). The patients' risk scores were calculated from the expression levels and regression coefficients: $\text{SPRrisk} = (-0.0290 * \text{CLEC3B expression}) + (0.1830 * \text{C1QTNF6 expression}) + (0.0020 * \text{TCN1 expression}) + (0.0123 * \text{F2 expression}) + (0.0522 * \text{FETUB expression}) + (0.0381 * \text{IGFBP1 expression}) + (0.0185 * \text{ANGPTL4 expression}) + (0.0107 * \text{IFNE expression}) + (0.0051 * \text{CCL20 expression})$. The expression level of each gene was \log_2 -normalized.



To assess whether risk score was an independent prognostic factor for LUAD, we performed univariate and multivariate Cox regression analyses for the SPRisk and other risk factor variables (age, TNM classification, gender, stage, KRAS mutation, TP53 mutation, and EGFR mutation) in TCGA training cohort. The

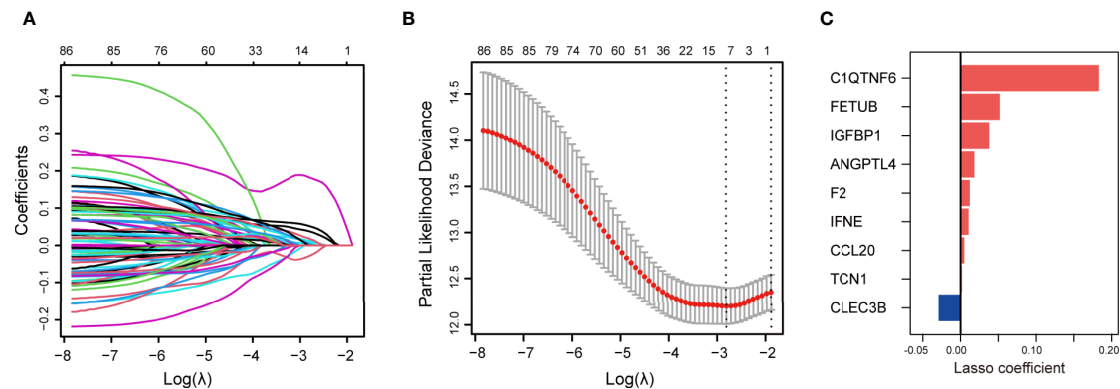


FIGURE 2 | Construction of secreted protein-related gene (SPRG)-related model for patients with lung adenocarcinoma (LUAD). **(A)** Least absolute shrinkage and selection operator method (LASSO) coefficient spectrum of differentially expressed secretory protein genes in The Cancer Genome Atlas LUAD cohort. **(B)** Cross-validation fit curve calculated by LASSO regression method. **(C)** Distribution of LASSO coefficients of the selected SPRGs.

univariate Cox regression results indicated that the patients' risk scores were significantly associated with overall survival (OS) (HR = 5.546, 95%CI = 3.735–8.234, $P < 0.001$) (**Figure 3F**). In the multivariate Cox regression analysis, SPRrisk was proved to be an independent risk factor for OS in TCGA training cohort (HR = 4.224, 95%CI = 2.727–6.541, $P < 0.001$) (**Figure 3G**). Furthermore, the results of the time-ROC analysis showed that SPRrisk was the most accurate predictor for OS (**Figure 3H**).

SPRrisk Is Closely Related to Different Clinicopathological Features

The expression levels of the nine SPRGs in the high- and low-risk groups in TCGA cohort LUAD dataset are presented in a heat map (**Figure 4A**). We also analyzed the association between the patients' risk scores and other pathological features in TCGA LUAD dataset. There were significant differences in the risk scores between patients of different gender ($P < 0.05$), clinical stage ($P < 0.01$), T stage ($P < 0.01$), KRAS mutation status ($P < 0.05$), and TP53 mutation status ($P < 0.01$) (**Figures 4B–G** and **Table 1**). The correlations among the SPRrisk, clinical stage, T stage, and TP53 mutation status partly revealed why SPRrisk could be a better prognostic marker in predicting OS for LUAD patients. Moreover, the incidence of LUAD is higher in women, and lung cancer in women is a severe health problem globally (43). LUAD is considered a different disease in women and men (43). However, the effect of sex on LUAD patients' survival is still controversial. In our study, men had a higher SPRrisk than women in TCGA cohort; this result may be due to differences in secretion environments between males and females or due to a greater exposure to risk factors in men. A more conclusive explanation requires further studies.

High SPRrisk Reflects the Low Level of Immune Infiltration in LUAD

Since many cytokines are secreted proteins and participate in the regulation of the tumor microenvironment, we analyzed the differences in their composition between the high- and low-risk

groups. Firstly, using the ESTIMATE algorithm, we observed that the low-risk group had higher ESTIMATE, immune, and stromal scores and lower tumor purity than the high-risk group, suggesting that the tumor cells in the low-risk group had more immune cell infiltration (**Figures 5A–D**). Moreover, we used CIBERSORT on RNA-seq gene expression profiles to quantify the relative abundance of 22 different immune cell types in the tumor immune microenvironment. The results revealed that the SPRrisk-low group had high levels of multiple antitumor immune components, including M1 macrophages and CD8⁺ T cells, while the proportion of M2 macrophages was higher in the SPRrisk-high group (**Figure 5E**). We also examined the expression of immunomodulatory genes between the high- and low-risk groups and found that high-risk patients had higher levels of pro-tumorigenic immunomodulatory molecules (including CD274 and CD276) and lower levels of anti-tumorigenic immunomodulatory molecules (including CD40LG and TNFRSF14) (**Figure 5F**). Human leucocyte antigen (HLA) complexes control the adaptive immunity by delivering defined fractions of intracellular and extracellular protein content to immune cells and have been shown to play important roles in anti-tumor immunity (44). In the present study, we also observed lower levels of HLA complexes in high-risk patients, including *HLA-DPB2*, *HLA-DQB1*, *HLA-DMA*, and *HLA-DRA* (**Supplementary Figure S4**). Similar results were also observed in GSE72094 (**Supplementary Figures S5A–C**) and GSE31210 (**Supplementary Figures S6A–C**). To explore the potential response of patients with LUAD to immunotherapy, we compared the TIDE scores and IPS across different SPRrisk groups in TCGA LUAD dataset. The results indicated that the patients in the low-SPRrisk group had a lower TIDE score and a higher IPS than those in the high-SPRrisk group (**Supplementary Figures S7A–E**). Similarly, the TIDE score distribution plots in two independent datasets (GSE72094 and GSE31210) yielded consistent findings (**Supplementary Figures S7F, G**). Collectively, these results suggested that risk scores may predict the effectiveness of immunotherapy in patients with LUAD.

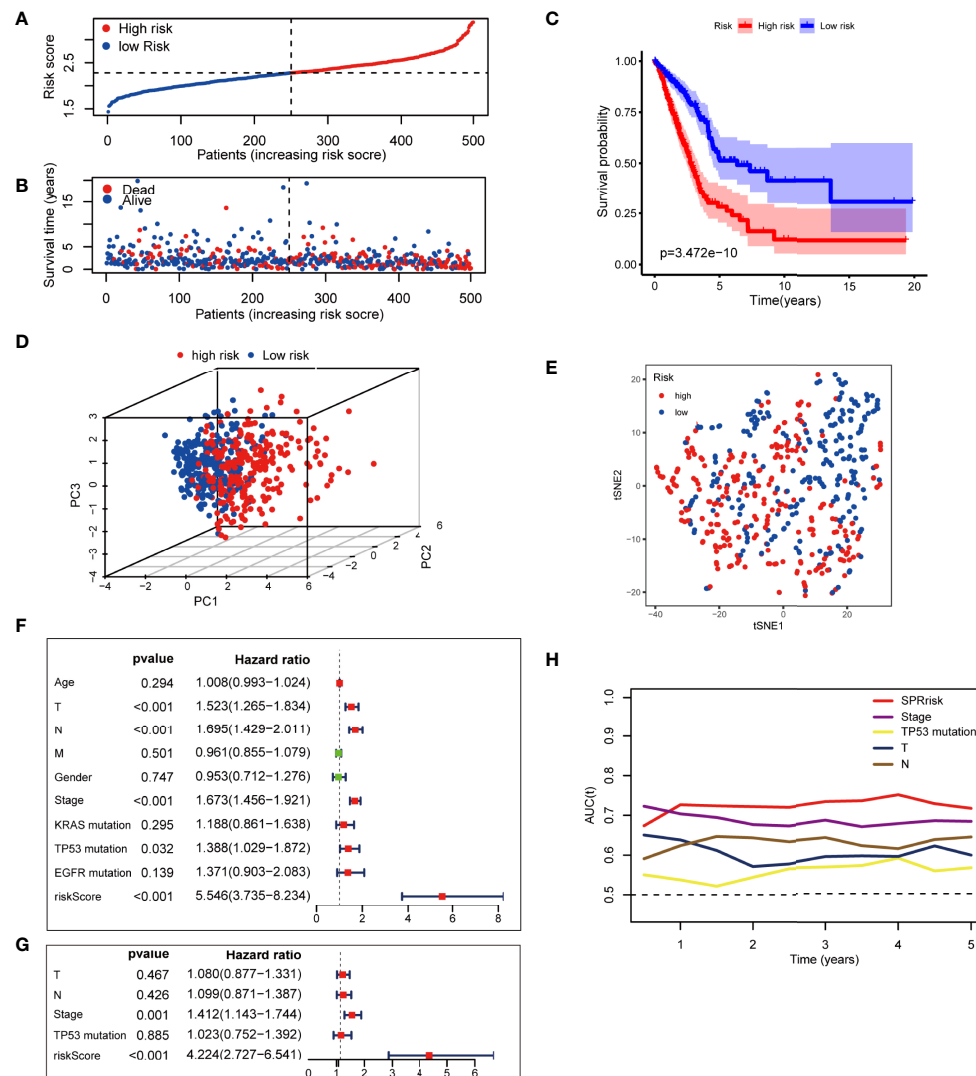


FIGURE 3 | Assessment of the prognostic signature in The Cancer Genome Atlas (TCGA) testing cohort. **(A)** Distribution of risk score and survival time of patients with lung adenocarcinoma (LUAD). **(B)** Scatter plot of survival status and risk score in patients with LUAD. **(C)** Kaplan–Meier curves of overall survival time between high- and low-risk groups using the log-rank test in TCGA LUAD dataset. **(D)** Principal component analysis of the 9 SPRG expression profiles of the high- and low-risk groups. **(E)** t-SNE analysis of the 9 SPRG expression profiles of the high- and low-risk groups as indicated by different colors. **(F, G)** Forest plot with hazard ratios from the univariate and multivariable Cox proportional hazards regression analysis in TCGA cohort. **(H)** The areas under the curve of time-dependent receiver operating characteristic curves verified the prognostic performance of the risk score in TCGA cohort.

Validation of the Nine Secreted Protein-Related Gene Signatures in the Independent Validation Sets

The baseline characteristics of the patients in the different risk groups in the GSE72094 and GSE31210 datasets are shown in **Supplementary Table S3**. To examine the accuracy of the model constructed based on TCGA testing cohort, we calculated the risk score of each patient in the validation sets according to the formula presented above. Additionally, the patients were divided into high-risk and low-risk groups according to the median risk

score. The results were consistent with those of TCGA testing cohort and indicated that the patients in the high-risk group had shorter survival times than those in the low-risk group (**Figures 6A, B**). In the multivariate analyses of the two independent sets, both SPRisk and stage were independent prognostic risk factors, suggesting the presence of a complementary mechanism (**Figures 6C, D**). Furthermore, in TCGA and independent validation sets, the SPRisk was consistently an independent prognostic factor after adjustments for the different clinicopathological characteristics (**Supplementary Table S4**). The ROC analysis also showed

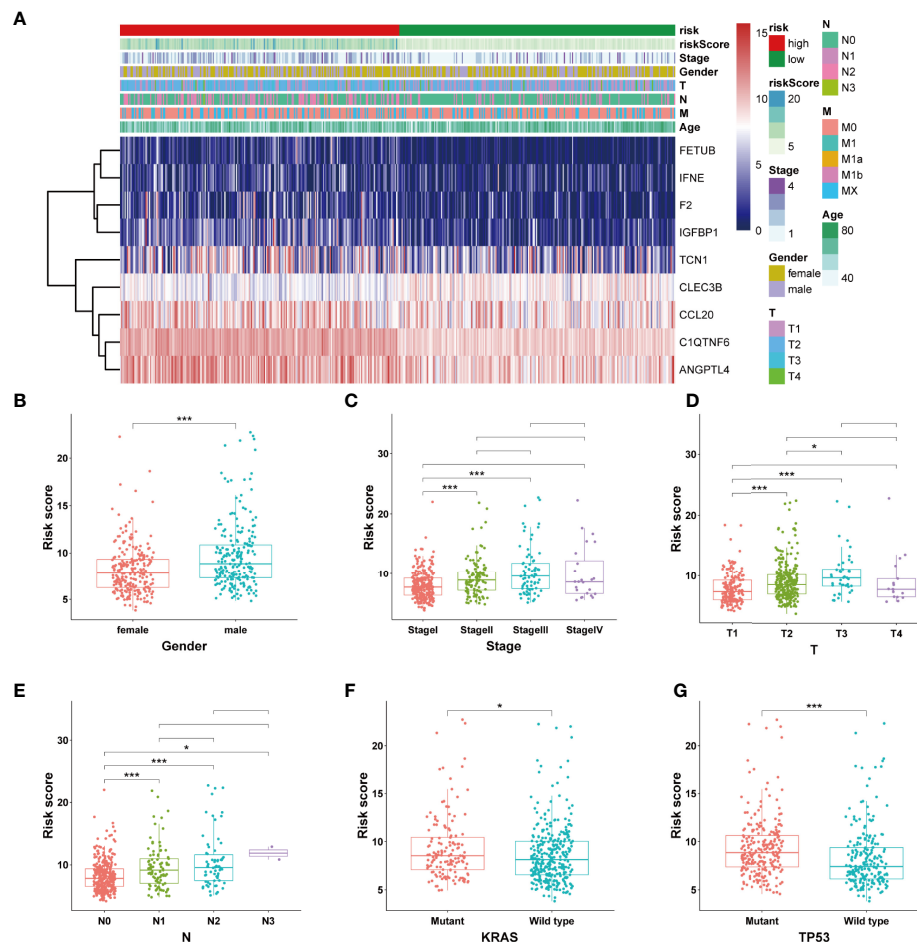


FIGURE 4 | Estimation of the correlation between the SPRisk and different clinicopathological features. **(A)** Heat map showing the association of the expression levels of 9 selected secreted protein-related genes and clinicopathologic features. **(B–G)** The different levels of risk scores in lung adenocarcinoma patients were stratified by gender, clinical stage, T stages, N stages, KRAS mutation status, and TP53 mutation status. * $P < 0.05$, *** $P < 0.001$.

higher AUCs for SPRisk, which highlighted the strong prognosis-predicting ability of SPRisk (**Figures 6E, F**).

Combination of the Secreted Protein-Related Signature and Clinicopathological Features Improves Survival Prediction

A nomogram was established using the SPRisk and stage as independent prognostic factors in TCGA cohort (**Figure 7A**). Through calculation, the C-index of the prognostic model developed, using TCGA cohort, was 0.784, indicating that the consistency of the model was satisfactory. The calibration curve results demonstrated that the survival status predicted by the prognostic model was in good agreement with the actual survival status (**Figures 7B–D**). In addition, multivariate ROC curves were plotted to compare the AUC of different prognostic factors, and the results showed that the nomogram presented the highest accuracy in predicting 1-, 3-, and 5-year OS (AUC = 0.838, 0.832, and 0.841, respectively) (**Figures 7E–G**). Similarly, we evaluated the ability of the prognostic model to predict the survival status

of patients with LUAD in the validation cohorts (GSE72094 and GSE31210). Two nomograms were generated based on the independent prognostic factors to predict the probability of OS (**Supplementary Figures 8A, B**). The C-index values for the two nomograms were 0.753 and 0.768, respectively, and the calibration plots indicated that the predicted survival of the model matched the actual survival (**Supplementary Figures 8C, D**). Additionally, the accuracy of this nomogram in predicting OS was the highest (**Supplementary Figures 8E, F**). Overall, these results suggested that the nomogram has a great potential for predicting the survival and prognosis of patients with LUAD.

A few laboratories have published studies in which they constructed prognostic models to achieve a more accurate evaluation of the prognosis for patients with LUAD—for example, the hypoxia-related risk score (HRisk) and the tumor microenvironment-related risk score (TMRisk) (45, 46). We compared our model with these existing predictive models in the validation sets (TCGA, GSE72094, and GSE31210) using multivariate Cox regression analysis and time-ROC analysis. In

TABLE 1 | Clinical and pathological characteristics of high- and low-risk patients in The Cancer Genome Atlas lung adenocarcinoma data set.

Parameter	SPRrisk-high	SPRrisk-low	P-value
	(N = 250)	(N = 250)	
Age (years)			
Mean (SD)	63.8 (10.4)	66.7 (10.4)	0.0015
Median (min., max.)	64 (33.0, 87.0)	68.0 (41.0, 88.0)	
Pathologic T			
T1	59 (23.6%)	108 (43.2%)	<0.001
T2	150 (60.0%)	117 (46.8%)	
T3	32 (12.8%)	13 (5.2%)	
T4	8 (3.2%)	10 (4.0%)	
Pathologic N			
N0	137 (54.8%)	187 (74.8%)	<0.001
N1	60 (24.0%)	34 (13.6%)	
N2	48 (19.2%)	21 (8.4%)	
N3	2 (0.8%)	0 (0%)	
Pathologic M			
M0	164 (65.6%)	168 (67.2%)	0.699
M1	14 (5.6%)	10 (4.0%)	
MX	70 (28.0%)	70 (28.0%)	
Gender			
Male	118 (47.2%)	112 (44.8%)	0.654
Female	132 (52.8%)	138 (55.2%)	
Stage			
Stage I	103 (41.2%)	165 (66.0%)	<0.001
Stage II	76 (30.4%)	43 (17.2%)	
Stage III	53 (21.2%)	27 (10.8%)	
Stage IV	14 (5.6%)	11 (4.4%)	
KRAS mutation			
WT	163 (65.2%)	175 (70.0%)	0.281
MUT	81 (32.4%)	69 (27.6%)	
TP53 mutation			
WT	88 (35.2%)	145 (58.0%)	<0.001
MUT	156 (62.4%)	99 (39.6%)	
EGFR mutation			
WT	216 (86.4%)	210 (84.0%)	0.497
MUT	28 (11.2%)	34 (13.6%)	

the GSE72094 dataset, the SPRrisk was still confirmed as an independent prognostic factor (**Supplementary Figures S9A, B**). New nomograms were constructed using the SPRrisk and the existing predictive gene signatures (**Supplementary Figures S9C, D**), and the addition of SPRrisk resulted in further improvements in the model's predictive ability of LUAD prognosis (**Supplementary Figures S9E, F**). In TCGA dataset, only SPRrisk and TMErisk were independent prognostic factors, and the nomogram and the ROC curves revealed that modeling outperformed both separately (**Supplementary Figures S10A–D**). In the GSE31210 dataset, HRrisk and TMErisk were not statistically significant in the multivariate Cox regression analysis (**Supplementary Figures S10E, F**).

Validation of the Expression Levels of Selected SPRGs

To assess the expression levels of the selected SPRGs within the lung tissue, lung tumor tissues ($n = 25$) and normal lung tissues ($n = 25$) were analyzed by qPCR. Compared to those in the non-tumor lung tissues, the expression levels of *C1QTNF6*, *TCN1*, *F2*, *FETUB*, *IGFBP1*, *ANGPTL4*, *IFNE*, and *CCL20* were upregulated in lung cancer tissues, while the expression level of *CLEC3B* was

downregulated (**Figures 8A–I**). To further validate the predictive power of the SPRG signature, ELISA was used to measure the plasma levels of the secreted proteins in our clinical dataset. The baseline characteristics of the patients are shown in **Supplementary Table S5**. Consistent with the qPCR results, the levels of *TCN1*, *FETUB*, *IGFBP1*, *ANGPTL4*, and *CCL20* were significantly elevated in the plasma of patients with LUAD (**Figures 8J–N**). The KM survival analysis showed that stage I patients had an extended survival rate than stage II patients (**Supplementary Figure S11**). Accordingly, differences in plasma levels of secreted proteins between stage I and stage II patients were analyzed using Students *t*-tests, and the results indicated that patients with stage II LUAD had higher plasma protein levels of *TCN1*, *FETUB*, *IGFBP1*, *ANGPTL4*, and *CCL20* (**Table 2**).

DISCUSSION

Secreted proteins are first synthesized in the cell and then actively secreted to other organelles or the extracellular environment. Secreted proteins include cytokines, growth factors, complement, degradation enzymes, antibodies, peptide hormones, and immunoglobulins, all of

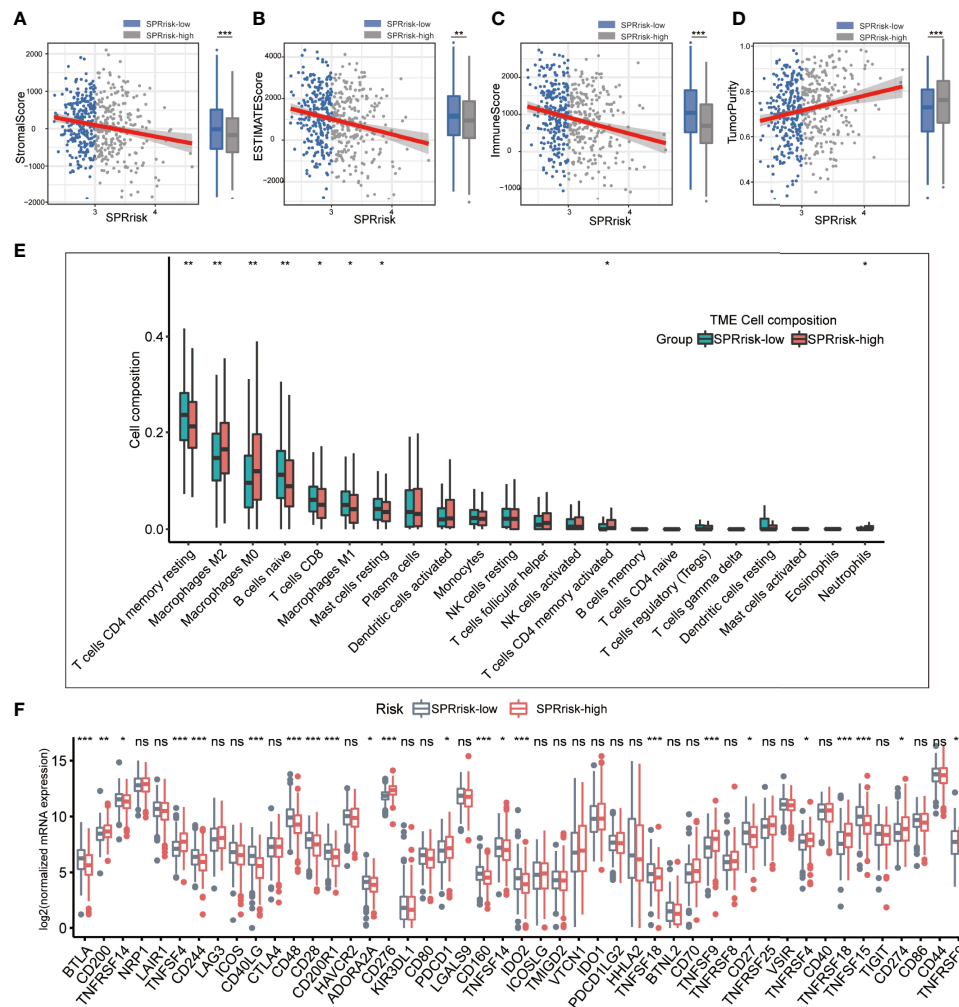


FIGURE 5 | Landscape of immune and stromal cell infiltrations in the low- and high-risk groups. **(A–D)** Comparison of ESTIMATE scores, immune scores, stromal scores, and tumor purity between the high- and low-risk groups. **(E)** The immune cell infiltration levels of 22 immune cell types between the low- and high-risk groups for patients with lung adenocarcinoma. **(F)** Analyses for the expression of immune checkpoint genes in the high- and low-risk groups. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and ns (no significance).

which have important physiological functions (47, 48). It is estimated that more than 2,000 proteins in human cells are secreted, and these protein molecules are critical in regulating the physiology and development of organisms. However, the biological functions of these proteins have remained poorly understood (49). Secreted proteins can be classified into classical secreted proteins and non-classical secreted proteins based on whether the N-terminal signal peptides are involved in the protein secretion process or not (50). In our study, we focused on the genes encoding proteins secreted into the plasma; hence, we selected nine SPRGs from the expression profiles of patients with LUAD to construct a prognostic model with good predictive power and specificity. It is worth mentioning that our research is the first to screen differentially expressed genes based on secreted proteins and to build a prognosis model based on these SPRGs for patients with LUAD. Sun *et al.* reported that *CLEC3B*, which encodes tetranectin in humans, was significantly

downregulated in patients with lung cancer compared with that in nontumor control groups according to database analysis and patient tissue sample detection (51). Indeed the plasma levels of *CLEC3B* are altered in the blood samples of patients with COVID-19 infection or acute coronary syndrome (52, 53). *CIQTNF6*, encoding C1q/tumor necrosis factor-related protein 6, is a newly identified adiponectin paralog associated with inflammation (54). Zhang *et al.* found that the inhibition of *CIQTNF6* attenuated cell proliferation, migration, and invasion and promoted apoptosis *in vitro* and *in vivo* in NSCLC (55). TCN1 generates a transcobalamin–cobalamin (vitamin B12) complex and regulates cobalamin homeostasis. It was reported that high levels of TCN1 in human serum are associated with leukemia, hepatocellular carcinoma, and phylloides of breast tumors (56, 57). As a member of the cysteine protease inhibitor family, FETUB is a glycoprotein. It has been reported that the levels of FETUB are altered in human serum in the process of ischemic stroke or severe COVID-

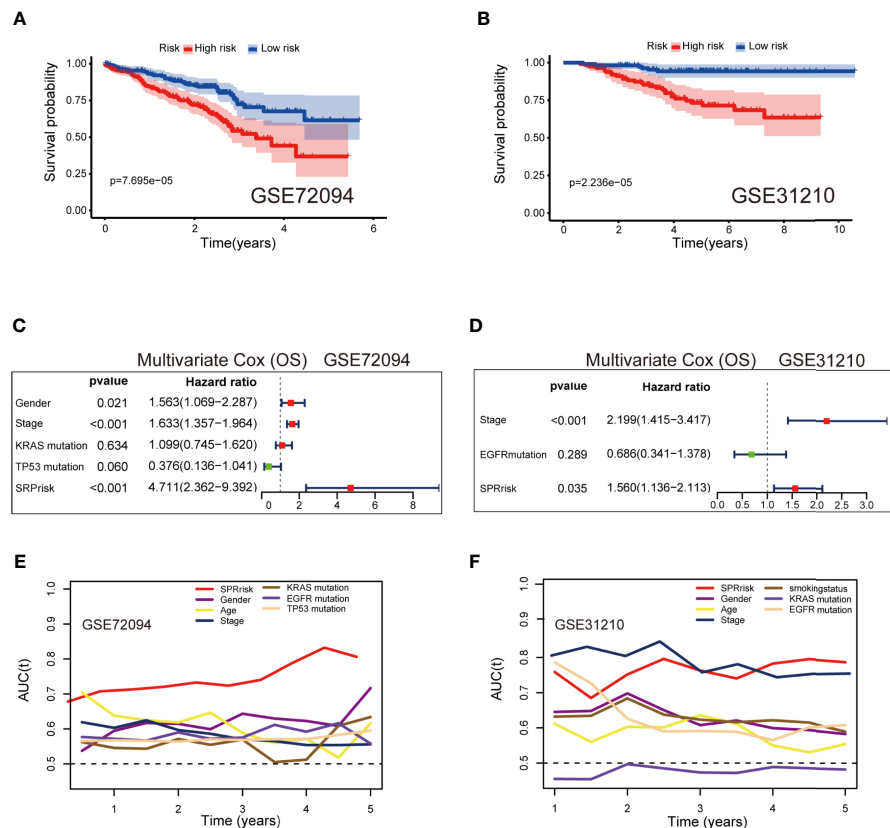


FIGURE 6 | Validation of the 9-gene signature in the independent validation sets. **(A, B)** Kaplan–Meier curves of overall survival time between the high- and low-risk groups using the log-rank test in the GSE72094 and GSE31210 datasets. **(C, D)** Forest plot with hazard ratios from the multivariable Cox proportional hazards regression analysis in the GSE72094 and GSE31210 datasets. **(E, F)** Areas under the curve of time-dependent receiver operating characteristic curves of risk factors in the GSE72094 and GSE31210 datasets.

19 (58, 59). IFNE is a type I interferon with unusual patterns of expression and function. Nevertheless, *in vivo* experiments indicated its efficacy in regulating mucosal immune responses and fighting bacterial and viral infections. ANGPTL4 and IGFBP1 are secreted into the plasma and are involved in cell energy metabolism (60–62).

During the development of malignant tumors, tumor cells secrete a variety of proteins, such as cytokines and proteolytic enzymes. The secreted proteins display an altered composition compared to the normal tissue, and their expression levels may change during different tumor stages (63). Consequently, secreted proteins have become the main source of potential tumor markers (64, 65). Since the expression levels of many secretory proteins are altered in tumors and these altered levels can be easily detected in body fluids, secretory proteins have good diagnostic and prognostic values. Some well-known secreted tumor markers include AFP, cell surface-associated protein (MUC1 or CA15-3), gastrin-releasing peptide, and prostate-specific antigen (or KLK3) (66–69). In our study, we conducted ELISA to examine the levels of several secreted proteins in the peripheral blood of patients with LUAD. We found that the levels of candidate secreted proteins were positively correlated with the clinical stage of the patients with

LUAD and agreed with the model results. Our findings highlighted the potential for the selected secreted proteins to serve as a prognostic marker for human LUAD. Detecting secreted protein levels in body fluids is economical, quantitative, and minimally invasive compared with RNA-seq, and more people may benefit from our study.

In addition, some secreted proteins play a significant role in regulating the immune microenvironment, which makes them potential targets for tumor therapy. Gelsolin (GSN) was reportedly secreted by cancer cells, which suppressed the killing activity of CD8⁺ T cells against tumor cells. Moreover, lower levels of intratumoral GSN transcripts are associated with signatures of anti-cancer immunity and increased patient survival (70). Tumor cells also secrete proteins, such as IL-10 and TGF- β , to remodel the immune microenvironment and promote tumor progression (71). Chemotherapy or radiotherapy can also induce senescence in tumor cells by modifying their secretome to a “senescence-associated secretory phenotype”, which also affects the immune response (72). In our study, we grouped the patients with LUAD into high- and low-risk groups based on the risk score. We found increased CD8⁺ T cell and M1 macrophage cell infiltration in the low-risk group, while the

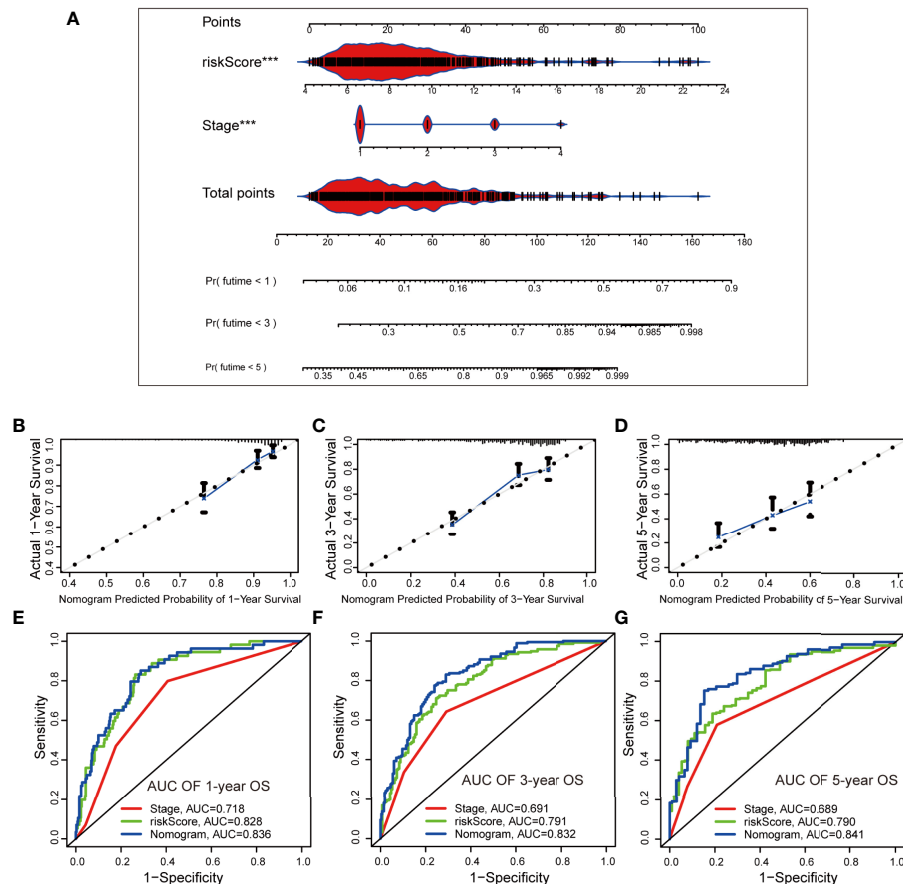


FIGURE 7 | Establishment and evaluation of a nomogram based on independent prognostic factors in The Cancer Genome Atlas cohort. **(A)** The nomogram generated from independent prognostic factors predicts the overall survival (OS) of patients with lung adenocarcinoma. **(B–D)** Calibration plot analyses for the predictive value of prognostic factors. **(E–G)** Comparison of receiver operating characteristic curves of independent prognostic factors in predicting 1-, 3-, and 5-year OS. *** $P < 0.001$.

high-risk group showed a higher M2 macrophage cell infiltration. CD8⁺ T cells are the primary mediators of anticancer immunity, and the modulation of the CD8⁺ T cell response has been a central focus of immunotherapy to treat cancer (73). Macrophages within the tumor stroma are tumor-associated macrophages and can be categorized as either classically activated M1 or alternatively activated M2 macrophages (74). M1 macrophages are considered anti-tumorous as they kill tumor cells by producing pro-inflammatory cytokines, such as IL-1 β and IL-12. In contrast, M2 macrophages are considered pro-tumorous since they stimulate the secretion of anti-inflammatory cytokines, such as IL-10, IL-13, and TGF- β (75). By estimating multiple published transcriptomic biomarkers based on pre-treatment tumor expression profiles, TIDE scores can predict patient response to immunotherapies (39). Our current findings also revealed that the low-risk group achieved a higher TIDE score than the high-risk group. The IPS was a superior predictor of response to anti-CTLA-4 and anti-PD-1 antibodies (38). Interestingly, significant differences in different IPS between the high- and low-risk groups were indicated. Thus, our risk model based on SPRGs

could be used to predict the immunotherapy response rates and present the most appropriate therapeutic options for patients with LUAD—for example, for the low-risk group with increased infiltration of CD8⁺ T cells and M1 macrophages, the immune checkpoint inhibitors may turn out to be effective treatments.

It is incontrovertible that this study has some limitations. First, various deficiencies in clinical information led to the incomplete validation of the partial results in TCGA LUAD training set and GEO verification set. Second, the number of available samples and clinical specimens was insufficient for conducting ELISA and comprehensive molecular studies, respectively. More tissue samples will be needed in further studies for validation. Third, the SPRGs were identified and validated using retrospective data from public databases. However, validation using a larger number of cases in a prospective cohort study is needed in the future. Finally, the molecular mechanism has not been characterized, and additional experiments are needed to explore the mechanistic roles of SPRGs in tumor progression.

In summary, our study enriches the current knowledge on the use of SPRGs for the prognostic prediction of LUAD. The

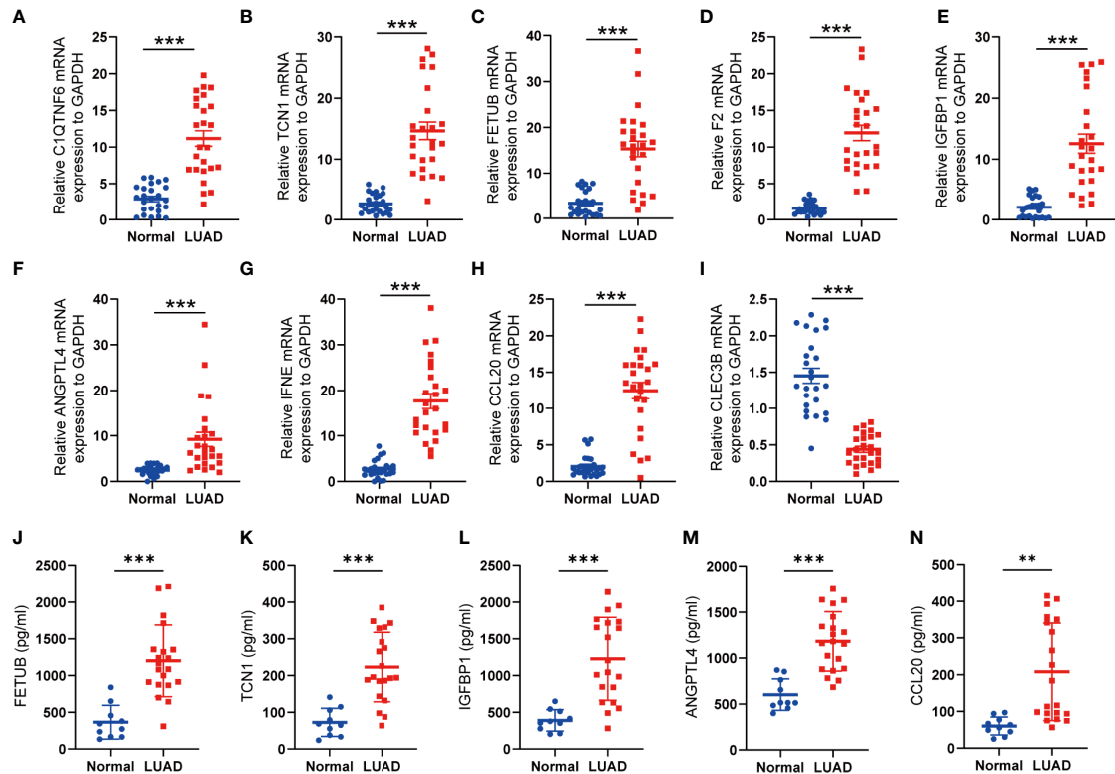


FIGURE 8 | Validation of the expression of SPRGs by qRT-PCR and ELISA. **(A–I)** Relative 9 SPRG mRNA expression between the normal and lung adenocarcinoma. **(J–N)** Plasma levels of the 5 secreted proteins between the normal and lung adenocarcinoma. $^{**}P < 0.01$, $^{***}P < 0.001$ by Student's *t*-test.

TABLE 2 | Correlation between the plasma levels of the secreted proteins and the clinical stage.

Parameter	Stage I	Stage II	P-value
	(N = 13)	(N = 7)	
Gender			
Female	6 (46.2%)	5 (71.4%)	0.5402
Male	7 (53.8%)	2 (28.6%)	
Smoking status			
Ever-smoker	7 (53.8%)	1 (14.3%)	0.2131
Never-smoker	6 (46.2%)	6 (85.7%)	
Age (years)			
Mean (SD)	59.1 (7.39)	58.4 (6.29)	0.839
Median (min., max.)	58 (49.0, 73.0]	58 (50.0, 69.0)	
ANGPTL4 (pg/ml)			
Mean (SD)	897 (214)	1,270 (391)	0.0451
Median (min., max.)	887 (582, 1,340)	1,160 (798, 1880)	
IGFBP1 (pg/ml)			
Mean (SD)	1,010 (455)	1,680 (421)	0.0052
Median (min., max.)	986 (441, 1,730)	1,770 (847, 2,140)	
CCL20 (pg/ml)			
Mean (SD)	169 (116)	294 (124)	0.0496
Median (min., max.)	95.3 (56.7, 357)	342 (117, 415)	
TCN1 (pg/ml)			
Mean (SD)	183 (81.9)	296 (79.6)	0.0102
Median (min., max.)	185 (48.2, 329)	332 (189, 393)	

prognostic SPRG model constructed in our study exhibited a robust capacity in predicting the survival outcomes of patients with LUAD and was correlated with the immune landscape of the LUAD microenvironment. We hope that these findings will offer useful insights for future studies and clinical practices.

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 Medical Ethics Committee of Union Hospital, Tongji Medical College, Huazhong University of Science and Technology. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SC and JZ conceived and designed the study. QL and LX collected and analyzed the related data. XF, QN, and LZ edited and wrote the draft. WM and HY revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (numbers 81973991 and 91643101 to WLM and numbers 82070066 and 81873401 to HY).

ACKNOWLEDGMENTS

All authors would like to thank the contributions of The Cancer Genome Atlas and Gene Expression Omnibus.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.870328/full#supplementary-material>.

Supplementary Figure 1 | Forest plot depicting the result of the univariate Cox regression analysis.

Supplementary Figure 2 | Survival analysis of overall survival in patients with lung adenocarcinoma. **(A–I)** Kaplan–Meier survival analysis of 9 selected secreted protein-related genes (SPRGs), respectively. The patients were stratified into high- and low-expression subgroups using the medians of the SPRGs.

Supplementary Figure 3 | Prognostic value of SPRS in predicting disease-free interval (DFI), disease-specific survival (DSS), and progression-free interval (PFI).

(A–C) Kaplan–Meier survival analysis of SPRS for FI, DSS, and PFI in TCGA cohort.

Supplementary Figure 4 | Expression of HLA family genes. Box plots depicting the HLA family genes' expression of high- and low-risk groups in TCGA cohort. The red box plots indicate the SPRrisk-high group, while the blue box plots indicate the SPRrisk-low group.

Supplementary Figure 5 | Landscape of immune cell infiltrations in GSE72094.

(A) Immune cell infiltration levels of 22 immune cell types between the SPRrisk-high and SPRrisk-low groups for patients with lung adenocarcinoma. **(B)** Analyses for the expression of immune checkpoint genes in the SPRrisk-high and SPRrisk-low groups. **(C)** Analyses for the expression of human leukocyte antigen family genes in the SPRrisk-high and SPRrisk-low groups.

Supplementary Figure 6 | Landscape of immune cell infiltrations in GSE31210.

(A) Immune cell infiltration levels of 22 immune cell types between these risk-high and SPRrisk-low groups for patients with lung adenocarcinoma. **(B)** Analyses for the expression of immune checkpoint genes in the SPRrisk-high and SPRrisk-low groups. **(C)** Analyses for the expression of human leukocyte antigen family genes in the SPRrisk-high and SPRrisk-low groups.

Supplementary Figure 7 | Distribution of the tumor immune dysfunction and exclusion (TIDE) scores and immunophenoscore (IPS) scores across different SPRrisk groups. **(A–D)** IPS score, IPS–CTLA4 blocker score, IPS–CTLA4 blocker score, and IPS–CTLA4 and PD1/PDL1/PDL2 blocker score distribution plots in The Cancer Genome Atlas (TCGA) training dataset. **(E)** TIDE score distribution plot in TCGA lung adenocarcinoma dataset. **(F)** TIDE score distribution plot in GSE72094 dataset. **(G)** TIDE score distribution plot in GSE31210 dataset.

Supplementary Figure 8 | Nomogram based on independent prognostic factors for overall survival (OS) of patients with lung adenocarcinoma (LUAD) in the independent validation sets. **(A, B)** The nomogram generated from independent prognostic factors predicts the OS of patients with LUAD in GSE72094 and GSE31210. **(C, D)** Calibration plot analyses for the predictive value of prognostic factors in the GSE72094 and GSE31210 datasets. **(E, F)** Comparison of receiver operating characteristic curves of independent prognostic factors in predicting the OS in the GSE72094 and GSE31210 datasets.

Supplementary Figure 9 | Assessment of the predictive ability of SPRrisk with the existing predictive models in GSE72094. **(A)** Multivariable Cox proportional hazards regression analysis of SPRrisk and TMERisk in the GSE72094 dataset. **(B)** Multivariable Cox proportional hazards regression analysis of SPRrisk and HRrisk in the GSE72094 dataset. **(C)** The nomogram generated from SPRrisk and TMERisk predicts the overall survival (OS) of patients in GSE72094. **(D)** The nomogram generated from SPRrisk and HRrisk predicts the OS of patients in GSE72094. **(E)** The areas under the curve (AUCs) of time-dependent receiver operating characteristic (ROC) curves verified the prognostic performance of SPRrisk and TMERisk in GSE72094. **(F)** The AUCs of time-dependent ROC curves verified the prognostic performance of the SPRrisk and HRrisk in GSE72094.

Supplementary Figure 10 | Assessment of the predictive ability of SPRrisk with the existing predictive models in The Cancer Genome Atlas (TCGA) and GSE31210 datasets. **(A)** Multivariable Cox proportional hazards regression analysis of SPRrisk and HRrisk in TCGA dataset. **(B)** Multivariable Cox proportional hazards regression analysis of SPRrisk and TMERisk in TCGA dataset. **(C)** The nomogram generated from SPRrisk and TMERisk predicts the overall survival of patients in TCGA dataset. **(D)** The areas under the curve of time-dependent receiver operating characteristic curves verified the prognostic performance of the SPRrisk and TMERisk in TCGA dataset. **(E)** Multivariable Cox proportional hazards regression analysis of SPRrisk and HRrisk in the GSE31210 dataset. **(F)** Multivariable Cox proportional hazards regression analysis of SPRrisk and TMERisk in the GSE31210 dataset.

Supplementary Figure 11 | Survival analysis of overall survival (OS) in patients with lung adenocarcinoma in our dataset. Kaplan–Meier survival analysis of clinical stage for OS in our dataset.

REFERENCES

- Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics 2021. *CA Cancer J Clin* (2021) 71(1):7–33. doi: 10.3322/caac.21654
- Travis WD, Brambilla E, Riely GJ. New Pathologic Classification of Lung Cancer: Relevance for Clinical Practice and Clinical Trials. *J Clin Oncol* (2013) 31(8):992–1001. doi: 10.1200/JCO.2012.46.9270
- Hirsch FR, Scagliotti GV, Mulshine JL, Kwon R, Curran WJ, Wu Y-L, et al. Lung Cancer: Current Therapies and New Targeted Treatments. *Lancet* (2017) 389(10066):299–311. doi: 10.1016/s0140-6736(16)30958-8
- Walters S, Maringe C, Coleman MP, Peake MD, Butler J, Young N, et al. Lung Cancer Survival and Stage at Diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: A Population-Based Study 2004–2007. *Thorax* (2013) 68(6):551–64. doi: 10.1136/thoraxjnl-2012-202297
- Zhang C, Zhang Z, Zhang G, Zhang Z, Luo Y, Wang F, et al. Clinical Significance and Inflammatory Landscapes of a Novel Recurrence-Associated Immune Signature in Early-Stage Lung Adenocarcinoma. *Cancer Lett* (2020) 479:31–41. doi: 10.1016/j.canlet.2020.03.016
- Wang Z, Wang Y, Yang T, Xing H, Wang Y, Gao L, et al. Machine Learning Revealed Stemness Features and a Novel Stemness-Based Classification With Appealing Implications in Discriminating the Prognosis, Immunotherapy and Temozolomide Responses of 906 Glioblastoma Patients. *Brief Bioinform* (2021) 22(5):bbab032. doi: 10.1093/bib/bbab032
- Gonzalez R, Jennings LL, Knuth M, Orth AP, Klock HE, Ou W, et al. Screening the Mammalian Extracellular Proteome for Regulators of Embryonic Human Stem Cell Pluripotency. *Proc Natl Acad Sci USA* (2010) 107(8):3552–7. doi: 10.1073/pnas.0914019107
- Lin H, Lee E, Hestir K, Leo C, Huang M, Bosch E, et al. Discovery of a Cytokine and Its Receptor by Functional Screening of the Extracellular Proteome. *Science* (2008) 320(5877):807–11. doi: 10.1126/science.1154370
- Liu T, Jia P, Ma H, Reed SA, Luo X, Larman HB, et al. Construction and Screening of a Lentiviral Secretome Library. *Cell Chem Biol* (2017) 24(6):767–771.e763. doi: 10.1016/j.chembiol.2017.05.017
- Uhlen M, Karlsson MJ, Hober A, Svensson AS, Scheffel J, Kotol D, et al. The Human Secretome. *Sci Signal* (2019) 12(609):eaaz0274. doi: 10.1126/scisignal.aaz0274
- Jang E, Lee S, Kim JH, Kim JH, Seo JW, Lee WH, et al. Secreted Protein Lipocalin-2 Promotes Microglial M1 Polarization. *FASEB J* (2013) 27(3):1176–90. doi: 10.1096/fj.12-222257
- Wu T, Zhang Q, Wu S, Hu W, Zhou T, Li K, et al. CILP-2 is a Novel Secreted Protein and Associated With Insulin Resistance. *J Mol Cell Biol* (2019) 11(12):1083–94. doi: 10.1093/jmcb/mjz016
- Yuh DY, Maekawa T, Li X, Kajikawa T, Bdeir K, Chavakis T, et al. The Secreted Protein DEL-1 Activates a Beta3 Integrin-FAK-ERK1/2-RUNX2 Pathway and Promotes Osteogenic Differentiation and Bone Regeneration. *J Biol Chem* (2020) 295(21):7261–73. doi: 10.1074/jbc.RA120.013024
- Jones SA, Jenkins BJ. Recent Insights Into Targeting the IL-6 Cytokine Family in Inflammatory Diseases and Cancer. *Nat Rev Immunol* (2018) 18(12):773–89. doi: 10.1038/s41577-018-0066-7
- Brychtova V, Vojtesek B, Hrstka R. Anterior Gradient 2: A Novel Player in Tumor Cell Biology. *Cancer Lett* (2011) 304(1):1–7. doi: 10.1016/j.canlet.2010.12.023
- Chevet E, Fessart D, Delom F, Mulet A, Vojtesek B, Hrstka R, et al. Emerging Roles for the Pro-Oncogenic Anterior Gradient-2 in Cancer Development. *Oncogene* (2013) 32(20):2499–509. doi: 10.1038/ncr.2012.346
- Di Maro G, Salerno P, Unger K, Orlandella FM, Monaco M, Chiappetta G, et al. Anterior Gradient Protein 2 Promotes Survival, Migration and Invasion of Papillary Thyroid Carcinoma Cells. *Mol Cancer* (2014) 13:160. doi: 10.1186/1476-4598-13-160
- Ellingsgaard H, Hauselmann I, Schuler B, Habib AM, Baggio LL, Meier DT, et al. Interleukin-6 Enhances Insulin Secretion by Increasing Glucagon-Like Peptide-1 Secretion From L Cells and Alpha Cells. *Nat Med* (2011) 17(11):1481–9. doi: 10.1038/nm.2513
- Choy EH, Kavanaugh AF, Jones SA. The Problem of Choice: Current Biologic Agents and Future Prospects in RA. *Nat Rev Rheumatol* (2013) 9(3):154–63. doi: 10.1038/nrrheum.2013.8
- Hunter CA, Jones SA. IL-6 as a Keystone Cytokine in Health and Disease. *Nat Immunol* (2015) 16(5):448–57. doi: 10.1038/ni.3153
- Johnson DE, O'Keefe RA, Grandis JR. Targeting the IL-6/JAK/STAT3 Signalling Axis in Cancer. *Nat Rev Clin Oncol* (2018) 15(4):234–48. doi: 10.1038/nrclinonc.2018.8
- Nishimoto N, Kanakura Y, Aozasa K, Johkoh T, Nakamura M, Nakano S, et al. Humanized Anti-Interleukin-6 Receptor Antibody Treatment of Multicentric Castleman Disease. *Blood* (2005) 106(8):2627–32. doi: 10.1182/blood-2004-12-4602
- Schett G, Elewaut D, McInnes IB, Dayer JM, Neurath MF. How Cytokine Networks Fuel Inflammation: Toward a Cytokine-Based Disease Taxonomy. *Nat Med* (2013) 19(7):822–4. doi: 10.1038/nm.3260
- Li Z, Wu W, Pan X, Li F, Zhu Q, He Z, et al. Serum Tumor Markers Level and Their Predictive Values for Solid and Micropapillary Components in Lung Adenocarcinoma. *Cancer Med* (2022). doi: 10.1002/cam4.4645
- Matsui A, Yokoo H, Negishi Y, Endo-Takahashi Y, Chun NA, Kadouchi I, et al. CXCL17 Expression by Tumor Cells Recruits CD11b+Gr1 High F4/80-Cells and Promotes Tumor Progression. *PLoS One* (2012) 7(8):e44080. doi: 10.1371/journal.pone.0044080
- Wang B, Shi L, Sun X, Wang L, Wang X, Chen C. Production of CCL20 From Lung Cancer Cells Induces the Cell Migration and Proliferation Through PI3K Pathway. *J Cell Mol Med* (2016) 20(5):920–9. doi: 10.1111/jcmm.12781
- Wang Z, Zhu J, Wang T, Zhou H, Wang J, Huang Z, et al. Loss of IL-34 Expression Indicates Poor Prognosis in Patients With Lung Adenocarcinoma. *Front Oncol* (2021) 11:639724. doi: 10.3389/fonc.2021.639724
- Pang B, Wu N, Guan R, Pang L, Li X, Li S, et al. Overexpression of RCC2 Enhances Cell Motility and Promotes Tumor Metastasis in Lung Adenocarcinoma by Inducing Epithelial-Mesenchymal Transition. *Clin Cancer Res* (2017) 23(18):5598–610. doi: 10.1158/1078-0432.CCR-16-2909
- Schabath MB, Welsh EA, Fulp WJ, Chen L, Teer JK, Thompson ZJ, et al. Differential Association of STK11 and TP53 With KRAS Mutation-Associated Gene Expression, Proliferation and Immune Surveillance in Lung Adenocarcinoma. *Oncogene* (2016) 35(24):3209–16. doi: 10.1038/ncr.2015.375
- Welsh EA, Eschrich SA, Berglund AE, Fenstermacher DA. Iterative Rank-Order Normalization of Gene Expression Microarray Data. *BMC Bioinf* (2013) 14:153. doi: 10.1186/1471-2105-14-153
- Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, Iwakawa R, et al. Identification of Genes Upregulated in ALK-Positive and EGFR/KRAS/ALK-Negative Lung Adenocarcinomas. *Cancer Res* (2012) 72(1):100–11. doi: 10.1158/0008-5472.CAN-11-1403
- Hubbell E, Liu WM, Mei R. Robust Estimators for Expression Analysis. *Bioinformatics* (2002) 18(12):1585–92. doi: 10.1093/bioinformatics/18.12.1585
- McCarthy DJ, Chen Y, Smyth GK. Differential Expression Analysis of Multifactor RNA-Seq Experiments With Respect to Biological Variation. *Nucleic Acids Res* (2012) 40(10):4288–97. doi: 10.1093/nar/gks042
- Tibshirani R. The Lasso Method for Variable Selection in the Cox Model. *Stat Med* (1997) 16(4):385–95. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::Aid-sim380>3.0.Co;2-3
- Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw* (2011) 39(5):1–13. doi: 10.18637/jss.v039.i05
- Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and Comparing Time-Dependent Areas Under Receiver Operating Characteristic Curves for Censored Event Times With Competing Risks. *Stat Med* (2013) 32(30):5381–97. doi: 10.1002/sim.5958
- Yoshihara K, Shahmoradgol M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring Tumour Purity and Stromal and Immune Cell Admixture From Expression Data. *Nat Commun* (2013) 4:2612. doi: 10.1038/ncomms3612
- Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-Cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep* (2017) 18(1):248–62. doi: 10.1016/j.celrep.2016.12.019
- Fu J, Li K, Zhang W, Wan C, Zhang J, Jiang P, et al. Large-Scale Public Data Reuse to Model Immunotherapy Response and Resistance. *Genome Med* (2020) 12(1):21. doi: 10.1186/s13073-020-0721-z
- Gao F, Zhu H-K, Zhu Y-B, Shan Q-N, Ling Q, Wei X-Y, et al. Predictive Value of Tumor Markers in Patients With Recurrent Hepatocellular Carcinoma in

- Different Vascular Invasion Pattern. *Hepatobiliary Pancreatic Dis Int* (2016) 15(4):371–7. doi: 10.1016/s1499-3872(16)60095-4
41. Nijmeh J, El-Chemaly S, Henske EP. Emerging Biomarkers of Lymphangioleiomyomatosis. *Expert Rev Respir Med* (2018) 12(2):95–102. doi: 10.1080/17476348.2018.1409622
 42. Lin BC, Desnoyers LR. FGF19 and Cancer. *Adv Exp Med Biol* (2012) 728:183–94. doi: 10.1007/978-1-4614-0887-1_12
 43. Rodriguez-Lara V, Avila-Costa MR. An Overview of Lung Cancer in Women and the Impact of Estrogen in Lung Carcinogenesis and Lung Cancer Treatment. *Front Med (Lausanne)* (2021) 8:600121. doi: 10.3389/fmed.2021.600121
 44. Marcu A, Bichmann L, Kuchenbecker L, Kowalewski DJ, Freudenmann LK, Backert L, et al. HLA Ligand Atlas: A Benign Reference of HLA-Presented Peptides to Improve T-Cell-Based Cancer Immunotherapy. *J Immunother Cancer* (2021) 9(4):e002071. doi: 10.1136/jitc-2020-002071
 45. Sun J, Zhao T, Zhao D, Qi X, Bao X, Shi R, et al. Development and Validation of a Hypoxia-Related Gene Signature to Predict Overall Survival in Early-Stage Lung Adenocarcinoma Patients. *Ther Adv Med Oncol* (2020) 12:1758835920937904. doi: 10.1177/1758835920937904
 46. Wu J, Li L, Zhang H, Zhao Y, Zhang H, Wu S, et al. A Risk Model Developed Based on Tumor Microenvironment Predicts Overall Survival and Associates With Tumor Immunity of Patients With Lung Adenocarcinoma. *Oncogene* (2021) 40(26):4413–24. doi: 10.1038/s41388-021-01853-y
 47. Huang Z, Ma L, Huang C, Li Q, Nice EC. Proteomic Profiling of Human Plasma for Cancer Biomarker Discovery. *Proteomics* (2017) 17(6):1600240. doi: 10.1002/pmic.201600240
 48. Rieckmann JC, Geiger R, Hornburg D, Wolf T, Kveler K, Jarrossay D, et al. Social Network Architecture of Human Immune Cells Unveiled by Quantitative Proteomics. *Nat Immunol* (2017) 18(5):583–93. doi: 10.1038/ni.3693
 49. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-Based Map of the Human Proteome. *Science* (2015) 347(6220):1260419. doi: 10.1126/science.1260419
 50. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. Feature-Based Prediction of Non-Classical and Leaderless Protein Secretion. *Protein Eng Des Sel* (2004) 17(4):349–56. doi: 10.1093/protein/gzh037
 51. Sun J, Xie T, Jamal M, Tu Z, Li X, Wu Y, et al. CLEC3B as a Potential Diagnostic and Prognostic Biomarker in Lung Cancer and Association With the Immune Microenvironment. *Cancer Cell Int* (2020) 20:106. doi: 10.1186/s12935-020-01183-1
 52. Maguire PB, Parsons ME, Szklanna PB, Zdanyte M, Munzer P, Chatterjee M, et al. Comparative Platelet Releasate Proteomic Profiling of Acute Coronary Syndrome Versus Stable Coronary Artery Disease. *Front Cardiovasc Med* (2020) 7:101. doi: 10.3389/fcvm.2020.00101
 53. Li Y, Hou G, Zhou H, Wang Y, Tun HM, Zhu A, et al. Multi-Platform Omics Analysis Reveals Molecular Signature for COVID-19 Pathogenesis, Prognosis and Drug Target Discovery. *Signal Transduct Target Ther* (2021) 6(1):155. doi: 10.1038/s41392-021-00508-4
 54. Yan S, Ding J, Zhang Y, Wang J, Zhang S, Yin T, et al. C1QTNF6 Participates in the Pathogenesis of PCOS by Affecting the Inflammatory Response of Granulosa Cellsdouble Dagger. *Biol Reprod* (2021) 105(2):427–38. doi: 10.1093/biolre/iaab094
 55. Zhang W, Feng G. C1QTNF6 Regulates Cell Proliferation and Apoptosis of NSCLC *In Vitro* and *In Vivo*. *Biosci Rep* (2021) 41(1):BSR20201541. doi: 10.1042/BSR20201541
 56. Chong LY, Cheok PY, Tan WJ, Thihe AA, Allen G, Ang MK, et al. Keratin 15, Transcobalamin I and Homeobox Gene Hox-B13 Expression in Breast Phyllodes Tumors: Novel Markers in Biological Classification. *Breast Cancer Res Treat* (2012) 132(1):143–51. doi: 10.1007/s10549-011-1555-6
 57. Liu GJ, Wang YJ, Yue M, Zhao LM, Guo YD, Liu YP, et al. High Expression of TCN1 is a Negative Prognostic Biomarker and can Predict Neoadjuvant Chemosensitivity of Colon Cancer. *Sci Rep* (2020) 10(1):11951. doi: 10.1038/s41598-020-68150-8
 58. Maglinger B, Frank JA, McLouth CJ, Trout AL, Roberts JM, Grupke S, et al. Proteomic Changes in Intracranial Blood During Human Ischemic Stroke. *J Neurointerv Surg* (2021) 13(4):395–9. doi: 10.1136/neurintsurg-2020-016118
 59. Vollmy F, van den Toorn H, Zenezini Chiozzi R, Zucchetti O, Papi A, Volta CA, et al. A Serum Proteome Signature to Predict Mortality in Severe COVID-19 Patients. *Life Sci Alliance* (2021) 4(9):e202101099. doi: 10.26508/lsa.202101099
 60. Wang X, Wei W, Krzeszinski JY, Wang Y, Wan Y. A Liver-Bone Endocrine Relay by IGFBP1 Promotes Osteoclastogenesis and Mediates FGF21-Induced Bone Resorption. *Cell Metab* (2015) 22(5):811–24. doi: 10.1016/j.cmet.2015.09.010
 61. Marucci A, Antonucci A, De Bonis C, Mangiacotti D, Scarale MG, Trischitta V, et al. GALNT2 as a Novel Modulator of Adipogenesis and Adipocyte Insulin Signaling. *Int J Obes* (2019) 43(12):2448–57. doi: 10.1038/s41366-019-0367-3
 62. Yang L, Wang Y, Sun R, Zhang Y, Fu Y, Zheng Z, et al. ANGPTL4 Promotes the Proliferation of Papillary Thyroid Cancer via AKT Pathway. *Oncol Targets Ther* (2020) 13:2299–309. doi: 10.2147/OTT.S237751
 63. Etkebest-Mitxelorena M, Del Rincon-Loza I, Martin-Antonio B. Tumor Secretome to Adoptive Cellular Immunotherapy: Reduce Me Before I Make You My Partner. *Front Immunol* (2021) 12:717850. doi: 10.3389/fimmu.2021.717850
 64. Makridakis M, Vlahou A. Secretome Proteomics for Discovery of Cancer Biomarkers. *J Proteomics* (2010) 73(12):2291–305. doi: 10.1016/j.jpro.2010.07.001
 65. Lin LL, Huang HC, Juan HF. Discovery of Biomarkers for Gastric Cancer: A Proteomics Approach. *J Proteomics* (2012) 75(11):3081–97. doi: 10.1016/j.jpro.2012.03.046
 66. Caram ME, Skolarus TA, Cooney KA. Limitations of Prostate-Specific Antigen Testing After a Prostate Cancer Diagnosis. *Eur Urol* (2016) 70(2):209–10. doi: 10.1016/j.eururo.2015.12.045
 67. Kasprzak A, Siodla E, Andrzejewska M, Szmaja J, Seraszek-Jaros A, Cofta S, et al. Differential Expression of Mucin 1 and Mucin 2 in Colorectal Cancer. *World J Gastroenterol* (2018) 24(36):4164–77. doi: 10.3748/wjg.v24.i36.4164
 68. Baratto L, Duan H, Macke H, Iagaru A. Imaging the Distribution of Gastrin-Releasing Peptide Receptors in Cancer. *J Nucl Med* (2020) 61(6):792–8. doi: 10.2967/jnumed.119.234971
 69. Li Z, Li H, Deng D, Liu R, Lv Y. Mass Spectrometric Assay of Alpha-Fetoprotein Isoforms for Accurate Serological Evaluation. *Anal Chem* (2020) 92(7):4807–13. doi: 10.1021/acs.analchem.9b03995
 70. Giampazolias E, Schulz O, Lim KHJ, Rogers NC, Chakravarty P, Srinivasan N, et al. Secreted Gelsolin Inhibits DNGR-1-Dependent Cross-Presentation and Cancer Immunity. *Cell* (2021) 184(15):4016–31.e4022. doi: 10.1016/j.cell.2021.05.021
 71. Noy R, Pollard JW. Tumor-Associated Macrophages: From Mechanisms to Therapy. *Immunity* (2014) 41(1):49–61. doi: 10.1016/j.immuni.2014.06.010
 72. Battram AM, Bachiller M, Martin-Antonio B. Senescence in the Development and Response to Cancer With Immunotherapy: A Double-Edged Sword. *Int J Mol Sci* (2020) 21(12):4346. doi: 10.3390/ijms21124346
 73. Waldman AD, Fritz JM, Lenardo MJ. A Guide to Cancer Immunotherapy: From T Cell Basic Science to Clinical Practice. *Nat Rev Immunol* (2020) 20(11):651–68. doi: 10.1038/s41577-020-0306-5
 74. Goswami KK, Ghosh T, Ghosh S, Sarkar M, Bose A, Baral R. Tumor Promoting Role of Anti-Tumor Macrophages in Tumor Microenvironment. *Cell Immunol* (2017) 316:1–10. doi: 10.1016/j.cellimm.2017.04.005
 75. Mantovani A, Sica A, Sozzani S, Allavena P, Vecchi A, Locati M. The Chemokine System in Diverse Forms of Macrophage Activation and Polarization. *Trends Immunol* (2004) 25(12):677–86. doi: 10.1016/j.it.2004.09.015

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.

Copyright © 2022 Chen, Zhang, Li, Xiao, Feng, Niu, Zhao, Ma and Ye. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Prediction of Transcription Factor Binding Sites Using a Combined Deep Learning Approach

Linan Cao[†], Pei Liu[†], Jialong Chen[†] and Lei Deng^{*}

School of Computer Science and Engineering, Central South University, Changsha, China

OPEN ACCESS

Edited by:

Rui Guo,
Harvard Medical School, United States

Reviewed by:

Yongchun Zuo,
Inner Mongolia University, China
Wei Chen,
North China University of Science and
Technology, China

*Correspondence:

Lei Deng
leideng@csu.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 10 March 2022

Accepted: 11 April 2022

Published: 03 June 2022

Citation:

Cao L, Liu P, Chen J and Deng L
(2022) Prediction of Transcription
Factor Binding Sites Using a
Combined Deep Learning Approach.
Front. Oncol. 12:893520.
doi: 10.3389/fonc.2022.893520

In the process of regulating gene expression and evolution, such as DNA replication and mRNA transcription, the binding of transcription factors (TFs) to TF binding sites (TFBS) plays a vital role. Precisely modeling the specificity of genes and searching for TFBS are helpful to explore the mechanism of cell expression. In recent years, computational and deep learning methods searching for TFBS have become an active field of research. However, existing methods generally cannot meet high performance and interpretability simultaneously. Here, we develop an accurate and interpretable attention-based hybrid approach, DeepARC, that combines a convolutional neural network (CNN) and recurrent neural network (RNN) to predict TFBS. DeepARC employs a positional embedding method to extract the hidden embedding from DNA sequences, including the positional information from OneHot encoding and the distributed embedding from DNA2Vec. DeepARC feeds the positional embedding of the DNA sequence into a CNN-BiLSTM-Attention-based framework to complete the task of finding the motif. Taking advantage of the attention mechanism, DeepARC can gain greater access to valuable information about the motif and bring interpretability to the work of searching for motifs through the attention weight graph. Moreover, DeepARC achieves promising performances with an average area under the receiver operating characteristic curve (AUC) score of 0.908 on five cell lines (A549, GM12878, Hep-G2, H1-hESC, and Hela) in the benchmark dataset. We also compare the positional embedding with OneHot and DNA2Vec and gain a competitive advantage.

Keywords: transcription factor binding sites, attention mechanism, positional embedding, deep learning, DNA

INTRODUCTION

The interaction between protein and DNA plays a pivotal role in *in vitro* life activities, such as mRNA transcription, DNA replication, and immune response (1). Transcription factors (TFs) are proteins that bind to regulatory DNA sequences and mediate gene expression. TF binding sites (TFBSs), also called motifs, typically range from a few to about 20 base pairs (bps) and are a type of DNA functional site. TF binds specifically to TFBS. Accurately finding the TFBS in the DNA sequence is essential for deciphering the mechanism of gene expression and understanding the life expression *in vitro* and drug design (2).

Studying the characteristics of TFBSs is a process of searching for subsequences with binding characteristics from the massive DNA sequence data. Unfortunately, traditional biological experiments are not only challenging to process massive amounts of data but also expensive and time-consuming. With the development of high-throughput technology, massive amounts of reliable experimental data can be obtained through *in vitro* experiments. These data contain potential TFBS sequences and provide convenience for obtaining TFBSs based on computational methods (3–5). MEME (4) searches for TFBS in DNA sequences by scoring the DNA sequences and then recursively selecting the sequences most likely to have motifs. AlignACE (5) computes possible sequences of TFBS based on Gibbs sampling. The common point of these algorithms is to use ChIP-seq high-throughput experimental data and statistical calculation methods to find potential TFBS, which has the characteristics of a large deviation of calculation accuracy. Because high-throughput experiments cannot accurately find the DNA subsequences where TFBSs are located under high-precision requirements, some sequence-based feature extraction methods have been proposed to solve the first step in motif searching. In past decades, position weight matrix (PWM) (1), OneHot (6), and K-mer (7) are all DNA representation methods that have achieved good results.

For the past few years, deep learning methods have been widely applied quite in many fields like computer vision, natural language processing, and speech recognition, and these fields have achieved good results, etc. (8–10). Predicting the interaction of biological sequences such as DNA/RNA sequences and protein sequences, as a new subject, has continuously been a very active research field, in which deep learning also plays a decisive role (11–19). Deep learning approaches can learn features from large amounts of data. DeepBind (6) is an earlier deep learning-based model in the field of gene sequencing. It miraculously adopts CNN to extract gene features predicted by protein binding sites, thus reshaping the entire era of using convolution kernels to capture features. In (12, 14), by fine-tuning the network architecture of CNN, the validity of various networks to verify TFBS has been evaluated in terms of overall. DanQ (13), the one who tactfully used long short-term memory (LSTM) to improve the before-and-after dependency in gene features, further enhanced the performance in the task of quantifying gene sequence functions.

Although the methods based on deep learning have achieved significant results in discovering TFBS, at this stage, a more in-depth and comprehensive application still needs great improvement: 1) accurately embedding the DNA sequence has been decisive to promote the model's performance. In previous studies, the traditional method such as OneHot (6) for encoding has been proposed as a promising, relatively achieved good performance, but it is difficult to improve due to its explosion of the consumption of computing resources when the OneHot embedding size increases. In (20), the NLP method has many applications in the field of DNA sequence and realizes distributed embedding representation. However, it was found that the position information contained in the DNA sequence

was lost during use. 2) It has proven effective as an emerging method at predicting capabilities in successfully applying the attention mechanism for NLP. However, there is still an uneasy process with multiple challenges that need to be addressed before acquiring practical application potential, such as limited knowledge of an outstanding method to integrate it into the field of genes. In this work, we develop a combined deep learning approach that uses OneHot and DNA2Vec embedding to extract the hidden embedding from DNA sequences and apply CNN and bidirectional LSTM network (BiLSTM) with an attention mechanism to build the prediction model. Experimental results show that our proposed method predicts better than existing state-of-the-art methods and has good interpretability.

MATERIALS AND METHODS

Datasets

High-throughput experiments produce a mass of protein-DNA binding datasets. We use ENCODE (Encyclopedia of DNA Elements), which offers TF cell type binding data analyzed by the ChIPseq method (21) to train and test our model. Zeng's works (12) have completed the preprocessing part of the work. In the preprocessing work, the positive samples consisting of 101 bps were generated in the central region of each ChIP-seq peak. The negative sample is obtained by recombining the positive sequence with the matching length. We distinguish positive samples from negative samples based on whether TFBS can be found in the sequence. So the positive samples represent TFBSs, while negative samples do not have binding sites with a TF in the sequence. In this study, we adopted 50 datasets that were selected at random from 690 ChIP-seq datasets, including five cell lines (GM12878, H1-hESC, Hep-G2, HeLa, and A549) as training sets and testing sets to measure the model performance. Of these data, 60% are used as the training set, 30% as the test set, and 10% as the verification set. In this article, our method runs as follows: first, we embed each DNA sample to get the position information and DNA sequence content features at the same time. Then we feed the DNA embedding to the attention-based model to get the final prediction.

Problem Statements

The problem of TFBS prediction can be expressed as follows. First of all, we divided all gene sequences into two categories based on whether TFBS could be found in the DNA sequence. The two categories are represented by label 0 or 1, which means that there is no TFBS or TFBS in the gene sequences, respectively. The embedded DNA sequences are expressed by $\{X^{(i)}, y^{(i)}\}_{i=1}^n$ and input into the model where $X^{(i)}$ is the input DNA sequence data and $y^{(i)}$ shows the type of gene sequence. After that, we train the model DeepARC (Figure 1) on the training sets. Our goal is to obtain high-accuracy classification results in the testing sets and extract the consistent sequence features from a tremendous amount of gene information.

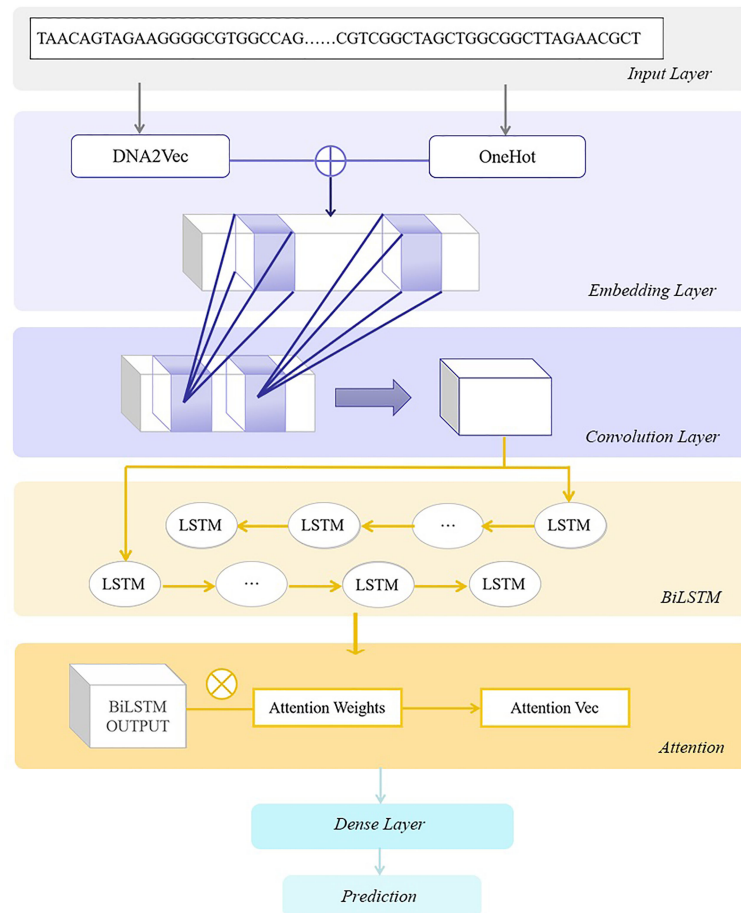


FIGURE 1 | The architecture of DeepARC. In the embedding layer, OneHot and k-mer encoding are used to generate position-based feature embedding from DNA sequences. Then convolution kernels are utilized to extract non-linear features. In the BiLSTM layer, we use a bidirectional long short-term memory network (BiLSTM) to capture the contextual dependencies of DNA sequences. Next, we use the attention mechanism to enhance the model's prediction performance, and finally, the prediction results are obtained through the dense layer.

Positional Embedding

1) OneHot encoding, also called one-bit efficient encoding, uses the N-bit status register to encode N states, each of which has its own independent register bit, and at any given time, only one bit in the code is valid and can be used to map characters to a unique encoding. As a simple and effective coding method, OneHot encoding has been widely used for indicating the state of a state machine, and there are many applications in bioinformatics and natural language processing (6, 14). In DeepBind, each fragment of gene sequence is regarded as a feature and encoded by OneHot in a special way. However, there are some drawbacks to the OneHot encoding. Due to the simple and sparse characteristics of OneHot encoding and the assumption that different features are independent, the mutual relationship between different coding units will be lost, and the distance relationship between coding units will not be reflected. For data with some kind of continuous relationship, encoding with the OneHot method may result in a situation where the accuracy rate will be significantly reduced. In

addition, the OneHot encoding dimension of each word is the size of the entire vocabulary. With the growth of embedded data, the dimension will become huge, and the coding will turn sparse, which will make the calculation cost very terrible.

2) In the field of molecular biology, mer represents a monomeric unit, and k-mer means a set of nucleotide strings with a length of k. Extracting k-mer from L-length DNA sequence can generate $L - k + 1$ fragments, and the association between different sequences after k-mer division can be preserved in these fragments. In WSCNNLSTM (17), instead of using OneHot coding, k-mer features of sequences are extracted. The association between sequences is still maintained in k-mer after the gene sequence is divided into k-mer. Therefore, WSCNNLSTM has better performance in TFBS classification work. A new method is proposed in DNA2Vec that can calculate the distribution representations of k-mer with variable length (20), which apply to NLP to biological sequence information.

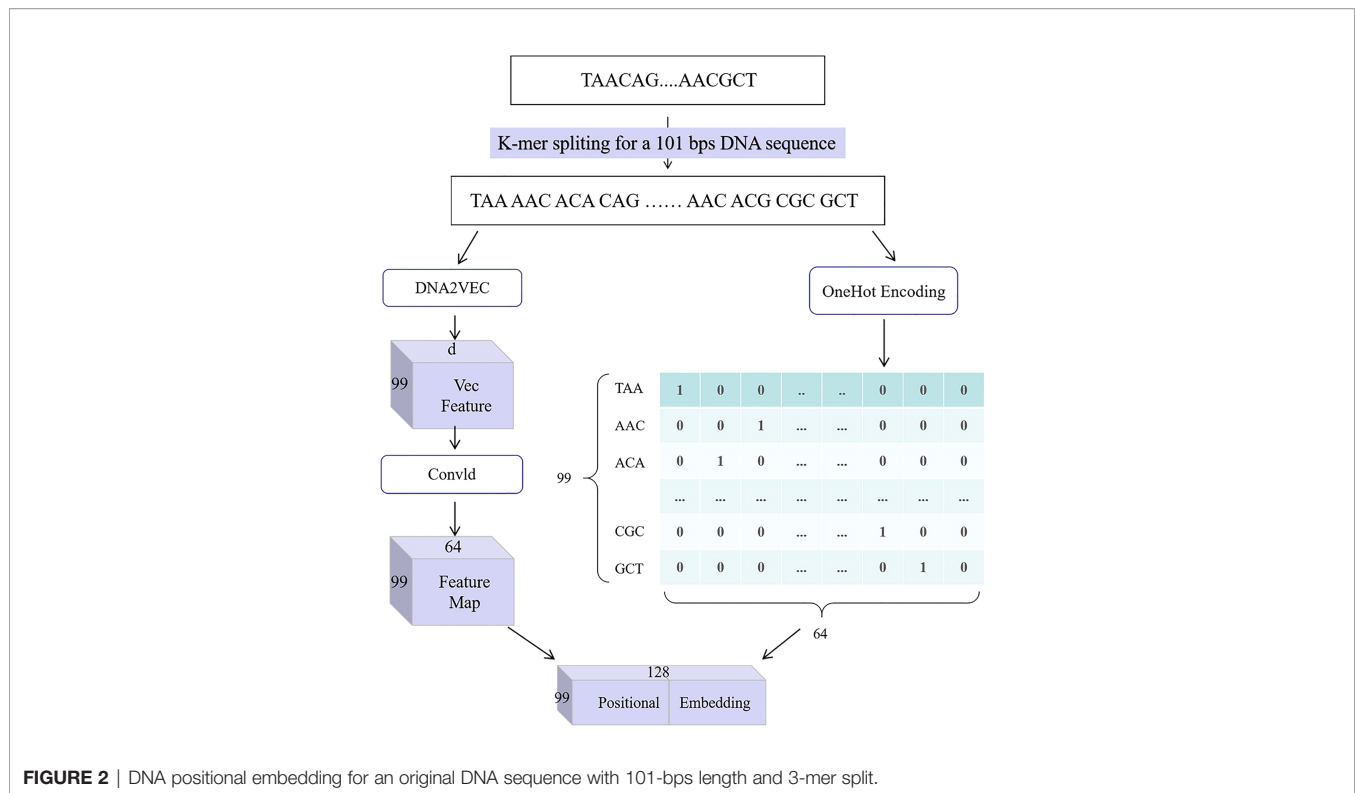


FIGURE 2 | DNA positional embedding for an original DNA sequence with 101-bps length and 3-mer split.

3) In this paper, we present a gene position embedding method, which transforms gene sequence into a characteristic matrix, as shown in **Figure 2**. It uses a combination of the OneHot encoding method, the DNA2Vec method, and the convolutional module. The workflow of this method is as follows: first, the input gene sequence, $S = (s_1, s_2, \dots, s_l)$, is divided into $L-k+1$ sequence features by k -mer cutting method, which is called $Z_{mer} = (z_1, z_2, \dots, z_{L-k+1})$. Next, we use OneHot encoding on the mer to get the unique location information called, as well as adopt DNA2Vec to get the context feature called $Z_{vec} \in \mathbb{R}^{(L-k+1) \times d}$, where d is the dimension of word embedding in the DNA2Vec. Because the feature dimension obtained by DNA2Vec is too high, we try to decrease the word embedding dimension to 4^k by extracting features through a convolution module. Finally, we linked the high-order dependent features Z_{OH} and Z_{vec} to obtain DNA position embeddings $Z_{pe} \in \mathbb{R}^{(L-k+1) \times (2 \times 4^k)}$.

Convolutional Neural Network

The convolutional neural network (CNN) is a sort of feedforward neural network with convolution calculation and depth architecture (18). It is successfully used in image recognition, video analysis, natural language processing, drug discovery, and other fields and has achieved good results (22, 23). The working process of CNN is usually to input image information; then pass a battery of convolutional layers, non-linear layers, pooling layers, and complete connection layers; and then get the final output result. Among them, the convolutional layer mainly has the function of feature extraction through the scan of the convolutional kernel, while the pooling layer primarily plays the role of feature selection

and information filtering. Therefore, CNN greatly reduces network parameters and has translational invariant properties. In the field of bioinformatics, CNN was initially applied to deal with DNA sequence information in the DeepBind. After embedding the DNA sequence in some manner, it is disposed of in the shape of a graph in the network. CNN is able to extract multiple features through scanning different convolution kernels so that it can handle various downstream works.

Because of its strong feature extraction ability, CNN was used to capture TFBS in this study. The embedded gene sequence Z_{pe} was input into the CNN model to get the extracted feature C . In our experiment, there are two submodules in the CNN model, and each submodule is composed of a convolution layer and a non-linear activation layer. Among them, the convolutional layer mainly plays the function of detecting TFBS to obtain features similar to TF-motif. The parameters of the model are set as follows: the size of the convolution kernel is 5, the padding is 2, and the step size is 1. The purpose of such a setting is to make the dimension of feature C to represent the constant length of the sequence and to be able to learn the features of the whole DNA sequence through BiLSTM. In order to prevent over-fitting of the model, the ReLU function is used in the non-linear activation layer. Finally, we extracted feature C through the CNN module.

Bidirectional Long Short-Term Memory Networks

BiLSTM is a particular type of recurrent neural network (RNN), which has parameter sharing, Turing-complete, and memorability, so it has some advantages in learning non-linear

characteristics of sequences. Compared with RNN, BiLSTM can handle the long-term dependency problem existing in RNN and can realize the real context-based consideration, so it also has higher accuracy (24). In terms of structure, based on the traditional RNN model, LSTM also adds a gate structure of forgetting gate, input gate, and output gate to control the information passing through the model, through the gate structure to control the input and output information flow, so as to solve the problem of long-term dependence in RNN. BiLSTM combines forward LSTM and backward LSTM for accurate context analysis. The following is the operation formula of the LSTM memory unit:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \quad (1)$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \quad (2)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \quad (4)$$

$$h_t = o_t \odot \tanh(c_t) \quad (5)$$

where i , f , o , c , and h represent the input gate, forget gate cell vector and hidden vector, respectively. W is the gate matrix, and b is the bias. The index t refers to the time step σ is the logistic sigmoid function \tanh is the active function to force the values to be between -1 and 1, and \odot denotes element-wise multiplication.

The DNA sequence is a series of letters used to represent the actual or hypothetical primary structure of DNA molecules carrying genetic information, which can be considered the mystery of life's sequence language to some extent. The LSTM model has been introduced by DeeperBind (14) and DeepTF (25) to analyze the long-term dependence of DNA sequences. However, in this paper, we adopt BiLSTM to capture associations between successive gene sequences. BiLSTM is composed of forward propagating LSTM and backward propagating LSTM and can analyze forward and backward sequence information. Therefore, it has higher accuracy than LSTM. The input of the BiLSTM model is the feature generated by passing the convolution layer, and the output includes the output feature P and the hidden state information h_n . It is worth noting that the sum output by the forward model and the backward model of BiLSTM is the final output feature at the i th position, and the formula is as follows:

$$h_i = \left[\vec{h}_i \oplus \overleftarrow{h}_i \right] \quad (6)$$

Attention Mechanism

In essence, the attention mechanism in deep learning is analogous to the selective visual attention mechanism in mankind, and the major objective is to select more important information for the current work goal from massive details. The attention mechanism is a kind of resource allocation scheme,

which is the principal method to deal with the trouble of information overload, and it is very suitable in the case of limited computing power. It allocates computing resources to more critical work and improves the utilization of resources. The self-attention mechanism has been used by BERT (26) to train natural language and has also obtained excellent results in text classification, machine translation, and other works. Attention mechanism has been used in the field of deep learning far and wide and achieved good results in named entity recognition, machine translation, and other fields (8–10, 19). Therefore, a soft attention mechanism was adopted in our experiment to focus attention on the TFBS we were looking for. The feature P and the hidden state h_n after the BiLSTM module are used as the input of the soft attention mechanism. The mer-level feature is merged into a sentence-level feature vector to generate the attention-weight vector (27). Finally, the DNA attention vector for the classified prediction can be calculated. The formula is as follows:

$$M = \tanh(H), \quad (7)$$

$$\alpha = \text{softmax}(\omega^T M), \quad (8)$$

$$\gamma = H\alpha^T, \quad (9)$$

$$h^* = \tanh(\gamma) \quad (10)$$

where $H \in \mathbb{R}^{d^w \times T}$, d^w is the dimension of the word vectors, ω is a trained parameter vector, and ω^T is a transpose. The dimension of ω , α , and γ is d^w , T and d^w , respectively.

Dense Module

The dense module constituted by two layers of a fully connected neural network, one dropout layer core and one sigmoid function, is the last module of the whole model. The full connection layer mainly acts as a classifier to classify the input into several categories. However, since the full connection layer has too many parameters, we added a dropout layer to the back of the full connection layer to prevent the over-fitting of the model from improving the generalization ability of the model (28). We take the binary cross-entropy loss calculated by the prediction and the goal as the cost function of the model, and the formula is as follows:

$$\text{loss} = -\frac{1}{N} \sum_{n=1}^N [y_n \log x_n + (1 - y_n) \log (1 - x_n)] \quad (11)$$

where x_n is the prediction and y_n is the goal.

RESULTS

DeepARC is an attention mechanism-based model for predicting the presence or absence of TFBSs on gene sequences. In the experiment, we randomly selected 50 datasets from ENCODE to conduct model training. To demonstrate the advantages of the model architecture and location embedding used in this article,

we also compare it to similar approaches that are currently popular. So as to test the property of DeepARC, we will use the three most advanced algorithms in this field to carry out comparative experiments on the same dataset. In the following content, we will analyze the experimental results in detail. First, we introduce the evaluation indicators used in this experiment. Second, the advantages of our model and the advantages of our location-embedding approach are presented. Third, we mainly introduce the performance comparison of our method with existing excellent predictors. Finally, we explain the attention mechanism used in the article.

Evaluation Measurements

Due to the characteristics of this experiment, we decided to select five evaluation measurements—sensitivity (*Sen*), specificity (*Spe*), accuracy (*Acc*), Mathew's correlation coefficient (*MCC*), and the area under the receiver operating characteristic curve (*AUC*)—to evaluate the prediction ability of our model (29). Their formula is as follows:

$$Sen = TP / (TP + FN) \quad (12)$$

$$Spe = TN / (TN + FP) \quad (13)$$

$$Acc = (TP + TN) / (TP + FP + TN + FN) \quad (14)$$

$$MCC = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \quad (15)$$

where *TP*, *TN*, *FP*, and *FN* are the number of true positives, true negatives, false positives, and false negatives, respectively.

Performance Comparison With Other Model Frameworks

In this experiment, we used the data encoded by OneHot as input, adopted the Adam optimizer (30), and set the learning rate to 0.001 and the step size to 20 to show the performance of the model architecture. In addition, in order to reflect the excellence of the model mechanism, we compare the performance differences of CNN-BiLSTM, BiLSTM-Att, and DeepARC in the same input set. CNN-BiLSTM represents a model whose model architecture is CNN+BiLSTM, BiLSTM-ATT represents a model whose architecture is BiLSTM+attention mechanism, and CNN-BiLSTM-Att is the model architecture used by DeepARC. The results of the three cross-validation tests are considered to be the final model performance. **Table 1** shows the final performance comparison results for each model, and **Table 2** describes the other detailed parameter settings.

As shown in **Table 1**, in the five datasets, our model (CNN-BiLSTM-Att) generally has the best performance and always has the highest score in *Acc* value, *MCC* value, and *AUC* value. First of all, by comparing the performance of CNN-BiLSTM-Att and BiLSTM-Att on the five datasets, it can be found that except for the *Spe*, CNN-BiLSTM-Att has better performance on other evaluation values, namely, *Sen*, *Acc*, and *MCC* were 5.43%, 1.36%, and 1.4% higher, as compared with BiLSTM-Att. It can

be seen that CNN is used to find TFBS features and has a good effect. Next, we compared CNN-BiLSTM-Att with CNN-BiLSTM and found that CNN-BiLSTM-Att has *Sen*, *Acc*, and *MCC* of 0.22%, 0.95%, and 1.26% higher than CNN-BiLSTM on the 5 datasets. It can also be seen that the attention mechanism enhances the weight of the model on motif to effectively promote the performance of the model. Compared with other popular model frameworks, our model architecture obviously has higher performance and certain advantages. However, it can also be seen from the table that our model has a poor performance in *SPE*. The lower *SPE* may be due to the tendency of the model's predicted samples to be positive.

Performance Comparison Among Positional Embedding and Other Methods

In this part, we primarily analyze the property of positional embedding in the DNA embedding part. To more intuitively observe the advantages and disadvantages of the performance, we compare the positional embedding method with DNA2Vec and OneHot. On the basis of the research (7), we set 3-mer to implicitly capture the binding information, 3-mer splitting, and one stride in the embedding. Consistency is maintained by using the CNN-BiLSTM-Att in the previous section as the model for the experiment. Moreover, the hyperparameters and prediction methods of the three methods are consistent with the above methods. **Table 3** shows the experimental performance results.

As can be seen from **Table 3**, on the five datasets, our positional embedding method obviously has the best performance, always having the highest *AUC* value and *MCC* value in the DNA embedding methods. This means that our positional embedding method has significantly better performance than the current popular embedding methods. Beyond that, the OneHot method has better performance in the model on many evaluation values, namely, *Sen*, *Acc*, and *MCC* were 1.47%, 0.276%, and 0.6% higher than has DNA2Vec. However, the best performance is positional embedding, which is higher than the OneHot method in all the evaluation indexes of the five datasets, 1.14%, 2.664%, 1.7%, 0.39%, and 0.16% in *Sen*, *Spe*, *Acc*, *MCC*, and *AUC*, respectively.

In our opinion, the reason why the location embedding method can achieve better results is that it combines the advantages of OneHot encoding and DNA2Vec encoding. It has a distributed representation of the content encoded by DNA2Vec, as well as location information in the OneHot encoding. Therefore, with a suitable model, the position embedding method can show better performance.

Performance Comparison With Other Existing Predictors

In this part, in order to analyze the performance of DeepARC, we compare it with several other prediction methods (DeepTF, DeepBind, and CNN-Zeng) on 50 randomly selected datasets. The comparison results obtained are shown in **Table 4**. As can be seen from the table, among the four methods, DeepARC has the best performance in each evaluation measurement. In addition, compared with DeepTF, which has the best performance among

TABLE 1 | Performance comparison of CNN-BiLSTM, BiLSTM-Attention, and CNN-BiLSTM-Att with OneHot embedding.

Dataset	Model	Sen (%)	Spe (%)	Acc (%)	MCC	AUC
A549	CNN-BiLSTM-Att	80.66	83.66	82.16	0.644	0.901
	CNN-BiLSTM	79.42	82.61	81.01	0.625	0.896
	BiLSTM-Att	74.13	86.96	80.55	0.616	0.887
GM12878	CNN-BiLSTM-Att	81.05	83.02	82.04	0.641	0.902
	CNN-BiLSTM	75.63	86.51	81.07	0.625	0.891
	BiLSTM-Att	73.90	85.58	79.74	0.600	0.880
Hela	CNN-BiLSTM-Att	79.95	74.32	77.13	0.545	0.860
	CNN-BiLSTM	80.17	73.81	76.99	0.543	0.858
	BiLSTM-Att	76.32	75.94	76.13	0.524	0.845
Hep-G2	CNN-BiLSTM-Att	81.47	84.50	82.98	0.660	0.908
	CNN-BiLSTM	85.87	77.99	81.93	0.641	0.906
	BiLSTM-Att	76.78	86.18	81.48	0.634	0.897
H1-hESC	CNN-BiLSTM-Att	81.26	82.13	82.72	0.636	0.891
	CNN-BiLSTM	82.19	81.31	81.25	0.629	0.883
	BiLSTM-Att	76.11	81.52	82.32	0.612	0.876

CNN, convolutional neural network; BiLSTM, bidirectional long short-term memory network; Sen, sensitivity; Spe, specificity; Acc, accuracy; MCC, Mathew's correlation coefficient; AUC, area under the receiver operating characteristic curve.

The bold section indicates the best performing indicators in each dataset.

TABLE 2 | Parameters setting of different models.

	CNN	CNN	-
Parameter	BiLSTM	BiLSTM	BiLSTM
	Att	–	Att
Learning rate	0.001	0.001	0.001
Epochs	20	20	20
Batch size	64	64	64
CNN layers	2	2	–
Kernel size	5	5	–
BiLSTM hidden size	16	16	32
Attention vec size	16	–	32
Dense neurons	16	32	32
Dropout	0.2	0.2	0.2
Optimizer	Adam	Adam	Adam

CNN, convolutional neural network; BiLSTM, bidirectional long short-term memory network.

the other competitive two methods, DeepARC has higher Sen, Spe, Acc, and MCC evaluation indexes of 4.58%, 2.83%, 2.12%, and 3.2%, respectively. The average AUC values of each method

on five cell lines are shown in **Figure 3**. It can also be seen from **Figure 3** that DeepARC has the highest accuracy and is 1.8% higher than the second-place method DeepTF.

According to the results, DeepARC has better predictive performance on the dataset compared with other methods. By analyzing the model architecture of DeepARC and several other methods, we reasonably believe that the position embedding method and model architecture of DeepARC, especially the use of the attention mechanism, play a promoting role in the experiment, thus improving the prediction performance of the model.

Attention Mechanism Brings Interpretation

CNN is one of the representative deep learning algorithms that are widely used at present, and it has a robust feature learning ability. But because of the high complexity of its architecture, it is often difficult to understand and explain the decisions that these networks make (31–34). Therefore, a layer of attention

TABLE 3 | Performance comparison of OneHot, DNA2Vec, and positional embedding.

Dataset	Model	Sen (%)	Spe (%)	Acc (%)	MCC	AUC
A549	Positional embedding	83.17	83.74	83.46	0.671	0.909
	DNA2Vec	77.58	85.64	81.61	0.635	0.896
	OneHot	80.66	83.66	82.16	0.644	0.901
GM12878	Positional embedding	81.31	83.48	82.40	0.650	0.905
	DNA2Vec	80.81	81.25	81.03	0.623	0.895
	OneHot	81.05	83.02	82.04	0.641	0.902
Hela	Positional embedding	80.36	84.64	82.50	0.652	0.906
	DNA2Vec	77.53	75.48	76.51	0.534	0.853
	OneHot	79.95	74.32	77.13	0.545	0.860
Hep-G2	Positional embedding	82.25	86.51	84.38	0.690	0.919
	DNA2Vec	80.91	85.14	83.25	0.661	0.908
	OneHot	81.47	84.50	82.98	0.660	0.908
H1-hESC	Positional embedding	83.00	82.58	82.79	0.658	0.905
	DNA2Vec	80.19	81.31	83.25	0.639	0.896
	OneHot	81.26	82.13	82.72	0.636	0.891

Sen, sensitivity; Spe, specificity; Acc, accuracy; MCC, Mathew's correlation coefficient; AUC, area under the receiver operating characteristic curve.

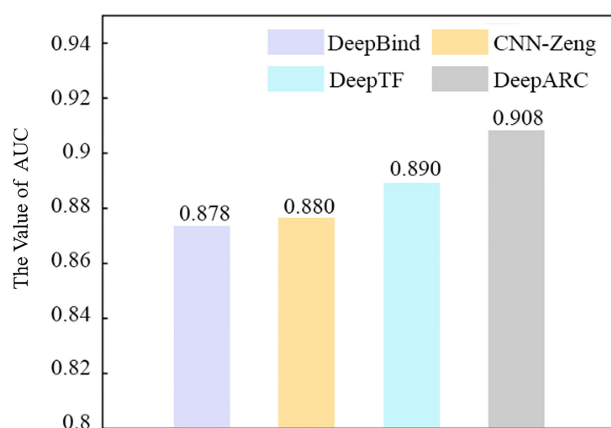
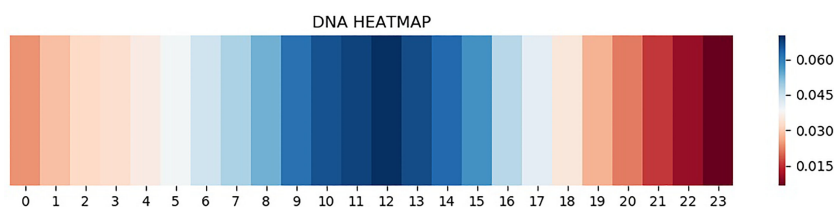
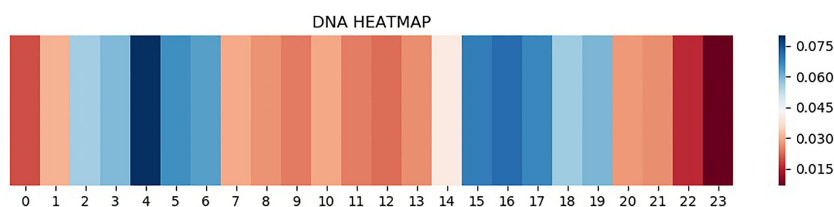
The bold section indicates the best performing indicators in each dataset.

TABLE 4 | Performance comparison of DeepARC and three existing predictors.

Model	Sen (%)	Spe (%)	Acc (%)	MCC
DeepARC	82.02	84.19	83.10	0.664
DeepTF	77.44	81.36	80.98	0.632
CNN-Zeng	72.12	81.96	79.92	0.619
DeepBind	72.64	81.44	79.82	0.609

Sen, sensitivity; Spe, specificity; Acc, accuracy; MCC, Mathew's correlation coefficient.

The bold part indicates the index with the best performance.

**FIGURE 3** | Performance of DeepARC and three existing predictors in ROC-AUC. ROC, receiver operating characteristic; AUC, area under the ROC curve.**FIGURE 4** | Heatmap of H1-hESC.**FIGURE 5** | Heatmap of A5493.

mechanism was added in DeepARC to enhance the weight of attention on the motif. In addition, we also visually display the average weights of different datasets in the model to strengthen the interpretability of the model (**Figures 4, 5**).

As can be seen from **Figure 4**, attention is a major concentration in the intermediate region. In other words, the model determines whether there is a TFBS in the input gene information mainly by sensing the peak value of the gene

sequence. The theory of peeling existing binding site sequences from the peak is consistent with this. However, there are two attention peaks shown in **Figure 5**, indicating that the model recognizes the presence of two TFBSs in the sequence.

CONCLUSIONS

In this work, we describe a novel attention-based network model named DeepARC to predict TFBSs. Driven by its beneficial strength of combining DNA2Vec and OneHot encoding, DeepARC could embed the gene information into a distributed positional representation and then predict the output using attention-based CNN-BiLSTM network architecture. The comparative work shows that DeepARC is superior to the existing state-of-the-art methods. To demonstrate the interpretability of DeepARC, we visualized the attention weights and found that the attention weight was concentrated in the peak region of the ChIP-seq. Although our method achieves good results, there is still room for improvement. DeepARC only uses DNA sequence information for feature embedding. Evolution information, physical-chemical properties, and embedding from language models can be integrated to improve performance in the future. On the other hand, the attention mechanism can be optimized to mark the accurate TFBS fragments directly.

REFERENCES

1. Stormo GD. Dna Binding Sites: Representation and Discovery. *Bioinformatics* (2000) 16:16–23. doi: 10.1093/bioinformatics/16.1.16
2. Stormo GD. Consensus Patterns in Dna. *Methods Enzymol* (1990) 183:211–21. doi: 10.1016/0076-6879(90)83015-2
3. Huang H, Kao MC, Zhou X, Liu JS, Wong WH. Determination of Local Statistical Significance of Patterns in Markov Sequences With Application to Promoter Element Identification. *J Comput Biol* (2004) 11:1–14. doi: 10.1089/106652704773416858
4. Bailey TL, Mikael B, Buske FA, Martin F, Grant CE, Luca C, et al. Meme Suite: Tools for Motif Discovery and Searching. *Nucleic Acids Res* (2009) 37:W202–8. doi: 10.1093/nar/gkp335
5. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational Identification of Cis-Regulatory Elements Associated With Groups of Functionally Related Genes in *Saccharomyces Cerevisiae*. *J Mol Biol* (2000) 296:1205–14. doi: 10.1006/jmbi.2000.3519
6. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the Sequence Specificities of Dna-and Rna-Binding Proteins by Deep Learning. *Nat Biotechnol* (2015) 33:831–8. doi: 10.1038/nbt.3300
7. Deng L, Wu H, Liu H. D2vcb: A Hybrid Deep Neural Network for the Prediction of In-Vivo Protein-Dna Binding From Combined Dna Sequence. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE)*, (2019) 74–77. doi: 10.1109/BIBM47256.2019.8983051
8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. In: I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, editors. *Advances in Neural Information Processing Systems*, vol. 30. Long Beach, California, USA: Curran Associates, Inc (2017)
9. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Show, Attend and Tell: Neural Image Caption Generation With Visual Attention. *Comput Sci* (2015) 37:2048–57. doi: 10.48550/arXiv.1502.03044
10. Bahdanau D, Chorowski J, Serdyuk D, Brakel P, Bengio Y. End-to-End Attention-Based Large Vocabulary Speech Recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (2016) 4945–9. doi: 10.1109/ICASSP.2016.7472618

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

Conceptualization: LC, JC, and LD. Methodology: LC, PL, JC, and LD. Validation: LC, JC, and LD. Writing—original draft preparation, LC, PL, and JC. Writing—review and editing: LD. Supervision: LD. Project administration: LD. Funding acquisition, LD. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China under grant No. 61972422 and No. 61672541. Publication costs are funded by the National Natural Science Foundation of China under grant No. 61972422. The funding body has not played any role in the design of the study and collection, analysis, and interpretation of data in writing the manuscript.

11. Zhou J, Troyanskaya OG. Predicting Effects of Noncoding Variants With Deep Learning-Based Sequence Model. *Nat Methods* (2015) 12:931–4. doi: 10.1038/nmeth.3547
12. Zeng H, Edwards M, Liu G, Gifford DK. Convolutional Neural Network Architectures for Predicting Dna-Protein Binding. *Bioinformatics* (2016) 32: i121–7. doi: 10.1093/bioinformatics/btw255
13. Quang D, Xie X. Danq: A Hybrid Convolutional and Recurrent Deep Neural Network for Quantifying the Function of Dna Sequences. *Nucleic Acids Res* (2016) 44:e107. doi: 10.1093/nar/gkw226
14. Hassanzadeh HR, Wang MD. Deeperbind: Enhancing Prediction of Sequence Specificities of Dna Binding Proteins. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE)*, (2016) 178–183. doi: 10.1109/BIBM.2016.7822515
15. Chen C, Hou J, Shi X, Yang H, Birchler JA, Cheng J, et al. Interpretable Attention Model in Transcription Factor Binding Site Prediction With Deep Neural Networks. *bioRxiv*, (2019) 648691. doi: 10.1101/648691
16. Chauhan S, Ahmad S. Enabling Full-Length Evolutionary Profiles Based Deep Convolutional Neural Network for Predicting Dna-Binding Proteins From Sequence. *Proteins: Struct Func Bioinform* (2020) 88:15–30.
17. Zhang Q, Shen Z, Huang D-S. Modeling In-Vivo Protein-Dna Binding by Combining Multiple-Instance Learning With a Hybrid Deep Neural Network. *Sci Rep* (2019) 9:1–12.
18. Gupta A, Rush AM. Dilated Convolutions for Modeling Long-Distance Genomic Dependencies. *arXiv preprint* (2017) arXiv:1710.01278
19. Park S, Koh Y, Jeon H, Kim H, Yeo Y, Kang J, et al. Enhancing the Interpretability of Transcription Factor Binding Site Prediction Using Attention Mechanism. *Sci Rep* (2020) 10:1–10.
20. Ng P. Dna2vec: Consistent Vector Representations of Variable-Length K-Mers. *arXiv preprint* (2017).
21. The ENCODE Project Consortium. An Integrated Encyclopedia of Dna Elements in the Human Genome. *Nature* (2012) 489:57.
22. Sutskever I, Vinyals O, Le QV. Sequence to Sequence Learning With Neural Networks. In: Z Ghahramani, M Welling, C Cortes, N Lawrence, KQ Weinberger, editors. *Advances in Neural Information Processing Systems*, vol. 27. Montreal, Canada: Curran Associates, Inc (2014)

23. Sun Y, Wang X, Tang X. Deep Learning Face Representation From Predicting 10,000 Classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2014) 1891–1898.
24. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput* (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735
25. Bao XR, Zhu YH, Yu DJ. (2019) Deeptf: Accurate Prediction of Transcription Factor Binding Sites by Combining Multi-Scale Convolution and Long Short-Term Memory Neural Network, in: nanjing, China: *International Conference on Intelligent Science and Big Data Engineering*. pp. 126–38. Springer.
26. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint* (2018) 4171–86.
27. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (2016) (volume 2: Short papers). 207–212.
28. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *J Mach Learn Res* (2014) 15:1929–58. doi: 10.5555/2627435.2670313
29. Trabelsi A, Chaabane M, Ben-Hur A. Comprehensive Evaluation of Deep Learning Architectures for Prediction of Dna/Rna Sequence Binding Specificities. *Bioinformatics* (2019) 35:i269–77. doi: 10.1093/bioinformatics/btz339
30. Kingma D, Ba J. Adam: A Method for Stochastic Optimization, in: San Diego, CA, USA: *International Conference on Learning Representations* (2015) doi: 10.48550/arXiv.1412.6980
31. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, in: Banff, Canada: *International Conference on Learning Representations* (2013) doi: 10.48550/arXiv.1312.6034
32. Singh R, Lanchantin J, Sekhon A, Qi Y. Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin. *Adv Neural Inf Process Syst* (2017) 30:6785–95. doi: 10.1101/329334
33. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding Neural Networks Through Deep Visualization, in: Vauban: *Deep Learning Workshop, International Conference on Machine Learning* (2015) doi: 10.48550/arXiv.1506.06579
34. Lanchantin J, Singh R, Wang B, Qi Y. Deep Motif Dashboard: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks. In *Pacific Symposium on Biocomputing 2017 (World Scientific)*, (2017) 22:254–65.

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.

Copyright © 2022 Cao, Liu, Chen and Deng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Heterogeneity Analysis of Bladder Cancer Based on DNA Methylation Molecular Profiling

Shuyu Wang^{1†}, Dali Xu^{1†}, Bo Gao², Shuhan Yan¹, Yiwei Sun¹, Xinxing Tang¹, Yanjia Jiao¹, Shan Huang^{3*} and Shumei Zhang^{1*}

¹ College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, ² Department of Radiology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China, ³ Department of Neurology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:

Rui Guo,
Brigham and Women's Hospital and
Harvard Medical School, United States

Reviewed by:

Ran Su,
Tianjin University, China
Hongmin Cai,
South China University of Technology,
China

*Correspondence:

Shumei Zhang
zhangshumei@nefu.edu.cn
Shan Huang
hmhuangshan@163.com

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 08 April 2022

Accepted: 13 May 2022

Published: 07 June 2022

Citation:

Wang S, Xu D, Gao B, Yan S, Sun Y,
Tang X, Jiao Y, Huang S and Zhang S
(2022) Heterogeneity Analysis of
Bladder Cancer Based on DNA
Methylation Molecular Profiling.
Front. Oncol. 12:915542.
doi: 10.3389/fonc.2022.915542

Bladder cancer is a highly complex and heterogeneous malignancy. Tumor heterogeneity is a barrier to effective diagnosis and treatment of bladder cancer. Human carcinogenesis is closely related to abnormal gene expression, and DNA methylation is an important regulatory factor of gene expression. Therefore, it is of great significance for bladder cancer research to characterize tumor heterogeneity by integrating genetic and epigenetic characteristics. This study explored specific molecular subtypes based on DNA methylation status and identified subtype-specific characteristics using patient samples from the TCGA database with DNA methylation and gene expression were measured simultaneously. The results were validated using an independent cohort from GEO database. Four DNA methylation molecular subtypes of bladder cancer were obtained with different prognostic states. In addition, subtype-specific DNA methylation markers were identified using an information entropy-based algorithm to represent the unique molecular characteristics of the subtype and verified in the test set. The results of this study can provide an important reference for clinicians to make treatment decisions.

Keywords: bladder cancer, DNA methylation, molecular subtypes, subtype specific biomarkers, heterogeneity analysis

INTRODUCTION

In recent years, cancer has become an important killer of human health, and seriously threatens people's life and health. It is generally believed that cancer is caused by the accumulation of mutations in cancer susceptibility genes and resulting abnormal cell growth, but a large number of recent studies have shown that in addition to genetic variation, abnormal DNA methylation also plays an important role in the occurrence and development of cancer (1). DNA methylation is the most extensively documented epigenetic modification that can influence cell fate and gene expression (2, 3), which finally leads to the inhibition of gene expression through formation of heterochromatin in the gene regulatory region (4).

There are many studies have demonstrated the importance of DNA methylation (5–9). Numerous studies have shown that global hypomethylation of DNA and hypermethylation of cytosine-phosphate-guanine (CpG)-enriched regions are common in cancers (10–13). Methylation

of promoters inhibits gene transcription, and abnormal methylation is one of the main causes of genomic instability, oncogene activation and tumor suppressor gene suppression. For example, abnormal methylation in colorectal tumors is characterized by hypermethylation in promoters and transcriptional silencing of tumor suppressors or DNA repair genes (14–17), coexisting with global methylation loss that leads to chromosomal and microsatellite instability and oncogene activation (18). Both promoter hypermethylation and global hypomethylation are markers of the early stage of colorectal cancer (19–22). Therefore, abnormal methylation may contribute greatly to the pathogenesis and progression of cancer. In addition, there are many studies of disease based on computational methods (23–25).

Bladder cancer is one of the most common malignant tumors in urology, and its incidence is increasing year by year. About 70% of newly discovered bladder cancer is non-invasive bladder cancer, but nearly 70% of patients relapse after surgical resection of the primary tumor, and 30% of them progress to invasive bladder cancer. Invasive bladder cancer has a poor prognosis and is the main cause of eventual metastasis and death of bladder cancer patients (26). Bladder cancer can be divided into two categories according to the invasion degree of tumor and whether it invades muscle layer. Nearly 70% of these cancers are non-muscle invasive bladder cancer (NMIBC). The main treatment methods for NMIBC are transurethral resection of the bladder tumor and local perfusion therapy of bladder (TURBT). TURBT is a minimally invasive surgery with little trauma and fast recovery. Patients have a relatively good prognosis (27). About 20% ~ 30% are muscle-infiltrating bladder cancer, which is prone to recurrence and distant metastasis after operation due to its high degree of malignancy and complicated treatment. Therefore, early identification of cancer types in patients with bladder cancer is of great significance for cancer treatment.

A large number of studies have focused on abnormalities in DNA methylation and its important role in the occurrence and development of bladder cancer. Kawakami et al. reported for the first time that MSH3 epigenetic regulation by means of DNA methylation might contribute to gene silencing, being implicated in bladder cancer carcinogenesis (28). In addition, there are many researchers aimed at the prognosis of bladder cancer at the level of DNA methylation. Recently, with BLCA sample transcriptome data and methylation data from The Cancer Genome Atlas (TCGA), 18 target genes were identified and the signature based on them was considered an effective and independent prognostic factor (29). However, the existence of tumor heterogeneity leads to the inconsistency of tumor phenotype, and the efficacy and prognosis of different patients are also significantly different. These differences not only pose great challenges to the clinical treatment of bladder cancer, but also reflect the importance of precision medicine. Genetic variation is the core of tumor heterogeneity. There is a wide range of genetic diversity in tumors, and genomic instability leads to a large number of mutations, which is the main cause of genetic heterogeneity in tumors (30). But epigenetic changes, including DNA methylation, also play an important role in

cancer development and perhaps in the molecular heterogeneity of cancer. A previous study showed that BRCA1 promoter methylation was correlated with clinical breast cancer stages (31). Thus, DNA methylation status may be used as a marker for cancer molecular subtyping.

In fact, a large number of studies have been devoted to the analysis of molecular subtypes and DNA methylation heterogeneity of bladder cancer. Attempts have been made to unravel the complexity and refine these molecular subtypes based on biomarkers and pathways, mutations and copy number aberrations, or protein abundance (32). Linskrog et al. performed an integrative multi-omics analysis of patients diagnosed with NMIBC and identified four classes reflecting tumor biology and disease aggressiveness (33). A comprehensive analysis of 412 muscle-invasive bladder cancers characterized by multiple TCGA analytical platforms, clustering by mRNA, lncRNA, and miRNA expression converged to identify subsets with differential epithelial-mesenchymal transition status, carcinoma-*in-situ* scores, histologic features, and survival (34). Recently, Ye et al. used DNA methylation to predict tumor molecular subtypes and efficacy of immunotherapy in bladder cancer (35). One previous study used DNA methylation profiling of bladder cancer samples obtained from the Illumina GoldenGate Methylation Bead Array and unsupervised clustering of those loci with the greatest change in methylation between tumor and non-diseased tissue was performed to defined molecular subgroups of bladder cancer (36). However, most of these analyses did not integrate DNA methylation and gene expression into a detailed classification of bladder cancer at the molecular level, nor did they provide specific biomarkers for individual molecular subtypes.

In this study, we addressed bladder tumor classification based on DNA methylation profiles of BLCA from The Cancer Genome Atlas (TCGA) database. The classification characteristics were obtained by integrating gene expression and DNA methylation data, then the molecular subtypes of bladder cancer were identified based on consistent clustering, and specific prognostic differences among these subgroups were analyzed. This classification system may help find new bladder cancer markers or molecular subtypes and more accurately subdivide bladder cancer patients. Additionally, our criteria will provide more targets for bladder cancer precision medicine by finding specific molecular markers for each subtype. Finally, the new molecular subtypes and subtype-specific molecular markers identified in this study were validated in an independent cohort from GEO database.

MATERIALS AND METHODS

Data Acquisition and Processing

The Illumina Infinium HumanMethylation450 Bead Chip DNA methylation profile data and RNA-seq data of bladder cancer patients as well as clinical information and survival data of the samples were obtained from TCGA database (37), including 408 tumor samples, 14 normal control samples. The expression data

were processed as follows: zero-valued entries were replaced by the minimal positive value of the dataset; the expression values were logarithmically transformed (base 2) to normalize the data. The methylation level of each probe was represented by β -value, which ranges from 0 to 1, corresponding to unmethylated and fully methylated, respectively. Probes with missing data in more than 70% of the samples were removed. The remaining probes with not available (NAs) were imputed using the k-nearest neighbors (KNN) imputation procedure. Unstable genomic sites, including CpGs in sex chromosomes and single nucleotide polymorphisms were removed. Because DNA methylation in promoter regions strongly influences gene expression (38), we selected CpGs within promoter regions. Promoter regions were defined as 2 kb upstream to 0.5 kb downstream from transcription start sites.

Identification of Differentially Expressed Genes and Differentially Methylated CpG Sites

In this study, the differences of gene expression and DNA methylation were combined to classify patients, so the data sets were first used to screen differentially expressed genes and differentially methylated CpG sites between cancer samples and adjacent control samples of bladder cancer.

Differentially expressed genes were screened by samr R package. Genes that meet the following two conditions are identified as differentially expressed genes: foldchange > 2, $q < 1$. Differential methylated CpG sites were screened by minfi R package. The CpGs whose adjusted p value were less than 0.05 and the difference of the average β values were more than 20 percent were considered differentially methylated CpGs between cancer patients and adjacent control tissues. The Differentially expressed genes and differentially methylated CpG sites were displayed using heat maps, which were completed using heatmap.2 function. All processes were programmed using R software.

Correlation Analysis of Gene Expression and DNA Methylation

Since hypermethylation in the gene promoter regions usually inhibits the expression of downstream genes, the methylation level of the gene promoter regions should be negatively correlated with the expression level of corresponding gene, that is, the higher the methylation level, the lower the corresponding gene expression level. Therefore, Pearson Correlation Coefficients between the differentially methylated CpG sites within promoter regions and differentially expressed genes were calculated, and CpGs whose DNA methylation levels significant negatively correlated (Pearson Correlation Coefficient less than 0, $p < 0.05$) with the corresponding gene expression levels were selected as classification characteristics, these CpG sites are the regulators of gene expression. Pearson Correlation Coefficient is calculated as follows:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\delta_X} \right) \left(\frac{Y_i - \bar{Y}}{\delta_Y} \right) \quad (1)$$

Molecular Subtypes of BLCA Were Obtained by Consistent Clustering

Consensus clustering was performed using the Consensus ClusterPlus package (39) to determine subgroups of BLCA based on the characteristic CpG sites obtained in the previous step. The algorithm began by subsampling a proportion of items and features from the data matrix, where each subsample was partitioned into up to k groups by a user-specified clustering algorithm such as k-means, hierarchical clustering or a custom algorithm. This process was repeated for a user-specified number of repetitions, providing a method of representing the consensus across multiple runs of the clustering algorithm and assessing the stability of the discovered clusters. Pairwise consensus values, defined as ‘the proportion of clustering runs in which two items are grouped together’, were calculated and stored in a consensus matrix for each k. Then, for each k, a final agglomerative hierarchical consensus clustering using distance of 1-consensus value was completed and pruned to k groups, which were called consensus clusters. This algorithm determined “consensus” clusters by measuring the stability of clustering results from the application of a given clustering method to random subsets of data. In each iteration, 80% of the tumors were sampled, and the k-means algorithm, with the Euclidean squared distance metric, i.e.

$$d = \sum_{k=1}^N (x_{11} - y_{11})^2 + \dots + (x_{1k} - y_{1k})^2 + \dots + (x_{1N} - y_{1N})^2 \quad (2)$$

was used with $k = 2$ to $k = 10$ groups; these results were compiled over 100 iterations. After executing ConsensusClusterPlus, the cluster-consensus and item-consensus results were obtained. The graphical output results included heatmaps of the consensus matrices, which displayed the clustering results, consensus cumulative distribution function (CDF) plot and delta area plot, which allow us to determine an approximate number of clusters. The criteria to determine the number of clusters we considered were that the consistency within the clusters was relatively high, the coefficient of variation was relatively low and that there was no appreciable increase in the area under the CDF curve. The coefficient of variation was calculated according to the following formula:

$$CV = \left(\frac{SD}{MN} \right) \times 100\% \quad (3)$$

in which SD represents the standard deviation, and MN represents the average of samples.

Differential Prognostic Analysis of Molecular Subtypes

In order to test the differences except DNA methylation levels among the bladder cancer subgroups obtained, survival analysis was performed on the patients in these subgroups. Kaplan–Meier plots were used to illustrate overall survival among BLCA subgroups defined by DNA methylation profiles. The log-rank test was used to evaluate the significance difference among the clusters, $p < 0.05$ was considered significant. Survival analyses were performed using the survival package in R.

Identification of Subgroup Specific DNA Methylation Biomarkers in Bladder Cancer

In this analysis, a quantitative approach for quantitative differentially methylated regions (QDMRs), which quantify methylation differences and identify DMRs from genome-wide methylation profiles by adapting Shannon entropy (40), was used to find the specific DNA methylation CpGs that were specifically hypermethylated or hypomethylated within particular bladder tumor subgroup. The quantification of DNA methylation difference across large numbers of samples and the identification of sample specificity plays important roles in genomic functional analyses. DMRs with different methylation statuses among multiple samples were regarded as possible epigenetic functional regions involved in transcriptional regulation. Thus, the identification of DMRs among multiple samples provided a more comprehensive survey for this study. With the rapid development of high-throughput detection technology, there have been considerable efforts in identifying DMRs from methylation profiles. However, the development of DNA methylation measurements proposed significant challenges for concurrent DMR methods. Shannon entropy, a quantitative measure of differences and uncertainty in data sets, has been widely applied in quantitative biology, such as identifying potential drug targets and tissue-specific genes. To quantify methylation differences and further identify DMRs across multiple samples, Zhang et al. adapted the Shannon entropy model and developed an improved approach, termed quantitative differentially methylated region (QDMR). QDMR was an effective tool for quantifying methylation differences and identifying DMRs across multiple samples. This approach can give a reasonable quantitative measure of methylation differences across multiple samples as well. We used the threshold that was determined by QDMR from the methylation probability model. Furthermore, QDMR can also measure the sample specificity of each DMR. For each DMR r , the entropy H_Q represents the methylation difference across all samples. For each sample S , the entropy is $H_{Q/S}$ the difference across samples that do not include sample S . Thus, the contribution of sample S to the whole methylation difference can be reflected by the entropy difference as:

$$\Delta H_{r/S} = H_{Q/S} - H_Q \quad (4)$$

And the categorical sample-specificity $CS_{r/S}$ can be defined as:

$$CS_{r/S} = \begin{cases} \Delta H_{r/S} \times \text{sign}_{r/S}, & \Delta H_{r/S} > 0 \\ 0, & \Delta H_{r/S} \leq 0 \end{cases} \quad (5)$$

where $\text{sign}_{r,S}$ is the sign of the difference between methylation level $m_{r/S}$ in sample S and the median methylation level of vector m_r in region r , as described by Zhang et al. (40). Thus, the subgroup with the maximal absolute of the categorical sample-specificity $CS_{r/S}$ was determined as the specific subgroup corresponding to the particular CpG site.

Functional Enrichment Analysis of Genes Corresponding to Specific CpGs

In this study, using DAVID (41, 42), a database used for annotation, visualization and integration of discoveries, we

conducted a GO (Gene Ontology) biological functions enrichment analysis and a KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways enrichment analysis towards the list of genes corresponding to specific CpGs, with p controlled within 0.05, which could find out the biological characteristics involved by our specific CpGs.

Construction and Verification of Classification Model

In order to verify the accuracy of classification characteristics, robustness of subtypes and accuracy of subtype-specific CpG sites of the bladder cancer DNA methylation subtypes classified in this study, another set of external test data set (GSE52955) was retrieved from GEO database. The classification features (characteristics CpG sites) were used as the features of the model, TCGA data used in this study was used as the training set to build a SVM classifier model, and the accuracy of the model was verified by the ten-fold cross-validation method. The external test set data is then entered into the built model, which is used to classify the new samples.

RESULTS

Acquisition of Classification Features

To obtain characteristics for molecular subtype classification, we first identify genes and CpG sites that differ between cancer and normal samples which are associated with cancer development. First, samr R package was used to screen differentially expressed genes as described above. 408 differentially expressed genes were obtained, and these differentially expressed genes were displayed by heat map. In the heat map, genes were represented by rows and patients were represented by columns, the red bars represent cancer patients, the blue bars represent normal tissue samples adjacent cancer samples, and the middle area were gene expression levels (Figure 1A). As can be seen from the heat map, these differentially expressed genes can clearly separate the bladder cancer patient samples from the para-cancer control samples.

Next, minfi R package was used to screen differentially methylated CpG sites. Through the processes mentioned above, 9702 differentially methylated CpGs between bladder cancer patients and control samples were identified. The differentially methylated CpGs were also shown in the heat map (Figure 1B). The heat map displays the methylation levels of differentially methylated CpGs in cancer samples and adjacent control samples. The rows represent CpG sites, the columns represent patients, and the colors represent the levels of DNA methylation. As can be seen from the heat map, these differentially methylated CpG sites can also clearly separate the bladder cancer patient samples from the para-cancer control samples.

Since the methylation level of gene promoter region was negatively correlated with the expression level of corresponding gene, the CpG loci which significant negative correlation with gene expression were extracted (Pearson Correlation Coefficient less than 0, $p < 0.05$). Finally, 986 CpG loci were obtained and analyzed as the classification features.

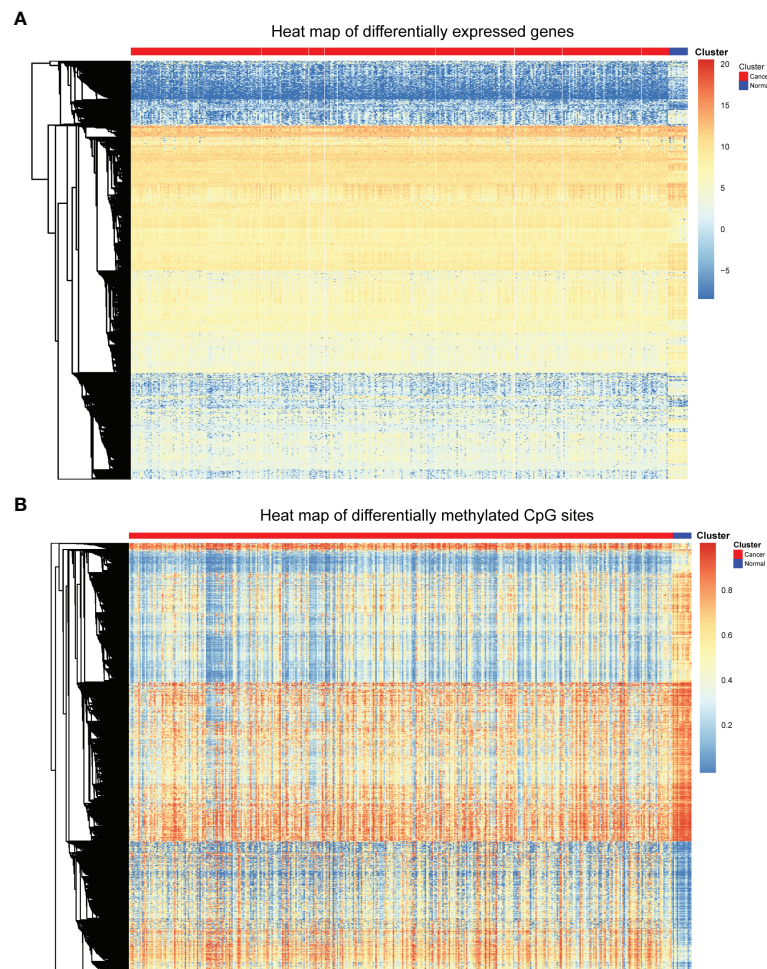


FIGURE 1 | Heat map of differentially expressed genes and differentially methylated CpGs. **(A)** Heat map of differentially expressed genes. **(B)** Heat map of differential methylated CpG sites.

Different Molecular Subtypes of Bladder Cancer Were Obtained Based on Consistent Clustering Algorithm

Next, consensus clustering based on the β values of the 986 CpG sites obtained was performed to obtain distinct DNA methylation molecular subtypes of bladder cancer. To determine the appropriate number of subgroups, the average cluster consensus and the coefficient of variation among clusters were calculated for each category number. In this study, the cluster number selection criteria we considered were relatively high average consistency within the clusters, relatively low coefficient of variation, and maximum area change under the CDF curve. The consensus matrix was naturally a better visualization tool to help assess the clusters' composition and number. We associated a color gradient from 0–1, with white corresponding to 0 and dark blue corresponding to 1, and assume the matrix is arranged so that items belonging to the same cluster are adjacent to each other. In this arrangement, a matrix corresponding to a perfect consensus will show a color-

coded heatmap characterized by blue blocks along the diagonal on a white background. The color-coded heatmap corresponding to the consensus matrix obtained by applying consensus clustering to these cases is shown in **Figure 2A**, and represents the consensus for $k = 4$, which displays a well-defined 4-block structure. It has the largest area change under CDF curve, the highest average consistency within the class, and the lowest consistency coefficient of variation (**Figures 2C, D**). Therefore, we determine the appropriate number of categories as 4. So, all bladder cancer patients were divided into four DNA methylation molecular subtypes.

Prognostic Analysis of Different Molecular Subtypes

After consistent clustering was used to identify DNA methylation subgroups in bladder cancer, we then examined whether there were differences among the subgroups in addition to DNA methylation levels. We examined the differences in prognosis among the four DNA methylation subgroups.

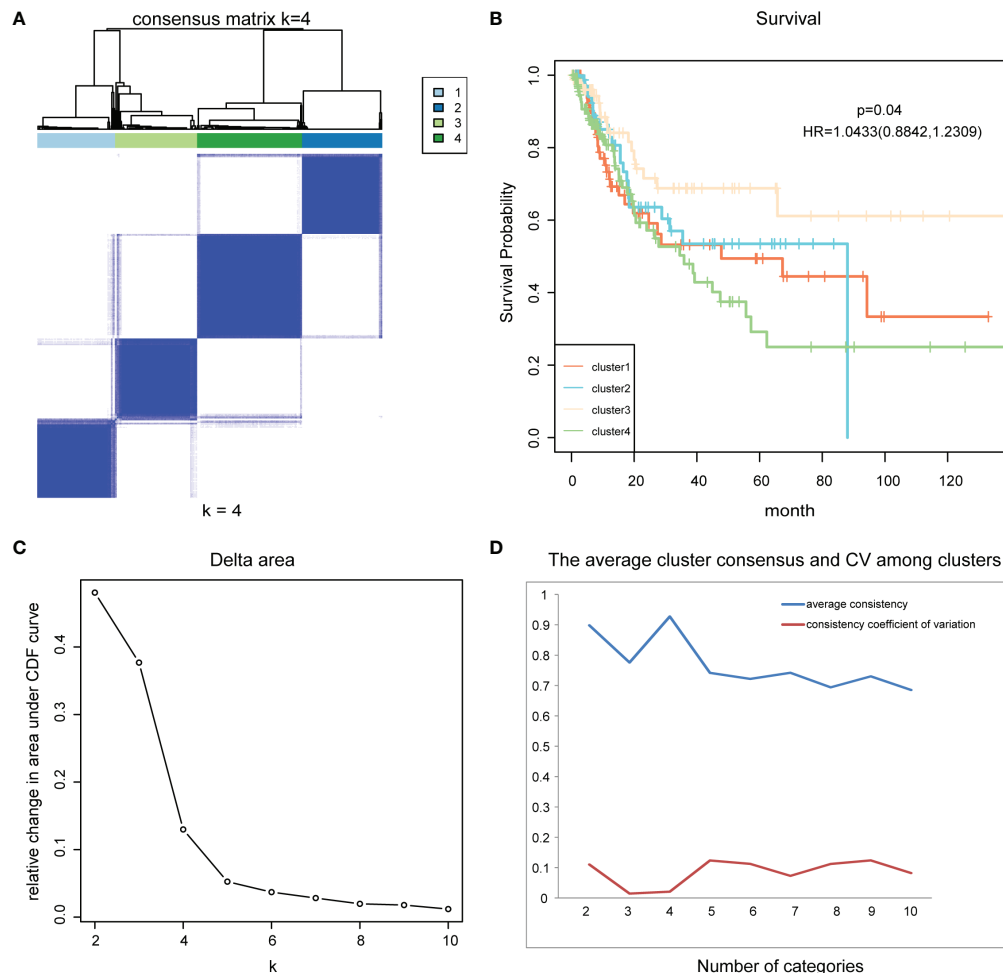


FIGURE 2 | Consensus clustering and survival analysis. **(A)** The color-coded heatmap corresponding to the consensus matrix for $k = 4$. **(B)** The survival curves of four DNA methylation subtypes of bladder cancer. **(C)** Delta area curve of consensus clustering. **(D)** The average cluster consensus and coefficient of variation among clusters for each category number k .

Kaplan-Meier survival analysis was performed for these four subtypes using functions `survfit()` and `survdiff()` in R package “Survival”, and log rank test was used to determine the statistical significance of survival differences. Results showed significant prognostic differences among the four subgroups ($p = 0.04$) (Figure 2B). This indicates that there are significant differences in the prognostic status of patients among the four DNA methylation molecular subtypes of bladder cancer, which can provide an important reference for clinicians to predict the survival status of patients and timely change the treatment plan.

Identification and Analysis of DNA Methylation Biomarkers Specific to DNA Methylation Subtypes

After identifying the DNA methylation molecular subtypes of bladder cancer using unsupervised consistent clustering, the present study focused on DNA methylation markers specific to each subtype. These markers could represent the unique

characteristics of each subtype, and their screening can provide a basis for the diagnosis of DNA methylation subtypes, and facilitate the better translation of research results into clinical application.

The QDMR software developed as a quantitative method described above was used in this study. 986 CpG loci across four DNA methylation subgroups (the classification features used in this study) were used as candidate features to screen for specific CpG markers in each subgroup. Since the methylation levels of these 986 CpG sites were used to distinguish the DNA methylation subgroups in this study, in each subgroup, these features should have similar methylation levels and there was very little variability between samples. Therefore, for each of the four DNA methylation subgroups, the average DNA methylation level of the 986 CpG sites in the samples was calculated to represent the DNA methylation pattern of that subgroup, and the 986×4 -dimensional result matrix was used as the input of QDMR. Finally, 52 specific hyper/hypomethylated CpG loci

were identified, corresponding to 38 genes. They can be used as specific DNA methylation markers for different DNA methylation subgroups in bladder cancer, representing the unique DNA methylation patterns of that subgroup. The results showed that cluster1 and cluster2 specific CpG sites were found, and the number of CpG sites was 9 and 43, respectively. The specific sites of cluster1 screened out by our study are all hypomethylated, while the specific sites of cluster2 are all hypermethylated (**Figure 3A**).

Cluster1 specific CpG loci mapped 3 genes, and cluster2 specific CpG loci mapped 35 genes (**Supplementary Material 1**). Next, DAVID bioinformatics tool was employed to complete the Gene Ontology and KEGG pathway enrichment analysis on cluster2 specific genes to further explore the biological processes or pathways involved. Results as shown in **Figure 3B**, these genes are involved in biological processes such as positive transcriptional regulation of RNA polymerase II promoters. But we did not find pathways in which these genes were significantly enriched.

Validation of DNA Methylation Molecular Subtypes and Subtype-Specific CpG Sites

To verify the robustness of the DNA methylation molecular subtypes of bladder cancer obtained in this study and the

accuracy of the subtype specific CpG sites screened, we searched the GEO database and obtained a set of Illumina Infinium HumanMethylation450 Bead Chip DNA methylation profile data of urinary tumors (GSE52955), which included 25 patients with bladder cancer. Data from these 25 patients were used as a test set to verify the molecular subtypes and subtype-specific CpG sites obtained in this study.

Firstly, we used 986 CpG loci previously screened as classification features, and constructed a support vector machine (SVM) classifier model using TCGA data set with classification labels (i.e., four DNA methylation molecular subtypes divided in this study). Here we conducted functional analysis of the genes corresponding to the 986 characteristic CpG loci, and found that they were enriched in regulation of transcription from RNA polymerase II promoter, cell differentiation, positive regulation of cell migration, cell-cell adhesion, signal transduction, negative regulation of cell proliferation, cAMP signaling pathway, vascular smooth muscle contraction and many other biological processes and pathways involved in cancer genesis and development. The model was verified using tenfold cross validation. The results showed that the classification accuracy of the model was 96%, sensitivity 96.1%, precision 96.1%, and area under ROC curve

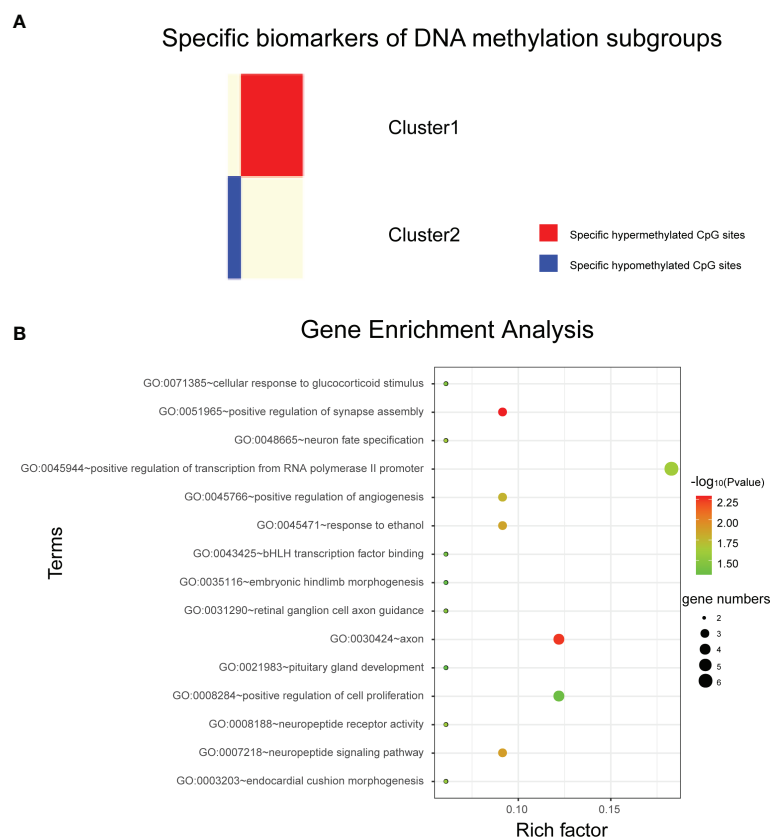


FIGURE 3 | Analysis of subtype specific biomarkers. **(A)** Specific hyper/hypo methylated CpG sites for DNA methylation cluster1 and cluster2. **(B)** Gene enrichment analysis of genes corresponding to specific hypermethylated CpG sites in cluster2.

(AUC) reached 0.968 (**Figure 4**). This further proved the accuracy of the classification features screened in this study and the robustness of the DNA methylation molecular subtypes of bladder cancer.

Next, we input the test set obtained from GEO database into the constructed classifier model, which is used to predict the test set samples into the four DNA methylation molecular subtypes divided in this study. The 25 samples in the test set were predicted to be cluster1, cluster2, cluster3 and cluster4 with 3, 5, 11 and 6 samples respectively. Next, we tested the DNA methylation level of the subtype-specific CpG sites screened in this study in the test set. The results showed that 6 of the 9 CpG sites with cluster1-specific hypomethylation in the test dataset

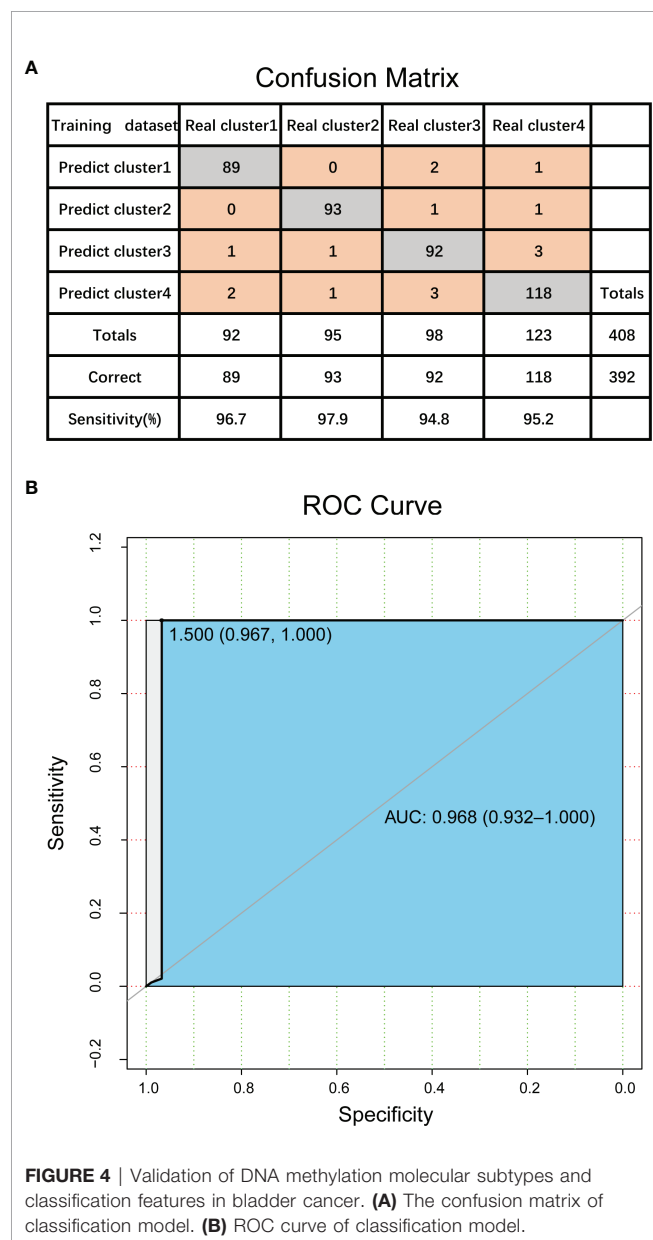
still had the lowest average methylation level in the four subtypes. The methylation level of the other three sites was not the lowest (close to the methylation level of cluster3), but significantly lower than that of the other two subtypes. In cluster2, all of the 43 CpG sites specific hypermethylated are still the CpG sites with the highest average methylation level among the four subtypes and significantly higher than the other three subtypes (**Supplementary Material 2**). This proves the accuracy and portability of the subtype-specific CpG sites screened in our study.

DISCUSSION

Cancer is a disease with high mortality rate and a serious threat to people's health. Previous studies focused only on the effect of genetic sequence changes on cancer, or malignancy. Recently, a relationship between cancer and the level of DNA methylation has been found. TCGA database is a publicly available resource covering a wide variety of data types in a variety of cancers. The Infinium HumanMethylation450 BeadChip array dataset of bladder cancer contains a large number of samples that were downloaded from TCGA for our classification analysis. The large sample sizes allowed us to explore the molecular subtypes of bladder cancer more comprehensively.

Precision medicine in cancer treatment is based on the assumption that every patient has a unique variation of genetic alterations and should be treated accordingly. Thus, for personalized medicine to be effective, it is necessary to achieve a detailed classification of the cancer genome and epigenome. Many studies have suggested that epigenetic modifications (DNA methylation) play a pivotal role in early detection, and improved molecular classification, prognosis and adjuvant treatment of bladder cancer. These opinions suggested that the level of analysis could have important biological and clinical implications in the era of precision medicine (43, 44). Moreover, classifications based solely on the tissue of origin or pathological features have shown their limitations. To this end, we conducted this study to obtain molecular classifications of bladder cancer epigenomes based on DNA methylation.

In this study, DNA methylation and gene expression profile data of TCGA were integrated to screen differentially expressed genes and differentially DNA methylated CpG sites. CpG sites with significant negative correlation between methylation level and gene expression level were extracted as classification features. Then bladder cancer samples were classified according to the classification features, and four DNA methylation molecular subtypes were obtained. The prognostic difference analysis of these four subtypes showed that there were significant differences in the molecular level and prognostic status among these subtypes. Furthermore, subtype-specific biomarkers were identified using information entropy-based algorithm to represent the unique molecular characteristics of each subtype. These results suggest that there are significant differences in epigenetics and prognosis among subpopulations of patients with the same cancer, and clinicians may be able to develop personalized and timely



treatment changes based on their prognostic status. However, the specific characteristic biomarkers for only two of the four DNA methylation molecular subtypes, namely cluster1 and cluster2, were identified. The specific characteristic biomarkers for cluster3 and cluster4 were not identified. This indicates that these two DNA methylated molecular subtypes are more similar at the molecular level, and their differentiation is not as obvious as the other two subtypes, which also brings certain limitations to this study. We hope that future studies can focus on further differentiation of these two subgroups.

In conclusion, our research identified four different molecular subgroups using the data of bladder tumors in TCGA. This is a more detailed explanation of the molecular heterogeneity of bladder cancer. The specific CpG sites and genes for particular subgroups can serve as biomarkers for personalized treatments. Changes in DNA methylation (hypo/hypermethylation) can be used as markers to diagnose particular subgroups, and clinicians can develop personalized treatments according to these prognoses. Additionally, our methods can also be used to study other tumors with high molecular heterogeneity.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: All data analyzed in this study are from open data (freely available to anyone) at TCGA database and GEO database.

REFERENCES

1. Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor Origin Detection With Tissue-Specific miRNA and DNA Methylation Markers. *Bioinformatics* (2018) 34(3):398–406. doi: 10.1093/bioinformatics/btx622
2. Jones PA, Baylin SB. The Epigenomics of Cancer. *Cell* (2007) 128(4):683–92. doi: 10.1016/j.cell.2007.01.029
3. Li H, Gong Y, Liu Y, Lin H, Wang G. Detection of Transcription Factors Binding to Methylated DNA by Deep Recurrent Neural Network. *Brief Bioinform* (2022) 23(1):bbab533. doi: 10.1093/bib/bbab533
4. Li E, Zhang Y. DNA Methylation in Mammals. *Cold Spring Harbor Perspect Biol* (2014) 6(5):a019133. doi: 10.1101/cshperspect.a019133
5. Luo X, Zhang T, Zhai Y, Wang F, Zhang S, Wang G. Effects of DNA Methylation on TFs in Human Embryonic Stem Cells. *Front Genet* (2021) 12:639461. doi: 10.3389/fgene.2021.639461
6. Mo F, Luo Y, Fan DA, Zeng H, Zhao YN, Luo M, et al. Integrated Analysis of mRNA-Seq and miRNA-Seq to Identify C-MYC, YAP1 and miR-3960 as Major Players in the Anticancer Effects of Caffeic Acid Phenethyl Ester in Human Small Cell Lung Cancer Cell Line. *Curr Gene Ther* (2020) 20(1):15–24. doi: 10.2174/1566523220666200523165159
7. Zuo Y, Song M, Li H, Chen X, Cao P, Zheng L, et al. Analysis of the Epigenetic Signature of Cell Reprogramming by Computational DNA Methylation Profiles. *Curr Bioinf* (2020) 15(6):589–99. doi: 10.2174/1574893614666190919103752
8. Tanaka E, Uchida D, Shiraha H, Kato H, Ohya A, Iwamuro M, et al. Promising Gene Therapy Using an Adenovirus Vector Carrying REIC/Dkk-3 Gene for the Treatment of Biliary Cancer. *Curr Gene Ther* (2020) 20(1):64–70. doi: 10.2174/1566523220666200309125709
9. Zhang S, Zhang J, Zhang Q, Liang Y, Du Y, Wang G. Identification of Prognostic Biomarkers for Bladder Cancer Based on DNA Methylation Profile. *Front Cell Dev Biol* (2021) 9:817086. doi: 10.3389/fcell.2021.817086
10. Karsli-Ceppioglu S, Dagdemir A, Judes G, Ngollo M, Penault-Llorca F, Pajon A, et al. Epigenetic Mechanisms of Breast Cancer: An Update of the Current Knowledge. *Epigenomics* (2014) 6(6):651–64. doi: 10.2217/epi.14.59
11. Jaenisch R, Bird A. Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals. *Nat Genet* (2003) 33 Suppl:245–54. doi: 10.1038/ng1089
12. Yalcin D, Otu HH. An Unbiased Predictive Model to Detect DNA Methylation Propensity of CpG Islands in the Human Genome. *Curr Bioinf* (2021) 16(2):179–96. doi: 10.2174/1574893615999200724145835
13. Sahu R, Pattanayak SP. Strategic Developments & Future Perspective on Gene Therapy for Breast Cancer: Role of mTOR and Brk/PTK6 as Molecular Targets. *Curr Gene Ther* (2020) 20(4):237–58. doi: 10.2174/1566523220999200731002408
14. Bariol C, Suter C, Cheong K, Ku SL, Meagher A, Hawkins N, et al. The Relationship Between Hypomethylation and CpG Island Methylation in Colorectal Neoplasia. *Am J Pathol* (2003) 162(4):1361–71. doi: 10.1016/S0002-9440(10)63932-6
15. Oster B, Thorsen K, Lamy P, Wojdacz TK, Hansen LL, Birkenkamp-Demtroder K, et al. Identification and Validation of Highly Frequent CpG Island Hypermethylation in Colorectal Adenomas and Carcinomas. *Int J Cancer* (2011) 129(12):2855–66. doi: 10.1002/ijc.25951
16. Shen Z, Zou Q. Basic Polar and Hydrophobic Properties are the Main Characteristics That Affect the Binding of Transcription Factors to Methylation Sites. *Bioinformatics* (2020) 36(15):4263–8. doi: 10.1093/bioinformatics/btaa492
17. Cheng L, Qi C, Yang H, Lu M, Cai Y, Fu T, et al. Gutmgene: A Comprehensive Database for Target Genes of Gut Microbes and Microbial Metabolites. *Nucleic Acids Res* (2021) 50(D1):D795–800. doi: 10.1093/nar/gkab786
18. Beggs AD, Jones A, El-Bahrawy M, Abulafi M, Hodgson SV, Tomlinson IP. Whole-Genome Methylation Analysis of Benign and Malignant Colorectal Tumours. *J Pathol* (2013) 229(5):697–704. doi: 10.1002/path.4132

AUTHOR CONTRIBUTIONS

SZ and SH conceived and designed the experiments. SW, DX, YS, and SY conducted all the data processing work described in the section of methods and performed the analysis. SZ and BG prepared and edited the manuscript. XT and YJ checked and proofread the entire manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Natural Science Foundation of China (62002057, 62002087, 62172129); Fundamental Research Funds for the Central Universities (2572020BH02); Innovation and Entrepreneurship Training Program for Students of Northeast Forestry University (DC2020141).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.915542/full#supplementary-material>

Supplementary Material 1 | Subtype-specific CpG sites and the corresponding genes.

Supplementary Material 2 | Categorical sample specificity of each subtype-specific site.

19. Lao VV, Grady WM. Epigenetics and Colorectal Cancer. *Nat Rev Gastroenterol Hepatol* (2011) 8(12):686–700. doi: 10.1038/nrgastro.2011.173
20. Luo Y, Wong CJ, Kaz AM, Dzieciatkowski S, Carter KT, Morris SM, et al. Differences in DNA Methylation Signatures Reveal Multiple Pathways of Progression From Adenoma to Colorectal Cancer. *Gastroenterology* (2014) 147(2):418–429.e418. doi: 10.1053/j.gastro.2014.04.039
21. Mohammed M, Mwambi H, Omolo B. Colorectal Cancer Classification and Survival Analysis Based on an Integrated RNA and DNA Molecular Signature. *Curr Bioinform* (2021) 16(4):583–600. doi: 10.2174/1574893615999200711170445
22. Gao J, Zhang L, Yu G, Qu G, Li Y, Yang X. Model With the GBDT for Colorectal Adenoma Risk Diagnosis. *Curr Bioinform* (2020) 15(9):971–9. doi: 10.2174/1574893614666191120142005
23. Liu Q, Wan J, Wang G. A Survey on Computational Methods in Discovering Protein Inhibitors of SARS-CoV-2. *Brief Bioinform* (2022) 23(1):bbab416. doi: 10.1093/bib/bbab416
24. Li Y, Qiao G, Wang K, Wang G. Drug-Target Interaction Predication via Multi-Channel Graph Neural Networks. *Brief Bioinform* (2022) 23(1):bbab346. doi: 10.1093/bib/bbab346
25. Li Y, Wang K, Wang G. Evaluating Disease Similarity Based on Gene Network Reconstruction and Representation. *Bioinformatics* (2021) btab252. doi: 10.1093/bioinformatics/btab252
26. Gerlinger M, Catto JW, Orntoft TF, Real FX, Zwarthoff EC, Swanton C. Intratumour Heterogeneity in Urologic Cancers: From Molecular Evidence to Clinical Implications. *Eur Urol* (2015) 67(4):729–37. doi: 10.1016/j.eururo.2014.04.014
27. Yuk HD, Kim JK, Jeong CW, Kwak C, Kim HH, Ku JH. Differences in Pathologic Results of Repeat Transurethral Resection of Bladder Tumor (TURBT) According to Institution Performing the Initial TURBT: Comparative Analyses Between Referred and Nonreferred Group. *BioMed Res Int* (2018) 2018:9432606. doi: 10.1155/2018/9432606
28. Kawakami T, Shiina H, Igawa M, Deguchi M, Nakajima K, Ogishima T, et al. Inactivation of the Hmsh3 Mismatch Repair Gene in Bladder Cancer. *Biochem Biophys Res Commun* (2004) 325(3):934–42. doi: 10.1016/j.bbrc.2004.10.114
29. Liu Z, Sun T, Zhang Z, Bi J, Kong C. An 18-Gene Signature Based on Glucose Metabolism and DNA Methylation Improves Prognostic Prediction for Urinary Bladder Cancer. *Genomics* (2021) 113(1 Pt 2):896–907. doi: 10.1016/j.ygeno.2020.10.022
30. Burki TK. High Genetic Heterogeneity in Some Breast Cancer Tumours. *Lancet Oncol* (2015) 16(15):e529. doi: 10.1016/S1470-2045(15)00359-9
31. Chen Y, Zhou J, Xu Y, Li Z, Wen X, Yao L, et al. BRCA1 Promoter Methylation Associated With Poor Survival in Chinese Patients With Sporadic Breast Cancer. *Cancer Sci* (2009) 100(9):1663–7. doi: 10.1111/j.1349-7006.2009.01225.x
32. Tan TZ, Rouanne M, Tan KT, Huang RY, Thiery JP. Molecular Subtypes of Urothelial Bladder Cancer: Results From a Meta-Cohort Analysis of 2411 Tumors. *Eur Urol* (2019) 75(3):423–32. doi: 10.1016/j.eururo.2018.08.027
33. Linskrog SV, Prip F, Lamy P, Taber A, Groeneveld CS, Birkenkamp-Demtroder K, et al. An Integrated Multi-Omics Analysis Identifies Prognostic Molecular Subtypes of non-Muscle-Invasive Bladder Cancer. *Nat Commun* (2021) 12(1):2301. doi: 10.1038/s41467-021-22465-w
34. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* (2017) 171(3):540–556.e525. doi: 10.1016/j.cell.2017.09.007
35. Ye F, Liang Y, Hu J, Hu Y, Liu Y, Cheng Z, et al. DNA Methylation Modification Map to Predict Tumor Molecular Subtypes and Efficacy of Immunotherapy in Bladder Cancer. *Front Cell Dev Biol* (2021) 9:760369. doi: 10.3389/fcell.2021.760369
36. Wilhelm-Benartzi CS, Koestler DC, Houseman EA, Christensen BC, Wiencke JK, Schned AR, et al. DNA Methylation Profiles Delineate Etiologic Heterogeneity and Clinically Important Subgroups of Bladder Cancer. *Carcinogenesis* (2010) 31(11):1972–6. doi: 10.1093/carcin/bgq178
37. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat Genet* (2013) 45(10):1113–20. doi: 10.1038/ng.2764
38. Wu H, Zhang Y. Reversing DNA Methylation: Mechanisms, Genomics, and Biological Functions. *Cell* (2014) 156(1-2):45–68. doi: 10.1016/j.cell.2013.12.019
39. Wilkerson MD, Hayes DN. ConsensusClusterPlus: A Class Discovery Tool With Confidence Assessments and Item Tracking. *Bioinformatics* (2010) 26(12):1572–3. doi: 10.1093/bioinformatics/btq170
40. Zhang Y, Liu H, Lv J, Xiao X, Zhu J, Liu X, et al. QDMR: A Quantitative Method for Identification of Differentially Methylated Regions by Entropy. *Nucleic Acids Res* (2011) 39(9):e58. doi: 10.1093/nar/gkr053
41. Huang D, Sherman BT, Lempicki RA. Bioinformatics Enrichment Tools: Paths Toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Res* (2009) 37(1):1–13. doi: 10.1093/nar/gkn923
42. Huang da W, Sherman BT, Lempicki RA. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat Protoc* (2009) 4(1):44–57. doi: 10.1038/nprot.2008.211
43. Pasculli B, Barbano R, Parrella P. Epigenetics of Breast Cancer: Biology and Clinical Implication in the Era of Precision Medicine. *Semin Cancer Biol* (2018) 51:22–35. doi: 10.1016/j.semcancer.2018.01.007
44. Hu WL, Zhou XH. Identification of Prognostic Signature in Cancer Based on DNA Methylation Interaction Network. *BMC Med Genomics* (2017) 10 (Suppl 4):63. doi: 10.1186/s12920-017-0307-9

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.

Copyright © 2022 Wang, Xu, Gao, Yan, Sun, Tang, Jiao, Huang and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Prediction of Gastric Cancer-Related Genes Based on the Graph Transformer Network

Yan Chen, Xuan Sun and Jiaxing Yang*

Department of Gastrointestinal Surgery, The First Hospital of Jilin University, Changchun, China

OPEN ACCESS

Edited by:

Liang Cheng,
Harbin Medical University, China

Reviewed by:

Xiaoke Ma,
Xidian University, China
Hui Liu,
School of Computer Science and
Technology, China

*Correspondence:

Jiaxing Yang
Jiaxingyang@jlu.edu.cn

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 23 March 2022

Accepted: 26 April 2022

Published: 30 June 2022

Citation:

Chen Y, Sun X and Yang J (2022)
Prediction of Gastric Cancer-
Related Genes Based on the
Graph Transformer Network.
Front. Oncol. 12:902616.
doi: 10.3389/fonc.2022.902616

Gastric cancer is a complex multifactorial and multistage process that involves a large number of tumor-related gene structural changes and abnormal expression. Therefore, knowing the related genes of gastric cancer can further understand the pathogenesis of gastric cancer and provide guidance for the development of targeted drugs. Traditional methods to discover gastric cancer-related genes based on biological experiments are time-consuming and expensive. In recent years, a large number of computational methods have been developed to identify gastric cancer-related genes. In addition, a large number of experiments show that establishing a biological network to identify disease-related genes has higher accuracy than ordinary methods. However, most of the current computing methods focus on the processing of homogeneous networks, and do not have the ability to encode heterogeneous networks. In this paper, we built a heterogeneous network using a disease similarity network and a gene interaction network. We implemented the graph transformer network (GTN) to encode this heterogeneous network. Meanwhile, the deep belief network (DBN) was applied to reduce the dimension of features. We call this method "DBN-GTN", and it performed best among four traditional methods and five similar methods.

Keywords: gastric cancer, susceptibility gene, graph transformer network, deep belief network, heterogeneous network

INTRODUCTION

Gastric cancer is a malignant tumor originated from gastric mucosal epithelial cells (1). At present, due to the increase of work pressure, the change of diet structure, and *Helicobacter pylori* infection, gastric cancer is gradually showing a younger trend (2, 3). Patients with early gastric cancer often have no obvious symptoms, or only nonspecific symptoms such as abdominal discomfort and flatulence (4). These symptoms are often similar to chronic gastric symptoms such as dyspepsia, gastritis, and gastric ulcer (5). Most patients with early-stage cancer find their condition through gastroscopy. Under reasonable medical measures, the 5-year survival rate of patients with early-stage gastric cancer can reach 90% (6). However, most patients with gastric cancer are in the middle and late stage of gastric cancer when they are diagnosed. The tumor has invaded the outside of the stomach and is complicated by lymph node metastasis; thus, the odds of being cured is low. Screening related genes closely related to gastric cancer can be used as molecular targets for diagnosis (7). Different gene combinations can reflect the early diagnosis, incidence, effectiveness of

treatment, and prognosis of gastric cancer. The early stage of gastric cancer generally only contains a few gene changes. These changes are potential molecular targets for early diagnosis (8). If these changes can be detected, gastric cancer can be detected as soon as possible, which can greatly improve the cure rate of gastric cancer. The typing of gastric cancer susceptibility genes and related genes can also provide some information for the prediction of the disease (9), so as to take preventive measures as soon as possible to prevent the deterioration of the disease. With the continuous in-depth post-genome studies, more genotypes will be found to be related to the occurrence, development, and prognosis of gastric cancer. The final conclusion can provide a new theoretical basis for the discussion of the molecular mechanism of gastric cancer (10, 11).

Gastric cancer is a complex and multifactorial disease. Environmental and genetic factors play an important role in the occurrence of gastric cancer (12, 13). MiRNA precisely regulates the occurrence of gastric cancer by participating in a network system composed of a series of important biological processes such as cell proliferation, apoptosis, and differentiation (14). A large number of studies have shown that according to the difference in expression level, specific miRNAs have become a potential biomarker of malignant cancer and have an impact similar to carcinogenic or tumor suppressor genes. For example, the expression of miR-21 and miR-155 is usually increased in gastric cancer, which can promote cell proliferation and induce the occurrence of malignant cancer (15), and the expression of mir-449 is usually reduced. It can inhibit cell proliferation and inhibit the further development of gastric cancer (16). To a large extent, miRNA is almost involved in the whole process of gastric cancer pathogenesis. Therefore, with the deepening of research, it can enrich the biological function of miRNA, show a new vision for the in-depth study of the molecular mechanism of the occurrence and development of gastric cancer, and show a broader platform for the medical field. The application of gene chips can further extend the research on gastric cancer into the gene regulation network, making it possible to explore the gene expression profile of gastric cancer in different pathological stages. Gene chips have become a powerful tool to study the molecular regulation mechanism and pathway of gastric cancer progress, and they have been widely used in the field of gastric cancer research. In recent years, tumor genomics and proteomics have been widely used in biomedical and clinical research. Since the rise of gene chips and microarray technology, people have used these technologies to find new disease subclasses (17, 18), identify new tumor markers (19, 20), distinguish tumor grades (21), and predict the prognosis of the disease. For example, Wang et al. found that the increased expression of INHBA was related to the low survival rate of patients with gastric cancer through gene enrichment analysis (22). Liu et al. confirmed that extracellular matrix receptors and cell cycle signaling pathways may play an important role in gastric cancer (23). Wnt signaling pathway may lead to carcinogenesis by stimulating the migration and invasion of gastric cancer cells (24). β -Catenin is frequently mutated in gastric cancer (25). Fze3 is overexpressed in 75% of gastric cancer tissues and hsrp is downregulated in 16% of

gastric cancer tissues, indicating that the expression of fze3 and hsrp in this pathological tissue is often changed (26). Highly recombinant Shh induces the migration and invasion of gastric cancer cells by regulating tissue growth factor (TGF), which plays a role in the alk5-smad3 pathway (27). LOXL2 can promote tumor invasion through the Src/FAK signaling pathway, and its expression in gastric cancer is significantly increased (28). The loss of embryonic liver cell lining protein (ELF) can destroy the TGF-mediated signal pathway by interfering with the localization of Smad3 and Smad4 and lead to gastric cancer (29). The increase of BMP-2 concentration can significantly improve the motility and invasiveness of gastric cancer cells (30). The upregulation of cycox-61 may lead to the progression of gastric cancer. Interleukin-6 induces the invasion of gastric cancer cell line AGS cells through the activation of the c-Src/RhoA/ROCK signaling pathway (31).

Although the cost of large-scale sequencing data is decreasing and the speed is increasing, the number of clear gastric cancer-related genes remains small. A large number of multi omics data of gastric cancer have been accumulated. It is an important means to fully understand the genetic mechanism of gastric cancer to preliminarily screen potential genes through large-scale data mining algorithms and then verify them one by one through biological experiments. Systems biology aims to study the interaction of various molecules with different structures and functions at the overall level of organisms, and then add computational methods to describe and predict biological functions (32, 33), phenotypes, and behaviors. Most of these methods are based on networks (34, 35). These computational methods have been widely used in the discovery of disease-related genes (33, 35–39), genetic mechanism (40, 41), gene expression (37, 40), protein function (42, 43), metabolic association (44, 45), and drug target (46, 47). Therefore, in this paper, we developed a novel method named “DBN-GTN” to identify gastric cancer-related genes in a large scale. This method is based on the thought of systems biology. It used multiple features of genes and gastric cancer to identify the patterns of gastric cancer-related genes, which can be used to find more gastric cancer-related genes.

METHOD

Workflow

We firstly constructed a disease similarity network and a gene interaction network. We connected the two networks together based on the known relationship between diseases and genes. For example, the public databases have shown that the EGFR gene has a relationship with gastric cancer. Then, the node “gastric cancer” can be connected to the node “EGFR”. Finally, we can obtain a heterogeneous network. Then, we should extract the features of diseases and genes, respectively. We used the relationship between miRNAs and both diseases and genes as the features. Therefore, gene feature is the regulatory relationship between the gene and all miRNAs. Disease feature is the known relationship between the disease and all miRNAs. Then, the deep

belief network (DBN) was applied to reduce the dimension of features. Finally, the graph transformer network was implemented to train the model and predict gastric cancer-related genes. The workflow is shown in **Figure 1**.

Construction of Heterogeneous Network

Firstly, we need to calculate the similarity of diseases. Disease ontology (DO) was applied to explore the relationship between different diseases. Every disease term in DO is related to some molecular components (such as genes, proteins, small molecules, and drugs), which are usually called annotation entities of diseases. The similarity between the two diseases is also related to their common ancestors. The similarity of two diseases from the same ancestor node is usually greater than that of two diseases that do not belong to the same ancestor node. Therefore, the similarity of two diseases can be calculated by calculating the amount of information of two disease ancestor nodes. Similarly, in DO, each disease is related to its annotation entity. The similarity between the two diseases can also be calculated by calculating the relationship between their annotation entities.

Then, we need to obtain gene interaction information. We downloaded gene interaction information from HumannetV2.0 (48). The genes that can interact with each other can be connected in the gene network.

Finally, we need to connect these two networks based on the known relationship between diseases and genes. The DisGeNet database (49) was used to obtain the associations between diseases and genes. Based on the information reported by DisGeNet, we can build a heterogeneous network of diseases and genes.

Feature of Diseases and Genes

DIANA-TarBase v8 (50) collected decade-long experimentally supported miRNA–gene interactions. Using this database, we obtained the relationship between genes known to be related to disease and miRNAs. Each miRNA is one dimension of a gene feature. If a gene is reported to be regulated by the miRNA, the feature value of this gene in this characteristic dimension is 1.

Mir2disease (51) contains 349 miRNAs, 163 diseases, 3,273 miRNAs, and the association information between diseases. Using this database, we obtained the relationship between miRNAs and diseases similar to gastric cancer. Each miRNA is one dimension of a disease feature. If a disease is reported to be related to the miRNA, the feature value of this disease in this characteristic dimension is 1.

Dimensionality Reduction by Deep Belief Network

In order to reduce the feature dimension of miRNA, we constructed a DBN network architecture based on Restricted Boltzmann Machine (RBM) for miRNA feature encoding. Each RBM is a layer in the DBN network architecture, and the DBN-based miRNA feature encoding method contains a total of 3 layers of RBMs.

First, the variables in RBM are divided into hidden variables and observable variables. Among them, the observable variables are the features of miRNAs. The observable and hidden

variables are represented by the observable layer and the hidden layer, respectively. The nodes in the RBM layer are not connected, and all the nodes in the adjacent RBM layers are connected to each other. This connection method is consistent with the fully connected neural network.

Unsupervised learning is difficult because the distribution of input miRNA features is unknown. Based on the conclusions of statistical mechanics, we describe the probability distribution with an energy-based model. An RBM is composed of miRNA features and latent variables, whose energy function is defined as:

$$E(v, h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i w_{ij} h_j = -a^T v - b^T h - v^T W h \quad (1)$$

Where the feature of genes can be represented $v = [v_1, v_2, \dots, v_m]^T$; h is the random vector $h = [h_1, h_2, \dots, h_n]^T$ W is the matrix of weight. Both a and b are bias.

With the energy function, the joint probability between the original feature of a gene and the feature after dimensionality reduction can be defined, and the conversion from the visualized node to the hidden node can be realized. Denote the joint probability distribution as $p(v, h)$, which is calculated as follows:

$$p(v, h) = \frac{1}{Z} \exp(-E(v, h)) = \frac{1}{Z} \exp(a^T v) \exp(b^T h) \exp(v^T W h) \quad (2)$$

$$\text{Where } Z = \sum_{v, h} \exp(-E(v, h))$$

is the partition function and can also be called normalization coefficient.

Prediction of Gastric Cancer-Related Genes by the Graph Transformer Network

Since our network is a heterogeneous network of diseases and genes, there are multiple types of meta-paths in it. The first step is to select edge types from the adjacency matrix A . Then, we need to do matrix multiplication of two selected adjacency matrices to learn a novel meta-path network $A^{(1)}$. This new adjacency matrix can be calculated as the sum of candidate adjacency matrices based on weight. The addition process is based on 1*1 convolution with the activation function softmax.

$$Q = F(A, W) = \sigma(A, \text{softmax}(W)) \quad (3)$$

$\sigma()$ represents a convolutional layer and W is the weight of it.

In this way, GTN can generate new meta-path adjacency matrices (52). Then, we can implement graph convolutional network (GCN) on these adjacency matrices. Each GCN layer can be calculated as:

$$H^{(l+1)} = \sigma(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^l W^l) \quad (4)$$

Finally, each node in GTN can be encoded as:

$$Z = \|\sigma(D^{-1} A X W)\| \quad (5)$$

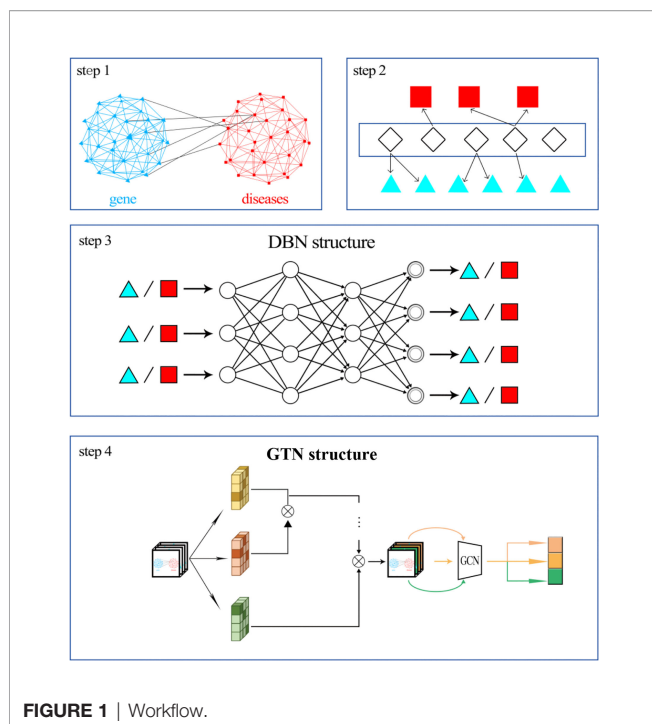


FIGURE 1 | Workflow.

RESULTS

Compare With Traditional Methods

We obtained a total of 435 genes that are reported to be related to gastric cancer. We randomly selected 435 genes as the positive samples and selected part of the remaining genes as negative samples to train the model. We compared DBN-GTN with several traditional methods, which include support vector machine (SVM), back-propagation artificial neural networks (BP-ANN), naive Bayes, and random forest. Since these methods do not have the ability to encode a network, we simply combined the features of genes and diseases to construct a disease-gene pair. We input these disease-gene pairs into these traditional methods and build models to predict gastric cancer-related genes. The performance of these methods is shown in **Table 1**.

As we can see in **Table 1**, DBN-GTN performed best among these methods. The main reason why the accuracy of our analysis of DBN-GTN is significantly higher than other methods is that it considers the association between diseases and the interaction between genes, while other traditional methods are limited by their own shortcomings and cannot incorporate this information into the models.

TABLE 1 | AUC and AUPR of traditional methods and DBN-GTN.

Method	AUC	AUPR
DBN-GTN	0.93	0.86
SVM	0.78	0.68
BP-ANN	0.80	0.73
Naive Bayes	0.72	0.63
Random Forest	0.75	0.69

Compare With Similar Methods

Two methods make up the DBN-GTN, and we try to replace the two methods with similar methods to test whether the accuracy of the method is the highest. DBN mainly plays the function of dimensionality reduction, and principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) have a similar function. Therefore, we try to use these two methods to replace DBN and test the performance. In addition, GCN can be used to encode a homogeneous network. To compare the difference between encoding a heterogeneous network and encoding two homogeneous networks separately, we used GCN to replace GTN. GCN was implemented to encode a gene interaction network and a disease similarity network, respectively. Then, the features of genes and diseases are combined together to train the GCN model. The experimental results are shown in **Figure 2**.

As we can see from **Figure 2**, DBN-GTN performed best among these methods and t-SNE-GCN performed worst. From the impact of dimensionality reduction on accuracy, DBN outperforms t-SNE and PCA, and t-SNE has the worst accuracy. This is because PCA can manually select the amount of information contained after dimensionality reduction, while t-SNE can only reduce the data to 2 to 3 dimensions. From the perspective of the influence of the coding network method on accuracy, the performance of GTN is better than that of GCN. This is because GTN can encode heterogeneous networks and obtain more information than two homogeneous networks by GCN.

CONCLUSION

Biologists discovered some genes related to gastric cancer through large-scale transcriptome and genome sequencing. These results suffer from sample heterogeneity and insufficient sample size. At the same time, these experiments also cost a lot of time and money. Therefore, from the perspective of systems biology, this paper mines the association patterns between diseases and genes, and establishes a model through deep learning algorithms to identify large-scale gastric cancer-related genes. Although a large number of previous studies have used computational methods to identify gastric cancer-related genes, most of these methods focus on extracting information from homogeneous networks and cannot fully incorporate the association between diseases and genes into the model. In this paper, we established a disease similarity network and a gene interaction network, and connected the two networks through the known correlation between the two to form a disease-gene heterogeneous network. At the same time, we extracted the features of the disease and gene based on their relationship with miRNAs. In other words, a bridge between diseases and genes is established through miRNAs. We employ deep belief networks for feature dimensionality reduction and GTN for heterogeneous network encoding. We call this method DBN-GTN. We compare the accuracy of this method with four traditional methods and five similar methods. Experimental results show that DBN-GTN outperforms our chosen

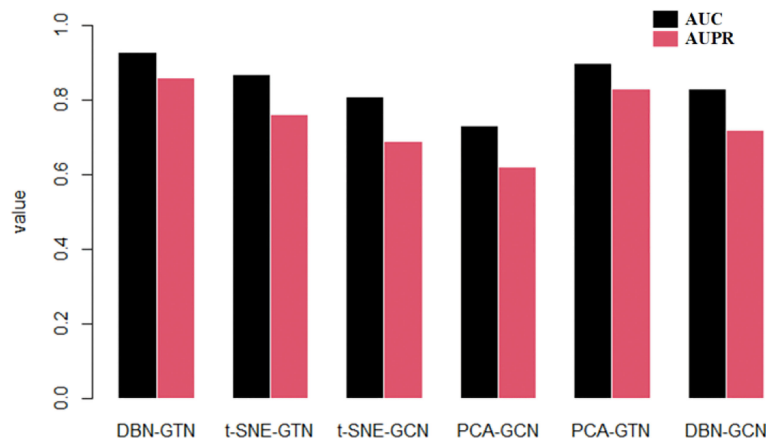


FIGURE 2 | AUC and AUPR of DBN-GTN and similar methods.

traditional method and similar methods, which shows that DBN-GTN is superior in the task of large-scale identification of gastric cancer genes. This paper provides support to further explain the genetic risk, susceptibility, and drug screening of gastric cancer.

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

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and

institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

Experimental design: YC, XS, and JY. Data analysis: YC and XS. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.902616/full#supplementary-material>

REFERENCES

- Thrift AP, El-Serag HB. Burden of Gastric Cancer. *Clin Gastroenterol Hepatol* (2020) 18:534–42. doi: 10.1016/j.cgh.2019.07.045
- Pyo JH, Lee H, Min YW, Min B-H, Lee JH, Kim K-M, et al. Young Age and Risk of Lymph Node Metastasis in Differentiated Type Early Gastric Cancer. *Ann Surg Oncol* (2018) 25:2713–9. doi: 10.1245/s10434-018-6659-3
- Sonkar C, Doharey PK, Rathore AS, Singh V, Kashyap D, Sahoo AK, et al. Repurposing of Gastric Cancer Drugs Against COVID-19. *Comput Biol Med* (2021) 137:104826. doi: 10.1016/j.compbiomed.2021.104826
- Gao F, Li M, Xiang R, Zhou X, Zhu L, Zhai Y. Expression of CLDN6 in Tissues of Gastric Cancer Patients: Association With Clinical Pathology and Prognosis. *Oncol Lett* (2019) 17:4621–5. doi: 10.3892/ol.2019.10129
- Miyamoto R, Inagawa S, Sano N, Tadano S, Adachi S, Yamamoto M. The Neutrophil-to-Lymphocyte Ratio (NLR) Predicts Short-Term and Long-Term Outcomes in Gastric Cancer Patients. *Eur J Surg Oncol* (2018) 44:607–12. doi: 10.1016/j.ejso.2018.02.003
- Shafabakhsh R, Yousefi B, Asemi Z, Nikfar B, Mansournia MA, Hallajzadeh J. Chitosan: A Compound for Drug Delivery System in Gastric Cancer—a Review. *Carbohydr Polymer* (2020) 242:116403. doi: 10.1016/j.carbpol.2020.116403
- Kahroba H, Hejazi MS, Samadi N. Exosomes: From Carcinogenesis and Metastasis to Diagnosis and Treatment of Gastric Cancer. *Cell Mol Life Sci* (2019) 76:1747–58. doi: 10.1007/s00018-019-03035-2
- Biagioni A, Skalamera I, Peri S, Schiavone N, Cianchi F, Giommoni E, et al. Update on Gastric Cancer Treatments and Gene Therapies. *Cancer Metastasis Rev* (2019) 38:537–48. doi: 10.1007/s10555-019-09803-7
- Yusefi AR, Lankarani KB, Bastani P, Radinmanesh M, Kavosi Z. Risk Factors for Gastric Cancer: A Systematic Review. *Asian Pacific J Cancer Prevent: APJCP* (2018) 19:591–603. doi: 10.22034/APJCP.2018.19.3.591
- Li Z, Zhang T, Lei H, Wei L, Liu Y, Shi Y, et al. Research on Gastric Cancer's Drug-Resistant Gene Regulatory Network Model. *Curr Bioinf* (2020) 15:225–34. doi: 10.2174/1574893614666190722102557
- Machlowska J, Baj J, Sitarz M, Maciejewski R, Sitarz R. Gastric Cancer: Epidemiology, Risk Factors, Classification, Genomic Characteristics and Treatment Strategies. *Int J Mol Sci* (2020) 21:4012. doi: 10.3390/ijms21114012
- Guggenheim DE, Shah MA. Gastric Cancer Epidemiology and Risk Factors. *J Surg Oncol* (2013) 107:230–6. doi: 10.1002/jso.23262
- Gu Y, Gao Y, Tang X, Xia H, Shi K. Bioinformatics Analysis Identifies CPZ as a Tumor Immunology Biomarker for Gastric Cancer. *Curr Bioinf* (2021) 16:98–105. doi: 10.2174/1574893615999200707145643

14. Cunningham D, Allum WH, Stenning SP, Thompson JN, Van De Velde CJ, Nicolson M, et al. Perioperative Chemotherapy Versus Surgery Alone for Resectable Gastroesophageal Cancer. *N Engl J Med* (2006) 355:11–20. doi: 10.1056/NEJMoa055531
15. Sasaki CT, Vageli DP. miR-21, miR-155, miR-192, and miR-375 Deregulations Related to NF-kappaB Activation in Gastroduodenal Fluid-Induced Early Preneoplastic Lesions of Laryngeal Mucosa *In Vivo*. *Neoplasia* (2016) 18:329–38. doi: 10.1016/j.neo.2016.04.007
16. Jang S-G, Yoo CW, Park SY, Kang S, Kim HK. Low Expression of miR-449 in Gynecologic Clear Cell Carcinoma. *Int J Gynecol Cancer* (2014) 24:1558–63. doi: 10.1097/IGC.0000000000000267
17. Lau SK, Boutros PC, Pintilie M, Blackhall FH, Zhu C-Q, Strumpf D, et al. Three-Gene Prognostic Classifier for Early-Stage non-Small-Cell Lung Cancer. *J Clin Oncol* (2007) 25:5562–9. doi: 10.1200/JCO.2007.12.0352
18. Yoshihara K, Tajima A, Komata D, Yamamoto T, Kodama S, Fujiwara H, et al. Gene Expression Profiling of Advanced-Stage Serous Ovarian Cancers Distinguishes Novel Subclasses and Implicates ZEB2 in Tumor Progression and Prognosis. *Cancer Sci* (2009) 100:1421–8. doi: 10.1111/j.1349-7006.2009.01204.x
19. Li B-S, Zhao Y-L, Guo G, Li W, Zhu E-D, Luo X, et al. Plasma microRNAs, miR-223, miR-21 and miR-218, as Novel Potential Biomarkers for Gastric Cancer Detection. (2012) 7:e41629. doi: 10.1371/journal.pone.0041629
20. Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor Origin Detection With Tissue-Specific miRNA and DNA Methylation Markers. *Bioinformatics* (2018) 34:398–406. doi: 10.1093/bioinformatics/btx622
21. Fèvre-Montange M, Champier J, Durand A, Wierinckx A, Honnorat J, Guyotat J, et al. Microarray Gene Expression Profiling in Meningiomas: Differential Expression According to Grade or Histopathological Subtype. *Int J Oncol* (2009) 35:1395–407. doi: 10.3892/ijo_00000457
22. Wang Q, Wen Y-G, Li D-P, Xia J, Zhou C-Z, Yan D-W, et al. Upregulated INHBA Expression Is Associated With Poor Survival in Gastric Cancer. *Med Oncol* (2012) 29:77–83. doi: 10.1007/s12032-010-9766-y
23. Liu P, Wang X, Hu C, Hu T. Bioinformatics Analysis With Graph-Based Clustering to Detect Gastric Cancer-Related Pathways. *Genet Mol Res* (2012) 11:3497–504. doi: 10.4238/2012.September.26.5
24. Kurayoshi M, Oue N, Yamamoto H, Kishida M, Inoue A, Asahara T, et al. Expression of Wnt-5a Is Correlated With Aggressiveness of Gastric Cancer by Stimulating Cell Migration and Invasion. *Cancer Res* (2006) 66:10439–48. doi: 10.1158/0008-5472.CAN-06-2359
25. Clements WM, Wang J, Sarnaik A, Kim OJ, Macdonald J, Fenoglio-Preiser C, et al. β -Catenin Mutation Is a Frequent Cause of Wnt Pathway Activation in Gastric Cancer. *Cancer Res* (2002) 62:3503–6. doi: 10.1002/cncr.10589
26. To K, Chan MW, Leung W, Yu J, Tong JH, Lee T, et al. Alterations of Frizzled (FzE3) and Secreted Frizzled Related Protein (hsFRP) Expression in Gastric Cancer. *Life Sci* (2001) 70:483–9. doi: 10.1016/S0024-3205(01)01422-9
27. Yoo YA, Kang MH, Kim JS, Oh SC. Sonic Hedgehog Signaling Promotes Motility and Invasiveness of Gastric Cancer Cells Through TGF- β -Mediated Activation of the ALK5–Smad 3 Pathway. *Carcinogenesis* (2008) 29:480–90. doi: 10.1093/carcin/bgm281
28. Peng L, Ran Y-L, Hu H, Yu L, Liu Q, Zhou Z, et al. Secreted LOXL2 is a Novel Therapeutic Target That Promotes Gastric Cancer Metastasis via the Src/FAK Pathway. *Carcinogenesis* (2009) 30:1660–9. doi: 10.1093/carcin/bgp178
29. Kim SS, Shetty K, Katuri V, Kitisin K, Baek HJ, Tang Y, et al. TGF- β Signaling Pathway Inactivation and Cell Cycle Deregulation in the Development of Gastric Cancer: Role of the β -Spectrin, ELF. *Biochem Biophys Res Commun* (2006) 344:1216–23. doi: 10.1016/j.bbrc.2006.03.236
30. Kang MH, Kim JS, Seo JE, Oh SC, Yoo YA. BMP2 Accelerates the Motility and Invasiveness of Gastric Cancer Cells via Activation of the Phosphatidylinositol 3-Kinase (PI3K)/Akt Pathway. *Exp Cell Res* (2010) 316:24–37. doi: 10.1016/j.yexcr.2009.10.010
31. Lin MT, Lin BR, Chang CC, Chu CY, Su HJ, Chen ST, et al. IL-6 Induces AGS Gastric Cancer Cell Invasion via Activation of the C-Src/RhoA/ROCK Signaling Pathway. *Int J Cancer* (2007) 120:2600–8. doi: 10.1002/ijc.22599
32. Zhao T, Hu Y, Peng J, Cheng L. DeepLGP: A Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* (2020) 36:4466–72. doi: 10.1093/bioinformatics/btaa428
33. Li D, Zhang S, Ma X. Dynamic Module Detection in Temporal Attributed Networks of Cancers. *IEEE/ACM Trans Comput Biol Bioinform* (2021) Epub ahead of print. doi: 10.1109/TCBB.2021.3069441
34. Ma X, Sun PG, Gong M. An Integrative Framework of Heterogeneous Genomic Data for Cancer Dynamic Modules Based on Matrix Decomposition. *IEEE/ACM Trans Comput Biol Bioinform* (2020).
35. Huang Z, Wang Y, Ma X. Clustering of Cancer Attributed Networks by Dynamically and Jointly Factorizing Multi-Layer Graphs. *IEEE/ACM Trans Comput Biol Bioinform* (2021). Epub ahead of print. doi: 10.1109/TCBB.2021.3090586
36. Nguyen TM, Kim N, Kim DH, Le HL, Piran MJ, Um S-J, et al. Deep Learning for Human Disease Detection, Subtype Classification, and Treatment Response Prediction Using Epigenomic Data. *Biomedicine* (2021) 9:1733. doi: 10.3390/biomedicine9111733
37. Wu W, Liu Z, Ma X. jsRC: A Flexible and Accurate Joint Learning Algorithm for Clustering of Single-Cell RNA-Sequencing Data. *Brief Bioinf* (2021) 22:bbaa433. doi: 10.1093/bib/bbaa433
38. Zhao T, Lyu S, Lu G, Juan L, Zeng X, Wei Z, et al. SC2disease: A Manually Curated Database of Single-Cell Transcriptome for Human Diseases. *Nucleic Acids Res* (2021) 49:D1413–9. doi: 10.1093/nar/gkaa838
39. Ma X, Sun P, Gong M. An Integrative Framework of Heterogeneous Genomic Data for Cancer Dynamic Modules Based on Matrix Decomposition. *IEEE/ACM Trans Comput Biol Bioinform* (2022) 19:305–16. doi: 10.1109/TCBB.2020.3004808
40. Peng J, Zhao T. Reduction in TOM1 Expression Exacerbates Alzheimer's Disease. *Proc Natl Acad Sci* (2020) 117:3915–6. doi: 10.1073/pnas.1917589117
41. Zhao T, Hu Y, Zang T, Cheng L. MRTFB Regulates the Expression of NOMO1 in Colon. *Proc Natl Acad Sci* (2020) 117:7568–9. doi: 10.1073/pnas.2000499117
42. Yao S, You R, Wang S, Xiong Y, Huang X, Zhu S. NetGO 2.0: Improving Large-Scale Protein Function Prediction With Massive Sequence, Text, Domain, Family and Network Information. *Nucleic Acids Res* (2021) 49:W469–75. doi: 10.1093/nar/gkab398
43. Zhao T, Liu J, Zeng X, Wang W, Li S, Zang T, et al. Prediction and Collection of Protein–Metabolite Interactions. *Brief Bioinf* (2021) 22:bbab014. doi: 10.1093/bib/bbab014
44. Kim Y, Kim GB, Lee SY. Machine Learning Applications in Genome-Scale Metabolic Modeling. *Curr Opin Syst Biol* (2021) 25:42–9. doi: 10.1016/j.coisb.2021.03.001
45. Zhao T, Hu Y, Cheng L. Deep-DRM: A Computational Method for Identifying Disease-Related Metabolites Based on Graph Deep Learning Approaches. *Brief Bioinf* (2021) 22:bbaa212. doi: 10.1093/bib/bbaa212
46. Abbasi K, Razzaghi P, Poso A, Ghanbari-Ara S, Masoudi-Nejad A. Deep Learning in Drug Target Interaction Prediction: Current and Future Perspectives. *Curr Med Chem* (2021) 28:2100–13. doi: 10.2174/0929867327666200907141016
47. Zhao T, Hu Y, Valsdottir LR, Zang T, Peng J. Identifying Drug–Target Interactions Based on Graph Convolutional Network and Deep Neural Network. *Briefings Bioinf* (2021) 22:2141–50. doi: 10.1093/bib/bbaa044
48. Hwang S, Kim CY, Yang S, Kim E, Hart T, Marcotte EM, et al. HumanNet V2: Human Gene Networks for Disease Research. *Nucleic Acids Res* (2019) 47:D573–80. doi: 10.1093/nar/gky1126
49. Piñero J, Bravo JJ, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants. *Nucleic Acids Res* (2017) 45(D1):D833–9. doi: 10.1093/nar/gkw943
50. Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, et al. DIANA-TarBase V8: A Decade-Long Collection of Experimentally Supported miRNA–Gene Interactions. *Nucleic Acids Res* (2018) 46:D239–45. doi: 10.1093/nar/gkx1141
51. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, et al. Mir2disease: A Manually Curated Database for microRNA Deregulation in Human Disease. *Nucleic Acids Res* (2009) 37:D98–D104. doi: 10.1093/nar/gkn714

52. Yun S, Jeong M, Kim R, Kang J, Kim HJ. Graph Transformer Networks. *Adv Neural Inf Process Syst* (2019) 32:1–11. doi: 10.48550/arXiv.1911.06455

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.

Copyright © 2022 Chen, Sun and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Multi-Omics Integration-Based Prioritisation of Competing Endogenous RNA Regulation Networks in Small Cell Lung Cancer: Molecular Characteristics and Drug Candidates

OPEN ACCESS

Edited by:

Tianyi Zhao,
Harbin Institute of Technology, China

Reviewed by:

Tuantuan Zhao,
Mayo Clinic, United States
Lin Hua,
Hebei General Hospital, China

*Correspondence:

Wei-Dong Hu
weidong_618@163.com
Min Zhang
sallyzhangmin@126.com
Jing Gao
jing.gao@helsinki.fi

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 26 March 2022

Accepted: 02 June 2022

Published: 04 July 2022

Citation:

Wang X-J, Gao J, Yu Q, Zhang M
and Hu W-D (2022) Multi-Omics
Integration-Based Prioritisation
of Competing Endogenous
RNA Regulation Networks in
Small Cell Lung Cancer: Molecular
Characteristics and Drug Candidates.
Front. Oncol. 12:904865.
doi: 10.3389/fonc.2022.904865

Xiao-Jun Wang^{1,2†}, Jing Gao^{1,2,3,4*†}, Qin Yu², Min Zhang^{5*} and Wei-Dong Hu^{1*}

¹ Department of Respiratory Medicine, Gansu Provincial Hospital, Lanzhou, China, ² The First School of Clinical Medicine, Lanzhou University, Lanzhou, China, ³ Respiratory Medicine Unit, Department of Medicine, Karolinska Institute, Stockholm, Sweden, ⁴ Department of Pulmonary Medicine, University of Helsinki and Helsinki University Hospital, Helsinki, Finland, ⁵ Department of Pathology, Gansu Provincial Hospital, Lanzhou, China

Background: The competing endogenous RNA (ceRNA) network-mediated regulatory mechanisms in small cell lung cancer (SCLC) remain largely unknown. This study aimed to integrate multi-omics profiles, including the transcriptome, regulome, genome and pharmacogenome profiles, to elucidate prioritised ceRNA characteristics, pathways and drug candidates in SCLC.

Method: We determined the plasma messenger RNA (mRNA), microRNA (miRNA), long noncoding RNA (lncRNA) and circular RNA (circRNA) expression levels using whole-transcriptome sequencing technology in our SCLC plasma cohort. Significantly expressed plasma mRNAs were then overlapped with the Gene Expression Omnibus (GEO) tissue mRNA data (GSE 40275, SCLC tissue cohort). Next, we applied a multistep multi-omics (transcriptome, regulome, genome and pharmacogenome) integration analysis to first construct the network and then to identify the lncRNA/circRNA-miRNA-mRNA ceRNA characteristics, genomic alterations, pathways and drug candidates in SCLC.

Results: The multi-omics integration-based prioritisation of SCLC ceRNA regulatory networks consisted of downregulated mRNAs (CSF3R/GAA), lncRNAs (AC005005.4-201/DLX6-AS1-201/NEAT1-203) and circRNAs (hsa_HLA-B_1/hsa_VEGFC_8) as well as upregulated miRNAs (hsa-miR-4525/hsa-miR-6747-3p). lncRNAs (lncRNA-AC005005.4-201 and NEAT1-203) and circRNAs (circRNA-hsa_HLA-B_1 and hsa_VEGFC_8) may regulate the inhibited effects of hsa-miR-6747-3p for CSF3R expression in SCLC, while lncRNA-DLX6-AS1-201 or circRNA-hsa_HLA-B_1 may neutralise the negative regulation of hsa-miR-4525 for GAA in SCLC. CSF3R and GAA were present in the genomic alteration, and further identified as targets of Favld and

Trastuzumab deruxtecan, respectively. In the SCLC-associated pathway analysis, CSF3R was involved in the autophagy pathways, while GAA was involved in the glucose metabolism pathways.

Conclusions: We identified potential lncRNA/cirRNA-miRNA-mRNA ceRNA regulatory mechanisms, pathways and promising drug candidates in SCLC, providing novel potential diagnostics and therapeutic targets in SCLC.

Keywords: small cell lung cancer (SCLC), multi-omics integration, competing endogenous RNA (ceRNA), long noncoding RNA (lncRNA), circular RNA (circRNA), microRNA (miRNA)

INTRODUCTION

Small cell lung cancer (SCLC) is a highly heterogeneous malignancy of neuroendocrine origin accounting for approximately 15% of all cases of lung cancer. SCLC is characterised by the early development of metastases, rapid recurrence and a low survival rate (1–4). The 5-year overall survival rate in SCLC barely reaches 5%, while average overall survival reaches only 2 to 4 months in untreated patients (1, 5, 6). Early diagnosis of SCLC remains quite challenging given its nonspecific symptoms and fast-growing tumours (7). Currently, chemotherapy and immunotherapy represent the most common treatment for SCLC, whereby chemotherapy alone remains the basis of standard treatment for the management of SCLC (7). While the initial response rate for first-line chemotherapy reaches approximately 60% in SCLC, patients may still quickly succumb given rapid recurrence following chemotherapy, primary or secondary drug resistance and ineffective second-line treatment options (8–10). Thus, limited effective therapies remain the primary reason for poor outcomes in SCLC (7, 8). The mechanisms behind the pathogenesis of SCLC are complex, and as yet unexplained by a single biomarker or specific mechanism (11). As such, an increased and comprehensive understanding of SCLC characteristics is crucial to guiding both diagnosis and treatment. Omics studies are emerging rapidly and offer tremendous potential to better understand the underlying disease mechanisms, as well as advancing early diagnostics and identifying potential drug targets.

Competitive endogenous RNA (ceRNA) is a novel layer of gene regulation in diseases, regulating each other at the post-transcription level by competing for shared microRNAs (miRNAs) (12). ceRNA networks link the function of protein-coding messenger RNA (mRNA) with noncoding RNAs (ncRNAs), which primarily include long noncoding RNAs (lncRNAs), circular RNAs (circRNAs) and miRNAs (12–15). The integrative assessment of the expressions of lncRNAs, circRNAs, miRNAs and mRNAs construct ceRNA networks (14–18). Several studies demonstrated that lung cancer associates with the dysregulation of the expression of ncRNAs including both lncRNAs and miRNAs, and the expression of several signalling pathways and oncogenes, while circRNAs may play a key role in lung cancer tumorigenesis, progression, invasion and metastasis (14, 18). miRNAs could control the target genes involved in cellular processes by downregulating gene expression through repressing or degrading mRNA targets (19–21). In addition, the majority of lncRNAs

compete with miRNAs to prevent miRNA binding to their target mRNA, leading to the transcriptional activation of target genes (22, 23). Furthermore, after binding to several sites for a particular miRNA or RNA-binding proteins (RBPs), circRNAs regulate alternative splicing and gene transcription through interaction (15, 23, 24). Consequently, these aberrantly expressed transcripts in the ceRNA network may represent potential therapeutic targets, diagnostic markers and prognostic markers in SCLC. In addition to transcriptomics, gene mutations play significant roles in new drug development in cancer. For instance, gene mutation profiles have facilitated the development of targeted agents in therapeutics for adenocarcinomas of the lung (25). Drug databases are developing rapidly, and the integrative analysis of omics data and drug databases provide us with excellent opportunities for drug development such as through pharmacogenomics (26). The rapidly expanding field of systems biology has proven reasonably effective at summarising knowledge related to cancer pathways, perhaps most importantly using the cancer literature to elucidate the molecular networks *via* which cancer develops. Thus, methodology which employs an integrative analysis of the literature could contribute to understanding the SCLC pathways (27).

In an attempt to understand the complexity and heterogeneity of SCLC, our study aimed to identify plasma mRNAs and compare them with the expression levels found in tissue to identify SCLC-specific mRNAs (28, 29) and, further, to evaluate the lncRNA/circRNA-miRNA-mRNA ceRNA regulatory network. Next, we applied a multi-omics integration analysis (transcriptome, regulome, genome and pharmacogenome) to discuss ceRNA regulation, genomic alterations, pathways and drug candidates in SCLC (see **Figure 1**) (30–32). Understanding the characteristics of the ceRNA regulatory network can potentially shed light on the screening of SCLC biomarkers, particularly those related to genomic alterations and novel therapeutic targets.

MATERIALS AND METHODS

In-House SCLC Plasma Cohort and SCLC Lung Tissue Cohort

In this study, we analysed two SCLC cohorts: an in-house SCLC plasma cohort ($n = 12$) and an SCLC lung tissue cohort (from GSE40275, $n = 62$) (33). The mRNA data in the SCLC tissue cohort were obtained from the lung tissue samples of SCLCs and adjacent nontumour regions. Our in-house SCLC plasma cohort includes

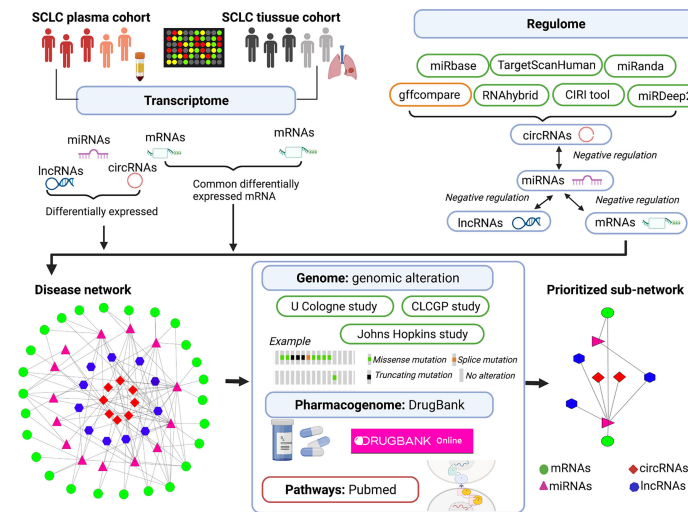


FIGURE 1 | Illustration of multi-omics-based prioritisation of ceRNAs and pathways. CLCGP, Clinical Lung Cancer Genome Project; CIRI, circRNA identifier; ceRNA, competitive endogenous RNA; circRNA, circular RNAs; DE, differentially expressed; SCLC, small cell lung cancer; lncRNA, long noncoding RNA; miRNA, microRNA; mRNA, messenger RNA; cBioPortal database (<https://www.cbioportal.org/datasets>); DrugBank database (<https://go.drugbank.com/>); Genecards database (<https://www.genecards.org/>); PubMed (<https://pubmed.ncbi.nlm.nih.gov/>).

eight SCLC patients and four healthy controls, collected between August and November 2020 at Gansu Provincial Hospital, China. The inclusion criteria of patients in our SCLC plasma cohort consisted of a histologically or cytologically confirmed initial SCLC without previous chemotherapy, radiotherapy, molecular-targeted therapy, immunotherapy or surgery. We excluded patients from our SCLC plasma cohort based on the following: (1) presence of other combined cancers; (2) pregnant or lactating patient; and (3) presentation with cardiopulmonary insufficiency, serious cardiovascular disease, a serious infection or severe malnutrition (34, 35). The mRNA data in the SCLC tissue cohort were obtained from the lung tissue samples of SCLCs and adjacent nontumour regions. In addition, the tissue mRNA expression levels were evaluated in the Gene Expression Omnibus database (GEO, <https://www.ncbi.nlm.nih.gov/gds/>) using the term “small cell lung cancer” with “homo sapiens”, “series” and “expression profiling by array”. The 19 SCLC lung tissue datasets were obtained, and no suitable plasma SCLC dataset could be extracted. Finally, we selected the GSE40275 tissue dataset of SCLC for further analysis, since this dataset was obtained from a single-sequencing platform, thereby avoiding a potential bias from inconsistencies in probes stemming from different sequencing platforms. This cohort study received ethical approval from the Ethics Committee of the Gansu Provincial Hospital, China (27 July 2020, No. 2020-183). Informed consent was obtained from all participants in the whole-transcriptome sequencing experiment, and the research adhered to the principles of the Declaration of Helsinki.

Whole-Transcriptome Sequencing Analysis in the Plasma SCLC Cohort

We determined the plasma messenger RNA (mRNA), microRNA (miRNA), long noncoding RNA (lncRNA) and circular RNA (circRNA) expression levels using the whole-

transcriptome sequencing technology in our SCLC plasma cohort. The extraction of total RNA from the plasma samples relied on the miRNeasy Mini Kit (Qiagen, Hilden, Germany) following the manufacturer’s protocol. The details appear in **Supplemental File 1**. A total of 1.5-μg RNA per sample was used as the input material for the lncRNA sequencing analysis, and a total of 2.5-ng RNA was used as the input material for the miRNA sequencing analysis. The details of the lncRNA and miRNA sequencing appear in **Supplemental File 2**. In addition, our SCLC plasma data were uploaded to a public platform [uploaded to the Sequence Read Archive (SRA) database (BioProject PRJNA 759049 (miRNA data) and BioProject PRJNA 762578 (mRNA, lncRNA and circRNA data)].

Identification of Differentially Expressed mRNA, miRNA, circRNA and lncRNA in SCLC

The significant differentially expressed mRNAs (DEmRNAs) in the SCLC tissue cohort were identified by comparing SCLC lung tissue and adjacent nontumour tissue from SCLC using the GEO2R tools from the R package “limma” in GSE40275 [fold change (FC) > 1.5, $p < 0.05$, and false discovery rate (FDR) < 0.2]. DEmRNAs in the SCLC plasma cohort were identified by comparing SCLC and healthy samples using the likelihood ratio test (LRT) in the R package “DESeq” (|FC| > 1.5, $p < 0.05$). Then, the commonly expressed DEmRNAs (Co-DEmRNAs, SCLC-specific mRNAs) were defined as the overlapping DEmRNAs between the SCLC plasma cohort and the SCLC lung tissue cohort (|FC| > 1.5, $p < 0.05$). The significant DE miRNAs, DE circRNAs and DE lncRNAs in the SCLC plasma cohort were identified by comparing SCLC and healthy plasma samples using

LRT in the R package “DESeq” ($|FC| > 1.5$, $p < 0.05$, and $FDR < 0.2$). FDR was computed using the methodology described by Benjamini and Hochberg (36). The volcano plots were created using the R package “ggplot2”. Finally, the Co-DEmRNAs, DEmiRNAs, DEcircRNAs and DELncRNAs were subsequently used in the ceRNA network construction.

Construction of the lncRNA/circRNA-miRNA-mRNA ceRNA-Mediated Regulatory Network

The previous step identifying the DEmiRNAs, DELncRNAs, DEcircRNAs and Co-DEmRNAs in SCLC was used to construct the lncRNA/circRNA-miRNA-mRNA ceRNA regulatory network. The regulome analysis was based on the targeted mRNA-miRNA, lncRNA-miRNA and circRNA-miRNA prediction using online analytical software tools. The targeted mRNAs of the miRNAs were predicted using two online analytical software tools: miRanda (version 3.3.a) (37) and TargetScanHuman database (version 5.0) (38). The targeted lncRNAs of the miRNAs were predicted using the online analytical software tools from the miRbase database (version 22.0) (37). The targeted circRNAs of the miRNAs were predicted using three online analytical software tools: RNAhybrid database (version 2.1.1) (39), miRanda (version 3.3.a) (40) and TargetScanHuman database (version 5.0) (38). The negative regulation of mRNA-miRNA, lncRNA-miRNA and circRNA-miRNA was selected in the further ceRNA network construction. Next, the lncRNAs, circRNAs and miRNAs were identified as known or novel using several analytical software tools: the gffcompare program (41), the circRNA identifier (CIRI) tool (42), the miRbase database (version 22.0) (37) and the miRDeep2 tools (43). Based on these results, we constructed the lncRNA/circRNA-miRNA-mRNA ceRNA regulatory network using the Cytoscape software (version 3.7.0) (44). Next, the differentially expressed lncRNA, circRNAs, miRNAs and mRNAs in the SCLC ceRNA network were analysed using the gene ontology (GO) analysis and the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway analysis. For the GO analysis, the differentially expressed lncRNA, circRNAs, miRNAs and mRNAs were classified into three categories: biological process (BP), cellular component (CC) and molecular function (MF). The KEGG pathway analysis was performed to analyse the potential pathways enriched by the differentially expressed lncRNA, circRNAs, miRNAs and mRNAs. The enrichment analysis was evaluated using the R package ClusterProfiler (45), for which we considered an adjusted $p < 0.05$ as statistically significant (46).

Evaluation of Genomic Alterations, Drug Candidates/Repurposing and Pathway Analysis in SCLC ceRNA Networks

The genomic alterations of mRNAs in the SCLC ceRNA network were determined through three datasets (47–49) from the cBioPortal database (<https://www.cbioportal.org/datasets>), including the Clinical Lung Cancer Genome Project (CLCGP) study (47), the Johns Hopkins study (48) and the University of Cologne study (U Cologne study) (49). The pharmacogenomics

data were downloaded from the DrugBank database (release 5.0) (<https://go.drugbank.com/>), including the rich drugs data and the drug-target genes data (50). The results obtained from the pharmacogenomics DrugBank database were further mined through the “Targets” tool using manual searches. The pathways of the mRNAs were first evaluated and annotated using the Genecards database (<https://www.genecards.org/>) (51), then the SCLC-associated pathways were further filtered through a literature search from PubMed (<https://pubmed.ncbi.nlm.nih.gov/>) using the terms “small cell lung cancer [Title/Abstract] OR SCLC [Title/Abstract] OR small cell lung cancer [MeSH Terms]” and “pathways [Title/Abstract]”.

RESULTS

Identification of Differentially Expressed mRNA, miRNA, circRNA and lncRNA in SCLC

We identified eight SCLC patients (62.5% male, median age of 62 years, 100% Asian and 50.0% advanced stage) and four healthy controls (75.0% male, median age of 66 years) in our SCLC plasma cohort, and 19 SCLC patients (84.2% male, median age of 66 years and 100% European) in the SCLC tissue cohort (GSE40275) (Table 1). Through our in-house whole-transcriptome sequencing data comparing SCLC plasma samples and healthy plasma samples, we harvested a total of 652 DE mRNAs (326 upregulated and 326 downregulated), 281 DE miRNAs (178 upregulated and 103 downregulated), 286 DE circRNAs (166 upregulated and 120 downregulated) and 1753 DE lncRNAs (1036 upregulated and 717 downregulated) for subsequent analysis. Overall, 8429 DE mRNAs (4808 upregulated and 3621 downregulated) were identified in the SCLC tissue cohort, ultimately resulting in 135 DE mRNAs (32 upregulated and 103 downregulated) expressed in two cohorts as common DE mRNAs (Co-DE mRNAs), and also identified as SCLC-specific mRNAs (Figure 2).

Construction of the lncRNA/circRNA-miRNA-mRNA ceRNA Network

The obtained 281 DE miRNAs, 1753 DE lncRNAs, 286 DE circRNAs and 135 Co-DE mRNAs in SCLC were initially involved in the ceRNA regulatory network construction. Integrating the selection rules described in the methods section, the SCLC lncRNA/circRNA-miRNA-mRNA ceRNA regulatory network was constructed, which included 58 mRNAs (4 upregulated and 54 downregulated), 301 lncRNAs (40 upregulated and 261 downregulated), 16 circRNAs (5 upregulated and 11 downregulated) and 24 miRNAs (20 upregulated and 4 downregulated) (Figures 3 and 4; Supplemental Tables 1 and 2). The lncRNA-miRNA-mRNA ceRNA regulatory network consisted of 381 nodes (301 lncRNAs, 23 miRNAs and 57 mRNAs) with 707 edges (Figure 3). In the lncRNA-miRNA-mRNA ceRNA network, the expression levels of 53 mRNAs and 261 lncRNAs decreased in SCLC and the expression levels of 19 miRNAs

TABLE 1 | Patient characteristics for the in-house SCLC plasma cohort (n = 12) and SCLC lung tissue cohort (from GSE40275, n = 62).

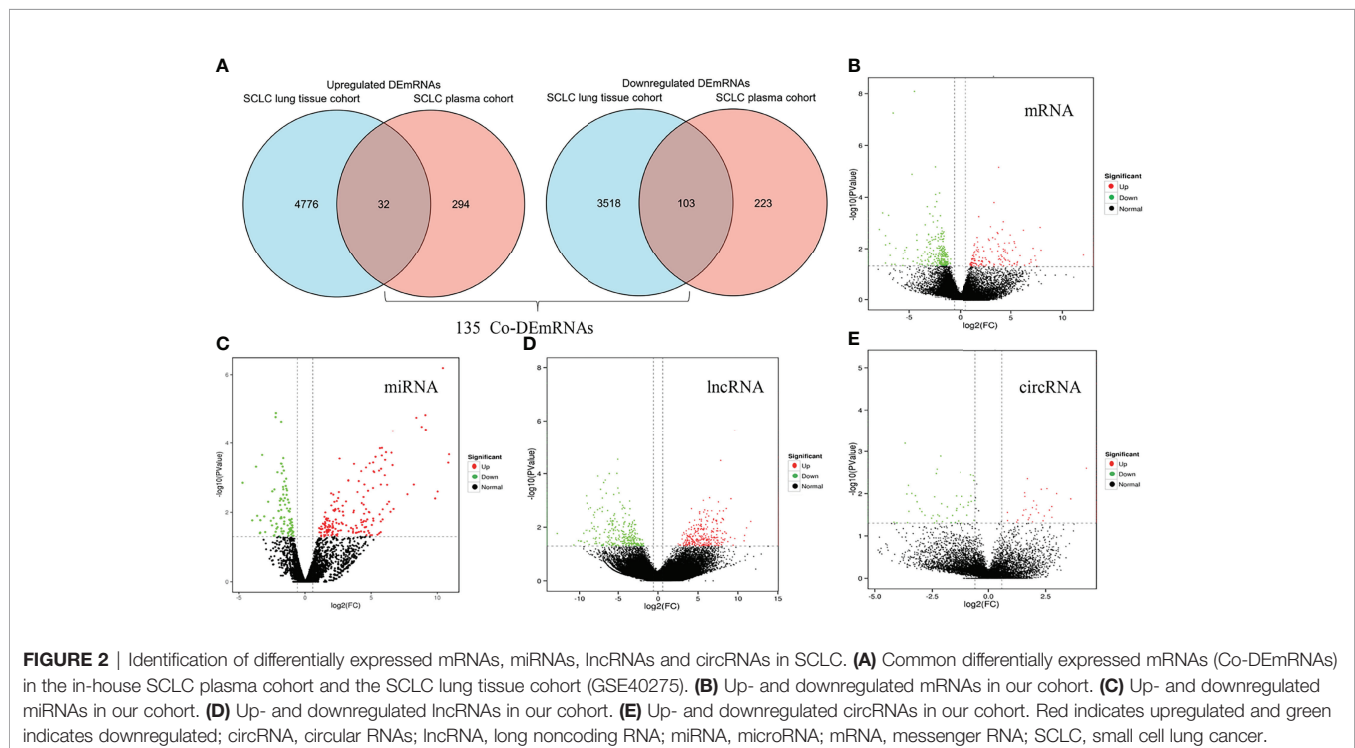
Patient characteristics		SCLC lung tissue cohort (GSE40275)		In-house SCLC plasma cohort	
		normal	SCLC patients	normal	SCLC patients
Age (median, in years)		66	70	66	61.5
Sex (males, %)		19 (44.2%)	16 (84.2%)	3 (75.0%)	5 (62.5%)
Country		Austria	Austria	China	China
Ethnicity		Austrian	Austrian	Asian	Asian
AJCC stage					
	Stage I	—	9 (47.4%)	—	0
	Stage II	—	4 (20.1%)	—	1 (12.5%)
	Stage III	—	6 (31.6%)	—	3 (37.5%)
	Stage IV	—	0	—	4 (50%)
VALSG stage					
	Extended stage	—	0	—	4 (50%)
	Limited stage	—	16 (100%)	—	4 (50%)
Outcome					
	Dead	—	NA	—	8 (100%)
	Living	—	NA	—	0

SCLC, small cell lung cancer; AJCC, American Joint Committee on Cancer; VALSG, Veterans Administration Lung Study Group.
NA, not available.

increased in SCLC, while the expression levels of 4 mRNAs and 40 lncRNAs increased in SCLC and the expression levels of 4 miRNAs decreased in SCLC (**Supplemental Table 1**). The circRNA-miRNA-mRNA ceRNA network consisted of 82 nodes (16 circRNAs, 19 miRNAs and 47 mRNAs) with 165 edges (**Figure 4**). In the circRNA-miRNA-mRNA ceRNA network, the expression levels of 43 mRNAs and 11 circRNAs decreased in SCLC and the expression levels of 16 miRNAs increased in SCLC, while the expression levels of four mRNAs and five circRNAs increased in SCLC and the expression levels of three miRNAs decreased in SCLC (**Supplemental Table 2**).

Functional Enrichment Analysis of mRNA, miRNA, circRNA and lncRNA in the ceRNA Network in SCLC

The differentially expressed levels of 58 mRNAs in the ceRNA network appear in **Table 2**. In the SCLC plasma cohort, the top three downregulated genes in the fold change (FC) were early growth response 1 (EGR1), complement factor D (CFD) and FosB proto-oncogene AP-1 transcription factor subunit (FOSB), while the top three upregulated genes in FC were zinc finger protein 704 (ZNF704), NOVA alternative splicing regulator 1 (NOVA1) and attractin like 1 (ATRNL1) (**Table 2**). **Table 3**



LncRNA-miRNA-mRNA ceRNA network in SCLC

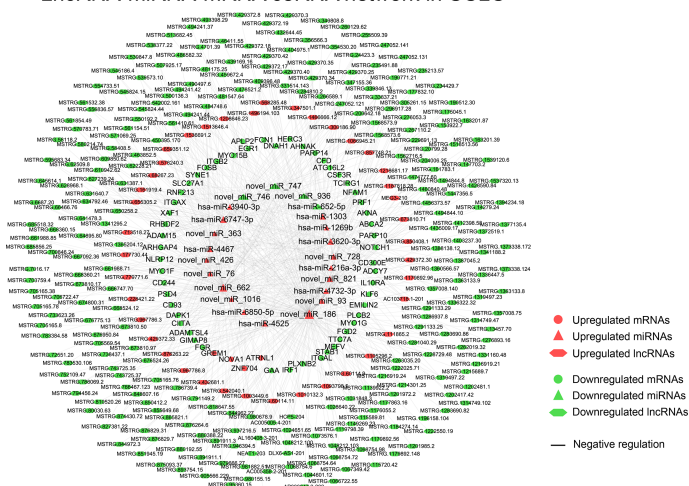


FIGURE 3 | The lncRNA-miRNA-mRNA ceRNAs network in SCLC. lncRNA, long noncoding RNA; miRNA, microRNA; mRNA, messenger RNA; SCLC, small cell lung cancer.

summarises 23 results from 58 mRNAs in the ceRNA network included in the GO analysis. This GO analysis indicated that the DEMRNAs were associated with numerous important biological processes and cellular components. The present study indicated that the biological processes of DEMRNAs primarily included processes such as neutrophil degranulation, neutrophil activation involved in the immune response, neutrophil activation, neutrophil-mediated immunity and an integrin-mediated signalling pathway among others. These biological functions associate with the protumour/prometastatic roles of inflammatory cells in cancer development and metastasis (Table 3) (52, 53). In terms of the cellular components, they mainly included the protein complex involved in cell adhesion and the integrin complex (Table 3), functions associated with

tumorigenesis (54, 55). In addition, no results were obtained from the molecular function of the GO analysis and the KEGG pathways analysis, given that adjusted $p > 0.05$ in these functional analyses. In addition, we also reported the differentially expressed levels of lncRNAs, circRNAs and miRNAs in the ceRNA network (Supplemental Tables 3-5). The functional GO analyses primarily revealed cell survival and proliferation in 42 functional results from 301 lncRNAs, the inflammatory and immune response function in 32 functional results from 32 circRNAs and inflammatory and immune response and cell proliferation in 66 functional results from 24 miRNAs, respectively (Tables 4-6). Among these functions, many tumour-related terms were significantly enriched, such as regulating the cell cycle, the negative regulation of cell growth,

circRNA-miRNA-mRNA ceRNA network in SCLC

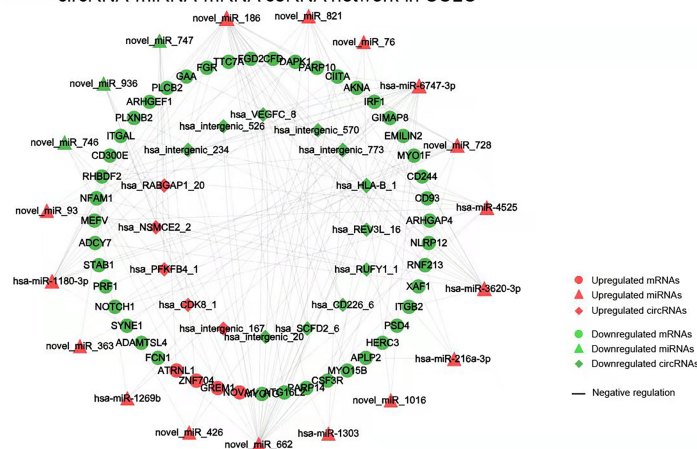


FIGURE 4 | The circRNA-miRNA-mRNA ceRNAs network in SCLC. circRNA, circular RNA; miRNA, microRNA; mRNA, messenger RNA; SCLC, small cell lung cancer.

TABLE 2 | Differentially expressed levels and genomic alterations of mRNAs in the ceRNA regulatory network in SCLC.

Gene symbol	Gene full name	Differentially expressed levels					Genomic alterations		
		In-house SCLC plasma cohort		SCLC lung tissue cohort (GSE40275)		Regulated	CLCGP, <i>Nat Genet</i> 2012	Johns Hopkins, <i>Nat Genet</i> 2012	U Cologne, <i>Nature</i> 2015
		log2FC	p value	log2FC	p value				
Genomic alterations (n = 50)									
EGR1	Early Growth Response 1	-3.232	2.30E-04	-2.611	3.42E-16	down	3.0%	0	0
CFD	Complement Factor D	-2.898	2.11E-02	-2.472	2.01E-25	down	0	0	0.8%
ABCA2	ATP Binding Cassette Subfamily A Member 2	-2.814	3.54E-03	-1.923	3.00E-04	down	3.0%	1.3%	2.5%
PRF1	Perforin 1	-2.699	5.32E-04	-2.038	5.60E-20	down	3.0%	0	1.7%
STAB1	Stabilin 1	-2.484	2.34E-04	-1.151	2.39E-14	down	3.0%	0	4.0%
AHNAK	AHNAK Nucleoprotein	-2.443	6.74E-06	-2.761	8.93E-29	down	7.0%	4.0%	6.0%
CD300E	CD300e Molecule	-2.428	7.77E-05	-1.009	3.72E-15	down	0	0	0.8%
CD244	CD244 Molecule	-2.332	2.57E-02	-0.751	3.88E-15	down	3.0%	0	0.8%
SLC27A1	Solute Carrier Family 27 Member 1	-2.331	3.90E-02	-0.76	6.68E-14	down	7.0%	1.3%	1.7%
PARP10	Poly (ADP-Ribose) Polymerase Family Member 10	-2.12	2.62E-02	-0.601	4.27E-10	down	3.0%	0	0.8%
MEFV	MEFV Innate Immuity Regulator, Pyrin	-2.051	9.52E-03	-1.031	2.62E-19	down	0	1.3%	1.7%
RHBDF2	Rhomboid 5 Homolog 2	-2.028	3.29E-02	-0.981	5.62E-14	down	3.0%	0	0
DNAH1	Dynein Axonemal Heavy Chain 1	-2.02	1.14E-02	-0.653	2.03E-18	down	0	0	7.0%
TCIRG1	T Cell Immune Regulator 1, ATPase H+-Transporting V0 Subunit A3	-1.998	9.10E-03	-1.421	4.46E-17	down	0	0	1.7%
NFAM1	NFAT Activating Protein With ITAM Motif 1	-1.976	4.41E-02	-0.708	2.61E-13	down	0	1.3%	0.8%
GIMAP8	GTPase, IMAP Family Member 8	-1.902	1.10E-02	-2.078	2.10E-31	down	10.0%	0	3.0%
PLXNB2	Plexin B2	-1.896	3.39E-03	-1.094	1.38E-09	down	7.0%	1.3%	3.0%
FGD2	FYVE, RhoGEF And PH Domain Containing 2	-1.885	3.07E-03	-1.384	7.07E-19	down	0	0	0.8%
NLRP12	NLR Family Pyrin Domain Containing 12	-1.862	2.96E-02	-0.852	4.45E-16	down	7.0%	1.3%	4.0%
NOTCH1	Notch Receptor 1	-1.848	2.58E-02	-1.497	3.76E-22	down	10.0%	1.3%	13.0%
FCN1	Ficolin 1	-1.844	7.66E-03	-1.675	3.00E-23	down	0	0	2.5%
CSF3R	Colony-stimulating factor 3 receptor	-1.801	2.63E-03	-2.469	2.01E-29	down	7.0%	1.3%	2.5%
GAA	Acid alpha-glucosidase	-1.789	3.85E-02	-1.108	5.29E-13	down	3.0%	1.3%	2.5%
ITGB2	Integrin Subunit Beta 2	-1.756	9.89E-03	-1.813	1.26E-11	down	3.0%	0	2.5%
EMILIN2	Elastin Microfibril Interfacer 2	-1.748	8.86E-03	-1.372	1.79E-18	down	0	2.5%	2.5%
ARHGAP4	Rho GTPase Activating Protein 4	-1.741	1.37E-02	-0.624	3.80E-07	down	3.0%	1.3%	4.0%
CD93	CD93 Molecule	-1.722	2.15E-02	-2.668	4.54E-34	down	3.0%	0	1.7%
DAPK1	Death Associated Protein Kinase 1	-1.707	1.97E-02	-1.123	5.50E-05	down	3.0%	2.5%	4.0%

(Continued)

TABLE 2 | Continued

Gene symbol	Gene full name	Differentially expressed levels				Genomic alterations			
		In-house SCLC plasma cohort		SCLC lung tissue cohort (GSE40275)		Regulated	CLCGP, <i>Nat Genet</i> 2012	Johns Hopkins, <i>Nat Genet</i> 2012	U Cologne, <i>Nature</i> 2015
		log2FC	p value	log2FC	p value				
TTC7A	Tetratricopeptide Repeat Domain 7A	-1.651	2.83E-02	-1.265	3.85E-20	down	0	1.3%	2.5%
PSD4	Pleckstrin And Sec7 Domain Containing 4	-1.632	1.74E-02	-0.802	3.80E-11	down	3.0%	1.3%	3.0%
CIITA	Class II Major Histocompatibility Complex Transactivator	-1.624	2.17E-03	-1.777	3.50E-17	down	0	1.3%	0
SYNE1	Spectrin Repeat Containing Nuclear Envelope Protein 1	-1.606	3.16E-03	-1.689	1.78E-19	down	28.0%	11.0%	23.0%
ITGAX	Integrin Subunit Alpha X	-1.592	1.15E-02	-2.083	9.75E-18	down	3.0%	1.3%	3.0%
ADAMTSL4	ADAMTS Like 4	-1.555	3.60E-02	-1.606	2.64E-22	down	0	0	2.5%
XAF1	XIAP Associated Factor 1	-1.552	1.88E-02	-1.445	3.44E-10	down	3.0%	1.3%	0
FGR	FGR Proto-Oncogene, Src Family Tyrosine Kinase	-1.488	2.02E-02	-2.179	5.88E-22	down	0	2.5%	0.8%
PLCB2	Phospholipase C Beta 2	-1.474	1.89E-02	-1.634	8.74E-19	down	0	1.3%	0
APLP2	Amyloid Beta Precursor Like Protein 2	-1.47	2.22E-02	-0.935	5.30E-17	down	5.0%	2.5%	0
AKNA	AT-Hook Transcription Factor	-1.467	4.69E-02	-1.126	7.79E-20	down	7.0%	2.5%	1.7%
RNF213	Ring Finger Protein 213	-1.452	1.46E-02	-0.714	8.47E-06	down	0	4.0%	2.5%
HERC3	HECT And RLD Domain Containing E3 Ubiquitin Protein Ligase 3	-1.45	4.01E-02	-0.725	1.92E-16	down	0	0	0.8%
ARHGEF1	Rho Guanine Nucleotide Exchange Factor 1	-1.443	3.72E-02	-0.724	6.28E-09	down	0	1.3%	2.5%
MYO1F	Myosin 1F	-1.394	4.04E-02	-2.014	2.90E-22	down	3.0%	1.3%	2.5%
MYO1G	Myosin 1G	-1.314	3.45E-02	-1.526	3.29E-20	down	3.0%	0	1.7%
ADCY7	Adenylate Cyclase 7	-1.314	3.64E-02	-1.626	7.20E-23	down	3.0%	0	4.0%
PARP14	Poly(ADP-Ribose) Polymerase Family Member 14	-1.233	4.09E-02	-1.178	2.66E-08	down	0	0	2.5%
ITGAL	Integrin Subunit Alpha L	-1.225	3.83E-02	-1.893	6.38E-17	down	3.0%	0	5.0%
ZNF704	Zinc Finger Protein 704	Inf	2.00E-02	1.059	8.34E-13	up	0	0	0.8%
NOVA1	NOVA Alternative Splicing Regulator 1	Inf	3.63E-02	1.039	2.61E-16	up	0	1.3%	1.7%
ATRNL1	Attractin Like 1	Inf	3.34E-02	0.878	6.12E-09	up	7.0%	4.0%	3.0%
No genomic alterations (n = 8)									
FOSB	FosB Proto-Oncogene, AP-1 Transcription Factor Subunit	-3.723	4.45E-02	-3.385	2.01E-15	down	0	0	0
ADAM15	ADAM Metallopeptidase Domain 15	-3.479	1.29E-02	-0.586	4.13E-08	down	0	0	0
KLF6	Kruppel Like Factor 6	-1.999	4.21E-03	-1.665	3.59E-18	down	0	0	0
IL10RA	Interleukin 10 Receptor Subunit Alpha	-1.921	2.54E-03	-1.772	8.82E-14	down	0	0	0
ATG16L2	Autophagy Related 16 Like 2	-1.851	0.01755	-0.665	1.89E-12	down	0	0	0

(Continued)

TABLE 2 | Continued

Gene symbol	Gene full name	Differentially expressed levels					Genomic alterations		
		In-house SCLC plasma cohort		SCLC lung tissue cohort (GSE40275)		Regulated	CLCGP, <i>Nat Genet</i> 2012	Johns Hopkins, <i>Nat Genet</i> 2012	U Cologne, <i>Nature</i> 2015
		log2FC	<i>p</i> value	log2FC	<i>p</i> value				
MYO15B	Myosin XVB	-1.814	4.40E-02	-0.837	1.26E-12	down	0	0	0
IRF1	Interferon Regulatory Factor 1	-1.589	8.63E-03	-1.775	4.68E-11	down	0	0	0
GREM1	Gremlin 1, DAN Family BMP Antagonist	Inf	4.60E-02	1.011	1.16E-08	up	0	0	0

SCLC, small cell lung cancer; circRNA, circular RNA; lncRNA, long noncoding RNA; miRNA, microRNA; mRNA, messenger RNA; ceRNA, competing endogenous RNA; FC, fold change; Inf, infinity; CLCGP, Clinical Lung Cancer Genome Project; U Cologne, University of Cologne study.

TABLE 3 | Functional enrichment analysis of mRNAs in the ceRNA network in SCLC.

ID	Description	Ontology	Bg Ratio	p value	Adjusted p	Genes symbol*	Count
GO:0043312	neutrophil degranulation	BP	485/18670	1.843E-06	9.114E-04	CFD/FCN1/FGR/GAA/ITGAL/ITGAX/ITGB2/TCIRG1/CD93/NFAM1	10
GO:0002283	neutrophil activation involved in immune response	BP	488/18670	1.948E-06	9.114E-04	CFD/FCN1/FGR/GAA/ITGAL/ITGAX/ITGB2/TCIRG1/CD93/NFAM1	10
GO:0042119	neutrophil activation	BP	498/18670	2.336E-06	9.114E-04	CFD/FCN1/FGR/GAA/ITGAL/ITGAX/ITGB2/TCIRG1/CD93/NFAM1	10
GO:0002446	neutrophil-mediated immunity	BP	499/18670	2.378E-06	9.114E-04	CFD/FCN1/FGR/GAA/ITGAL/ITGAX/ITGB2/TCIRG1/CD93/NFAM1	10
GO:0007229	integrin-mediated signalling pathway	BP	103/18670	1.545E-05	4.738E-03	FGR/ITGAL/ITGAX/ITGB2/ADAM15	5
GO:0050663	cytokine secretion	BP	240/18670	8.892E-05	2.272E-02	FCN1/FGR/NOTCH1/TCIRG1/CD244/NLRP12	6
GO:0030198	extracellular matrix organization	BP	368/18670	1.237E-04	2.467E-02	ITGAL/ITGAX/ITGB2/NOTCH1/ADAM15/GREM1/ADAMTSL4	7
GO:0050900	leukocyte migration	BP	499/18670	1.287E-04	2.467E-02	CSF3R/ITGAL/ITGAX/ITGB2/GREM1/CD244/MYO1G/NLRP12	8
GO:0043062	extracellular structure organization	BP	422/18670	2.861E-04	4.873E-02	ITGAL/ITGAX/ITGB2/NOTCH1/ADAM15/GREM1/ADAMTSL4	7
GO:0101003	ficolin-1-rich granule membrane	CC	61/19717	8.873E-07	6.558E-05	GAA/ITGAX/ITGB2/TCIRG1/CD93	5
GO:0101002	ficolin-1-rich granule	CC	185/19717	1.017E-06	6.558E-05	CFD/FCN1/GAA/ITGAX/ITGB2/TCIRG1/CD93	7
GO:0030667	secretory granule membrane	CC	298/19717	2.154E-06	9.263E-05	APLP2/GAA/ITGAL/ITGAX/ITGB2/TCIRG1/CD93/NFAM1	8
GO:0070821	tertiary granule membrane	CC	73/19717	5.888E-05	1.899E-03	GAA/ITGAX/ITGB2/CD93	4
GO:0008305	integrin complex	CC	31/19717	9.721E-05	2.370E-03	ITGAL/ITGAX/ITGB2	3
GO:0070820	tertiary granule	CC	164/19717	1.106E-04	2.370E-03	GAA/ITGAX/ITGB2/TCIRG1/CD93	5
GO:0098636	protein complex involved in cell adhesion	CC	34/19717	1.286E-04	2.370E-03	ITGAL/ITGAX/ITGB2	3
GO:0005774	vascular membrane	CC	412/19717	1.184E-03	1.910E-02	ABCA2/GAA/TCIRG1/AHNAK/ATG16L2/NFAM1	6
GO:0031256	leading edge membrane	CC	170/19717	1.476E-03	1.988E-02	FGR/PSD4/MYO1G/FGD2	4
GO:0001726	Ruffle	CC	172/19717	1.541E-03	1.988E-02	FGR/MEFV/PSD4/FGD2	4
GO:0035579	specific granule membrane	CC	91/19717	2.324E-03	2.725E-02	ITGAL/ITGB2/CD93	3

(Continued)

TABLE 3 | Continued

ID	Description	Ontology	Bg Ratio	p value	Adjusted p	Genes symbol*	Count
GO:0032587	ruffle membrane	CC	94/19717	2.549E-03	2.740E-02	FGR/PSD4/FGD2	3
GO:0005765	lysosomal membrane	CC	354/19717	3.536E-03	3.297E-02	ABCA2/GAA/TCIRG1/AHNAK/NFAM1	5
GO:0098852	lytic vacuole membrane	CC	355/19717	3.578E-03	3.297E-02	ABCA2/GAA/TCIRG1/AHNAK/NFAM1	5

GO, gene ontology; BP, biological process; CC, cellular component; KEGG, Kyoto Encyclopaedia of Genes and Genomes; ceRNA, competing endogenous RNA; circRNA, circular RNAs; lncRNA, long noncoding RNA; miRNA, microRNA; mRNA, messenger RNA; SCLC, small cell lung cancer; Bg, background. *The full name of gene symbols is available in **Table 2**.

DNA recombination and the MyD88-independent toll-like receptor signalling pathway, as well as the regulation of dendritic cell differentiation. In the KEGG pathways analyses, five pathways were identified in the lncRNAs, consisting of olfactory transduction, the neuroactive ligand–receptor interaction, nicotine addiction, carbohydrate digestion and absorption, and the protein digestion and absorption pathway (**Table 4**). The 60 pathways found in the miRNAs and mainly tumour-related pathways were significantly enriched, including the cAMP signalling pathway, focal adhesion, the MAPK signalling pathway, the Hippo signalling pathway and the ECM–receptor interaction (**Table 7**).

Evaluation of Genomic Alterations, Drug Candidates/Repurposing and Pathways in SCLC ceRNA Network

In total, 50 of 58 mRNAs in the ceRNA network presented genomic alterations, with the percentage of genomic alterations ranging from 0.8% to 28% (**Table 2**). The drug–target gene pharmacogenomics analysis showed that three [colony-stimulating factor 3 receptor (CSF3R) (alterations range 1.3–7.0%, FC (in plasma cohort): -1.801 , $p = 2.63 \times 10^{-3}$), acid alpha-glucosidase (GAA) (alterations range 1.3–3.0%, FC: -1.789 and $p = 3.85 \times 10^{-2}$), FGR proto-oncogene Src family tyrosine kinase (FGR) (alterations range 0–2.5%, FC: -1.488 , $p = 2.02 \times 10^{-2}$)] of 50 mRNAs in the ceRNA network were identified as potential drug targets (**Tables 2 and 8**). CSF3R and GAA were identified as targets of FavId and Trastuzumab deruxtecan, respectively, while FGR was confirmed as a target of Dasatinib and Zanubrutinib (**Table 8**). Next, the pathway analysis found that CSF3R, GAA and FGR were annotated in the 13 pathways in the Genecards database (**Table 9**). The SCLC-associated pathways were further identified through a literature review (56–58). We concluded that CSF3R was involved in the autophagy pathway and GAA was involved in the glucose metabolism pathway, while these two pathways were involved in SCLC occurrence and progression from the literature (**Table 9**) (56–58).

Identification of Multi-Omics Integration-Based Prioritisation of the ceRNA SCLC Network

The multi-omics integration-based prioritisation of the ceRNA regulatory network in SCLC consisted of two mRNAs, two miRNAs, three lncRNAs and two circRNAs (**Figure 5**). In this ceRNA network, the expression levels of mRNAs (CSF3R/GAA),

lncRNAs (AC005005.4-201/DLX6-AS1-201/NEAT1-203) and circRNAs (hsa_HLA-B_1/hsa_VEGFC_8) decreased in SCLC, while the expression levels of miRNAs (hsa-miR-4525/hsa-miR-6747-3p) increased in SCLC. The primary regulatory axes in the ceRNA network were identified as follows: 1) lncRNA-miRNA-mRNA: AC005005.4-201/NEAT1-203-hsa-miR-6747-3p-CSF3R and DLX6-AS1-201-hsa-miR-4525-GAA; and 2) circRNA-miRNA-mRNA: hsa_HLA-B_1/hsa_VEGFC_8-hsa-miR-6747-3p-CSF3R and hsa_HLA-B_1-hsa-miR-4525-GAA (**Figure 5**). Thus, lncRNAs (lncRNA-AC005005.4-201 and NEAT1-203) and circRNAs (circRNA-hsa_HLA-B_1 and hsa_VEGFC_8) may regulate the inhibited effects of hsa-miR-6747-3p for CSF3R expression in SCLC, and lncRNA-DLX6-AS1-201 or circRNA-hsa_HLA-B_1 may neutralise the negative regulation of hsa-miR-4525 for GAA in SCLC.

DISCUSSION

Here, we integrated our own omics data (transcriptome and regulome) and public omics data (genome and pharmacogenome) to elucidate the multi-omics integration-based prioritisation of ceRNA-mediated network characteristics, pathways and drug candidates in SCLC. The prioritisation of the SCLC ceRNA regulatory network consisted of two mRNAs (CSF3R/GAA), two miRNAs (hsa-miR-4525/hsa-miR-6747-3p), three lncRNAs (AC005005.4-201/DLX6-AS1-201/NEAT1-203) and two circRNAs (hsa_HLA-B_1/hsa_VEGFC_8). The expression levels of mRNAs, lncRNAs and circRNAs decreased in SCLC, while the expression levels of miRNAs increased in SCLC. In addition, lncRNAs (lncRNA-AC005005.4-201 and NEAT1-203) and circRNAs (circRNA-hsa_HLA-B_1 and hsa_VEGFC_8) may regulate the inhibited effects of hsa-miR-6747-3p for CSF3R expression in SCLC, and lncRNA-DLX6-AS1-201 or circRNA-hsa_HLA-B_1 may neutralise the negative regulation of hsa-miR-4525 related to GAA in SCLC. The pharmacogenomics analysis identified CSF3R and GAA as targets of FavId and Trastuzumab deruxtecan, respectively. The SCLC-associated pathway analysis revealed that CSF3R was involved in the autophagy pathway, while GAA was involved in the glucose metabolism pathway. These findings may contribute to understanding the molecular pathogenesis of SCLC, supporting the development of novel diagnostics and therapeutic compounds for SCLC patients in clinical settings.

TABLE 4 | Functional enrichment analysis and pathway results of lncRNAs in the ceRNA network.

ID	Description	Ontology	Bg Ratio	p value	Adjusted p
GO:0050911	Detection of chemical stimulus involved in sensory perception of smell	BP	0.0252	2.5259E-19	1.5494E-15
GO:0032199	Reverse transcription involved in RNA-mediated transposition	BP	0.0486	2.0772E-15	6.3707E-12
GO:0090305	Nucleic acid phosphodiester bond hydrolysis	BP	0.058	2.5433E-14	5.2002E-11
GO:0007186	G-protein coupled receptor signalling pathway	BP	0.0406	7.6252E-13	1.1693E-09
GO:0097252	Oligodendrocyte apoptotic process	BP	0.039	1.6472E-11	2.0208E-08
GO:0006289	Nucleotide-excision repair	BP	0.0402	4.2385E-11	4.3332E-08
GO:0090200	Positive regulation of release of cytochrome c from mitochondria	BP	0.0399	8.0612E-11	6.7949E-08
GO:0000733	DNA strand renaturation	BP	0.0395	8.8620E-11	6.7949E-08
GO:0007569	Cell aging	BP	0.0397	1.1273E-10	7.6831E-08
GO:0030308	Negative regulation of cell growth	BP	0.0447	1.8945E-10	1.1621E-07
GO:0007275	Multicellular organism development	BP	0.0664	2.2891E-08	1.2765E-05
GO:0006310	DNA recombination	BP	0.0325	3.6143E-08	1.8475E-05
GO:0006278	RNA-dependent DNA biosynthetic process	BP	0.0088	4.2840E-08	2.0214E-05
GO:0032197	Transposition, RNA-mediated	BP	0.0081	6.9777E-06	3.0572E-03
GO:0009987	Cellular process	BP	0.003	1.2068E-05	4.9352E-03
GO:0006259	DNA metabolic process	BP	0.0054	1.4735E-05	5.6492E-03
GO:0007156	Homophilic cell adhesion via plasma membrane adhesion molecules	BP	0.0098	7.3718E-05	2.6599E-02
GO:0016043	Cellular component organisation	BP	0.0064	7.8691E-05	2.6816E-02
GO:0044238	Primary metabolic process	BP	0.0027	1.3610E-04	4.3939E-02
GO:0048741	Skeletal muscle fibre development	BP	0.0141	1.6591E-04	4.8714E-02
GO:0003338	Metanephros morphogenesis	BP	0.001	1.7472E-04	4.8714E-02
GO:0070307	Lens fibre cell development	BP	0.001	1.7472E-04	4.8714E-02
GO:0044424	Intracellular part	CC	0.007	1.7884E-07	1.6189E-04
GO:0043229	Intracellular organelle	CC	0.0019	1.2468E-06	4.2980E-04
GO:0005886	Plasma membrane	CC	0.1378	1.4243E-06	4.2980E-04
GO:0044446	Intracellular organelle part	CC	0.0032	2.8392E-06	6.4257E-04
GO:0098588	Bounding membrane of organelle	CC	0.0086	5.0429E-05	9.1302E-03
GO:0044456	Synapse part	CC	0.0013	1.3285E-04	1.9921E-02
GO:0005739	Mitochondrion	CC	0.0821	1.6615E-04	1.9921E-02
GO:0005796	Golgi lumen	CC	0.0066	1.7604E-04	1.9921E-02
GO:0005578	Proteinaceous extracellular matrix	CC	0.0098	2.3813E-04	2.3264E-02
GO:0016021	Integral component of membrane	CC	0.2479	2.5699E-04	2.3264E-02
GO:0097546	Ciliary base	CC	0.0041	5.4612E-04	4.4944E-02
GO:0005887	Integral component of plasma membrane	CC	0.066	6.2753E-04	4.7340E-02
GO:0003964	RNA-directed DNA polymerase activity	MF	0.0534	7.1861E-20	1.0143E-16
GO:0004984	Olfactory receptor activity	MF	0.0249	1.0694E-19	1.0143E-16
GO:0004930	G-protein coupled receptor activity	MF	0.0316	4.2156E-17	2.6656E-14
GO:0009036	Type II site-specific deoxyribonuclease activity	MF	0.0479	1.2775E-16	6.0586E-14
GO:0005507	Copper ion binding	MF	0.0408	1.0171E-10	3.8588E-08
GO:0005488	Binding	MF	0.0105	1.9980E-10	6.3171E-08
GO:0043167	Ion binding	MF	0.0116	3.3737E-07	9.1428E-05
GO:0005549	Odorant binding	MF	0.0056	1.5499E-06	3.6752E-04
hsa04740	Olfactory transduction	KEGG	0.0598	8.0910E-40	2.2399E-37
hsa04080	Neuroactive ligand-receptor interaction	KEGG	0.0385	5.7920E-08	8.0173E-06
hsa05033	Nicotine addiction	KEGG	0.0054	2.1224E-04	1.9585E-02
hsa04973	Carbohydrate digestion and absorption	KEGG	0.0076	5.8461E-04	3.3090E-02
hsa04974	Protein digestion and absorption	KEGG	0.012	5.9763E-04	3.3090E-02

GO, gene ontology; BP, biological process; CC, cellular component; KEGG, Kyoto Encyclopaedia of Genes and Genomes; ceRNA, competing endogenous RNA; circRNA, circular RNAs; lncRNA, long noncoding RNA; miRNA, microRNA; mRNA, messenger RNA; Bg, background.

TABLE 5 | Functional enrichment analysis of circRNAs in the ceRNA network.

ID	Description	Ontology	Bg Ratio	p value	Adjusted p
GO:0032655	regulation of interleukin-12 production	BP	0.0001	2.963E-04	3.119E-04
GO:0032675	regulation of interleukin-6 production	BP	0.0001	2.963E-04	3.119E-04
GO:2001198	regulation of dendritic cell differentiation	BP	0.0001	2.963E-04	3.119E-04
GO:0002667	regulation of T cell anergy	BP	0.0002	8.888E-04	7.017E-04

(Continued)

TABLE 5 | Continued

ID	Description	Ontology	Bg Ratio	p value	Adjusted p
GO:0002486	antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway, TAP-independent	BP	0.0004	1.481E-03	8.311E-04
GO:0015031	protein transport	BP	0.0175	1.784E-03	8.311E-04
GO:0001916	positive regulation of T cell-mediated cytotoxicity	BP	0.0005	2.073E-03	8.311E-04
GO:0016045	detection of bacterium	BP	0.0006	2.369E-03	8.311E-04
GO:0042270	protection from natural killer cell-mediated cytotoxicity	BP	0.0006	2.369E-03	8.311E-04
GO:0002480	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	BP	0.0007	2.664E-03	8.414E-04
GO:0030100	regulation of endocytosis	BP	0.0010	3.847E-03	1.104E-03
GO:0006904	vesicle docking involved in exocytosis	BP	0.0011	4.438E-03	1.168E-03
GO:0060337	type I interferon signalling pathway	BP	0.0022	8.861E-03	2.152E-03
GO:0002479	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	BP	0.0030	1.180E-02	2.546E-03
GO:0060333	interferon gamma-mediated signalling pathway	BP	0.0030	1.210E-02	2.546E-03
GO:0051726	regulation of cell cycle	BP	0.0074	2.931E-02	5.784E-03
GO:0006367	transcription initiation from RNA polymerase II promoter	BP	0.0108	4.257E-02	7.908E-03
GO:0006468	protein phosphorylation	BP	0.0122	4.801E-02	8.423E-03
GO:0031901	early endosome membrane	CC	0.0062	1.137E-04	5.983E-04
GO:0042612	MHC class I protein complex	CC	0.0010	2.892E-03	6.082E-03
GO:0016592	mediator complex	CC	0.0017	4.954E-03	6.082E-03
GO:0071556	integral component of the luminal side of endoplasmic reticulum membrane	CC	0.0017	4.954E-03	6.082E-03
GO:0012507	ER to Golgi transport vesicle membrane	CC	0.0019	5.778E-03	6.082E-03
GO:0030670	phagocytic vesicle membrane	CC	0.0028	8.454E-03	7.415E-03
GO:0046977	TAP binding	MF	0.0004	1.093E-03	3.107E-03
GO:0008353	RNA polymerase II carboxy-terminal domain kinase activity	MF	0.0007	1.967E-03	3.107E-03
GO:0004693	cyclin-dependent protein serine/threonine kinase activity	MF	0.0020	6.113E-03	5.857E-03
GO:0042605	peptide antigen binding	MF	0.0025	7.419E-03	5.857E-03
GO:0051087	chaperone binding	MF	0.0039	1.155E-02	6.307E-03
GO:0008565	protein transporter activity	MF	0.0040	1.198E-02	6.307E-03
GO:0008289	lipid binding	MF	0.0061	1.826E-02	8.239E-03
GO:0005102	receptor binding	MF	0.0102	3.031E-02	1.197E-02

GO, gene ontology; BP, biological process; CC, cellular component; ceRNA, competing endogenous RNA; circRNA, circular RNAs; Bg, background.

TABLE 6 | Functional enrichment analysis of miRNAs in the ceRNA network.

ID	Description	Ontology	Bg Ratio	p value	Adjusted p
GO:0006355	regulation of transcription, DNA-templated	BP	0.0921	8.5782E-11	5.7934E-07
GO:000122	negative regulation of transcription from RNA polymerase II promoter	BP	0.0565	1.5250E-08	5.1498E-05
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	BP	0.0664	2.6517E-08	5.9696E-05
GO:0060348	bone development	BP	0.0174	6.6166E-08	1.0284E-04
GO:0017144	drug metabolic process	BP	0.0187	8.6063E-08	1.0284E-04
GO:0017187	peptidyl-glutamic acid carboxylation	BP	0.0179	9.1359E-08	1.0284E-04
GO:0042373	vitamin K metabolic process	BP	0.0177	2.6291E-07	2.5366E-04
GO:0007250	activation of NF-kappa-inducing kinase activity	BP	0.0124	3.1140E-07	2.5645E-04
GO:0007156	hemophilic cell adhesion via plasma membrane adhesion molecules	BP	0.0098	3.4174E-07	2.5645E-04
GO:0032743	positive regulation of interleukin 2 production	BP	0.0125	6.9433E-07	4.6893E-04
GO:2000679	positive regulation of transcription regulatory region DNA binding	BP	0.017	1.4950E-06	9.1787E-04
GO:0031293	membrane protein intracellular domain proteolysis	BP	0.0124	2.1866E-06	1.2306E-03
GO:0002756	MyD88-independent toll-like receptor signalling pathway	BP	0.0116	3.2308E-06	1.6785E-03
GO:0000187	activation of MAPK activity	BP	0.0213	5.9088E-06	2.8504E-03
GO:0002726	positive regulation of T cell cytokine production	BP	0.0124	9.3027E-06	4.1885E-03
GO:0070555	response to interleukin 1	BP	0.0098	1.0189E-05	4.2601E-03
GO:0051865	protein auto-ubiquitination	BP	0.0143	1.0723E-05	4.2601E-03
GO:0045672	positive regulation of osteoclast differentiation	BP	0.0125	1.3074E-05	4.7509E-03
GO:0001932	regulation of protein phosphorylation	BP	0.003	1.4009E-05	4.7509E-03
GO:0070534	protein K63-linked ubiquitination	BP	0.0182	1.4069E-05	4.7509E-03
GO:0031398	positive regulation of protein ubiquitination	BP	0.0121	2.6984E-05	8.2836E-03
GO:0034162	toll-like receptor 9 signalling pathway	BP	0.0121	2.6984E-05	8.2836E-03
GO:0070423	nucleotide-binding oligomerisation domain containing signalling pathway	BP	0.0175	2.8472E-05	8.3605E-03
GO:0043507	positive regulation of JUN kinase activity	BP	0.0139	4.0463E-05	1.1386E-02
GO:0030574	collagen catabolic process	BP	0.0023	4.2766E-05	1.1553E-02
GO:0071222	cellular response to lipopolysaccharide	BP	0.0118	5.4294E-05	1.4070E-02
GO:0002755	MyD88-dependent toll-like receptor signalling pathway	BP	0.0181	5.6249E-05	1.4070E-02
GO:0046513	ceramide biosynthetic process	BP	0.0096	6.7342E-05	1.6134E-02
GO:0035019	somatic stem cell population maintenance	BP	0.0075	6.9279E-05	1.6134E-02
GO:0001707	mesoderm formation	BP	0.0013	8.3053E-05	1.8697E-02
GO:0007596	blood coagulation	BP	0.0236	9.1112E-05	1.9850E-02

(Continued)

TABLE 6 | Continued

ID	Description	Ontology	Bg Ratio	p value	Adjusted p
GO:0050870	positive regulation of T cell activation	BP	0.0067	1.5077E-04	3.1820E-02
GO:0007155	cell adhesion	BP	0.0111	1.5634E-04	3.1997E-02
GO:0015886	heme transport	BP	0.0035	1.6785E-04	3.2879E-02
GO:0043065	positive regulation of apoptotic process	BP	0.028	1.7039E-04	3.2879E-02
GO:0045059	positive thymic T cell selection	BP	0.0023	1.9247E-04	3.6108E-02
GO:0035023	regulation of Rho protein signal transduction	BP	0.0039	2.1384E-04	3.9032E-02
GO:0051092	positive regulation of NF-kappa B transcription factor activity	BP	0.026	2.2701E-04	4.0346E-02
GO:0031410	cytoplasmic vesicle	CC	0.014	8.1251E-07	6.2029E-04
GO:0005789	endoplasmic reticulum membrane	CC	0.0602	1.2645E-06	6.2029E-04
GO:0010008	endosome membrane	CC	0.0227	1.8113E-06	6.2029E-04
GO:0005829	cytosol	CC	0.1935	7.5017E-06	1.9267E-03
GO:0034704	calcium channel complex	CC	0.0008	5.8805E-05	1.2083E-02
GO:0005811	lipid droplet	CC	0.012	7.5151E-05	1.2868E-02
GO:0035631	CD40 receptor complex	CC	0.0098	1.0108E-04	1.3848E-02
GO:0009898	cytoplasmic side of plasma membrane	CC	0.0116	1.0783E-04	1.3848E-02
GO:0005667	transcription factor complex	CC	0.0095	3.8115E-04	4.3509E-02
GO:0003700	transcription factor activity, sequence-specific DNA binding	MF	0.0684	1.2784E-14	2.7816E-11
GO:0000977	RNA polymerase II regulatory region sequence-specific DNA binding	MF	0.0261	6.2083E-13	6.7539E-10
GO:0046872	metal ion binding	MF	0.1355	1.8538E-08	1.3445E-05
GO:0031996	thioesterase binding	MF	0.0128	1.6534E-07	7.5276E-05
GO:0031624	ubiquitin conjugating enzyme binding	MF	0.0136	1.7299E-07	7.5276E-05
GO:0042826	histone deacetylase binding	MF	0.0191	3.5880E-07	1.3011E-04
GO:0047057	vitamin-K-epoxide reductase (warfarin-sensitive) activity	MF	0.0174	4.3664E-07	1.3572E-04
GO:0043422	protein kinase B binding	MF	0.0122	5.3369E-07	1.4515E-04
GO:0031435	mitogen-activated protein kinase binding	MF	0.0125	2.4614E-06	5.9505E-04
GO:0005164	tumour necrosis factor receptor binding	MF	0.0133	4.8342E-06	1.0518E-03
GO:0050291	sphingosine N-acyltransferase activity	MF	0.0083	2.3602E-05	4.6685E-03
GO:0003682	chromatin binding	MF	0.0168	3.5302E-05	6.4008E-03
GO:0001077	transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	MF	0.0219	4.3610E-05	7.2989E-03
GO:0005096	GTPase activator activity	MF	0.0123	1.0248E-04	1.5927E-02
GO:0031625	ubiquitin protein ligase binding	MF	0.0317	1.2898E-04	1.8708E-02

(Continued)

TABLE 6 | Continued

ID	Description	Ontology	Bg Ratio	p value	Adjusted p
GO:0001078	transcriptional repressor activity, RNA polymerase II core promoter proximal region sequence-specific binding	MF	0.009	1.5448E-04	2.1007E-02
GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding	MF	0.0242	1.9703E-04	2.5217E-02
GO:0008270	zinc ion binding	MF	0.0636	2.6201E-04	3.1672E-02
GO:0001047	core promoter binding	MF	0.0109	4.3480E-04	4.9792E-02

GO, gene ontology; BP, biological process; CC, cellular component; ceRNA, competing endogenous RNA; miRNA, microRNA; Bg, background.

TABLE 7 | Pathway results of miRNAs in the ceRNA network.

ID	Description	Bg Ratio	p value	Adjusted p
hsa04921	Oxytocin signalling pathway	0.0212	1.3332E-08	2.9190E-06
hsa04261	Adrenergic signalling in cardiomyocytes	0.0208	6.2595E-07	6.5092E-05
hsa04024	cAMP signalling pathway	0.0273	8.9189E-07	6.5092E-05
hsa04510	Focal adhesion	0.0298	1.5422E-05	7.4010E-04
hsa04750	Inflammatory mediator regulation of TRP channels	0.0137	1.6901E-05	7.4010E-04
hsa04713	Circadian entrainment	0.0125	2.1580E-05	7.8746E-04
hsa04360	Axon guidance	0.0239	2.5931E-05	8.1108E-04
hsa04015	Rap1 signalling pathway	0.03	4.7767E-05	1.3073E-03
hsa05200	Pathways in cancer	0.055	6.2554E-05	1.5218E-03
hsa04611	Platelet activation	0.0165	7.0689E-05	1.5477E-03
hsa04010	MAPK signalling pathway	0.0381	1.1422E-04	2.2735E-03
hsa04724	Glutamatergic synapse	0.0149	1.3881E-04	2.3601E-03
hsa04725	Cholinergic synapse	0.0151	1.4013E-04	2.3601E-03
hsa05206	MicroRNAs in cancer	0.0193	2.6751E-04	4.1690E-03
hsa04728	Dopaminergic synapse	0.0165	2.8562E-04	4.1690E-03
hsa04925	Aldosterone synthesis and secretion	0.0108	3.4051E-04	4.6596E-03
hsa01522	Endocrine resistance	0.0132	3.9237E-04	4.6724E-03
hsa04722	Neurotrophin signalling pathway	0.0168	3.9400E-04	4.6724E-03
hsa04720	Long-term potentiation	0.0089	4.0547E-04	4.6724E-03
hsa04390	Hippo signalling pathway	0.0209	4.5090E-04	4.9362E-03
hsa04512	ECM–receptor interaction	0.0112	5.2201E-04	5.4425E-03
hsa04512	Wnt signalling pathway	0.0195	6.5195E-04	6.4883E-03
hsa04915	Oestrogen signalling pathway	0.0137	7.9656E-04	7.5828E-03

(Continued)

TABLE 7 | Continued

ID	Description	Bg Ratio	p value	Adjusted p
hsa04924	Renin secretion	0.0086	9.2945E-04	8.4792E-03
hsa04022	cGMP–PKG signalling pathway	0.0247	1.2704E-03	1.1126E-02
hsa04923	Regulation of lipolysis in adipocytes	0.0082	1.3531E-03	1.1192E-02
hsa05210	Colorectal cancer	0.009	1.4528E-03	1.1192E-02
hsa04014	Ras signalling pathway	0.0325	1.4684E-03	1.1192E-02
hsa04912	GnRH signalling pathway	0.0124	1.5787E-03	1.1192E-02
hsa04727	GABAergic synapse	0.0114	1.5828E-03	1.1192E-02
hsa04911	Insulin secretion	0.0116	1.5846E-03	1.1192E-02
hsa00512	Mucin type O-Glycan biosynthesis	0.0039	2.0450E-03	1.3992E-02
hsa04910	Insulin signalling pathway	0.0212	2.2097E-03	1.4661E-02
hsa00514	Other types of O-glycan biosynthesis	0.0042	2.3080E-03	1.4862E-02
hsa04012	ErbB signalling pathway	0.0121	3.0528E-03	1.8829E-02
hsa04270	Vascular smooth muscle contraction	0.017	3.0960E-03	1.8829E-02
hsa01212	Fatty acid metabolism	0.0068	4.0784E-03	2.3933E-02
hsa04020	Calcium signalling pathway	0.0302	4.2385E-03	2.3933E-02
hsa04930	Type II diabetes mellitus	0.0081	4.4096E-03	2.3933E-02
hsa04931	Insulin resistance	0.0158	4.4262E-03	2.3933E-02
hsa04971	Gastric acid secretion	0.0097	4.4817E-03	2.3933E-02
hsa04152	AMPK signalling pathway	0.018	4.7769E-03	2.4902E-02
hsa04211	Longevity regulating pathway	0.0135	5.2272E-03	2.6447E-02
hsa04916	Melanogenesis	0.0132	5.3149E-03	2.6447E-02
hsa04340	Hedgehog signalling pathway	0.0069	6.1564E-03	2.9698E-02
hsa04213	Longevity regulating pathway – multiple species	0.009	6.3540E-03	2.9698E-02
hsa05221	Acute myeloid leukaemia	0.0082	6.3751E-03	2.9698E-02
hsa04550	Signalling pathways regulating pluripotency of stem cells	0.0196	6.6773E-03	3.0458E-02
hsa05410	Hypertrophic cardiomyopathy (HCM)	0.0115	7.8953E-03	3.5279E-02
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.0097	8.6718E-03	3.7274E-02
hsa04962	Vasopressin-regulated water reabsorption	0.006	8.6824E-03	3.7274E-02
hsa04144	Endocytosis	0.0376	9.4078E-03	3.9612E-02
hsa04068	FoxO signalling pathway	0.0201	9.9201E-03	4.0816E-02
hsa04350	TGF-beta signalling pathway	0.0119	1.0253E-02	4.0816E-02

(Continued)

TABLE 7 | Continued

ID	Description	Bg Ratio	p value	Adjusted p
hsa05222	Small cell lung cancer	0.0119	1.0253E-02	4.0816E-02
hsa01521	EGFR tyrosine kinase inhibitor resistance	0.0116	1.0447E-02	4.0847E-02
hsa00531	Glycosaminoglycan degradation	0.0026	1.1704E-02	4.4958E-02
hsa04723	Retrograde endocannabinoid signalling	0.0133	1.1967E-02	4.5173E-02
hsa04142	Lysosome	0.0175	1.2975E-02	4.8149E-02

KEGG, *Kyoto Encyclopaedia of Genes and Genomes*; ceRNA, *competing endogenous RNA*; miRNA, *microRNA*; Bg, *background*.

TABLE 8 | Potential drug candidates of mRNAs in the ceRNA networks in SCLC.

mRNAs	Drug candidate	Type*	Therapy*	Main roles*	Data resource
Colony-stimulating factor 3 receptor (CSF3R)	Favld	an active immunotherapy	Tumour therapy	based upon unique genetic information extracted from a patient's tumour	https://go.drugbank.com/drugs/DB05249
	Pegfilgrastim	a recombinant human granulocyte colony stimulating factor	Adjuvant therapy	stimulate the production of neutrophils and prevent febrile neutropenia or infections after myelosuppressive chemotherapy	https://go.drugbank.com/drugs/DB00019
	Filgrastim	a form of recombinant human granulocyte colony stimulating factor	Adjuvant therapy	induce the production of granulocytes and lower infection risk after myelosuppressive therapy	https://go.drugbank.com/drugs/DB00099
	Lenograstim	a granulocyte colony-stimulating factor	Adjuvant therapy	reduce the duration of neutropenia in bone marrow transplant and cytotoxic chemotherapy, as well as mobilizing hematopoietic stem cells in healthy donors	https://go.drugbank.com/drugs/DB13144
	Lipegfilgrastim	a medication	Adjuvant therapy	reduce the duration of chemotherapy-induced neutropenia and incidence of febrile neutropenia in cytotoxic chemotherapy	https://go.drugbank.com/drugs/DB13200
Acid alpha-glucosidase (GAA)	Trastuzumab deruxtecan	an antibody	Tumour therapy	treat certain types of unresectable or metastatic HER-2 positive breast cancer	https://go.drugbank.com/drugs/DB14962
	Acarbose	an alpha-glucosidase inhibitor	Other therapy	adjunctly with diet and exercise for the management of glycaemic control in patients with type 2 diabetes mellitus.	https://go.drugbank.com/drugs/DB00284
	AT2220	pharmacological chaperones	Other therapy	increase GAA activity in cell lines derived from Pompe patients and in transfected cells expressing misfolded forms of GAA	https://go.drugbank.com/drugs/DB05200
	Miglitol	an oral alpha-glucosidase inhibitor	Other therapy	improve glycaemic control by delaying the digestion of carbohydrates	https://go.drugbank.com/drugs/DB00491
FGR Proto-Oncogene, Src Family Tyrosine Kinase (FGR)	Dasatinib	a tyrosine kinase inhibitor	Tumour therapy	treat lymphoblastic or chronic myeloid leukaemia with resistance or intolerance to prior therapy	https://go.drugbank.com/drugs/DB01254
	Zanubrutinib	a kinase inhibitor	Tumour therapy	treat mantle cell lymphoma, a type of B-cell non-Hodgkin lymphoma, in adults who previously received therapy.	https://go.drugbank.com/drugs/DB015035
	Fostamatinib	a spleen tyrosine kinase inhibitor	Other therapy	treat chronic immune thrombocytopenia after attempting one other treatment.	https://go.drugbank.com/drugs/DB12010

ceRNA, *competing endogenous RNA*; SCLC, *small cell lung cancer*; HER-2, *human epidermal growth factor receptor-2*; *, *the information is from Drugbank (<https://go.drugbank.com/>)*.

In this study, we first reported the multi-omics integration-based prioritisation of the lncRNA/circRNA-miRNA-mRNA ceRNA disease network, as well as the molecular characteristics and drug candidates or repurposed drugs in SCLC. The ceRNA is a layer of gene regulation in diseases, and the transcripts can regulate each other at the post-transcription level by competing for shared

miRNAs (12, 16, 17). Here, we found that two lncRNAs (lncRNA-AC005005.4-201 and NEAT1-203) and two circRNAs (circRNA-hsa_HLA-B_1 and hsa_VEGFC_8) may regulate the inhibiting effects of hsa-miR-6747-3p for CSF3R expression, while lncRNA-DLX6-AS1-201 or circRNA-hsa_HLA-B_1 may neutralise the negative regulation of hsa-

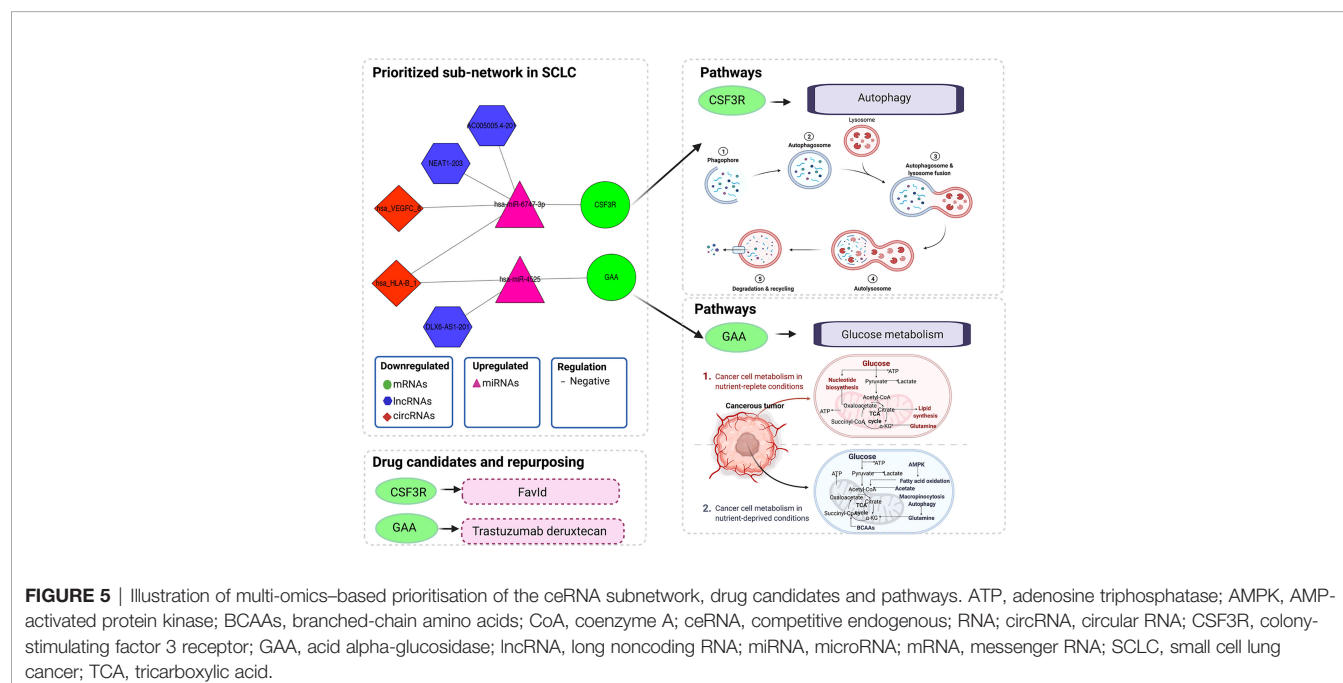
TABLE 9 | Pathways of mRNAs in the ceRNA networks in SCLC.

mRNAs	Gene ontology (GO) based on molecular function	Pathways	Associated to SCLC pathway
Colony-stimulating factor 3 receptor (CSF3R)	Cytokine binding (GO:0019955)	Autophagy pathway	Güçlü E, et al. (56); Liu H, et al. (57)
	Cytokine receptor activity (GO:0004896)	Akt signalling	na
	Protein binding (GO:0005515)	PEDF-induced signalling	na
	Signalling receptor activity (GO:0038023)	Cytokine signalling in the immune system	na
	Granulocyte colony-stimulating factor binding (GO:0051916)	Hematopoietic cell lineage	na
Acid alpha-glucosidase (GAA)	Catalytic activity (GO:0003824)	Glucose metabolism	Yan X, et al. (58)
	Hydrolase activity, hydrolyzing O-glycosyl compounds (GO:0004553)	Innate immune system	na
	Alpha-1,4-glucosidase activity (GO:0004558)	Galactose metabolism	na
	Hydrolase activity (GO:0016787)	Metabolism	na
	Hydrolase activity, acting on glycosyl bonds (GO:0016798)	Lysosome	na
FGR Proto-Oncogene, Src Family Tyrosine Kinase (FGR)	Nucleotide binding (GO:0000166)	Innate immune system	na
	Phosphotyrosine residue binding (GO:0001784)	Platelet homeostasis	na
	Protein kinase activity (GO:0004672)	Tyrosine kinases/adaptors	na
	Protein tyrosine kinase activity (GO:0004713)	CCR5 pathway in macrophages	na
	Transmembrane receptor protein tyrosine kinase activity (GO:0004714)	Integrin pathway	na

ceRNA, competing endogenous RNA; SCLC, small cell lung cancer; Akt, protein kinase B; CCR5, chemokine-CC motif-receptor-5; GO, gene ontology; PEDF, pigment epithelium derived factor; na, not available.

miR-4525 for GAA. Consistent with our findings for dysregulated lncRNAs in SCLC, previous studies found that lncRNAs DLX6-AS1 and NEAT1 were significantly dysregulated in non-SCLC, gastric cancer and pancreatic cancer (59–62). Specifically, upregulated DLX6-AS1 in gastric cancer tissue associated with distant metastasis and a poor clinical prognosis, while siRNA-DLX6-AS1 may inhibit gastric cancer cell proliferation, migration, invasion and the epithelial-mesenchymal transition *in vitro* (18). In addition, our study

identified the regulatory axis in lncRNA-DLX6-AS1-201/hsa-miR-4525/GAA, which associated with the glucose metabolism pathway in SCLC. Interestingly, Qian et al. reported that sh-DLX6-AS1 may modulate glucose metabolism and cell growth *via* miR-4290/3-phosphoinositide-dependent protein kinase 1 in gastric cancer cells (63). Considering the role of DLX6-AS1 in glucose metabolism, we inferred that DLX6-AS1 could affect the occurrence and progression of SCLC *via* glucose metabolism through modulating hsa-miR-4525/GAA in SCLC. Similar to the



other dysregulated lncRNA reports (59–62), Xu et al. found that lncRNA-NEAT1 may promote gastric cancer angiogenesis by enhancing the proliferation, migration and tube formation ability of endothelial cells through the miR-17-5p/transforming growth factor- β receptor 2 (TGF β R2) pathway (61), while lncRNA-NEAT1 may play a vital role in tumorigenesis and the development of SCLC through the hsa-miR-6747-3p/CSF3R axis. Importantly, in addition to lncRNA-DLX6-AS1 and NEAT1, we are the first to report another potential regulatory axis of ceRNA, while the regulatory mechanisms require further exploration through *in vivo* and *in vitro* studies. Our findings, however, suggest that the promising lncRNA/circRNA-miRNA-mRNA ceRNA regulatory characteristics in SCLC may provide new potential mechanisms and therapeutic targets.

To the best of our knowledge, this is also the first study to investigate the roles of CSF3R and GAA in the SCLC ceRNA regulation networks, pathways and drug candidates. CSF3R is a type 1 cytokine receptor, encoding the receptor for granulocyte colony-stimulating factor (G-CSF) and playing a crucial role in granulocyte proliferation and differentiation (64, 65). The altered CSF3R expression or activating heterozygous variants in CSF3R have been identified as risk factors in the development of multiple malignancies, such as colorectal cancer, myeloid malignancies and lymphoid malignancies (65–67). This is particularly the case for mutations in CSF3R commonly present in chronic neutrophilic leukaemia or atypical chronic myeloid leukaemia (68). Given the roles of CSF3R reported in chronic neutrophilic leukaemia or atypical chronic myeloid leukaemia (66, 68), our findings suggest that CSF3R might play a pivotal role in the occurrence and development of SCLC. Furthermore, our results suggest that CSF3R might modulate the autophagy pathway, which associated with SCLC (57, 58). The functions of autophagy in cancer may involve an anticancer or a cancer effect (69). Previous studies suggested that a hypoxia-HIF1A-AS2-autophagy interaction may play a role in drug sensitivity in SCLC, while a high expression of secreted phosphoprotein 1 (SPP1) inhibited autophagy and apoptosis, promoting the development of SCLC (57, 58). In addition, Rupniewska et al. found that SCLC cells may be more sensitive to autophagy inhibitors (70). In our study, CSF3R was identified as the potential drug target of FavId. FavId is an active immunotherapy with stimulating tumour-specific T cells and humoral immunity (71, 72). Alissafi et al. reported that autophagy-deficient therapy exhibited a mediated suppression of antitumour immunity *via* the efficient activation of tumour-specific CD4⁺ T cells (73), which was consistent with the mechanism of FavId in a tumour. Thus, our results suggest that genetic alterations or an altered expression of CSF3R may serve as a risk factor in SCLC development and associate with the autophagy pathway, while FavId could serve as a potential drug therapy through the CSF3R target to treat SCLC, even though additional *in vivo* or *in vitro* studies are needed to clarify these associations in SCLC. GAA, as one of the lysosomal enzymes, was the other key gene in our study. This is the first study to find that GAA might participant in the occurrence and development of SCLC *via* glucose metabolism. Similarly, Hamura et al.

reported that the modulation of GAA could affect cell proliferation and apoptosis and manipulate chemoresistance in pancreatic cancer cells *via* malfunctioned mitochondria (74). The dysregulated metabolism of glucose in mitochondria is known as an adverse microenvironment in solid tumours, referred to as the Warburg effect, including glucose deprivation and lactic acidosis, potentially resulting in an elevated glycolytic activity in tumour cells (75–78). Yan et al. showed that glucose metabolic reprogramming improves SCLC cell proliferation and metastasis, suggesting it could be a potential regulatory strategy interfering with glucose metabolism in SCLC (56). Considering the function of GAA, which catalyses the production of glucose from glycogen in lysosomes, altering the GAA expression or genetic status could inhibit tumorigenesis in SCLC through the lysosome pathway (56, 74–78). Interestingly, the DrugBank analysis showed that the drug targeting GAA was Trastuzumab-deruxtecan. Trastuzumab-deruxtecan is primarily used for patients with human epidermal growth factor receptor 2 (HER2)-mutant tumours including non-SCLC and in the absence of SCLC (79–81). Upon binding to HER2, Trastuzumab-deruxtecan disrupts the HER2 signalling, undergoes internalisation and intracellular linker cleavage by lysosomal enzymes and ultimately causes DNA damage and apoptotic cell death (80). In addition, Martinho et al. found that the inhibitors of the HER family (mainly HER2) reduced cervical cancer aggressiveness by blocking glucose metabolism (82). Combined with the roles of the glucose metabolism pathway in SCLC and the antitumour roles of Trastuzumab-deruxtecan *via* the glucose metabolism pathway, our findings suggest that Trastuzumab-deruxtecan may be a promising drug candidate *via* GAA in SCLC through the glucose metabolism pathway. However, further *in vivo* or *in vitro* studies are needed to clarify these promising drug candidates' ability to treat SCLC.

The strength of this study is our use of network-based multi-omics integration to prioritise ceRNA characteristics and drug candidates in SCLC from two well-characterised study cohorts, including newly tested whole-transcriptome sequencing data in the SCLC study, and the data were uploaded to a public platform [the Sequence Read Archive (SRA) database]. In addition to these strengths, we also note several limitations. First, our study included our own omics data and public data. In addition, the relatively small size of our cohort represents a limitation to our findings, although the results of the mRNA study were validated in a relatively large cohort. Second, the ceRNA characteristics and drug candidates and repurposing are quite promising, although further mechanistic studies from cells and animal models, as well as clinical validation studies, are needed. In addition, we performed no survival analysis in this study, since no available and suitable survival data were obtained from public databases, including the Cancer Genome Atlas (TCGA) and Kaplan–Meier plotter databases. Finally, the survival data in our SCLC plasma cohort were incapable of producing useful results for the prognostic analysis given the relatively small sample sizes and quite limited follow-up time.

In conclusion, we report primary findings related to a multi-omics integration-based prioritisation of the lncRNA/circRNA-

miRNA-mRNA ceRNA regulatory network, pathways and promising drug candidates in SCLC. These findings indicate novel, potential diagnostic and therapeutic targets in SCLC.

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

This study received ethical approval from the Ethics Committee of the Gansu Provincial Hospital, China (27 July 2020, No. 2020-183). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

W-DH, MZ, X-JW and JG contributed to the design of the study. X-JW and W-DH performed the sample collection, analysis and downloaded the data. X-JW and JG contributed to the data analysis and to writing the manuscript. W-DH, MZ, QY, X-JW and JG revised the manuscript. All authors approved the final version of the manuscript.

FUNDING

This study was supported by the Science-Technology Foundation for Young Scientist of the Gansu Province of China (Grant no.18JR3RA059), the Science-Technology Foundation for Scientists of the Gansu Province of China (Grant no.21JR7RA595), the Science-Technology Foundation for Lanzhou City of China (Grant no.2018-4-65) and the Scientists Fund of the Gansu Provincial Hospital of China (Grant no.18GSS4-25). Jing Gao was also supported by the Swedish Heart-Lung Foundation, the Swedish Asthma and

Allergy Foundation, the Sigrid Jusélius Foundation and the Väinö and Laina Kivi Foundation.

ACKNOWLEDGMENTS

We extend our deepest gratitude to all of the patients who volunteered to participate in our study. We thank Jin Li, from the Faculty of Information Technology and Communication Sciences, Tampere University (Finland), for assistance with the tables and figures. We also extend our gratitude to Vanessa L Fuller, from Language Services at the University of Helsinki (Finland), for assistance with the initial English-language revision of this manuscript. In addition, we thank the Biomarker Technologies Corporation (Beijing, China) for sequencing technology and support. Figures were created using the BioRender software (©biorender.com).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.904865/full#supplementary-material>.

Supplementary Table 1 | The lncRNAs, miRNAs and mRNAs in the lncRNA-miRNA-mRNA ceRNA network.

Supplementary Table 2 | The circRNAs, miRNAs and mRNAs in the circRNA-miRNA-mRNA ceRNA network.

Supplementary Table 3 | Differentially expressed levels of 301 lncRNAs in the ceRNA network.

Supplementary Table 4 | Differentially expressed levels of 16 circRNAs in the ceRNA network.

Supplementary Table 5 | Differentially expressed levels of 24 miRNAs in the ceRNA network.

Supplemental File 1 | The whole-transcriptome sequencing process.

Supplemental File 2 | Data cleaning process.

REFERENCES

- Byers LA, Rudin CM. Small Cell Lung Cancer: Where do We Go From Here? *Cancer Am Cancer Soc* (2015) 121(5):664–72. doi: 10.1002/cncr.29098
- Bernhardt EB, Jalal SI. Small Cell Lung Cancer. *Cancer Treat Res* (2016) 170:301–22. doi: 10.1007/978-3-319-40389-2_14
- Wang Y, Zou S, Zhao Z, Liu P, Ke C, Xu S. New Insights Into Small-Cell Lung Cancer Development and Therapy. *Cell Biol Int* (2020) 44(8):1564–76. doi: 10.1002/cbin.11359
- Johal S, Hettle R, Carroll J, Maguire P, Wynne T. Real-World Treatment Patterns and Outcomes in Small-Cell Lung Cancer: A Systematic Literature Review. *J Thorac Dis* (2021) 13(6):3692–707. doi: 10.21037/jtd-20-3034
- Demedts IK, Vermaelen KY, van Meerbeeck JP. Treatment of Extensive-Stage Small Cell Lung Carcinoma: Current Status and Future Prospects. *Eur Respir J* (2010) 35(1):202–15. doi: 10.1183/09031936.00105009
- Stinchcombe TE, Gore EM. Limited-Stage Small Cell Lung Cancer: Current Chemoradiotherapy Treatment Paradigms. *Oncologist* (2010) 15(2):187–95. doi: 10.1634/theoncologist.2009-0298
- Yang S, Zhang Z, Wang Q. Emerging Therapies for Small Cell Lung Cancer. *J Hematol Oncol* (2019) 12(1):47. doi: 10.1186/s13045-019-0736-3
- Koinis F, Kotsakis A, Georgoulas V. Small Cell Lung Cancer (SCLC): No Treatment Advances in Recent Years. *Transl Lung Cancer Res* (2016) 5(1):39–50. doi: 10.3978/j.issn.2218-6751.2016.01.03
- Rudin CM, Awad MM, Navarro A, Gottfried M, Peters S, Csoszi T, et al. Pembrolizumab or Placebo Plus Etoposide and Platinum as First-Line Therapy for Extensive-Stage Small-Cell Lung Cancer: Randomized, Double-Blind, Phase III KEYNOTE-604 Study. *J Clin Oncol* (2020) 38(21):2369–79. doi: 10.1200/JCO.20.00793
- Hiddinga BI, Raskin J, Janssens A, Pauwels P, Van Meerbeeck JP. Recent Developments in the Treatment of Small Cell Lung Cancer. *Eur Respir Rev* (2021) 30(161):210079. doi: 10.1183/16000617.0079-2021

11. Amini A, Byers LA, Welsh JW, Komaki RU. Progress in the Management of Limited-Stage Small Cell Lung Cancer. *Cancer Am Cancer Soc* (2014) 120 (6):790–8. doi: 10.1002/cncr.28505
12. Qi X, Zhang DH, Wu N, Xiao JH, Wang X, Ma W. ceRNA in Cancer: Possible Functions and Clinical Implications. *J Med Genet* (2015) 52(10):710–8. doi: 10.1136/jmedgenet-2015-103334
13. Matsui M, Corey DR. Non-Coding RNAs as Drug Targets. *Nat Rev Drug Discov* (2017) 16(3):167–79. doi: 10.1038/nrd.2016.117
14. 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
15. Ghafouri-Fard S, Shoori H, Branicki W, Taheri M. Non-Coding RNA Profile in Lung Cancer. *Exp Mol Pathol* (2020) 114:104411. doi: 10.1016/j.yexmp.2020.104411
16. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell* (2011) 146(3):353–8. doi: 10.1016/j.cell.2011.07.014
17. Zhou RS, Zhang EX, Sun QF, Ye ZJ, Liu JW, Zhou DH, et al. Integrated Analysis of lncRNA-miRNA-mRNA ceRNA Network in Squamous Cell Carcinoma of Tongue. *BMC Cancer* (2019) 19(1):779. doi: 10.1186/s12885-019-5983-8
18. Liang Y, Zhang CD, Zhang C, Dai DQ. DLX6-AS1/miR-204-5p/OCT1 Positive Feedback Loop Promotes Tumor Progression and Epithelial-Mesenchymal Transition in Gastric Cancer. *Gastric Cancer* (2020) 23 (2):212–27. doi: 10.1007/s10120-019-01002-1
19. Di Leva G, Garofalo M, Croce CM. MicroRNAs in Cancer. *Annu Rev Pathol* (2014) 9:287–314. doi: 10.1146/annurev-pathol-012513-104715
20. Azizi M, Othman I, Naidu R. The Role of MicroRNAs in Lung Cancer Metabolism. *Cancers (Basel)* (2021) 13(7):1716. doi: 10.3390/cancers13071716
21. Wang XJ, Gao J, Wang Z, Yu Q. Identification of a Potentially Functional microRNA-mRNA Regulatory Network in Lung Adenocarcinoma Using a Bioinformatics Analysis. *Front Cell Dev Biol* (2021) 9:641840. doi: 10.3389/fcell.2021.641840
22. Wu T, Du Y. lncRNAs: From Basic Research to Medical Application. *Int J Biol Sci* (2017) 13(3):295–307. doi: 10.7150/ijbs.16968
23. Kumar S, Pandey M, Sharawat SK. Biological Functions of Long Noncoding RNAs and Circular RNAs in Small-Cell Lung Cancer. *Epigenomics-Uk* (2020) 12(19):1751–63. doi: 10.2217/epi-2020-0214
24. Barrett SP, Wang PL, Salzman J. Circular RNA Biogenesis can Proceed Through an Exon-Containing Lariat Precursor. *Elife* (2015) 4:e7540. doi: 10.7554/eLife.07540
25. Stewart PA, Welsh EA, Slebos R, Fang B, Izumi V, Chambers M, et al. Proteogenomic Landscape of Squamous Cell Lung Cancer. *Nat Commun* (2019) 10(1):3578. doi: 10.1038/s41467-019-11452-x
26. Gong L, Zhang D, Dong Y, Lei Y, Qian Y, Tan X, et al. Integrated Bioinformatics Analysis for Identifying the Therapeutic Targets of Aspirin in Small Cell Lung Cancer. *J BioMed Inform* (2018) 88:20–8. doi: 10.1016/j.jbi.2018.11.001
27. Kuenzi BM, Ideker T. A Census of Pathway Maps in Cancer Systems Biology. *Nat Rev Cancer* (2020) 20(4):233–46. doi: 10.1038/s41568-020-0240-7
28. Chakraborty S, Andrieux G, Hasan A, Ahmed M, Hosen MI, Rahman T, et al. Harnessing the Tissue and Plasma lncRNA-Peptidome to Discover Peptide-Based Cancer Biomarkers. *Sci Rep* (2019) 9(1):12322. doi: 10.1038/s41598-019-48774-1
29. Ciojocneanu R, Braicu C, Raduly L, Jurj A, Zanoaga O, Magdo L, et al. Plasma and Tissue Specific miRNA Expression Pattern and Functional Analysis Associated to Colorectal Cancer Patients. *Cancers (Basel)* (2020) 12(4):843. doi: 10.3390/cancers12040843
30. Chen X, Li C, Li Y, Wu S, Liu W, Lin T, et al. Characterization of METTL7B to Evaluate TME and Predict Prognosis by Integrative Analysis of Multi-Omics Data in Glioma. *Front Mol Biosci* (2021) 8:727481. doi: 10.3389/fmolb.2021.727481
31. Lu Y, Jin J, Du Q, Hu M, Wei Y, Wang M, et al. Multi-Omics Analysis of the Anti-Tumor Synergistic Mechanism and Potential Application of Immune Checkpoint Blockade Combined With Lenvatinib. *Front Cell Dev Biol* (2021) 9:730240. doi: 10.3389/fcell.2021.730240
32. Ponzi E, Thoresen M, Haugdahl NT, Mollersen K. Integrative, Multi-Omics, Analysis of Blood Samples Improves Model Predictions: Applications to Cancer. *BMC Bioinf* (2021) 22(1):395. doi: 10.1186/s12859-021-04296-0
33. Kastner S, Voss T, Keuerleber S, Glockel C, Freissmuth M, Sommergruber W. Expression of G Protein-Coupled Receptor 19 in Human Lung Cancer Cells is Triggered by Entry Into S-Phase and Supports G(2)-M Cell-Cycle Progression. *Mol Cancer Res* (2012) 10(10):1343–58. doi: 10.1158/1541-7786.MCR-12-0139
34. Wang L, Meng L, Wang XW, Ma GY, Chen JH. Expression of RRM1 and RRM2 as a Novel Prognostic Marker in Advanced non-Small Cell Lung Cancer Receiving Chemotherapy. *Tumour Biol* (2014) 35(3):1899–906. doi: 10.1007/s13277-013-1255-4
35. Li L, Zhang L, Tian Y, Zhang T, Duan G, Liu Y, et al. Serum Chemokine CXCL7 as a Diagnostic Biomarker for Colorectal Cancer. *Front Oncol* (2019) 9:921. doi: 10.3389/fonc.2019.00921
36. Feser WJ, Fingerlin TE, Strand MJ, Glueck DH. Calculating Average Power for the Benjamini-Hochberg Procedure. *J Stat Theory Appl* (2009) 8(3):325–52.
37. Kozomara A, Birgaoanu M, Griffiths-Jones S. Mirbase: From microRNA Sequences to Function. *Nucleic Acids Res* (2019) 47(D1):D155–62. doi: 10.1093/nar/gky1141
38. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of Mammalian microRNA Targets. *Cell* (2003) 115(7):787–98. doi: 10.1016/s0092-8674(03)01018-3
39. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and Effective Prediction of microRNA/Target Duplexes. *Rna* (2004) 10(10):1507–17. doi: 10.1261/rna.5248604
40. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.Org Resource: Targets and Expression. *Nucleic Acids Res* (2008) 36(Database issue):D149–53. doi: 10.1093/nar/gkm995
41. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. *F1000Res* (2020) 9:ISC Comm J-304. doi: 10.12688/f1000research.23297.2
42. Gao Y, Wang J, Zhao F. CIRI: An Efficient and Unbiased Algorithm for *De Novo* Circular RNA Identification. *Genome Biol* (2015) 16:4. doi: 10.1186/s13059-014-0571-3
43. Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. Mirdeep2 Accurately Identifies Known and Hundreds of Novel microRNA Genes in Seven Animal Clades. *Nucleic Acids Res* (2012) 40(1):37–52. doi: 10.1093/nar/gkr688
44. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* (2003) 13(11):2498–504. doi: 10.1101/gr.1239303
45. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *Omic* (2012) 16(5):284–7. doi: 10.1089/omi.2011.0118
46. Yu K, Kuang L, Fu T, Zhang C, Zhou Y, Zhu C, et al. CREM Is Correlated With Immune-Suppressive Microenvironment and Predicts Poor Prognosis in Gastric Adenocarcinoma. *Front Cell Dev Biol* (2021) 9:697748. doi: 10.3389/fcell.2021.697748
47. Peifer M, Fernandez-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, et al. Integrative Genome Analyses Identify Key Somatic Driver Mutations of Small-Cell Lung Cancer. *Nat Genet* (2012) 44(10):1104–10. doi: 10.1038/ng.2396
48. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, et al. Comprehensive Genomic Analysis Identifies SOX2 as a Frequently Amplified Gene in Small-Cell Lung Cancer. *Nat Genet* (2012) 44(10):1111–6. doi: 10.1038/ng.2405
49. George J, Lim JS, Jang SJ, Cun Y, Ozretic L, Kong G, et al. Comprehensive Genomic Profiles of Small Cell Lung Cancer. *Nature* (2015) 524(7563):47–53. doi: 10.1038/nature14664
50. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res* (2018) 46(D1):D1074–82. doi: 10.1093/nar/gkx1037
51. Wang Y, Yang R, Yan F, Jin Y, Liu X, Wang T. Medcarpin Protects Cerebral Microvascular Endothelial Cells Against Oxygen-Glucose Deprivation/Reoxygenation-Induced Injury via the PI3K/Akt/FoxO Pathway: A Study of Network Pharmacology Analysis and Experimental Validation. *Neurochem Res* (2022) 47(2):347–57. doi: 10.1007/s11064-021-03449-0
52. Wu M, Ma M, Tan Z, Zheng H, Liu X. Neutrophil: A New Player in Metastatic Cancers. *Front Immunol* (2020) 11:565165. doi: 10.3389/fimmu.2020.565165
53. Rawat K, Syeda S, Shrivastava A. Neutrophil-Derived Granule Cargoes: Paving the Way for Tumor Growth and Progression. *Cancer Metastasis Rev* (2021) 40(1):221–44. doi: 10.1007/s10555-020-09951-1

54. Hou S, Wang J, Li W, Hao X, Hang Q. Roles of Integrins in Gastrointestinal Cancer Metastasis. *Front Mol Biosci* (2021) 8:708779. doi: 10.3389/fmolb.2021.708779
55. Roehrig AE, Klupsch K, Oses-Prieto JA, Chaib S, Henderson S, Emmett W, et al. Cell-Cell Adhesion Regulates Merlin/NF2 Interaction With the PAF Complex. *PLoS One* (2021) 16(8):e254697. doi: 10.1371/journal.pone.0254697
56. Yan X, Li F, Dozmorov I, Frank MB, Dao M, Centola M, et al. External Qi of Yan Xin Qigong Induces Cell Death and Gene Expression Alterations Promoting Apoptosis and Inhibiting Proliferation, Migration and Glucose Metabolism in Small-Cell Lung Cancer Cells. *Mol Cell Biochem* (2012) 363(1-2):245–55. doi: 10.1007/s11010-011-1176-8
57. Liu H, Wei S, Zhang L, Yuan C, Duan Y, Wang Q. Secreted Phosphoprotein 1 Promotes the Development of Small Cell Lung Cancer Cells by Inhibiting Autophagy and Apoptosis. *Pathol Oncol Res* (2019) 25(4):1487–95. doi: 10.1007/s12253-018-0504-7
58. Guclu E, Eroglu GC, Kurar E, Vural H. Knockdown of lncRNA HIF1A-AS2 Increases Drug Sensitivity of SCLC Cells in Association With Autophagy. *Med Oncol* (2021) 38(9):113. doi: 10.1007/s12032-021-01562-2
59. Fu X, Deng X, Xiao W, Huang B, Yi X, Zou Y. Downregulation of NEAT1 Sensitizes Gemcitabine-Resistant Pancreatic Cancer Cells to Gemcitabine Through Modulation of the miR-506-3p/ZEB2/EMT Axis. *Am J Cancer Res* (2021) 11(8):3841–56. doi: 10.1186/s13059-014-0571-3
60. Li K, Yao T, Zhang Y, Li W, Wang Z. NEAT1 as a Competing Endogenous RNA in Tumorigenesis of Various Cancers: Role, Mechanism and Therapeutic Potential. *Int J Biol Sci* (2021) 17(13):3428–40. doi: 10.7150/ijbs.62728
61. Xu Y, Li Y, Qiu Y, Sun F, Zhu G, Sun J, et al. lncRNA NEAT1 Promotes Gastric Cancer Progression Through miR-17-5p/TGFbetaR2 Axis Up-Regulated Angiogenesis. *Front Cell Dev Biol* (2021) 9:705697. doi: 10.3389/fcell.2021.705697
62. Zheng Q, Gu X, Yang Q, Chu Q, Dai Y, Chen Z. DLX6-AS1 is a Potential Biomarker and Therapeutic Target in Cancer Initiation and Progression. *Clin Chim Acta* (2021) 517:1–8. doi: 10.1016/j.cca.2021.02.006
63. Qian Y, Song W, Wu X, Hou G, Wang H, Hang X, et al. DLX6 Antisense RNA 1 Modulates Glucose Metabolism and Cell Growth in Gastric Cancer by Targeting microRNA-4290. *Dig Dis Sci* (2021) 66(2):460–73. doi: 10.1007/s10620-020-06223-4
64. Touw IP, van de Geijn GJ. Granulocyte Colony-Stimulating Factor and its Receptor in Normal Myeloid Cell Development, Leukemia and Related Blood Cell Disorders. *Front Biosci* (2007) 12:800–15. doi: 10.2741/2103
65. Trotter AM, Druhan LJ, Kraft IL, Lance A, Feurstein S, Helgeson M, et al. Heterozygous Germ Line CSF3R Variants as Risk Alleles for Development of Hematologic Malignancies. *Blood Adv* (2020) 4(20):5269–84. doi: 10.1182/bloodadvances.2020002013
66. Rashid M, Alasiri A, Al BM, Alkhalidi A, Alsuhailani A, Alsultan A, et al. Identification of CSF3R Mutations in B-Lineage Acute Lymphoblastic Leukemia Using Comprehensive Cancer Panel and Next-Generation Sequencing. *Genes (Basel)* (2021) 12(9):1326. doi: 10.3390/genes12091326
67. Saunders AS, Bender DE, Ray AL, Wu X, Morris KT. Colony-Stimulating Factor 3 Signaling in Colon and Rectal Cancers: Immune Response and CMS Classification in TCGA Data. *PLoS One* (2021) 16(2):e247233. doi: 10.1371/journal.pone.0247233
68. Maxson JE, Gotlib J, Polley DA, Fleischman AG, Agarwal A, Eide CA, et al. Oncogenic CSF3R Mutations in Chronic Neutrophilic Leukemia and Atypical CML. *N Engl J Med* (2013) 368(19):1781–90. doi: 10.1056/NEJMoa1214514
69. Wang Z, Zhou C, Yang S. The Roles, Controversies, and Combination Therapies of Autophagy in Lung Cancer. *Cell Biol Int* (2022) 46(1):3–11. doi: 10.1002/cbin.11704
70. Rupniewska E, Roy R, Mauri FA, Liu X, Kaliszczak M, Bellezza G, et al. Targeting Autophagy Sensitises Lung Cancer Cells to Src Family Kinase Inhibitors. *Oncotarget* (2018) 9(44):27346–62. doi: 10.18632/oncotarget.25213
71. Hurvitz SA, Timmerman JM. Recombinant, Tumour-Derived Idiotype Vaccination for Indolent B Cell non-Hodgkin's Lymphomas: A Focus on FavId. *Expert Opin Biol Ther* (2005) 5(6):841–52. doi: 10.1517/14712598.5.6.841
72. Reinis M. Drug Evaluation: FavId, a Patient-Specific Idiotypic Vaccine for non-Hodgkin's Lymphoma. *Curr Opin Mol Ther* (2007) 9(3):291–8.
73. Alissafi T, Hatzioannou A, Mintzas K, Barouni RM, Banos A, Sormendi S, et al. Autophagy Orchestrates the Regulatory Program of Tumor-Associated Myeloid-Derived Suppressor Cells. *J Clin Invest* (2018) 128(9):3840–52. doi: 10.1172/JCI120888
74. Hamura R, Shirai Y, Shimada Y, Saito N, Taniai T, Horiuchi T, et al. Suppression of Lysosomal Acid Alpha-Glucosidase Impacts the Modulation of Transcription Factor EB Translocation in Pancreatic Cancer. *Cancer Sci* (2021) 112(6):2335–48. doi: 10.1111/cas.14921
75. Vanhove K, Graulus GJ, Mesotten L, Thomeer M, Derveaux E, Noben JP, et al. The Metabolic Landscape of Lung Cancer: New Insights in a Disturbed Glucose Metabolism. *Front Oncol* (2019) 9:1215. doi: 10.3389/fonc.2019.01215
76. Chisari A, Golan I, Campisano S, Gelabert C, Moustakas A, Sancho P, et al. Glucose and Amino Acid Metabolic Dependencies Linked to Stemness and Metastasis in Different Aggressive Cancer Types. *Front Pharmacol* (2021) 12:723798. doi: 10.3389/fphar.2021.723798
77. Qi X, Li Q, Che X, Wang Q, Wu G. The Uniqueness of Clear Cell Renal Cell Carcinoma: Summary of the Process and Abnormality of Glucose Metabolism and Lipid Metabolism in ccRCC. *Front Oncol* (2021) 11:727778. doi: 10.3389/fonc.2021.727778
78. Shin E, Koo JS. Glucose Metabolism and Glucose Transporters in Breast Cancer. *Front Cell Dev Biol* (2021) 9:728759. doi: 10.3389/fcell.2021.728759
79. Grieb BC, Agarwal R. HER2-Directed Therapy in Advanced Gastric and Gastroesophageal Adenocarcinoma: Triumphs and Troubles. *Curr Treat Options Oncol* (2021) 22(10):88. doi: 10.1007/s11864-021-00884-7
80. Indini A, Rijavec E, Grossi F. Trastuzumab Deruxtecan: Changing the Destiny of HER2 Expressing Solid Tumors. *Int J Mol Sci* (2021) 22(9):4774. doi: 10.3390/ijms22094774
81. Riudavets M, Sullivan I, Abdayem P, Planchard D. Targeting HER2 in non-Small-Cell Lung Cancer (NSCLC): A Glimpse of Hope? An Updated Review on Therapeutic Strategies in NSCLC Harbouring HER2 Alterations. *ESMO Open* (2021) 6(5):100260. doi: 10.1016/j.esmoop.2021.100260
82. Martinho O, Silva-Oliveira R, Cury FP, Barbosa AM, Granja S, Evangelista AF, et al. HER Family Receptors are Important Theranostic Biomarkers for Cervical Cancer: Blocking Glucose Metabolism Enhances the Therapeutic Effect of HER Inhibitors. *Theranostics* (2017) 7(3):717–32. doi: 10.7150/thno.17154

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.

Copyright © 2022 Wang, Gao, Yu, Zhang and Hu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Exome Sequencing Reveals Genetic Variability and Identifies Chronic Prognostic Loci in Chinese Sarcoidosis Patients

Qian Zhang¹, Hui Huang¹, Meijun Zhang², Chuling Fang¹, Na Wang¹, Xiaoyan Jing¹, Jian Guo¹, Wei Sun¹, Xiaoyu Yang¹ and Zuojun Xu^{1*}

¹ Department of Respiratory Medicine, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China, ² ANNOROAD Co., Beijing, China

OPEN ACCESS

Edited by:

Liang Cheng,
Harbin Medical University, China

Reviewed by:

Jujuan Zhuang,
Dalian Maritime University, China
Taigang Liu,
Shanghai Ocean University, China

*Correspondence:

Zuojun Xu
xuzj@hotmail.com

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 01 April 2022

Accepted: 30 May 2022

Published: 04 July 2022

Citation:

Zhang Q, Huang H, Zhang M, Fang C, Wang N, Jing X, Guo J, Sun W, Yang X and Xu Z (2022) Exome Sequencing Reveals Genetic Variability and Identifies Chronic Prognostic Loci in Chinese Sarcoidosis Patients. *Front. Oncol.* 12:910227. doi: 10.3389/fonc.2022.910227

Background: Sarcoidosis is an inflammatory disease characterized by non-caseating granuloma formation in various organs, with several recognized genetic and environmental risk factors. Despite substantial progress, the genetic determinants associated with its prognosis remain largely unknown.

Objectives: This study aimed to identify the genetic changes involved in sarcoidosis and evaluate their clinical relevance.

Methods: We performed whole-exome sequencing (WES) in 116 sporadic sarcoidosis patients (acute sarcoidosis patients, n=58; chronic sarcoidosis patients, n=58). In addition, 208 healthy controls were selected from 1000 G East Asian population data. To identify genes enriched in sarcoidosis, Fisher exact tests were performed. The identified genes were included for further pathway analysis using Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG). Additionally, we used the STRING database to construct a protein network of rare variants and Cytoscape to identify hub genes of signaling pathways.

Results: WES and Fisher's exact test identified 1,311 variants in 439 protein-coding genes. A total of 135 single nucleotide polymorphisms (SNPs) on 30 protein-coding genes involved in the immunological process based on the GO and KEGG enrichment analysis. Pathway enrichment analysis showed osteoclast differentiation and cytokine-cytokine receptor interactions. Three missense mutations (rs76740888, rs149664918, and rs78251590) in two genes (PRSS3 and CNN2) of immune-related genes showed significantly different mutation frequencies between the disease group and healthy controls. The correlation of genetic abnormalities with clinical outcomes using multivariate analysis of the clinical features and mutation loci showed that the missense variant (rs76740888, Chr9:33796673 G>A) of PRSS3 [$p=0.04$, odds ratio (OR) = 2.49] was significantly associated with chronic disease prognosis. Additionally, the top two hub genes were CCL4 and CXCR4 based on protein-protein interaction (PPI) network analysis.

Conclusion: Our study provides new insights into the molecular pathogenesis of sarcoidosis and identifies novel genetic alterations in this disease, especially PRSS3, which may be promising targets for future therapeutic strategies for chronic sarcoidosis.

Keywords: whole-exome sequencing, sarcoidosis, non-synonymous mutations, adaptive immune response, chronic sarcoidosis prognosis

INTRODUCTION

Sarcoidosis (MIM 609464) is an immune-mediated disease affecting multiple organs and is characterized by non-caseating necrotizing granulomatous lesions with an elusive etiology (1). The disease is characterized variably across races, and the prognosis and course of the disease depend on the phenotypic characteristics. Compared to white Americans (10.9/100,000), African Americans (35.5/100,000) are affected more frequently, and African Americans tend to develop chronic, severe disease prognoses (2, 3). Some patients may experience spontaneous remission, but others may suffer from a chronic course, ultimately leading to death in severe cases (4). Epidemiological studies suggest that the disease has a racial predisposition and family clustering characteristics (5). Most researchers agree that the etiology of sarcoidosis is due to environmental exposure, genetic factors, and immune system dysregulation. Sarcoidosis is a multiple-gene-affected disease; many published studies have reported the candidate genes of sarcoidosis. Genetic predisposition plays a vital role in the etiology of sarcoidosis and contributes to the heterogeneity of clinical manifestations and prognosis (6).

The classification of sarcoidosis prognosis is based on the duration of the disease course: acute sarcoidosis (≤ 2 years) and chronic sarcoidosis (> 2 years) (7). Up to 40% of patients develop chronic disease with persistent lung inflammation and tissue fibrosis, which contribute to the majority of sarcoidosis mortality (8). Several genetic mutations have been associated with the clinical course of sarcoidosis, and various distinct ethnic groups argue for genetic influence (3). Previous studies suggested that class II HLA-DRB1*03:01 is associated with resolving disease more than the persistent group in Finnish, Croatia, and Czech sarcoidosis patients (9–11). Other non-HLA gene polymorphisms associated with clinical course, including TLR3 (L412F, rs3775291), promoted a persistent clinical phenotype in Irish and American Caucasian patients (12); also, tumor necrosis factor- β (TNF- β) alleles TNF- β 1 and TNF- β 3 were found to be associated with prolonged clinical course in Japan and Dutch sarcoidosis patients, respectively (12, 13).

In addition to classical candidate-gene filtering methods, genome-wide association analysis (GWAS) also contributes to identifying suspected genes associated with disease etiology. Hofmann et al. reported using GWAS to identify the ANXA11 gene as a new susceptibility locus for sarcoidosis from over 440,000 single nucleotide polymorphisms (SNPs) among 500 patients and controls (14). Additionally, Franke et al. found that the C10ORF67 gene was significantly associated with sarcoidosis

and Crohn's disease among over 83,000 SNPs using the GWAS method (15). Using the whole-exome sequencing method, Elisa Lahtela et al. reported that variations in AADACL3 and C1orf158, located on chromosome 1p36.21, were associated with resolved disease prognosis among 72 Finnish patients (16).

However, limited studies have reported the association between genetic markers and chronic sarcoidosis prognosis in the Chinese population, which requires in-depth research to diagnose and treat sarcoidosis. We present a strategy using whole-exome sequencing data of sarcoidosis patients to evaluate which genetic variants distinguish chronic sarcoidosis prognosis. We identified sequence variations in a sample of 116 Chinese sarcoidosis cases, acute and chronic prognosis, to pinpoint the genetic variety of sarcoidosis prognosis.

MATERIALS AND METHODS

Study Population

One hundred sixteen sarcoidosis patients who were consecutive cases from January 2016 to December 2017 in Peking Union Medical College Hospital and 208 healthy controls were selected for the whole-exome sequencing (WES) study. The patients who underwent WES were diagnosed based on the American Thoracic Society (ATS)/European Respiratory Society (ERS)/World Association of Sarcoidosis and Other Granulomatous Disorders (WASOG) criteria (17). The inclusion criteria included clinical manifestation, radiological characteristics, and pathological evidence. The stages of sarcoidosis were determined following the "Scadding" classification for sarcoidosis. Radiological evaluation of sarcoidosis in the outpatient clinic at Peking Union Medical College Hospital was performed by two physicians with expertise in the respiratory department. All patients who had a clinical follow-up of at least 4 years participated. The diagnosis of the patients with sarcoidosis was confirmed by transbronchial lung biopsy (TBLB). The clinical outcomes of sarcoidosis patients were classified into the acute group (resolve within 2 years, $n=58$ patients, 50%) and the chronic group (persisting over 2 years, $n=58$ patients, 50%) (17). All resources were investigated in the Electronic Health Record database of the Peking Union Medical College Hospital. This study was conducted in accordance with the Declaration of Helsinki, and the protocol used to collect human blood samples and clinical resources was approved by the Ethics Committee of Peking Union Medical College Hospital. Written informed consent was obtained from all subjects.

Whole-Exome Sequencing

DNA was extracted from the blood samples using the QIAamp™ DNA and Blood Mini Kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. WES was performed by ANNOROAD Co. (Beijing, China) using the SureSelectXTTarget Enrichment System (G7530-90000) method for exon capture, and a library was constructed. Then, the paired-end sequencing program was run on the Illumina NovaSeq S2 sequencing platform, and 150-bp reads were obtained. CASAVA 1.8 was used to complete imaging analysis and base detection of the high-throughput sequencing image files with data filtering. Burrows–Wheeler Aligner (BWA) v0.7.17 software was used to compare the sequencing results and the human genome reference sequence (UCSC GRCh37/hg19). Then, The Genome Analysis Toolkit (GATK) v3.8 was used to perform variant calling and to identify SNPs and insertions and deletions (InDels).

Initial Variants and Sample Quality Control

We performed the initial variant and sample quality control (shown in **Figure 1**). All QC steps were analyzed using the software package PLINK v1.09. After variant quality control of the raw data, 5,771,425 SNPs passed SNP quality control with the recommendation from GATK. Among them, 3,981,961 (69%) variants passed Hardy–Weinberg equilibrium (HWE) quality control ($p < 1e-6$). Finally, 508,403 SNPs passed the SNP sample

missing rate (<5%) quality control. In addition, 1,223,109 InDels passed quality control with the recommendation from GATK. Of the InDels, 1,000,293 (82%) passed HWE quality control. Additionally, 39,161 InDels passed the sample missingness rate (<5%) quality control. Then, we manipulated the sample quality control and found that all 116 samples and 208 healthy controls passed heterozygosity (mean \pm 4 SD), sample missing rate (<5%), and familiar relationships (π -hat<0.2), meaning that all samples and controls could be used for further evaluation (shown in **Supplementary Figure S1**).

Mutation Site Filtering and Annotation

The SNPs and InDels identified in 116 sarcoidosis patients were tested with Fisher's exact test. In this disease research, sample collection was challenging, and therefore, we had a smaller sample size. However, the samples were randomly collected, and to control for false positives, we used different methods to obtain true positive sites and genes. For the Fisher's exact test, to control population stratification, the genomic inflation factor was used to adjust the chi-2 value and recomputed p -value; to control false positives in multiple comparisons, the p -values were subjected to Bonferroni multiple corrections (Q -value ≤ 0.05). With Fisher's exact test results, the Multimarker Analysis on GenoMic Annotation (MAGMA) v1.6 software package (18) was used to perform SNP-wide mean model for gene-based association analysis with the default setting. SNPs were

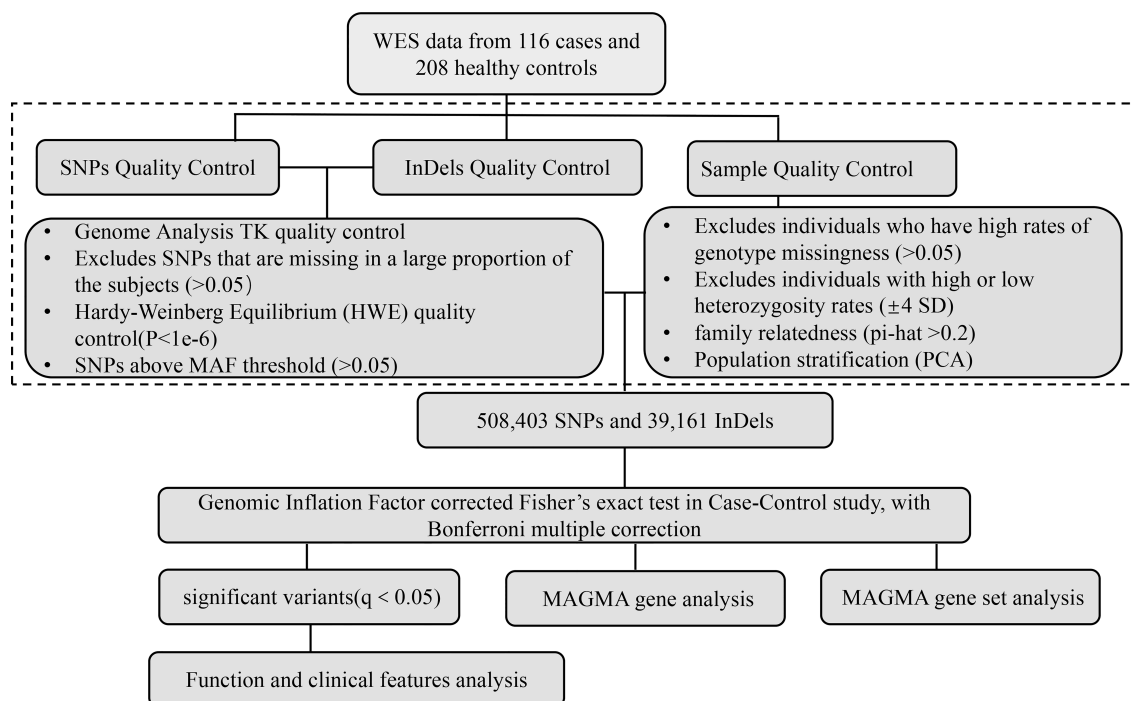


FIGURE 1 | Analytical strategy workflow for variant filtration and candidate gene selection. A schematic overview of the steps involved in whole-exome sequencing analysis with pathogenesis candidate gene detection is shown. SNPs, single nucleotide polymorphisms; InDels, insertions and deletions; WES, whole-exome sequencing; MAF, minor allele frequency; MAGMA, Multimarker Analysis on GenoMic Annotation.

assigned to the genes obtained from Ensembl build 85 (only protein-coding genes). Genome-wide significance was set at 0.05/(the number of tested genes). Genes whose *p*-value reached genome-wide significance can be labeled in the Manhattan plot. Using the result of gene analysis (gene-level *p*-value), gene-set analysis was also performed with default parameters of MAGMA v1.6. The gene sets were obtained from sigdb v7.0 for “Curated gene sets” and “GO terms.” The R package “clusterProfiler” was used to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis (19). The expression characteristics of the immune-related candidate genes were determined based on the data obtained from Genotype-Tissue Expression (GTEx; gtexportal.org). The characteristics of mutations in immune-related candidate genes were identified using Maftools in the R package (20). The STRING online database (STRING, <https://www.string-db.org/>) and PPI pairs with a combined score of ≥ 0.4 were used to construct a PPI network. Cytoscape software v3.7.2 was used to predict the regulatory relationship between genes and analyze the topological parameters of the network. The Genome Reference Consortium Human Build 37 (GRCh37) of *Homo sapiens* in the NCBI database was utilized for SNP description.

Statistical Analysis

Statistical analyses were performed using SPSS version 26 and GraphPad[®] Prism Version 8.0.0 for Mac OS X (San Diego, CA, USA). The normality of the variables was estimated using the Shapiro–Wilk normality test. Non-normally distributed continuous variables were expressed as medians and interquartile ranges [M, (Q1, Q3)], and normally distributed continuous variables were described as the means and standard deviations. Categorical variables are shown as counts and percentages. The independent samples t-test was used for comparing variables with normal distribution. The non-parametric test (Mann–Whitney U test) was used to compare non-normally distributed continuous variables. Pearson’s two test or Fisher’s exact test was used to analyze the categorical variables. Binary logistic regression (backwards method) was used to explore independent factors (age, sex, Lofgren syndrome, radiology stage, rs76740888, and rs78251590) that were statistically significant predictors of the binary dependent variable (disease prognosis). The variables with the highest *p*-values were removed from the model until all *p*-values for the remaining variables were ≤ 0.05 . The logistic models calculated odds ratios (ORs) and their respective 95% confidence intervals (CIs).

RESULTS

Functional Analyses of the Sarcoidosis-Related 439 Candidate Genes

WES, mutation site filtering, and annotation analyses of 116 sarcoidosis patients and 208 healthy controls revealed that 1,311 variants were significant and were allotted to 439 candidate genes

(shown in **Figure 1**). We further performed GO analyses for these candidate genes filtered from the case–control Fisher’s exact test on Metascape.org (21). The most enriched GO terms were the epoxygenase P450 pathway [count = 5 (1.34%), $\log_{10}(p) = -5.26$], keratinization [count = 13 (3.48%), $\log_{10}(p) = (-4.93)$], and defective GALNT3 causing familial hyperphosphatemic tumoral calcinosis (HFTC) [count = 4 (1.07%), $\log_{10}(p) = -4.19$] (shown in **Supplementary Figure S2A**). The top-level Gene Ontology biological processes comprised metabolic process, developmental process, response to stimulus, and cell proliferation (shown in **Supplementary Figure S2B**). To capture the relationships among the terms, we analyzed the network of enriched terms where terms with a similarity >0.3 were connected by edges and selected the terms with the best *p*-value from every 20 clusters using Cytoscape on Metascape.org (shown in **Supplementary Figures S2C, D**).

Further KEGG pathway enrichment of 439 genes in the “cluster profiler” revealed four significant pathway aggregations, including “caffeine metabolism” (gene ratio: 3/90, adjusted *p*-value = 0.0073227), “drug metabolism—other enzymes” (gene ratio: 6/90, adjusted *p*-value = 0.0073227), “retinol metabolism” (gene ratio: 6/90, adjusted *p*-value = 0.01552589), and “drug metabolism—cytochrome P450” (gene ratio: 6/90, adjusted *p*-value = 0.02368937; shown in **Supplementary Figures S3A, B**). In addition, a molecular complex detection (MCODE) analysis was performed to identify the modules within the protein–protein interaction (PPI) network (parameter degree cut-off ≥ 2 and the MCODE score ≥ 1.0) using Cytoscape software (22, 23). We found that 439 candidate genes were significantly clustered into three groups presented in green, red, and blue nodes. MCODE 1 (red nodes, MCODE score = 3.4) has 10 genes, i.e., CUL5, RBX1, EFTUD2, HSPA8, CAND1, RPA1, STAU1, ATAD3A, PABPC1, and SLC25A5. MCODE 2 (blue nodes, MCODE score = 1) contains RB1, ZNF99, and ZNF208. MCODE 3 (green nodes, MCODE score = 1) has three genes, including CYP2A7, CYP2F1, and CYP4F2 (shown in **Supplementary Figures S3C, D**).

In addition, the genes filtered by Fisher’s exact test with the genomic inflation factor are presented in **Supplementary Table S2**. The Manhattan plot (shown in **Supplementary Figure S4A**) showed visual identification of statistically significant data points with $p < 0.05$. We tested for GO term (biological processes) enrichment to assess the gene-set covered biological functions and pathways. Four significantly enriched GO terms were detected, including loneliness (MATG) (*p* adjusted=0.000134505), loneliness (*p* adjusted=0.000742134), extremely high intelligence (*p* adjusted=0.032706257), and Plasminogen activator inhibitor type 1 levels (PAI-1) (*p* adjusted=0.038615263), see **Supplementary Figure S4B**.

Selection and Functional Analysis of Immune-Related Genes

Previous studies suggested that sarcoidosis is an immune-related granulomatous disease associated with genetic susceptibility (24, 25). To distinguish the immune-related pathogenic genes in 439 candidate genes identified in our study, we searched the

“Immune” term among “GO biological process” and found 36 immune-related GO terms (see **Table 1**) covering 135 variants of 30 immune-related genes. The SNP details of the 30 immune-related candidate genes are listed in **Supplementary Table S1**.

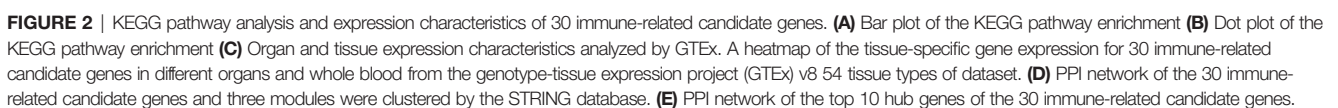
Two significant KEGG pathways enrichments, “hsa04380: osteoclast differentiation” (p -adjust = 0.0000587, gene count = 6) and “hsa04060: cytokine–cytokine receptor interaction” (p -adjust = 0.019641002, gene count = 5), were revealed by analysis of the 30 immune-related genes. Six genes in “osteoclast differentiation” enrichment include CSF1R, LILRA3, LILRA6, LILRB2, LILRB3, and LILRB5. The genes involved in “cytokine–cytokine receptor interaction” included CCL4, CSF1R, CXCR4, FLT1, and IL7 (shown in **Figures 2A, B**). Immune-related

candidate gene expression characteristics were also analyzed using genotype-tissue expression (GTEx) data to investigate the potential role of these genes in multiple organs and tissues. **Figure 2C** shows that CXCR4 and MSTN had the highest expression in the lung among the 30 immune-related candidate genes.

We further analyzed the PPI of 30 immune-related genes using the STRING database. When a “medium confidence = 0.400” was defined as the cutoff criterion of the minimum required interaction score, three clusters were identified from the PPI network (shown in **Figure 2D**). The largest cluster comprised 27 nodes and 24 edges, with an average node degree of 1.78 (PPI enrichment p -value = 2.27×10^{-5}). The hub genes

TABLE 1 | Gene Ontology terms associated with “immune” in Cluster Profiler analyses.

GO ID	GO term	Count	Genes
GO:0002376	Immune system process	30	APOL1 AZGP1P1 BSG CCL4 CNN2 COCH CSF1R CXCR4 EDN1 EZR FLT1 GZMB IL7 KCTD7 KIFAP3 KRT16P3 KRT6A LILRA3 LILRA6 LILRB2 LILRB3 LILRB5 MARCH1 MSTN NR1D1 PRSS3 RB1 RPA1 SAA1 SWAP70
GO:0006955	Immune response	17	APOL1 AZGP1P1 CCL4 COCH CSF1R EDN1 EZR GZMB IL7 KCTD7 KRT16P3 LILRB2 MARCH1 NR1D1 PRSS3 SAA1 SWAP70
GO:0002682	Regulation of immune system process	15	CCL4 COCH CSF1R EDN1 EZR IL7 KCTD7 KRT6A LILRA6 LILRB2 LILRB3 MSTN NR1D1 RB1 SWAP70
GO:0045087	Innate immune response	10	APOL1 CCL4 COCH CSF1R EDN1 GZMB KRT16P3 NR1D1 PRSS3 SAA1
GO:0002520	Immune system development	9	CNN2 CSF1R IL7 LILRA6 LILRB2 LILRB3 RB1 RPA1 SWAP70
GO:0002684	Positive regulation of immune system process	9	CCL4 COCH EDN1 EZR IL7 KCTD7 LILRB2 NR1D1 SWAP70
GO:0002252	Immune effector process	5	GZMB KCTD7 KRT6A MSTN SWAP70
GO:0050776	Regulation of immune response	5	COCH EZR KCTD7 LILRB2 NR1D1
GO:0002683	Negative regulation of immune system process	4	EZR KCTD7 LILRB2 NR1D1
GO:0002764	Immune response-regulating signaling pathway	4	EZR KCTD7 LILRB2 NR1D1
GO:0050778	Positive regulation of immune response	4	COCH EZR KCTD7 NR1D1
GO:0002697	Regulation of immune effector process	3	KCTD7 KRT6A MSTN
GO:0002768	Immune-response-regulating cell surface receptor signaling pathway	3	EZR KCTD7 LILRB2
GO:0002757	Immune-response-activating signal transduction	3	EZR KCTD7 NR1D1
GO:0002253	Activation of immune response	3	EZR KCTD7 NR1D1
GO:0002366	Leukocyte activation involved in immune response	2	KCTD7 SWAP70
GO:0002263	Cell activation involved in immune response	2	KCTD7 SWAP70
GO:0045089	Positive regulation of innate immune response	2	COCH NR1D1
GO:0045088	Regulation of innate immune response	2	COCH NR1D1
GO:0002429	Immune-response-activating cell surface receptor signaling pathway	2	EZR KCTD7
GO:0002767	Immune-response-inhibiting cell surface receptor signaling pathway	1	LILRB2
GO:0002765	Immune-response-inhibiting signal transduction	1	LILRB2
GO:0002279	Mast cell activation involved in immune response	1	KCTD7
GO:0002312	B-cell activation involved in immune response	1	SWAP70
GO:0002562	Somatic diversification of immune receptors via germline recombination within a single locus	1	SWAP70
GO:0002200	Somatic diversification of immune receptors	1	SWAP70
GO:0050777	Negative regulation of immune response	1	LILRB2
GO:0002698	Negative regulation of immune effector process	1	KCTD7
GO:0002285	Lymphocyte activation involved in immune response	1	SWAP70
GO:0002699	Positive regulation of immune effector process	1	KCTD7
GO:0002758	Innate immune response-activating signal transduction	1	NR1D1
GO:0002218	Activation of innate immune response	1	NR1D1
GO:0016064	Immunoglobulin mediated immune response	1	SWAP70
GO:0002460	Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	1	SWAP70
GO:0006959	Humoral immune response	1	IL7
GO:0002250	Adaptive immune response	1	SWAP70



were determined by overlapping the genes according to the top 10 nodes selected by the degree in cytoHubba (26). The identification of hub genes and module interactions is helpful in selecting the key genes that reveal the underlying molecular mechanisms of sarcoidosis pathogenesis-associated immune-related candidate genes (27). The top 10 hub genes were selected, and they were arranged by rank degree and presented in different colors (higher rank degree labeled red, lower rank degree marked yellow). The genes with the most significant rank were CCL4 and CXCR4, which had the most interrelation with other associated immune-associated genes (shown in **Figure 2E**).

Three Missense Mutations Suggested Immune-Related Pathogenesis of Sarcoidosis

We next investigated non-synonymous SNPs, which have been thought to play a more critical role in pathogenesis, as they have different alleles encoding different amino acids. A total of nine non-synonymous SNPs were found in four immune-related genes (shown in **Table 2**). Among them, seven non-synonymous variants in three genes showed a significant difference in the frequency between 116 sarcoidosis and 208 healthy control groups, with much higher frequencies in the sarcoidosis group than in the control group ($p < 0.001$, odds ratio ≥ 1 ; see **Table 2**), including PRSS3, LILRA6 (LILRB3), and CNN2.

Furthermore, the mutation characteristics of 135 SNPs from the 30 immune-related candidate genes were analyzed using Maftools in R software. Among the three immune-related genes with significant mutation frequencies, the PRSS3 and CNN2 genes were detected, which contained the highest missense mutation ratios, at 100% and 97%, respectively, among the 116 sarcoidosis patients. Furthermore, non-synonymous SNPs in

LILRB2, GZMB, APOL1, CNN2, SWAP70, and CSF1R were shown in over 50% of sarcoidosis patients. The SAA1 gene showed multiple hit and splice site mutations among 116 sporadic sarcoidosis patients (shown in **Figure 3A**). Two non-synonymous variants (NC_000009.11:g.33796673G>A and NC_000009.11:g.33797969T>A) in the PRSS3 gene were located on exon 3 of Trypsin-3 isoform 3 and exon 2 of Trypsin-3 isoforms 1, 2, and 4. A non-synonymous SNP in CNN2 (NC_000019.9: g.1037871C>A) was found in exon 7 of calponin-2 isoforms a, c, and d and exon 6 of calponin-2 isoform b (shown in **Figures 3B-F**).

Univariate Analysis and Multivariate Logistic Regression Analysis Detected the Risk Factors for Disease Prognosis

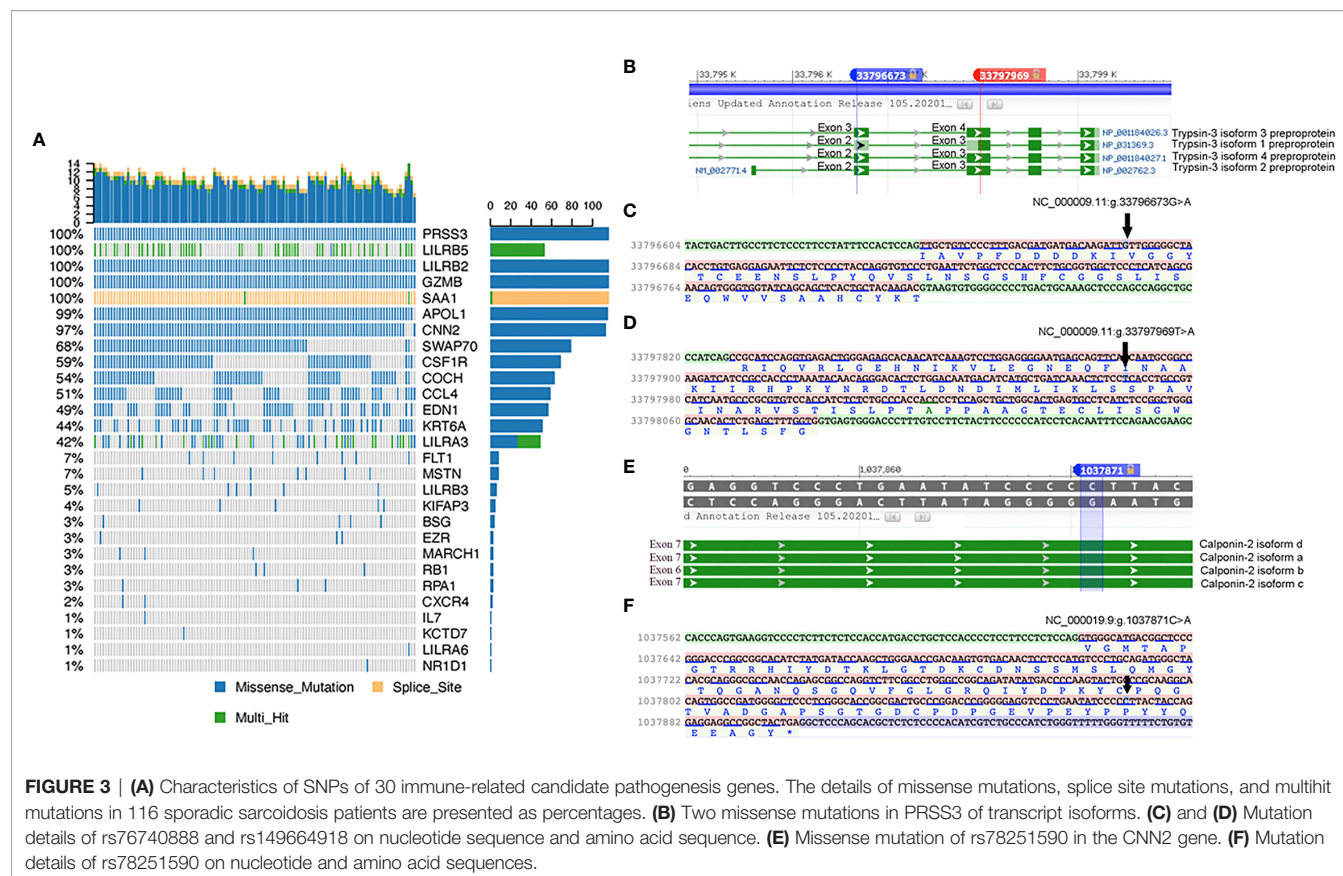
This WES study included 58 (50%) acute sarcoidosis patients and 58 (50%) chronic prognosis sarcoidosis patients (shown in **Table 3**). Univariate analysis of the acute and chronic prognosis groups showed that the acute prognosis group was younger than the chronic prognosis group ($p=0.016$). Additionally, rs76740888 (G to A) and rs78251590 (C to A) mutations were associated with disease prognosis ($p=0.034$ and $p<0.001$, respectively).

The related risk factors with $p<0.1$, including age, sex, Lofgren syndrome, radiology stage, rs76740888, and rs78251590, were set as the independent variables and included in multivariate logistic analyses. The disease prognosis was determined as the dependent variable. After adjustments for the founding variables, the binary logistic regression analysis showed that only age ($p=0.037$), radiology stage II ($p=0.03$), and rs76740888 ($p=0.038$) were retained as significant predictors of sarcoidosis prognosis. **Table 4** lists the variables and parameters that were finally screened into the model. Only age older than 50 (OR, 0.41),

TABLE 2 | Immune-associated non-synonymous variation details in the Fisher's exact test.

Chr	Position	Chromosomal location	Ref	Alt	Variation ID	Frequency of case group	Frequency of control group	p-value	OR	Functional annotation	Gene detail	Exonic function
9	33796673	9p13.3	G	A	rs76740888	0.3147	0.01683	1.83E-28	26.83	Exonic	PRSS3	Non-synonymous
9	33797969	9p13.3	T	A	rs149664918	0.2241	0.05769	1.16E-09	4.719	Exonic	PRSS3	Non-synonymous
17	38253621	17q21.1	A	G	rs201066687	0	0.2428	4.58E-22	0	Exonic	NR1D1	Non-synonymous
19	54744710	19q13.42	C	T	rs1132600	0.1724	0.009615	4.27E-15	21.46	Exonic	LILRA6, LILRB3	Non-synonymous
19	54744711	19q13.42	C	G	rs1132599	0.1724	0.009615	4.27E-15	21.46	Exonic	LILRA6, LILRB3	Non-synonymous
19	54744722	19q13.42	T	C	rs1132597	0.181	0.009615	5.16E-16	22.77	Exonic	LILRA6, LILRB3	Non-synonymous
19	54745989	19q13.42	G	C	rs1052963	0.6336	0.3077	9.31E-16	3.891	Exonic	LILRA6, LILRB3	Non-synonymous
19	1037871	19p13.3	C	A	rs78251590	0.3621	0.06731	1.24E-20	7.865	Exonic	CNN2	Non-synonymous
19	1037640	19p13.3	C	T	rs200303627	0	0.1875	9.56E-17	0	Exonic	CNN2	Non-synonymous

Chr, chromosome; Ref, reference genome base type; Alt, alteration of sample base type; p-value, p-value of Fisher's exact test between sarcoidosis case group and healthy control group; function annotation, region of mutation site annotation by refGene database; gene detail, annotation of transcripts related to mutation sites based on refGene database; exonic function, annotation of exome region from refGene database; AA change, annotation of amino acid changes of mutation sites based on refGene database.



radiology stage II classification (OR, 0.25), and rs76740888 G to A mutation (OR, 2.49) were identified as independent factors correlated with an increase risk of chronic disease prognosis.

DISCUSSION

Sarcoidosis is an inflammatory disease characterized by granulomatosis present in multiple organs and triggered by environmental factors that interact with environmental triggers to result in the innate immune activation of macrophages and dendritic cells, which further upregulates the expression of the major histocompatibility complex (MHC) and cytokines that induce the activation of the adaptive immune response (3). Previous GWAS on the sporadic and familial aggregation of sarcoidosis patients showed that the candidate genes are strongly associated with disease severity, including HLA and non-HLA genes (9, 12). Most of the genes were related to T-cell regulation and T-cell activation during antigen presentation by APCs (antigen-presenting cells) (5). Other genes associated with immune regulation of sarcoidosis, including NOTCH4, TNF α , NOD2, and ANXA11, were also detected by GWAS analysis in different races (28–31). Multiple factors, including genetic composition and the context of antigen presentation, could impact the inflammatory immune response, resulting in a self-limiting or a chronic relapse type of sarcoidosis prognosis. The

current study points out that the abundance of SNPs is associated with disease susceptibility and prognosis in sarcoidosis patients. A few papers have reported that human leukocyte antigen (HLA) DRB1*15 positivity is associated with an increased risk for a chronic course of sarcoidosis (32, 33). However, the susceptibility variants and the signaling pathways are different among different races of sarcoidosis patients (34–36). To our knowledge, this result is the first report on the genetics of Chinese sporadic sarcoidosis patients, which has significance for understanding immunogenetic pathogenesis and the development of chronic disease prognosis.

The present study revealed that 1,311 variants in 439 genes were present in 116 sporadic Chinese sarcoidosis patients compared to 208 healthy controls. Enrichment analysis with GO biological process terms revealed that 135 variants in 30 genes were related to the “Immune” associated GO term. The PRSS3 and CNN2 genes were detected with the highest missense mutation ratios (100% and 97%, respectively).

The PRSS3 (serine protease 3) gene product, trypsin-3, is a trypsinogen of the trypsin family of serine proteases and is expressed in multiple organs, such as the lung. The PRSS3 gene is located on the locus of T-cell receptor beta variable orphan on chromosome 9 [cytogenetic location: 9p13.3; genomic coordinates (GRCh37/hg19) 33750677–33799229] and is associated with thyroiditis and Rickettsialpox (37, 38). This gene is localized to the locus of T-cell receptor beta variable

TABLE 3 | Univariate analysis of three missense mutations and clinical characteristics of 116 sarcoidosis patients.

Characteristic	Acute disease (<2 years)		Chronic disease (≥2 years)		P
No. of patients (%)	58	(%)	58	(%)	
Age					0.016 ^a
<50 years	37	64%	24	41%	
≥50 years	21	36%	34	59%	
Gender					0.077 ^a
Female	34	59%	43	74%	
Male	24	41%	15	26%	
Syndrome					0.079 ^a
Lofgren syndrome	6		13		
Extrapulmonary involvement	8		8		1 ^a
Laboratory tests					0.343 ^a
ACE					
<68 (U/L)	49	84%	45	78%	
≥68 (U/L)	9	16%	13	22%	
ESR					0.709 ^a
<15 (mm/h)	33	57%	31	53%	
≥15 (mm/h)	25	43%	27	47%	
hsCRP					0.576 ^a
<3 (mg/L)	30	52%	33	57%	
≥3 (mg/L)	28	48%	25	43%	
Ca (mmol/L)					0.611 ^a
<2.70 (mmol/L)	57	98%	55	95%	
≥2.70 (mmol/L)	1	2%	3	5%	
ALT (U/L)					0.488 ^a
<40 (U/L)	55	95%	52	90%	
≥40 (U/L)	3	5%	6	10%	
NLR (X ± SD)	2.68 (1.92, 4.00)		2.57 (2.02, 3.35)		0.359 ^b
BALF					0.793 ^a
CD4/CD8 ratio					
<2.0	9	16%	8	14%	
≥2.0	49	84%	50	86%	
PFT					0.964 ^b
FEV1/FVC (%) [M, (Q1, Q3)]	79.43 (75.59, 82.02)		78.95 (73.87, 84.29)		
DLCO (% pred) (X ± SD)	83.45 ± 13.22		84.34 ± 14.17		0.922 ^c
CPI [M, (Q1, Q3)]	12.29 (2.21, 19.95)		14.8 (5.45, 23.41)		0.316 ^b
Radiology stage ^c					0.09 ^a
Stage I	19	33%	10	17%	
Stage II	30	52%	32	55%	
Stage III	9	16%	16	28%	
SNPs					0.034 ^a
rs76740888					
GG	16		27		
GA	42		31		
rs149664918					0.709 ^a
TT	33		31		
TA	25		27		
rs78251590					<0.001 ^a
CC	44		14		
CA	14		44		

^aχ² test.^bMann–Whitney U test.^cIndependent samples t-test.

X ± SD, mean ± standard deviation; M, (Q1, Q3), median, first quartile, and the third quartile.

Radiology stage was according to the Scadding classification.

ACE, angiotensin converting enzyme; ESR, erythrocyte sedimentation rate; hsCRP, hypersensitive C-reactive protein; Ca, calcium in serum; ALT, alanine aminotransferase; NLR, neutrophil–lymphocyte ratio; BALF, bronchoalveolar lavage fluid; PFT, pulmonary function test; FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity; DLCO, diffusing capacity of the lung for carbon monoxide for single-breath method; CPI, complex physiological index.

orphans. It has been suggested to be involved in the proteolytic processing of proteins, digestion, blood coagulation, immune response, and development (39). Mesotrypsin/PRSS3 is an atypical isoform of trypsin that is expressed in the brain and other organs and is involved in the process of antimicrobial humoral response, cobalamin metabolic process, digestion,

endothelial cell migration, neutrophil degranulation, proteolysis, and zymogen activation *via* calcium ion binding, protein binding, and serine-type endopeptidase activity signaling pathway (40). Pathways related to the PRSS3 gene on KEGG are “Influenza A,” “neuroactive ligand–receptor interaction,” “pancreatic secretion,” and “protein digestion and absorption”

TABLE 4 | Results of logistic regression analysis of risk factors for disease prognosis.

Variables	B	SE	Wald	p-value	OR value	95% CI for Exp(B)
Age (≥50)	−0.89	0.43	4.34	0.04	0.41	0.18–0.95
Sex (female)	−0.75	0.46	2.74	0.1	0.47	0.19–1.15
With Löfgren syndrome	−1.03	0.58	3.14	0.08	0.36	0.12–1.12
Radiology stage						
Stage I	–	–	–	–	Ref	
Stage II	−1.4	0.63	4.93	0.03	0.25	0.07–0.85
Stage III	−0.27	0.53	0.25	0.62	0.77	0.27–2.17
SNPs						
rs76740888						
GG	–	–	–	–	Ref	
GA	0.91	0.44	4.33	0.04	2.49	1.05–5.89

(41). Diseases associated with PRSS3 include thyroiditis and Hashimoto thyroiditis. The elevated expression of PRSS3 is associated with a poor prognosis for multiple cancers, including lung adenocarcinoma, gastric cancer, ductal carcinoma of the breast, and pancreatic cancer (37, 42–44). Two protein-coding SNPs on the PRSS3 gene identified in our research have not yet been published. The G to A missense mutation (rs76740888) in the exon of PRSS3 on chromosome 9p13.3 could cause an mRNA allele change and an amino acid change in four trypsin-3 isoforms, causing 10 coding sequence variants at each codon and amino acid. Protease imbalances have been found in another interstitial lung disease. Shanna Ashley et al. suggested that trypsin-3 was a potential biomarker for idiopathic pulmonary pneumonia (IPF) by proteomic analysis of plasma from IPF patients (45). How trypsin-3 influences sarcoidosis is still unknown.

CNN2 is located on chromosome 19p13.3 and functions as an actin cytoskeleton-associated protein and modifies the innate immune system pathways, including the inhibitory regulation of macrophage migration and phagocytosis (46). CNN2 is expressed in many organ tissues and cells, including epidermal keratinocytes, lung alveolar cells, and fibroblasts. In our study, we identified a C>A missense mutation (rs78251590) on CNN2 that may participate in the regulation of immune regulation of sarcoidosis.

We also attempted to identify biological pathways by inputting 30 candidate genes from the immune-related GO category. Two pathways were extracted from the KEGG analysis, including “hsa04380: osteoclast differentiation” (p-adjust = 0.0000587, gene count = 6) and “hsa04060: cytokine–cytokine receptor interaction” (p-adjust = 0.019641002, gene count = 5). A recent investigation of sarcoidosis illustrated that the differentially expressed genes (DEGs) identified by comparing the microarray datasets between sarcoidosis patients and healthy controls were significantly enriched in the positive regulation of protein kinase activity, osteoblast differentiation, and inflammatory response (47). The osteoclast differentiation and cytokine–cytokine receptor interaction pathways identified in our study may provide new ideas for understanding the role of immune-related gene pathways in Chinese sarcoidosis patients. Meanwhile, the “hsa04060: cytokine–cytokine receptor interaction” pathway, including the CCL4, CSF1R, CXCR4, FLT1, and IL7 genes, could be highly

associated with immune regulation and is strongly suspected to be involved in the pathogenesis of sarcoidosis (5, 48, 49). Interleukin (IL)-7 is essential for T-cell generation and plays a pivotal role in the proliferation and survival of memory and naive T cells and T helper type 17 (Th17) cells. Elliott Crouser et al. reported that IL-7 gene transcripts and transcript networks were highly engaged in pulmonary sarcoidosis biological processes and observed overexpression of the IL-7 protein in sarcoidosis patients. Similarly, Patterson et al. observed that the circulating cytokine IL-7 was increased in sarcoidosis patients compared to the control group (50). Keiichiro Yoshioka et al. used Gene Ontology enrichment analysis with RNA sequencing datasets. They revealed several biological processes related to the pathogenesis of sarcoidosis, such as cellular response to IL-1 and interferon gamma (IFN- γ), regulation of IL-6 production, and response to lipopolysaccharide. Meanwhile, they confirmed that the tumor necrosis factor (TNF), toll-like receptor signaling, and IL-17 signaling pathways were involved in the sarcoidosis pathobiology from KEGG pathway enrichment analysis (51).

The leading hub genes with variants are also essential regulators due to their changes in the activities of proteins and regulation mechanisms (47). Based on the analysis of the top 10 hub genes among 30 immune-related candidate genes, we found that CCL4 and CXCR4 were the most significant interrelated genes. C–C motif chemokine ligand 4 (CCL4) encodes a mitogen-inducible monokine involved in PEDF-induced signaling and the Akt signaling pathway, which could be secreted and involved in inflammatory functions. Barczyk et al. reported that the release of CCL4 chemokine was found to play a significant role in the recruitment of CD8+ T cells and CD4+ T cells to the inflammation sites in sarcoidosis patients (52). Another hub gene that we investigated was CXCR4, which encodes the C–X–C chemokine receptor type 4 protein and is characterized as the receptor for the C–X–C chemokine CXCL12/SDF-1 that transduces a signal by increasing intracellular calcium ion levels and contributes to enhancing MAPK1/MAPK3 activation. Katerina Antoniou et al. suggested that a significant increase in CXCR4 mRNA levels has been detected in sarcoidosis patients compared with healthy controls (53). CXCR4 has a functional relationship with sarcoidosis. The binding of bacterial lipopolysaccharide (LPS) mediates the LPS-induced inflammatory response and affects TNF secretion by monocytes, which are involved in excessive cytokine responses

and induce the development of pulmonary sarcoidosis (54). LPS is mainly detected as a potential non-tuberculosis-associated pathogen-associated molecular pattern (PAMP) in sarcoidosis patients, which is an essential factor for the pathogenesis of sarcoidosis (55).

Interestingly, according to GO analysis of genes filtered from Fisher's exact test with genomic inflation factor adjustment, we identified that the GO term "plasminogen activator inhibitor type 1 levels (PAI-1)" showed the highest proportion of overlapping genes in gene sets. PAI-1, also called "serpin family E member 1 (SERPINE1)," a member of the serine proteinase inhibitor (serpin) superfamily, has been shown to promote fibrosis in multiple organ systems and function as a component of innate antiviral immunity. Florence Jeny et al. identified that hypoxia increased the profibrotic response with PAI-1 secretion associated with human lung fibroblast migration inhibition in monocyte-derived (MD) macrophages among highly active sarcoidosis patients (56).

Finally, the correlation of genetic profiles with clinical outcomes through multivariate analysis showed that the missense variant (rs76740888, Chr9:33796673 G>A) of PRSS3 [$p=0.04$, odds ratio (OR)=2.49] was significantly associated with chronic prognosis. However, this candidate gene should be further analyzed to explore its potential and contribution to sarcoidosis prognosis. Furthermore, in keeping with prior reports, individuals with Stage II radiological classification had a more severe prognosis than those seen for the other stages. Manuel Rubio-Rivas and colleagues conducted a retrospective cohort study of 691 sarcoidosis patients. They suggested that stage II radiological classification at diagnosis was one of the risk factors related to the chronic trend of sarcoidosis (57).

Therefore, according to this study, the identified GO and KEGG pathways and immune candidate genes may act as pathogenesis and prognosis impactors for sarcoidosis. We acknowledge some limitations in our study. First, we need to validate the mechanisms that underlie the association between all genetic variants and sarcoidosis outcomes and the mediating pathway. Second, the findings need to be evaluated in larger cohorts before generalization due to the result being based on patients from a single center who developed sarcoidosis.

CONCLUSION

Our WES study identified 135 SNPs in 30 candidate genes enriched in immune-related GO and KEGG pathways. Of these genes, we found that patients who carried missense mutations of rs76740888 (Chr9:33796673 G to A) on the PRSS3 gene had a higher probability of a chronic sarcoidosis

prognosis. In addition, through a rigorous interrogation of candidate mutations in genes using available informatic data resources, we envisaged that the highly ranked hub genes among 30 immune-related candidate genes could also contribute to the pathogenesis of sarcoidosis, including CCL4 and CXCR4. Taken together, our data support the further understanding of the role of genetic mutations in immune regulation leading to the pathogenesis of sarcoidosis.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the NCBI repository, accession number PRJNA848857.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Bioethics Committee of the Medical University of Peking Union Medical College Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

QZ led the study at all stages and drafted the manuscript. CF, HH, and ZX designed the project and are involved at all stages. MZ and QZ designed and performed the data analyses. NW, WS, and XJ collected the clinical resources collection. JG and XY contributed to revising the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by The National Natural Science Foundation of China (Grant 82070067) and The Beijing Municipal Natural Science Foundation (Grant 7212076).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.910227/full#supplementary-material>

REFERENCES

- Spagnolo P, Rossi G, Trisolini R, Sverzellati N, Baughman RP, Wells AU. Pulmonary Sarcoidosis. *Lancet Respir Med* (2018) 6(5):389–402. doi: 10.1016/S2213-2600(18)30064-X
- Iannuzzi MC, Rybicki BA, Teirstein AS. Sarcoidosis. *N Engl J Med* (2007) 357(21):2153–65. doi: 10.1056/NEJMra071714
- Iannuzzi MC, Rybicki BA. Genetics of Sarcoidosis: Candidate Genes and Genome Scans. *Proc Am Thorac Soc* (2007) 4(1):108–16. doi: 10.1513/pats.200607-141JG
- Carmona EM, Kalra S, Ryu JH. Pulmonary Sarcoidosis: Diagnosis and Treatment. *Mayo Clin Proc* (2016) 91(7):946–54. doi: 10.1016/j.mayocp.2016.03.004
- Moller DR, Rybicki BA, Hamzeh NY, Montgomery CG, Chen ES, Drake W, et al. Genetic, Immunologic, and Environmental Basis of Sarcoidosis. *Ann Am*

- Thorac Soc* (2017) 14(Supplement_6):S429–S36. doi: 10.1513/AnnalsATS.201707-565OT
6. Grunewald J, Spagnolo P, Wahlstrom J, Eklund A. Immunogenetics of Disease-Causing Inflammation in Sarcoidosis. *Clin Rev Allergy Immunol* (2015) 49(1):19–35. doi: 10.1007/s12016-015-8477-8
 7. Valeyre D, Prasse A, Nunes H, Uzunhan Y, Brillet PY, Muller-Quernheim J. Sarcoidosis. *Lancet* (2014) 383(9923):1155–67. doi: 10.1016/S0140-6736(13)60680-7
 8. Ungprasert P, Ryu JH, Matteson EL. Clinical Manifestations, Diagnosis, and Treatment of Sarcoidosis. *Mayo Clin Proc Innov Qual Outcomes* (2019) 3(3):358–75. doi: 10.1016/j.mayocpiqo.2019.04.006
 9. Wennerstrom A, Pietinalho A, Vauhkonen H, Lahtela L, Palikhe A, Hedman J, et al. HLA-DRB1 Allele Frequencies and C4 Copy Number Variation in Finnish Sarcoidosis Patients and Associations With Disease Prognosis. *Hum Immunol* (2012) 73(1):93–100. doi: 10.1016/j.humimm.2011.10.016
 10. Grubic Z, Zuncic R, Peros-Golubicic T, Tekavec-Trkanjec J, Martinez N, Alilovic M, et al. HLA Class I and Class II Frequencies in Patients With Sarcoidosis From Croatia: Role of HLA-B8, -DRB1*0301, and -DQB1*0201 Haplotype in Clinical Variations of the Disease. *Tissue Antigens* (2007) 70(4):301–6. doi: 10.1111/j.1399-0039.2007.00904.x
 11. Mrazek F, Holla LI, Hutyrova B, Znojil V, Vasku A, Kolek V, et al. Association of Tumour Necrosis Factor-Alpha, Lymphotoxin-Alpha and HLA-DRB1 Gene Polymorphisms With Lofgren's Syndrome in Czech Patients With Sarcoidosis. *Tissue Antigens* (2005) 65(2):163–71. doi: 10.1111/j.1399-0039.2005.00370.x
 12. Cooke G, Kamal I, Strengert M, Hams E, Mawhinney L, Tynan A, et al. Toll-Like Receptor 3 L412F Polymorphism Promotes a Persistent Clinical Phenotype in Pulmonary Sarcoidosis. *QJM* (2018) 111(4):217–24. doi: 10.1093/qjmed/hcx243
 13. Yamaguchi E, Itoh A, Hizawa N, Kawakami Y. The Gene Polymorphism of Tumor Necrosis Factor-Beta, But Not That of Tumor Necrosis Factor-Alpha, is Associated With the Prognosis of Sarcoidosis. *Chest* (2001) 119(3):753–61. doi: 10.1378/chest.119.3.753
 14. Hofmann S, Franke A, Fischer A, Jacobs G, Nothnagel M, Gaede KI, et al. Genome-Wide Association Study Identifies ANXA11 as a New Susceptibility Locus for Sarcoidosis. *Nat Genet* (2008) 40(9):1103–6. doi: 10.1038/ng.198
 15. Franke A, Fischer A, Nothnagel M, Becker C, Grabe N, Till A, et al. Genome-Wide Association Analysis in Sarcoidosis and Crohn's Disease Unravels a Common Susceptibility Locus on 10p12.2. *Gastroenterology* (2008) 135(4):1207–15. doi: 10.1053/j.gastro.2008.07.017
 16. Lahtela E, Kankainen M, Sinisalo J, Selroos O, Lokki ML. Exome Sequencing Identifies Susceptibility Loci for Sarcoidosis Prognosis. *Front Immunol* (2019) 10:2964. doi: 10.3389/fimmu.2019.02964
 17. Committee tEE. Statement on Sarcoidosis. Joint Statement of the American Thoracic Society (ATS), the European Respiratory Society (ERS) and the World Association of Sarcoidosis and Other Granulomatous Disorders (WASOG) Adopted by the ATS Board of Directors and by the ERS Executive Committee, February 1999. *Am J Respir Crit Care Med* (1999) 160(2):736–55. doi: 10.1164/ajrccm.160.2.ats4-99
 18. de Leeuw CA, Mooij JM, Heskies T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput Biol* (2015) 11(4):e1004219. doi: 10.1371/journal.pcbi.1004219
 19. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* (2012) 16(5):284–7. doi: 10.1089/omi.2011.0118
 20. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: Efficient and Comprehensive Analysis of Somatic Variants in Cancer. *Genome Res* (2018) 28(11):1747–56. doi: 10.1101/gr.239244.118
 21. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets. *Nat Commun* (2019) 10(1):1523. doi: 10.1038/s41467-019-09234-6
 22. Roeder HG, Manke T, O'Keefe S, Vingron M, Haas SA. PASTAA: Identifying Transcription Factors Associated With Sets of Co-Regulated Genes. *Bioinformatics* (2009) 25(4):435–42. doi: 10.1093/bioinformatics/btn627
 23. Bader GD, Hogue CW. An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks. *BMC Bioinf* (2003) 4:2. doi: 10.1186/1471-2105-4-2
 24. Starshinova AA, Malkova AM, Basantsova NY, Zinchenko YS, Kudryavtsev IV, Ershov GA, et al. Sarcoidosis as an Autoimmune Disease. *Front Immunol* (2019) 10:2933. doi: 10.3389/fimmu.2019.02933
 25. Wolin A, Lahtela EL, Anttila V, Petrek M, Grunewald J, Van Moersel CHM, et al. SNP Variants in Major Histocompatibility Complex Are Associated With Sarcoidosis Susceptibility-A Joint Analysis in Four European Populations. *Front Immunol* (2017) 8:422. doi: 10.3389/fimmu.2017.00422
 26. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, et al. Evidence for Dynamically Organized Modularity in the Yeast Protein-Protein Interaction Network. *Nature* (2004) 430(6995):88–93. doi: 10.1038/nature02555
 27. Das S, Meher PK, Rai A, Bhar LM, Mandal BN. Statistical Approaches for Gene Selection, Hub Gene Identification and Module Interaction in Gene Co-Expression Network Analysis: An Application to Aluminum Stress in Soybean (Glycine Max L.). *PLoS One* (2017) 12(1):e0169605. doi: 10.1371/journal.pone.0169605
 28. Casanova NG, Gonzalez-Garay ML, Sun B, Bime C, Sun X, Knox KS, et al. Differential Transcriptomics in Sarcoidosis Lung and Lymph Node Granulomas With Comparisons to Pathogen-Specific Granulomas. *Respir Res* (2020) 21(1):321. doi: 10.1186/s12931-020-01537-3
 29. Ramos-Casals M, Retamozo S, Siso-Almirall A, Perez-Alvarez R, Pallares L, Brito-Zeron P. Clinically-Useful Serum Biomarkers for Diagnosis and Prognosis of Sarcoidosis. *Expert Rev Clin Immunol* (2019) 15(4):391–405. doi: 10.1080/1744666X.2019.1568240
 30. Chen X, Zhou Z, Zhang Y, Cheng X, Guo X, Yang X. NOD2/CARD15 Gene Polymorphisms and Sarcoidosis Susceptibility: Review and Meta-Analysis. *Sarcoidosis Vasc Diffuse Lung Dis* (2018) 35(2):115–22. doi: 10.36141/svdl.v35i2.6257
 31. Karakaya B, van der Vis JJ, Veltkamp M, Biesma DH, Grutters JC, van Moersel CHM. ANXA11 rs1049550 Associates with Lofgren's Syndrome and Chronic Sarcoidosis Patients. *Cells* (2022) 11(9). doi: 10.3390/cells11091557
 32. Berlin M, Fogdell-Hahn A, Olerup O, Eklund A, Grunewald J. HLA-DR Predicts the Prognosis in Scandinavian Patients With Pulmonary Sarcoidosis. *Am J Respir Crit Care Med* (1997) 156(5):1601–5. doi: 10.1164/ajrccm.156.5.9704069
 33. Grunewald J, Eklund A, Olerup O. Human Leukocyte Antigen Class I Alleles and the Disease Course in Sarcoidosis Patients. *Am J Respir Crit Care Med* (2004) 169(6):696–702. doi: 10.1164/rccm.200303-459OC
 34. Grunewald J, Grutters JC, Arkema EV, Saketkoo LA, Moller DR, Muller-Quernheim J. Sarcoidosis. *Nat Rev Dis Primers* (2019) 5(1):45. doi: 10.1038/s41572-019-0096-x
 35. Levin AM, Adrianto I, Datta I, Iannuzzi MC, Trudeau S, Li J, et al. Association of HLA-DRB1 With Sarcoidosis Susceptibility and Progression in African Americans. *Am J Respir Cell Mol Biol* (2015) 53(2):206–16. doi: 10.1165/rccm.2014-0227OC
 36. Suzuki H, Ota M, Meguro A, Katsuyama Y, Kawagoe T, Ishihara M, et al. Genetic Characterization and Susceptibility for Sarcoidosis in Japanese Patients: Risk Factors of BTNL2 Gene Polymorphisms and HLA Class II Alleles. *Invest Ophthalmol Vis Sci* (2012) 53(11):7109–15. doi: 10.1167/iovs.12-10491
 37. Wang F, Hu YL, Feng Y, Guo YB, Liu YF, Mao QS, et al. High-Level Expression of PRSS3 Correlates With Metastasis and Poor Prognosis in Patients With Gastric Cancer. *J Surg Oncol* (2019) 119(8):1108–21. doi: 10.1002/jso.25448
 38. Hayashi H, Kubo Y, Izumida M, Takahashi E, Kido H, Sato K, et al. Enterokinase Enhances Influenza A Virus Infection by Activating Trypsinogen in Human Cell Lines. *Front Cell Infect Microbiol* (2018) 8:91. doi: 10.3389/fcimb.2018.00091
 39. Schilling O, Biniossek ML, Mayer B, Elsasser B, Brandstetter H, Goettig P, et al. Specificity Profiling of Human Trypsin-Isoenzymes. *Biol Chem* (2018) 399(9):997–1007. doi: 10.1515/hsz-2018-0107
 40. Rolland-Fourcade C, Denadai-Souza A, Cirillo C, Lopez C, Jaramillo JO, Desormeaux C, et al. Epithelial Expression and Function of Trypsin-3 in Irritable Bowel Syndrome. *Gut* (2017) 66(10):1767–78. doi: 10.1136/gutjnl-2016-312094

41. Wu J, Li Z, Zeng K, Wu K, Xu D, Zhou J, et al. Key Genes Associated With Pancreatic Cancer and Their Association With Outcomes: A Bioinformatics Analysis. *Mol Med Rep* (2019) 20(2):1343–52. doi: 10.3892/mmr.2019.10321
42. Ma H, Hockla A, Mehner C, Coban M, Papo N, Radisky DC, et al. PRSS3/Mesotrypsin and Kallikrein-Related Peptidase 5 are Associated With Poor Prognosis and Contribute to Tumor Cell Invasion and Growth in Lung Adenocarcinoma. *Sci Rep* (2019) 9(1):1844. doi: 10.1038/s41598-018-38362-0
43. Qian L, Gao X, Huang H, Lu S, Cai Y, Hua Y, et al. PRSS3 is a Prognostic Marker in Invasive Ductal Carcinoma of the Breast. *Oncotarget* (2017) 8(13):21444–53. doi: 10.18632/oncotarget.15590
44. Jiang G, Cao F, Ren G, Gao D, Bhakta V, Zhang Y, et al. PRSS3 Promotes Tumour Growth and Metastasis of Human Pancreatic Cancer. *Gut* (2010) 59(11):1535–44. doi: 10.1136/gut.2009.200105
45. Ashley SL, Xia M, Murray S, O'Dwyer DN, Grant E, White ES, et al. Six-SOMAmer Index Relating to Immune, Protease and Angiogenic Functions Predicts Progression in IPF. *PLoS One* (2016) 11(8):e0159878. doi: 10.1371/journal.pone.0159878
46. Hossain MM, Crish JF, Eckert RL, Lin JJ, Jin JP. H2-Calponin is Regulated by Mechanical Tension and Modifies the Function of Actin Cytoskeleton. *J Biol Chem* (2005) 280(51):42442–53. doi: 10.1074/jbc.M509952200
47. Tazyeen S, Ahmed MM, Farooqui A, Alam A, Malik MZ, Saeed M, et al. Identification of Key Regulators in Sarcoidosis Through Multidimensional Systems Biological Approach. *Sci Rep* (2022) 12(1):1236. doi: 10.1038/s41598-022-05129-7
48. Fischer A, Grunewald J, Spagnolo P, Nebel A, Schreiber S, Muller-Quernheim J. Genetics of Sarcoidosis. *Semin Respir Crit Care Med* (2014) 35(3):296–306. doi: 10.1055/s-0034-1376860
49. Calender A, Weichhart T, Valeyre D, Pacheco Y. Current Insights in Genetics of Sarcoidosis: Functional and Clinical Impacts. *J Clin Med* (2020) 9(8):2633. doi: 10.3390/jcm9082633
50. Patterson KC, Franek BS, Muller-Quernheim J, Sperling AI, Sweiss NJ, Niewold TB. Circulating Cytokines in Sarcoidosis: Phenotype-Specific Alterations for Fibrotic and non-Fibrotic Pulmonary Disease. *Cytokine* (2013) 61(3):906–11. doi: 10.1016/j.cyt.2012.12.016
51. Yoshioka K, Sato H, Kawasaki T, Ishii D, Imamoto T, Abe M, et al. Transcriptome Analysis of Peripheral Blood Mononuclear Cells in Pulmonary Sarcoidosis. *Front Med (Lausanne)* (2022) 9:822094. doi: 10.3389/fmed.2022.822094
52. Barczyk A, Pierzchala E, Caramori G, Sozanska E. Increased Expression of CCL4/MIP-1beta in CD8+ Cells and CD4+ Cells in Sarcoidosis. *Int J Immunopathol Pharmacol* (2014) 27(2):185–93. doi: 10.1177/039463201402700205
53. Antoniou KM, Soufla G, Proklou A, Margaritopoulos G, Choulaki C, Lymbouridou R, et al. Different Activity of the Biological Axis VEGF-Flt-1 (Fms-Like Tyrosine Kinase 1) and CXCL Chemokines Between Pulmonary Sarcoidosis and Idiopathic Pulmonary Fibrosis: A Bronchoalveolar Lavage Study. *Clin Dev Immunol* (2009) 2009:537929. doi: 10.1155/2009/537929
54. Stopinsek S, Ihan A, Salobir B, Terelj M, Simcic S. Fungal Cell Wall Agents and Bacterial Lipopolysaccharide in Organic Dust as Possible Risk Factors for Pulmonary Sarcoidosis. *J Occup Med Toxicol* (2016) 11:46. doi: 10.1186/s12995-016-0135-4
55. Mortaz E, Adcock IM, Abedini A, Kiani A, Kazempour-Dizaji M, Movassaghi M, et al. The Role of Pattern Recognition Receptors in Lung Sarcoidosis. *Eur J Pharmacol* (2017) 808:44–8. doi: 10.1016/j.ejphar.2017.01.020
56. Jeny F, Bernaudin JF, Valeyre D, Kambouchner M, Pretolani M, Nunes H, et al. Hypoxia Promotes a Mixed Inflammatory-Fibrotic Macrophages Phenotype in Active Sarcoidosis. *Front Immunol* (2021) 12:719009. doi: 10.3389/fimmu.2021.719009
57. Rubio-Rivas M, Franco J, Corbella X. Sarcoidosis Presenting With and Without Lofgren's Syndrome: Clinical, Radiological and Behavioral Differences Observed in a Group of 691 patients. *Joint Bone Spine* (2020) 87(2):141–7. doi: 10.1016/j.jbspin.2019.10.001

Conflict of Interest: Author MZ was employed by ANNOROAD 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.

Copyright © 2022 Zhang, Huang, Zhang, Fang, Wang, Jing, Guo, Sun, Yang and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Evaluating the Microsatellite Instability of Colorectal Cancer Based on Multimodal Deep Learning Integrating Histopathological and Molecular Data

Wenjing Qiu^{1,2†}, Jiasheng Yang^{1†}, Bing Wang¹, Min Yang^{1,2}, Geng Tian^{2,3}, Peizhen Wang^{1*} and Jialiang Yang^{2,3*}

OPEN ACCESS

Edited by:

Liang Cheng,
Harbin Medical University, China

Reviewed by:

Guangzhou Xiong,
Huazhong University of Science and
Technology, China
Man Liu,
Beijing Jiaotong University, China

*Correspondence:

Peizhen Wang
pzhwang@ahut.edu.cn
Jialiang Yang
yangjl@geneis.cn

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 21 April 2022

Accepted: 30 May 2022

Published: 05 July 2022

Citation:

Qiu W, Yang J, Wang B, Yang M,
Tian G, Wang P and Yang J (2022)
Evaluating the Microsatellite Instability
of Colorectal Cancer Based on
Multimodal Deep Learning Integrating
Histopathological and Molecular Data.
Front. Oncol. 12:925079.
doi: 10.3389/fonc.2022.925079

¹ School of Electrical and Information Engineering, Anhui University of Technology, Maanshan, China, ² Science System Department, Geneis Beijing Co., Ltd., Beijing, China, ³ Qingdao Genesis Institute of Big Data Mining and Precision Medicine, Qingdao, China

Microsatellite instability (MSI), an important biomarker for immunotherapy and the diagnosis of Lynch syndrome, refers to the change of microsatellite (MS) sequence length caused by insertion or deletion during DNA replication. However, traditional wet-lab experiment-based MSI detection is time-consuming and relies on experimental conditions. In addition, a comprehensive study on the associations between MSI status and various molecules like mRNA and miRNA has not been performed. In this study, we first studied the association between MSI status and several molecules including mRNA, miRNA, lncRNA, DNA methylation, and copy number variation (CNV) using colorectal cancer data from The Cancer Genome Atlas (TCGA). Then, we developed a novel deep learning framework to predict MSI status based solely on hematoxylin and eosin (H&E) staining images, and combined the H&E image with the above-mentioned molecules by multimodal compact bilinear pooling. Our results showed that there were significant differences in mRNA, miRNA, and lncRNA between the high microsatellite instability (MSI-H) patient group and the low microsatellite instability or microsatellite stability (MSI-L/MSS) patient group. By using the H&E image alone, one can predict MSI status with an acceptable prediction area under the curve (AUC) of 0.809 in 5-fold cross-validation. The fusion models integrating H&E image with a single type of molecule have higher prediction accuracies than that using H&E image alone, with the highest AUC of 0.952 achieved when combining H&E image with DNA methylation data. However, prediction accuracy will decrease when combining H&E image with all types of molecular data. In conclusion, combining H&E image with deep learning can predict the MSI status of colorectal cancer, the accuracy of which can further be improved by integrating appropriate molecular data. This study may have clinical significance in practice.

Keywords: microsatellite instability, H&E images, multi-omics data, multimodal deep learning, compact bilinear pooling

1 INTRODUCTION

Colorectal cancer (CRC) is a common digestive tract malignancy. CRC is the third largest cancer in the world, and the second leading cause of cancer-related death; the incidence rate and mortality rate of CRC were third and fifth, respectively, among all cancers in China, with more than 250,000 new patients and 140,000 deaths annually (1–3). Sporadic colorectal cancer (SCRC) accounts for about 85%, and hereditary nonpolyposis colorectal cancer (HNPCC) accounts for about 10%–15% of all CRC patients (4). SCRC is mainly affected by environment, diet, living habits, and chronic inflammation, which leads to the mutations of the “administrator gene” and “guard gene”; the mutations disrupt the mechanisms for inhibiting cell growth, promoting cell death, and maintaining cell stability. Among them, microsatellite instability (MSI) is involved in the occurrence of SCRC, with an incidence of 12%–15% (5). The value of MSI in the diagnosis, treatment response, and prognosis of CRC has attracted global attention (6–8).

MSI refers to the change in the length of normal microsatellites caused by the deletion or insertion of repeated bases compared with normal tissue cells (9). In 2001, Fukushima and Takenoshita (10) found that MSI significantly increased the random mutation rate of genes, especially the mutation of tumor-related genes, which is an important mechanism of tumorigenesis.

There is some evidence to support the use of pre-diagnostic MSI in clinical decision-making. First, MSI detection is recommended for the diagnosis of Lynch syndrome. Lynch syndrome is the most common hereditary colon cancer syndrome, which is associated with germline mutations in the MMR gene (MLH1, MSH2, MSH6, or PMS2) (11). MSI status helps to identify families with the syndrome. Second, MSI is one of the key factors affecting the prognosis of CRC, especially in early cases (12, 13). In general, patients with stage II CRC with high MSI (MSI-H)/MMR deficiency (d MMR) have a better prognosis than patients with microsatellite stability (MSS) and low MSI (MSI-L)/MMR (p MMR) (13). Third, MSI status can be used to evaluate therapeutic response, including fluoropyrimidine-based chemotherapy (14) and immunotherapy (15). Fluoropyrimidine (5-FU or capecitabine) is the pillar of the CRC chemotherapy strategy. It plays an important role not only in neoadjuvant therapy but also in prognosis treatment (16, 17). However, patients with MSI-H status are usually resistant to 5-FU-based chemotherapy (18). Immunotherapy is an emerging and promising treatment for CRC because MSI-H tumors have a large number of mutant neoantigens, which makes them sensitive to immune checkpoint inhibitors (19). Therefore, MSI status is crucial for selecting CRC treatment and evaluating the response to treatment (20).

In recent years, the deep learning method has become a newly developing method, which has shown excellent performance in the fields of computer vision (21, 22), speech recognition (23), and bioinformatics (24–27). Deep learning technology has the characteristics of end-to-end training, and can also represent abstract concepts or patterns level by level through deep neural networks (28). At the same time, researchers use the technology of transfer learning to transfer the network model pre-trained by Image Net to the classification task of pathological image segmentation by fine-tuning the classifier layer of convolutional neural network. In

the 2016 CAMELYON breast cancer lymph node metastasis challenge, 25 of the 32 algorithms submitted by the contestants used convolution neural networks (CNNs) (29) including VGG-16 (30), GoogLeNet (31), and other well-known models such as (32). Xu et al. used pre-trained AlexNet to extract the features of brain tumor pathological image blocks and achieved 97.5% classification accuracy on the small-sample MICCAI 2014 brain tumor digital pathology challenge dataset. Yang et al. proposed a multimodal deep learning method to predict the recurrence and metastasis risk of Her2-positive breast cancer by integrating pathological image with clinical information (33). Ye et al. developed a deep convolution network to evaluate prognosis of cervical cancer (34). Ke et al. (35) used the knowledge distillation model of multistage CNN to classify MSI-H and MSS, and obtained an AUC = 0.802; Kather et al. (36) used ResNet18 to predict the histopathological sections of CRC, and the AUC obtained by MSI was 0.84.

With the increasing availability of high-throughput genomic and transcriptional data, there are several molecular biomarkers in The Cancer Genome Atlas (TCGA), including somatic mutation, copy number variation, gene expression, microRNA expression, and DNA methylation, which were used to track cancer (37–39) and predict cancer recurrence and metastasis (40). Hayes identified relevant microRNA and mRNA features that predict high-risk and low-risk patients with glioblastoma (GBM). Sun et al. integrated gene expression profile, CNA spectrum, and clinical data to predict the prognosis of breast cancer, achieving a good performance of AUC = 0.843.

Based on the feasibility of cancer prediction and multimodal fusion from the pathological image level, our goal was to compare these unimodal data and combinations to predict the MSI ability of CRC in a unified context and to explore whether multimodal data fusion can significantly improve prediction accuracy compared with single-mode data.

2 MATERIALS AND METHODS

2.1 Data Description

We overlapped the H&E images data and omics data to obtain 353 sample sizes, of which 63 were labeled MSI-Hs, which were marked as 1; 290 cases were labeled MSSs, which were labeled as 0.

Pathological image. We used the method of Kather et al. to publish the CRC with hematoxylin and eosin stabilized (CRC-HE) dataset, including 100,000 pieces of 224×224 pixel H&E-stained pathological images that were divided into blocks; each pixel in the block corresponds to $0.5 \mu\text{m} \times 0.5 \mu\text{m}$ organization. To eliminate the color difference of slices from different data sources in the process of production and scanning, all H&E images have been dyed and standardized according to the method of Macenko et al. (41).

Multi-omics data. Multi-omics data of CRC were downloaded from the TCGA database, including messenger RNA (mRNA), microRNA expression (miRNA), long non-coding RNA (lncRNA), DNA methylation (Met), and gene copy number variation (CNV). Their forms include Counts and FPKM. The difference between FPKM and Counts is that Counts is the original expression quantity that is not processed in

the data background, although FPKM and Counts are data processing methods. In the analysis of this paper, the difference analysis part adopts the form of Counts, and the modeling analysis part adopts the form of FPKM. **Table 1** shows the characteristic dimensions of each omics data.

2.2 Feature Extraction

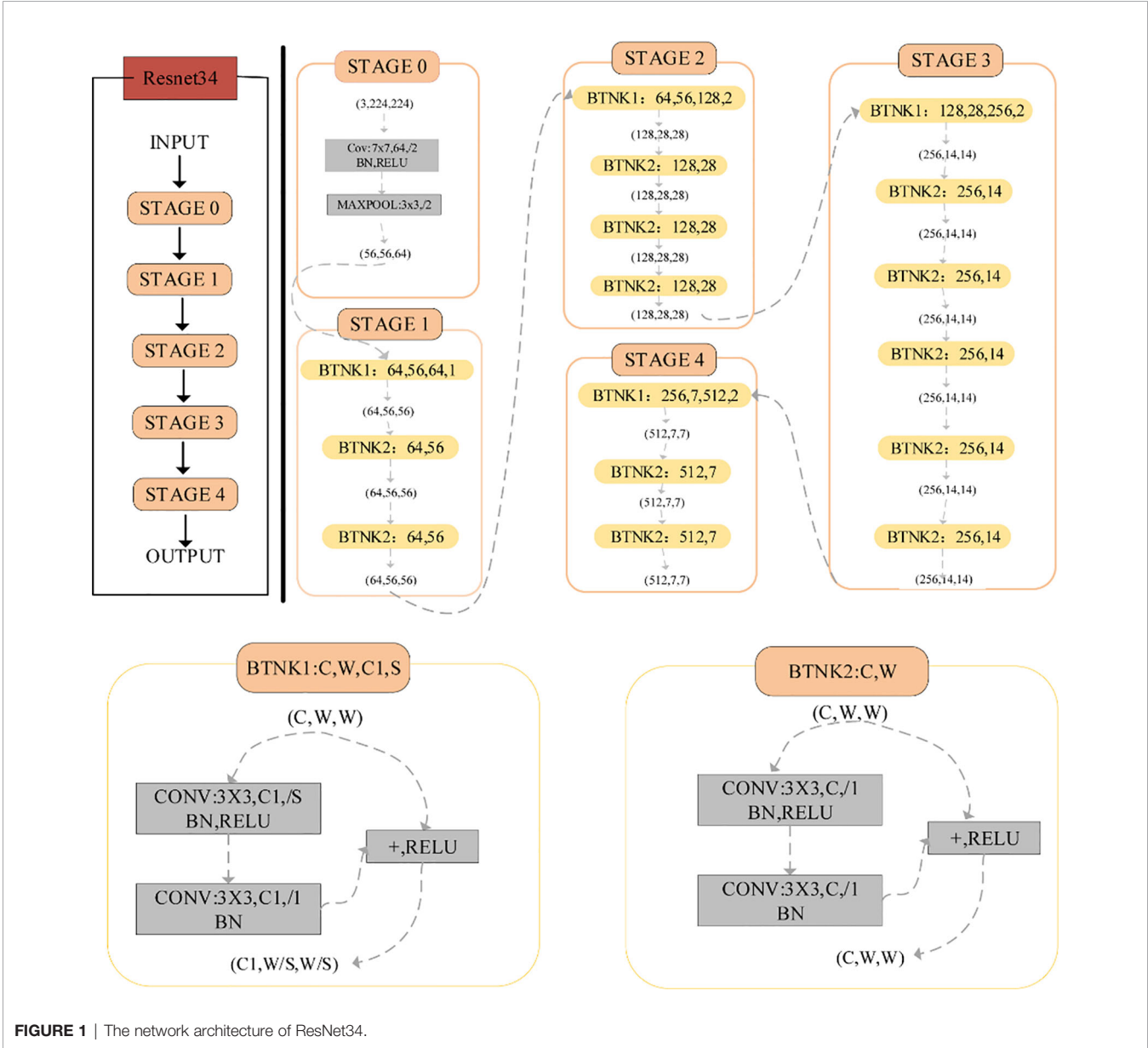
2.2.1 H&E Image Feature Representation
Based on ResNet34

CNN is the latest algorithm for image recognition and classification because of its stable learning performance (42). CNN includes an input layer, a middle hidden layer, and an output layer. The middle-hidden layer is composed of multiple convolution layers, pooling layers, and full connection layers. CNN can be optimized through error backpropagation and

TABLE 1 | The properties of the dataset.

Data Category	Abbreviation	Number of features
Messenger RNA	mRNA	19,531
MicroRNAs	miRNA	1,881
Long non-coding RNA	lncRNA	7,308
DNA methylation	Met	27,578
Copy number variation	CNV	60,483

gradient descent algorithm. However, after reaching a certain depth, increasing the number of layers of CNN cannot further improve the classification performance. Due to the vanishing gradient problem, the network convergence speed is slow and the classification accuracy is negative. ResNet is used to solve this problem. The difference between residual network and ordinary network is that jump connection is introduced, which can help



the information of the previous residual block enter the next block stream unimpeded, improve the information flow, and avoid the problem of vanishing gradient and the degradation caused by the great depth of the network.

ResNet is a large-scale CNN constructed from residual blocks. We used ResNet34 (Figure 1) to extract H&E image features. The architecture of ResNet34 is divided into four stages. Every Resnet architecture performed the initial convolution and max-pooling using 7 x 7 and 3 x 3 kernel sizes, respectively. The residual structure of BTNK1 can reduce the dimension, and the dimension is reduced by a 1 x 1 convolution kernel on the shortcut branch. It is worth noting that in Stage 2, Stage 3, and Stage 4, it is executed with stride 2; therefore, the size of the input will be halved in height and width, but the channel width will be doubled. When the image advances from one stage to another, the channel width will be doubled and the input size will be reduced by half. Finally, the network has an average pool layer, followed by a full connection layer containing 1,000 neurons.

2.2.2 Feature Extraction of Multi-Omics Data

A common problem with high-throughput sequencing datasets is the so-called “Curse of dimensionality” (40). Variable selection is very important for interpretation and prediction, especially for high-dimensional datasets. In this work, we used the characteristic importance attribute of Random forest (Gini-index) (43) to deal with high-dimensional variables in omics data. Features with Gini-index greater than or equal to 0.005 were the most important features. Then, the multimodal data are simply spliced from the important features obtained from the single group data. Then, select according to the feature importance of random forest, and the feature with a Gini-index greater than 0.005 is regarded as the most important feature.

2.3 Feature Fusion

The most common fusion methods are concatenation, element-wise product, and element-wise sum. These simple operations are not as effective as the outer product, and complex relationships can be established between the two modes. However, the complexity of outer product calculation is too

high. The n -dimensional vector calculated the outer product to obtain the n^2 -dimensional vector. In this work, our fusion method was the multimodal compact bilinear (MCB) model. MCB maps the result of the outer product to low-dimensional space without explicit calculation of the outer product.

2.4 Screening of Differentially Expressed Genes

The R package “Deseq2” was used to identify differentially expressed genes (DEGs) in mRNA, miRNA, and lncRNA gene expression profiles. Genes with an adjusted p -value < 0.1 and a \log_2 foldchange (LFC) > 0 were classified as upregulated genes, whereas those with an adjusted p -value < 0.1 and an LFC < 0 were classified as downregulated genes. Taking $|\log_2(\text{foldchange})| \geq 1$ and the corrected p -value < 0.05 as the threshold, the genes with significant differences were selected. The R-Pack “heat map” shows significantly different genes. The R-Pack “cluster analyzer” is used for Gene Ontology (GO) enrichment analysis and calculation. R-Pack ggplot2 is used to generate enrichment pathways in significantly different genes.

2.5 Evaluation Metrics

Fivefold cross validation (5-f cv) is used to evaluate the accuracy of the algorithm. 5-k cv: Divide the dataset into five equally, and take turns using four of them as training data and one as test data. The performance of the classification algorithm is estimated by averaging 5 test sets. For binary classification, the area under the subject operating characteristic curve (AUC), Accuracy (Acc), Precision, Recall, and F1_score are used to evaluate the performance of the model.

3 RESULTS

3.1 The Overall Framework of This Study

In this work, we studied the data in two parts. In the first part, the differences of mRNA, miRNA, and lncRNA were analyzed. In the second part, in the modeling analysis, we conducted two experiments (Figure 2). First, only the H&E image data were used to build the model and predict the classification (Figure 2A). Second,

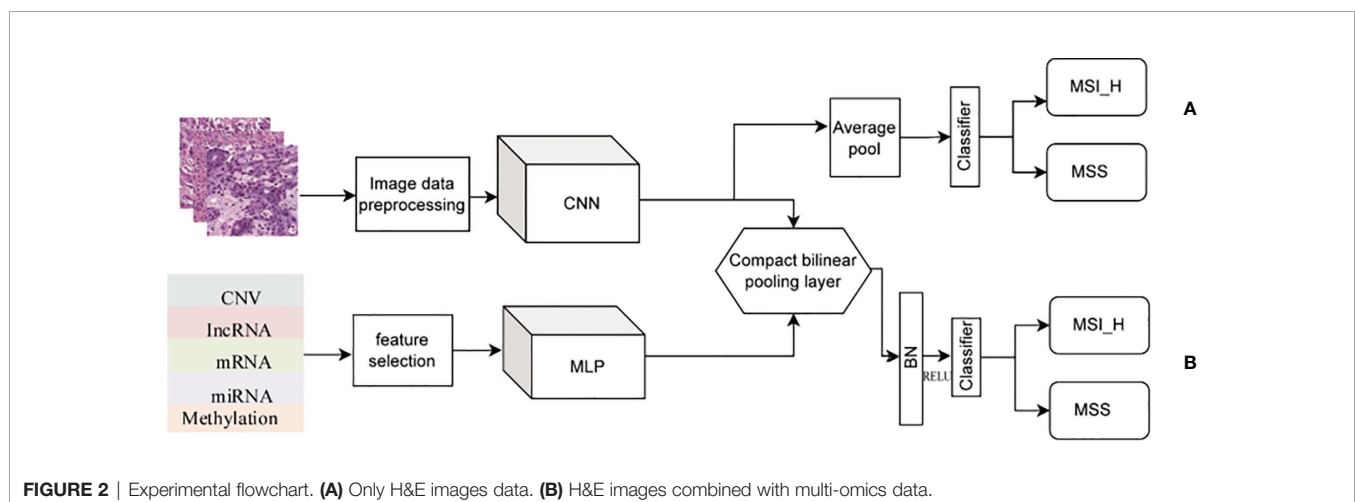


FIGURE 2 | Experimental flowchart. (A) Only H&E images data. (B) H&E images combined with multi-omics data.

the H&E image was combined with omics for prediction and classification (**Figure 2B**), including H&E image combined with single omics data and H&E images combined with multi-omics data.

3.2 mRNA, lncRNA, and miRNAs Differ Significantly Between MSI-H and MSI-L/MSS Groups

We comprehensively analyzed the differential expression of mRNA, lncRNA, and miRNA between MSI-L/MSI-H and MSS

groups. In the lncRNA group, we obtained 1,130 upregulated expressions and 631 downregulated expressions. A total of 172 upregulated expressions and 125 downregulated expressions were obtained in miRNA. In the mRNA group, 5,210 upregulated genes and 5,466 downregulated genes were obtained. After strictly restricting the adjusted *p*-value, we obtained 663 significantly differentially expressed lncRNAs, 61 significantly differentially expressed miRNAs, and 1,898 significantly different mRNA genes (see **Supplementary**

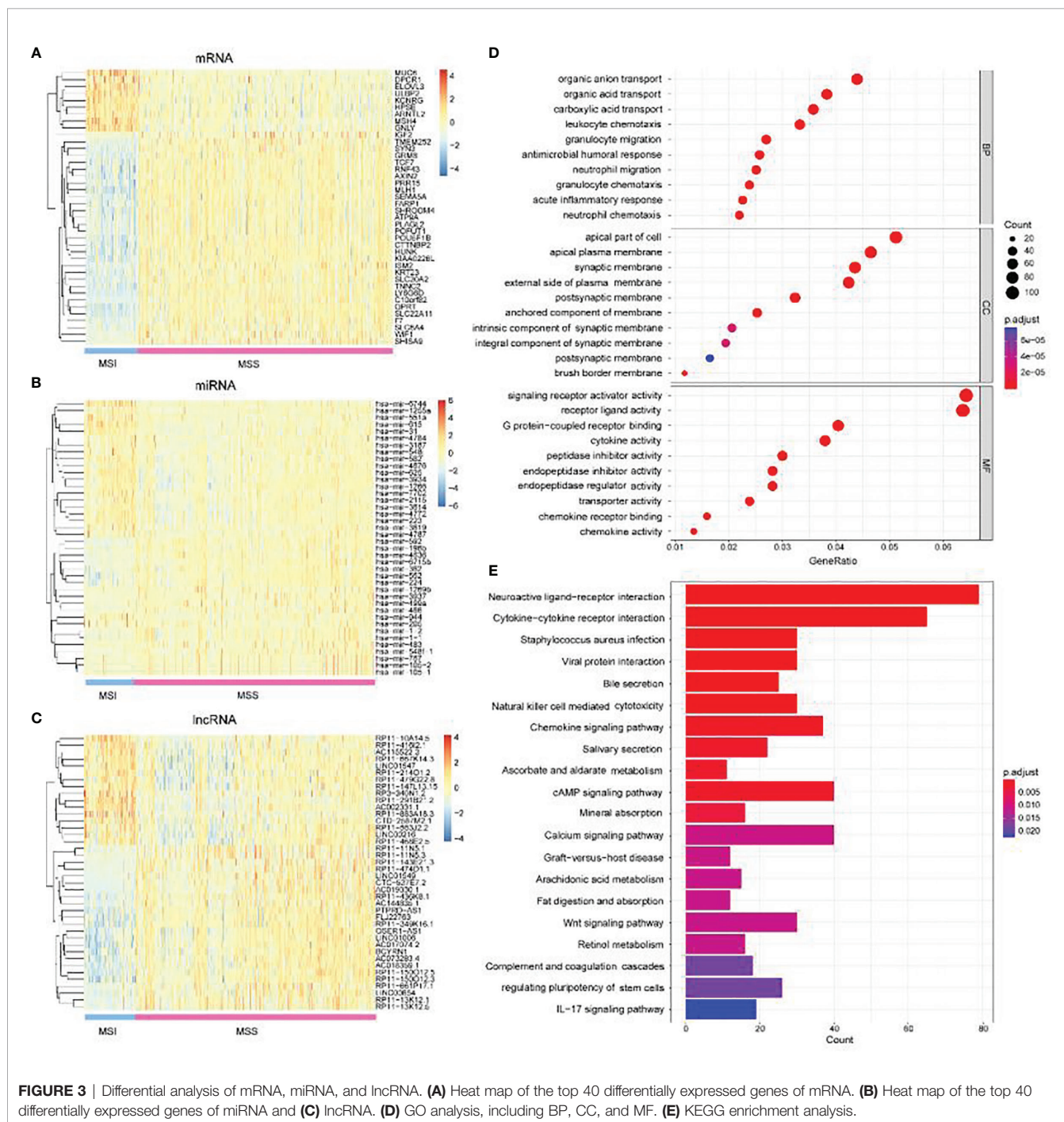


FIGURE 3 | Differential analysis of mRNA, miRNA, and lncRNA. **(A)** Heat map of the top 40 differentially expressed genes of mRNA. **(B)** Heat map of the top 40 differentially expressed genes of miRNA and **(C)** lncRNA. **(D)** GO analysis, including BP, CC, and MF. **(E)** KEGG enrichment analysis.

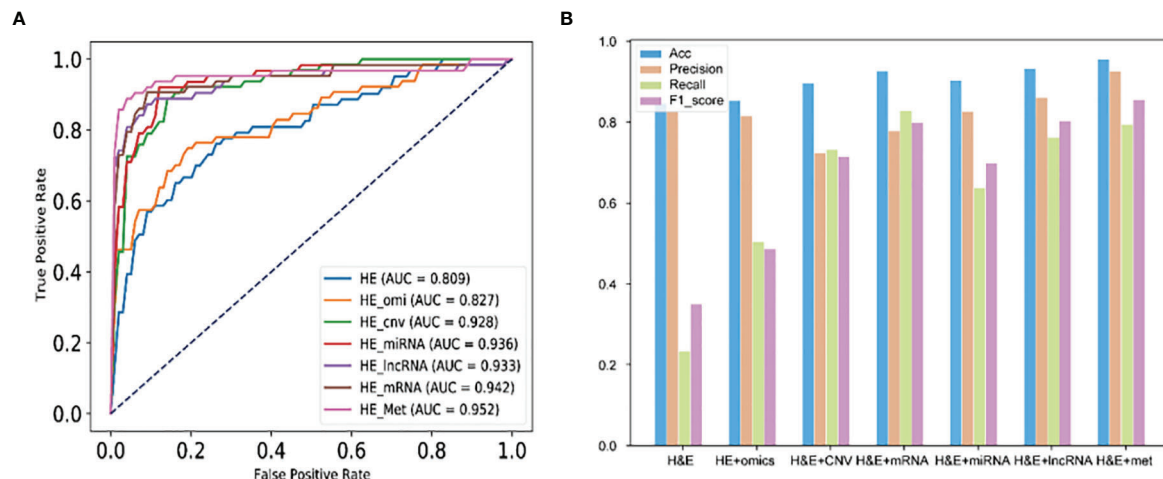


FIGURE 4 | Performance of H&E images and images combined with omics data. **(A)** The AUC score of image and image combined with omics data. **(B)** Performance of each mode in Accuracy, Precision, Recall, and F1_score index. HE_omi: H&E image features combined with multi-omics features.

Tables 1–3). As shown in **Figures 3A–C**, we used the first 40 significant difference expressions to draw the heat map.

GO analysis was used to annotate the function of DEGs between MSI-H and MSI-L/MSS. In the biological process (BP) category, genes with significant differences were mainly enriched in organic acid, organic anion, and carboxylic acid transport. For cell component (CC) categories, genes with significant differences were mainly clustered in the apical part of the cell. In the binding molecular function (MF), significantly different genes were mainly involved in signaling receptor activator activity and receptor–ligand activity (**Figure 3D**). Further KEGG enrichment analysis was carried out to explore the potential pathological pathway of cancer. As shown in **Figure 3E**, the first two significant enrichment pathways were neuroactive ligand–receptor interaction and cytokine receptor interaction. Our significantly different genes were involved in these pathways, which may also contribute to the diagnosis of cancer. For example, the *MUC6* gene is one of the mucin genes that make up the gastric mucosa, and its expression is downregulated in precancerous lesions and gastric cancer tissues (44). *Dpcr1* *DPCR1*(*Muc3MUC3*) is a protein-coding gene. *MUC3* may regulate NF kappa B signaling and play a role in cell growth.

3.3 H&E Images Combined With DNA Methylation Performed Best in Predicting MSI of Colorectal Cancer

We evaluated the performance of images combined with omics data in predicting the MSI of CRC. 5-f cv was used to train ResNet34. As shown in **Figure 4A**, the prediction result of H&E images combined with DNA methylation (ROC = 0.952) was higher than that of H&E images, H&E images combined with multi-omics, and image combined with other omics data. In addition to H&E images combined with

methylation, H&E images combined with other omics was lower than the prediction result of image in precision index. In Acc, Recall, and F1_score index, the prediction results of image combined with omics were higher than those of image (**Figure 4B**).

4 DISCUSSION

As we all know, MSI is widely considered as an indicator of prediction and prognosis. It has been well studied in several types of human cancers. In CRC, about 15% to 20% of CRC cases are found to be associated with MSI-H. Therefore, MSI states that detection is particularly important for CRC and is recommended by current clinical guidelines (6, 45). With the continuous development of computer deep learning technology, computer-aided diagnosis and prognosis prediction based on H&E staining images has attracted more and more attention because of its advantages of high speed, low cost, and no trauma. Multimodal fusion is a typical interdisciplinary field and has gradually become a research hotspot. In many studies, some results have been achieved (46–48). In conclusion, the accuracy of the image-based prognosis prediction model needs to be further improved.

In this study, we systematically analyzed the differences in mRNA, lncRNA, and miRNA omics data between MSI-H and MSI-L/MSS groups, and compared the classification performance of image and image data combined with omics data to predict the MSI of CRC. In this experiment, by comparing the results of ROC, we found that H&E image combined with Met had the best performance in predicting the MSI of CRC. The result of H&E image combined with all omics data was lower than that of image combined with single omics data and higher than that of H&E images.

Our study has some limitations. First, the selected omics data were the cancer sample construction and evaluation model, not the adjacent data. Only the differences between MSI-L/MSS and MSI-H in cancer samples were studied. Second, we do not have independent datasets for validation, because we cannot find other databases to provide the required data except for the TCGA database. Finally, our multi-omics feature was just simple splicing of different single omics. It is best to test the effects of interactions between omics because the genes of each omics are not completely independent. Therefore, in our follow-up study, we will try to include para-cancerous samples, including independent test samples, and add interactive items and new classification models to improve the prediction accuracy.

5 CONCLUSION

To sum up, we integrated molecular biological information and images to classify and predict the MSI of CRC. This is the first study to compare the ability of different modes in predicting the MSI of CRC under the same conditions, including the same dataset, the same preprocessing scheme, and the same classification algorithm. There were significant differences in mRNA, lncRNA, and miRNA omics data between MSI-H and MSI-L/MSS groups. By comparing the results of ROC, we found that H&E images combined with Met had the best performance in predicting the MSI of CRC. The result of image combined with all omics data was lower than that of image combined with single omics data and higher than that of H&E images.

REFERENCES

- Chen W, Zheng R, Zheng S, Ceng H, Zuo T, Jia M. Analysis of Malignant Tumor Incidence and Death in China in 2012. *China Cancer* (2016) 1(8). doi: 10.11735/j.issn.1004-0242.2015.01.A001
- Liu H, Qiu C, Wang B, Bing P, Tian G, Zhang X, et al. Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor Tissue-Of-Origin. *Front Cell Dev Biol* (2021) 9:619330. doi: 10.3389/fcell.2021.619330
- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660
- Romanowicz-Makowska H, Smolarz B, Langner E, Kozłowska E, Kulig A, Dzik A. Analysis of Microsatellite Instability and BRCA1 Mutations in Patients From Hereditary Nonpolyposis Colorectal Cancer (HNPCC) Family. *Pol J Pathol* (2005) 56(1):21–6.
- Pancione M, Remo A, Colantuoni V. Genetic and Epigenetic Events Generate Multiple Pathways in Colorectal Cancer Progression. *Patholog Res Int* (2012) 2012:509348. doi: 10.1155/2012/509348
- Kawakami H, Zaanani A, Sinicrope FA. Microsatellite Instability Testing and its Role in the Management of Colorectal Cancer. *Curr Treat Options Oncol* (2015) 16(7):30. doi: 10.1007/s11864-015-0348-2
- Liu W, Zhang D, Tan SA, Liu X, Lai J. Sigmoid Colon Adenocarcinoma With Isolated Loss of PMS2 Presenting in a Patient With Synchronous Prostate Cancer With Intact MMR: Diagnosis and Analysis of the Family Pedigree. *Anticancer Res* (2018) 38(8):4847–52. doi: 10.21873/anticancer.12796
- Evrard C, Tachon G, Randrian V, Karayan-Tapon L, Tougeron D. Microsatellite Instability: Diagnosis, Heterogeneity, Discordance, and

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/>.

AUTHOR CONTRIBUTIONS

JLY and PW designed the study. WQ, JSY, BW, MY, and GT performed the study, analyzed the data, and interpreted data. WQ and JLY wrote the manuscript. JSY, BW, MY, GT, and PW reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Natural Science Foundation of China (numbers 51574004 and 62172004), the Natural Science Foundation of the Higher Education Institutions of Anhui Province, China (KJ2019A0085), the Academic Foundation for Top Talents of the Higher Education Institutions of Anhui Province (gxbjZD2016041), and the Educational Commission of Anhui Province (KJ2019ZD05).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.925079/full#supplementary-material>

- Clinical Impact in Colorectal Cancer. *Cancers (Basel)* (2019) 11(10):1567. doi: 10.3390/cancers11101567
- Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M. Ubiquitous Somatic Mutations in Simple Repeated Sequences Reveal a New Mechanism for Colonic Carcinogenesis. *Nature* (1993) 363(6429):558–61. doi: 10.1038/363558a0
- Fukushima T, Takenoshita S. Colorectal Carcinogenesis. *Fukushima J Med Sci* (2001) 47(1):1–11. doi: 10.5387/fms.47.1
- Vasen HF. Clinical Description of the Lynch Syndrome [Hereditary Nonpolyposis Colorectal Cancer (HNPCC)]. *Fam Cancer* (2005) 4(3):219–25. doi: 10.1007/s10689-004-3906-5
- Popat S, Hubner R, Houlston RS. Systematic Review of Microsatellite Instability and Colorectal Cancer Prognosis. *J Clin Oncol* (2005) 23(3):609–18. doi: 10.1200/jco.2005.01.086
- Merok MA, Ahlquist T, Røyrvik EC, Tufteland KF, Hektoen M, Sjø OH, et al. Microsatellite Instability has a Positive Prognostic Impact on Stage II Colorectal Cancer After Complete Resection: Results From a Large, Consecutive Norwegian Series. *Ann Oncol* (2013) 24(5):1274–82. doi: 10.1093/annonc/mts614
- Li LS, Morales JC, Veigl M, Sedwick D, Greer S, Meyers M, et al. DNA Mismatch Repair (MMR)-Dependent 5-Fluorouracil Cytotoxicity and the Potential for New Therapeutic Targets. *Br J Pharmacol* (2009) 158(3):679–92. doi: 10.1111/j.1476-5381.2009.00423.x
- Mandal R, Samstein RM, Lee KW, Havel JJ, Wang H, Krishna C, et al. Genetic Diversity of Tumors With Mismatch Repair Deficiency Influences Anti-PD-1 Immunotherapy Response. *Science* (2019) 364(6439):485–91. doi: 10.1126/science.aau0447
- Ludmir EB, Palta M, Willett CG, Cizto BG. Total Neoadjuvant Therapy for Rectal Cancer: An Emerging Option. *Cancer* (2017) 123(9):1497–506. doi: 10.1002/cnrc.30600

17. Tomasello G, Petrelli F, Ghidini M, Russo A, Passalacqua R, Barni S. FOLFOXIRI Plus Bevacizumab as Conversion Therapy for Patients With Initially Unresectable Metastatic Colorectal Cancer: A Systematic Review and Pooled Analysis. *JAMA Oncol* (2017) 3(7):e170278. doi: 10.1001/jamaoncol.2017.0278
18. Fischer F, Baerenfaller K, Jiricny J. 5-Fluorouracil is Efficiently Removed From DNA by the Base Excision and Mismatch Repair Systems. *Gastroenterology* (2007) 133(6):1858–68. doi: 10.1053/j.gastro.2007.09.003
19. Chalabi M, Fanchi LF, Dijkstra KK, Van den Berg JG, Aalbers AG, Sikorska K, et al. Neoadjuvant Immunotherapy Leads to Pathological Responses in MMR-Proficient and MMR-Deficient Early-Stage Colon Cancers. *Nat Med* (2020) 26(4):566–76. doi: 10.1038/s41591-020-0805-8
20. Diagnosis, and Treatment Guidelines for Colorectal Cancer Working Group, C. Chinese Society of Clinical Oncology (CSCO) Diagnosis and Treatment Guidelines for Colorectal Cancer 2018 (English Version). *Chin J Cancer Res* (2019) 31(1):117–34. doi: 10.21147/j.issn.1000-9604.2019.01.07
21. Cireşan D, Meier U, Schmidhuber J. "Multi-Column Deep Neural Networks for Image Classification". In: *Computer Vision & Pattern Recognition*.
22. Wu M, Li C. "Image Recognition Based on Deep Learning". 2015 *Chinese Automation Congress (CAC)* IEEE (2015) pp: 542–6. doi: 10.1109/CAC.2015.7382560.
23. Kaker E, Heittola T, Huttunen H, Virtanen T. "Polyphonic Sound Event Detection Using Multi Label Deep Neural Networks". In: 2015 *International Joint Conference on Neural Networks (IJCNN)* IEEE pp: 3642–49. doi: 10.1109/CVPR.2012.6248110.
24. Quang D, Chen Y, Xie X. DANN: A Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants. *Bioinformatics* (2015) 5(7):761–3. doi: 10.1093/bioinformatics/btu703
25. Yifei C, Yi L, Rajiv N, Aravind S, Xiaohui X. Gene Expression Inference With Deep Learning. *Bioinf (Oxford England)* (2016) 32(12):1832–9. doi: 10.1093/bioinformatics/btw074.
26. Xu J, Cai L, Liao B, Zhu W, Yang J. CMF-Impute: An Accurate Imputation Tool for Single-Cell RNA-Seq Data. *Bioinformatics* (2020) 36(10):3139–47. doi: 10.1093/bioinformatics/btaa109
27. Meng Y, Lu C, Jin M, Xu J, Zeng X, Yang J. A Weighted Bilinear Neural Collaborative Filtering Approach for Drug Repositioning. *Brief Bioinform* (2022) 23(2):bbab581. doi: 10.1093/bib/bbab581
28. Zhang A, Lipton ZC, Li M, Smola AJ. *Dive Into Deep Learning*. Beijing: Posts and Telecommunications Press (2021).
29. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *Jama* (2017) 318(22):2199–210. doi: 10.1001/jama.2017.14585
30. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Comput Sci* (2014) arXiv:1409.1556.
31. Szegedy C, Liu W, Jia Y, Sermanet P, Rabinovich A. Going Deeper With Convolutions. *IEEE Comput Society* (2014) pp:1–9. doi: 10.1109/CVPR.2015.7298594
32. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *IEEE* (2016) pp: 770–778. doi: 10.1109/CVPR.2016.90
33. Yang J, Ju J, Guo L, Ji B, Shi S, Yang Z, et al. Prediction of HER2-Positive Breast Cancer Recurrence and Metastasis Risk From Histopathological Images and Clinical Information via Multimodal Deep Learning. *Comput Struct Biotechnol J* (2022) 20:333–42. doi: 10.1016/j.csbj.2021.12.028
34. Ye Z, Zhang Y, Liang Y, Lang J, Zhang X, Zang G, et al. Cervical Cancer Metastasis and Recurrence Risk Prediction Based on Deep Convolutional Neural Network. *Curr Bioinf* (2022) 17(2):164–73. doi: 10.2174/1574893616666210708143556
35. Ke J, Shen Y, Wright JD, Jing N, Shen D. "Identifying Patch-Level MSI From Histological Images of Colorectal Cancer by a Knowledge Distillation Model". In: *IEEE* 2020 1043-6. doi: 10.1109/BIBM49941.
36. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep Learning can Predict Microsatellite Instability Directly From Histology in Gastrointestinal Cancer. *Nat Med* (2019) 25(7):1054–6. doi: 10.1038/s41591-019-0462-y
37. Li Y, Kang K, Krahn JM, Croutwater N, Lee K, Umbach DM, et al. A Comprehensive Genomic Pan-Cancer Classification Using The Cancer Genome Atlas Gene Expression Data. *BMC Genomics* (2017) 18(1):508. doi: 10.1186/s12864-017-3906-0
38. Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor Origin Detection With Tissue-Specific miRNA and DNA Methylation Markers. *Bioinformatics* (2018) 34(3):398–406. doi: 10.1093/bioinformatics/btx622
39. He B, Lang J, Wang B, Liu X, Lu Q, He J, et al. TOOme: A Novel Computational Framework to Infer Cancer Tissue-Of-Origin by Integrating Both Gene Mutation and Expression. *Front Bioeng Biotechnol* (2020) 8:394. doi: 10.3389/fbioe.2020.00394
40. Sun D, Wang M, Li A. A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data. *IEEE/ACM Transactions on Computational Biology & Bioinformatics* (2018). 16(3):841–50. doi: 10.1109/TCBB.2018.2806438.
41. Macenko M, Niethammer M, Marron JS, Borland D, Thomas NE. "A Method for Normalizing Histology Slides for Quantitative Analysis", in: *Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE (2009) pp: 1107–10. doi: 10.1109/ISBI.2009.5193250.
42. Egmont-Petersen M, Ridder DD, Handels H. Image Processing With Neural Networks—A Review. *Pattern Recognition* (2002) 35(10):2279–301. doi: 10.1016/S0031-3203(01)00178-9
43. Prusa JD, Khoshgoftaar TM, Napolitano A. Using Feature Selection in Combination with Ensemble Learning Techniques to Improve Tweet Sentiment Classification Performance. 2015 *IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)* IEEE (2015) pp: 186–93. doi: 10.1109/ICTAI.2015.39.
44. Wang R, Fang D, Liu W, Men R. Expression of MUC6 Apomucin in the Tissues of Precancerous Lesion and Gastric Carcinoma and its Significance. *J Third Military Med Univ* (2001) 23(1):3. doi: 10.3321/j.issn:1000-5404.2001.01.004
45. Eso Y, Shimizu T, Takeda H, Takai A, Marusawa H. Microsatellite Instability and Immune Checkpoint Inhibitors: Toward Precision Medicine Against Gastrointestinal and Hepatobiliary Cancers. *J Gastroenterol* (2020) 55(1):15–26. doi: 10.1007/s00535-019-01620-7
46. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting Cancer Outcomes From Histology and Genomics Using Convolutional Networks. *Proc Natl Acad Sci* (2018) 115(13):E2970–e2979. doi: 10.1073/pnas.1717139115
47. Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, Lindeman NI, et al. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE Trans Med Imaging* (2020) pp: 99:1–1. doi: 10.1109/tmi.2020.3021387
48. Subramanian V, Syeda-Mahmood T, Do MN. *Multimodal Fusion Using Sparse CCA for Breast Cancer Survival Prediction 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE (2021) pp: 1429–32. doi: 10.1109/ISBI48211.2021.9434033

Conflict of Interest: Authors WQ, JLY, GT, and MY were employed by Geneis Beijing Co., Ltd., Beijing.

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.

Copyright © 2022 Qiu, Yang, Wang, Yang, Tian, Wang and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Computational Characterizing Necroptosis Reveals Implications for Immune Infiltration and Immunotherapy of Hepatocellular Carcinoma

OPEN ACCESS

Edited by:

Liang Cheng,
Harbin Medical University, China

Reviewed by:

Leyi Wei,
Shandong University, China
Mengyuan Yang,
Zhengzhou University, China

*Correspondence:

Lei Liu
liulei84207@tmmu.edu.cn
Yungang Xu
yungang.xu@uth.tmc.edu

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

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 30 April 2022

Accepted: 06 June 2022

Published: 07 July 2022

Citation:

Zhu J, Han T, Zhao S, Zhu Y, Ma S,
Xu F, Bai T, Tang Y, Xu Y and Liu L
(2022) Computational Characterizing
Necroptosis Reveals Implications
for Immune Infiltration
and Immunotherapy of
Hepatocellular Carcinoma.
Front. Oncol. 12:933210.
doi: 10.3389/fonc.2022.933210

Jun Zhu^{1,2†}, Tenghui Han^{3†}, Shoujie Zhao⁴, Yejing Zhu⁴, Shouzheng Ma⁵, Fenghua Xu¹,
Tingting Bai¹, Yuxin Tang¹, Yungang Xu^{6,7*} and Lei Liu^{1*}

¹ Department of Gastroenterology, Daping Hospital, Army Medical University, Chongqing, China, ² Department of General Surgery, The Southern Theater Air Force Hospital, Guangzhou, China, ³ Department of Neurology, Xijing Hospital, Fourth Military Medical University, Xi'an, China, ⁴ Department of General Surgery, Tangdu Hospital, Fourth Military Medical University, Xi'an, China, ⁵ Department of Surgery, Tangdu Hospital, Fourth Military Medical University, Xi'an, China, ⁶ Department of Cell Biology and Genetics, School of Basic Medical Sciences, Xi'an Jiaotong University Health Science Center, Xi'an, China, ⁷ Centre for Computational Systems Medicine, School of Biomedical Informatics, The University of Texas Health Science Centre at Houston, Houston, TX, United States

Necroptosis is a programmed form of necrotic cell death in regulating cancer ontogenesis, progression, and tumor microenvironment (TME) and could drive tumor-infiltrating cells to release pro-inflammatory cytokines, incurring strong immune responses. Nowadays, there are few identified biomarkers applied in clinical immunotherapy, and it is increasingly recognized that high levels of tumor necroptosis could enhance the response to immunotherapy. However, comprehensive characterization of necroptosis associated with TME and immunotherapy in Hepatocellular carcinoma (HCC) remains unexplored. Here, we computationally characterized necroptosis landscape in HCC samples from TCGA and ICGA cohorts and stratified them into two necroptosis clusters (A or B) with significantly different characteristics in clinical prognosis, immune cell function, and TME-landscapes. Additionally, to further evaluate the necroptosis levels of each sample, we established a novel necroptosis-related gene score (NRGscore). We further investigated the TME, tumor mutational burden (TMB), clinical response to immunotherapy, and chemotherapeutic drug sensitivity of HCC subgroups stratified by the necroptosis landscapes. The NRGscore is robust and highly predictive of HCC clinical outcomes. Further analysis indicated that the high NRGscore group resembles the immune-inflamed phenotype while the low score group is analogous to the immune-exclusion or metabolism phenotype. Additionally, the high NRGscore group is more sensitive to immune checkpoint blockade-based immunotherapy, which was further

validated using an external HCC cohort, metastatic melanoma cohort, and advanced urothelial cancer cohort. Besides, the NRGscore was demonstrated as a potential biomarker for chemotherapy, wherein the high NRGscore patients with more tumor stem cell composition could be more sensitive to Cisplatin, Doxorubicin, Paclitaxel-based chemotherapy, and Sorafenib therapy. Collectively, a comprehensive characterization of the necroptosis in HCC suggested its implications for predicting immune infiltration and response to immunotherapy of HCC, providing promising strategies for treatment.

Keywords: necroptosis, tumor microenvironment, immunotherapy, chemotherapy, tumor-infiltrating cells, hepatocellular carcinoma

INTRODUCTION

Hepatocellular carcinoma (HCC) is acknowledged to be one of the most common malignant tumors globally, accounting for one-third of cancer mortalities (1). Risk factors of HCC progression contain metabolic disorders, viral infections by hepatitis B virus and hepatitis C virus, absorption of the aflatoxin-contaminated meal, and cirrhosis out of alcoholic hepatitis (2–4). Though early examination and intervention have achieved success, the ratio of diagnosis at the early stage remains low (5). The majority of HCC patients are diagnosed at an advanced stage (6). Although plentiful efforts were made in both diagnosis and treatment of HCC, its overall survival (OS) rate remains unfavorable. Seventy percent of HCC patients have recurrent neoplasm 5 years after resection (7). Consequently, the 5-year survival rate of HCC patients turns out unsatisfactory (8).

Chemotherapy failure has been a major obstacle during cancer treatment, among which apoptosis resistance (innate or acquired) is widely accepted mechanistically. How to bypass the apoptosis pathway and induce effective cell death pathways are becoming crucial in the treatments of HCC. Like apoptosis and necrosis, necroptosis is a novel programmed form of necrotic cell death that is mainly manipulated by receptor-interacting Protein Kinase 1 (RIPK1), RIPK3, and Mixed Lineage Kinase domain-like (MLKL) (9). Accumulating data demonstrate that necroptosis plays a crucial role in the regulation of cancer biology, including cancer progression (10, 11), cancer metastasis (12), cancer Immunosurveillance (13), and cancer subtypes (14, 15).

With the advent of immunotherapy, such as immune checkpoint blockades (ICBs), only a few HCC patients have been reported to gain clinical benefits from ICBs (16, 17). To date, Some predictors of response to ICBs were identified, including PD-L1 expression (18), the degree of cytotoxic T cell infiltration (19), tumor mutational burden (TMB) (20), mismatch repair deficiency (21), and activated Wnt/ β -catenin signaling (22), et al. However, to our best knowledge, these above biomarkers have been confirmed in HCC (23). There is a rationale supporting the development of some novel biomarkers for ICBs in HCC patients. Recently, a few studies revealed crosstalk between necroptosis and antitumor immunity (24). Necroptosis not only has direct interaction with immune cells like dendritic cells (DCs) and natural killer T cells (NKT) but also could initiate an adaptive immune system by promoting

DC cells and macrophages to release pro-inflammatory cytokines into the tumor microenvironment (TME), incurring strong immune responses (25). In *in vivo* and *in vitro* experiments, necroptotic tumor cells were shown to induce antitumor immunogenicity through the cross-priming and proliferation of CD8⁺ T cells (26). Furthermore, with the advance in nanomedicine, necroptotic cancer cells could boost antitumor immunity by administrating cell-mimicry nanovaccine with a tailored immunostimulatory modality (26). In addition, the effect of tumor regression by nanovaccine could be enhanced by combination with ICBs (26). However, the discovery of a specific necroptosis marker, a thorough investigation of the molecular mechanism, and a clarification of its crosstalk with other cell death machinery and its interaction with the immune system should be urgently further investigated. Taken together, necroptosis has a close relationship with TME and antitumor immunity, suggesting necroptosis could become a novel biomarker for chemotherapy and immunotherapy of HCC. However, it's rarely depicted for necroptosis characteristics from a multi-omics perspective and its correlation with ICBs and chemotherapy in HCC.

In this study, we uncovered that necroptosis regulators were clinically predictive and independent factors for HCC patients by survival analysis and unsupervised clustering analysis. Necroptosis regulators could distinguish patients into two necroptosis clusters (Nclusters A and B) and different Nclusters were enriched into several TME-related pathways and various immune infiltration pathways. Therefore, we proposed a hypothesis: necroptosis could become a new and indispensable factor for HCC prognosis and TME characterization. A novel necroptosis-related gene score (NRGscore) was established by the Lasso algorithm and multiple Cox regression analysis. In TME cell infiltration of HCC, NRGscore had a negative relation with activated NK cells and macrophage M1 cells while positively associated with Treg cells. NRGscore had meaningful guidance for patients, where the high NRGscore group could be more sensitive to ICB treatment, which was validated in other independent cancer cohorts, as well. As for chemotherapy and targeted therapy, due to more stem cells and enhanced cell proliferation, the high NRGscore group could be more sensitive to Cisplatin, Doxorubicin, Paclitaxel-based chemotherapy, and Sorafenib targeted therapy than its counterparts. Taken together, the prognostic NRGscore system

could help to dissect the TME characterization of HCC and to interpret the clinical responses to chemotherapies and immunotherapies, providing new target molecules for the treatment of cancers.

METHODS

Data Sources

The workflow of this study is shown in **Figure S1**. Necroptosis-related genes were collected based on the GESA necroptosis gene set and published literature (**Table S1**). Clinical information and mRNA expression matrixes of HCC patients were acquired from The Cancer Genome Atlas (TCGA, <https://portal.gdc.cancer.gov/>) database and International Cancer Genome Consortium (ICGC, <https://dcc.icgc.org/>) database. Then we transformed FPKM values into transcripts per kilobase million (TPM) values. The basic clinical information of HCC datasets in the study is summarized in **Table S2**. The somatic mutation data of the TCGA and ICGC cohort were downloaded from the UCSC Xena (<https://gdc.xenahubs.net/>). Additionally, the TCGA cohort was applied to Copy Number Variation (CNV) analysis. Another HCC validated cohort (GSE54236) (27, 28) and corresponding clinical features were collected to assess predictive power in our study.

Identification of Significant Mutational Genes

The “maftools” R package was applied to process the mutation annotation format (maf) data, and the “MutSigCV” algorithm was implemented to screen the significant mutational genes (SMGs) (29). The significance of nonsilent somatic mutations in a gene was measured based on the background mutation rates by silent mutation. The false discovery rates (FDR (30)) were then calculated, and genes with statistical significance ($FDR \leq 0.1$) were set as SMGs. Then, waterfall plots were employed to visualize the mutation information of these significant SMGs in the TCGA cohort. Besides, Fisher’s test was applied to detect the mutually exclusive or co-occurring ratio of necroptosis-related genes. By adopting the “ExtractSignatures” function that applies the Bayesian nonnegative matrix factorization-based framework, we determined the mutational signatures using the genomic data. The optimal number of mutational signatures for the TCGA cohort could be detected by the “SignatureEnrichment” function and then it automatically assigned a given signature to each sample.

Unsupervised Cluster Analysis

Based on the expression of necroptosis-related genes, unsupervised clustering analysis was performed to stratify patients into different clusters. The “ConsensusClusterPlus” R package was adopted to determine the number of clusters and guarantee the stability of classification (31). The principal component analysis (PCA) was utilized to investigate gene-expression arrays among distinct clusters.

Tumor Infiltration Cell and Immune-Related Function Analysis

To estimate the abundance and activity of tumor infiltration cells (TICs) in HCC, the single-sample gene set enrichment analysis (ssGSEA) algorithm was implemented by using the “GSVA” R package (32). In addition, CIBERSORT (33), an analysis algorithm based on the immune gene set, was also used to evaluate the TICs levels of HCC. The algorithm was run for 1000 permutations and HCC samples with an output $P < 0.05$ were selected as previously reported (34, 35). Twenty-three types of TICs were comprised of adaptive immune cells (B cells, T cells, CD8 T cells, T follicular helper (Tfh), Th1, Th2, Th17, and Treg cells), and innate immune cells (NK cells, CD56 dim NK cells, CD56 bright NK cells, DCs, plasmacytoid DCs, immature DCs, neutrophils, mast cells, and macrophages). Besides, immune-related pathways (such as cytolytic activity, T-cell costimulation, inflammation-promoting, and para-inflammation) were also calculated *via* ssGSEA. The biosimilarity of infiltrating immune cells and immune-related functions were estimated by the Gaussian fitting model.

Functional Annotation Analysis

To investigate the enrichments of biological processes, cell components, and molecular function pathways, Gene Ontology (GO) analysis, and Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway analysis were conducted using the “clusterProfiler” R package. Additionally, Gene Set Enrichment Analysis (GSEA) (36) was adopted to identify the differences in DEGs between the distinct clusters in the enrichment of the KEGG pathway. Permutations were performed 1000 times for each analysis. P -value < 0.05 and adjusted P -value (Q value) < 0.05 were considered statistically significant.

Differentially Expressed Genes Analysis

To identify necroptosis-related genes, patients were classified into distinct groups according to sample types, necroptosis clusters, and NRGscore, respectively. The “limma” R package was utilized to determine DEGs between different groups (37). The significance filtering criteria for determining DEGs were set as adjusted P -value < 0.001 and fold change > 1.5 .

Dimension Reduction and Construction of NRGscore

Prognostic genes were identified from DEGs by performing the univariate Cox regression. The least absolute shrinkage and selection operator (Lasso) regression was conducted to necroptosis gene signature *via* utilizing the “glmnet” R package. Responding coefficients (β) of the signature were verified. Additionally, the signature was calculated by the following equation: $NRGscore = \sum(\exp(\text{gene}) * \beta)$, where \exp indicated RNA expression of HCC samples.

Clinical Characteristic Evaluation of NRGscore

The survival curves for the different subgroups were generated by Kaplan-Meier (K-M) methods. The areas under the curve (AUC)

of the receiver operating characteristic (ROC) curve were applied to assess the predictive value of gene signature. The nomogram was built based on the NRGscore and clinicopathologic characteristics including age, gender, T stage, and N stage to predict the survival probability of 1-, 3-, and 5-year OS of HCC patients. The calibration curve of the nomogram was plotted to estimate the prediction possibilities according to the observed survival rates. The nomogram and calibration plots were generated based on the “rms” R package.

Estimation of Tumor Immune Microenvironment and Tumor Mutational Burden

To further dissect the immune landscape of HCC, the “ESTIMATE” package in R was used to evaluate the immune, stromal, and ESTIMATE scores, which reflect the ratio of the immune/stromal components of the tumor immune microenvironment (TME). To determine the tumor mutational burden (TMB) of each patient, we also counted the nonsynonymous and synonymous mutation counts in the TCGA cohort.

Assessment of Clinical Response to the Immunotherapy

TIDE (38) algorithm and immunophenoscore (IPS) function (39) were prevalently recognized to be effective methods to predict cancer patients' response to immunotherapy. There are two main mechanisms to immune escape and resistance to ICB-based immunotherapy (40, 41): (1) high levels of dysfunctional CTL; (2) immunosuppressive factors to exclude T cells from the tumor region. TIDE algorithm integrates these two immune-escape mechanisms and was used to predict patients' response to ICB-based therapy based on the transcriptome. According to characterizing the determinant factors of cancer immunogenicity and antigenomes, we stratified HCC patients into different IPS groups. HCC sample is more immunogenic when the z-score of IPS is higher (42). Two immunotherapy cohorts, metastatic melanoma cohort treated with Nivolumab (19), and advanced urothelial cancer cohort with the intervention of atezolizumab (IMvigor210, <http://research-pub.gene.com/IMvigor210> CoreBiologies/) (43) were downloaded to evaluate the predictive value of the necroptosis score system for immunotherapy. In addition, the TIDE website (<http://tide.dfci.harvard.edu/login/>) was used to further evaluate the predictive power of other cancer cohorts by inputting gene coefficients. The raw gene expression data of all cohorts were normalized according to previous literature (44, 45).

Dissection of Cancer Stem Cell And Prediction of Chemotherapeutic Drugs Sensitivity

To evaluate the cancer stem cell ratio and differentiation degree of HCC, the “limma” and “corrplot” R packages were applied. We adopted Spearman's method to explore the correlation between

NRGscore and cancer stem cells. In addition, based on the information retrieved from the Genomics of Drug Sensitivity in Cancer database (<https://www.cancerrxgene.org/>), we estimate the sensitivity of different chemotherapeutic drugs between high and low NRGscore subgroups. The prediction process used was the “pRRophetic” package (46) where the half-maximal inhibitory concentration was calculated by ridge regression model based on gene expression profiles.

Statistical Analysis

All statistical analyses were conducted by R software (<http://www.R-project.org>, version 4.1.2), PERL programming language (version 5.32.1.1, <https://www.perl.org/>). K-M curve analysis with a log-rank test was utilized to compare OS between diverse subgroups. Mann-Whitney and Kruskal-Wallis test with adjusted P-values were employed to compare either ssGSEA scores of immune cells or functions of the distinct clusters as indicated in the article. P-value < 0.05 was statistically significant.

RESULTS

Hepatocellular Carcinoma Was Well Characterized by Necroptosis Genes

Firstly, we investigated 67 necroptosis genes (Table S1) regarding their expression and genetic variations between tumor and paratumor samples in TCGA cohorts. Fifty-one necroptosis genes presented significantly differential expression, of which 42 genes were upregulated and nine genes were downregulated in tumors (Figure 1A). Many differentially expressed necroptosis genes have been reported to be involved in modifying the microenvironment of HCC (15), including upregulated SIRT1/2 (47) and RIPK1 (48). It is well known that genetic variations, such as copy number variations (CNV) and mutations, result in perturbations of gene expression during tumorigenesis of HCC. The investigation of CNV alternation frequency showed a prevalent change in 67 necroptosis genes, and we observed that 44 genes exhibited the amplification in copy numbers and 23 genes showed depletion of CNVs (Figures 1B, C). Additionally, a total of 108 of 364 (29.67%) patients exhibited at least one type of mutation (Figure 1D). The most highly mutated genes, such as MYC, TERT, FASLG, RIPK1, TNFSF10, TARDBP, and CDKN2A, have been well characterized in liver cancer (15, 48), diffuse large B-cell lymphoma (49), neuroblastoma (50), and kidney renal clear cell carcinoma (51, 52). Especially, as the three most important necroptosis-driving molecules (53), RIPK1 and MLKL were upregulated in tumors and CNV of RIPK1 revealed amplification more than depletion, while RIPK3 presented no significant alternation in genomic expression. Extrapolating from all the above results, necroptosis genes present high heterogeneity among RNA expression, CNV, and mutations in HCC samples, which shows promise in tumorigenesis and development in HCC.

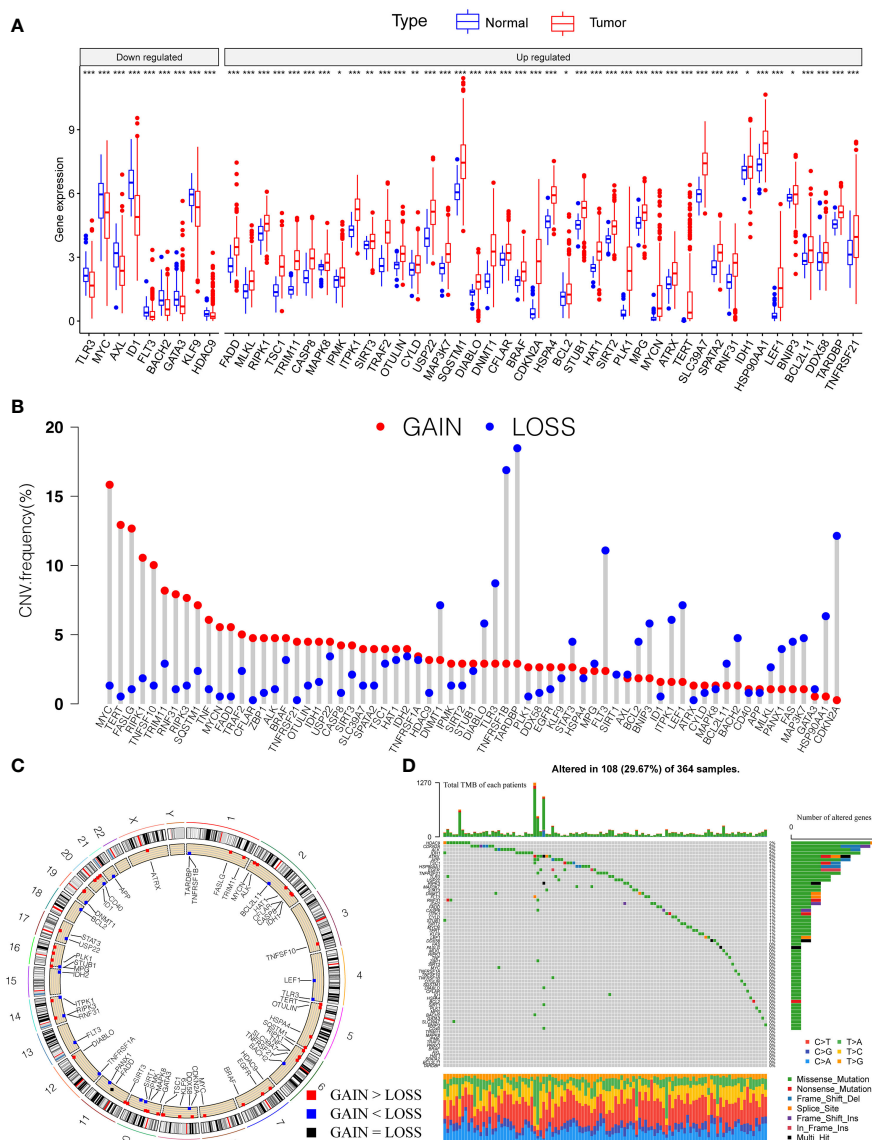


FIGURE 1 | Genetic landscape and expression variation of necroptosis regulators in HCC. **(A)** The differential expression level of necroptosis regulators between tumor and normal tissues. **(B)** CNV alteration of necroptosis regulators in tumor tissues. Column represented the frequency of the variations. The green dot represented the deletion of CNV. The red dot represented the amplification of CNV. **(C)** Location of CNV of necroptosis regulators in chromosomes. Red dots represent genes gain than loss, blue dot presents genes loss than gain and the black dot means loss equal gain. **(D)** One hundred eight patients (29.67%) exhibited various genetic alterations, including missense, nonsense, splice, frameshift, and multiple mutations. Each column indicated individual HCC patients, and the upper bar diagram exhibited the TMB of HCC patients. The right number represented the mutation frequency, and the bar diagram on the right exhibited the proportion of each genetic alteration. $P < 0.05$ *; $P < 0.01$ **; $P < 0.001$ ***. HCC, hepatocellular carcinoma; CNV, copy number variation; TMB, tumor mutation burden.

Necroptosis Clusters Exhibit Distinct Prognosis, Biological Functions, and Immune Characteristics

To integrate the prognostic value of these necroptosis genes, we sought to construct a prognosis and correlation network according to the clinical outcomes of HCC patients (**Figure 2A**). Three necroptosis genes were favorable factors while 22 genes were risk factors, which disclosed that necroptosis may play a tumor-promoting role in HCC. Interestingly, we found that a positive

correlation between prognostic necroptotic genes occurred more frequently than negative connections. To further explore the clinical role of necroptosis genes in HCC, we collected 599 clinical samples from TCGA and ICGC datasets and the basic clinical parameters are shown in **Table S2**. Based on expression profiles of necroptosis genes, we implemented unsupervised clustering to analyze the HCC samples from two cohorts and classified patients into qualitatively different subgroups. Two distinct subgroups were ultimately identified, including 244

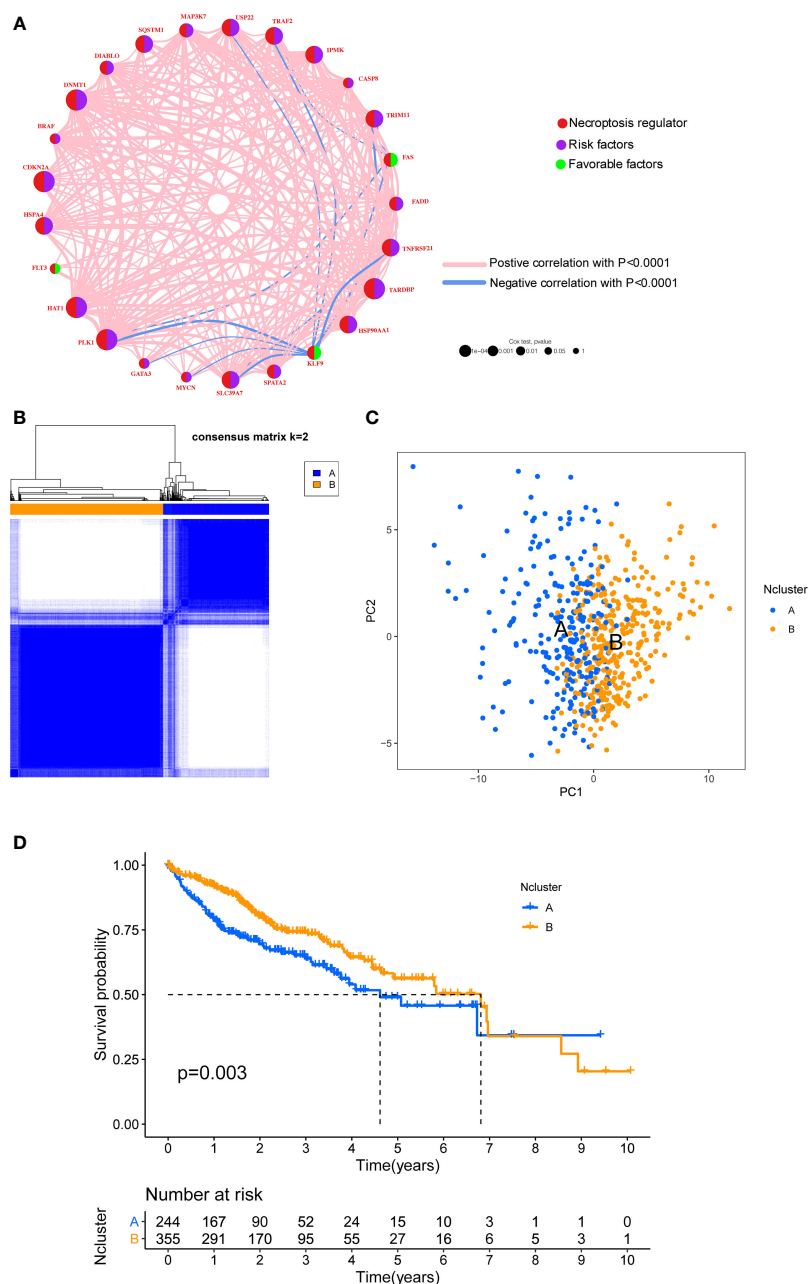


FIGURE 2 | Unsupervised clustering of necroptosis regulators. **(A)** Interaction network of necroptosis regulators in HCC. Necroptosis regulators were indicated by red circles. Risk factors were indicated by purple circles, and favorable factors were indicated by green circles. Different sizes of the circle represented the different P-values. Red lines showed the positive correlation, and blue lines showed the negative correlation. **(B)** Consensus clustering matrix for $k=2$. **(C)** The feature distribution between different two Nclusters was plotted via PCA. **(D)** Kaplan-Meier curves of the OS for the two Nclusters of HCC patients. Ncluster, necroptosis cluster; PCA, principal component analysis; OS, overall survival; HCC, hepatocellular carcinoma.

cases in subgroup A and 355 cases in subgroup B, which were referred to as necroptosis clusters (Ncluster) A and B, respectively (**Figure 2B**). Additionally, PCA results showed necroptosis genes could vividly distinguish one subgroup from another (**Figure 2C**). K-M curves for the two subgroups revealed the prominent survival advantage in Ncluster B compared to A (**Figure 2D**). Moreover, correlations between two clusters and other clinicopathological

parameters, including age, gender, TNM stage, tumor stage, and tumor grade, are shown in the heatmap (**Figure S2**). Ncluster A was characterized by higher levels of necroptosis and more advanced clinical stage and grade, thus resulting in a poor prognosis.

To further explore the potential biological behavior of necroptotic patterns, we performed GSEA analysis between

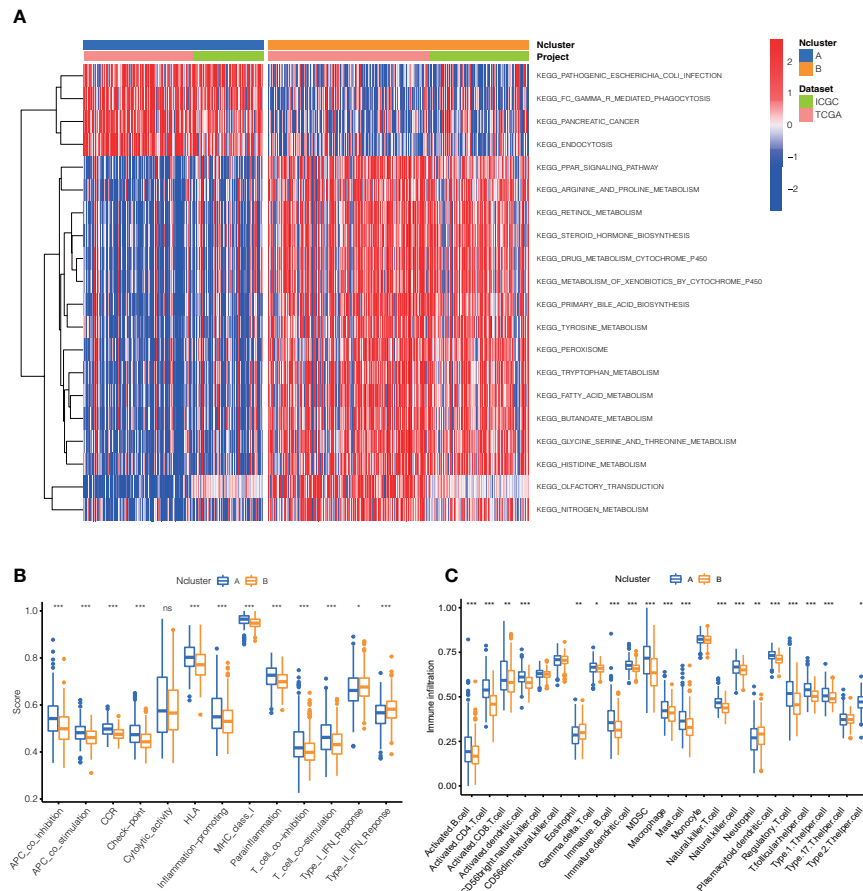


FIGURE 3 | Dramatic difference of biological features between Ncluster A and B. **(A)** Heatmap showed various KEGG pathways were enriched in Ncluster A and B. **(B)** Immune function and infiltrating immune cells **(C)** altered in two Nclusters shown by violin diagram. The heatmap was used to visualize these biological processes, and red represented activated pathways and blue represented inhibited pathways. Ncluster, necroptosis cluster.

two Nclusters. In KEGG signal pathways, Ncluster B was characterized by enhanced metabolism pathways, such as drugs, steroid hormone, bile acid, and fatty acid metabolism, while Ncluster A was characterized by upregulated cell endocytosis and phagocytosis pathways (**Figure 3A**). The tumor-infiltrating cells and immune functions play a critical role in the process of tumorigenesis and migration in HCC (54). In immune pathways of GSVA, we found Ncluster A possess more pathways for activated immune cell than Ncluster B (**Figure S3A**), suggesting N cluster A had a trend to immune-inflamed phenotype and Ncluster B was inclined to immune-excluded phenotype. Next, we investigated the difference between TICs and immune-related functions in two clusters by ssGSEA analysis. As vividly shown in **Figure 3B**, 10 immune functions were markedly downregulated and only IFN-response function upregulated in Ncluster B, which revealed that Ncluster B had less complicated TME. Consistent with results of immune pathways, there was significantly different TIC abundance in two subgroups, among which 19/23 TICs varied immensely, including 17 highly expressed types and only two downregulated

kinds of TICs in Ncluster A (**Figure 3C**). In addition, we also determined 2988 Ncluster-related DEGs (**Table S3**) by limma R packages and conducted GO and KEGG enrichment analysis based on the Ncluster DEGs. Likely, DEGs were enriched in cell adhesion (EMT pathways) and T cell activation pathways (**Figure S3B**), PI3K-AKT pathways, and abundant immune cell regulation pathways (**Figure S3C**). Collectively, necroptosis could classify patients into two distinct subgroups and Ncluster A tends to have immune-inflamed phenotype and complex tumor-infiltrating patterns while Ncluster B was inclined to immune-excluded phenotype and metabolism phenotype.

Necroptosis Patterns Define Necroptosis-Related Gene Clusters and NRGscore

To further evaluate the underlying biological role of necroptosis in clinical outcomes and tumor-infiltrating traits, we screened 2988 Ncluster related DEGs and determined 1819 prognostic genes by univariate Cox analysis (**Table S4**). Similarly, an

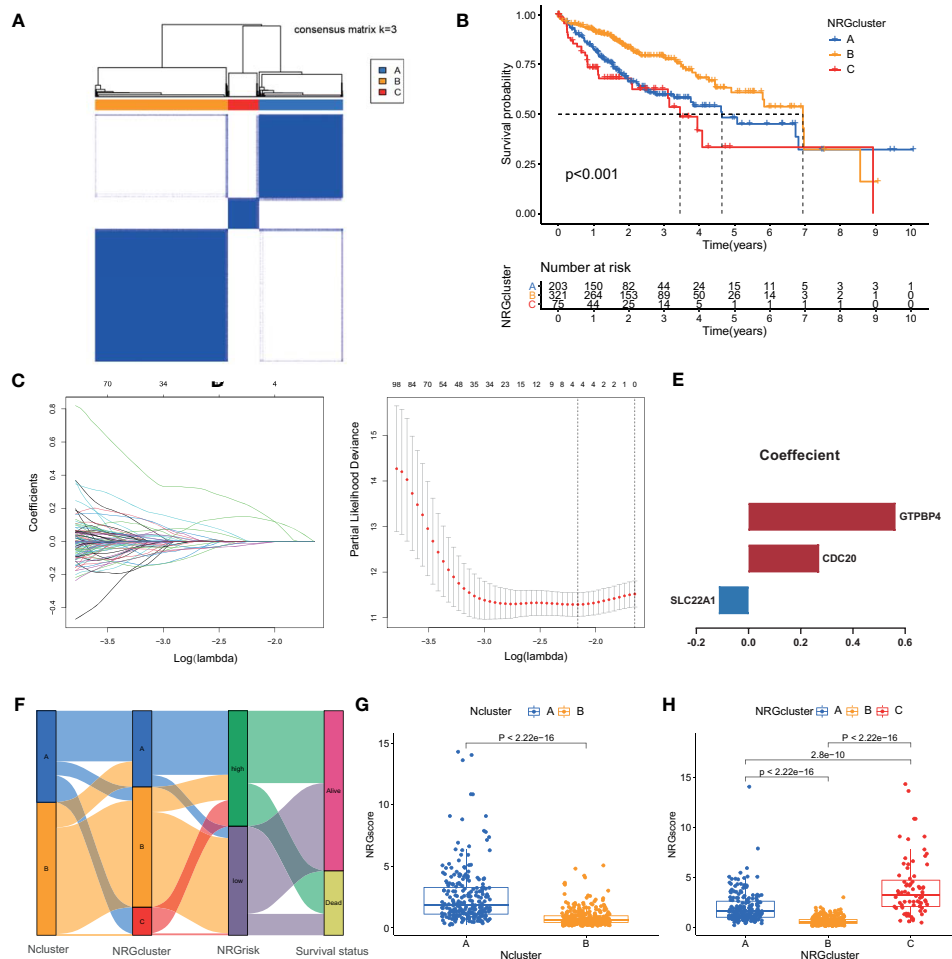


FIGURE 4 | Identification of necroptosis genomic classification and construction of NRGscore. **(A)** Consensus clustering matrix for $k=3$. **(B)** Kaplan-Meier curves indicated necroptosis genomic phenotypes were markedly related to the OS of 599 patients in TCGA and ICGC cohorts, of which 203 cases were in gene cluster A, 321 cases in gene cluster B, and 75 cases in gene cluster C. **(C)** The partial likelihood deviance plot. **(D)** The Lasso regression coefficient profiles. **(E)** Coefficient of three core necroptosis-related DEGs. **(F)** The alluvial diagram exhibited the correlation of Ncluster, NRGcluster, NRGscore, and survival status. **(G)** Differences in NRGscore between two Nclusters in TCGA and ICGC cohorts. The upper and lower ends of the boxes represented the interquartile range of values. The lines in the boxes represented median value, and dots showed outliers. **(H)** Differences in NRGscore among three NRGclusters in TCGC and ICGC cohorts. $P < 0.05$ *; $P < 0.01$ **; $P < 0.001$ ***. NRGscore, necroptosis-related gene score; OS, overall survival; TCGA, the cancer genome atlas; ICGC, international cancer genome consortium; Lasso, least absolute shrinkage and selection operator; DEGs, differentially expressed genes; Ncluster, necroptosis cluster; NRGclusters, necroptosis-related genes clusters.

unsupervised clustering method was used to accurately classify HCC patients into three stable gene clusters, which was termed as necroptosis-related gene cluster (NRGcluster), including 203 cases in NRGcluster A, 321 cases in NRGcluster B, and 75 cases in NRGcluster C (**Figure 4A**). The three NRGclusters had dramatic prognostic differences in HCC patients: NRGcluster B was proven to possess better prognostic outcomes, while patients in NRGcluster C were associated with poorer outcomes (**Figure 4B**). Furthermore, the three NRGclusters exhibited the different Nclusters, necroptosis gene expression (**Figure S4A**), among which NRGclusters had nearly Ncluster B subgroups and NRGcluster C was included in Ncluster A. Besides, the three NRGclusters had distinct expression profiles of necroptosis genes

(**Figure S4A, B**). Among them, NRGcluster A or C had more expression of 51 than NRGcluster B while only six necroptosis regulators were highly expressed in Ncluster B compared to A or C, which revealed that the expression level of necroptosis genes had a negative relationship with HCC clinical prognosis and NRGcluster could represent necroptosis patterns of each HCC sample.

Integrating the evidence above, we found that necroptosis played an irreplaceable role in affecting prognosis, remodeling TME, and regulating immune reaction in HCC. However, the Ncluster or NRGcluster could not accurately clarify individual necroptosis patterns, and to further explore the heterogeneity and complexity of necroptosis, the Lasso algorithm was

employed to quantify individual patients. Based on the optimal value of λ ($\lambda=3$), we constructed a necroptosis scoring system, termed as NRGscore (**Figures 4C, D**) by adopting multivariate COX regression analysis. Three hub genes were comprised, including GTPBP4 (Coefficient = 0.5632), CDC20 (Coefficient = 0.271), and SLC22A1 (Coefficient = -0.1136) (**Figure 4E**). The NRGscore was calculated by RNA expression multiplied by its corresponding coefficient.

The alluvial diagram visualized the quantification changes of patients and exhibited the interaction of Ncluster, NRGcluster, NRGscore, and survival state (**Figure 4F**). Results indicated that Ncluster A was mainly linked to a higher NRGscore, whereas Ncluster B exhibited a lower score (**Figure 4G**). Remarkably, NRGcluster C showed the highest NRGscore, followed by NRGcluster A, while NRGcluster B revealed the lowest scores (**Figure 4H**). Then, we investigated the relationship between NRGscore and necroptosis-driving regulators (RIPK1, RIPK2, and MLKL) and found NRGscore was positively linked to the expression of these three regulators (**Figure S5**). Additionally, we further stratified patients into high NRGscore and low NRGscore groups according to the median cut-off value derived from the R Survminer package. Next, we investigated the RNA expression levels of necroptosis genes, and the results exhibited that most (46/49) genes were highly expressed and only three (3/49) were downregulated in the high NRGscore subgroup (**Figure S6A**). Taken together, NRGcluster could accurately represent overall necroptosis levels and the NRGscore system could depict individual necroptosis patterns of HCC patients.

NRGscore Is Highly Predictive of Clinical Outcomes for HCC

Then, the prognostic value of the necroptosis scoring system in predicting patients' survival outcomes was also estimated. We divided HCC patients into two training and testing cohorts randomly, including 300 cases in the training set, and 299 cases in the test set. K-M curves demonstrated that HCC patients in the high NRGscore group had lower OS rates than their counterparts in overall, training, and test sets ($P < 0.001$) (**Figures 5A, C**). Next, ROC curves were plotted to appraise the accuracy of the scoring system in predicting survival at 1, 3, and 5 years. (Overall set: AUC at 1-, 3-, and 5-year is 0.766, 0.718, and 0.685; Training set: 0.774, 0.756, and 0.824, respectively; Test set: 0.762, 0.691, and 0.530, separately; **Figures 5D, F**). Besides, we assessed the NRGscore of each HCC case amidst the overall, training, and test sets, which implied that HCC patients in the low NRGscore group had better survival status and less ratio of dead status than the high NRGscore group (**Figures S6B, D**). Therefore, a high NRGscore was considered a high NRGrisk while a low NRGscore was deemed a low NRGrisk. In addition, to further evaluate the clinical application of the NRGscore system, we downloaded an external HCC dataset (GSE54236) and calculated the NRGscore of each HCC patient. The survival curve demonstrated that the high NRGrisk had a poorer prognosis than the low risk (**Figure S7A**), and the ROC curve indicated the NRGscore system had a robust predictive ability

(**Figure S7B**). We also found patients with high NRGrisk had advanced clinical stage and grade than low NRGrisk (**Table S5**). In general, the above results indicated that the necroptosis scoring system had a promising capacity to predict the survival of HCC patients.

To better build a convenient and applicable clinical prognosis evaluation method, we plotted a nomogram graph based on the NRGscore system and clinical features including gender, age, and pathologic stage (**Figure 5G**). The concordance index of the nomogram was 0.707 (95% CI 0.652–0.762). The calibration plot for the possibility of 1-, 3-, and 5-year survival exhibited good coincidence degrees of survival probability between the prediction and real observations (**Figure 5H**). These results demonstrated that the nomogram could be an effective approach to predicting the prognosis outcome of HCC patients for clinicians. All the above results disclosed that NRGscore could have a satisfying clinical prediction value in HCC.

NRGscore Accurately Predicts Immunotherapeutic Benefits for HCC

Based on the above results, Ncluster A was characterized by more abundant immune infiltrating cells and enhanced immune function than Ncluster B. Ncluster A could represent the immune-inflamed type and B represent the immune-excluded phenotype, which gave us a hint that the Ncluster A subgroup of patients could respond better to immunotherapy than B cluster. As shown in the alluvial diagram, high NRGscore had an overwhelming ratio of Ncluster A while low NRGscore were nearly derived from Ncluster B. Therefore, we speculated that patients with high NRGscore could be more sensitive to immunotherapy. First, we performed GSEA to investigate the biological of high and low NRGrisk, and the results (**Figure 6A**) revealed that high NRGrisk was characterized by cell proliferation and DNA repair-related pathways (cell cycle, DNA replication, nucleotide excision repair, and spliceosome, et al.) while low NRGrisk was characterized by enhanced metabolism pathways (drug metabolism of P450, fatty acid metabolism, and bile acid metabolism, et al.). Alternatively, the above results were also verified by other famous pathways, including hallmark pathways (**Figure S8A**) and reactome pathways (**Figure S8B**). Surprisingly, almost all immune pathways were enriched in the high NRGrisk group (**Figure S8C**), which further demonstrated high NRGrisk resembled Ncluster A group and could be more sensitive to clinical immunotherapy. Additionally, we further calculated the TME parts, including stromal, immune, and ESTIMATE parts. As expected, the stromal component was significantly downregulated in the high NRGrisk score group (**Figure 6B**). Previous research underlined the hub role of stromal activation in resistance to checkpoint immunotherapy and high NRGrisk patients with low stromal could get more clinical benefits from immunotherapy. To comprehensively evaluate the correlation between the scoring system and immunotherapy, we first estimated the TICs abundance based on the NRGscore. Results indicated that core genes in the scoring system had a close relation to various types of TICs. Among them, CDC20 was

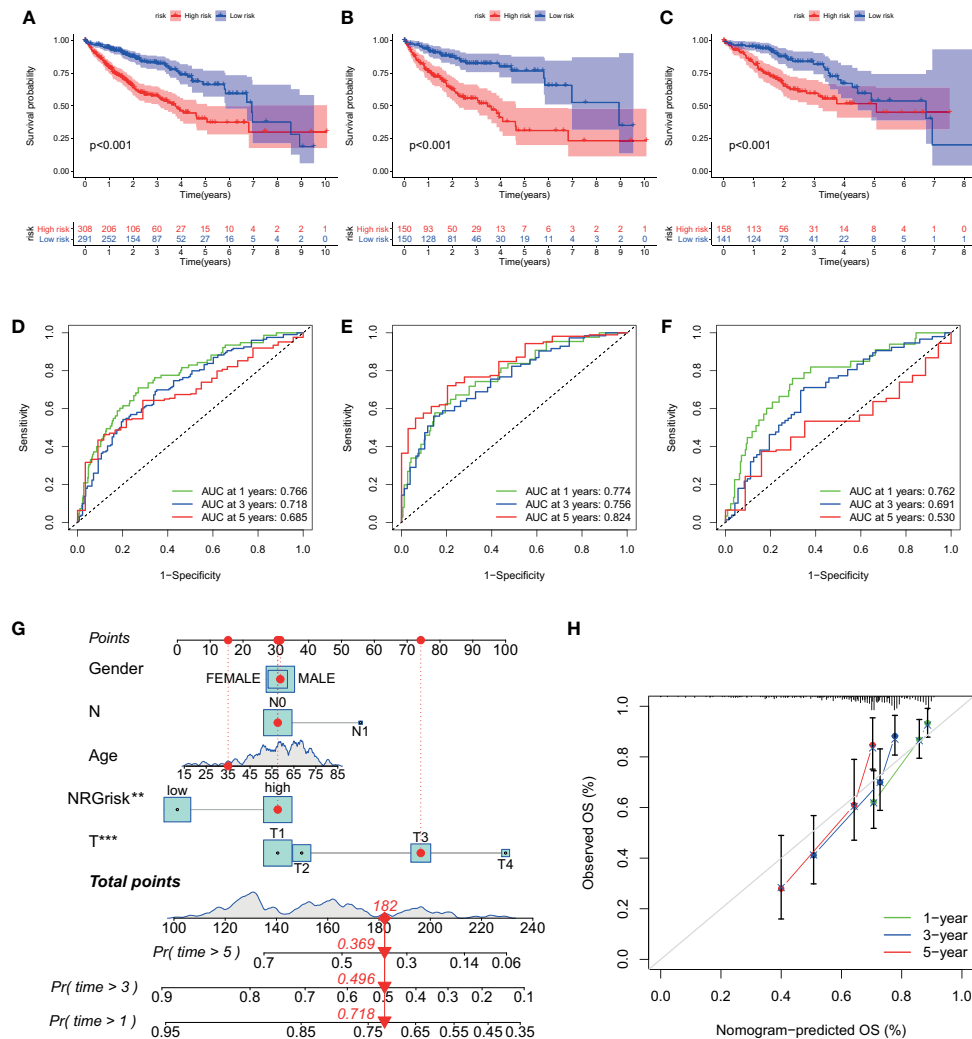


FIGURE 5 | Evaluation of predictive power of NRGscore in training and test cohorts. **(A, C)** Kaplan-Meier curves showed that the high NRGrisk group had a more inferior OS than the low NRGrisk group in TCGA and ICGC cohorts, of which 599 cases were in the overall set **(A)**, 300 cases were in training set **(B)**, and 299 cases were in the test set **(C)**. **(D-F)** ROC of NRGscore scheme: Areas under the curve of 1-, 3-, and 5-year OS in the overall set **(D)**, training set **(E)**, and test set **(F)**. **(G)** Nomogram predicting 1-, 3-, and 5-year OS of patients based on NRGscore and other clinical parameters, including gender, age, T stage, and N stage. The red dot presented the point of each parameter, and the length of the line segment reflected the contribution of factors to the outcome event. **(H)** Calibration plot of the nomogram for predicting the probability of 1-, 3-, and 5-year OS. The colored line was the fit line and represented the predicted value (the horizontal axis) corresponding to the actual value (the vertical axis). The gray diagonal was the ideal case. $P < 0.01^{**}$; $P < 0.001^{***}$. TCGA, the cancer genome atlas; ICGC, international cancer genome consortium; NRGscore, necroptosis-related gene score; NRGrisk, necroptosis-related gene risk; ROC, receiver operating characteristic curves; OS, overall survival.

mainly related to T immune cells; GTPBP4 had a tight correlation with NK immune cells; SLC22A1 was mostly correlated to macrophages and DC immune cells (**Figure 6C**). Various immune functions altered with the NRGscore system; the heatmap indicated that a higher NRGscore always followed more inhibition of immune response (APC co-inhibition, MHC-I class, T cell co-inhibition and Treg cell, IFN-response inhibition; **Figure S8D**). Besides, three TICs were significantly positively correlated with the NRGscore, including Mast cells activated ($R=0.22$), T cells follicular helper ($R=0.26$), and T cells CD4 memory activated ($R=0.32$), while five TICs were

significantly negatively correlated, including B cells naïve ($R=-0.24$), NK cells activated ($R=-0.24$), macrophages M1 ($R=-0.25$), Mast cells resting ($R=-0.3$), and T cells CD4 memory resting ($R=-0.36$) (**Figure S9**). To our knowledge, PD-L expression was considered as an important predictor of response to ICBs, and we found that NRGscore was positively associated with several known immune checkpoints (PD-1, PD-L1, PD-L2, CTLA4, TIM-3, IDO1, LAG3, TIGIT; **Table S6**). Integrating the results above, we found the NRGscore system could be dramatically correlated to the immune landscape of HCC, and its role in estimating the immunotherapy deserved further exploration.

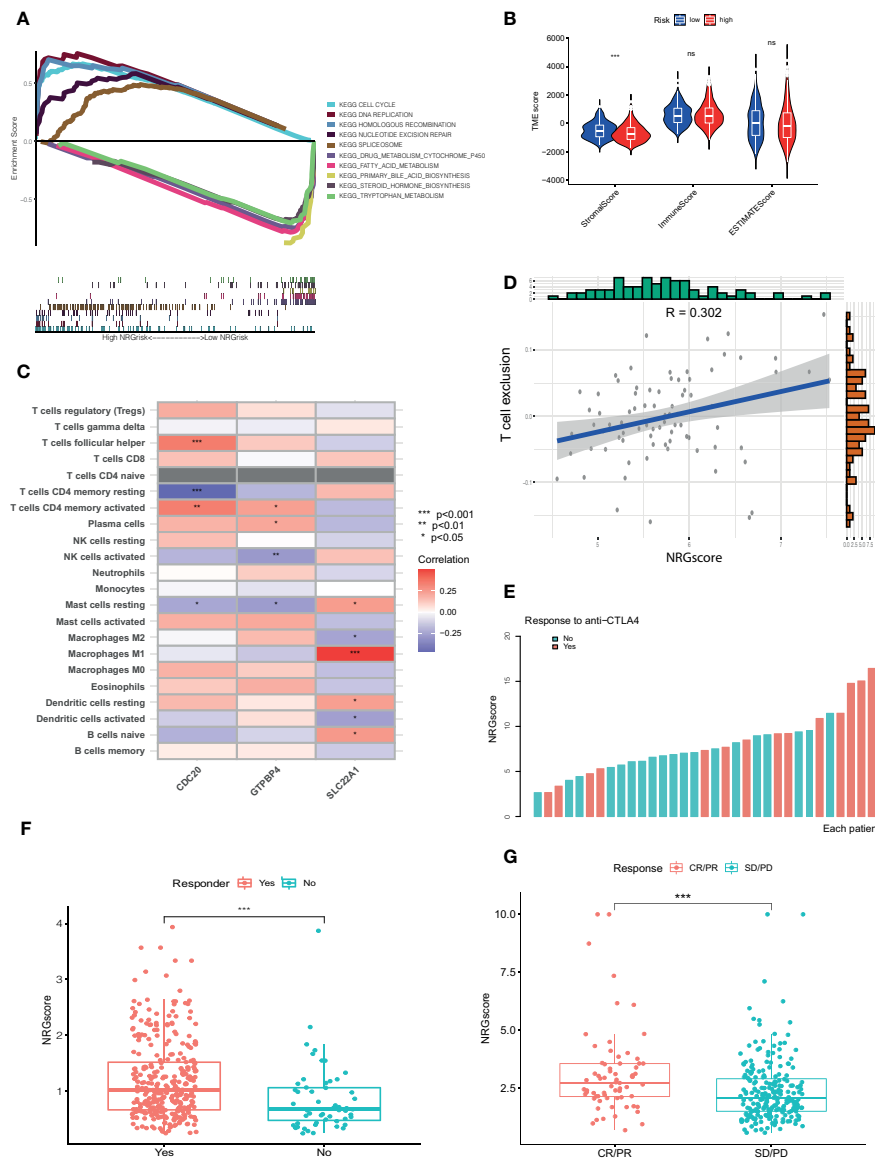


FIGURE 6 | NRGscore could accurately predict response to ICBs. **(A)** Different biological features in high and low NRGrisk groups by GSEA in KEGG pathways. **(B)** Violin diagrams exhibited the correlation of immune, stromal, and ESTIMATE scores with NRGrisk groups. **(C)** Heatmap indicated the relevance of TICs and three hub genes in NRGscore. Different colors represented the different degrees of correlation. **(D)** T cell exclusion score was positively associated with NRGscore in the GSE54236 HCC cohort. **(E)** Waterfall diagram of NRGscore with responder or non-responder to CTLA-4 cohort. **(F)** Patients who respond to ICB possess more NRGscore than non-responders in the TCGA cohort by the TIDE algorithm. **(G)** NRGscore of patients varied with different responses to immunotherapy in the advanced urothelial cancer cohort. CR/PR is short for complete response or partial response; SD/PD represented the stable disease or progressive disease. $P < 0.05$ *; $P < 0.01$ **; $P < 0.001$ ***. ICBs, immune checkpoint blockades; NRGscore, necroptosis-related gene score; TICs, tumor-infiltrating cells; NRGrisk, necroptosis-related gene risk; TIDE, tumor immune dysfunction, and exclusion.

For the present limitation research, the clinical application of immunotherapy in HCC was still not widespread and there are few biomarkers to predict the response to ICBs (23). Therefore, we evaluated the response of checkpoint immunotherapy to patients with different NRGscore. In the cohort of GSE54236, we found NRGscore was positively linked to the level of T cell exclusion, which indicated a high NRGscore with strong

immune inhibition and could respond to ICB therapy (**Figure 6D**). In another cohort of metastatic melanoma, patients with response to CTLA-4 therapy possessed a higher NRGscore (**Figure 6E**). In addition, we also found similar results in the TCGA HCC cohort (**Figure 6F**) and IMvigor210 cohort treated with atezolizumab (**Figure 6G**). IPS values were elevated in high NRGscore subgroups (**Table S7**), which indicated that

the high NRGscore subgroup was relatively responsive to immunotherapy. In other cancer cohorts with ICB therapy, the NRGscore system could also accurately predict ICB response (Figure S10). Collectively, the above results elucidated that the NRGscore system could play a non-negligible role in predicting the immunotherapy response in HCC patients, or even in other tumors.

Patients With High NRGrisk Were More Sensitive to Common Chemotherapy Regimens

As a widely accepted biomarker for prognosis and immunotherapy, tumor mutation burden (TMB) was also evaluated in our study.

Although there were few differences in TMB between high and low NRGrisk subgroups (Figure S11A), and there is no linear correlation of TMB with NRGscore (Figure S11B), the gene mutation type altered in two subgroups in the waterfall curve (Figures 7A, B). Results indicated that the high NRGscore subgroup presented more extensive TMB than the low score subgroup. Among them, TP53 was the first mutated gene with a 39% mutation alteration rate in the high score subgroup, while CTNNB1 was the top rank mutated gene with a 29% mutation alteration rate in the low score subgroup (Figures 7A,B). Consistent with a recent study, CTNNB1, TP53 mutation-associated pathways, and metabolism profiles were identified in HBV-related HCC (55). Besides, we found that the ratio of tumor stem cells in HCC was significantly positively related to NRGscore

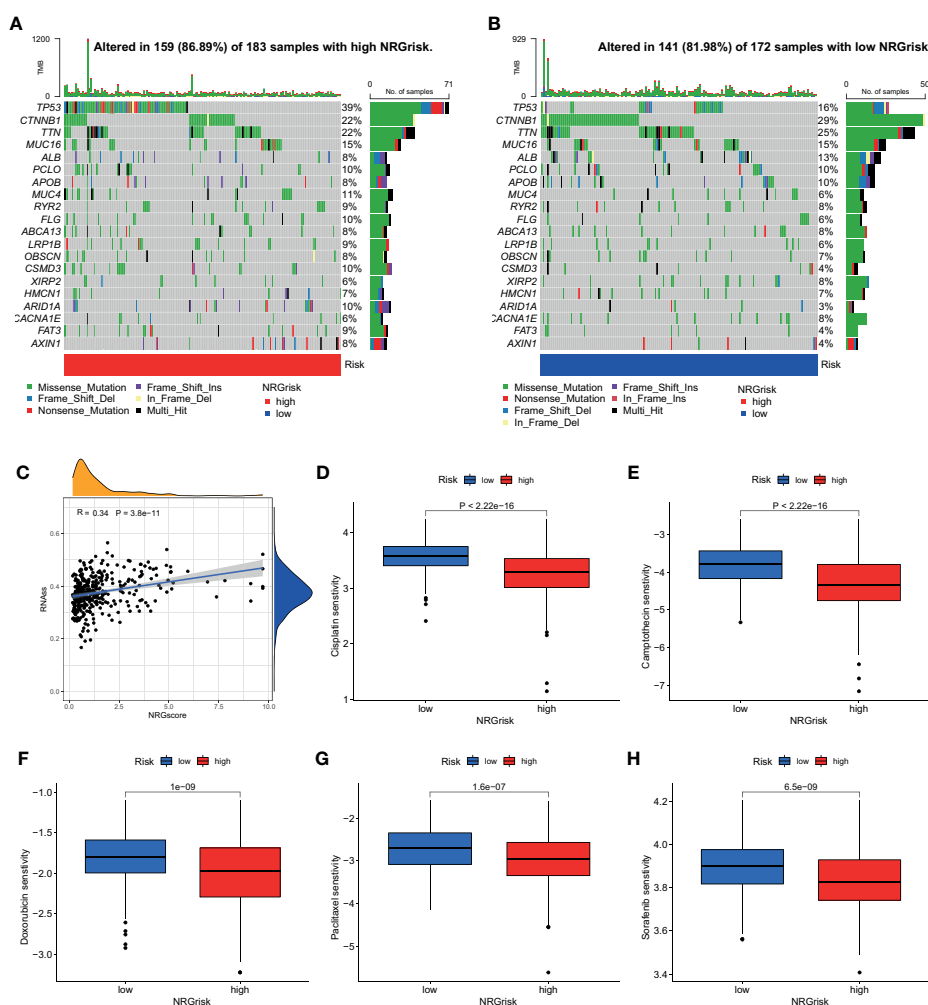


FIGURE 7 | Characteristics of NRGscore with tumor somatic mutation, stem cell, and chemotherapy sensitivity. (A, B) The waterfall plot of tumor somatic mutation was drawn in those with high NRGrisk group (A) and low NRGrisk (B) respectively. Each column indicated individual HCC patients, and the upper bar diagram exhibited TMB. The right number represented the mutation frequency, and the bar diagram on the right exhibited the proportion of each variant type, including missense, nonsense, splice, frameshift, and multiple mutations. (C) Correlation between NRGscore and component of stem cell. (D–H) Box diagrams showed the chemotherapy response between high NRGrisk and low NRGrisk groups: (D) Cisplatin, (E) Camptothecin, (F) Doxorubicin, (G) Paclitaxel, (H) Sorafenib. NRGscore, necroptosis-related gene score; NRGrisk, necroptosis-related gene risk; HCC, hepatocellular carcinoma; TMB, tumor mutation burden; IC50, half-maximal inhibitory concentration.

($R=0.34$) (**Figure 7C**). Considering the frequent use of chemotherapy in the treatment of HCC, we further explored the response of patients with 138 different types of drugs. Analyses of consequences revealed that several prevalent clinical chemotherapies of HCC exhibited low IC50 in the high NRGscore group, including Cisplatin, Camptothecin, Doxorubicin, Paclitaxel, and Sorafenib. The finding suggested that patients with high NRGscore were more sensitive to the treatment of chemotherapy drugs than those with low scores in HCC (**Figures 7D, H**).

DISCUSSION

The NRGscore, a tool designed to evaluate necroptosis patterns of each HCC patient, is a robust biomarker for predicting clinical outcomes and for guiding rational and effective immunotherapy. Our findings revealed a survival benefit trend was observed in Ncluster A compared to B, suggesting necroptosis played a non-negligible role in the progression and prognosis in HCC. GSVA analysis disclosed that Ncluster A had more TME-infiltrating cells and more complex immune-response than Ncluster B, substantiating Ncluster could be a surrogate and favorable biomarker for TME. However, individual necroptosis characteristics could not be elucidated dependent on Ncluster, and we built a novel NRGscore scheme, determined by three core genes (GTPBP4, CDC20, and SLC22A1). NRGscore could accurately and robustly predict 1-, 3-, and 5-year OS of HCC. More importantly, the high NRGscore group had more sensitivity to clinical chemotherapy and ICB-based immunotherapy than the low score group, which demonstrated the NRGscore could be meaningful guidance for HCC treatment.

Mounting evidence demonstrates that necroptosis plays a pivotal role in predicting clinical outcomes of cancer, regulating cancer progression and metastasis, remodeling TME, and thus affecting immunotherapy and chemotherapy (56). Although necroptosis had been shown to perform an antitumor function in cancers, much evidence also demonstrated that necroptosis may play a tumor-promoting role and trigger cancer metastasis (11, 57). The paradox phenomena could be explained by high inflammation TME caused by necroptosis and following elevated ROS levels could further promote cancer progression and metastasis (56). Our results uncovered that higher necroptosis patterns were, worse outcomes did HCC patients have due to most necroptotic regulators being risky genes in HCC. In addition, high NRGscore patients always had more mRNA levels of three necroptosis-driving genes (MLKL, RIPK1, and RIPK3), further validating that necroptosis could play a tumor-promoting role in HCC. In another research on liver cancer, necroptosis directly determines the cancer subtype by driving cell releasing damage-associated molecular patterns (DAMPs) which could reshape the TME (58), causing the switch from HCC to intrahepatic cholangiocarcinoma (ICC) (15). Here, we found that the Ncluster A subgroup had a cline to the high NRGscore group while Ncluster B resembled low NRGscore patients in the alluvial diagram. Ncluster A and high NRGscore were characterized by

more TME-infiltrating cells and enhanced APC-co-inhibitor and checkpoint gene expression, corresponding to immune-inflamed phenotype. Ncluster B and low NRGscore were characterized by enhanced metabolism and decreased stem cells and proliferation, corresponding to immune-exclusion or metabolism phenotype. TIDE results suggested that the high NRGscore group with lower stromal activation could achieve clinical benefits in immunotherapy, consistent with research that emphasized the irreplaceable role of stromal activation in resistance to PD-1/PD-L1 inhibitors (45, 59, 60).

Furthermore, we elucidated the predictive value in another three independent cohorts (HCC cohort with TIDE results; metastatic melanoma cohort treated with Nivolumab; advanced urothelial cancer cohort with the intervention of atezolizumab (IMvigor210)). In line with the results of TCGA HCC, we observed a significant positive trend between NRGscore and T cell exclusion in the GSE54236 HCC cohort. In the metastatic melanoma cohort, responders to anti-CTLA4 possessed a higher NRGscore than non-responders. Similarly, patients' sensitive to anti-PD-L1 had more NRGscore than non-responders. Furthermore, we input our NRGscore system into the TIDE website to predict other cancer cohorts undertaking ICBs and the results revealed that NRGscore could behave well in predicting response. We also found different somatic gene mutations in high and low NRGscore groups and that TP53 mutation was most common in high risk and CTNNB1 for low risk. The results may help us to gain a comprehensive understanding of precision immunotherapy and promoting the necroptosis could enhance the response to ICBs. Intriguingly, we observed that NRGscore could perform as a candidate biomarker for classic chemotherapy and target therapy (Sorafenib, Doxorubicin, Cisplatin, Camptothecin, and Paclitaxel) for HCC.

Additionally, we have validated the pivotal role of three necroptosis-related genes in HCC. It has been widely accepted that the anaphase-promoting complex (APC) drives and governs cell cycle progression (anaphase initiation and late mitosis exit (61, 62)) by interacting with two essential activators - CDC20 and CDH1 (63). However, the two proteins performed different roles in tumor development. CDC20 was considered as an oncogene while CDH1 was deemed as a suppressor in multicancer (64–66). CDC20 has been reported to be overexpressed in HCC tissues and positively related to the tumor, TNM stage, and ki-67 expression (67). In tumorigenesis of HCC, APC/CDC20 could stabilize the HIF-1 α by degrading oxygen-dependent prolyl hydroxylase enzymes3 (PHD3) (68). Suppressing CDC20 (depletion of endogenous or pharmacological inhibitors) in diverse cancer cell lines led to a mitotic arrest and apoptosis, suggesting targeting CDC20 might be a novel anti-cancer therapy, especially in cancer with high expression (69). GTPBP was a GTP-binding protein that is involved in 60S ribosome biogenesis (70). The previous report found that GTPBP4 could reduce TP53 accumulation and increased expression of GTPBP4 correlated with reduced survival (71). In colorectal cancer, GTPBP4 was demonstrated as an oncogene that could disrupt the actin cytoskeleton and promote metastasis of cancer (72). In addition, depletion of

GTPBP4 could inhibit cell proliferation, and high expression correlated with a worse prognosis than the low expression of GTPBP4 in breast cancer (73). In a study of HCC, as an RNA binding protein, GTPBP4 could stratify patients into low- and high-risk, which could predict survival outcomes (74). The upregulated expression of GTPBP4 promotes the proliferation of liver cancer cells and promotes the growth of tumors in mice, while the downregulated expression of GTPBP4 inhibits the proliferation of liver cancer cells and inhibits the growth of tumors in mice (75). SLC22A1 (OCT1) is one member of organic cation transporters, which could uptake intracellular inactivation, a broad spectrum of endogenous and exogenous substrates as well as anticancer drugs (76–78). The OCT1 activity was reported to correlate with the sensitivity of tyrosine kinase inhibitors (TKI) in patients with chronic myeloid leukemia (CML) (79). The downregulation of OCT1 is associated with tumor progression and a worse patient survival (80).

The results of our study should be further validated in a prospective cohort of HCC patients undergoing immunotherapy or chemotherapy. Also, these three hub genes in NRGscore and their relationship with necroptosis should be further explored in basic experiments. Since not all high NRGscore patients would benefit from immunotherapy, more clinical parameters and appropriate evaluations of immune infiltration and TMB should be taken into consideration.

In the current study, we found that necroptosis patterns could become a novel index to predict HCC patients' outcomes and high necroptosis levels might suggest a poor survival prognosis. Ncluster was built based on these necroptosis regulators, and two Nclusters had immense changes in prognosis; cell phagocytosis-, endocytosis-, and peroxisome-related pathways; TME cell infiltration characterization; and immune function. In addition, we designed a novel individual necroptosis pattern evaluation system: NRGscore scheme. NRGscore could accurately and robustly predict 1-, 3-, and 5-year OS, which was validated in another HCC cohort. The novel nomogram was comprised of NRGscore and other clinicopathological features to visualize the risk of HCC patients. In ICB-based immunotherapy of HCC, NRGscore had meaningful guidance for patients and the high NRGscore group could be more sensitive to ICB treatment, which was further validated in the metastatic melanoma cohort and advanced urothelial cancer cohort. The high NRGscore group had an increased ratio of cancer stem cells than its counterpart, which further demonstrates high NRGscore could be more sensitive to chemotherapy and targeted molecular therapy.

REFERENCES

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2018) 68:394–424. doi: 10.3322/caac.21492
- Kulik L, El-Serag HB. Epidemiology and Management of Hepatocellular Carcinoma. *Gastroenterology* (2019) 156:477–491.e1. doi: 10.1053/j.gastro.2018.08.065
- Singal AG, Lampertico P, Nahon P. Epidemiology and Surveillance for Hepatocellular Carcinoma: New Trends. *J Hepatol* (2020) 72:250–61. doi: 10.1016/j.jhep.2019.08.025
- McGlynn KA, Petrick JL, El-Serag HB. Epidemiology of Hepatocellular Carcinoma. *Hepatol (Baltimore Md.)* (2021) 73(Suppl 1):4–13. doi: 10.1002/hep.31288
- Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to Build a Bridge From a Population-Based to a More "Personalized" Approach to Cancer Staging. *CA: Cancer J Clin* (2017) 67:93–9. doi: 10.3322/caac.21388
- Bruix J, Reig M, Sherman M. Evidence-Based Diagnosis, Staging, and Treatment of Patients With Hepatocellular Carcinoma. *Gastroenterology* (2016) 150:835–53. doi: 10.1053/j.gastro.2015.12.041
- Llovet JM, Zucman-Rossi J, Pikarsky E, Sangro B, Schwartz M, Sherman M, et al. Hepatocellular Carcinoma. *Nat Rev Dis Primers* (2016) 2:16018. doi: 10.1038/nrdp.2016.18

CONCLUSIONS

In the study, we found high expression of necroptosis could affect the TME of HCC and be closely related to the immune-inflamed phenotype. In predicting response to ICB, we concluded that high NRGrisk patients could be more sensitive than low NRGrisk. We systematically characterized the landscapes of necroptosis in HCC patients and suggested that NRGscore could be a prognostic marker and help to interpret the responses of HCC to chemotherapies and immunotherapies, providing new strategies for the treatment of cancers.

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

LL and YX designed the study. JZ and TH performed most of the results and completed the manuscript together. SZ, YZ, and SM helped with all experiments and data analysis. FX, TB, and YT helped write the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (8217113876, 62171365).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2022.933210/full#supplementary-material>

8. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2019. *CA: Cancer J Clin* (2019) 69:7–34. doi: 10.3322/caac.21551
9. Degterev A, Hitomi J, Gerscheid M, Ch'en IL, Korkina O, Teng X, et al. Identification of RIP1 Kinase as a Specific Cellular Target of Necrostatins. *Nat Chem Biol* (2008) 4:313–21. doi: 10.1038/nchembio.83
10. Su Z, Yang Z, Xu Y, Chen Y, Yu Q. Apoptosis, Autophagy, Necroptosis, and Cancer Metastasis. *Mol Cancer* (2015) 14:48. doi: 10.1186/s12943-015-0321-5
11. Liu X, Zhou M, Mei L, Ruan J, Hu Q, Peng J, et al. Key Roles of Necroptotic Factors in Promoting Tumor Growth. *Oncotarget* (2016) 7:22219–33. doi: 10.18632/oncotarget.7924
12. Najafav A, Chen H, Yuan J. Necroptosis and Cancer. *Trends Cancer* (2017) 3:294–301. doi: 10.1016/j.trecan.2017.03.002
13. Moriawaki K, Balaji S, McQuade T, Malhotra N, Kang J, Chan FK. The Necroptosis Adaptor RIPK3 Promotes Injury-Induced Cytokine Expression and Tissue Repair. *Immunity* (2014) 41:567–78. doi: 10.1016/j.immuni.2014.09.016
14. Stoll G, Ma Y, Yang H, Kepp O, Zitvogel L, Kroemer G. Pro-Necrotic Molecules Impact Local Immunosurveillance in Human Breast Cancer. *Oncimmunology* (2017) 6:e1299302. doi: 10.1080/2162402X.2017.1299302
15. Seehawer M, Heinzmann F, D'Artista L, Harbig J, Roux PF, Hoenicke L, et al. Necroptosis Microenvironment Directs Lineage Commitment in Liver Cancer. *Nature* (2018) 562:69–75. doi: 10.1038/s41586-018-0519-y
16. Finn RS, Qin S, Ikeda M, Galle PR, Ducreux M, Kim TY, et al. Atezolizumab Plus Bevacizumab in Unresectable Hepatocellular Carcinoma. *N Engl J Med* (2020) 382:1894–905. doi: 10.1056/NEJMoa1915745
17. Yau T, Park JW, Finn RS, Cheng AL, Mathurin P, Edeline J, et al. CheckMate 459: A Randomized, Multi-Center Phase III Study of Nivolumab (NIVO) vs Sorafenib (SOR) as First-Line (1L) Treatment in Patients (Pts) With Advanced Hepatocellular Carcinoma (aHCC). *Ann Oncol* (2019) 30:874–+. doi: 10.1093/annonc/mdz394.029
18. Zhu AX, Finn RS, Cattani S, Edeline J, Ogasawara S, Palmer DH, et al. KEYNOTE-224: Pembrolizumab in Patients With Advanced Hepatocellular Carcinoma Previously Treated With Sorafenib. *J Clin Oncol* 36 (2018) 36:209–209. doi: 10.1200/JCO.2018.36.4_suppl.209
19. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic Correlates of Response to CTLA-4 Blockade in Metastatic Melanoma. *Science* (2015) 350:207–11. doi: 10.1126/science.aad0095
20. Havel JJ, Chowell D, Chan TA. The Evolving Landscape of Biomarkers for Checkpoint Inhibitor Immunotherapy. *Nat Rev Cancer* (2019) 19:133–50. doi: 10.1038/s41568-019-0116-x
21. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 Blockade in Tumors With Mismatch-Repair Deficiency. *New Engl J Med* (2015) 372:2509–20. doi: 10.1056/NEJMoa1500596
22. Ruiz de Galarreta M, Bresnahan E, Molina-Sánchez P, Lindblad KE, Maier B, Sia D, et al. β -Catenin Activation Promotes Immune Escape and Resistance to Anti-PD-1 Therapy in Hepatocellular Carcinoma. *Cancer Discovery* (2019) 9:1124–41. doi: 10.1158/2159-8290.CD-19-0074
23. Pinter M, Jain RK, Duda DG. The Current Landscape of Immune Checkpoint Blockade in Hepatocellular Carcinoma: A Review. *JAMA Oncol* (2021) 7:113–23. doi: 10.1001/jamaoncol.2020.3381
24. Tang R, Xu J, Zhang B, Liu J, Liang C, Hua J, et al. Ferroptosis, Necroptosis, and Pyroptosis in Anticancer Immunity. *J Hematol Oncol* (2020) 13:110. doi: 10.1186/s13045-020-00946-7
25. Kaczmarek A, Vandenabeele P, Krysko DV. Necroptosis: The Release of Damage-Associated Molecular Patterns and its Physiological Relevance. *Immunity* (2013) 38:209–23. doi: 10.1016/j.immuni.2013.02.003
26. Aaes TL, Kaczmarek A, Delvaeye T, De Craene B, De Koker S, Heyndrickx L, et al. Vaccination With Necroptotic Cancer Cells Induces Efficient Anti-Tumor Immunity. *Cell Rep* (2016) 15:274–87. doi: 10.1016/j.celrep.2016.03.037
27. Villa E, Critelli R, Lei B, Marzocchi G, Cammà C, Giannelli G, et al. Neoangiogenesis-Related Genes are Hallmarks of Fast-Growing Hepatocellular Carcinomas and Worst Survival. *Results prospective study Gut* (2016) 65:861–9. doi: 10.1136/gutjnl-2014-308483
28. Zubiete-Franco I, García-Rodríguez JL, Lopitz-Otsoa F, Serrano-Macia M, Simon J, Fernández-Tussy P, et al. SUMOylation Regulates LKB1 Localization and its Oncogenic Activity in Liver Cancer. *EBioMedicine* (2019) 40:406–21. doi: 10.1016/j.ebiom.2018.12.031
29. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: Efficient and Comprehensive Analysis of Somatic Variants in Cancer. *Genome Res* (2018) 28:1747–56. doi: 10.1101/gr.239244.118
30. Glickman ME, Rao SR, Schultz MR. False Discovery Rate Control is a Recommended Alternative to Bonferroni-Type Adjustments in Health Studies. *J Clin Epidemiol* (2014) 67:850–7. doi: 10.1016/j.jclinepi.2014.03.012
31. Wilkerson MD, Hayes DN. ConsensusClusterPlus: A Class Discovery Tool With Confidence Assessments and Item Tracking. *Bioinformatics* (2010) 26:1572–3. doi: 10.1093/bioinformatics/btq170
32. Hänzelmann S, Castelo R, Guinney J. GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinf* (2013) 14:7. doi: 10.1186/1471-2105-14-7
33. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells With CIBERSORT. *Methods Mol Biol (Clifton N.J.)* (2018) 1711:243–59. doi: 10.1007/978-1-4939-7493-1_12
34. Zhu J, Zhou Y, Wang L, Hao J, Chen R, Liu L, et al. CXCL5/CXCL8 is a Promising Potential Prognostic and Tumor Microenvironment-Related Cluster in Hepatocellular Carcinoma. *J Gastrointest Oncol* (2020) 11:1364–80. doi: 10.21037/jgo-20-556
35. Zhu J, Wang L, Zhou Y, Hao J, Wang S, Liu L, et al. Comprehensive Analysis of the Relationship Between Competitive Endogenous RNA (ceRNA) Networks and Tumor Infiltrating-Cells in Hepatocellular Carcinoma. *J gastrointestinal Oncol* (2020) 11:1381–98. doi: 10.21037/jgo-20-555
36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc Natl Acad Sci United States America* (2005) 102:15545–50. doi: 10.1073/pnas.0506580102
37. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res* (2015) 43:e47. doi: 10.1093/nar/gkv007
38. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, et al. Signatures of T Cell Dysfunction and Exclusion Predict Cancer Immunotherapy Response. *Nat Med* (2018) 24:1550–8. doi: 10.1038/s41591-018-0136-1
39. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-Cancer Immunogenomic Analyses Reveal Genotype-Immune-phenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep* (2017) 18:248–62. doi: 10.1016/j.celrep.2016.12.019
40. Gajewski TF, Schreiber H, Fu YX. Innate and Adaptive Immune Cells in the Tumor Microenvironment. *Nat Immunol* (2013) 14:1014–22. doi: 10.1038/ni.2703
41. Joyce JA, Fearon DT. T Cell Exclusion, Immune Privilege, and the Tumor Microenvironment. *Science* (2015) 348:74–80. doi: 10.1126/science.aaa6204
42. García-Mulero S, Alonso MH, Pardo J, Santos C, Sanjuan X, Salazar R, et al. Lung Metastases Share Common Immune Features Regardless of Primary Tumor Origin. *J Immunother Cancer* (2020) 8(1). doi: 10.1136/jitc-2019-000491
43. Necchi A, Joseph RW, Loriot Y, Hoffman-Censits J, Perez-Gracia JL, Petrylak DP, et al. Atezolizumab in Platinum-Treated Locally Advanced or Metastatic Urothelial Carcinoma: Post-Progression Outcomes From the Phase II IMvigor210 Study. *Ann Oncol* (2017) 28:3044–50. doi: 10.1093/annonc/mdx518
44. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* (2017) 168:542. doi: 10.1016/j.cell.2017.01.010
45. Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y, et al. Tgfb Attenuates Tumour Response to PD-L1 Blockade by Contributing to Exclusion of T Cells. *Nature* (2018) 554:544–8. doi: 10.1038/nature25501
46. Gleeleher P, Cox N, Huang RS. Prorhetic: An R Package for Prediction of Clinical Chemotherapeutic Response From Tumor Gene Expression Levels. *PLoS One* (2014) 9:e107468. doi: 10.1371/journal.pone.0107468
47. Carafa V, Nebbioso A, Cuomo F, Rotili D, Cobellis G, Bontempo P, et al. RIP1-HAT1-SIRT Complex Identification and Targeting in Treatment and Prevention of Cancer. *Clin Cancer Res an Off J Am Assoc Cancer Res* (2018) 24:2886–900. doi: 10.1158/1078-0432.CCR-17-3081
48. Schneider AT, Gautheron J, Feoktistova M, Roderburg C, Loosen SH, Roy S, et al. RIPK1 Suppresses a TRAF2-Dependent Pathway to Liver Cancer. *Cancer Cell* (2017) 31:94–109. doi: 10.1016/j.ccell.2016.11.009

49. Xiong J, Wang L, Fei XC, Jiang XF, Zheng Z, Zhao Y, et al. MYC is a Positive Regulator of Choline Metabolism and Impedes Mitophagy-Dependent Necroptosis in Diffuse Large B-Cell Lymphoma. *Blood Cancer J* (2017) 7:e0. doi: 10.1038/bcj.2017.61
50. Nicolai S, Pieraccioli M, Peschiaroli A, Melino G, Raschella G. Neuroblastoma: Oncogenic Mechanisms and Therapeutic Exploitation of Necroptosis. *Cell Death Dis* (2015) 6:e2010. doi: 10.1038/cddis.2015.354
51. Chen W, Lin W, Wu L, Xu A, Liu C, Huang P. A Novel Prognostic Predictor of Immune Microenvironment and Therapeutic Response in Kidney Renal Clear Cell Carcinoma Based on Necroptosis-Related Gene Signature. *Int J Med Sci* (2022) 19:377–92. doi: 10.7150/ijms.69060
52. Dey A, Mustafi SB, Saha S, Kumar Dhar Dwivedi S, Mukherjee P, Bhattacharya R. Inhibition of BMI1 Induces Autophagy-Mediated Necroptosis. *Autophagy* (2016) 12:659–70. doi: 10.1080/15548627.2016.1147670
53. Nicolè L, Sanavia T, Cappellesso R, Maffei V, Akiba J, Kawahara A, et al. Necroptosis-Driving Genes RIPK1, RIPK3 and MLKL-P are Associated With Intratumoral CD3(+) and CD8(+) T Cell Density and Predict Prognosis in Hepatocellular Carcinoma. *J Immunother Cancer* (2022) 10(3). doi: 10.1136/jitc-2021-004031
54. Ruf B, Heinrich B, Greten TF. Immunobiology and Immunotherapy of HCC: Spotlight on Innate and Innate-Like Immune Cells. *Cell Mol Immunol* (2021) 18:12–27. doi: 10.1038/s41423-020-00572-w
55. Gao Q, Zhu H, Dong L, Shi W, Chen R, Song Z, et al. Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. *Cell* (2019) 179:561–77.e22. doi: 10.1016/j.cell.2019.08.052
56. Gong Y, Fan Z, Luo G, Yang C, Huang Q, Fan K, et al. The Role of Necroptosis in Cancer Biology and Therapy. *Mol Cancer* (2019) 18:100. doi: 10.1186/s12943-019-1029-8
57. Seifert L, Werba G, Tiwari S, Gao LY NN, Allothman S, Alqunaibit D, et al. The Necrosome Promotes Pancreatic Oncogenesis via CXCL1 and Mincle-Induced Immune Suppression. *Nature* (2016) 532:245–9. doi: 10.1038/nature17403
58. Pasparakis M, Vandenabeele P. Necroptosis and its Role in Inflammation. *Nature* (2015) 517:311–20. doi: 10.1038/nature14191
59. Ravi R, Noonan KA, Pham V, Bedi R, Zhavoronkov A, Ozerov IV, et al. Bifunctional Immune Checkpoint-Targeted Antibody-Ligand Traps That Simultaneously Disable Tgβ Enhance the Efficacy of Cancer Immunotherapy. *Nat Commun* (2018) 9:741. doi: 10.1038/s41467-017-02696-6
60. Nagarsheth N, Wicha MS, Zou W. Chemokines in the Cancer Microenvironment and Their Relevance in Cancer Immunotherapy. *Nat Rev Immunol* (2017) 17:559–72. doi: 10.1038/nri.2017.49
61. Fung TK, Poon RY. A Roller Coaster Ride With the Mitotic Cyclins. *Semin Cell Dev Biol* (2005) 16:335–42. doi: 10.1016/j.semcdb.2005.02.014
62. Weinstein J, Jacobsen FW, Hsu-Chen J, Wu T, Baum LG. A Novel Mammalian Protein, P55cdc, Present in Dividing Cells is Associated With Protein Kinase Activity and has Homology to the Saccharomyces Cerevisiae Cell Division Cycle Proteins Cdc20 and Cdc4. *Mol Cell Biol* (1994) 14:3350–63. doi: 10.1128/mcb.14.5.3350-3363.1994
63. Wang Z, Wan L, Zhong J, Inuzuka H, Liu P, Sarkar FH, et al. Cdc20: A Potential Novel Therapeutic Target for Cancer Treatment. *Curr Pharm Des* (2013) 19:3210–4. doi: 10.2174/1381612811319180005
64. Zhou Z, He M, Shah AA, Wan Y. Insights Into APC/C: From Cellular Function to Diseases and Therapeutics. *Cell Div* (2016) 11:9. doi: 10.1186/s13008-016-0021-6
65. Chang DZ, Ma Y, Ji B, Liu Y, Hwu P, Abbruzzese JL, et al. Increased CDC20 Expression is Associated With Pancreatic Ductal Adenocarcinoma Differentiation and Progression. *J Hematol Oncol* (2012) 5:15. doi: 10.1186/1756-8722-5-15
66. Kim JM, Sohn HY, Yoon SY, Oh JH, Yang JO, Kim JH, et al. Identification of Gastric Cancer-Related Genes Using a cDNA Microarray Containing Novel Expressed Sequence Tags Expressed in Gastric Cancer Cells. *Clin Cancer Res* (2005) 11:473–82. doi: 10.1158/1078-0432.473.11.2
67. Li J, Gao JZ, Du JL, Huang ZX, Wei LX. Increased CDC20 Expression is Associated With Development and Progression of Hepatocellular Carcinoma. *Int J Oncol* (2014) 45:1547–55. doi: 10.3892/ijo.2014.2559
68. Shi M, Dai WQ, Jia RR, Zhang QH, Wei J, Wang YG, et al. APC(CDC20)-Mediated Degradation of PHD3 Stabilizes HIF-1α and Promotes Tumorigenesis in Hepatocellular Carcinoma. *Cancer Lett* (2021) 496:144–55. doi: 10.1016/j.canlet.2020.10.011
69. Wang L, Zhang J, Wan L, Zhou X, Wang Z, Wei W. Targeting Cdc20 as a Novel Cancer Therapeutic Strategy. *Pharmacol Ther* (2015) 151:141–51. doi: 10.1016/j.pharmthera.2015.04.002
70. Fuentes JL, Datta K, Sullivan SM, Walker A, Maddock JR. In Vivo Functional Characterization of the Saccharomyces Cerevisiae 60S Biogenesis GTPase Nog1. *Mol Genet Genomics* (2007) 278:105–23. doi: 10.1007/s00438-007-0233-1
71. Lunardi A, Di Minin G, Provero P, Dal Ferro M, Carotti M, Del Sal G, et al. A Genome-Scale Protein Interaction Profile of Drosophila P53 Uncovers Additional Nodes of the Human P53 Network. *Proc Natl Acad Sci U.S.A.* (2010) 107:6322–7. doi: 10.1073/pnas.1002447107
72. Yu H, Jin S, Zhang N, Xu Q. Up-Regulation of GTPBP4 in Colorectal Carcinoma is Responsible for Tumor Metastasis. *Biochem Biophys Res Commun* (2016) 480:48–54. doi: 10.1016/j.bbrc.2016.10.010
73. Pawitan Y, Bjöhle J, Amler L, Borg AL, Eghazi S, Hall P, et al. Gene Expression Profiling Spares Early Breast Cancer Patients From Adjuvant Therapy: Derived and Validated in Two Population-Based Cohorts. *Breast Cancer Res* (2005) 7:R953–64. doi: 10.1186/bcr1325
74. Wang L, Zhang Z, Li Y, Wan Y, Xing B. Integrated Bioinformatic Analysis of RNA Binding Proteins in Hepatocellular Carcinoma. *Aging (Albany NY)* (2020) 13:2480–505. doi: 10.18632/aging.202281
75. Chen J, Zhang J, Zhang Z. Upregulation of GTPBP4 Promotes the Proliferation of Liver Cancer Cells. *J Oncol* (2021) 2021:1049104. doi: 10.1155/2021/1049104
76. Zhang S, Lovejoy KS, Shima JE, Lagpagan LL, Shu Y, Lapuk A, et al. Organic Cation Transporters are Determinants of Oxaliplatin Cytotoxicity. *Cancer Res* (2006) 66:8847–57. doi: 10.1158/0008-5472.CAN-06-0769
77. Jonker JW, Schinkel AH. Pharmacological and Physiological Functions of the Polyspecific Organic Cation Transporters: OCT1, 2, and 3 (SLC22A1-3). *J Pharmacol Exp Ther* (2004) 308:2–9. doi: 10.1124/jpet.103.053298
78. Thomas J, Wang L, Clark RE, Pirmohamed M. Active Transport of Imatinib Into and Out of Cells: Implications for Drug Resistance. *Blood* (2004) 104:3739–45. doi: 10.1182/blood-2003-12-4276
79. Crossman LC, Druker BJ, Deininger MW, Pirmohamed M, Wang L, Clark RE. hOCT 1 and Resistance to Imatinib. *Blood* (2005) 106:1133–4. doi: 10.1182/blood-2005-02-0694
80. Heise M, Lautem A, Knapstein J, Schattenberg JM, Hoppe-Lotichius M, Foltys D, et al. Downregulation of Organic Cation Transporters OCT1 (SLC22A1) and OCT3 (SLC22A3) in Human Hepatocellular Carcinoma and Their Prognostic Significance. *BMC Cancer* (2012) 12:109. doi: 10.1186/1471-2407-12-109

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.

Copyright © 2022 Zhu, Han, Zhao, Zhu, Ma, Xu, Bai, Tang, Xu and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Pathogens and Pathogenesis in Wheezing Diseases in Children Under 6

Yongjun Tang¹, Yaxiong Yang¹, Ruohui He², Rong Huang^{1*},
Xiangrong Zheng¹ and Chentao Liu¹

¹ Department of Pediatrics, Xiangya Hospital, Central South University, Changsha, China, ² Department of Pharmacy, Ningyuan County of People's Hospital, Yongzhou, China

OPEN ACCESS

Edited by:

Liang Cheng,
Harbin Medical University, China

Reviewed by:

Guang Yang,
People's Liberation Army General
Hospital, China
Jingpu Zhang,
Henan University of Urban
Construction, China

*Correspondence:

Rong Huang
hrong61@csu.edu.cn

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 17 April 2022

Accepted: 17 June 2022

Published: 14 July 2022

Citation:

Tang Y, Yang Y, He R, Huang R,
Zheng X and Liu C (2022) Pathogens
and Pathogenesis in Wheezing
Diseases in Children Under 6.
Front. Oncol. 12:922214.
doi: 10.3389/fonc.2022.922214

Few studies have comprehensively assessed the roles of cytokine production in wheezing pathogenesis. Therefore, we undertook this study to determine the association between wheezing episodes and cytokines, and to provide further information on this topic. Firstly, we retrospectively collected 1176 children, including 122 subjects with first wheezing and 54 subjects with recurrent wheezing, to analyze the etiology and clinical characteristics of children with wheezing diseases. Then, we collected 52 children with wheezing diseases and 25 normal controls to detect the expression of interferon- γ (IFN- γ), interleukin-4 (IL-4), IFN- γ /IL-4, IL-17A, IL-17E, IgE, matrix metalloproteinase-3 (MMP-3), and MMP-9 in serum or plasma. The results showed that boys under 3 years old with history of allergies were more likely to develop wheezing diseases. In our cohort, *M. pneumoniae* caused a greater proportion of wheezing in children than expected. The expression of IgE [18.80 (13.65-31.00) vs. 17.9 (10.15-21.60)], IL-4 [24.00 (24.00-48.00) vs. 23.00 (9.50-27.00)], IFN- γ [70.59 (41.63-116.46) vs. 49.83 (29.58-81.74)], MMP3 [53.40 (20.02-128.2) vs. 30.90 (13.80-50.95)], MMP9 [148.10 (99.30-276.10) vs. 122.10 (82.20-162.35)], IL-17A [80.55 (54.46-113.08) vs. 61.11 (29.43-93.87)], and IL-17E [1.75 (0.66-2.77) vs. 1.19 (0.488-2.1615)] were significantly increased in the wheezing group ($p < 0.05$) compared to normal controls, while the level of IFN- γ /IL-4 had no significant difference between the two groups (1.24 ± 1.88 vs 0.68 ± 0.74 , $p > 0.05$). There was altered cytokine production in children with wheezing diseases which was quite similar to asthma pathogenesis. Sex, age, pathogen infection, and inflammation in our study were also risk factors for wheezing diseases.

Keywords: wheezing, infant, IgE, IL-4, IFN- γ , MMP, IL-17

INTRODUCTION

Wheezing respiratory illness is one of the most common diseases of childhood, with waves of prevalence in winter and summer. Wheezing infants always presented clinical manifestations such as a cough, fever, or shadows in chest X-rays. Approximately one-third of children under 3 years of age had had wheezing experiences, and over half of the participants who had been hospitalized for wheezing episodes had asthma in young adulthood (1, 2). Recurrent wheezing episodes greatly

influence the maturation of children's immune and respiratory systems, increase the risk of developing asthma in the future, and place a heavy burden on families and society.

There are many causes of wheezing diseases, such as infections, allergies, congenital anatomical deformities, and environmental and genetic factors. Infections can be caused by viruses, mycoplasmas (MP), bacteria, fungi, and other pathogens alone or in combination. Wheezing caused by different etiologies show different disease rules, and corresponding treatment plans and prognoses are also different. Martinez et al. (3) enrolled 725 children and found that patients with an allergic disease were more likely to develop another kind of allergic disease. Van's follow-up study of a total of 6491 patients with allergic rhinitis underwent in 8.4 years showed that allergic rhinitis was an independent risk factor for asthma as patients with allergic rhinitis were five times more likely to develop asthma than people without (4). Ronmark and his colleagues carried out a survey with a questionnaire with 30,000 subjects in Switzerland that suggested that the co-existence of asthma, rhinitis, and eczema was common. Allergies, a history of asthma in the family, and smoking were all risk factors for eczema and the patients were characterized by wheezing. Therefore, finding out the risk factors of children with wheezing diseases is very important.

Several studies have underlined the association between pathogen infections and wheezing episodes, as children who had a virus, MP, bacteria, or fungal infection are more vulnerable to developing wheezing diseases (5). The altered cytokine production was also reported to be involved in the pathogenesis of MP-associated asthma. Hahn and his colleagues (6) identified that one-third of newborn infants with MP infection will develop asthma and the MP-specific IgE was detected in blood, nasopharyngeal, and bronchial secretion in half of the children with asthma. Another study undertaken on mice showed that bronchial hyperresponsiveness was alleviated on the third day of MP infection. On the fifth day, the ratio of interferon- γ (IFN- γ)/interleukin-4(IL-4) increased, but it decreased during 9-16 days indicating that the helper T lymphocyte 1 (Th1) cells played a major role at the initial stage of infection and Th2 cells at the later stage (7). The imbalance of CD4 helper T cell (Th) function was thought to be the most important mechanism, in which Th1 specific IFN- γ was decreased while the Th2 specific cytokine, IL-4, was observed to be increased. IL-4 functioned to enhance the proliferation of B lymphocytes and activate the synthesis of IgE in serum which was the mediator of rapid-type allergy reaction. Besides, the level of other cytokines like matrix metalloproteinase-3 (MMP-3), matrix metalloproteinase-9 (MMP-9), and interleukin-17 (IL-17) were reported to participate in Th1/Th2 imbalance. However, whether they were involved in the pathogenesis of wheezing diseases was still unknown (8, 9).

To understand the differences and provide suggestions for the treatment of wheezing in children with different phenotypes, we conducted a retrospective study on 176 infants with wheezing diseases enrolled from Pediatrics department of Xiangya Hospital. In addition, we collected 52 subjects and 25 control subjects who displayed no wheezing from Xiangya Hospital and

Hunan Children's Hospital and aimed to reveal the association between cytokines (including IgE, IL-4, IFN- γ , IL-4/IFN- γ , MMP3, MMP9, IL-17A, and IL-17E) and wheezing illness.

MATERIALS AND METHODS

Study Population

Cases From Xiangya Hospital

We recruited 176 children with wheezing diseases between September 2013 and April 2014 in the department of pediatrics, Xiangya Hospital. To be included in the study, had to have been hospitalized and diagnosed with asthmatic bronchitis and asthmatic pneumonia (10). Subjects were excluded from the study if: ① they had been born prematurely or born with respiratory malformation and heart disease; ② they had a primary immunodeficiency disease; ③ they had used hormones or immunosuppressive drugs less than 4 weeks before admission. The patients were aged from 1 month to 6 years (Figure 1).

Fifty-two subjects with wheezing illness were collected between December 2014 and February 2015. The inclusion and exclusion criteria were similar to the above. Twenty-five children who displayed no wheezing symptoms were enrolled as a control in this study. The control was defined as infants who had no history of wheezing, allergies, recurrent respiratory infections, use of glucocorticoids and immunosuppressive agents, other severe diseases, or disability.

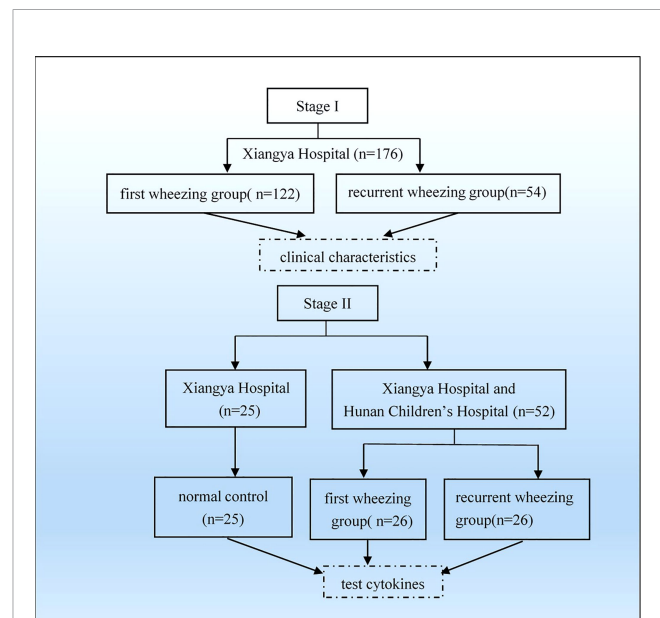


FIGURE 1 | Flowchart of grouping and analysis. Firstly, we collected 176 patients from Xiangya Hospital and grouped as first wheezing ($n = 122$) and recurrent wheezing ($n = 54$) to analyze clinical characteristics. Then we collected 52 patients from Xiangya Hospital and Hunan Children's Hospital and 25 normal control from Xiangya Hospital to test cytokines.

Data

The questionnaire was designed to collect clinical data including name, sex, age, telephone number, admission time, clinical symptoms, allergy history of the individual and their family, auxiliary measure of inspection, chest X-rays, and the treatment programs. A chest CT was conducted on specific patients.

Viral Diagnostics

Blood samples were analyzed for respiratory viruses including coxsackie virus, respiratory syncytial virus (RSV), cytomegalovirus (CMV), Epstein-Barr virus (EB virus), rubella virus, and adenovirus by using standard techniques. The presence of IgM antibodies was considered indicative of viral infection.

Bacterial Diagnostics

Bacterial antibodies were performed on serum, bone marrow, and sputum samples. Details of bacterial culture were reported previously and the species were identified according to standard methods (11).

MP Diagnostics

Specimens were obtained from subjects using throat swabs and serum. *M. pneumoniae* was measured as previously reported and the patients were diagnosed with MP infection when anti-mycoplasma antibody titers were $\geq 1:80$ (12).

Cytokine Assays

IL-4, IFN- γ , and IL-17 were detected on frozen serum samples using Human Cytokine/Chemokine Magnetic Bead Panel (Millipore, USA, Billerica, Massachusetts). The well was pre-wet in 200 μ L Assay Buffer, then mixed on a plate shaker for 10 min. We removed the wash buffer thoroughly and added 25 μ L of control or standard to the wells regarding Assay buffer as 0 pg/ml standard. 25 μ L was added to the wells of samples. Next, we add 25 μ L serum matrix to the background, standards, and control wells and 25 μ L serum sample to sample wells. We vortexed the mixing bottle and added 25 μ L of the mixed beads to each well. Hereafter, we sealed and wrapped the plates and incubated them at 4°C for 16–18 h. We removed the well content and washed the plate with 200 μ L Assay buffer per time and added 25 μ L detection antibodies afterward. The plates were incubated at room temperature for 1 h and 25 μ L Streptavidin-Phycoerythrin was added per well and then incubated at room temperature. The well content was removed and washed with 200 μ L wash buffer two times then 150 μ L sheath fluid or drive fluid was added per well, and the concentrations of cytokine were

detected on Luminex® 200™, HTS, FLEXMAP 3D®, or MAGPIX® with xPONENT® software.

MMP3 and MMP9 Measurements

The concentrations of MMP3 and MMP9 in serum or plasma samples were measured using enzyme immunoassay kits (R&D Systems, USA). Assays were performed following the manufacturer's protocol. The data were analyzed using the standard curve-fitting method for calculating MMP3 and MMP9 concentrations in samples.

Data Analysis

SPSS 26.0 statistical software was used for data analysis. Data are shown as mean \pm SD. χ^2 test for proportions was used to assess the comparability between different sex and age groups. The data of IgE, IL-4, IFN- γ , MMP3, MMP9, IL-17A, and IL-17E were analyzed using the Rank sum test. The comparability of IL-4/IFN- γ between control and wheezing patients used an analysis of variance. Values of $p < 0.05$ are considered significant.

RESULTS

Characteristics of Children With Wheezing

A total of 176 children were enrolled in Xiangya Hospital cases, including 129 males and 47 females. The ratios of male to female in the first wheezing group and the recurrent group were 2.6:1 and 3.2:1, respectively. Boys in the first wheezing group and the recurrent wheezing group had more morbidity than girls, but there was no significant difference in gender between the two groups ($p > 0.05$, **Table 1**).

The cases from Xiangya Hospital and Hunan Children's Hospital included 26 first wheezing patients (19 males and seven females) with an average age of 11.35 ± 8.91 months, recurrent wheezing children (21 males and five females) with an average age of 13.96 ± 9.67 months. As for the normal control group, there were 19 male children and six female children with an average age of 13.92 ± 9.67 months. There is no significant difference in gender and age ratio in cases from Xiangya Hospital and cases from Xiangya Hospital and Hunan Children's Hospital ($p > 0.05$, **Table 2**).

Association Between Allergy History and Wheezing

Fifty-one infants in cases from Xiangya Hospital had allergy history (mainly referred to as eczema). Among them, one was diagnosed with allergic rhinitis, one had urticaria, one was

TABLE 1 | Clinical data of 176 wheezing children from Xiangya Hospital.

	Total	First wheezing	Recurrent wheezing	P-value
Subjects (M/F)	176 (129/47)	122 (88/34)	54 (41/13)	0.600
Positive history of allergy	51	26 (21.3%)	25 (46.3%)	0.001
Positive family history of asthma	25	17 (13.9%)	8 (14.8%)	0.877
MP infection	38 (21.6%)	30 (24.6%)	8 (14.8%)	0.146

M, male; F, female; P-value was based on χ^2 test.

TABLE 2 | Characteristics of cases and control from Xiangya Hospital and Hunan Children's Hospital.

	No-wheezing control	Wheezing infants			p-values
		Total	First wheezing	Recurrent wheezing	
Subjects (M/F)	25 (19/6)	52 (40/12)	26 (19/7)	26 (21/5)	ns ^a
Age, months	13.92 ± 9.67	12.98 ± 9.29	11.35 ± 8.91	13.96 ± 9.67	ns ^a

^a χ^2 test for comparison of first wheezing group versus recurrent wheezing group.

allergic to milk, and one was allergic to medicine (Penicillin G). Besides, in the first wheezing group, 26 children were allergy history-positive, while in the recurrent wheezing group, 25 children were allergy history-positive. Incidence of wheezing illness was strongly related to a history of allergy (21.3% vs. 43.6%, $p=0.001$) (Table 1).

Wheezing and Family History of Asthma

By analyzing the cases in Xiangya Hospital, we found that 17 (13.9%) and eight (14.8%) children were asthma family history-positive in the first wheezing group and recurrent wheezing group, respectively, which indicated that there was no significant association between wheezing and family history of asthma. (Table 1).

Wheezing and Pathogen

Of the samples collected in Xiangya Hospital, 97 (55.11% of all the individuals) children were infected with a virus, MP, bacteria, or fungus. Fifty-five (31.3%) patients were infected with a virus, 38 (21.6%) were infected with MP, 29 (16.5%) patients were infected with bacteria, and 32 (18.2%) patients had more than one infection, respectively. Additionally, only seven (4.0%) patients in our study were infected with fungus. Twenty-seven (15.3%) infants were infected with coxsackie virus, 13 (7.4%) subjects were infected with RSV, 12 (6.8%) individuals were infected with cytomegalovirus, three (1.7%) patients were infected with EB virus, two (1.1%) patients were infected with rubella virus and two with adenovirus (1.1%), respectively. Only one (0.6%) individual was infected with influenza B virus. Moreover, three (1.7%) patients were infected with both the coxsackie virus and cytomegalovirus, one (0.6%) infant was infected with RSV and cytomegalovirus, and one infant was infected with RSV and EB virus (Figure 2).

Thirty subjects in the first wheezing group were infected with MP, along with eight patients in the recurrent wheezing group who had MP infection. However, no significant difference (24.6% vs. 14.8%, $p>0.05$) was found between the two groups (Table 1). Of the patients aged from 1 month to 1 year, 22 of the samples were infected with MP, 12 of the subjects who were 1 year old to 3 years old had MP infection, and four of the samples aged from 3 years old to 6 years old were infected with MP (Table 3). There was no significant difference among three age groups (20.0% vs. 25.0% vs. 22.2%, $p>0.05$). There were 17 children with MP infection in the allergy history-positive group and 21 children with MP infection in the allergy history-negative group. There was a significant difference (33.3% vs. 16.8%, $p<0.05$) between the two groups (Table 4).

Of the 29 patients with bacterial infection, 10 (5.6%) patients had streptococcus pneumoniae, five (2.8%) patients had klebsiella pneumoniae, five (2.8%) patients had staphylococcus, three (1.7%) patients had E. coli, three (1.7%) patients had staphylococcus aureus, one (0.6%) had acid-producing bacillus, one (0.6%) had haemophilus influenzae, and one (0.6%) had acinetobacter baumannii. Seven (4.0%) patients were infected with a fungus, including six (3.4%) children with candida albicans and one with candida glabrata (Figure 2).

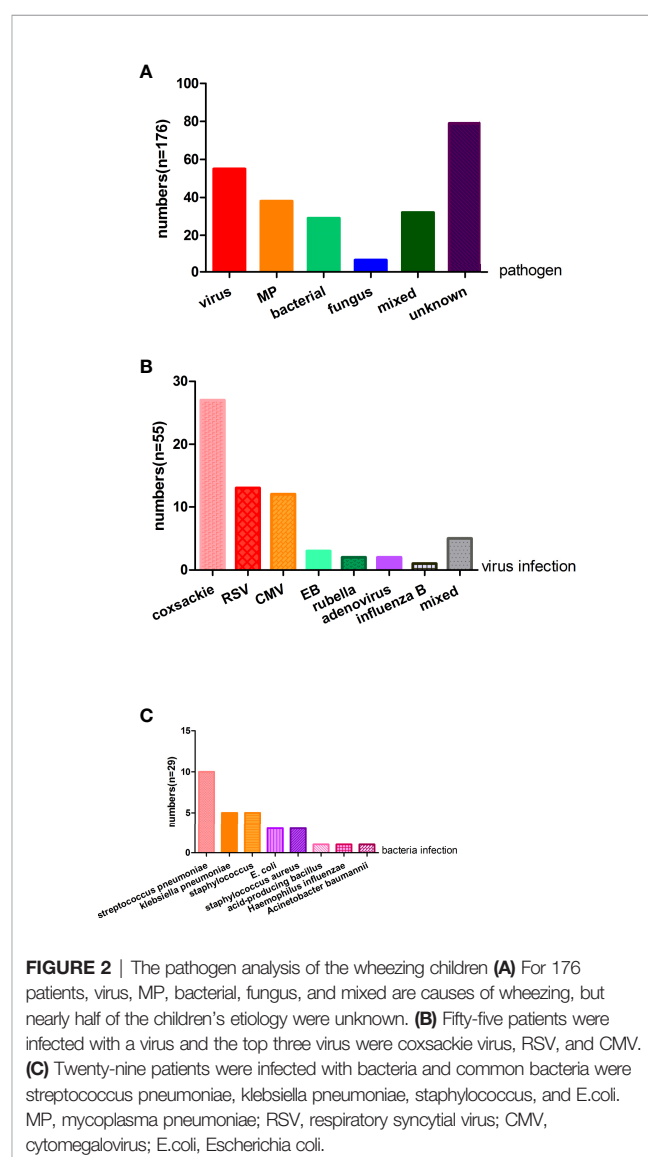


FIGURE 2 | The pathogen analysis of the wheezing children (A) For 176 patients, virus, MP, bacterial, fungus, and mixed are causes of wheezing, but nearly half of the children's etiology were unknown. (B) Fifty-five patients were infected with a virus and the top three virus were coxsackie virus, RSV, and CMV. (C) Twenty-nine patients were infected with bacteria and common bacteria were streptococcus pneumoniae, klebsiella pneumoniae, staphylococcus, and E.coli. MP, mycoplasma pneumoniae; RSV, respiratory syncytial virus; CMV, cytomegalovirus; E.coli, Escherichia coli.

TABLE 3 | Age distribution of all and MP-infected subjects from Xiangya Hospital.

Age groups	All subjects	MP-infected subjects
1month to 1 year old	110 (62.5%)	22 (20.0%)
1~3 years old	48 (27.3%)	12 (25.0%)
3~6 years old	18 (10.2%)	4 (22.2%)
Total	176 (100%)	38 (21.6%)

$\chi^2 = 0.4937$, $p > 0.05$.

Chest X-Ray Results and Wheezing

Chest X-rays at the time of evaluation were normal in five of 176 subjects from Xiangya Hospital. Forty subjects showed increased bronchovesicular shadows, of which one case was diagnosed with combined emphysema. One hundred twenty-two patients showed streaky shadows in X-ray examination results: two individuals combined with axillary lymph node, one subject combined with plural effusion, one had plural reaction, one had emphysematous left lung, and one had left-sided diaphragmatic hernia.

Comparison of IgE, IL-4, IFN- γ , IL-4/IFN- γ , MMP3, MMP9, IL-17A, and IL-17E Levels in Different Groups of Cases From Xiangya Hospital and Hunan Children's Hospital

The children with wheezing diseases had a slightly elevated serum IgE level compared with the control group [18.80(13.65-31.00) vs. 17.9(10.15-21.60), $p = 0.029$]. There was no difference between the first wheezing group and the recurrent wheezing group [19.05 (14.28-36.78) vs. 21.30(15.40-49.03), $p > 0.05$]. Wheezing diseases were associated with significantly higher levels of IL-4 [24.00 (24.00-48.00) vs. 23.00(9.50-27.00), $p = 0.0001$] and IFN- γ [70.59 (41.63-116.46) vs. 49.83(29.58-81.74), $p = 0.004$] than the normal control. But there was no statistical difference in IL-4/IFN- γ (1.24 ± 1.88 vs. 0.68 ± 0.74 , $p > 0.05$) in the above groups. We also analyzed the difference between the first cases and recurrent wheezing cases, but they showed no difference in the level of IL-4 and IFN- γ and the ratio of IL-4/IFN- γ ($p > 0.05$). Moreover, infants with wheezing diseases had significantly higher levels of MMP3 [53.40(20.02-128.2) vs. 30.90(13.80-50.95)] and MMP9 [148.10 (99.30-276.10) vs. 122.10(82.20-162.35)] than the no-wheezing control ($p = 0.001$). The subjects with recurrent wheezing showed higher levels of MMP9 than the samples with first wheezing [254.30(188.00-577.95) vs. 145.55(93.70-279.08), $p = 0.009$], but there were no differences in the levels of MMP3. The difference of IL-17A [80.55(54.46-113.08) vs. 61.11(29.43-93.87)] and IL-17E [1.75(0.66-2.77) vs. 1.19(0.488-2.1615)] were only found in the comparison between wheezing groups and the control group ($p = 0.005$, $p = 0.031$) (Table 5).

DISCUSSION

Wheezing disease was one of the most common respiratory diseases with greater prevalence in winter and spring. Patients can develop heart failure and the illnesses are life threatening if they were not treated properly. Most of the wheezing children were under 3 years old and boys were predominant in the proportion of patients. Previous studies showed that wheezing infants usually had a positive history of allergy (eczema, allergic rhinitis, urticaria, and food allergies) individually or in their family, and a similar result was found in our research (13).

It was of note that children with eczema and papule-like urticaria were more likely to develop asthma. Zhao et al. had noted allergic history and asthma history in the family as risk factors for treating asthma (14). Ronmark et al. carried out a survey with a questionnaire with 30,000 subjects in Swiss suggesting that the co-existence of asthma, rhinitis, and eczema is common. Allergy, asthma history in the family, and smoking were all risk factors for eczema and the patients were characterized by wheezing. Based on our own result and Ronmark's research, allergy history was closely associated with wheezing diseases and was a risk factor for wheezing episodes (15). Fifty-one of 176 children in our research had allergy history such as eczema, allergic rhinitis, urticaria, milk allergy, and drug allergy. Of the first wheezing patients, 21.3% had an allergy history and the rate was significantly higher in the subgroup of recurrent wheezing infants. Therefore, it is conceivable that allergy history was an independent risk factor for wheezing disease.

In our cohort, over half of the patients had an airway infection, including 29.6% of the patients tested had a viral infection, 21.6% of them had MP infection, and 32 subjects had more than one infection. Dominant amounts of major pathogens were restricted to mycoplasma. In our study, 21.6% of the patients who were hospitalized for asthma had MP infection, which was a basic coincidence with the previous report (16). We also conducted a study to determine the association between MP infection and allergy history. Our results showed a different proportion of MP infection between subjects with and without allergy history.

TABLE 4 | The comparison of the ratio of MP infection between groups with and without a history of allergy.

Groups	Number of subjects with MP infection	The ratio of MP infection (%)
With history of allergy (n = 51)	17	33.3
Without history of allergy (n = 125)	21	16.8
Total	38	21.6

$\chi^2 = 5.849$, $p < 0.05$.

TABLE 5 | IgE, Interleukin-4 (IL-4), interferon- γ (IFN- γ), IL-4/IFN- γ , MMP3, MMP-9, IL-17A, and IL-17E expression in 52 wheezing cases and 25 controls.

	No-wheezing control (n=25)	Wheezing children			p-values	
		Total (n = 52)	First wheezing (n = 26)	Recurrent wheezing (n = 26)		
IgE [M (P ₂₅ -P ₇₅) IU/mL]	17.9 (10.15-21.60)	18.80 (13.65-31.00)	19.05 (14.28-36.78)	21.30 (15.40-49.03)	0.627 ^a	0.029 ^b
IL-4 [M (P ₂₅ -P ₇₅) pg/mL]	23.00 (9.50-27.00)	24.00 (24.00-48.00)	24.00 (24.00-155.00)	24.00 (24.00-109.50)	0.623 ^a	0.0001 ^b
IFN- γ [M (P ₂₅ -P ₇₅) pg/mL]	49.83 (29.58-81.74)	70.59 (41.63-116.46)	92.39 (59.71-125.43)	90.68 (36.42-131.60)	0.770 ^a	0.004 ^b
IL-4/IFN- γ (x \pm s)	0.68 \pm 0.74	1.24 \pm 1.88	0.98 \pm 1.00	1.49 \pm 2.47	0.194 ^c	0.160 ^d
MMP3 [M (P ₂₅ -P ₇₅) pg/mL]	30.90 (13.80-50.95)	53.40 (20.02-128.2)	65.72 (24.85-137.75)	94.15 (38.81-213.23)	0.564 ^a	0.001 ^b
MMP9 [M (P ₂₅ -P ₇₅) pg/mL]	122.10 (82.20-162.35)	148.10 (99.30-276.10)	145.55 (93.70-279.08)	254.30 (188.00-577.95)	0.009 ^a	0.001 ^b
IL-17A [M (P ₂₅ -P ₇₅) pg/mL]	61.11 (29.43-93.87)	80.55 (54.46-113.08)	96.07 (66.66-130.13)	83.11 (54.60-135.46)	0.301 ^a	0.005 ^b
IL-17E [M (P ₂₅ -P ₇₅) pg/mL]	1.19 (0.488-2.1615)	1.75 (0.66-2.77)	1.83 (1.04-3.47)	2.13 (0.48-3.96)	0.687 ^a	0.031 ^b

^a Rank sum test for comparison of first wheezing group versus recurrent wheezing group.

^b Rank sum test for comparison of wheezing group versus no-wheezing control.

^c ANOVA for comparison between first wheezing and recurrent wheezing patients.

^d ANOVA for comparison between no-wheezing control and wheezing infants.

In recent years, the incidence of wheezing is increasing year by year. Over half of children wheezing break out repeatedly, and breathing and the immune system to mature in infants and young children period, the period of recurrent wheezing may help children to adversely affect the body's immune system and respiratory system, after treatment the most infant wheezing can alleviate, but about 50% of children with recurrent wheezing can develop bronchial asthma for children. As the symptoms of wheezing and the corresponding risk factors in children change over time, there are certain limitations. Therefore, it is important to explore the association between the pathogenesis of wheezing diseases and the pathogenesis of asthma. The essence of pediatric bronchial asthma is chronic airway inflammation, and Th1/Th2 imbalance is a widely accepted theory. MMP3, MMP9, and IL-17 are involved in airway remodeling of childhood asthma, and in our study they are also involved in the pathogenesis of infant wheezing disease, which provide more experimental basis for prevention and treatment of infant wheezing diseases.

In our study, 29.6% of all the subjects were infected with a virus, 27 patients had coxsackie virus (15.3% of all the samples), 13 patients had RSV infection (6.8% of all the samples), and 12 children had cytomegalovirus (6.8% of all the samples). The discrepancy observed between our result and the previous study may result from the different times of sample collection, as our samples were collected in winter.

Viral infection at an early age is linked to recurrent wheezing diseases and asthma in children. Viral infection causes the destruction of airway epithelial cells and results in the decrease of the airway immune response, and indirectly exacerbates allergic inflammation. It is suggested there were more eosinophils and lymphocytes infiltration in the lung tissues within RSV-infected mice. Besides, the increased level of cytokine TNF- α , IFN- γ , IL-5, and IL-2 were also observed (17). RSV infection had an influence on Th1/Th2 compartment toward Th1 and facilitated the development of airway inflammation. Some studies have reported that RSV-infected airway epithelial cells secrete cytokines and chemokines including thymic stromal lymphopoietin which work to activate Th2 and lead to Th2 factor secretion. RSV infection also activates T lymphocyte like Th2, Th17, regulatory T cells, and cytotoxic T cells (18). The results of Tourdot's research showed that RSV infection in mice would not induce airway inflammation but would thicken the mouse bronchial basement membrane and result in an increased amount of collagen in lung tissue. In addition, the amount of fibroblastic growth factor (FGF) was increased with RSV infection (19).

Previous studies have identified the association between wheezing diseases and viral infection, but the role of bacterial infection has not been shown. One study determined *Haemophilus influenzae*, *Moraxella catarrhalis*, and *Streptococcus pneumoniae* as the major causes of bacterial infection associated with wheezing. In this research, gram-negative bacilli were found in most of the samples, but the number of white blood cells and neutrophils was sometimes not increased. The bacterial airway

colonization was reported to increase the risk of wheezing and continuous wheezing leading to the exacerbation of acute wheezing diseases (20). Of all the samples in our present study, 16.5% were infected with bacterial-like *streptococcus pneumoniae*, *klebsiella pneumoniae*, *staphylococcus*, *Escherichia coli*, *staphylococcus aureus*, acid-producing *bacillus*, *haemophilus influenzae*, and *acinetobacter baumannii*, indicating that bacterial infection was associated with wheezing diseases of infants.

A lower respiratory tract viral infection with an upper respiratory tract bacterial infection was prevalent among wheezing patients. In the study cohort of Hishiki's research (21), 43.6% of all the patients had bacterial infections, *haemophilus influenzae* (43.9% of all the samples), *streptococcus pneumoniae* (36.6% of all the samples), and *moraxella catarrh* (29.3% of all the samples) were the common causative bacteria. RSV-infected wheezing children were often co-infected with a drug-resistant lower respiratory tract infection (21). A previous study, which enrolled 165 RSV-infected children who lived in an intensive care unit, suggested that gram-positive bacteria was found in the lower respiratory tract secretion in 42.4% of all the samples. The RSV-infected children tended to develop bacterial pneumonia (22). Besides virus and bacterial coinfection, the coinfection of different kinds of viruses was also observed in patients.

The wheezing patients were often during mixed infection; however, the mechanism causing wheezing was not thoroughly studied. One possible explanation is that patients with MP or virus may suffer airway mucosal cell damage, thus making them more likely to be infected with another pathogen.

In our study, 32 subjects were infected with more than two kinds of pathogens (18.2% of all the patients), six individuals were infected with three pathogens (3.4% of all the samples), and 18.8% of the patients had mixed infection means "had more than one infection. Virus and MP infected all six patients. It reminded us of the importance of combination therapy for these patients.

There were reports that showed that the mechanism of immune response in children with allergic rhinitis was similar

to that of asthma. Recurrent eczema was a common allergic disease characterized by Th1/Th2 imbalance. Increased levels of IgE was detected in patients with eczema and the level of IgE was positively correlated with the severity of the disease. In our cohort from Hunan Children's Hospital, the level of IgE was significantly higher in the wheezing group but was not associated with the number of wheezing episodes.

IL-4 was specifically secreted by Th2 cells and worked to promote the proliferation, differentiation, and activation of B cells. It was critical in the synthesis of IgE by B cells and was associated with immune diseases. Another important cytokine in Th1/Th2 balance was IFN- γ , which was a cytokine signature of activated T cells and NK cells. IFN- γ functioned to activate macrophages, neutrophils, and NK cells, as well as to suppress the proliferation of Th2 cells by promoting the transition of Th0 cells to Th1 cells. Some experts thought IFN- γ and IL-12 could prevent damage from RSV infection (23), but there were studies showing that the level of IFN- γ varied with age because increased IFN- γ expression was observed in patients who were older than 1 year old compared to a healthy control, comparing with age-matched controls, IFN- γ levels were significantly higher in RSV group ≥ 12 months of age (1 year old). Our result showed that elevated IFN- γ levels may be age-related (average age in the wheezing group was 12.65 ± 8.68 months) (24). Our results showed that the level of IL-4 and IL-4/IFN- γ was significantly increased in wheezing children, indicating the activation of Th1 and Th2 cells. However, in our study, the level of IFN- γ was increased in wheezing children. To our knowledge, the discrepancy may relate to age.

Biopsy examination for severe asthma airway tissue showed the increased level of MMP9 (25). MMP9 could stimulate the secretion of growth factor like TGF- β 1 and FGF and induce cell proliferation and differentiation. MMP9 was also an essential player in IL-13 mediating TGF- β 1 secretion and was involved in pulmonary fibrosis and degradation of extracellular matrix. MMP served as a marker of airway remodeling (26). A previous study suggests that RSV infection may be associated with bronchial hyper-responsiveness, as leukotriene, neutrophil, and lymphocyte levels increased in alveolar lavage fluid of infected mice and bronchial hyper-responsiveness is observed in mice (27). MMP3 cannot only synergistically degrade the extracellular matrix, but also accelerate the decomposition of the basement membrane of pulmonary blood vessels, and participate in the remodeling of pulmonary blood vessels, playing an indispensable role in the occurrence and prognosis of lung diseases. In mouse models, it was found that MMP3 can also mediate the occurrence of acute lung injury by interacting with neutrophils and macrophages (28). We evaluated the expression of MMP3 and MMP9 and found a significant increase in wheezing patients. Moreover, the level of MMP9 was associated with the number of wheezing episodes as there was a difference in our study between first and recurrent wheezing children.

IL-17 was another altered cytokine in wheezing children compared to the control group. In our analysis, both IL-17A and IL-17E were significantly increased in the wheezing group, but there was no difference between first and recurrent wheezing patients. IL-17 was a major pro-inflammation cytokine produced

by Th17 cells. IL-17 had been recognized to induce the production of Th2-specific cytokine-like IL-4 and IL-13 and was also involved in Th1/Th2 imbalance (29). Our study revealed that IL-17 is involved in the pathogenesis of wheezing disease in young children, and children with a wheezing disease may experience airway inflammation and airway remodeling.

It should be noted that limitations exist in this study. Firstly, the sample size was relatively small to analyze the association between cytokines and wheezing disease, which may lead to low statistical power for data analysis. Thus it is necessary to strengthen cooperation between hospitals. Secondly, our study was preliminary, and more research is needed to explore the mechanisms of wheezing disease, especially in immune response.

CONCLUSIONS

In our study, a higher incidence of wheezing disease was observed in boys than girls and wheezing disease was prevalent in children under 3 years old. Children with allergy history were more likely to develop wheezing diseases. MP and viral infection were the common causative pathogens of infants with wheezing diseases and infants with allergy history were more vulnerable to MP infection. MMP3, MMP9, and IL-17 are involved in the pathogenesis of children wheezing disease, which provide a more experimental basis for the prevention and treatment of wheezing disease.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Chinese Clinical Trial Registry. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

YJT and RHH prepared original draft preparation, YJT and YXY analyzed data, XRZ and CTL collected sample, RH supervised the whole study, reviewing and editing the final approval of the version submitted for publication.

FUNDING

This work was supported by the Hunan Provincial Natural Science Foundation of China (2016JJ4109 and 2020JJ4909).

REFERENCES

- Heikkilä P, Korppi M, Ruotsalainen M, Backman K. Viral Wheezing in Early Childhood as a Risk Factor for Asthma in Young Adulthood: A Prospective Long-Term Cohort Study. *Health Sci Rep* (2022) 5(2):e538. doi: 10.1002/hsr2.538
- Oksel C, Granell R, Mahmoud O, Custovic A, Henderson AJ, Stelar, et al. Causes of Variability in Latent Phenotypes of Childhood Wheeze. *J Allergy Clin Immunol* (2019) 143:1783–1790.e11. doi: 10.1016/j.jaci.2018.10.059
- Martinez FD. Viruses and Atopic Sensitization in the First Years of Life. *Am J Respir Crit Care Med* (2000) 162:S95–9. doi: 10.1164/ajrccm.162.supplement_2.ras-8
- van den Nieuwenhof L, Schermer T, Bosch Y, Bousquet J, Heijdra Y, Bor H, et al. Is Physician-Diagnosed Allergic Rhinitis a Risk Factor for the Development of Asthma? *Allergy* (2010) 65:1049–55. doi: 10.1111/j.1398-9995.2009.02316.x
- Beigelman A, Bacharier LB. Infection-Induced Wheezing in Young Children. *J Allergy Clin Immunol* (2014) 133:603–4. doi: 10.1016/j.jaci.2013.12.001
- Hahn DL, Azenabor AA, Beatty WL and Byrne GI. Chlamydia Pneumoniae as a Respiratory Pathogen. *Front Biosci* (2002) 7:e66–76. doi: 10.2741/hahn
- Chu HW, Honour JM, Rawlinson CA, Harbeck RJ and Martin RJ. Effects of Respiratory Mycoplasma Pneumoniae Infection on Allergen-Induced Bronchial Hyperresponsiveness and Lung Inflammation in Mice. *Infect Immun* (2003) 71:1520–6. doi: 10.1128/IAI.71.3.1520-1526.2003
- Gern JE, Calatroni A, Jaffee KF, Lynn H, Dresen A, Cruikshank WW, et al. Patterns of Immune Development in Urban Preschoolers With Recurrent Wheeze and/or Atopy. *J Allergy Clin Immunol* (2017) 140:836–44. doi: 10.1016/j.jaci.2016.10.052
- Ronchi A, Doern C, Brock E, Pugni L and Sanchez PJ. Neonatal Adenoviral Infection: A Seventeen Year Experience and Review of the Literature. *J Pediatr* (2014) 164:529–35. doi: 10.1016/j.jpeds.2013.11.009
- American Academy of Pediatrics Subcommittee on Diagnosis and Management of Bronchiolitis. Diagnosis and Management of Bronchiolitis. *Pediatrics* (2006) 118(4):1774–93. doi: 10.1542/peds.2006-2223
- Nagayama Y, Tsubaki T, Nakayama S, Sawada K, Taguchi K, Toba T, et al. Bacterial Colonization in Respiratory Secretions From Acute and Recurrent Wheezing Infants and Children. *Pediatr Allergy Immunol* (2007) 18:110–7. doi: 10.1111/j.1399-3038.2006.00492.x
- Jeong YC, Yeo MS, Kim JH, Lee HB and Oh JW. Mycoplasma Pneumoniae Infection Affects the Serum Levels of Vascular Endothelial Growth Factor and Interleukin-5 in Atopic Children. *Allergy Asthma Immunol Res* (2012) 4:92–7. doi: 10.4168/aa.2012.4.2.92
- Carraro S, Bozzetto S, Giordano G, El MD, Stocchero M, Pirillo P, et al. Wheezing Preschool Children With Early-Onset Asthma Reveal a Specific Metabolomic Profile. *Pediatr Allergy Immunol* (2018) 29:375–82. doi: 10.1111/pai.12879
- Zhao J, He Q, Zhang G, Chen Q, Bai J, Huang Y, et al. Status of Asthma Control in Children and the Effect of Parents' Knowledge, Attitude, and Practice (KAP) in China: A Multicenter Study. *Ann Allergy Asthma Immunol* (2012) 109:190–4. doi: 10.1016/j.ana.2012.07.005
- Ronmark EP, Ekerljung L, Lotvall J, Wennergren G, Ronmark E, Toren K, et al. Eczema Among Adults: Prevalence, Risk Factors and Relation to Airway Diseases. Results From a Large-Scale Population Survey in Sweden. *Br J Dermatol* (2012) 166:1301–8. doi: 10.1111/j.1365-2133.2012.10904.x
- Lieberman D, Lieberman D, Printz S, Ben-Yaakov M, Lazarovich Z, Ohana B, et al. Atypical Pathogen Infection in Adults With Acute Exacerbation of Bronchial Asthma. *Am J Respir Crit Care Med* (2003) 167:406–10. doi: 10.1164/rccm.200209-996OC
- Becnel D, You D, Erskin J, Dimina DM and Cormier SA. A Role for Airway Remodeling During Respiratory Syncytial Virus Infection. *Respir Res* (2005) 6:122. doi: 10.1186/1465-9921-6-122
- Hansel TT, Johnston SL and Openshaw PJ. Microbes and Mucosal Immune Responses in Asthma. *Lancet* (2013) 381:861–73. doi: 10.1016/S0140-6736(12)62202-8
- Tourdou S, Mathie S, Hussell T, Edwards L, Wang H, Openshaw PJ, et al. Respiratory Syncytial Virus Infection Provokes Airway Remodelling in Allergen-Exposed Mice in Absence of Prior Allergen Sensitization. *Clin Exp Allergy* (2008) 38:1016–24. doi: 10.1111/j.1365-2222.2008.02974.x
- Bisgaard H, Hermansen MN, Bonnelykke K, Stokholm J, Baty F, Skjott NL, et al. Association of Bacteria and Viruses With Wheezy Episodes in Young Children: Prospective Birth Cohort Study. *BMJ* (2010) 341:c4978. doi: 10.1136/bmj.c4978
- Hishiki H, Ishiwada N, Fukasawa C, Abe K, Hoshino T, Aizawa J, et al. Incidence of Bacterial Coinfection With Respiratory Syncytial Virus Bronchopulmonary Infection in Pediatric Inpatients. *J Infect Chemother* (2011) 17:87–90. doi: 10.1007/s10156-010-0097-x
- Thorburn K, Harigopal S, Reddy V, Taylor N, van Saene HK. High Incidence of Pulmonary Bacterial Co-Infection in Children With Severe Respiratory Syncytial Virus (RSV) Bronchiolitis. *Thorax* (2006) 61:611–5. doi: 10.1136/thx.2005.048397
- Chen ZM, Mao JH, Du LZ and Tang YM. Association of Cytokine Responses With Disease Severity in Infants With Respiratory Syncytial Virus Infection. *Acta Paediatr* (2002) 91:914–22. doi: 10.1111/j.1651-2227.2002.tb02877.x
- Chung HL, Park HJ, Kim SY and Kim SG. Age-Related Difference in Immune Responses to Respiratory Syncytial Virus Infection in Young Children. *Pediatr Allergy Immunol* (2007) 18:94–9. doi: 10.1111/j.1399-3038.2006.00501.x
- Araujo BB, Dolnikoff M, Silva LF, Elliot J, Lindeman JH, Ferreira DS, et al. Extracellular Matrix Components and Regulators in the Airway Smooth Muscle in Asthma. *Eur Respir J* (2008) 32:61–9. doi: 10.1183/09031936.00147807
- Matsumoto H, Niimi A, Takemura M, Ueda T, Minakuchi M, Tabuena R, et al. Relationship of Airway Wall Thickening to an Imbalance Between Matrix Metalloproteinase-9 and Its Inhibitor in Asthma. *Thorax* (2005) 60:277–81. doi: 10.1136/thx.2004.028936
- Fullmer JJ, Khan AM, Elidemir O, Chiappetta C, Stark JM and Colasurdo GN. Role of Cysteinyl Leukotrienes in Airway Inflammation and Responsiveness Following RSV Infection in BALB/c Mice. *Pediatr Allergy Immunol* (2005) 16:593–601. doi: 10.1111/j.1399-3038.2005.00248.x
- Warner RL, Beltran L, Younkin EM, Lewis CS, Weiss SI, Varani J, et al. Role of Stromelysin 1 and Gelatinase B in Experimental Acute Lung Injury. *Am J Respir Cell Mol Biol* (2001) 24(5):537–44. doi: 10.1165/ajrcmb.24.5.4160
- Isailovic N, Daigo K, Mantovani A and Selmi C. Interleukin-17 and Innate Immunity in Infections and Chronic Inflammation. *J Autoimmun* (2015) 60:1–11. doi: 10.1016/j.jaut.2015.04.006

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.

Copyright © 2022 Tang, Yang, He, Huang, Zheng and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

EDITED BY
Liang Cheng,
Harbin Medical University, China

REVIEWED BY
Pranjal Kumar,
Indian Institute of Technology
Dharwad, India
Tienan Feng,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Meng Xu
xumengjnu@foxmail.com

[†]These authors have contributed
equally to this work

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

RECEIVED 24 April 2022
ACCEPTED 27 June 2022
PUBLISHED 22 July 2022

CITATION
Mo X, Hu D, Yang P, Li Y, Bashir S,
Nai A, Ma F, Jia G and Xu M (2022) A
novel cuproptosis-related prognostic
lncRNA signature and lncRNA
MIR31HG/miR-193a-3p/TNFRSF21
regulatory axis in lung
adenocarcinoma.
Front. Oncol. 12:927706.
doi: 10.3389/fonc.2022.927706

COPYRIGHT
© 2022 Mo, Hu, Yang, Li, Bashir, Nai,
Ma, Jia and Xu. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution
or reproduction in other forums is
permitted, provided the original author
(s) and the copyright owner(s) are
credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

A novel cuproptosis-related prognostic lncRNA signature and lncRNA MIR31HG/miR-193a-3p/TNFRSF21 regulatory axis in lung adenocarcinoma

Xiacong Mo^{1†}, Di Hu^{2†}, Pingshan Yang³, Yin Li¹,
Shoaib Bashir¹, Aitao Nai¹, Feng Ma¹, Guoxia Jia¹
and Meng Xu^{1*}

¹Department of Oncology, The First Affiliated Hospital of Jinan University, Jinan University, Guangzhou, China, ²Department of Neurology and Stroke Centre, The First Affiliated Hospital of Jinan University, Guangzhou, China, ³Department of Thoracic Surgery, The First Affiliated Hospital of Jinan University, Guangzhou, China

Lung adenocarcinoma (LUAD) remains the most common subtype of lung malignancy. Cuproptosis is a newly identified cell death which could regulate tumor cell proliferation and progression. Long non-coding RNAs (lncRNAs) are key molecules and potential biomarkers for diagnosing and treating various diseases. However, the effects of cuproptosis-related lncRNAs on LUAD are still unclear. In our study, 7 cuproptosis-related lncRNAs were selected to establish a prognostic model using univariate Cox regression analysis, LASSO algorithm, and multivariate analysis. Furthermore, we evaluated AC008764.2, AL022323.1, ELN-AS1, and LINC00578, which were identified as protective lncRNAs, while AL031667.3, AL606489.1, and MIR31HG were identified as risk lncRNAs. The risk score calculated by the prognostic model proved to be an effective independent factor compared with other clinical features by Cox regression analyses [univariate analysis: hazard ratio (HR) = 1.065, 95% confidence interval (CI) = 1.043–1.087, $P < 0.001$; multivariate analysis: HR = 1.067, 95% CI = 1.044–1.091, $P < 0.001$]. In addition, both analyses (ROC and nomogram) were used to corroborate the accuracy and reliability of this signature. The correlation between cuproptosis-related lncRNAs and immune microenvironment was elucidated, where 7 immune cells and 8 immune-correlated pathways were found to be differentially expressed between two risk groups. Furthermore, our results also identified and verified the ceRNA of cuproptosis-related lncRNA MIR31HG/miR-193a-3p/TNFRSF21 regulatory axis using bioinformatics tools. MIR31HG was highly expressed in LUAD specimens and some LUAD cell lines. Inhibition of MIR31HG clearly reduced the proliferation, migration, and invasion of the LUAD cells. MIR31HG showed oncogenic features *via* sponging miR-193a-3p and tended to positively

regulate TNFRSF21 expression. In a word, lncRNA MIR31HG acts as an oncogene in LUAD by targeting miR-193a-3p to modulate TNFRSF21, which may be beneficial to the gene therapy of LUAD.

KEYWORDS

lung adenocarcinoma, cuproptosis, MIR31HG, miR-193a-3p, TNFRSF21

Introduction

Lung adenocarcinoma (LUAD) is one of the most common subtypes of lung malignancy, and it ranges from initially non-invasive tumors to high-mortality-specific invasive tumors (1, 2). Although a number of biomarkers or diagnostic tools that may be employed to predict the prognosis of LUAD have been discovered in recent years (3, 4), they are still in the stage of molecular research (5). Therefore, it will be indispensable to identify promising biomarkers and prognostic models to reveal the prognostic genetic characteristics of LUAD and obtain the most accurate clinical information.

Cuproptosis is a novel mode of cell death that is relevant to copper and mitochondrial respiration. The pathological mechanism is that copper interacts directly with the fatty acylated components of the tricarboxylic acid (TCA) cycle, leading to excessive aggregation of fatty acylated proteins and loss of iron-sulfur cluster proteins, which stimulates proteotoxic stress and cell death (6). Recent studies have found a close relationship between copper cell death and human cancer (7, 8), which proved that cuproptosis is closely related to the development of cancers, but it remains unclear in LUAD.

lncRNA is a newly discovered functional lncRNA which has the ability to mediate various mechanisms through their multiple functions and plays vital roles in a large number of cancer processes (9, 10). Recent studies have confirmed that lncRNAs regulated the early development of LUAD through different signaling pathways—for example, lncRNA JPX promoted the tumorigenesis in LUAD by sponging miR-33a-5p (11), and lncRNA GMDS-AS1 could inhibit LUAD *via* regulating the miR-96-5p/CYLD pathway (12). Besides this, lncRNA FAM83A-AS1 accelerated LUAD migration, proliferation, and invasion (13). However, the precise involvement of cuproptosis-related lncRNAs in LUAD is still ambiguous.

Our current work was designed to delve deeper into the expression profiles of cuproptosis-related lncRNAs and their relationship with the immune microenvironment as well as validate potentially relevant regulatory mechanisms in LUAD. Our findings would provide valuable references for efficient prognostic biomarkers and the diagnosis of LUAD.

Materials and methods

Data collection and processing

The data were all obtained from The Cancer Genome Atlas (TCGA; <https://portal.gdc.cancer.gov/>). The LUAD microarray gene profiling dataset in this research was obtained from the GEO NCBI web server (<https://www.ncbi.nlm.nih.gov/geo/>). Patients with missing survival information were excluded. Perl (<https://www.perl.org>, version 5.32.1) was used to collate the clinical details.

Identification of cuproptosis-related genes

Cuproptosis-correlated genes were identified based on previous reports (6, 14, 15). In total, 39 mRNAs were finally selected as differentially expressed genes (DEGs) by the packages of “limma” and “pheatmap”. Then, a PPI network (related to protein-protein interaction) of the 39 DEGs was established by a search tool to retrieve interacting genes (STRING, <https://string-db.org/>).

KEGG and GO enrichment analysis

The Gene Ontology (GO; <http://www.geneontology.org/>) and the Kyoto Encyclopedia of Genes and Genomes (KEGG; <http://www.genome.jp/kegg/>) enrichment analyses were conducted using the “ggplot2” R package. The GO database was performed to analyze the biological characteristics of these cuproptosis-related genes. KEGG was performed to detect the signaling pathway of cuproptosis-related genes.

Identification of cuproptosis-related lncRNAs and prognosis model construction

lncRNAs related to cuproptosis-related DEGs were screened out based on Pearson correlation analysis. We randomly classified the

included cases ($n = 504$) into training and validation cohorts at a 1:1 ratio. Cuproptosis-related lncRNAs were selected following univariate Cox regression analysis, LASSO Cox algorithm, and multivariate analysis. These cuproptosis-related lncRNAs were chosen to establish a prognostic model (the risk score = $\text{expression}_{\text{lncRNA1}} \times \text{coefficient}_{\text{lncRNA1}} + \text{expression}_{\text{lncRNA2}} \times \text{coefficient}_{\text{lncRNA2}} + \dots + \text{expression}_{\text{lncRNA}_n} \times \text{coefficient}_{\text{lncRNA}_n}$). We then analyzed the hazard ratio (HR) of prognostic factors for distinguishing between protective lncRNA (HR >1) and risk lncRNA (HR <1). These cuproptosis-related lncRNAs were further visualized *via* Cytoscape and Sankey diagram. Furthermore, the patients were classified into two risk groups (high and low).

Clinical meaning of the prognostic model

Univariate and multivariate Cox regressions were performed to identify whether the risk score and clinical features (age, gender, grade, *etc.*) were valuable prognostic indicators for LUAD patients. Nomograms were used to show the clinical features and risk scores of survival rates.

Competing endogenous RNA network construction

To further elucidate the potential mechanism of cuproptosis-related lncRNAs in LUAD, we constructed a ceRNA network. Mircode (www.mircode.org) was utilized to predict the miRNA targets connecting to cuproptosis-related lncRNAs. After miRNA identification, TargetScan (http://www.targetscan.org/vert_72/) and miRDB databases (<http://mirdb.org/>) were utilized to predict mRNA targets interacting with miRNAs.

Cell culture and clinical specimens

The LUAD cell lines (A549, NCI-H2009, and PC9) and bronchial epithelioid cells (HBE) were generous gifts from Dr. Feng Ma. The LUAD cells were cultured in RPMI-1640 medium (Hyclone; GE Healthcare) and maintained in a humidified incubator at 37°C in 5% CO₂. In total, 34 paired LUAD tissues (T) and normal specimens (N) were collected from the First Affiliated Hospital of Jinan University from April 2020 to May 2021. Our research concerning human tissues was reviewed and approved by the Ethics Committee of The First Affiliated Hospital of Jinan University. All patients signed informed consents in the present study.

Cell transfection

si-MIR31HG, miR-193a-3p inhibitor, miR-193a-3p mimics, and their negative control (si-ctrl) were synthesized by GenePharma (Shanghai, China). NCI-H2009 and A549 cells

were evenly plated in 96-well plates. When the two cells reached about 80–90% confluence, they were transfected with plasmid by Lipo3000 (Invitrogen, Carlsbad, CA, USA). NCI-H2009 and A549 cells were harvested for the subsequent experiments following incubation at 37°C for 48 h.

Cell proliferation assay

96-well plates were taken to seed LUAD cells (2×10^5 cells/well) (A549 and NCI-H2009). After incubation at 37°C and 5% CO₂ for various times, a CCK-8 reagent test kit, which was provided from Tiangen (Hangzhou, China), was mixed at 10 µl/well, and LUAD cells were evenly incubated for 3 h at 37°C and 5% CO₂. Finally, we read the absorbance at 450 nm on the enzyme labeling instrument (Thermo Fisher Scientific, Inc.)

Wound healing assay

The migration was identified by wound healing assay. Transfected cells were seeded in individual 6-well dishes and incubated at 37°C until reaching about 80–90% confluence. Then, the two cells were scratched with a constant-diameter stripe from the bottom of the wells by a sterile 200-µl pipette tip. Filming was performed at 0 and 24 h after wounding. A total of 10 areas were randomly selected to mark and measure.

Transwell assay

The transfected LUAD cell suspensions (200 µl) were moved to the upper chamber of the transwell module (Corning, Inc.) and incubated for 24 h at 37°C and 5% CO₂. The cells would invade the bottom chamber that contained the prepared medium (with 10% fetal bovine serum added). The invaded NCI-H2009 and A549 cells that existed in the lower chamber were thereby treated with methanol and 0.1% crystal violet. The cell invasion rate was measured by eluting the crystal violet which existed in the transwell by 33% acetic acid. Finally, we measured the OD 570 nm value in the eluted liquid.

Dual-luciferase assay

The online tool TargetScan was applied to identify the potential binding sites. The wild-type site (wt) and mutant site (mut) sequences of MIR31HG (MIR31HG wt and MIR31HG mut) and TNFRSF21 (TNFRSF21 wt and TNFRSF21 mut), including the homologous binding sites of miR-193a-3p, were amplified and uniformly plugged in the vector pGL3 (Promega, Madison, WI, USA). Then, miR-193a-3p mimics were co-transfected with MIR31HG wt, MIR31HG mt,

TNFRSF21 wt, or TNFRSF21 mut using Lipo3000. After 48 h, Dual-Luciferase Reporter Assay System provided from Promega was used to detect the luciferase activity.

RNA immunoprecipitation assay

EZ-Magna RIPTM RNA-Binding Protein Immunoprecipitation Kit provided from Labbiotech was applied to execute the RNA immunoprecipitation (RIP) assay. NCI-H2009 and A549 were uniformly mixed with RIP buffer owning beads stuck by anti-Ago2 or anti-IgG (negative control) and incubated overnight. Finally, the obtained immunoprecipitated complexes were measured by real-time PCR.

Western blot

RIP assay lysis buffer (Beyotime, Shanghai, China) was used to extract the protein from LUAD cells or tissues. After measuring the density of all proteins, the target proteins (30 µg/lane) were separated by SDS-PAGE (10%) and then carefully transferred onto polyvinylidene fluoride membranes (Bio-Rad, Hercules, CA, USA). Later on, the transferred membranes were blocked with silk milk (5%) for 1 h at 37°C and incubated with the primary anti-TNFRSF21 (catalogue no. ab8417; 1:800; Abcam; USA) and β-actin (catalog no. ab8226; 1:3,000; Abcam; USA) overnight at 4°C. On the second day, the membranes were then treated with a corresponding secondary antibody (catalog no. ab6721; 1:5,000; catalog no. ab6728; 1:5,000; Abcam; USA). Finally, the protein signals on the membrane were visualized by enhanced ECL detection kit (Beyotime, Shanghai, China).

RT-qPCR analysis

Trizol (Beyotime, Shanghai, China) was employed to extract the total RNA from LUAD cells and tissues. RNA was reverse-transcribed into complementary DNA (cDNA) using SuperScript VILO cDNA Kit (Thermo Fisher Scientific, Inc.). SYBR Green qPCR Master Mix (Applied Biosystems, USA) was applied to detect the quantitative PCR from the $2^{-\Delta\Delta C_q}$ method. The primers are listed in Table 1.

Statistical analysis

One-way analysis of variance (ANOVA) and paired samples *t*-test were used to assess differences between groups. Pearson's correlation test analyzed the correlations. SPSS 25.0 software and GraphPad Prism 8.0.1 were performed for statistical analyses. All experiments were performed independently and repeated three times. *P* < 0.05 was considered statistically significant.

Results

Identification of the expression of cuproptosis-related genes in LUAD

The flow chart of the study is shown in Figure 1. The expression degrees of 49 genes linked to cell cuproptosis were compared in LUAD and normal tissues from the TCGA dataset, and 39 cuproptosis-related genes were identified as DEGs. Then, 26 genes (CLU, PDHB, BCL2, COMMD1, etc.) were detected to be enriched, while 13 genes (CD36, TLR4, TNFRSF21, ABCA1, etc.) were decreased in the LUAD group relative to normal tissues (Figure 2A). PPI showed the interaction among 39 DEGs (Figure 2B).

Biological functional enrichment research of cuproptosis-related DEGs

GO and KEGG databases were used to analyze the potential and meaningful function of 39 DEGs. The GO analysis suggested that these genes in the biological processes were enriched in “neuron death” and “neuron apoptotic process”. These genes in cell component were enriched in “mitochondrial matrix” and “oxidoreductase complex”. Alterations in molecular function were brimming with “amide binding” and “oxidoreductase activity” (Figure 2C). Moreover, the KEGG analysis indicated that these DEGs were enriched in “pathways of neurodegeneration multiple diseases”, “apoptosis”, “necroptosis”, and “p53 signaling pathway” (Figure 2D).

Identification of cuproptosis-related lncRNAs and co-expression network construction

Firstly, this work classified the included cases (*n* = 504) into training (*n* = 252) and validation (*n* = 252) cohorts at a 1:1 ratio.

TABLE 1 Primer list.

Gene	Primers
MIR31HG	Forward: 5'-TCCCAGTTTCAGACCACC-3' Reverse: 5'-CCAGGCTATGTCITTCCTCTAT-3'
TNFRSF21	Forward: 5'-ATTCGCCAGGCTGAGGACAAAC-3' Reverse: 5'-ACACACACACACCCCAAC-3'
GADPH	Forward: 5'-ACCACAGTCCATGCCATCAC-3' Reverse: 5'-TCCACCACCTGTTGCTGTA-3'
U6	Forward: 5'-CTCGCTTCGGCAGCACA-3' Reverse: 5'-AACGCTTCACGAATTTGCGT-3'
miR-193a-3p	Forward: 5'-CGCGAACTGGCCTACAAAGTG-3' Reverse: 5'-AGTGCAGGGTCCGAGGTATT-3'

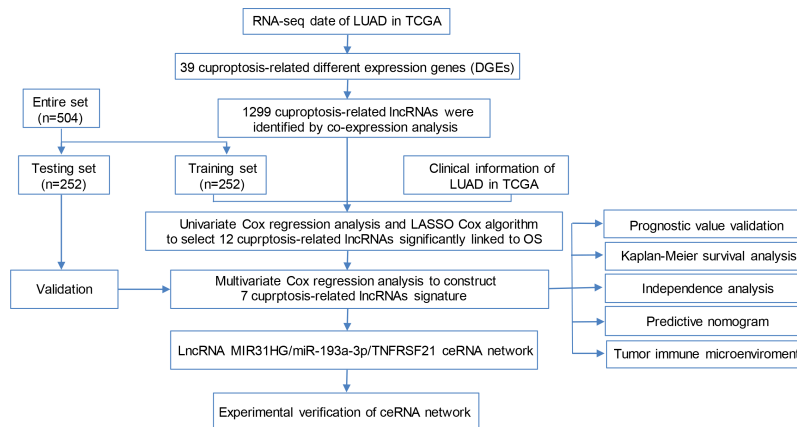


FIGURE 1
Flow chart of the study.

The clinicopathologic characteristics of LUAD patients are listed in Table 2. A total of 1,299 lncRNAs related to 39 DEGs were screened out for future analysis according to Pearson correlation method. Secondly, we operated the univariate Cox regression analysis and LASSO Cox algorithm to reduce multicollinearity,

and we found 12 cuproptosis-related lncRNAs (Figures 3A, B). Finally, 7 lncRNAs, including AL031667.3, ELN-AS1, LINC00578, AL022323.1, AL606489.1, AC008764.2, and MIR31HG, were selected through subsequent multivariate analysis to construct the risk model, and the global p -value

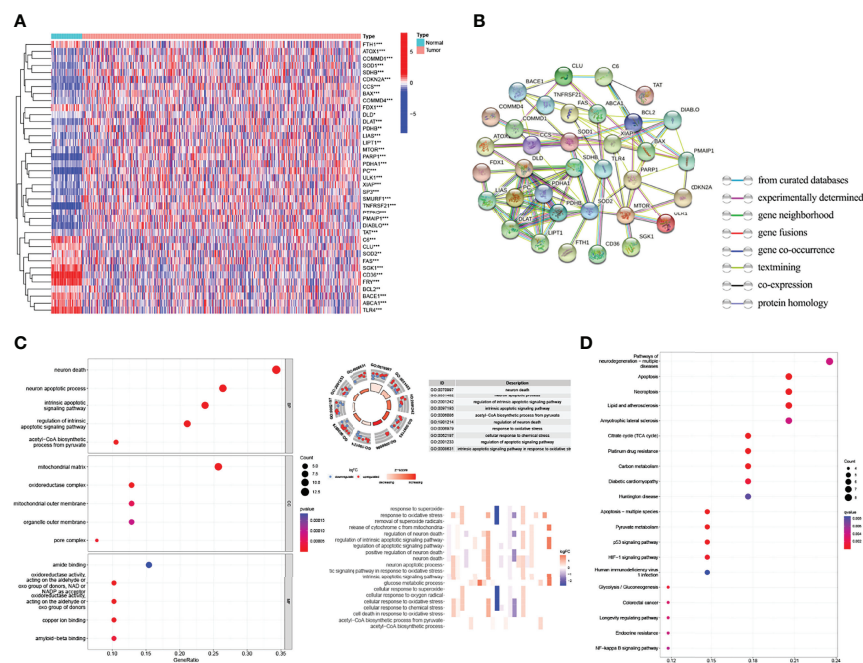


FIGURE 2
Identification of the expression and functional enrichment of cuproptosis-related genes in lung adenocarcinoma (LUAD). (A) Heat map reflecting the distribution of the 39 cuproptosis-related genes in LUAD and normal tissues. Red: upregulation; green: downregulation. (B) PPI network showing the interaction among 39 cuproptosis-related genes. (C) GO enrichment of cuproptosis-related genes. (D) Enriched KEGG pathways of cuproptosis-related genes. PPI, protein-protein interaction network; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

was 1.3927×10^{-12} (Figure 3C). The co-expression network between cuproptosis-related lncRNAs and genes were constructed in Figure 3D. Among these 7 cuproptosis-related lncRNAs, AC008764.2, AL022323.1, ELN-AS1, and LINC00578 were identified as protective lncRNAs, while AL031667.3, AL606489.1, and MIR31HG were identified as risk lncRNAs (Figure 3E).

Construction of a predictive risk model in LUAD patients

The 7 lncRNAs were identified in the risk model with “risk score” = $AL022323.1 \times (-0.379446) + AC008764.2 \times (-0.231185) + LINC00578 \times (-0.218454) + ELN-AS1 \times (-0.097027) + AL031667.3 \times (0.145940) + AL606489.1 \times 0.180937 + MIR31HG \times (0.235981)$. The LUAD patients in both training ($n = 252$) and testing ($n = 252$) cohorts were divided into high- and low-risk groups based on the median risk score (Figures 4A–D). The expression degrees of 7 lncRNAs in the two different groups were exhibited by a heat map (Figures 4E, F). Interestingly, patients with high risk had worse overall survival (OS) compared with patients with low risk as revealed by Kaplan–Meier analysis (Figures 4G, H).

Prognosis value of model lncRNAs in LUAD

Cox regression analyses pointed out that the score calculated by the corresponding model was an independent factor to predict the OS in LUAD relative to other clinical factors (Figures 5A, B). Meanwhile, the outcome of the ROC curve analysis suggested that the risk score tended to show more sensitivity and specificity than the other clinical features (risk score: AUC = 0.740) (Figure 5C).

Construction and detection of the predictive nomogram in LUAD

The nomogram models were established to validate the LUAD patients’ survival value of 1, 3, and 5 years (Figure 5D). Risk score was identified as independent prognostic factor ($***p < 0.001$). The results of the calibration curve revealed that the nomogram model showed a significant accuracy in predicting the LUAD patients’ OS (Figure 5E). The model likewise showed high sensitivity and efficacy (1-year AUC = 0.773, 3-year AUC = 0.753, and 5-year AUC = 0.805) (Figure 5F). In short, these results identified that both risk

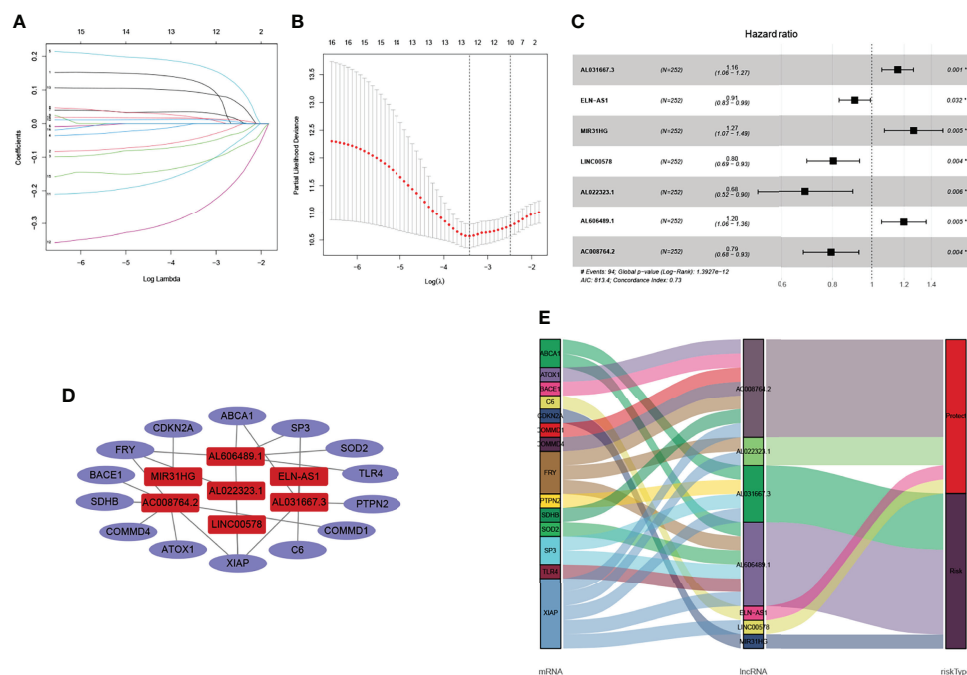


FIGURE 3

Identification of cuproptosis-related lncRNAs and co-expression network construction. (A, B) LASSO Cox algorithm was employed to establish a prognosis model. The color depth of the nodes depicted the corrected P -value of ontologies. The size of the nodes depicted the number of genes that are engaged in the ontologies. (C) A total of 7 lncRNAs were selected through subsequent multivariate analysis to construct the risk model. (D) Co-expression structure between cuproptosis-related lncRNAs and genes. (E) Sankey diagram was used to visualize the co-expression network.

TABLE 2 The clinicopathologic characteristics of 504 lung adenocarcinoma patients in The Cancer Genome Atlas.

Characteristics	Training cohort (n = 252)	Validation cohort (n = 252)	Entire set (n = 504)
Age			
≤65	134 (53.17%)	122 (48.41%)	256 (50.79%)
>65	118 (46.83%)	130 (51.59%)	248 (49.21%)
Gender			
Female	131 (51.98%)	139 (55.16%)	270 (53.57%)
Male	121 (48.02%)	113 (44.84%)	234 (46.43%)
T			
T1–T2	214 (84.92%)	223 (88.49%)	437 (86.71%)
T3–T4	37 (14.68%)	27 (10.71%)	64 (12.70%)
Unknown	1 (0.40%)	2 (0.80%)	3 (0.59%)
N			
N0	166 (65.87%)	159 (63.10%)	325 (64.48%)
N1–N3	80 (31.75%)	88 (34.92%)	168 (33.33%)
Unknown	6 (2.38%)	5 (1.98%)	11 (2.18%)
M			
M0	170 (67.46%)	167 (66.27%)	337 (66.87%)
M1	12 (4.76%)	14 (5.56%)	26 (5.16%)
Unknown	70 (27.78%)	71 (28.17%)	141 (27.98%)
Stage			
Stage I–stage II	202 (80.16%)	193 (76.59%)	395 (78.37%)
Stage III–stage IV	50 (19.84%)	59 (23.41%)	109 (21.63%)

model and nomogram model revealed that the overall survival rate can be predicted relatively well.

Correlation analysis of different groups and tumor microenvironment infiltration

To evaluate the role of the risk model in the immune microenvironment of LUAD, the CIBERSORT algorithm was performed to compare 22 different immune cell types in LUAD, finding that 7 of these immune cell types were differentially expressed in two risk groups ($*p < 0.05$ and $**p < 0.01$) (Figure 6A). Meanwhile, we discovered that, in the TCGA project, the low-risk group of the TCGA cohort had a significantly higher score of most immune-correlated pathways than the high-risk group (Figure 6B). In addition, the relationship between the risk model and infiltration of immune cells is described in Figure 6C, suggesting that there was a positive correlation between the survival outcome of LUAD patients and the high degrees of M0 macrophages, M1 macrophages, CD4 memory-activated T cells, and CD8 T cells, while there was a negative correlation between the survival outcome of LUAD patients and the high degrees of activated dendritic cells, resting dendritic cells, resting mast cells, monocytes, and resting CD4 memory T cells. To sum up, these results highlight the immunomodulatory effects of the risk model.

Construction of the lncRNA MIR31HG/miR-193a-3p/TNFRSF21 regulatory axis

To further understand the potential mechanism of cuproptosis-related lncRNAs in LUAD, we constructed the network of lncRNA–miRNA–mRNA interaction regulatory axis. According to Mircode database, we found that lncRNA MIR31HG bound miRNAs using “Perl” software (Figure 7A). Among these miRNAs, 2 miRNAs (miR-206 and miR-193a-3p) were identified to be less expressed in lung cancer (16, 17), which was contrary to the expression of the target lncRNA MIR31HG. Based on this result, we then explored its downstream mRNA targets to construct the miRNA–mRNA axis. According to the miRDB and TargetScan databases, cuproptosis-related mRNA (TNFRSF21) was identified as the downstream target of miR-193a-3p (Figures 7B, C). We then found that only TNFRSF21 had an upregulated expression in LUAD tissues according to the GEPIA (<http://gepia.cancer-pku.cn/>) database (Figure 7D).

lncRNA MIR31HG is overexpressed in LUAD tissues and cell lines

Figure 7E shows that lots of H3K27Ac marks existed in the promoter region of lncRNA MIR31HG from the UCSC web server. To understand the role of lncRNA MIR31HG in LUAD, qRT-PCR analysis, TCGA database, and GEO dataset

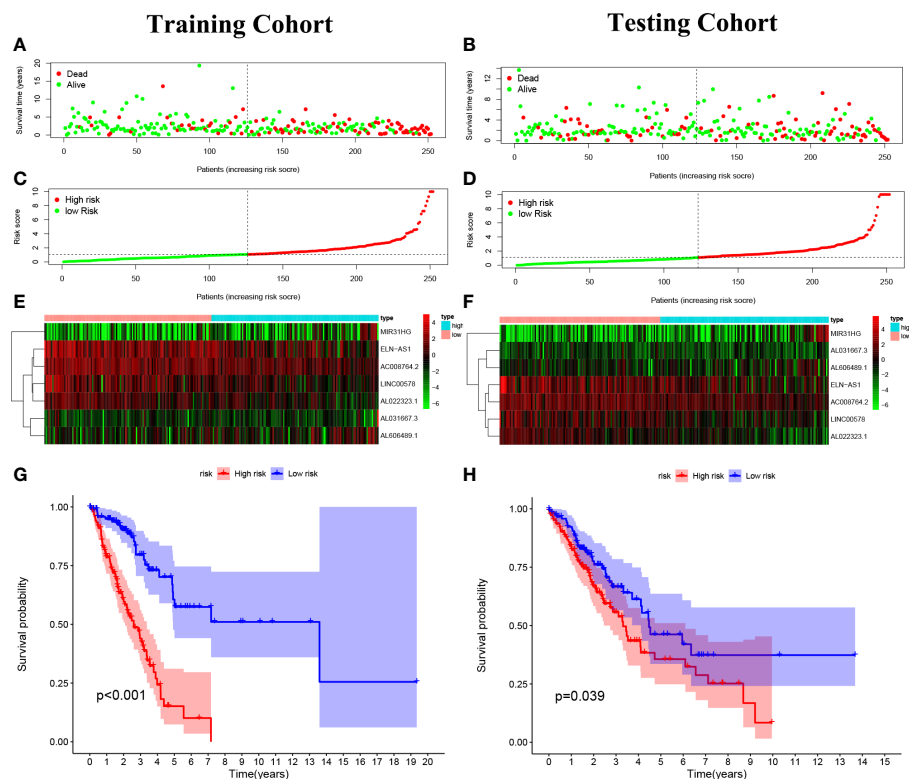


FIGURE 4

Construction of the predictive risk model in lung adenocarcinoma (LUAD) patients. (A, B) The risk score was calculated by 7 cuproptosis-related lncRNAs in the two cohorts (training and validation), and two risk groups were formed in LUAD patients. Green, low risk; red, high risk. (C, D) Survival status of LUAD patients. Green: survival; red: death. (E, F) The heat map indicates the expression degrees of 7 cuproptosis-related lncRNAs. (G, H) The Kaplan–Meier analysis revealed the overall survival in the two risk groups.

(GSE:130740) were used and identified the high level of MIR31HG in LUAD specimens relative to normal tissues (Figures 7F–H). We likewise selected several lung cancer cell lines (A549, NCI-H2009, and PC9) for experimental validation *in vitro*, with bronchial epithelioid cell (HBE) as the control group, and the result suggested that the expression of MIR31HG was also enhanced in lung cancer cell lines compared with the control group (Figure 7I). Additionally, we identified that LUAD patients with a higher expression level of lncRNA MIR31HG had a shorter OS time than those with a lower MIR31HG expression level (Figure 7J).

lncRNA MIR31HG mainly locate in the cytoplasm and its knockdown inhibits LUAD cell proliferation, migration, and invasion

The subcellular localization analysis of lncRNA plays an essential role in exploring the functional mechanism of lncRNA (18). Based on this, we adopted a one-step method to completely

isolate cytoplasmic RNA and nuclear RNA. The reverse transcription and assay analysis of lncRNA MIR31HG transcript levels by qRT-PCR revealed that the transcripts of lncRNA MIR31HG were mainly distributed in the cytoplasm of NCI-H2009 and A549 cells, which was consistent with the predicted result of “lncLocator” (<http://www.csbio.sjtu.edu.cn/bioinf/lncLocator/>) (Figures 8B–D). Subsequently, in order to determine whether lncRNA MIR31HG is involved in the initiation and progression of LUAD, functional interference techniques were used to evaluate the behavioral effects of lncRNA MIR31HG deletion. Figure 8A shows that the transfection was clearly successful in LUAD cell lines. The results of CCK-8 detection suggested that lncRNA MIR31HG interference could significantly inhibit the proliferation activity of NCI-H2009 and A549 cells (Figures 8E, F). Similarly, the colony formation assays revealed that the clone capacity of LUAD cells was remarkably inhibited by the silencing of lncRNA MIR31HG (Figures 8I, J). The data from wound healing and transwell analyses showed that NCI-H2009 and A549 cell lines exhibited significantly attenuated migration and invasion functions after interference with MIR31HG

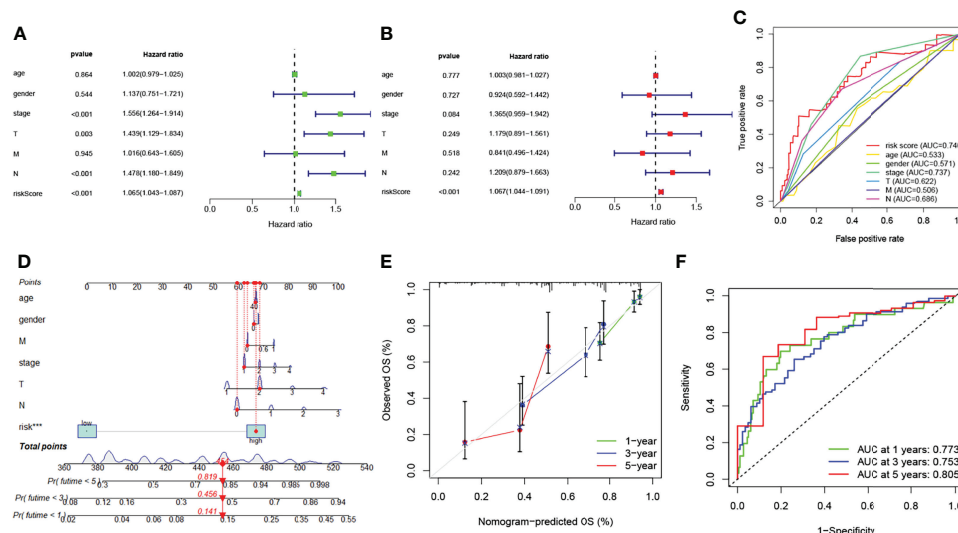


FIGURE 5

Prognosis value of model lncRNAs in lung adenocarcinoma (LUAD). (A) Univariate analysis of various clinical features and risk score. (B) Multivariate analysis of various clinical features and risk score. (C) ROC curves of the risk model (risk score: AUC = 0.740). (D) A nomogram was performed to predict the 1-, 3-, and 5-year survival. (E) Calibration curve of the nomogram model. (F) The results of ROC curves in predicting the LUAD survival rates. ROC, receiver operator characteristic; AUC, area under the curve.

(Figures 8G, H, K, L). These findings demonstrate that lncRNA MIR31HG played a key role in stimulating the progression of LUAD.

miR-193a-3p is sponged by lncRNA MIR31HG

To further explore the specific mechanism of lncRNA MIR31HG as ceRNA in LUAD, we identified the target miR-193a-3p by Mircode database (Figure 7A). In addition, we noticed that the expression level of miR-193a-3p was low in both LUAD tumor tissues and cell lines compared with normal groups, respectively (Figures 9A, B). qRT-PCR was investigated following si-MIR31HG or miR-193a-3p mimic transfection, revealing that silencing of lncRNA MIR31HG significantly promoted the expression of miR-193a-3p relative to “si-ctrl” (Figure 9C), whereas miR-193a-3p over-expression clearly decreased MIR31HG expression (Figure 9D). Figure 9E shows some binding sites between miR-193a-3p and MIR31HG 3' UTR. The luciferase assay clearly verified the potential relationship between lncRNA MIR31HG and miR-193a-3p (Figures 9F, G). The RIP analysis suggested that lncRNA MIR31HG and miR-193a-3p had obvious immunoprecipitation in Ago2 complex in LUAD cell lines NCI-H2009 and A549 (Figures 9H, I). In addition, we further found that there existed a negative correlation between lncRNA MIR31HG and miR-193a-3p (Figure 9J).

Interference of miR-193a-3p reverses the effect of lncRNA MIR31HG on LUAD cells

To further ascertain whether the effect of lncRNA MIR31HG on LUAD cells is affected by miR-193a-3p, we co-transfected miR-193a-3p inhibitor with si-MIR31HG. Figure 10A shows that the interference effect of miR-193a-3p is successful. The proliferation activity of LUAD cell lines NCI-H2009 and A549 was blocked by si-MIR31HG and rescued by the addition of miR-193a-3p inhibitor (Figures 10B, C). Similarly, colony formation assays established that the clone capacity of LUAD cells was remarkably inhibited by si-MIR31HG but returned by adding miR-193a-3p inhibitor (Figures 10D, E). Moreover, the ability of cell migration and invasion was simultaneously decreased by si-MIR31HG but restored by transfecting miR-193a-3p inhibitor (Figures 10F–J). These findings proved that the MIR31HG interference inhibited the malignant activities of LUAD cells *via* upregulating miR-193a-3p.

miR-193a-3p targets downstream TNFRSF21

The cuproptosis-related mRNA (TNFRSF21) was identified as the downstream target of miR-193a-3p from miRDB and TargetScan databases (Figure 7C). Firstly, we noticed that TNFRSF21 was clearly upregulated in both LUAD tumor tissues and NCI-H2009 and A549 cells compared with normal



result of qRT-PCR identified that lncRNA MIR31HG inhibitor significantly inhibited the expression of TNFRSF21, but these alterations were reversed by miR-193a-3p inhibitor (Figure 11K), which was consistent with the outcome of the Western blot (Figure 11L). Figure 11M shows the schematic diagram of the mechanism of the lncRNA MIR31HG/miR-193a-3/TNFRSF21 regulatory axis.

Discussion

Cuproptosis, as a new type of death, is receiving more and more attention. Our study systematically identified cuproptosis-related lncRNAs based on the involvement of lncRNAs and cuproptosis-related mRNAs in LUAD. After that, a unique risk

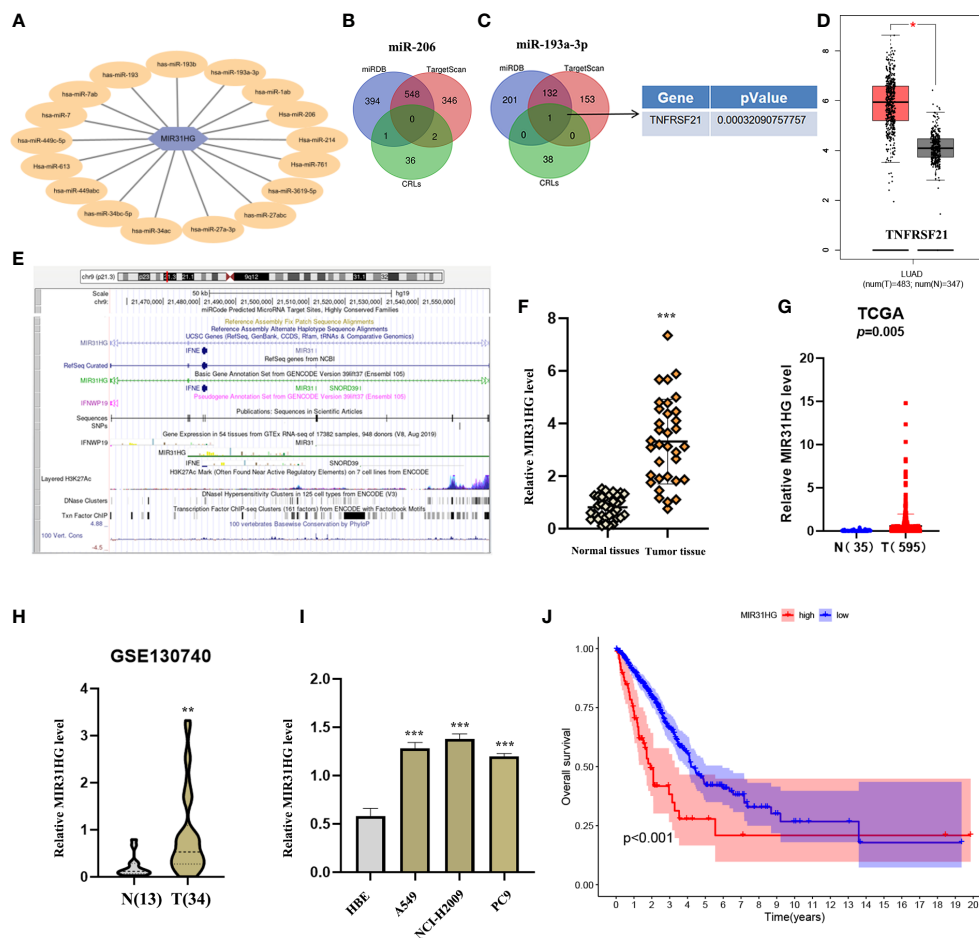


FIGURE 7

Construction of a regulatory axis of lncRNA-miRNA-mRNA. (A) lncRNA MIR31HG bound to 17 miRNAs. (B, C) Venn diagram identifying the downstream targets in miR-206 and miR-193a-3p, respectively, from miRDB and TargetScan databases. (D) The expression of TNFRSF21 was investigated by GEPIA database. (E) Lots of H3K27Ac marks existed in the promoter region of MIR31HG from the UCSC web server. (F, G) MIR31HG expression in lung adenocarcinoma (LUAD) specimens relative to normal tissues as detected by qRT-PCR and The Cancer Genome Atlas cohort. (H) MIR31HG expression in LUAD specimens relative to normal tissues as detected by GSE130740 cohort. (I) MIR31HG expression in different LUAD cell lines (A549, NCI-H2009, and PC9) compared with bronchial epithelioid cell (HBE) was estimated by qRT-PCR. (J) The overall survival time of patients with LUAD was measured by Kaplan-Meier analysis. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

model with various features was constructed to predict the prognosis of LUAD patients. Firstly, 39 differentially expressed cuproptosis-related genes were identified as “DEGs” from all cuproptosis-related genes by comparing LUAD and normal tissues. Both GO and KEGG databases indicated that these “DEGs” were enriched in the pathways of neurodegenerative diseases. In the nervous system, copper is involved in myelination, excitotoxic cell death, synaptic activity, and neurotrophic factor-induced signaling cascades (19). Aberrant copper homeostasis could lead to multiple pathological sequelae, including cancer, inflammation, and neurodegeneration. Correcting disturbed copper homeostasis is a promising therapeutic strategy for neurodegenerative diseases (20). Secondly, a risk model consisting of 7 lncRNAs was built to

form two risk groups of LUAD patients. Thirdly, the risk score was proved to be an independent prognostic factor with a high degree of sensitivity and specificity compared with other clinical factors. Moreover, the analysis of the infiltration of immune cells and the risk model highlighted the immunomodulatory effects of cuproptosis-related lncRNAs in these two risk groups. Finally, we predicted and verified the network of lncRNA MIR31HG/miR-193a-3p/TNFRSF21, which may play a potential role in the progression of LUAD.

Multiple studies have confirmed that lncRNAs played important roles in different aspects of LUAD development—for example, Deng *et al.* reported that the lncRNA LINC00472 inhibits the migration and invasion of LUAD by regulating YBX1 (21). Qu *et al.* discovered that PD-L1 lncRNA splice

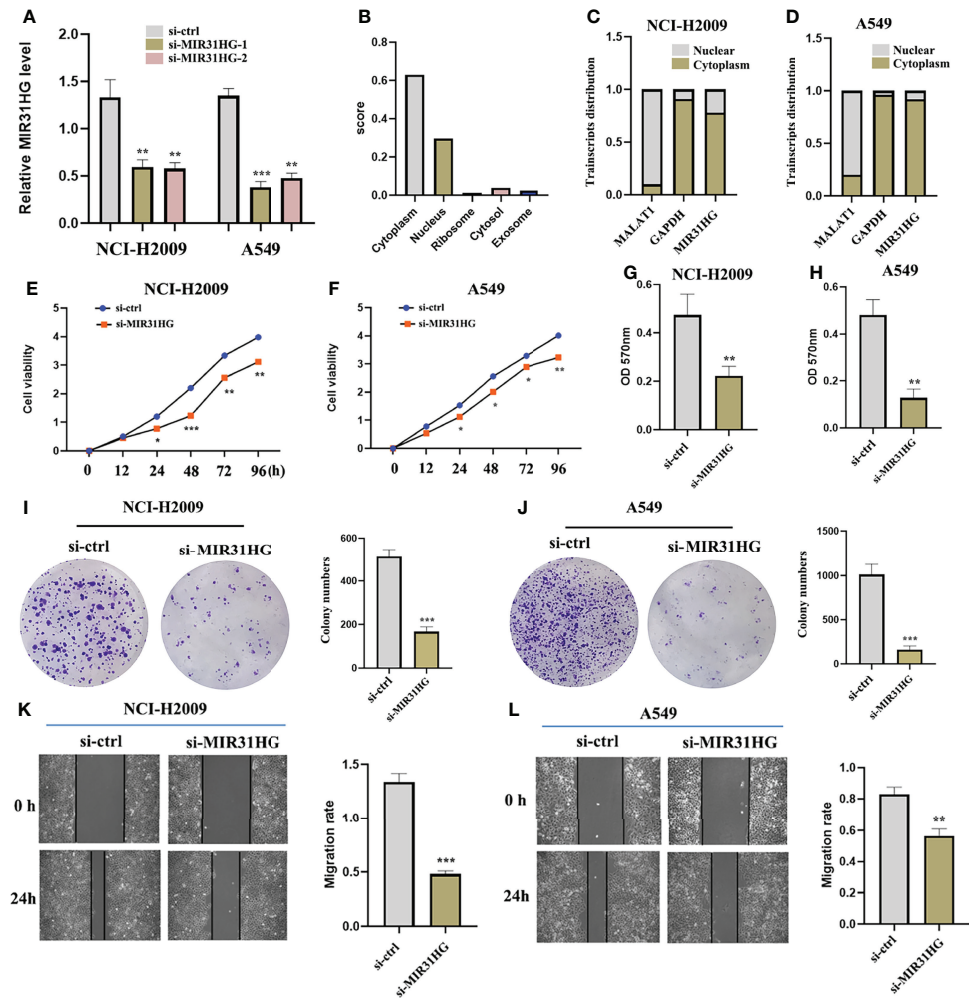


FIGURE 8

The effect of lncRNA MIR31HG on lung adenocarcinoma LUAD cell viability, migration, and invasion. (A) The efficiency of MIR31HG knockdown (si-MIR31HG-1 and si-MIR31HG-2) was assessed by qRT-PCR. (B) The subcellular localization of MIR31HG was predicted by "LncLocator". (C, D) The relative lncRNA MIR31HG expression level both in the cytoplasm and the nucleus of the NCI-H2009 and A549 cell lines were simultaneously measured by qRT-PCR. (E, F) The proliferation of NCI-H2009 and A549 cells was detected by CCK-8 assays. (G, H) The invasion of NCI-H2009 and A549 cells was investigated by transwell assays. (I, J) The clone capacity of NCI-H2009 and A549 cells was identified by colony formation assay. (K, L) The migration of NCI-H2009 and A549 cells was determined by wound healing assays. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

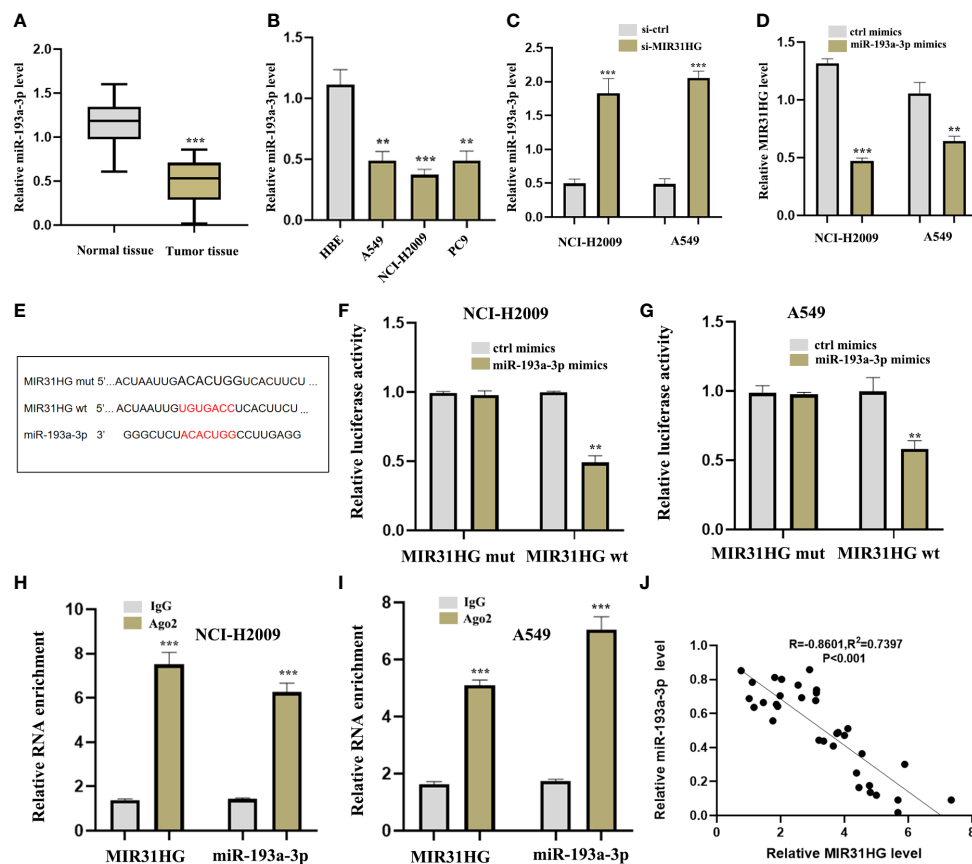


FIGURE 9

IncRNA MIR31HG acted as ceRNA for miR-193a-3p. (A) miR-193a-3p expression in lung adenocarcinoma (LUAD) specimens compared with normal tissues as detected by qRT-PCR. *** $p < 0.001$ vs. normal tissue. (B) miR-193a-3p expression in different LUAD cell lines (A549, NCI-H2009, and PC9) relative to bronchial epithelioid cell (HBE) was estimated by qRT-PCR. (C) miR-193a-3p expression following si-MIR31HG transfection was assessed by qRT-PCR. (D) MIR31HG expression following miR-193a-3p overexpression was measured by qRT-PCR. (E) Schematic diagram of the predicted interacting sites. (F, G) The relationship between MIR31HG and miR-193a-3p in NCI-H2009 and A549 cells was performed by dual-luciferase reporter assay. (H, I) The immunoprecipitation of MIR31HG and miR-193a-3p in NCI-H2009 and A549 cells was determined by RNA immunoprecipitation experiment. (J) The relationship between MIR31HG and miR-193a-3p was investigated by Pearson's analysis. ** $p < 0.01$; *** $p < 0.001$.

isoform stimulated the progression of LUAD via the c-Myc axis (22). However, studies on cuproptosis-related lncRNAs in LUAD, especially their potential ability to predict prognosis in these LUAD patients, are fairly inadequate. Therefore, our work established a predictive model on the basis of 7 cuproptosis-related lncRNAs, including AL031667.3, ELN-AS1, LINC00578, AL022323.1, AL606489.1, AC008764.2, and MIR31HG. This means that, among these lncRNAs, some lncRNAs have been reported to be involved in the pathogenesis of several tumor diseases—for instance, Zheng *et al.* manifested that lncRNA AL031667.3 played a risk biomarker role in the prognosis of lung adenocarcinoma (23). Wang *et al.* detected that lncRNA ELN-AS1 was identified as a protective factor of endometrial cancer patients (24). Moreover, lncRNA LINC00578 could inhibit the tumor proliferation of pancreatic cancer (25) and lung adenocarcinoma (26). In addition, lncRNA AL022323.1 was

identified as a protective factor of colorectal cancer with low aggression (27). Guo *et al.* revealed that lncRNA AL606489.1 was referred to as a prognostic marker and dangerous effector in lung adenocarcinoma (28). lncRNA MIR31HG has been reported to promote glioblastoma progression by regulating Wnt/ β -catenin signaling (29). These reports likewise supported our conclusion that lncRNAs AL031667.3, AL606489.1, and MIR31HG were the dangerous factors of LUAD, while ELN-AS1, LINC00578, and AL022323.1, as protective factors, played an important role in prolonging LUAD patients' survival time, which were consistent with what we found in this research.

Immunity therapy is closely related to the prognosis of cancer patients, and its positive response usually depends on the dynamic regulation between tumor cells and immune modulators in the tumor microenvironment (TME) (30).

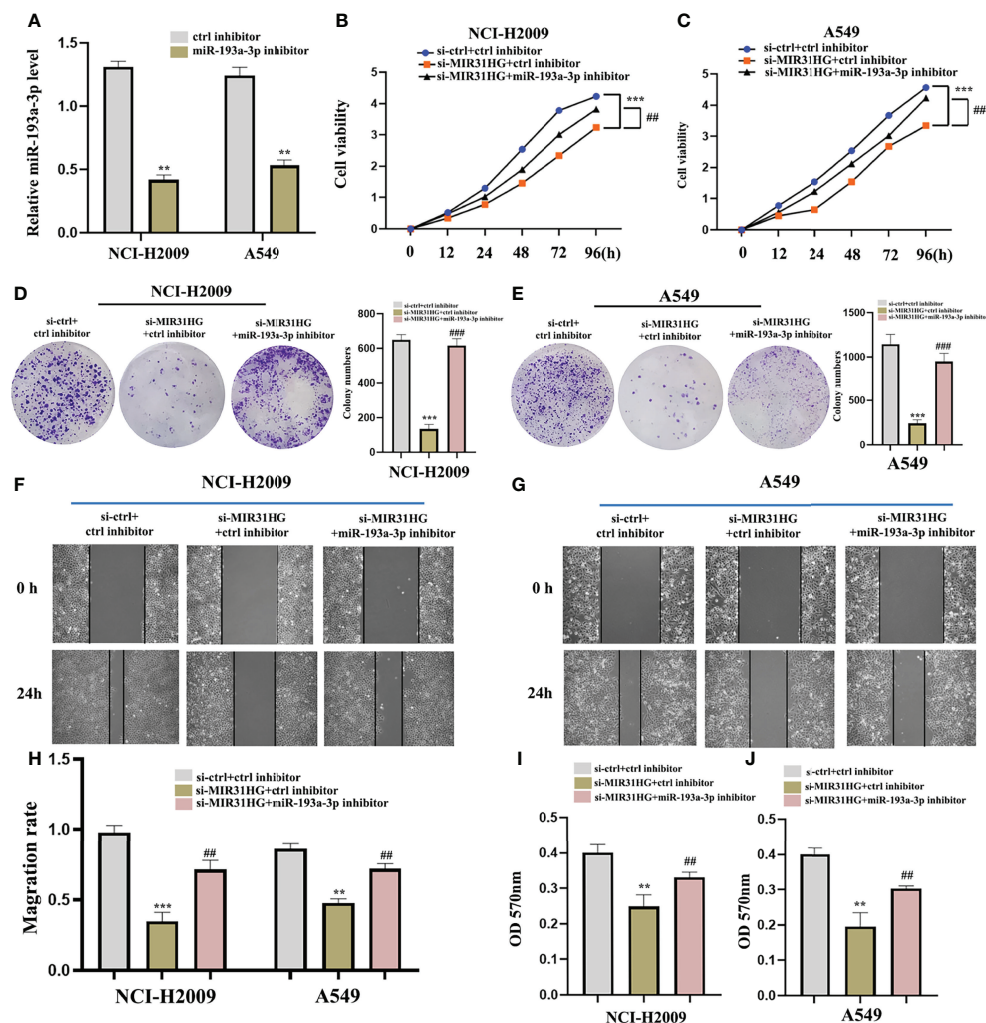


FIGURE 10

miR-193a-3p reversed the lncRNA MIR31HG knockdown effects on lung adenocarcinoma cells. (A) The expression of miR-193a-3p was investigated by qRT-PCR. (B, C) The proliferation of NCI-H2009 and A549 cells were detected by CCK-8 assays. (D, E) The clone capacity of NCI-H2009 and A549 cells was identified by colony formation assay. (F–H) The migration of NCI-H2009 and A549 cells was determined by wound healing assays. (I, J) The invasion of NCI-H2009 and A549 cells was investigated by transwell assays. ** $p < 0.01$; *** $p < 0.001$ vs. ctrl inhibitor, ## $p < 0.01$; ### $p < 0.01$ vs. si-MIR31HG+miR-193a-3p inhibitor.

Therefore, effective exploration of the immunological characteristics of the TME will be conducive to a rapid discovery of a variety of new immunity therapy strategies and identification of more potential clinical prognostic biomarkers (31, 32). From the results of our study, immune-correlated infiltrating cells and immune-interrelated pathways were found to be mostly concentrated in the low-risk group, indicating that immune-suppressive therapy might be more effective in the low-risk group of LUAD patients.

A huge number of studies have proved that lncRNAs could sponge microRNA (miRNA) loci and regard as competing endogenous RNAs (ceRNAs), thus effecting and adjusting the biological activities of downstream mRNAs (33, 34). The

lncRNA-correlated ceRNAs have been recently elucidated to play an irreplaceable role in the development of various cancers—for example, lncRNA-CDC6 promoted breast cancer progression by regulating the axis of microRNA-215/CDC6 (35). lncRNA HOXD-AS1 may, as a ceRNA, stimulate liver cancer metastasis (36). lncRNA MT1JP functioned as a ceRNA to regulate miR-92a-3p/FBXW7 in gastric cancer (37). These results discovered that the axis had been extensively explored and reported in many diseases. Therefore, the novel network of lncRNA MIR31HG/miR-193a-3p/TNFRSF21 in LUAD was constructed by using biological tools.

In non-small cell lung cancer, lncRNA MIR31HG could sponge miR-241 to upregulate SP1, thus stimulating tumor

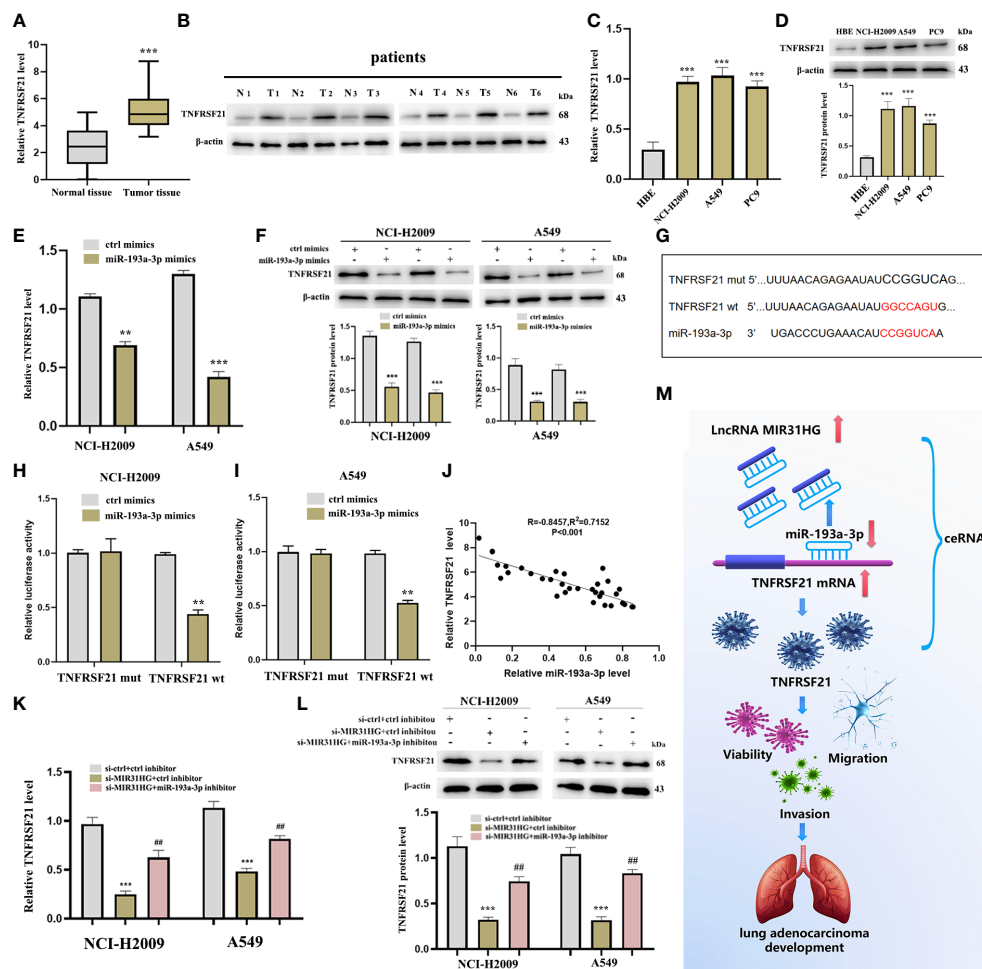


FIGURE 11

TNFRSF21 was a downstream target of miR-193a-3 and lncRNA MIR31HG sponged miR-193a-3p to upregulate TNFRSF21. (A) TNFRSF21 expression in lung adenocarcinoma (LUAD) specimens compared with normal tissues as detected by qRT-PCR. *** $p < 0.001$ vs. normal tissue. (B) TNFRSF21 expression in 6 LUAD patients relative to normal tissues as detected by western blot. (C, D) TNFRSF21 expression in different LUAD cell lines (A549, NCI-H2099, and PC9) relative to bronchial epithelioid cell (HBE) was estimated by qRT-PCR or western blot. *** $p < 0.001$ vs. HBE. (E, F) TNFRSF21 expression following miR-193a-3p overexpression was assessed by qRT-PCR or western blot. ** $p < 0.01$; *** $p < 0.001$ vs. ctrl mimics. (G) Schematic diagram of the predicted interacting sites. (H, I) The relationship between TNFRSF21 and miR-193a-3p in NCI-H2099 and A549 cells was performed by dual-luciferase reporter assay. ** $p < 0.01$ vs. ctrl mimics. (J) Relationship between TNFRSF21 and miR-193a-3p investigated by Pearson's analysis. (K, L) TNFRSF21 expression following si-MIR31HG or si-MIR31HG + miR-193a-3p inhibitor was assessed by qRT-PCR or western blot. *** $p < 0.001$ vs. ctrl inhibitor; ## $p < 0.01$ vs. si-MIR31HG + miR-193a-3p inhibitor. (M) Schematic diagram of the mechanism of lncRNA MIR31HG/miR-193a-3p/TNFRSF21 regulatory axis.

progression (38). Upregulation of lncRNA MIR31HG acted as an oncogene by stimulating the Wnt/ β -catenin axis in lung cancer (39). In this work, we illustrated that lncRNA MIR31HG and TNFRSF21 were over-expressed, while miR-193a-3p was downregulated in human LUAD tumor tissues and cells. Moreover, functional experiments were executed using si-MIR31HG transfection, revealing that the proliferation, migration, and invasion of LUAD cells were inhibited following the interference of MIR31HG, which were consistent with previous reports on lncRNA MIR31HG. In addition, Liu *et al.* detected that miR-193a-3p expression was decreased and

could act as a tumor suppressor in lung cancer (40). This report supported our experimental results on the low expression of miR-193a-3p in LUAD. In our study, we predicted that lncRNA MIR31HG might interact with miR-193a-3p according to Mircode database. The RIP analysis suggested that lncRNA MIR31HG could directly integrate with miR-193a-3p in the level of Ago2 complex, and we also found a negative correlation between lncRNA MIR31HG and miR-193a-3p in LUAD. Moreover, the knockdown of miR-193a-3p could partly weaken the effect of lncRNA MIR31HG interference on LUAD cells. These findings not only further defined the tumor

suppressor properties of miR-193a-3p but also identified that miR-193a-3p was involved in the development of LUAD through the ceRNA regulatory pattern. Based on this result, we then explored miR-193a-3p' downstream mRNA target. Finally, the cuproptosis-related mRNA (TNFRSF21) was identified as the downstream target of miR-193a-3p from miRDB and TargetScan databases. Our work found that TNFRSF21 showed a high expression level in NCI-H2009 and A549 cells compared with the normal group, which was supported by RT-qPCR and western blot. More interestingly, the regulation between lncRNA MIR31HG and TNFRSF21 was mediated by miR-193a-3p. In short, we concluded that lncRNA MIR31HG upregulated TNFRSF21 through sponging miR-193a-3p, which might be the indispensable mechanism of lncRNA MIR31HG-regulated LUAD progression.

Unfortunately, there are still many defects in our current research that need further improvement. First of all, the risk model in this study was mainly established from the TCGA LUAD cohort, so it is best to use the GEO cohort to further verify the accuracy of the LUAD patients' prognosis. Secondly, there is a lack of analysis of the relationship between cuproptosis and lipid metabolism TCA. Moreover, we preliminarily predicted and verified that the network of lncRNA MIR31HG/miR-193a-3p/TNFRSF21 might play a potential role in LUAD *in vitro*, but more *in vivo* tests are still needed for deep verification, which is also the focus of our future work.

Conclusion

To summarize, we successfully identified 7 cuproptosis-related lncRNAs—AL031667.3, ELN-AS1, LINC00578, AL022323.1, AL606489.1, AC008764.2, and MIR31HG. On the basis of these lncRNAs, a valid predictive model was established for LUAD patients' clinical prognosis, which proved to be an effective independent factor compared with other clinical features. The correlation between cuproptosis-related lncRNAs and immune infiltration was elucidated based on the overall risk score of groups, which would lay a foundation to improve anti-tumor immunity and develop a new treatment system for LUAD. Interestingly, our research also predicted and verified the network of lncRNA MIR31HG/miR-193a-3p/TNFRSF21. We revealed the oncogenic function of lncRNA MIR31HG in the malignant progression of LUAD and remarkably identified its potential mechanism by regulating the miR-193a-3p/TNFRSF21 axis, which might be beneficial to further elucidate the pathogenesis of LUAD and provide new ideas for clinical treatment.

Data availability statement

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

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of The First Affiliated Hospital of Jinan University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

XM, DH, and MX conceived and designed the experiments. XM and DH conducted the research. XM, DH, PY, MX, and YL contributed materials and analysis tools. XM, DH, PY, MX, AN, FM, SB, and GJ analyzed the results. XM and DH wrote the paper. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (no. 81774376) and the Science and Technology Foundation of Guangzhou (no. 201803010059).

Acknowledgments

We thank the investigators and patients in the TCGA and GEO for providing data.

Conflict of interest

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

Publisher's note

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

- Kuhn E, Morbini P, Cancellieri A, Damiani S, Cavazza A, Comin CE. Adenocarcinoma classification: Patterns and prognosis. *Pathologica* (2018) 110(1):5–11.
- Zhang Y, Fu F, Chen H. Management of ground-glass opacities in the lung cancer spectrum. *Ann Thorac Surg* (2020) 110(6):1796–804. doi: 10.1016/j.athoracsur.2020.04.094
- Dong S, Men W, Yang S, Xu S. Identification of lung adenocarcinoma biomarkers based on bioinformatic analysis and human samples. *Oncol Rep* (2020) 43(5):1437–50. doi: 10.3892/or.2020.7526
- Li Z, Ma G, Pan Y. Five circRNAs serve as potential diagnostic and prognostic biomarkers in lung adenocarcinoma. *Clin Lab* (2021) 67(8). doi: 10.7754/Clin.Lab.2020.200523
- Li Y, Gu J, Xu F, Zhu Q, Chen Y, Ge D, et al. Molecular characterization, biological function, tumor microenvironment association and clinical significance of M6a regulators in lung adenocarcinoma. *Brief Bioinform* (2021) 22(4):bbaa225. doi: 10.1093/bib/bbaa225
- Tsvetkov P, Coy S, Petrova B, Dreishpoon M, Verma A, Abdusamad M, et al. Copper induces cell death by targeting lipoylated TCA cycle proteins. *Science* (2022) 375(6586):1254–61. doi: 10.1126/science.abf0529
- Jiang Y, Huo Z, Qi X, Zuo T, Wu Z. Copper-induced tumor cell death mechanisms and antitumor therapeutic applications of copper complexes. *Nanomedicine (Lond)* (2022) 17(5):303–24. doi: 10.2217/nnm-2021-0374
- Cui L, Gouw AM, LaGory EL, Guo S, Attarwala N, Tang Y, et al. Mitochondrial copper depletion suppresses triple-negative breast cancer in mice. *Nat Biotechnol* (2021) 39(3):357–67. doi: 10.1038/s41587-020-0707-9
- Isin M, Dalay N. LncRNAs and neoplasia. *Clin Chim Acta* (2015) 444:280–8. doi: 10.1016/j.cca.2015.02.046
- Peng WX, Koirala P, Mo YY. LncRNA-mediated regulation of cell signaling in cancer. *Oncogene* (2017) 36(41):5661–7. doi: 10.1038/ncr.2017.184
- Pan J, Fang S, Tian H, Zhou C, Zhao X, Tian H, et al. LncRNA JPX/miR-33a-5p/Twist1 axis regulates tumorigenesis and metastasis of lung cancer by activating wnt/ β -catenin signaling. *Mol Cancer* (2020) 19(1):9. doi: 10.1186/s12943-020-1133-9
- Zhao M, Xin XF, Zhang JY, Dai W, Lv TF, Song Y. LncRNA GMD5-AS1 inhibits lung adenocarcinoma development by regulating miR-96-5p/CYLD signaling. *Cancer Med* (2020) 9(3):1196–208. doi: 10.1002/cam4.2776
- Wang W, Zhao Z, Xu C, Li C, Ding C, Chen J, et al. LncRNA FAM83A-AS1 promotes lung adenocarcinoma progression by enhancing the pre-mRNA stability of FAM83A. *Thorac Cancer* (2021) 12(10):1495–502. doi: 10.1111/1759-7714.13928
- Kim BE, Nevitt T, Thiele DJ. Mechanisms for copper acquisition, distribution and regulation. *Nat Chem Biol* (2008) 4(3):176–85. doi: 10.1038/nchembio.72
- Ge EJ, Bush AI, Casini A, Cobine PA, Cross JR, DeNicola GM, et al. Connecting copper and cancer: From transition metal signalling to metalloplasia. *Nat Rev Cancer* (2022) 22(2):102–13. doi: 10.1038/s41568-021-00417-2
- Li W, Liu Y, Li ZJ, Shi Y, Deng J, Bai J, et al. Unravelling the role of LncRNA WT1-AS/miR-206/NAMPT axis as prognostic biomarkers in lung adenocarcinoma. *Biomolecules* (2021) 11(2):203. doi: 10.3390/biom11020203
- Fan Q, Hu X, Zhang H, Wang S, Zhang H, You C, et al. MiR-193a-3p is an important tumour suppressor in lung cancer and directly targets KRAS. *Cell Physiol Biochem* (2017) 44(4):1311–24. doi: 10.1159/000485491
- Yoon JH, Abdelmohsen K, Gorospe M. Posttranscriptional gene regulation by long noncoding RNA. *J Mol Biol* (2013) 425(19):3723–30. doi: 10.1016/j.jmb.2012.11.024
- Gromadzka G, Tarnacka B, Flaga A, Adamczyk A. Copper dyshomeostasis in neurodegenerative diseases-therapeutic implications. *Int J Mol Sci* (2020) 21(23):9259. doi: 10.3390/ijms21239259
- Choo XY, Alukaidey L, White AR, Grubman A. Neuroinflammation and copper in alzheimer's disease. *Int J Alzheimers Dis* (2013) 2013:145345. doi: 10.1155/2013/145345
- Deng X, Xiong W, Jiang X, Zhang S, Li Z, Zhou Y, et al. LncRNA LINC00472 regulates cell stiffness and inhibits the migration and invasion of lung adenocarcinoma by binding to YBX1. *Cell Death Dis* (2020) 11(11):945. doi: 10.1038/s41419-020-03147-9
- Qu S, Jiao Z, Lu G, Yao B, Wang T, Rong W, et al. PD-L1 lncRNA splice isoform promotes lung adenocarcinoma progression via enhancing c-myc activity. *Genome Biol* (2021) 22(1):104. doi: 10.1186/s13059-021-02331-0
- Zheng Z, Zhang Q, Wu W, Xue Y, Liu S, Chen Q, et al. Identification and validation of a ferroptosis-related long non-coding RNA signature for predicting the outcome of lung adenocarcinoma. *Front Genet* (2021) 12:690509. doi: 10.3389/fgene.2021.690509
- Wang Z, Liu Y, Zhang J, Zhao R, Zhou X, Wang H. An immune-related long noncoding RNA signature as a prognostic biomarker for human endometrial cancer. *J Oncol* (2021) 2021:9972454. doi: 10.1155/2021/9972454
- Zhang B, Li C, Sun Z. Long non-coding RNA LINC00346, LINC00578, LINC00673, LINC00671, LINC00261, and SNHG9 are novel prognostic markers for pancreatic cancer. *Am J Transl Res* (2018) 10(8):2648–58.
- Wang L, Zhao H, Xu Y, Li J, Deng C, Deng Y, et al. Systematic identification of lincRNA-based prognostic biomarkers by integrating lincRNA expression and copy number variation in lung adenocarcinoma. *Int J Cancer* (2019) 144(7):1723–34. doi: 10.1002/ijc.31865
- Wei J, Ge X, Tang Y, Qian Y, Lu W, Jiang K, et al. An autophagy-related long noncoding RNA signature contributes to poor prognosis in colorectal cancer. *J Oncol* (2020) 2020:4728947. doi: 10.1155/2020/4728947
- Guo Y, Qu Z, Li D, Bai F, Xing J, Ding Q, et al. Identification of a prognostic ferroptosis-related lncRNA signature in the tumor microenvironment of lung adenocarcinoma. *Cell Death Discovery* (2021) 7(1):190. doi: 10.1038/s41420-021-00576-z
- Zhang R, Wu D, Wang Y, Wu L, Gao G, Shan D. LncRNA MIR31HG is activated by STAT1 and facilitates glioblastoma cell growth via wnt/ β -catenin signaling pathway. *Neurosci Res* (2021) S0168-0102(21):00092–4. doi: 10.1016/j.neures.2021.04.008
- Wu T, Dai Y. Tumor microenvironment and therapeutic response. *Cancer Lett* (2017) 387:61–8. doi: 10.1016/j.canlet.2016.01.043
- Xiao Y, Yu D. Tumor microenvironment as a therapeutic target in cancer. *Pharmacol Ther* (2021) 221:107753. doi: 10.1016/j.pharmthera.2020.107753
- Sokratous G, Polyzoidis S, Ashkan K. Immune infiltration of tumor microenvironment following immunotherapy for glioblastoma multiforme. *Hum Vaccin Immunother* (2017) 13(11):2575–82. doi: 10.1080/21645515.2017.1303582
- Zhou RS, Zhang EX, Sun QF, Ye ZJ, Liu JW, Zhou DH, et al. Integrated analysis of lncRNA-miRNA-mRNA ceRNA network in squamous cell carcinoma of tongue. *BMC Cancer* (2019) 19(1):779. doi: 10.1186/s12885-019-5983-8
- Braga EA, Fridman MV, Moscovtsev AA, Filippova EA, Dmitriev AA, Kushlinskii NE. LncRNAs in ovarian cancer progression, metastasis, and main pathways: ceRNA and alternative mechanisms. *Int J Mol Sci* (2020) 21(22):8855. doi: 10.3390/ijms21228855
- Kong X, Duan Y, Sang Y, Li Y, Zhang H, Liang Y, et al. LncRNA-CDC6 promotes breast cancer progression and function as ceRNA to target CDC6 by sponging microRNA-215. *J Cell Physiol* (2019) 234(6):9105–17. doi: 10.1002/jcp.27587
- Wang H, Huo X, Yang XR, He J, Cheng L, Wang N, et al. STAT3-mediated upregulation of lncRNA HOXD-AS1 as a ceRNA facilitates liver cancer metastasis by regulating SOX4. *Mol Cancer* (2017) 16(1):136. doi: 10.1186/s12943-017-0680-1
- Zhang G, Li S, Lu J, Ge Y, Wang Q, Ma G, et al. LncRNA MT1JP functions as a ceRNA in regulating FBXW7 through competitively binding to miR-92a-3p in gastric cancer. *Mol Cancer* (2018) 17(1):87. doi: 10.1186/s12943-018-0829-6
- Dandan W, Jianliang C, Haiyan H, Hang M, Xuedong L. Long noncoding RNA MIR31HG is activated by SP1 and promotes cell migration and invasion by sponging miR-214 in NSCLC. *Gene* (2019) 692:223–30. doi: 10.1016/j.gene.2018.12.077
- Zheng S, Zhang X, Wang X, Li J. MIR31HG promotes cell proliferation and invasion by activating the wnt/ β -catenin signaling pathway in non-small cell lung cancer. *Oncol Lett* (2019) 17(1):221–9. doi: 10.3892/ol.2018.9607
- Liu X, Min S, Wu N, Liu H, Wang T, Li W, et al. miR-193a-3p inhibition of the slug activator PAK4 suppresses non-small cell lung cancer aggressiveness via the P53/Slug/L1CAM pathway. *Cancer Lett* (2019) 447:56–65. doi: 10.1016/j.canlet.2019.01.027



Deep-LC: A Novel Deep Learning Method of Identifying Non-Small Cell Lung Cancer-Related Genes

Mo Li[†], Guang xian Meng[†], Xiao wei Liu, Tian Ma, Ge Sun^{*} and HongMei He^{*}

Second Affiliated Hospital of Dalian Medical University, Dalian, China

OPEN ACCESS

Edited by:

Tianyi Zhao,
Harbin Institute of Technology, China

Reviewed by:

Ningyi Zhang,
Harbin Institute of Technology, China
Sheng Li,
Zhongnan Hospital, Wuhan University,
China

*Correspondence:

Ge Sun
sunge86@126.com
HongMei He
hongmeihe1976@163.com

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

Received: 21 May 2022

Accepted: 16 June 2022

Published: 22 July 2022

Citation:

Li M, Meng Gx, Liu Xw, Ma T,
Sun G and He H (2022) Deep-LC: A
Novel Deep Learning Method of
Identifying Non-Small Cell Lung
Cancer-Related Genes.
Front. Oncol. 12:949546.
doi: 10.3389/fonc.2022.949546

According to statistics, lung cancer kills 1.8 million people each year and is the main cause of cancer mortality worldwide. Non-small cell lung cancer (NSCLC) accounts for over 85% of all lung cancers. Lung cancer has a strong genetic predisposition, demonstrating that the susceptibility and survival of lung cancer are related to specific genes. Genome-wide association studies (GWASs) and next-generation sequencing have been used to discover genes related to NSCLC. However, many studies ignored the intricate interaction information between gene pairs. In the paper, we proposed a novel deep learning method named Deep-LC for predicting NSCLC-related genes. First, we built a gene interaction network and used graph convolutional networks (GCNs) to extract features of genes and interactions between gene pairs. Then a simple convolutional neural network (CNN) module is used as the decoder to decide whether the gene is related to the disease. Deep-LC is an end-to-end method, and from the evaluation results, we can conclude that Deep-LC performs well in mining potential NSCLC-related genes and performs better than existing state-of-the-art methods.

Keywords: non-small cell lung cancer, genome-wide association analysis, graph convolutional networks, convolutional neural network (CNN) accelerator, Deep-LC

INTRODUCTION

Statistics show that lung cancer causes 1.8 million deaths each year and remains the leading cause of cancer deaths all over the world (1). Small cell lung cancer (SCLC) and non-SCLC (NSCLC) are two main types. NSCLC accounts for almost 85% of all types of lung cancer (2). Lung cancer has a strong genetic predisposition, and the specific genes are responsible for enhanced risk (3), in addition to being affected by external incentives such as smoking, secondhand or passive smoking, alcohol, and air pollution (4).

Genome-wide association studies (GWASs) have been widely used to identify which genes are related to lung cancer. Hung et al. (5) first utilized GWAS to examine single-nucleotide polymorphisms (SNPs) and discovered a locus in chromosome region 15q25 that was substantially linked to lung cancer. Six genes are found, including three subunits of the nicotinic acetylcholine receptor (CHRNA5, CHRNA3, and CHRNB4). Hu et al. (6) did GWAS on 5,408 subjects and demonstrated that the 5p15 locus is specific to lung cancer. In addition, they found that

an independent locus, 22q12.2, may be linked to the susceptibility to lung cancer. Genes are associated not only with the susceptibility to lung cancer but also with lung cancer survival. The 9p21.3 locus was demonstrated to be linked to susceptibility (7) and survival (8).

In addition to GWAS, some studies discovered new variants through next-generation sequencing (NGS), like whole-exome sequencing (WES) and whole-genome sequencing (WGS). Xiong et al. (9) found an uncommon mutation in PARK2 that causes the tumor suppressor gene to lose function in a five-generation family with lung cancer. Exome sequencing of sporadic and familial lung cancer patients also revealed infrequent detrimental mutations in GWAS-nominated sites in DBH and CDC147 genes (10). In a family with a high prevalence of lung adenocarcinoma, it was found that a functional missense mutation in the oncogene YAP1 was linked to the likelihood of getting the illness through WGS (11).

With a more comprehensive understanding of genes, more and more studies take gene interaction information into account. Maurano et al. (12) demonstrated that the regulation relationship between genes plays a vital role in the disease research field. Although GWAS, WES, and WGS demonstrated the effectiveness of mining disease-related genes in previous studies, this method ignores a large amount of complex information about interactions between gene pairs. Interaction networks have proven effective in the field of biological information, like identifying disease-related molecules (13) and predicting protein–metabolite interactions (14). Graph convolutional network (GCN) (15) is one type of neural network architecture to learn nodes and edges of graphs. It has been proved that GCN enhances algorithms of abilities to mine information and make decisions in the bioinformatics field. For example, Deep-DRM was proposed to identify disease-related metabolites (16). In Deep-DRM, GCN was applied as an encoder to integrate features of metabolites and disease. In DeepLGP, GCN was applied to convolve a gene interaction network for encoding the features of genes and lncRNAs (17). Cheng et al. (18) proposed a deep learning method to predict cell type-specific genes of lung cancer based on SC2disease (19) and other databases. This task only inferred cell type-specific genes of lung cancer in 8 cell types, instead of directly demonstrating whether the gene is related to lung cancer.

Interaction relationships of genes can be translated into a graph network. We treated the task of identifying NSCLC-related genes as a binary classification and proposed a novel deep learning method named DEEP-LC to solve it. GCN was applied to learn and extract relevant features from gene interaction networks, and CNN was the classification module to identify target genes.

METHOD

The method called Deep-LC that we proposed includes two parts. The structure is shown in **Figure 1**. First, we constructed a graph network by gene interaction information related to lung

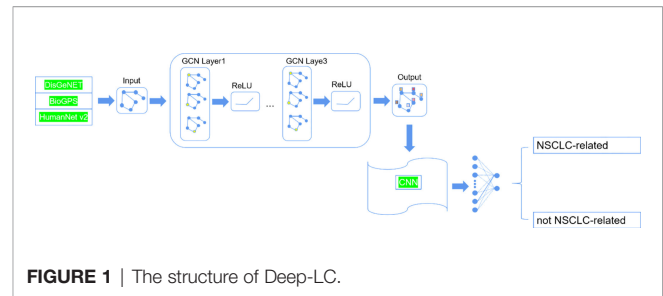


FIGURE 1 | The structure of Deep-LC.

cancer and used a GCN to extract features of interaction information between genes. Then, we constructed a small convolutional neural network (CNN) to identify potential lung cancer-related genes.

Construction of the Graph Network of Genes

The graph network of genes represents the genes interaction network. The graph network contains nodes and edges. In the study, the genes that we selected are the nodes, and interaction information between gene pairs is the edges. Interactions information was obtained from the public database. It should be noted that outliers that had no interaction with other genes were removed.

Extracting Features by Graph Convolutional Network

Since the interactions between genes were expressed by the gene network, we use a GCN to extract features from the gene network. The graph network we built can be expressed as $G = (V, E, W)$. V represents the nodes of the network, E represents the edges, and W represents the weighted matrix encoding the connection weight between vertices.

The Laplacian matrix is defined as

$$L = D - A \quad (1)$$

where D means the degree matrix of the network and A is the adjacency matrix.

Since the features of genes should contain not only connections between nodes but also the information itself, we can get

$$A' = A + I \quad (2)$$

where I is the identity matrix.

Then the inverse degree matrix D' can be obtained.

$$D' = \sum A' \quad (3)$$

Last, we can get the features, as follows:

$$X' = \sigma \left(D' \frac{1}{2} A' D' \frac{1}{2} X \right) \quad (4)$$

where X is the features map of each node and σ is the activation function. In the study, we use rectified linear unit (ReLU) function as the activation function. The expression is as follows:

$$\text{ReLU} = \max(x, 0) \quad (5)$$

Identifying Non-Small Cell Lung Cancer-Related Genes by Convolutional Neural Network

CNN excels at computer vision and is gaining traction in the field of bioinformatics. In comparison to a pure deep neural network, CNN performs better due to the following characteristics: 1) by utilizing the sparsity of connections and parameter sharing, the convolutional layer has fewer parameters. In other words, under the same amount of parameters, CNN is superior at mining and learning characteristics from nodes. 2) The convolutional layer gathers data from both global and local features. Because the features of disease-related genes focused on some specific areas, global features are redundant when it comes to identifying disease-related genes. As a result, studying local features can assist us in extracting crucial information from features. Therefore, CNN is applied as the supervised model to decide which genes are associated with NSCLC in the study.

The structure of the CNN is shown in **Table 1**. Our CNN module has four convolutional layers and a full-connected layer. We still used ReLU as the activation function the same as the GCN. Between layers, we added batch normalization (20) to avoid gradient disappearance and gradient explosion and avoid over-fitting. Both the above layers strengthen the ability of the features fusion learning and decision making.

It should be noted that the activation function we used after the full-connected layer is softmax function. Because our task is the binary classification task, we used binary cross-entropy as the loss function, as follows:

$$\text{Loss} = -y_i \log(p_i) - (1 - y_i) \log(1 - p_i) \quad (6)$$

where y_i means the true value and p_i means the predicted value.

For training details, we used dropout to avoid over-fitting, and we set the rate at 0.5. We used Adam with default parameters as the optimization algorithm. We trained our method 50 epochs. The initial learning rate is 0.01 and reduced to 1/10 after 40 epochs.

RESULT

Datasets

In the paper, we selected genes that are related to NSCLC disease from DisGeNET (21), which is a platform that integrates information on

gene–disease associations. NSCLC includes stage I, II, III, IIIA, and IIIB types; they are 115, 11, 16, 12, and 11 disease-related genes for the different types of NSCLC, respectively. After integrating the same genes with different types of NSCLC, we obtained 142 NSCLC-related genes. We obtained gene expression of different tissues from BioGPS (22). After deleting genes that lacked information on the probe set, we obtained 142 positive samples finally. Considering data balance, we randomly selected 142 genes that were reported as not being related to NSCLC as the negative samples. Then we obtained interactions between genes from the HumanNet database (23). In the gene interaction network, the nodes are genes that we selected, and the edges are interactions between gene pairs. In the paper, we used log likelihood score (LLS) as the weight of the edges because these scores can represent the interactions between genes.

Experiment Setup

Cross validation was used to demonstrate the performance of the algorithm in the study. The fold number was set to 10. Specifically, the dataset including the test set and the train set was divided into 10 subsets. One subset was randomly selected as the test set, and the remaining subsets were selected as the train set. In other words, every experiment was repeated 10 times totally in the paper.

The task of identifying lung cancer-related genes can be treated as a binary classification problem. The precision–recall curve is plotted based on different precision and recall, and the receiver operating characteristic curve (ROC curve) is based on different recall and false-positive rates (FPRs). Precision, recall, and FPR can be calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{FPR} = \frac{FP}{FP + TN} \quad (9)$$

where TP is a true positive, FP is a false positive, FN is a false negative, and TN is a true negative. We used the area under the precision–recall curve (AUPR) and the area under the ROC curve (AUC) as evaluating indicators. AUPR and AUC can help us demonstrate the effectiveness of the classification algorithm.

Performance

Stacking N-level GCN layers can convolve information from its N-order neighbors. Stacking too many GCN layers may lead to the vanishing gradient problem (24). Too little layers may cause feature learning insufficiency. So we evaluated the influence of different numbers of GCN layers on Deep-LC. The results are shown in **Table 2**. Deep-LC with three GCN layers has the best performance.

If the number of layers is more than three, both AUC and AUPR scores decrease. This result might be related to the gradient vanishing problem to some extent. We can conclude that the performance of the Deep-LC method is enhanced by stacking layers. This operation can strengthen the capability of feature fusion and be helpful for feature mining.

TABLE 1 | The structure of CNN.

Layers	Kernel size	The number of filters
Convolutional layer	3	32
Batch normalization/ReLU		
Convolutional layer	3	64
Batch normalization/ReLU		
Convolutional layer	3	32
Batch normalization/ReLU		
Convolutional layer	3	16
Batch normalization/ReLU		

CNN, convolutional neural network; ReLU, rectified linear unit.

Comparison Experiments

We compared Deep-LC with the other four methods, including GCN, CNN, random forest (RF), and K-nearest neighbor (KNN). **Table 3** shows the specific results, and **Figure 2** depicts the outcomes.

According to the results of the trial, Deep-LC outperforms all other approaches in terms of AUC and AUPR scores of 0.8017 and 0.7893. As compared to GCN, CNN, RF, and KNN, Deep-LC's AUC scores increase by 9.18%, 12.56%, 15.09%, and 30.63%, respectively, and AUPR scores rise by 12.31%, 15.13%, 15.49%, and 32.38%, respectively. KNN had the lowest results, with AUC and AUPR of 0.6137 and 0.5962, respectively. In conclusion, the results reveal that Deep-LC outperforms various state-of-the-art approaches in terms of identifying NSCLC-related genes. The performance of using GCN and CNN is better than using one alone.

TABLE 2 | The performance of Deep-LC with different of GCN layers.

Layers	AUC	AUPR
1	0.7051	0.7264
2	0.7895	0.7708
3	0.8017	0.7893
4	0.7643	0.7329

GCN, graph convolutional network; AUC, area under the receiver operating characteristic curve; AUPR, area under the precision–recall curve.

CASE STUDY

To further demonstrate the effectiveness of Deep-LC, we did case studies. We aimed to identify some genes that may be related to NSCLC disease and not a positive sample that we selected. At last, we found several genes and relevant papers to support them. **Table 4** lists the genes.

CONCLUSION

Lung cancer is the main cause of cancer mortality worldwide. NSCLC accounts for over 85% of all lung cancers. GWAS and NGS have been used to discover genes related to NSCLC. However, many studies ignored the intricate interaction information between gene pairs. In the paper, we proposed a novel deep learning method named Deep-LC for identifying

TABLE 3 | The AUC and AUPR scores of Deep-LC and other four methods.

Method	AUC	AUPR
Deep-LC	0.8017	0.7893
GCN	0.7343	0.7028
CNN	0.7122	0.6855
RF	0.6965	0.6834
KNN	0.6137	0.5962

AUC, area under the receiver operating characteristic curve; AUPR, area under the precision–recall curve; GCN, graph convolutional network; CNN, convolutional neural network; RF, random forest; KNN, K-nearest neighbor.

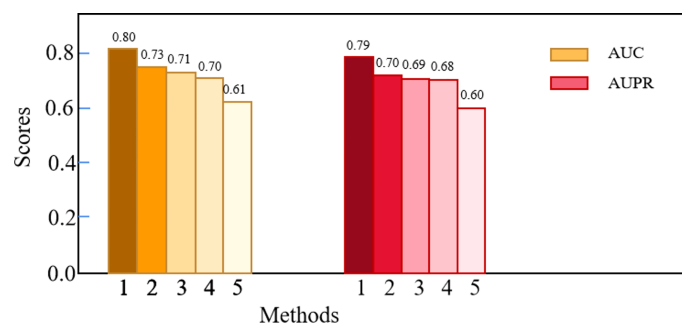


FIGURE 2 | The comparison results of Deep-LC and other four methods.

TABLE 4 | The details of genes that we mined by Deep-LC method.

Name	Entrez ID	References
KLK10	5655	Zhang et al. proved that KL10 was considerably downregulated in NSCLC compared to non-cancer samples. They concluded that KLK10 functions as a tumor suppressor gene in NSCLC, and epigenetic inactivation is a common occurrence in NSCLC pathogenesis that could be exploited as a biomarker (25).
DLEC1	9940	The study found that expression levels of DLEC1 were significantly different between tumor and normal tissues ($p = 0.0001$) (26).
EFEMP1	2202	EFEMP1 found a significantly higher frequency of methylation in NSCLC compared with the normal tissues ($p \leq 0.001$) (27).

NSCLC, non-small cell lung cancer.

NSCLC-related genes. We treated the task as a binary classification problem and integrated information to build a gene interaction network. GCNs were applied as an encoder to extract features of gene interactions network, and a simple CNN module was applied as the decoder to decide whether the gene is related to the disease. Deep-LC is an end-to-end method, and from the evaluation results, we can conclude that Deep-LC performs better than existing state-of-the-art methods.

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.

REFERENCES

- Zitnik M, Agrawal M, Leskovec J. Modeling Polypharmacy Side Effects With Graph Convolutional Networks. *Bioinformatics* (2018) 34(13):i457–66. doi: 10.1093/bioinformatics/bty294
- Navada S, Lai P, Schwartz AG, Kalemkerian GP. Temporal Trends in Small Cell Lung Cancer: Analysis of the National Surveillance, Epidemiology, and End-Results (SEER) Database. *J Clin Oncol* (2006) 24(18_suppl):7082–2. doi: 10.1200/jco.2006.24.18_suppl.7082
- Matakidou A, Eisen T, Houlston R. Systematic Review of the Relationship Between Family History and Lung Cancer Risk. *Br J Cancer* (2005) 93(7):825–33. doi: 10.1038/sj.bjc.6602769
- Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. *In Mayo Clinic Proc Elsevier* (2008) 83(5):584–94. doi: 10.1016/S0025-6196(11)60735-0
- Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A Susceptibility Locus for Lung Cancer Maps to Nicotinic Acetylcholine Receptor Subunit Genes on 15q25. *Nature* (2008) 452(7187):633–7. doi: 10.1038/nature06885
- Hu Z, Wu C, Shi Y, Guo H, Zhao X, Yin Z, et al. A Genome-Wide Association Study Identifies Two New Lung Cancer Susceptibility Loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat Genet* (2011) 43(8):792–6.
- Wang Z, Seow WJ, Shiraiishi K, Hsiung CA, Matsuo K, Liu J, et al. Meta-Analysis of Genome-Wide Association Studies Identifies Multiple Lung Cancer Susceptibility Loci in Never-Smoking Asian Women. *Hum Mol Genet* (2016) 25(3):620–9. doi: 10.1093/hmg/ddv494
- Hu L, Wu C, Zhao X, Heist R, Su L, Zhao Y, et al. Genome-Wide Association Study of Prognosis in Advanced Non-Small Cell Lung Cancer Patients Receiving Platinum-Based Chemotherapy. *Clin Cancer Res* (2012) 18(19):5507–14. doi: 10.1158/1078-0432.CCR-12-1202
- Xiong D, Wang Y, Kupert E, Simpson C, Pinney SM, Gaba CR, et al. A Recurrent Mutation in PARK2 Is Associated With Familial Lung Cancer. *Am J Hum Genet* (2015) 96(2):301–8. doi: 10.1016/j.ajhg.2014.12.016
- Liu Y, Kheradmand F, Davis CF, Scheurer ME, Wheeler D, Tsavachidis S, et al. Focused Analysis of Exome Sequencing Data for Rare Germline Mutations in Familial and Sporadic Lung Cancer. *J Thorac Oncol* (2016) 11(1):52–61. doi: 10.1016/j.jtho.2015.09.015
- Chen HY, Yu S-L, Ho B-C, Su K-Y, Hsu Y-C, Chang C-S, et al. R331W Missense Mutation of Oncogene YAP1 Is a Germline Risk Allele for Lung Adenocarcinoma With Medical Actionability. *J Clin Oncol* (2015) 33(20):2303. doi: 10.1200/JCO.2014.59.3590
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* (2012) 337(6099):1190–5. doi: 10.1126/science.1222794

AUTHOR CONTRIBUTIONS

ML and GM designed the experiments, analyzed the data, and wrote the manuscript. XL and TM analyzed the bioinformatic data. GS provided important ideas. This whole work is guided by HH. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Science and Technology funds from Liaoning Education Department (No. LZ2020038) and Wu Jieping Medical Foundation (No. 320.6750.2021-16-47).

- Peng J, Zhao T. Reduction in TOM1 Expression Exacerbates Alzheimer's Disease. *Proc Natl Acad Sci* (2020) 117(8):3915–6. doi: 10.1073/pnas.1917589117
- Zhao T, Liu J, Zeng W, Wang X, Li S, Zang T, et al. Prediction and Collection of Protein–Metabolite Interactions. *Briefings Bioinf* (2021) 22(5):bbab014. doi: 10.1093/bib/bbab014
- Kipf TN, Welling M. Semi-Supervised Classification With Graph Convolutional Networks. *arXiv* (2016). doi: 10.48550/arXiv.1609.02907
- Zhao T, Hu Y, Cheng L. Deep-DRM: A Computational Method for Identifying Disease-Related Metabolites Based on Graph Deep Learning Approaches. *Briefings Bioinformatics* (2021) 22(4):bbaa212. doi: 10.1093/bib/bbaa212
- Zhao T, Hu Y, Peng J, Cheng L. DeepLGP: A Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* (2020) 36(16):4466–72. doi: 10.1093/bioinformatics/btaa42818
- Cheng N, Chen C, Li C, Huang J. Inferring Cell-Type-Specific Genes of Lung Cancer Based on Deep Learning. *Curr Gene Ther* (2022). doi: 10.2174/1566523222666220324110914
- Zhao T, Lyu S, Lu G, Juan L, Zeng X, Wei Z, et al. SC2disease: A Manually Curated Database of Single-Cell Transcriptome for Human Diseases. *Nucleic Acids Res* (2021) 49(D1):D1413–9. doi: 10.1093/nar/gkaa838
- Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *PMLR* (2015). doi: 10.5555/3045118.304
- Piñero J, Bravo A, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants. *Nucleic Acids Res* (2016) 45(D1):833–9. doi: 10.1093/nar/gkw943
- Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, et al. BioGPS: An Extensible and Customizable Portal for Querying and Organizing Gene Annotation Resources. *Genome Biol* (2009) 10(11):1–8. doi: 10.1186/gb-2009-10-11-r130
- Hwang S, Kim CY, Yang S, Kim E, Hart T, Marcotte EM, et al. HumanNet V2: Human Gene Networks for Disease Research. *Nucleic Acids Res* (2019) 47(D1):D573–80. doi: 10.1093/nar/gky1126
- Li G, Muller M, Thabet A, Ghanem B. Deepgcns: Can Gcns Go as Deep as Cnns? *In Proc IEEE/CVF Int Conf Comput Vision* (2019), 9267–76. doi: 10.1109/ICCV.2019.00936
- Zhang Y, Song H, Miao Y, Wang R, Chen L. Frequent Transcriptional Inactivation of Kallikrein 10 Gene by CpG Island Hypermethylation in Non-Small Cell Lung Cancer. *Cancer Sci* (2010) 101(4):934–40. doi: 10.1111/j.1349-7006.2009.01486.x
- Pastuszak-Lewandoska D, Kordiak J, Antczak A, Migdalska-Sęk M, Czarnecka KH, Górski P, et al. Expression Level and Methylation Status of Three Tumor Suppressor Genes, DLEC1, ITGA9 and MLH1, in Non-Small Cell Lung Cancer. *Med Oncol* (2016) 33(7):1–8. doi: 10.1007/s12032-016-0791-3
- Zhang Y, Wang R, Song H, Huang G, Yi J, Zheng Y, et al. Methylation of Multiple Genes as a Candidate Biomarker in non-Small Cell Lung

Cancer. *Cancer Lett* (2011) 303(1):21–8. doi: 10.1016/j.canlet.2010.12.011

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.

Copyright © 2022 Li, Meng, Liu, Ma, Sun and He. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

EDITED BY

Liang Cheng,
Harbin Medical University, China

REVIEWED BY

Shuyuan Wang,
Harbin Medical University, China
Letizia Gnetti,
University Hospital of Parma, Italy

*CORRESPONDENCE

Changli Wang
wangchangli@tmuch.com
Bin Zhang
zhangbin_09@tmu.edu.cn
Dongsheng Yue
yuedongsheng@tmu.edu.cn

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

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 28 April 2022

ACCEPTED 11 July 2022

PUBLISHED 03 August 2022

CITATION

Huo Y, Sun L, Yuan J, Zhang H,
Zhang Z, Zhang L, Huang W, Sun X,
Tang Z, Feng Y, Mo H, Yang Z,
Zhang C, Yu Z, Yue D, Zhang B and
Wang C (2022) Comprehensive
analyses unveil novel genomic and
immunological characteristics of
micropapillary pattern in lung
adenocarcinoma.
Front. Oncol. 12:931209.
doi: 10.3389/fonc.2022.931209

COPYRIGHT

© 2022 Huo, Sun, Yuan, Zhang, Zhang,
Zhang, Huang, Sun, Tang, Feng, Mo,
Yang, Zhang, Yu, Yue, Zhang and Wang.
This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Comprehensive analyses unveil novel genomic and immunological characteristics of micropapillary pattern in lung adenocarcinoma

Yansong Huo^{1†}, Leina Sun^{2†}, Jie Yuan^{3†}, Hua Zhang^{1†},
Zhenfa Zhang¹, Lianmin Zhang¹, Wuhao Huang¹,
Xiaoyan Sun¹, Zhe Tang¹, Yingnan Feng¹, Huilan Mo¹,
Zuoquan Yang³, Chao Zhang³, Zicheng Yu³,
Dongsheng Yue^{1*}, Bin Zhang^{1*} and Changli Wang^{1*}

¹Department of Lung Cancer, Tianjin Lung Cancer Center, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China, ²Department of Pathology, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China, ³GenePlus-Shenzhen, Shenzhen, China

Lung adenocarcinoma (LUAD) usually contains heterogeneous histological subtypes, among which the micropapillary (MIP) subtype was associated with poor prognosis while the lepidic (LEP) subtype possessed the most favorable outcome. However, the genomic features of the MIP subtype responsible for its malignant behaviors are substantially unknown. In this study, eight FFPE samples from LUAD patients were micro-dissected to isolate MIP and LEP components, then sequenced by whole-exome sequencing. More comprehensive analyses involving our samples and public validation cohorts on the two subtypes were performed to better decipher the key biological and evolutionary mechanisms. As expected, the LEP and MIP subtypes exhibited the largest disease-free survival (DFS) differences in our patients. *EGFR* was found with the highest mutation frequency. Additionally, shared mutations were observed between paired LEP and MIP components from single patients, and recurrent mutations were verified in the Lung-Broad, Lung-OncoSG, and TCGA-LUAD cohorts. Distinct biological processes or pathways were involved in the evolution of the two components. Besides, analyses of copy number variation (CNV) and intratumor heterogeneity (ITH) further discovered the possible immunosurveillance escape, the discrepancy between mutation and CNV level, ITH, and the pervasive DNA damage response and WNT pathway gene alternations in the MIP component. Phylogenetic analysis of five pairs of LEP and MIP components further confirmed the presence of ancestral *EGFR* mutations. Through comprehensive analyses combining our samples and public cohorts, *PTP4A3*, *NAPRT*, and *RECQL4* were identified to be co-amplified. Multi-omics data also demonstrated the immunosuppression

prevalence in the MIP component. Our results uncovered the evolutionary pattern of the concomitant LEP and MIP components from the same patient that they were derived from the same initiation cells and the pathway-specific mutations acquired after *EGFR* clonal mutation could shape the subtype-specificity. We also confirmed the immunosuppression prevalence in the MIP subtype by multi-omics data analyses, which may have resulted in its unfavorable prognosis.

KEYWORDS

lung adenocarcinoma, histological subtypes, whole-exome sequencing, copy number alternation, intratumor heterogeneity

Introduction

Lung adenocarcinoma (LUAD) is the most common histological type of non-small-cell lung cancer (NSCLC) (1). Most cases of adenocarcinoma are composed of heterogeneous histological subtypes rather than a single one. In the year of 2015, the World Health Organization (WHO) proposed a novel definition of five LUAD subtypes to address the histologic heterogeneity, including the lepidic (LEP), acinar (ACI), papillary (PAP), micropapillary (MIP), and solid (SOL) pattern types. Patients presenting with MIP are prone to lymphovascular invasion and pleural invasion, as well as lymph node or intrapulmonary metastasis after surgical resection (2). Meanwhile, previous studies indicated that patients with a LEP growth pattern exhibited less aggressive behavior and had the most favorable outcomes among the predefined subtypes (3).

Aiming for the elucidation of the mechanisms beyond tumorigenesis and malignance discrepancy, several studies were conducted to evaluate the molecular and genetic features of LUAD subtypes, especially on MIP and LEP. As for the MIP subtype, a recent study observed the disruption of the catenin–cadherin complex (4), which possibly contributed to its poor intercellular adherence. At the genetic level, MIP/SOL tumors have a significantly higher tumor mutation burden (TMB) and fraction of the genome altered than other LUAD subtypes. Key oncogenes *BRAF* and *EGFR* were found with higher mutation frequency in LUAD with MIP in multiregional and multiracial cohorts (5). The gene and protein levels of *c-MET* were also found to be elevated in MIP and patients with a poor prognosis (5, 6). Although dysregulated oncogenes associated with poorer prognoses of MIP-predominant LUAD were identified, there remain key mechanisms that are uncharacterized. For example, the genetic association between subtypes and the evolutionary trajectory of the relatively malignant MIP subtype was scarcely discussed.

Noticing the recent emergence of lung cancer immunotherapy, studies assessing the efficacy of immune-related therapies on MIP-

predominant LUAD have emerged. Considering the abundance of programmed death-ligand 1 (PD-L1) and programmed cell death protein 1 (PD-1) as well as the tumor immunological microenvironment crucially influence the immunotherapy effectiveness, Francois et al. found the significant differences in PD-L1 expression levels between LUAD histological patterns (7), while Zhang et al. detected higher CD4+ and CD8+ T-cell infiltration as well as increased PD-L1 abundance in samples with a higher percentage of MIP components through the immunohistochemistry staining (8). Regarding the fact that both the studies focused on restricted components of the tumor microenvironment (TME), a more comprehensive analysis of the variation of TME in specific LUAD subtypes could fill the gap in optimal treatment determination, especially for MIP patients.

To address the abovementioned limitations, we retrospectively reviewed 286 patients with different histological subtype-predominance and compared their survival differences. Patients simultaneously possessing MIP and LEP components were further selected for whole-exome sequencing (WES) on both LEP and MIP components, and the genetic differences responsible for varied prognosis and the subtype-level genetic association was investigated. Multi-cohort analyses further discovered the genes specifically altered in MIP or LEP as well as the extent of immune infiltration. Our results expanded the evolution of cognition between the LUAD subtypes and offered therapeutic suggestions for MIP patients.

Materials and methods

Patient selection and histopathologic subtyping

We retrospectively reviewed patients diagnosed with LUAD at the Tianjin Cancer Hospital from 2011 to 2014. Among patients who underwent tumor resection, those with an MIP component

exceeding 5% of the area size were primarily selected. Patients receiving pre-surgery anticancer treatment, with stage IV disease or other malignancies were excluded. A total of 286 patients passed the selection criteria, and the resected tumors were restaged according to the eighth edition of the American Joint Committee on Cancer TNM staging system for lung cancer. For the LUAD histological subtyping, the formalin-fixed paraffin-embedded (FFPE) samples were first stained with hematoxylin and eosin (H&E) and reviewed by two pathologists. The percentage of each histological component was further calculated in 5% increments and the most dominant pattern was recorded. This study was approved by our institutional review board. Written informed consent was obtained from all patients.

Sample laser-capture micro-dissection and high-throughput sequencing data generation

Eight FFPE samples from LUAD patients who underwent surgery at the Tianjin Cancer Hospital between 2018 and 2020 were micro-dissected to isolate MIP and LEP components using a NIKON ECLIPSE TI2, Japan microscope. More specifically, 20 FFPE slides were cut into 10 μ m thick sections, baked for 1 h at 60 °C, stained with H&E, and immersed in xylene. The MIP and LEP areas were circumscribed electronically under the microscope and collected in a centrifugation tube with an adhesive-cap after ablating with a cold laser. Later, genomic DNA of the five pairs of MIP and LEP components plus one MIP component passing initial quality control was extracted by a Maxwell 16 FFPE Plus LEV DNA purification kit and fragmented by an ultra-sonicator UCD-200 (Diagenode, Seraing, Belgium) with length-based selection through Hieff NGS DNA selection beads. DNA quantity was assessed by a Qubit 2.0 Fluorometer with a Quanti-IT dsDNA HS Assay Kit (Thermo Fisher Scientific, MA, USA). The sequencing libraries were further constructed by a custom 53 M whole-exon capturing probe (IDT, IA, USA). The Geneplus-2000 sequencing platform (Geneplus, Beijing, China) further sequenced the libraries in a 100 bp paired-end manner.

Mutation calling, somatic copy number alteration detection and mutational signature analysis

Raw sequencing data were primarily filtered on the total read volume, GC content, Q30 percentage, and duplication rate. Later read alignment to the human genome (hg19) was performed by BWA (9) (version 0.7.10). After sample coverage filtration, MuTect (10) (version 1.1.4) from the GATK (version 4.0) pipeline identified single nucleotide variants (SNVs), small insertions and deletions (InDels), while segment-level somatic copy number alternations

(SCNAs) were detected by GATK. Several rounds of filtration on SNVs were conducted, including (1) retaining variants with low frequency (≤ 0.01) in a population from the 1,000 Genomes Project (<https://www.internationalgenome.org/>), the Genome Aggregation Database (gnomAD) (<https://gnomad.broadinstitute.org>), and the Exome Aggregation Consortium (ExAC); (2) keeping mutations with a number of supporting reads greater than 3; (3) keeping mutations with non-zero variant allele frequency (VAF) (≥ 0.01); and (4) only functional alternations were preserved. Cancer-associated genes and cancer driver genes were collected from the Cancer Gene Census in the COSMIC database (<https://cancer.sanger.ac.uk/census>) and two pan-cancer publications (11, 12) for further comparisons. As for SCNAs, an in-house script (13) employed statistical significance between tumor and normal tissues for focal level SCNA inference. Additionally, the mutational spectrum and absolute contribution of COSMIC v3 SBS (single base substitution) mutational signatures were derived by MutationalPatterns (14) on unfiltered somatic mutations, while Sigminer (15) quantified the absolute exposures of COSMIC v3 DBS (double base substitution) and ID (InDel) signatures.

Intratumor heterogeneity measurement and SNV/SCNA clonal architecture inference

We measured the intratumor heterogeneity (ITH) of samples on both SNV and SCNA. For filtered SNVs, the mutant-allele tumor heterogeneity (MATH) score (16) was calculated using VAF values. The ABSOLUTE (17) tool further estimated the cancer ploidy, tumor purity, rescaled copy ratio, and cancer cell fraction (CCF) by combining SNV and SCNA data. The clonal architectures of SNVs were derived from the higher clonal mutation probability and the CCF upper 95% confidence interval greater than 1. For SCNAs, copy-neutral LOH (CNLOH) segments were initially discarded, and clonal architectures were further annotated using allelic subclonal information from ABSOLUTE outputs. Additionally, by using an in-house script, we constructed the phylogenetic trees by identifying shared and private mutations or focal SCNAs in concomitant MIP and LEP components from one patient. Clonal mutations existing in both components constituted the trunk of the phylogenetic tree, while private mutations constituted the tree branches. Focal SCNA information was also considered in the tree construction procedure, i.e., marking the shared focal SCNA between MIP and LEP.

Pathway annotation and Gene Oncology analysis

For the integrative analysis of SNVs and SCNAs, DNA damage repair (DDR) related genes were collected from a

previous publication (18). The populational structures of mutations were identified on filtered SNVs and annotated SCNAs by PyClone-VI (19). These clone clusters were visualized by ClonEvol (20). The Enrichr (21) tool was used for pathway enrichment and Gene Oncology (GO) analysis. Enriched GO Biological Processes and Reactome (22) pathway entries were reported with P-values.

Public data curation for comparisons

For multi-omics data comparison, SNV, SCNA, transcriptomic, and proteomic data were retrieved from multiple LUAD datasets. More specifically, SNV and SCNA data from four datasets (Lung-Broad (23), Lung-MSKCC, Lung-OncoSG (24), and TCGA-LUAD) were downloaded from the cBioPortal for Cancer Genomics database (25) or the UCSC Xena database (26) and only non-synonymous mutations were retained. Survival information is also downloaded if available. Additionally, transcriptomic data from the Lung-OncoSG, TCGA-LUAD, and one GEO dataset GSE148801 (27) containing good-prognosis (e.g., LEP, ACI, and PAP histological subtypes) and poor-prognosis (e.g., MIP and SOL) samples were collected while proteomics data from the TCGA-LUAD were similarly curated. Only data from LEP and MIP subtypes were used for further comparisons.

Immune infiltration analysis by measuring the activity of cancer immunity cycle

Immune infiltration analysis was conducted on curated transcriptomic samples for comparison between LEP and MIP components. Regarding the recognition, response, and killing of cancer cells by the immune system in a step-wise manner, i.e., the cancer immunity cycle (28), we applied the single sample Gene Set Enrichment Analysis (ssGSEA) method from the GSVA R package (<http://bioconductor.org/packages/release/bioc/html/GSVA.html>) to assess the activities of these steps in the cycle. More specifically, gene signature sets for steps of the cancer immunity cycle were first downloaded from the TIP database (<http://biocc.hrbmu.edu.cn/TIP/>). Later, the gsva function in the GSVA R package was applied in the step activity quantification procedure.

Statistical methods

A two-sided Mann–Whitney test was used for evaluating group-level differences between LEP and MIP components. As for multiple comparisons, P-values were adjusted by the Benjamini–Hochberg method. When the comparisons were

conducted on categorical data, Fisher's exact test was used. As for the protein expression data, a one-sided Student's t-test was used for comparison. For all tests, a P-value (adjusted P-value) <0.05 was considered statistically significant. The Kaplan–Meier (K-M) survival curves were generated by the survminer package (<https://rpkgs.datanovia.com/survminer/>) and the P-values were calculated using the log-rank test.

Results

Clinicopathologic characteristics of selected patients

Among the 286 patients, 51 (17.8%) were LEP-predominant, 178 (62.2%) were ACI-predominant, 16 (5.6%) were PAP-predominant, 29 (10.1%) were MIP-predominant, and 12 (4.2%) were SOL-predominant adenocarcinoma. The clinicopathologic characteristics of all patients are summarized in [Supplementary Table 1](#). As shown in [Supplementary Table 2](#) and [Supplementary Figure 1](#), patients with the MIP-predominant subtype exhibited a significantly worse DFS than those with other adenocarcinoma subtypes ($P < 0.05$, [Supplementary Figure 1A](#)), and LEP and MIP subtypes exhibited the largest DFS difference, while SOL-predominant patients showed significantly worse OS ($P < 0.05$, [Supplementary Figure 1B](#)).

Mutational landscape exhibits the involvement of distinct biological processes in LEP and MIP lesions

Among eight micro-dissected samples, six cases passed quality control, but the quantity of LEP component in one case was inadequate. Six MIP and five LEP components were finally sequenced ([Supplementary Table 3](#) and [Figure 1A](#)). A total of 2,035 and 2,757 SNVs and InDels were identified, while 684/791 and 257/284 mutations were retained after quality and cancer-related gene filtration ([Supplementary Figure 2A](#) and [Supplementary Table 4](#)). Genes with the highest mutation frequency after quality filtering are shown in [Figure 1B](#). *EGFR* was identified to be the most frequently mutated drive gene, in line with the finding that LEP and MIP components possess significantly higher *EGFR* mutation frequencies (29). Besides, several cancer-associated genes including *TP53*, *TRIO*, *CEBPA*, *PCLO*, and *PDE4DIP* were concomitantly mutated ([Figure 1B](#)), denoting *p53*, WNT-beta-catenin signaling, *PI3K/AKT/mTOR* signaling, and DNA repair pathways were affected. Interestingly, shared mutations were observed between paired LEP and MIP components from single patients ([Figure 1B](#)), raising the presumption that the paired LEP and MIP components could be homogeneous. We also compared the mutation frequency of

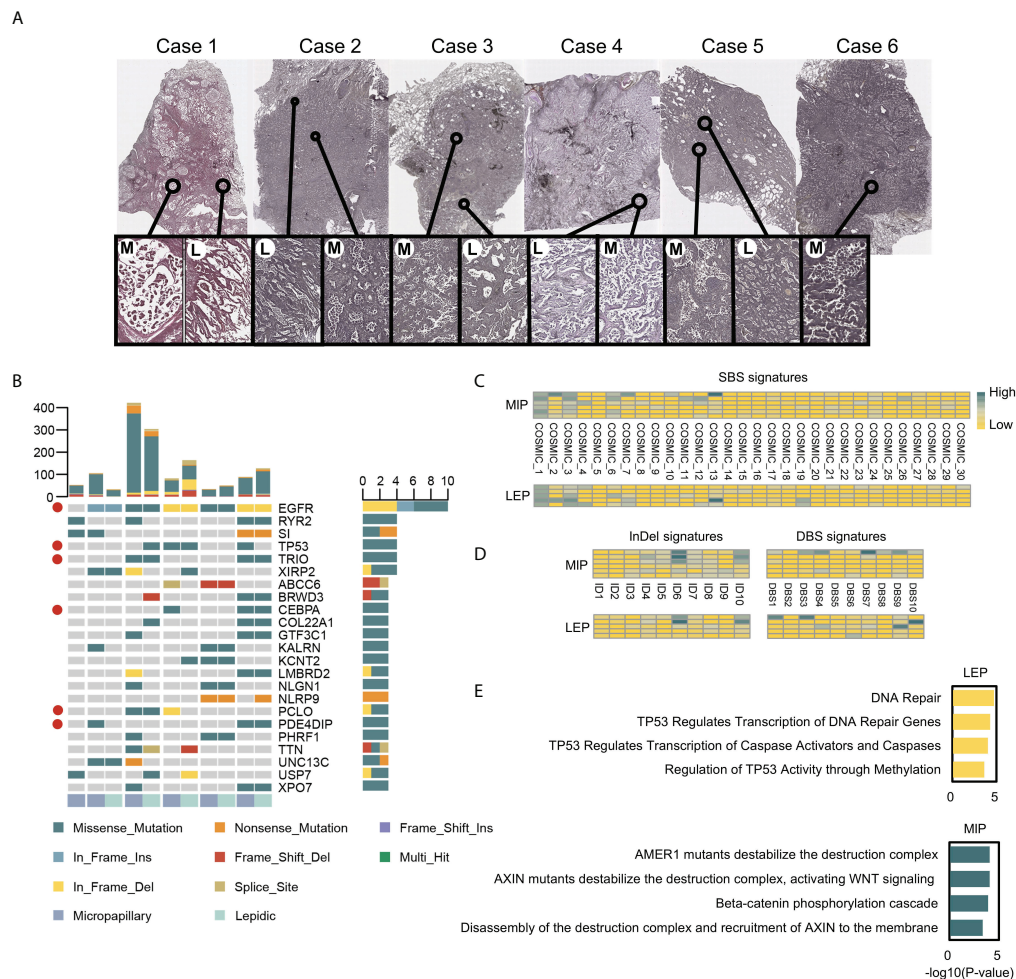


FIGURE 1

The H&E stained slides and the mutational landscape of MIP and LEP subtypes. (A) Scanner view (magnification $\times 20$) of the H&E-stained MIP (M) and LEP (L) subtypes from six patients. (B) Genes with top mutation frequency in samples. Cancer-associated genes were marked with red circles. (C) Quantifications of the SBS signatures in two histological subtypes. (D) Quantifications on the InDel and DBS signatures in subtypes. (E) Pathways enriched in mutated tumor suppressor genes (TSGs) for LEP and MIP subtypes.

these genes with public cohorts, including Lung-Broad, Lung-OncoSG, and TCGA-LUAD. Several cancer-associated genes, including *EGFR*, *TP53*, *TRIO*, *PCLO*, and *PDE4DIP*, were found recurrently mutated (Supplementary Figure 2B).

Mutation signature analysis was conducted on unfiltered mutations separately for LEP and MIP components. The point substitution spectrum plot displayed an insignificant difference between the two histological subtypes (Supplementary Figure 2C). Similarly, the SBS, InDel, and DBS signatures mapped to the COSMIC database (accessed in March 2021) were similar between the two subtypes (Figures 1C, D), indicating the histological differences between LEP and MIP components could be caused by alternations in specific key genes.

Of particular interest, mutated tumor suppressor genes (TSGs) were enriched in distinct pathways in the LEP and MIP subtypes (Figure 1E). TSGs from LEP components were enriched in DNA repair and *TP53*-related pathways, while mutated TSGs in MIP components were found to be enriched in pathways associated with beta-catenin destruction complex, *AXIN* mutation and WNT signaling, which followed report (4). When concerning the enriched pathways for mutated oncogenes, seven of the 10 top-enriched Reactome pathways from the two groups were identical, which were mainly associated with *EGFR* and *PI3K* signaling (Supplementary Figure 2D). By inspecting the mutated TSG pathway enrichment pattern in Lung-Broad (Supplementary Figures 3A, B), Lung-OncoSG (Supplementary Figures 3C, D),

and TCGA-LUAD (Supplementary Figures 3E, F) cohorts, similar entries were identified in both LEP and MIP samples, while *NOTCH1*-related pathways were additionally found in TCGA-LUAD MIP samples, which was unsurprising since the cross-talk between NOTCH and WNT pathways was previously unveiled (30). As for the oncogenes mutated in three public cohorts, shared terms were found between LEP and MIP components in Lung-Broad (Supplementary Figures 4A, B), Lung-OncoSG (Supplementary Figures 4C, D), and TCGA-LUAD (Supplementary Figures 4E, F) cohorts but with lower overlapping proportion, which endorsed the possible existence of a common mutational ancestor in the paired components.

Copy number alternation and clonality analysis uncovered distinct ITH characteristics in LEP and MIP subtypes

Through the segment-level copy number alternation identification procedures, multiple amplified and deleted segments were detected (Supplementary Figure 5A). Chromosomes including 3,4,5,10,15,17, and 18 exhibited different copy number alternation patterns between the two subtypes, and the MIP subtype showed both higher chromosome level (Supplementary Figure 5B) and arm-level copy number variation (CNV) burden (Supplementary Figure 5C), which followed the report (31). To further pinpoint the recurrent SCNAs at the focal level, we identified 1,116 genes with somatic copy number alternations through statistical testing on read coverages from all samples (details in *Materials and methods*), among which 159/80 genes were uniquely amplified/deleted in the LEP component, while 34/11 genes were uniquely amplified/deleted in the MIP component. By annotating the enriched pathways on these genes, 27 pathways were found to overlap between the enrichment results of uniquely amplified genes in LEP and deleted genes in MIP, which could be categorized into the immune system, innate immune system, interleukin signaling, *SHC1* events, *ERK* activation, and *FRS*-mediated signaling pathways (Figure 2A). When inspecting the number of genes, pathways related to the immune system and innate immune system got the highest gene number variated (37 genes amplified in LEP and four genes deleted in MIP subtype), indicating that MIP LUADs tend to have induced immunosurveillance escape. Additionally, two pathways were identical between the enrichment results of uniquely deleted genes in LEP and amplified genes in MIP (Figure 2B), which were associated with homology directed repair (HDR) and mRNA fate regulation, but the variated gene number was limited (six genes for LEP and four genes for the MIP group).

Intratumor heterogeneity (ITH) can depict the genetic and epigenetic tumor inner diversity and is proven to be closely related to cancer progression, therapeutic resistance, and

recurrences. To compare the ITH of the two histological subtypes at both the mutational and copy number level, we annotated the mutations/SCNAs with clonality. As shown in Figure 2C, there was no significant difference in the clonal tumor mutation burden (cTMB) and subclonal mutation proportion between the MIP and LEP groups (Figure 2D). The MATH score, which is widely used to measure the mutational ITH, exhibited a similar trend (Supplementary Figure 5D). As for the copy number variations, the MIP group possessed a significantly higher proportion of subclonal SCNAs (Figure 2E) as well as a trend of higher subclonal genome fraction (Figure 2F). Interestingly, the frequency of clonal mutations in DNA Damage Response (DDR) and WNT pathway genes was higher in the MIP subtype (Supplementary Figure 5E), which may possibly partially increase the subclonal genome alternations in immune-related genes since the association between canonical WNT-beta-catenin signaling and carcinogenesis as well as immune suppression was clear (32). Indeed, six MIP components showed a trend of a higher percentage of subclonal SCNA (Figure 2G) as well as a higher number of focal deletions (Figure 2H) on the genes related to the two immune pathways. We confirmed the results that genes significantly deleted in the MIP subgroup exhibited association with immune-related terms in the Lung-Broad and Lung-OncoSG cohorts (Supplementary Figure 6).

Evolutionary pattern exploration on the paired LEP and MIP components

To elaborate on the possible evolutionary process between LEP and MIP subtypes, we delineated the phylogenetic trees for each patient based on mutations as well as focal level SCNAs. As shown in Figure 3, all five patients possessed truncal mutations between paired LEP and MIP components, while no obvious bias on private mutation burden after truncal divergence was observed. Clonal mutations on cancer drivers including *EGFR*, *TP53*, and *CEBPA* were identified, and *EGFR* was the only gene coincident in five pairs, which confirmed the presence of ancestral mutations. The driver mutations private to LEP were enriched in chromatin organization, *TP53*-related and DNA double strand repair pathways (Supplementary Figure 7A), while mutations private to MIP were enriched in cellular signaling and beta-catenin-related pathways (Supplementary Figure 7B). We further annotated the shared mutations in Figure 3 with clonality to explore the clonal-subclonal transitions between the LEP and MIP subtypes. For the genes possessing mutations with increased clonality in MIP, GO terms related to neurogenesis were found enriched (Supplementary Figure 7C), denoting that tumor-induced neurogenesis and nerve-cancer crosstalk may account for the aggressiveness of the MIP subtype. Oppositely, genes with

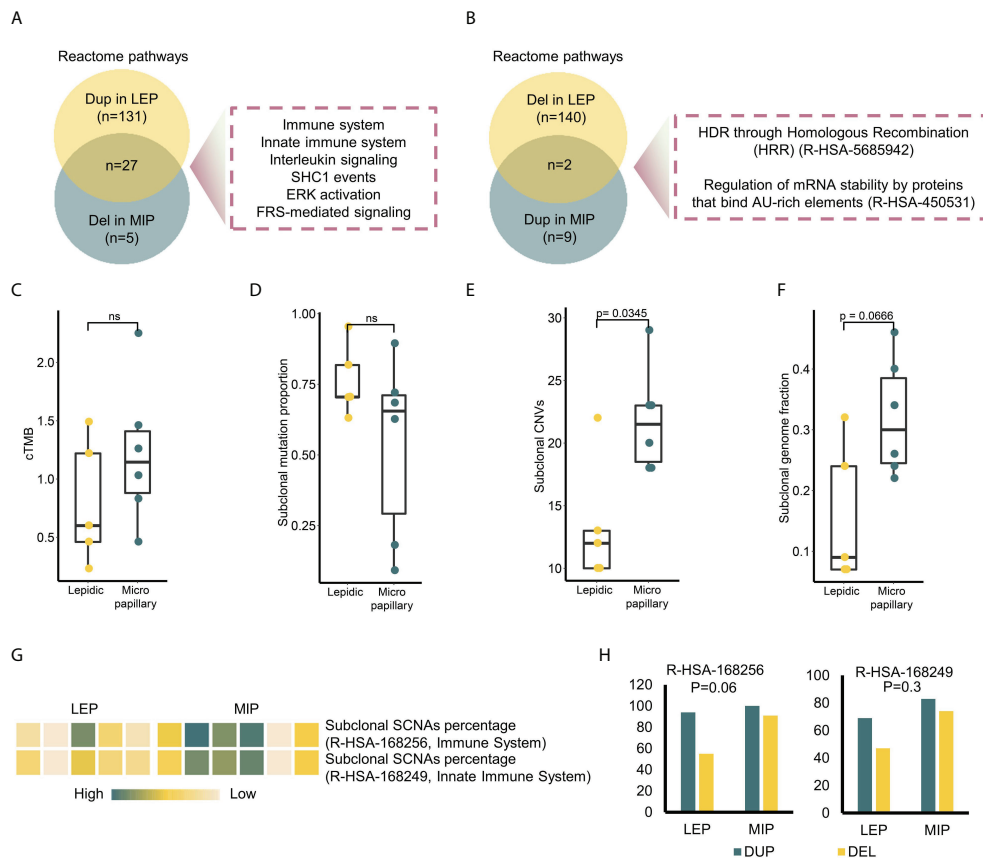


FIGURE 2

Multi-perspective investigation on intratumor heterogeneity (ITH) difference between two subtypes. **(A)** Intersection of the enriched pathways in genes uniquely amplified in LEP and deleted in MIP. **(B)** Intersection of the enriched pathways in genes uniquely deleted in LEP and amplified in MIP. **(C–F)** Clonal tumor mutation burden (cTMB), subclonal mutation proportion, subclonal CNV and subclonal genome fraction distribution in two subtypes. **(G)** Subclonal SCNA percentage of two Reactome immune pathways in sequenced samples. **(H)** Focal alteration number on the genes in two immune pathways. P-values on the alternation discrepancy between subtypes were calculated by Fisher's exact test.

mutations switched from subclonal to clonal in LEP were associated with cell-cycle related GO biological processes (Supplementary Figure 7D). As for the truncal focal SCNAs, several driver genes including *CSMD3*, *SPTAN1*, *BCORL1*, *CAMTA1*, *GRIN2A*, *MED12*, and *TRAF7* were concurrently amplified in the two subtypes (Figure 3), which were associated with developmental biology (R-HSA-1266738) and *EGFR*-related Reactome pathways (R-HSA-179812 and R-HSA-180336). Moreover, the deletion of *TP53*, *MUC4*, *ARID5B*, *ANK1*, *PTEN*, *SFPQ*, *FANCA*, *MAF*, and *ZFX3* were observed in the two subtypes. Interestingly, no driver gene showed concordant copy number variation in five pairs of samples, possibly due to the elevated SCNA-level ITH in the MIP group. We also scrutinized the genes with shared copy number variation in the paired samples. As shown in Supplementary Figure 8A, most shared deletions were on immune-related genes, while signal transduction and *PI3K/AKT* pathways,

which abnormality is highly associated with tumor progression and therapeutic resistance, were found uniformly amplified (Supplementary Figure 8B). To further derive the mutational transitions and evolutionary trajectory, we used PyClone-VI to infer the mutational populations and their evolution among paired components. As shown in Supplementary Figure 9A, numerous clone clusters were identified in 5 patients, which exhibited dynamic variant allele frequency (VAF) alternation. Clusters with drastically increased VAF in the LEP subtype were mainly enriched in mRNA splicing pathways (Supplementary Figure 9B), while clusters with increased VAF in the MIP subtype were associated with *ERBB2* functions (Supplementary Figure 9C). These data imply that LEP and MIP components from one patient were derived from the same initiation cells and the pathway-specific mutations acquired after the *EGFR* clonal mutation eventually shaped the subtype-specificity.

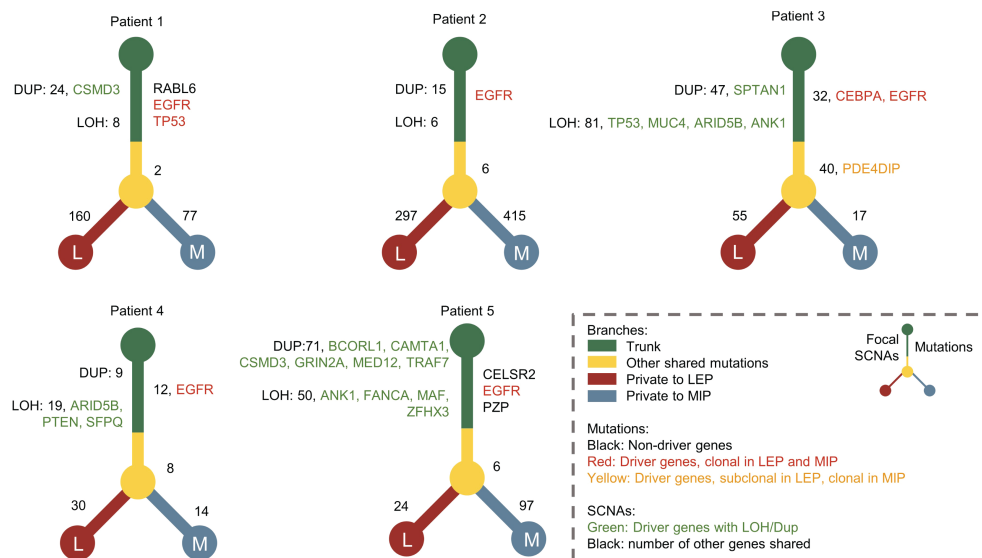


FIGURE 3

The phylogenetic trees constructed for patients with concomitant MIP and LEP components. Driver genes with mutations and focal CNV were marked with different colors. The numbers of shared mutations and focal CNVs were labeled beside the tree trunk, while mutation numbers private to MIP and LEP components were labeled beside the tree branches.

Group-wise comparison discovered co-amplified pattern of PTP4A3, NAPRT, and RECQL4

We next gathered SNV and SCNA data and identified the genes with an alternation frequency difference between the LEP and MIP groups. As shown in Table 1, mutation frequency difference was observed on nine genes, with three genes

specifically mutated in LEP group. Additionally, five genes were found with a distinct copy number alternation pattern, with one gene specifically amplified in the LEP group (Table 1). Similarly, mutation frequencies on the above nine genes and the key gene identified in phylogenetic analyses (*EGFR*) were inspected in four public cohorts (Table 2), and *EGFR* was the only gene with a significant mutational difference in non-east Asian public cohorts. Moreover, the five SCNA genes showed an

TABLE 1 Genes exhibited subtype-specific alternation frequency among the sequenced samples.

Genes	Number of samples altered in each subtype		Number of mutations in each subtype		Alternation Type
	LEP (a total of 5)	MIP (total 6)	LEP (a total of 5)	MIP (a total of 6)	
C10orf71	1	0	3	0	Mutation
SLC32A1	1	0	4	0	Mutation
DISC1	1	0	5	0	Mutation
AHCTF1	0	1	0	3	Mutation
PHRF1	1	2	1	3	Mutation
PLEC	0	1	0	3	Mutation
RYR2	1	3	1	3	Mutation
SI	1	3	0	3	Mutation
SYNE2	1	2	1	4	Mutation
RCSD1	3	0			Duplication
PTP4A3	1	5			Duplication
EZR	2	5			Deletion
NAPRT	2	5			Duplication
RECQL4	2	5			Duplication

alteration frequency difference in three non-east Asian cohorts (Table 2). Interestingly, the altered sample proportion or the alteration frequency for *PTP4A3*, *NAPRT*, and *RECQL4* was highly similar between our MIP group (Table 1) and public cohorts (Lung-Broad, Lung-OncoSG, and TCGA-LUAD, Table 2), implying the feasibility of their cooperative function through duplication in MIP components. Spearman's correlation coefficient (SCC) on the SCNA pattern of our 11 samples confirmed the association of the co-amplified genes (Supplementary Figure 10A). Such strong association was also observed on SCNA data for LEP and MIP adenocarcinoma from Lung-Broad (Supplementary Figure 10B), Lung-OncoSG (Supplementary Figure 10C, left), and TCGA-LUAD (Supplementary Figure 10D, left) cohorts. Concerning the fact that SCNA is highly related to the consequent gene expression alteration, we calculated the expressional SCC of the five genes in cohorts with available transcriptomic data. When compared to all samples (Supplementary Figures 10C, D, middle), the expressional associations between *PTP4A3*, *NAPRT*, and *RECQL4* transformed to a higher synergetic state for LEP and MIP samples in both the Lung-OncoSG (Supplementary Figure 10C, right) and TCGA-LUAD (Supplementary Figure 10D, right) cohorts. More explicitly, the correlation between SCNA and RNA expression was higher for the three genes in two public cohorts (Supplementary Figure 10E), and *NAPRT* as well as *PTP4A3* exhibited significantly higher LEP/MIP group-specificity. As an exemplification, the correlation

between the SCNA and RNA expression of the *PTP4A3* gene increased from 0.372 to 0.693 when narrowed to only LEP/MIP samples in the TCGA-LUAD cohort (Supplementary Figures 10F, G). All three co-amplified genes were significantly overexpressed in tumor samples (Supplementary Figures 10H–J, left) and *RECQL4* possessed significantly higher expression levels in the MIP subtype (Supplementary Figure 10J, right) in the TCGA-LUAD cohort. To conclude, our comprehensive analyses identified *PTP4A3*, *NAPRT*, and *RECQL4* were co-amplified and co-expressed specifically in LEP/MIP adenocarcinoma, and *RECQL4* was upregulated in the MIP group.

Immune-related analyses uncovered elevated immunosuppression in MIP subtype

The disparity of cancer immunity cycle activity was examined in the Lung-OncoSG, GEO, and TCGA-LUAD cohorts as described in *Materials and methods*. Activities of recurrent cancer immunity steps including the release of cancer cell antigen, CD8+ T-cell recruiting, dendritic cell recruiting, macrophage recruiting, T-helper 17 (Th17) cell recruiting, T-cell infiltration into tumors and killing of cancer cells were significantly higher in MIP subtype (Supplementary Figures 11A–C). By further examining the differentially expressed proteins between LEP and MIP subtypes

TABLE 2 Inspection on the alteration frequency of genes exhibited subtype-specific alterations using four public cohorts.

Genes	Number of samples altered in each subtype												Alternation Type
	Lung-Broad			Lung-MSKCC			Lung-OncoSG			TCGA-LUAD			
	LEP (a total of 13)	MIP (a total of 17)	Adj. P-value	LEP (a total of 88)	MIP or Solid (a total of 105)	P-value	LEP (a total of 10)	MIP (a total of 4)	Adj. P-value	LEP (a total of 12)	MIP (a total of 23)	Adj. P-value	
EGFR	4	2	0.00052	38	19	7.3E-05	6	3	0.525	2	4	0.00246	Mutation
SLC32A1	0	1	0.626							1	1	1	Mutation
DISC1										0	1	1	Mutation
AHCTF1	0	1	0.626							2	1	1	Mutation
PHRF1	0	1	0.13904							0	1	0.34034	Mutation
PLEC	0	3	1							1	2	1	Mutation
RYR2	1	7	0.80126				2	1	0.656	2	7	1	Mutation
SI	1	5	0.626				2	0	0.525	1	4	1	Mutation
SYNE2	0	1	0.13904							1	4	1	Mutation
RCSD1	4	4	1				6	3	0.5535	10	20	0.00278	Duplication
PTP4A3	3	1	0.0502				5	2	1	6	10	1	Duplication
EZR	0	1	0.00042				2	2	0.58	8	13	1	Deletion
NAPRT	3	1	0.01233				5	2	1	6	10	1	Duplication
RECQL4				2	1	1.4E-08	5	2	1	6	10	1	Duplication

from the TCGA-LUAD dataset, the identified proteins with MIP-specific elevation (Supplementary Figure 12A) were significantly enriched in *PD-L1* and *PD-1* checkpoint pathways in cancer (Supplementary Figure 12B). These data suggest that the MIP subtype could exist in an immune-suppressive microenvironment.

Discussion

Consistent with other studies, we confirmed the survival disparity between LUAD subtypes. By performing WES on micro-dissected LUAD tissue samples of MIP and LEP components, we explored the genetic features related to the LEP/MIP growth pattern and the evolutionary connection between LUAD subtypes. Our results revealed that LEP and MIP subtypes could be derived from the same initiation cells with *EGFR* mutation and the ultimate histological dissimilitude was shaped by the pathway-specific mutations acquired along evolution. Our results showed that the *EGFR* trunk mutation arose between pre-invasive and invasive LUAD and LEP/MIP components were evolved by a branched evolution model.

Through comprehensive comparisons of genetic alternations, the biological characteristics of the two LUAD subtypes were elucidated. As for mutational comparisons, TSG mutations in LEP were associated with DNA repair and *TP53* regulation, while genes related to WNT signaling and beta-catenin destruction complexes got both higher mutational frequency and clonality. Driver mutations private to MIP were also enriched in cellular signaling and beta-catenin-related pathways, while genes that possessed lower mutational heterogeneity in MIP were associated with neurogenesis and *ERBB2* functions. Aberrant WNT signaling pathway activation caused by gene mutations of intracellular components is associated with a higher rate of recurrence in early-stage NSCLC. On the other hand, being the critical downstream effector in the canonical WNT pathway, excessive intracellular beta-catenin promotes lung cancer aggression. Liang et al. further confirmed that the intracellular beta-catenin expression in MIP-predominant LUAD was higher than LEP-predominant LUAD (4). Besides, neurogenesis induced by tumors shapes an immunosuppressive microenvironment (33). Although cancer-related neurogenesis is considered to be associated with solid tumor metastasis, its role in LUAD remains poorly understood. Our results suggest an inner association between the MIP aggressive phenotype and neurogenesis. The activation of well-known proto-oncogene *ERBB2* signaling was associated with poor outcomes in NSCLC (34), coinciding with MIP characteristics. The copy number of genes related to the immune system, innate immune system, interleukin signaling, SHC1 events, ERK activation, and FRS-mediated signaling were also found to decrease in MIP. With the highest proportion of immune genes affected, the immunosuppression status in the

MIP subtype was confirmed. Apart from specified genomic alternations, ITH provides crucial information for drug responsiveness and clinical prognosis. Discordance between SNV and SCNA ITH was particularly observed in the MIP subtype. Subclonal genetic instability possibly facilitated MIP neoplastic cell proliferation (35) and the clonal mutations on key MIP-specific pathways contributed to its aggressive behavior.

We discovered three genes with co-amplification tendency, both in our discovery cohort and in three public validation cohorts. Previous studies proved that the knockdown of *PTP4A3* inhibited cell migration and invasion of lung cancer cell lines (36). It also induced microvascular and lymphatic vessel formation by increasing VEGF and VEGF-C expression in lung cancer tissues, which was in accordance with the clinical observations that the MIP component in LUAD increases the risk of distant and lymph node metastasis. A previous study also found that the loss of *NAPRT* promoted the epithelial-mesenchymal transition (EMT) by stabilizing beta-catenin (37). The elevated expression of *NAPRT* was conceivably associated with the disruption of the catenin-cadherin complex in MIP. Moreover, *RECQL4* could coordinate and regulate cell proliferation and cell cycle progression by protecting chromosome stability (38), and its protein expression was remarkably higher in LUAD (39). The biological mechanisms of these three genes further verified our discoveries.

Interestingly, we observed numerous genetic alternations associated with the immune status in the LEP and MIP subtypes. For example, the enrichment of immune-related pathways, including immune system and innate immune system, was observed on genes uniquely amplified in LEP and deleted in the MIP category. Along with the observation of a higher percentage of focal deletions on the immune pathway genes in our cohort and the deficiency in immune-related pathways for MIP group-specific deleted genes in Lung-Broad and Lung-OncoSG cohorts, we suspected that the MIP subtype could possess an immuno-suppression microenvironment, in other words, an induced immuno-surveillance escape. Regarding the increasing enthusiasm for lung cancer immunotherapy and our hypothesis, we next assessed and compared the TME landscape between the MIP and LEP components using the stepwise activities of the cancer immunity cycle. Steps including CD8+ T-cell recruiting, T-helper 17 (Th17) cell recruiting, T-cell infiltration into tumors and killing of cancer cells were significantly higher in MIP samples. However, elevated T-cell infiltration does not always indicate better clinical outcomes for patients. For instance, elevated expression of PD-L1 and PD-1 could inhibit the activation of T cells, conferring an immuno-evasion and immuno-suppression tumor status. Unsurprisingly, proteomic analysis further confirmed the activation of the PD1/PD-L1 pathway in the MIP subtype, partially elucidating the deteriorative survival of MIP patients. Generally, our work revealed the comprehensive TME situation of LEP/MIP

components and immuno-suppression features in MIP-predominant LUAD.

Our study has several limitations. Firstly, the cohort only included five pairs of LEP/MIP components detached from five LUAD patients and 11 samples. Further studies with a larger amount of patient involvement can better decipher the evolutionary trajectory between LUAD histological subtypes and identify subtype-specific genetic changes. Moreover, the three genes with co-amplification or co-expression tendency should be further experimentally validated, particularly on their protein expression status. Additionally, we portrayed the TME heterogeneity using bulk RNA-seq data. With the recent maturation of multiple advanced techniques, using methods including single-cell RNA-seq, spatial transcriptomics, and multiplexed immunohistochemistry could better dissect the TME in LUAD. Lastly, our analyses only focused on MIP and LEP subtypes. A more integrated study incorporating other LUAD histologic subtypes could better decode the disease.

To conclude, we identified subtype-specific genetic differences responsible for varied prognosis and the evolution trajectory of the MIP subtype. The subtype-specificity was possibly shaped by pathway-specific mutations acquired after the *EGFR* clonal mutation. The tumor microenvironment revealed the immunosuppression prevalence in MIP, which could contribute to its unfavorable prognosis. Immune checkpoint inhibitor treatments like anti-PD-1/anti-PD-L1 could maximize the therapeutic benefit for MIP-predominant LUAD patients.

Data availability statement

The raw sequencing datasets presented in this study can be found in the China National GeneBank DataBase (<https://db.cngb.org/>) with project number CNP0003191.

Ethics statement

The studies involving human participants were reviewed and approved by the Tianjin Medical University Cancer Institute and Hospital review board. The patients/participants provided their written informed consent to participate in this study.

References

1. Bray F, Ferlay J, Laversanne M, Brewster DH, Gombe Mbalawa C, Kohler B, et al. Cancer incidence in five continents: Inclusion criteria, highlights from volume X and the global status of cancer registration. *Int J Cancer* (2015) 137(9):2060–71. doi: 10.1002/ijc.29670
2. Hung JJ, Yeh YC, Jeng WJ, Wu YC, Chou TY, Hsu WH. Factors predicting occult lymph node metastasis in completely resected lung adenocarcinoma of 3

Author contributions

CW, BZ, and DY contributed to the study design. LS, ZZ, LZ, YH, WH, XS, ZT, YF, and HM contributed to data collection. JY, HZ, ZuY, CZ, and ZiY performed statistical analysis and data interpretation. YH and JY drafted the manuscript. CW and BZ revised the manuscript. CW and DY provided financial support and study supervision. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work was financially supported by the National Natural Science Foundation of China (grant number 81772484 to CW, grant number 82173038 to DY).

Conflict of interest

Authors JY, ZuY, CZ, and ZiY were employed by GenePlus-Shenzhen.

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.2022.931209/full#supplementary-material>

Cm or smaller. *Eur J Cardiothorac Surg* (2016) 50(2):329–36. doi: 10.1093/ejcts/ezv485

3. Makinen JM, Laitakari K, Johnson S, Makitaro R, Bloigu R, Lappi-Blanco E, et al. Nonpredominant lepidic pattern correlates with better outcome in invasive lung adenocarcinoma. *Lung Cancer* (2015) 90(3):568–74. doi: 10.1016/j.lungcan.2015.10.014

4. Zhu L, Yang S, Zheng L, Zhang G, Cheng G. Wnt/Beta-catenin pathway activation *Via* Wnt1 overexpression and Axin1 downregulation correlates with cadherin-catenin complex disruption and increased lymph node involvement in micropapillary-predominant lung adenocarcinoma. *J Thorac Dis* (2020) 12 (10):5906–15. doi: 10.21037/jtd-20-1495
5. Warth A, Penzel R, Lindenmaier H, Brandt R, Stenzinger A, Herpel E, et al. Egrf, kras, braf and alk gene alterations in lung adenocarcinomas: Patient outcome, interplay with morphology and immunophenotype. *Eur Respir J* (2014) 43(3):872–83. doi: 10.1183/09031936.00018013
6. Tsao MS, Marguet S, Le Teuff G, Lantuejoul S, Shepherd FA, Seymour L, et al. Subtype classification of lung adenocarcinoma predicts benefit from adjuvant chemotherapy in patients undergoing complete resection. *J Clin Oncol* (2015) 33 (30):3439–46. doi: 10.1200/JCO.2014.58.8335
7. Ng Kee Kwong F, Laggner U, McKinney O, Croud J, Rice A, Nicholson AG. Expression of pd-L1 correlates with pleomorphic morphology and histological patterns of non-Small-Cell lung carcinomas. *Histopathology* (2018) 72(6):1024–32. doi: 10.1111/his.13466
8. Zhang S, Xu Y, Zhao P, Bao H, Wang X, Liu R, et al. Integrated analysis of genomic and immunological features in lung adenocarcinoma with micropapillary component. *Front Oncol* (2021) 11:652193. doi: 10.3389/fonc.2021.652193
9. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013) 1303.3997. doi: 10.48550/arXiv.1303.3997
10. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* (2013) 31(3):213–9. doi: 10.1038/nbt.2514
11. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* (2018) 173(2):371–85. e18. doi: 10.1016/j.cell.2018.02.060
12. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science* (2013) 339(6127):1546–58. doi: 10.1126/science.1235122
13. McGranahan N, Rosenthal R, Hiley CT, Rowan AJ, Watkins TBK, Wilson GA, et al. Allele-specific hla loss and immune escape in lung cancer evolution. *Cell* (2017) 171(6):1259–71.e11. doi: 10.1016/j.cell.2017.10.001
14. Blokzijl F, Janssen R, van Bortel R, Cuppen E. Mutational patterns: Comprehensive genome-wide analysis of mutational processes. *Genome Med* (2018) 10(1):33. doi: 10.1186/s13073-018-0539-0
15. Wang S, Tao Z, Wu T, Liu XS. Sigflow: An automated and comprehensive pipeline for cancer genome mutational signature analysis. *Bioinformatics* (2021) 37 (11):1590–2. doi: 10.1093/bioinformatics/btaa895
16. Mroz EA, Rocco JW. Math, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncol* (2013) 49(3):211–5. doi: 10.1016/j.oraloncology.2012.09.007
17. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* (2012) 30(5):413–21. doi: 10.1038/nbt.2203
18. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, et al. Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome atlas. *Cell Rep* (2018) 23(1):239–54 e6. doi: 10.1016/j.celrep.2018.03.076
19. Gillis S, Roth A. Pyclone-vi: Scalable inference of clonal population structures using whole genome data. *BMC Bioinf* (2020) 21(1):571. doi: 10.1186/s12859-020-03919-2
20. Dang HX, White BS, Foltz SM, Miller CA, Luo J, Fields RC, et al. Clonevol: Clonal ordering and visualization in cancer sequencing. *Ann Oncol* (2017) 28 (12):3076–82. doi: 10.1093/annonc/mdx517
21. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* (2016) 44(W1):W90–7. doi: 10.1093/nar/gkw377
22. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. *Nucleic Acids Res* (2018) 46(D1):D649–55. doi: 10.1093/nar/gkx1132
23. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* (2012) 150(6):1107–20. doi: 10.1016/j.cell.2012.08.029
24. Chen J, Yang H, Teo ASM, Amer LB, Sherbaf FG, Tan CQ, et al. Genomic landscape of lung adenocarcinoma in East asians. *Nat Genet* (2020) 52(2):177–86. doi: 10.1038/s41588-019-0569-6
25. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioportal. *Sci Signaling* (2013) 6(269):pl1–pl. doi: 10.1126/scisignal.2004088
26. Goldman M, Craft B, Hastie M, Repecka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnol* (2020) 38(6):675–8.
27. Tone M, Tahara S, Nojima S, Motooka D, Okuzaki D, Morii E. Htr3a is correlated with unfavorable histology and promotes proliferation through erk phosphorylation in lung adenocarcinoma. *Cancer Sci* (2020) 111(10):3953–61. doi: 10.1111/cas.14592
28. Chen DS, Mellman I. Oncology meets immunology: The cancer-immunity cycle. *Immunity* (2013) 39(1):1–10. doi: 10.1016/j.immuni.2013.07.012
29. Li H, Pan Y, Li Y, Li C, Wang R, Hu H, et al. Frequency of well-identified oncogenic driver mutations in lung adenocarcinoma of smokers varies with histological subtypes and graduated smoking dose. *Lung Cancer* (2013) 79(1):8–13. doi: 10.1016/j.lungcan.2012.09.018
30. Krishnamurthy N, Kurzrock R. Targeting the Wnt/Beta-catenin pathway in cancer: Update on effectors and inhibitors. *Cancer Treat Rev* (2018) 62:50–60. doi: 10.1016/j.ctrv.2017.11.002
31. Caso R, Sanchez-Vega F, Tan KS, Mastrogiacomo B, Zhou J, Jones GD, et al. The underlying tumor genomics of predominant histologic subtypes in lung adenocarcinoma. *J Thorac Oncol* (2020) 15(12):1844–56. doi: 10.1016/j.jtho.2020.08.005
32. Chehrizi-Raffae A, Dorff T, Pal S, Lyou Y. Wnt/B-catenin signaling and immunotherapy resistance: Lessons for the treatment of urothelial carcinoma. *Cancers* (2021) 13:889. doi: 10.1038/s41392-020-0205-z
33. Cervantes-Villagrana RD, Albores-Garcia D, Cervantes-Villagrana AR, Garcia-Acevez SJ. Tumor-induced neurogenesis and immune evasion as targets of innovative anti-cancer therapies. *Signal Transduct Target Ther* (2020) 5(1):99. doi: 10.1038/s41392-020-0205-z
34. Riudavets M, Sullivan I, Abdayem P, Planchard D. Targeting Her2 in non-Small-Cell lung cancer (Nsccl): A glimpse of hope? an updated review on therapeutic strategies in nsccl harbouring Her2 alterations. *ESMO Open* (2021) 6 (5):100260. doi: 10.1016/j.esmoop.2021.100260
35. Dentre SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* (2021) 184(8):2239–54.e39. doi: 10.1016/j.cell.2021.03.009
36. Jian M, Nan L, Guocheng J, Qingfu Z, Xueshan Q, Enhua W. Downregulating prl-3 inhibit migration and invasion of lung cancer cell *Via* rhoa and Mdial. *Tumori* (2012) 98(3):370–6. doi: 10.1700/1125.12407
37. Lee J, Kim H, Lee JE, Shin SJ, Oh S, Kwon G, et al. Selective cytotoxicity of the nampt inhibitor Fk866 toward gastric cancer cells with markers of the epithelial-mesenchymal transition, due to loss of naprt. *Gastroenterology* (2018) 155(3):799–814 e13. doi: 10.1053/j.gastro.2018.05.024
38. Fang H, Niu K, Mo D, Zhu Y, Tan Q, Wei D, et al. Recql4-aurora b kinase axis is essential for cellular proliferation, cell cycle progression, and mitotic integrity. *Oncogenesis* (2018) 7(9):68. doi: 10.1038/s41389-018-0080-4
39. Jiang W, Xu J, Liao Z, Li G, Zhang C, Feng Y. Prognostic signature for lung adenocarcinoma patients based on cell-Cycle-Related genes. *Front Cell Dev Biol* (2021) 9:655950. doi: 10.3389/fcell.2021.655950



OPEN ACCESS

EDITED BY

Liang Cheng,
Harbin Medical University, China

REVIEWED BY

Yuhua Yao,
Hainan Normal University, China
Peng Zhang,
Shanghai University of Medicine and
Health Sciences, China

*CORRESPONDENCE

Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

[†]These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 23 June 2022

ACCEPTED 25 July 2022

PUBLISHED 11 August 2022

CITATION

Song J, Huang F, Chen L, Feng K,
Jian F, Huang T and Cai Y-D (2022)
Identification of methylation signatures
associated with CAR T cell in B-cell
acute lymphoblastic leukemia and
non-hodgkin's lymphoma.
Front. Oncol. 12:976262.
doi: 10.3389/fonc.2022.976262

COPYRIGHT

© 2022 Song, Huang, Chen, Feng, Jian,
Huang and Cai. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Identification of methylation signatures associated with CAR T cell in B-cell acute lymphoblastic leukemia and non-hodgkin's lymphoma

Jiwei Song^{1†}, FeiMing Huang^{2†}, Lei Chen^{3†}, KaiYan Feng⁴,
Fangfang Jian⁵, Tao Huang^{6,7*} and Yu-Dong Cai^{2*}

¹College of Life Science, Changchun Sci-Tech University, Shuangyang, China, ²School of Life Sciences, Shanghai University, Shanghai, China, ³College of Information Engineering, Shanghai Maritime University, Shanghai, China, ⁴Department of Computer Science, Guangdong AIB Polytechnic College, Guangzhou, China, ⁵Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, ⁶Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, ⁷CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

CD19-targeted CAR T cell immunotherapy has exceptional efficacy for the treatment of B-cell malignancies. B-cell acute lymphocytic leukemia and non-Hodgkin's lymphoma are two common B-cell malignancies with high recurrence rate and are refractory to cure. Although CAR T-cell immunotherapy overcomes the limitations of conventional treatments for such malignancies, failure of treatment and tumor recurrence remain common. In this study, we searched for important methylation signatures to differentiate CAR-transduced and untransduced T cells from patients with acute lymphoblastic leukemia and non-Hodgkin's lymphoma. First, we used three feature ranking methods, namely, Monte Carlo feature selection, light gradient boosting machine, and least absolute shrinkage and selection operator, to rank all methylation features in order of their importance. Then, the incremental feature selection method was adopted to construct efficient classifiers and filter the optimal feature subsets. Some important methylated genes, namely, *SERPINB6*, *ANK1*, *PDCD5*, *DAPK2*, and *DNAJB6*, were identified. Furthermore, the classification rules for distinguishing different classes were established, which can precisely describe the role of methylation features in the classification. Overall, we applied advanced machine learning approaches to the high-throughput data, investigating the mechanism of CAR T cells to establish the theoretical foundation for modifying CAR T cells.

KEYWORDS

CAR T cell, B-cell acute lymphocytic leukemia, B-cell acute non-Hodgkin's lymphoma, feature selection, classification algorithm, classification rule

Introduction

The chimeric antigen receptor (CAR) T cell immunotherapy is a type of pericyte therapy in which T cells are genetically modified to express chimeric antigen receptors that detect and kill tumor cells in patients (1). In the USA, over 70,000 people are diagnosed with non-Hodgkin's lymphoma (NHL) annually, with a 5-year survival rate of roughly 70% (2). Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer, accounting for approximately 25% of all pediatric cancer cases, and has a high recurrence rate (3). CD19-targeted CAR T-cell immunotherapy has a high response rate in B-cell ALL and B-cell NHL, especially in ALL, with a treatment effectiveness of 90% (4). However, CAR-T therapy is not effective for all tumor patients, and drug-resistant relapse occurs in approximately 50% of patients treated with CD19-targeted CAR-T (5). Some CAR T cells become exhausted, resulting in an increase in inhibitory receptors and a loss of effector function (6, 7). Creating a long-lasting therapeutic response is an essential problem that demands a better knowledge of the cellular and molecular processes that drive CAR T cell proliferation, contraction, and persistence in patients. Studying the specific functions of CAR-transduced T-cells at the molecular level, such as epigenetic level, can help in the understanding of the deeper mechanisms of CAR-T cell immunotherapy and clinical identification of potential targets for effective cancer treatment.

CAR consists of an antigen recognition domain, a co-stimulatory region, and a T cell activation region (8–10). Through multiple signaling cascades, the costimulatory region and T cell activation region activate the CAR T cells, which exhibit proliferative and cytotoxic properties (11). Activated CAR T cells have different gene expression patterns compared with regular T cells, which are influenced by epigenetic modifications (12). DNA may be modified in various ways, the most frequent of which is direct nucleotide methylation. Methylation of promoters results in a decrease in gene expression and suppression of transcription. High-expression genes have high levels of methylation at introns but low levels of methylation at the promoter or regulatory areas (13, 14). Epigenetic imprinting is emerging as a unifying subject in the study of immunological memory and the correlation of long-lasting antitumor responses.

Modifications in DNA methylation shape the overall immune response by altering the phenotype and function of CAR T cells. Zebley et al. showed that alterations in DNA methylation are linked to the proliferation and contraction of CAR T cells and that CD19-targeted CAR T cells acquire DNA methylation features over time. These results suggested that these cells are developing into a progenitor subset of exhausted T cells (15). Meanwhile, Wang et al. discovered that CAR T cells treated with low doses of the demethylating drug decitabine had stronger antitumor, proliferation, and cytokine release abilities.

This result indicates the presence of methylation in CAR T cells that inhibit their oncogenic functions (16). Among the large number of methylation sites, traditional biological experiments cannot meet the requirement of searching for methylation sites that affect the proliferation, failure, and oncogenic functions of CAR T cells. Therefore, this study was focused on how to combine advanced computational methods, such as machine learning, to mine CAR T cell-specific methylation sites to find potential sites for the sustained activation of CAR T cells.

Herein, we devised a process to rapidly screen CAR T cells for specific methylation sites. First, the methylation sites were analyzed and sorted by three feature ranking methods, namely, least absolute shrinkage and selection operator (LASSO) (17), light gradient boosting machine (LightGBM) (18), and Monte Carlo feature selection (MCFS) (19). Then, the incremental feature selection (IFS) (20) method was used to estimate the importance of feature subsets, which were constructed from three ranked methylation site lists, by evaluating the performance of classifiers on these subsets. One optimal feature subset was obtained from each list generated by one feature ranking method. The intersection of all obtained optimal feature subsets was investigated. The methylation sites that recurred multiple times were considered to be highly correlated with the specific functions of the CAR T cell, because the three feature ranking methods used different and independent concepts. Moreover, we also used decision trees (DTs) (21) to create quantitative classification rules that can accurately describe the composition of features for distinguishing each class. All in all, we identified the methylation sites associated with specific functions of the CAR T cells on a large scale using an efficient machine learning based framework and provided a functional description of highly ranked methylation sites in conjunction with the literature.

Materials and methods

Data and preprocessing

The T-cell methylation profiles of 157 patients with B-cell malignancies, including ALL and NHL, were downloaded from the GEO database under the accession number GSE179414 (22). The dataset comprised 77 ALL and 37 NHL cases, who were treated with CART19 cells. These two groups of patients were injected with CAR-transduced T cells and were referred to as ALL transduced and NHL transduced samples, respectively. Meanwhile, 13 ALL and 30 NHL cases were also included in the dataset, but they were not given CART19 cells. These patients were injected with CAR-untransduced T cells and were referred to as ALL untransduced and NHL untransduced

samples. Each group was deemed as a class in this study. We investigated their essential differences by studying the classification problem on these classes. Furthermore, each sample in the dataset was represented by 865,859 methylation sites. These sites were termed as features in this investigation.

Feature ranking methods

A large number of methylation sites were involved in the investigated methylation profiles, which were deemed as features in this study. Evidently, a small proportion of features were highly related to distinguish the CAR-untransduced and -transduced T cells. The powerful feature analysis method in machine learning was necessary. Here, three such methods were employed, including MCFS (19), LightGBM (18), and LASSO (17).

Monte Carlo feature selection

The MCFS algorithm is a DT-based method for determining the relevance of features. This method was first proposed by Micha et al. and has been widely used in tackling various complex medical and biological problems, showing promise in solving such problems (19, 23, 24).

The procedures of MCFS can be summarized as the following steps: (1) s feature subsets are randomly constructed from all features; (2) For each feature subset, t DTs are constructed by randomly selecting training and test samples from the original datasets; (3) After $t \times s$ DTs have been built, each feature g is evaluated by the relative importance (RI), which can be computed as follows:

$$RI_g = \sum_{\tau=1}^{st} (wAcc)^u \sum_{n_g(\tau)} IG(n_g(\tau)) \left(\frac{no.in \ n_g(\tau)}{no.in \ \tau} \right)^v, \quad (1)$$

where $wAcc$ is the weighted accuracy; $IG(n_g(\tau))$ stands for the information gain (IG) of $n_g(\tau)$ (a DT node with the feature g); $no.in \ n_g(\tau)$ stands for the number of samples in $n_g(\tau)$; $no.in \ \tau$ stands for the sample sizes in the tree root; and u and v are two settled positive integers. According to the RI values of all features, they are ranked in a feature list by the decreasing order of their RI values.

In this study, we adopted the MCFS program downloaded from <http://www.ipipan.eu/staff/m.draminski/mcfs.html>. Default parameters were used to execute such program, where u and v were set to one.

Light gradient boosting machine

LightGBM is an iterative boosting tree classifier proposed by Microsoft and is a modified version of the gradient boosting DT (18). LightGBM uses the total number of times (i.e., T_Split) that each feature is involved in the trees iteratively created and the gain (i.e., T_Gain) that a feature is utilized for splitting in all DTs

as measurements of feature relevance for the prediction. They are defined as

$$T_Split = \sum_{t=1}^K Split_t, \quad (2)$$

$$T_Gain = \sum_{t=1}^K Gain_t, \quad (3)$$

where K is the K DTs generated by K iterations. Here, we used T_Split as a metric to measure the importance of features, i.e., features were sorted in the decreasing order of their T_Split values.

This study adopted the LightGBM program retrieved from <https://lightgbm.readthedocs.io/en/latest/>. It was performed with its default parameters.

Least absolute shrinkage and selection operator

The LASSO algorithm is a feature selection method based on linear regression models that selects and compresses variables to prevent overfitting (17). This method uses the L1 paradigm to create a penalty function that selectively removes lower-correlation variables by imposing a bigger penalty on the larger value of the feature variables. This process results in a model with fewer feature variables and effectively avoiding overfitting. If the coefficients of the input features did not contribute positively to the training of the machine learning model, they were scaled down. As a result, the features could be ranked according to their coefficients.

Here, the LASSO package integrated in Scikit-learn (25) was used and its default parameters were adopted.

Incremental feature selection

By three feature ranking methods, all features were ranked in three lists. Evidently, top features in each list were important. However, determining the number of top features was still a problem. Thus, the IFS method (20, 26–28) was employed, which can determine the suitable number of top features. The procedures of IFS method can be divided as follows: (1) Several feature subsets are constructed based on the ranked feature list, which consists of some top features in the list; (2) A classifier is constructed on samples represented by features in each subset and its performance is evaluated by ten-fold cross-validation (29); (3) The classifier with the best performance can be found and the feature subset used in this classifier is picked up as the optimal feature subset. As three ranked feature list was produced in this study, IFS method was executed on each list. Three optimal feature subsets were obtained. We drew Venn diagrams for these three feature sets to display and analyze their intersection results.

Synthetic minority over sampling technique

As described in Section Data and preprocessing, the size of the class with the most samples (77) was about six times as large as that of the class with the least samples (13). Given the imbalance in sample size, when building and evaluating the classifiers, the predicted results would be biased toward the classes with a larger sample size, reducing the generalization ability for the model. In view of this, synthetic minority oversampling technique (SMOTE) algorithm was used in this study to effectively achieve data balance by enlarging the size of each minority class (30, 31). It generates new samples for each minority class by the linear combination of two samples in the same minority class, which are near enough. Finally, all classes have the same number of samples. This study adopted the SMOTE program obtained from <https://github.com/scikitlearn-contrib/imbalanced-learn>. Likewise, its default parameters were used.

Classification algorithm

In the IFS method, classifiers were built to evaluate the importance of constructed feature subsets. A certain classification algorithm was necessary to execute the IFS method. In this study, we applied four classification algorithms, namely, K-nearest neighbor (KNN) (32), support vector machine (SVM) (33), random forest (RF) (34), and DT (21). The purpose of employing these algorithms was to fully test the importance of each constructed feature subset and select the best one.

The KNN algorithm is one of the most classic classification algorithms. Its principle is quite simple. However, its performance is still acceptable in some cases. Given a test sample, KNN finds its k nearest neighbors in the training dataset. According to the labels of these neighbors, the label of the test sample can be determined.

The SVM is a classification algorithm based on statistical learning theory. It generally maps samples into a high-dimensional space by using a kernel function and linearly separates them by finding the maximum margin separating hyperplane. For a test sample, it is also mapped into the same high-dimensional space and its class is determined by the side of hyperplane that the test sample is located.

The RF is also a classic classification algorithm, which is quite different from SVM. In fact, it is an ensemble algorithm, which contains several DTs. Each DT is built by randomly selecting features and samples. For a test sample, each DT gives its prediction. RF integrates these predictions using majority voting.

For the above three classification algorithms, their classification principles are quite difficult for us to understand. Thus, few insights can be extracted from them. DT has its merits in this regard. It is a white-box algorithm, whose classification procedures are completely open, giving opportunities for us to uncover its principle learned from the given dataset, thereby access more knowledge from the dataset. A DT is a tree structure consisting of a series of nodes and branches that use logical operations. Two types of nodes are contained in a DT, they are branch and leaf nodes. The branch node is always related to one feature. According the threshold, samples in a branch node are classified into two groups. The leaf node stands for one class. Samples that reach such node are assigned the corresponding class label. During predictions, it starts at the root node and sorts the test samples down the tree according to the thresholds defined at each branch node. Furthermore, a DT can be represented by lots of if-then rules. Each rule is constructed by a path from the root node to one leaf node. From these rules, a clearer picture on each class can be uncovered.

Above algorithms have been applied to construct various classifiers in dealing with biological and medical problems (35–40). In this study, we used the corresponding Python Scikit-learn packages (25) of above four classification algorithms to implement them.

Performance evaluation

To evaluate the performance of all classifiers constructed in the IFS method, several measurements were employed. First, as multi-class classifiers, the overall accuracy (ACC) was adopted, which is the most accepted measurements. It is defined as the proportion of correctly predicted samples among all samples. However, such measurement is not perfect if the sizes of classes are quite different. Thus, we also employed the Matthew correlation coefficients (MCC) (41), which is deemed as a balanced measurement. As four classes were involved, the MCC in multi-class was adopted, which is defined as

$$MCC = \frac{cov(X, Y)}{\sqrt{cov(X, X)cov(Y, Y)}}, \quad (4)$$

where X and Y are two binary matrices, indicating the true and predicted class of each sample.

In addition, we computed the precision, recall and F1-score for each class, which is defined as

$$Precision = \frac{TP}{TP + FP}, \quad (5)$$

$$Recall = \frac{TP}{TP + FN}, \quad (6)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

where *TP* stands for the number of samples in such class which are correctly predicted, *FP* is the number of samples in other classes which are classified into this class, *FN* is the number of samples in such class which are wrongly predicted. According to F1-score on each class, the macro F1 and weighted F1 are further computed to give a whole evaluation on classifiers. For macro F1, it is defined as the mean of all F1-score values on all classes, whereas weighted F1 integrates all F1-score values by further considering the class sizes, that is, it is the weighted mean of F1-score values.

In this study, weighted F1 was picked up as the key measurement. Other measurements were provided as reference.

Functional enrichment analysis

By analyzing the T-cell methylation profiles downloaded from the GEO with several machine learning algorithms, the optimal feature subsets, containing several methylation probes, were obtained. After taking the union operation on these subsets and mapping features in the union set onto the genes, ClusterProfiler in R was used to calculate the enrichment of these genes on GO terms and KEGG pathways (42). The p-value was corrected with FDR, and 0.05 was chosen as the cutoff value. Only the GO terms and KEGG pathways with FDR<0.05 were considered statistically significant.

Results

We built a machine learning based framework for analyzing CAR-transduced and untransduced T cells in different B-cell

malignancies and further constructed efficient classifiers to discriminate CAR-transduced and untransduced T cells. The entire procedures are illustrated in Figure 1. The detailed results were listed in this section.

Results of feature selecting methods

Each sample was represented by a large number of features (methylation sites). They were deeply analyzed by three feature ranking methods (MCFS, LightGBM, and LASSO). Each method produced one feature list, which is provided in Table S1. It was necessary to pointed out that only features with evaluation score (RI for MCFS, T_Split for LightGBM and coefficient for LASSO) larger than zero were provided in Table S1. The top-ranked features are considered to be important because of their participation in the classification. Their biological significance and the reasons why they are important as core classification features would be discussed in Section Discussion.

Results of IFS method

The three ordered feature lists created by three feature ranking methods were fed into the IFS method one by one and four classification algorithms (DT, KNN, RF and SVM) were used in the IFS method. To save time, we only considered the top 1000 features in each list. For each list, 1000 possible feature subsets were constructed, on which 1000 classifiers with one give classification algorithm were built and evaluated by ten-fold cross-validation. The evaluated results, including measurements listed in Section Performance evaluation, are available in Table S2.

For the feature list yielded by MCFS method, we plotted an IFS curve for each classification algorithm to illustrate

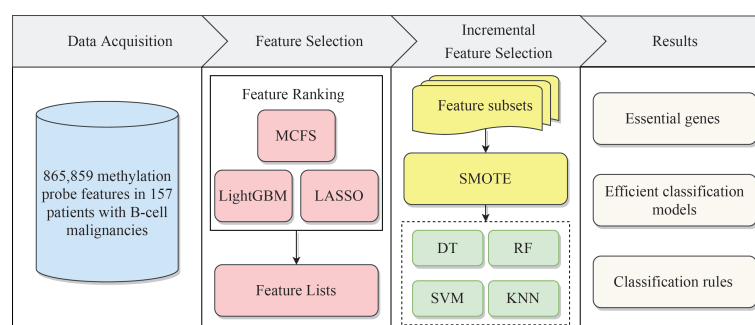


FIGURE 1

Flow chart of the entire analytical process. The 865,859 methylation probe signals on patients with B-cell malignancies were ranked according to feature importance by using three feature ranking algorithms, namely, MCFS, LightGBM, and LASSO. Then, three ordered feature lists were fed into the incremental feature selection (IFS) method, which incorporates four classification algorithms. Finally, based on the IFS results, the essential genes, efficient classification models and classification rules were extracted.

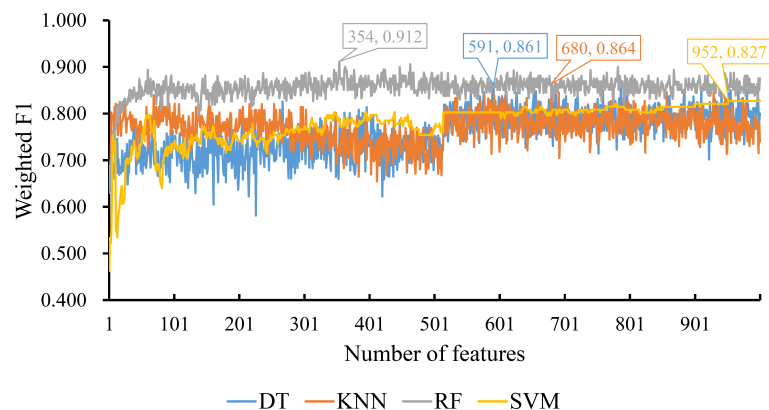


FIGURE 2

IFS curves for displaying the performance of four classification algorithms on the feature list yielded by MCFS method. The best classifiers on different algorithms yield the weight F1 values of 0.861, 0.864, 0.912 and 0.827, respectively, which use top 591, 680, 354 and 952, respectively, features in the list.

TABLE 1 Performance of the best classifiers using different classification algorithms and feature ranking methods.

Feature ranking method	Classification algorithm	Weighted F1	Macro F1	ACC	MCC
MCFS	DT	0.861	0.866	0.860	0.792
	KNN	0.864	0.897	0.860	0.801
	RF	0.912	0.914	0.911	0.866
	SVM	0.827	0.808	0.822	0.746
LightGBM	DT	0.956	0.965	0.955	0.935
	KNN	0.938	0.952	0.936	0.909
	RF	0.975	0.981	0.975	0.963
	SVM	0.950	0.953	0.949	0.927
LASSO	DT	0.912	0.921	0.911	0.869
	KNN	0.943	0.933	0.943	0.914
	RF	0.987	0.990	0.987	0.981
	SVM	0.987	0.990	0.987	0.981

its performance on different feature subsets, which is shown in Figure 2. It can be observed that DT, KNN, RF and SVM can yielded the highest weighted F1 values of 0.861, 0.864, 0.912 and 0.827, respectively. These values were obtained by using top 591, 680, 354 and 952, respectively, features in the list, which comprised the optimal feature subsets for these four classification algorithms. With the optimal feature subsets, we can build the best DT, KNN, RF and SVM classifiers. The values of macro F1, ACC and MCC of these classifiers are listed in Table 1. Furthermore, their performance on four classes is illustrated in Figure 3. Evidently, the best RF classifier was superior to other three best classifiers and the best DT classifier was only better than the best SVM classifier.

For the feature list generated by LightGBM method, four IFS curves were also drawn, which are shown in Figure 4. When top 181, 12, 140 and 43 features in the list were adopted, four classification algorithms produced the highest weighted F1 values of 0.956, 0.938, 0.975 and 0.950, respectively. These features constituted the optimal feature subset for each classification algorithm. Furthermore, the best DT/KNN/RF/SVM classifier was built with its corresponding optimal feature subset. The detailed performance of these best classifiers is listed in Table 1 and shown in Figure 3. Likewise, the best RF classifier still provided the highest performance. As for the best DT classifier, it was a little better than the best KNN and SVM classifiers.

For the last feature list generated by LASSO method, IFS curves were also plotted, as shown in Figure 5. The highest weighted F1 for DT, KNN, RF and SVM were 0.912, 0.943, 0.987 and 0.987, respectively. To reach such performance, top 9, 12, 28 and 111, respectively, features were used. These features comprised the optimal feature subset for each classification algorithm. Similarly, the best DT, KNN, RF and SVM classifiers were constructed with the corresponding optimal feature subsets. Table 1 and Figure 3 provide their detailed performance. The best RF classifier provided equal performance to the best SVM classifier. However, the best RF classifier adopted much less features than the best SVM classifier. Thus, this classifier was still deemed to be better than other three classifiers. On the other hand, the best DT classifier gave the lowest performance among all four best classifiers.

Based on the above arguments, the best RF classifier always provided better performance than other three best classifiers on each feature list. Among three RF classifiers built on three feature lists, the RF classifier on the list generated by LASSO provided the highest performance. Such classifier can be a useful tool to

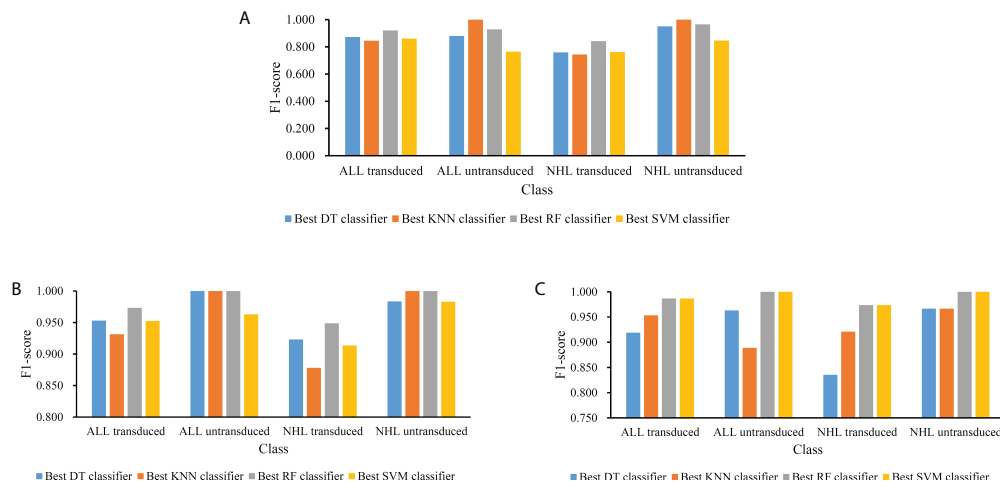


FIGURE 3 Performance of the best classifiers using different classification algorithms and feature lists on four classes. (A) Feature list generated by MCFS method; (B) Feature list generated by LightGBM method; (C) Feature list generated by LASSO method.

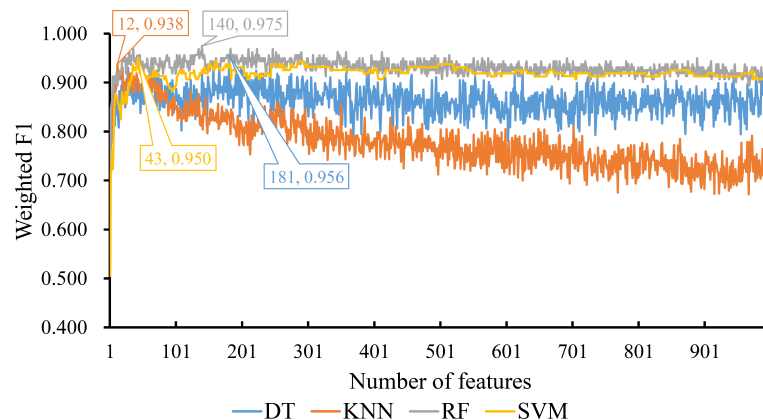


FIGURE 4 IFS curves for displaying the performance of four classification algorithms on the feature list yielded by LighGBM method. The best classifiers on different algorithms yield the weight F1 values of 0.956, 0.938, 0.975 and 0.950, respectively, which use top 181, 12, 140 and 43, respectively, features in the list.

discriminate CAR-transduced and untransduced T cells. On the other hand, the DT classifiers generally gave the low performance. However, they can provide more clues to uncover the differences between CAR-transduced and untransduced T cells.

Feature intersection

As mentioned above, on each feature list yielded by one feature ranking method, the best RF classifier was always better than other three best classifiers. Thus, its optimal feature subset was picked up as the optimal feature subset for one feature ranking method. In

detail, the optimal feature subsets for MCFS, LightGBM and Lasso consisted of the top 354, 140, and 28 features in the lists generated by MCFS, LightGBM, and LASSO, respectively.

For each above-mentioned optimal feature subset, features in such subset were mapped onto their related genes, which comprised the optimal gene subset. Concretely, the optimal gene set for MCFS, LightGBM and LASSO contained 231, 97 and 16 genes, respectively. Detailed genes in these sets are provided in Table S3. The intersection of these three gene sets was investigated and plotted in one Venn diagram, as shown in Figure 6. It can be observed that no genes were contained in all three optimal gene sets, three genes (*SERPINB6*, *ANK1*, *OST4*)

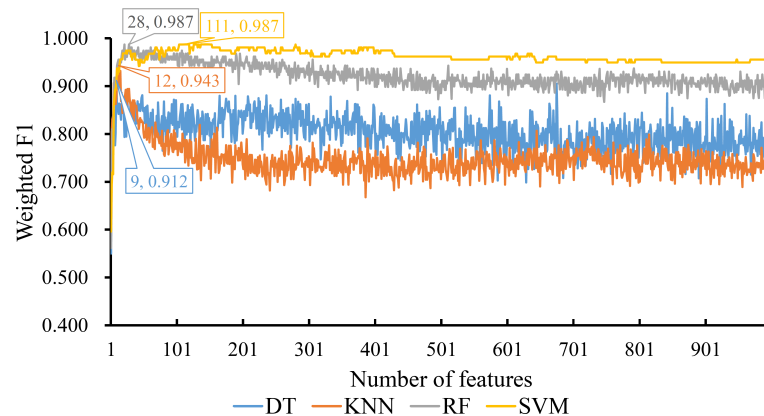


FIGURE 5

IFS curves for displaying the performance of four classification algorithms on the feature list yielded by LASSO method. The best classifiers on different algorithms yield the weight F1 values of 0.912, 0.943, 0.987 and 0.987, respectively, which use top 9, 12, 28 and 111, respectively, features in the list.

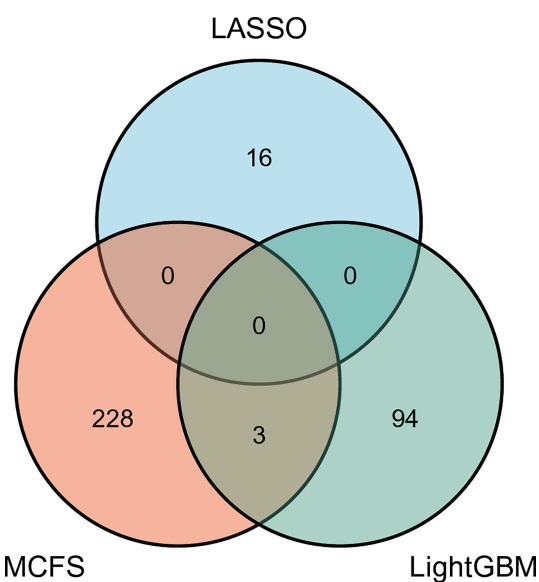


FIGURE 6

Venn diagram to show the intersection of the optimal gene sets for MCFS, LightGBM, and LASSO. Three genes are contained in two optimal gene sets, indicating their importance.

were in two optimal gene sets. Overlapped genes would be discussed in Section Discussion.

Classification rules

Although the DT classifiers were generally weaker than RF classifiers, they can provide clues hidden in the investigated

methylation profiles, which cannot be extracted by other classifiers. According to the IFS results on three feature lists, the best DT classifiers used top 591, 181 and 9 features in three lists, respectively. With these features, three DTs were learned on all samples. Each DT induced a rule set, which contained 17, 10 and 19 rules, respectively. Detailed rules are listed in Table S4. In each rule set, each class was assigned at least one rule, as shown in Figure 7. Following the rules in each rule set, we can determine the class of a test sample. Furthermore, their most contributions were the clear descriptions on different methylation patterns on CAR-transduced and untransduced T cells. This would be discussed in Section Discussion.

Results of enrichment analysis

The optimal feature subsets for three feature ranking methods were determined by the IFS method. We mapped the methylation probes in three optimal feature subsets to genes, yielding a total of 341 genes. Then, the functional enrichment analysis was performed on these genes. The enrichment results are provided in Table S5. Two GO terms were enriched by 341 genes, whereas no KEGG pathways were enriched by these genes with FDR<0.05. GO enrichment result indicated that 12 of these genes were involved in the splicing process, suggesting that the transcripts of these genes may be involved in regulating CAR T-cell processes.

Discussion

In this study, we applied several advanced machine learning algorithms to deeply mine the T-cell methylation profiles of

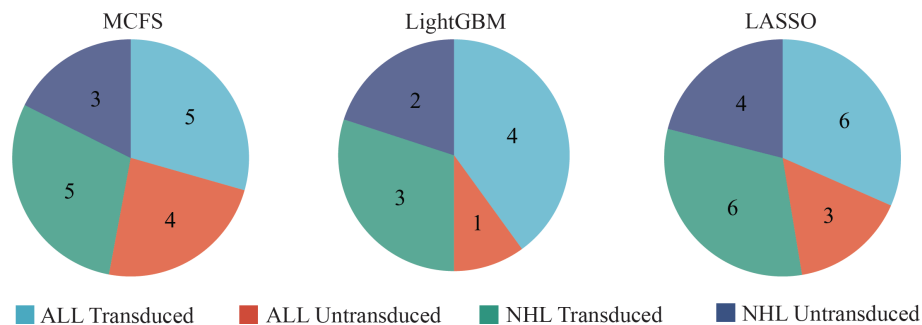


FIGURE 7

The number of rules extracted from the decision tree built on feature lists yielded by MCFS, LightGBM, and LASSO, respectively, on four classes.

patients with B-cell malignancies. Latent important genes were obtained and interesting classification rules were constructed. This section gave extensive analysis on these genes and rules.

Top ranked genes in multiple algorithms

The first gene was *SERPINB6* (targeted by probes cg27001747 and cg04181408), which encoded a member of the serpin superfamily and ovalbumin-serpin subfamily (43, 44). *SERPINB6* appeared in both LightGBM and MCFS in the subset of optimal features. Both methylation probes were linked to the promoter and found in the 5'-UTR region of *SERPINB6*, suggesting that they may affect the transcriptional regulation of this gene. Serpinb9, a homolog of *SERPINB6*, has been shown to protect T cells from Granzyme-B leaked from granules and also participates in T cell homeostasis (45, 46). Although the function of *SERPINB6* in T cells has yet to be established, this protein is important to other immune cells. In neutrophils and monocytes, *SERPINB6* inhibits Cathepsin G, thereby preventing programmed necrosis (47). Thus, *SERPINB6* may play a role in the normal functioning of CAR-T cells. However, more research is needed to confirm this concept. Furthermore, *SERPINB6* methylation has been linked to the risk of CLL pathogenicity (48). This result demonstrates the precision by which our method can identify CAR-T cell-specific genes and differential genes in B-cell malignancies.

The next probes identified were cg09405790 and cg02172579, which both targeted the gene body of *ANK1*. *ANK1* was found in the subset of optimum features in LightGBM and MCFS. *ANK1* is a modular adaptor protein that mediates the connection of integral membrane proteins to the spectrin cytoskeleton (49). *ANK1* methylation has been shown to regulate the expression of microRNA-486-5p, which inhibits Interleukin-22 production by helper T cells via the Dock1/NF-B/Snail signaling pathway. Such process results in cancer suppression (50, 51).

The cg18756060 and cg04001935 probes were designed to detect the DNA methylation status in a specific intergenic region on chromosome 2 (chr2:27294139-27294915) according to GRCh37. Such region has been shown to be the coding region of gene *OST4*, and the protein encoded by *OST4* is an important subunit of oligosaccharyltransferase (OST). Similar to *SERPINB6* and *ANK1*, *OST4* is an intersection feature of the optimal feature subsets of LightGBM and MCFS. Eukaryotic OSTs catalyze the N-glycosylation of nascent polypeptides in the lumen of the endoplasmic reticulum, a conserved biosynthetic process that diversifies the structure and function of proteins (52). Kumar et al. found that N-glycosylation activity remained elevated during the activation and expansion of human T cells, and lymphocytes in a resting state had lower N-glycosylation activity (53). These results suggest that OST was involved in T cell activation in transduced CARs, and that OST activity was influenced by methylation of *OST4*.

PDCD5 (also known as *TFAR19*), which is targeted by the optimal features cg13563193, has been generally reported to participate in immunoregulation. *PDCD5* is at the top of the list of feature rankings obtained with the LASSO method. *PDCD5* interacted with FOXP3 to promote FOXP3 acetylation, hence reducing effector cytokine production (54). Meanwhile, the methylation signal of *PDCD5* was primarily found in the promoter region, which negatively regulated the *PDCD5* expression and thus relaxed the immunosuppressive effect of Treg. This activity could explain the mechanism by which the CAR-T cells were activated and therefore appeared in our list. In addition, in hepatocellular carcinoma, the *PDCD5* overexpression stimulates the promoter activity of KLF9, and the upregulation of KLF9 inhibits cell migration and proliferation (55). This phenomenon also suggests that the cg13563193 methylation signature may suppress the expression level of *PDCD5*. Yuan et al. have discovered that *PDCD5* inhibits the production of proinflammatory mediators and promotes the secretion of anti-inflammatory cytokines by modifying the T-lymphocyte homeostasis (56). The two

hallmark clinical toxicities associated with CAR-T cell therapy are cytokine release syndrome (CRS) and neurotoxicity (57, 58). The characteristics of CRS produce massive inflammation, suggesting a possible involvement of PDCD5 in this process.

The probe cg07632860 was developed to detect the methylation status of the transcription start site of the *DAPK2*. *DAPK2* is at the top of the list of feature rankings obtained with LASSO. *DAPK2* encodes a member of the serine/threonine protein kinase family, which functions as a tumor suppressor and regulates autophagic and apoptotic processes in various cell types (59, 60). When T lymphocytes are activated, they secrete inflammatory cytokines, such as TNF- and IL-6. During this process, *DAPK2* is activated by T cell receptor, which inhibits T-cell activation (61, 62). We discovered that cg07632860 targeted the regulatory region of *DAPK2*, implying that it may limit the expression level of the protein. Meanwhile, *DAPK2* expression has been found to be downregulated in ALL and NHL (63). Low levels of *DAPK* lead to T-cell activation, which implies the CAR-T cell activation mode. Furthermore, the inflammatory cytokines IL-17 and IL-32 have been demonstrated to use *DAPK2* as a signaling mediator (64). Whether the production of cytokine storm, one of the side effects of CAR-T immunotherapy, is linked to *DAPK2* is worthy of investigation.

The next predicted gene, *DNAJB6*, targeted by cg18753341, encodes a member of the DNAJ protein family, which is one of two key groups of molecular chaperones involved in biological activities, such as protein folding and oligomeric protein complex assembly. Strict control of the cell cycle process is essential for the proper functioning of T lymphocytes. Slfn1 has been shown to play an important role in the establishment and maintenance of T lymphocyte quiescence (65). Overexpression of *DnaJB6* increases Slfn1 nuclear accumulation and causes cell-cycle arrest, whereas Slfn1 is mostly sequestered in the cytoplasm, and no cell-cycle arrest has been detected in *DnaJB6* knock-down cells (66). Furthermore, transgenic expression of *DNAJB6* in T cells blocks Slfn1 degradation, enhances its nuclear import, and results in T cell proliferation suppression when T cell receptors are activated (66). In addition, *DNAJB6* is neurotoxic when overexpressed in primary neurons, suggesting that it may be a potential locus for CAR-T treatment to eliminate side effects (67).

Analysis of classification rules

In addition to the functional analysis of the top-ranked features, we also mined the specific rules used to distinguish each class based on the classification tree structure of the DTs. The rules of each class consisted of methylation probes and their signal intensities, and each methylation probe was linked to a gene to describe its function in greater depth.

The first rule was aimed to distinguish T cells derived from patients with ALL that have been transduced with CAR. *MYCN*,

which is targeted by cg13799853, was an important site with low methylation, according to the classification rule based on LASSO results. In our classification rules, *MYCN* exhibited lower methylation levels. *MYCN* has been demonstrated to have lower methylation levels in relapsed children with B-cell acute lymphoblastic leukemia (B-ALL), which was consistent with the usage of *MYCN* in this study as a key feature to differentiate B-ALL (68). *MYCN* also downregulates *DKK3* expression and activates the Wnt/ β -catenin signaling pathway at the transcriptional level, boosting the development of B-ALL (69). Meanwhile, *MYCN* apparently decreases the interferon signaling, promoting a non-inflamed and T-cell infiltration-poor ("cool") tumor microenvironment (70). In the classification rule based on the MCFS results, *HDGF* targeted by cg18593717 was an important locus, which exhibited a lower methylation level. *HDGF* has been demonstrated to cause Foxp3 + Treg differentiation and that Tregs decrease CD8+ cytotoxic T cell activity (71). This phenomenon suggests that *HDGF* may act as a potential gene driving the activation of CAR-T cell.

The second rule was used to distinguish the T cells derived from patients with ALL without transduced CARs. After constructing the DT by using the optimal subset obtained after MCFS, the classification rules were established. Among them, hypomethylation of the *ZBTB7A*, also known as LRF, was an important quantitative rule. Many studies have shown that *ZBTB7A* is closely associated with B and T cell differentiation and plays an important role in their fate decisions (72, 73). Meanwhile, dysregulation in B-cell maturation can lead to the development of autoimmune syndromes and B-cell malignancies (73). *ZBTB7* was described in the rules in our study, because it plays an important role in both immune processes and cancer development.

The next two rules were used to distinguish between CAR-transduced and untransduced T cells derived from patients with NHL. *TP73* targeted by cg10654015 appeared in our rules and exhibited a higher methylation status. *TP73* has been demonstrated to be frequently methylated in NHLs (74). This result is consistent with the highly methylated results of *TP73* found in our study, indicating the accuracy of our method. Furthermore, *TP73* deletion has been shown to impact lymphoma formation by several mechanisms, such as altered gene expression patterns, defective early T-cell growth, impaired apoptosis, and chromosomal abnormality accumulation (75). This phenomenon suggests that *TP73* may be a potential target for the modification of CAR-T cells.

Conclusion

We applied a powerful computational strategy based on DNA methylation probe data to uncover the features of CAR T cells across diverse B-cell malignancies. The outcomes can be summarized in the three key components. First, a series of

methylation signatures and genes were extracted, which can be used to distinguish cells from four different origins. The findings provided a theoretical foundation to precisely modify CAR T cells and treat B-cell malignancies. Second, efficient multi-class classifiers were built to aid in a more accurate delineation of T cells prior to treatment. The delineation of T cells facilitated the screening for T cells that could efficiently suppress cancer *in vivo* and further improve those that were not successfully transduced. Finally, some classification rules were built to specifically distinguish a particular class of cells. These rules aided to better understand the specific functions of CAR T cells by describing the degree of gene methylation.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179414>.

Author contributions

TH and YD-C designed the study. LC and KYF performed the experiments. JS, FMH and FJ analyzed the results. JS, FMH and LC wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

Funding

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences [XDB38050200, XDA26040304], National Key R&D Program of China [2018YFC0910403], the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences [202002].

References

1. Singh AK, Mcguirk JP. CAR T cells: continuation in a revolution of immunotherapy. *Lancet Oncol* (2020) 21:e168–78. doi: 10.1016/S1470-2045(19)30823-X
2. Bethesda M. *SEER cancer stat facts: Non-Hodgkin lymphoma*. Available at: <https://seer.cancer.gov/statfacts/html/nhl.html> (Accessed Mar 2022).
3. Ward E, Desantis C, Robbins A, Kohler B, Jemal A. Childhood and adolescent cancer statistic. *CA Cancer J Clin* (2014) 64:83–103. doi: 10.3322/caac.21219
4. Maude SL, Frey N, Shaw PA, Aplenc R, Barrett DM, Bunin NJ, et al. Chimeric antigen receptor T cells for sustained remissions in leukemia. *N Engl J Med* (2014) 371:1507–17. doi: 10.1056/NEJMoa1407222
5. Shah NN, Fry TJ. Mechanisms of resistance to CAR T cell therapy. *Nat Rev Clin Oncol* (2019) 16:372–85. doi: 10.1038/s41571-019-0184-6
6. Wherry EJ, Kurachi M. Molecular and cellular insights into T cell exhaustion. *Nat Rev Immunol* (2015) 15:486–99. doi: 10.1038/nri3862
7. Schietinger A, Philip M, Krisnawan VE, Chiu EY, Delrow JJ, Basom RS, et al. Tumor-specific T cell dysfunction is a dynamic antigen-driven differentiation program initiated early during tumorigenesis. *Immunity* (2016) 45:389–401. doi: 10.1016/j.immuni.2016.07.011
8. Kalos M, Levine BL, Porter DL, Katz S, Grupp SA, Bagg A, et al. T Cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia. *Sci Transl Med* (2011) 3:95ra73. doi: 10.1126/scitranslmed.3002842
9. Brentjens RJ, Davila ML, Riviere I, Park J, Wang X, Cowell LG, et al. CD19-targeted T cells rapidly induce molecular remissions in adults with chemotherapy-refractory acute lymphoblastic leukemia. *Sci Transl Med* (2013) 5:177ra138. doi: 10.1126/scitranslmed.3005930
10. Brudno JN, Lam N, Vanasse D, Shen YW, Rose JJ, Rossi J, et al. Safety and feasibility of anti-CD19 CAR T cells with fully human binding domains in patients with b-cell lymphoma. *Nat Med* (2020) 26:270–80. doi: 10.1038/s41591-019-0737-3
11. Larson RC, Maus MV. Recent advances and discoveries in the mechanisms and functions of CAR T cells. *Nat Rev Cancer* (2021) 21:145–61. doi: 10.1038/s41568-020-00323-z
12. Akbari B, Ghahri-Saremi N, Soltantoyeh T, Hadjati J, Ghassemi S, Mirzaei HR. Epigenetic strategies to boost CAR T cell therapy. *Mol Ther* (2021) 29:2640–59. doi: 10.1016/j.ymthe.2021.08.003

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/fonc.2022.976262/full#supplementary-material>

SUPPLEMENTARY TABLE 1

Feature lists obtained by MCFS, LightGBM and LASSO.

SUPPLEMENTARY TABLE 2

Performance of IFS with different classification algorithms and feature lists.

SUPPLEMENTARY TABLE 3

Optimal gene sets based on MCFS, LightGBM, and LASSO methods, and their intersection results.

SUPPLEMENTARY TABLE 4

Classification rules generated by decision tree.

SUPPLEMENTARY TABLE 5

GO and KEGG enrichment results on the optimal gene sets of three feature ranking methods.

13. Ball MP, Li JB, Gao Y, Lee JH, Leproust EM, Park IH, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* (2009) 27:361–8. doi: 10.1038/nbt.1533
14. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* (2012) 13:484–92. doi: 10.1038/nrg3230
15. Zebley CC, Brown C, Mi T, Fan Y, Alli S, Boi S, et al. CD19-CAR T cells undergo exhaustion DNA methylation programming in patients with acute lymphoblastic leukemia. *Cell Rep* (2021) 37:110079. doi: 10.1016/j.celrep.2021.110079
16. Wang Y, Tong C, Dai H, Wu Z, Han X, Guo Y, et al. Low-dose decitabine priming endows CAR T cells with enhanced and persistent antitumor potential via epigenetic reprogramming. *Nat Commun* (2021) 12:409. doi: 10.1038/s41467-020-20696-x
17. Tibshirani RJ. Regression shrinkage and selection via the LASSO. *Journal of the royal statistical society. Ser B: Methodological* (1996) 73:273–82.
18. Ke G, Meng Q, Finely T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree; *Adv Neural Inf Process Syst* 30 (NIP2017) (2017).
19. Micha D, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J. Monte Carlo Feature selection for supervised classification. *Bioinformatics* (2008) 24:110–7. doi: 10.1093/bioinformatics/btm486
20. Liu HA, Setiono R. Incremental feature selection. *Appl Intell* (1998) 9:217–30. doi: 10.1023/A:1008363719778
21. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Trans systems man cybernetics* (1991) 21:660–74. doi: 10.1109/21.97458
22. Garcia-Prieto CA, Villanueva L, Bueno-Costa A, Davalos V, González-Navarro EA, Urbano-Ispizua JM, et al. Epigenetic profiling and response to CD19 chimeric antigen receptor T-cell therapy in B-cell malignancies. *J Natl Cancer Inst* (2021) 114:436–45. doi: 10.1093/jnci/djab194
23. Chen L, Li J, Zhang YH, Feng K, Wang S, Zhang Y, et al. Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J Cell Biochem* (2018) 119:3394–403. doi: 10.1002/jcb.26507
24. Chen X, Jin Y, Feng Y. Evaluation of plasma extracellular vesicle MicroRNA signatures for lung adenocarcinoma and granuloma with Monte-Carlo feature selection method. *Front Genet* (2019) 10:367. doi: 10.3389/fgene.2019.00367
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* (2011) 12:2825–30.
26. Chen L, Li Z, Zhang S, Zhang Y-H, Huang T, Cai Y-D. Predicting RNA 5-methylcytosine sites by using essential sequence features and distributions. *BioMed Res Int* (2022) 2022:4035462. doi: 10.1155/2022/4035462
27. Ding S, Wang D, Zhou X, Chen L, Feng K, Xu X, et al. Predicting heart cell types by using transcriptome profiles and a machine learning method. *Life* (2022) 12:228. doi: 10.3390/life12020228
28. Zhou X, Ding S, Wang D, Chen L, Feng K, Huang T, et al. Identification of cell markers and their expression patterns in skin based on single-cell RNA-sequencing profiles. *Life* (2022) 12:550. doi: 10.3390/life12040550
29. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *International joint conference on artificial intelligence*. San Francisco, CA, United States: Morgan Kaufmann Publishers Inc. (1995). p. 1137–45.
30. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* (2002) 16:321–57. doi: 10.1613/jair.953
31. Pan X, Chen L, Liu I, Niu Z, Huang T, Cai YD. Identifying protein subcellular locations with embeddings-based node2loc. *IEEE/ACM Trans Comput Biol Bioinform* (2022) 19:666–75. doi: 10.1109/TCBB.2021.3080386
32. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory* (1967) 13:21–7. doi: 10.1109/TIT.1967.1053964
33. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* (1995) 20:273–97. doi: 10.1007/BF00994018
34. Breiman L. Random forests. *Mach Learn* (2001) 45:5–32. doi: 10.1023/A:1010933404324
35. Chen W, Chen L, Dai Q. iMPT-FDNPL: identification of membrane protein types with functional domains and a natural language processing approach. *Comput Math Methods Med* (2021) 2021:7681497. doi: 10.1155/2021/7681497
36. Onesime M, Yang Z, Dai Q. Genomic island prediction via chi-square test and random forest algorithm. *Comput Math Methods Med* (2021) 2021:9969751. doi: 10.1155/2021/9969751
37. Wang Y, Xu Y, Yang Z, Liu X, Dai Q. Using recursive feature selection with random forest to improve protein structural class prediction for low-similarity sequences. *Comput Math Methods Med* (2021) 2021:5529389. doi: 10.1155/2021/5529389
38. Li X, Lu L, Chen L. Identification of protein functions in mouse with a label space partition method. *Math Biosci Eng* (2022) 19:3820–42. doi: 10.3934/mbe.2022176
39. Wu Z, Chen L. Similarity-based method with multiple-feature sampling for predicting drug side effects. *Comput Math Methods Med* (2022) 2022:9547317. doi: 10.1155/2022/9547317
40. Yang Y, Chen L. Identification of drug-disease associations by using multiple drug and disease networks. *Curr Bioinf* (2022) 17:48–59. doi: 10.2174/1574893616666210825115406
41. Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta (BBA)-Protein Structure* (1975) 405:442–51. doi: 10.1016/0005-2795(75)90109-9
42. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* (2021) 2:100141. doi: 10.1016/j.xinn.2021.100141
43. Gettins PG. Serpin structure, mechanism, and function. *Chem Rev* (2002) 102:4751–804. doi: 10.1021/cr010170+
44. Lin HY, Muller YA, Hammond GL. Molecular and structural basis of steroid hormone binding and release from corticosteroid-binding globulin. *Mol Cell Endocrinol* (2010) 316:3–12. doi: 10.1016/j.mce.2009.06.015
45. Hirst CE, Buzza MS, Bird CH, Warren HS, Cameron PU, Zhang M, et al. The intracellular granzyme b inhibitor, proteinase inhibitor 9, is up-regulated during accessory cell maturation and effector cell degranulation, and its overexpression enhances CTL potency. *J Immunol* (2003) 170:805–15. doi: 10.4049/jimmunol.170.2.805
46. Azzi J, Skartsis N, Mounayar M, Magee CN, Batal I, Ting C, et al. Serine protease inhibitor 6 plays a critical role in protecting murine granzyme b-producing regulatory T cells. *J Immunol* (2013) 191:2319–27. doi: 10.4049/jimmunol.1300851
47. Burgener SS, Leborgne NGF, Snipas SJ, Salvesen GS, Bird PI, Benarafa C. Cathepsin G inhibition by Serpinb1 and Serpinb6 prevents programmed necrosis in neutrophils and monocytes and reduces GSDMD-driven inflammation. *Cell Rep* (2019) 27:3646–3656.e3645. doi: 10.1016/j.celrep.2019.05.065
48. Berndt SI, Camp NJ, Skibola CF, Vijai J, Wang Z, Gu J, et al. Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. *Nat Commun* (2016) 7:10933. doi: 10.1038/ncomms10933
49. Cunha SR, Mohler PJ. Ankyrin protein networks in membrane formation and stabilization. *J Of Cell And Mol Med* (2009) 13:4364–76. doi: 10.1111/j.1582-4934.2009.00943.x
50. Tessema M, Yingling CM, Picchi MA, Wu G, Ryba T, Lin Y, et al. ANK1 methylation regulates expression of MicroRNA-486-5p and discriminates lung tumors by histology and smoking status. *Cancer Lett* (2017) 410:191–200. doi: 10.1016/j.canlet.2017.09.038
51. Li H, Mou Q, Li P, Yang Z, Wang Z, Niu J, et al. MiR-486-5p inhibits IL-22-induced epithelial-mesenchymal transition of breast cancer cell by repressing Dock1. *J Cancer* (2019) 10:4695–706. doi: 10.7150/jca.30596
52. Dumax-Vorzet A, Roboti P, High S. OST4 is a subunit of the mammalian oligosaccharyltransferase required for efficient N-glycosylation. *J Cell Sci* (2013) 126:2595–606. doi: 10.1242/jcs.115410
53. Kumar V, Heinemann FS, Ozols J. Interleukin-2 induces N-glycosylation in T-cells: characterization of human lymphocyte oligosaccharyltransferase. *Biochem Biophys Res Commun* (1998) 247:524–9. doi: 10.1006/bbrc.1998.8780
54. Xiao J, Liu C, Li G, Peng S, Hu J, Qu L, et al. PDCD5 negatively regulates autoimmunity by upregulating FOXP3(+) regulatory T cells and suppressing Th17 and Th1 responses. *J Autoimmun* (2013) 47:34–44. doi: 10.1016/j.jaut.2013.08.002
55. Fu DZ, Cheng Y, He H, Liu HY, Liu YF. The fate of krüppel-like factor 9-positive hepatic carcinoma cells may be determined by the programmed cell death protein 5. *Int J Oncol* (2014) 44:153–60. doi: 10.3892/ijo.2013.2147
56. Yuan F, Wang J, Zhang K, Li Z, Guan Z. Programmed cell death 5 transgenic mice attenuates adjuvant induced arthritis by 2 modifying the T lymphocytes balance. *Biol Res* (2017) 50:40. doi: 10.1186/s40659-017-0145-4
57. Hay KA, Hanafi LA, Li D, Gust J, Liles WC, Wurfel MM, et al. Kinetics and biomarkers of severe cytokine release syndrome after CD19 chimeric antigen receptor-modified T-cell therapy. *Blood* (2017) 130:2295–306. doi: 10.1182/blood-2017-06-793141
58. Neelapu SS, Tummala S, Kebriaei P, Wierda W, Gutierrez C, Locke FL, et al. Chimeric antigen receptor T-cell therapy - assessment and management of toxicities. *Nat Rev Clin Oncol* (2018) 15:47–62. doi: 10.1038/nrclinonc.2017.148
59. Bialik S, Kimchi A. The death-associated protein kinases: structure, function, and beyond. *Annu Rev Biochem* (2006) 75:189–210. doi: 10.1146/annurev.biochem.75.103004.142615

60. Michie AM, Mccaig AM, Nakagawa R, Vukovic M. Death-associated protein kinase (DAPK) and signal transduction: regulation in cancer. *FEBS J* (2010) 277:74–80. doi: 10.1111/j.1742-4658.2009.07414.x
61. Chuang YT, Fang LW, Lin-Feng MH, Chen RH, Lai MZ. The tumor suppressor death-associated protein kinase targets to TCR-stimulated NF-kappa b activation. *J Immunol* (2008) 180:3238–49. doi: 10.4049/jimmunol.180.5.3238
62. Lai MZ, Chen RH. Regulation of inflammation by DAPK. *Apoptosis* (2014) 19:357–63. doi: 10.1007/s10495-013-0933-4
63. Tur MK, Daramola AK, Gattenlöhner S, Herling M, Chetty S, Barth S. Restoration of DAP kinase tumor suppressor function: A therapeutic strategy to selectively induce apoptosis in cancer cells using immunokinase fusion proteins. *Biomedicine* (2017) 5:59. doi: 10.3390/biomedicine5040059
64. Turner-Brannen E, Choi KY, Arsenault R, El-Gabalawy H, Napper S, Mookherjee N. Inflammatory cytokines IL-32 and IL-17 have common signaling intermediates despite differential dependence on TNF-receptor 1. *J Immunol* (2011) 186:7127–35. doi: 10.4049/jimmunol.1002306
65. Schwarz DA, Katayama CD, Hedrick SM. Schlafen, a new family of growth regulatory genes that affect thymocyte development. *Immunity* (1998) 9:657–68. doi: 10.1016/S1074-7613(00)80663-9
66. Zhang Y, Yang Z, Cao Y, Zhang S, Li H, Huang Y, et al. The Hsp40 family chaperone protein DnaJB6 enhances Schlafen1 nuclear localization which is critical for promotion of cell-cycle arrest in T-cells. *Biochem J* (2008) 413:239–50. doi: 10.1042/BJ20071510
67. Smith C, D'mello SR. Cell and context-dependent effects of the heat shock protein DNAJB6 on neuronal survival. *Mol Neurobiol* (2016) 53:5628–39. doi: 10.1007/s12035-015-9452-3
68. Bhatia P, Singh M, Singh A, Sharma P, Trehan A, Varma N. Epigenetic analysis reveals significant differential expression of miR-378C and miR-128-2-5p in a cohort of relapsed pediatric b-acute lymphoblastic leukemia cases. *Int J Lab Hematol* (2021) 43:1016–23. doi: 10.1111/ijlh.13477
69. Kong D, Zhao L, Sun L, Fan S, Li H, Zhao Y, et al. MYCN is a novel oncogenic target in adult b-ALL that activates the wnt/ β -catenin pathway by suppressing DKK3. *J Cell Mol Med* (2018) 22:3627–37. doi: 10.1111/jcmm.13644
70. Seier JA, Reinhardt J, Saraf K, Ng SS, Layer JP, Corvino D, et al. Druggable epigenetic suppression of interferon-induced chemokine expression linked to MYCN amplification in neuroblastoma. *J Immunother Cancer* (2021) 9:e001335. doi: 10.1136/jitc-2020-001335
71. Sun AM, Li CG, Zhang YQ, Lin SM, Niu HR, Shi YS. Hepatocarcinoma cell-derived hepatoma-derived growth factor (HDGF) induces regulatory T cells. *Cytokine* (2015) 72:31–5. doi: 10.1016/j.cyto.2014.12.001
72. Maeda T, Merghoub T, Hobbs RM, Dong L, Maeda M, Zakrzewski J, et al. Regulation of b versus T lymphoid lineage fate decision by the proto-oncogene LRF. *Science* (2007) 316:860–6. doi: 10.1126/science.1140881
73. Sakurai N, Maeda M, Lee SU, Ishikawa Y, Li M, Williams JC, et al. The LRF transcription factor regulates mature b cell development and the germinal center response in mice. *J Clin Invest* (2011) 121:2583–98. doi: 10.1172/JCI45682
74. Martinez-Delgado B, Melendez B, Cuadros M, Garcia MJ, Nomdedeu J, Rivas C, et al. Frequent inactivation of the p73 gene by abnormal methylation or LOH in non-hodgkin's lymphomas. *Int J Cancer* (2002) 102:15–9. doi: 10.1002/ijc.10618
75. Nemajerova A, Palacios G, Nowak NJ, Matsui S, Petrenko O. Targeted deletion of p73 in mice reveals its role in T cell development and lymphomagenesis. *PloS One* (2009) 4:e7784. doi: 10.1371/journal.pone.0007784



OPEN ACCESS

EDITED BY

Tianyi Zhao,
Harbin Institute of Technology, China

REVIEWED BY

Yang Yang,
Inner Mongolia University, China
Jingyu Huang,
Department of Thoracic Surgery,
Wuhan University, China

*CORRESPONDENCE

Fan Meng
mengfan@gmu.edu.cn

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 30 June 2022

ACCEPTED 25 July 2022

PUBLISHED 16 August 2022

CITATION

He X, Li W-S, Qiu Z-G, Zhang L,
Long H-M, Zhang G-S, Huang Y-W,
Zhan Y-m and Meng F (2022) A
computational method for large-scale
identification of esophageal cancer-
related genes.
Front. Oncol. 12:982641.
doi: 10.3389/fonc.2022.982641

COPYRIGHT

© 2022 He, Li, Qiu, Zhang, Long, Zhang,
Huang, Zhan and Meng. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use,
distribution or reproduction is
permitted which does not comply with
these terms.

A computational method for large-scale identification of esophageal cancer-related genes

Xin He¹, Wei-Song Li², Zhen-Gang Qiu³, Lei Zhang⁴,
He-Ming Long³, Gui-Sheng Zhang⁵, Yang-Wen Huang⁵,
Yun-mei Zhan⁵ and Fan Meng^{4*}

¹Department of Respiratory and Critical Care, The First Affiliated Hospital of Gannan Medical University, Ganzhou, China, ²Department of pathology, The First Affiliated Hospital of Gannan Medical University, Ganzhou, China, ³Department of Oncology, The First Affiliated Hospital of Gannan Medical University, Ganzhou, China, ⁴Department of Gastroenterology, The First Affiliated Hospital of Gannan Medical University, Ganzhou, China, ⁵School of Basic Medicine, Gannan Medical University, Ganzhou, China

The incidence of esophageal cancer has obvious genetic susceptibility. Identifying esophageal cancer-related genes plays a huge role in the prevention and treatment of esophageal cancer. Through various sequencing methods, researchers have found only a small number of genes associated with esophageal cancer. In order to improve the efficiency of esophageal cancer genetic susceptibility research, this paper proposes a method for large-scale identification of esophageal cancer-related genes by computational methods. In order to improve the efficiency of esophageal cancer genetic susceptibility research, this paper proposes a method for large-scale identification of esophageal cancer-related genes by computational methods. This method fuses graph convolutional network and logical matrix factorization to effectively identify esophageal cancer-related genes through the association between genes. We call this method GCNLMF which achieved AUC as 0.927 and AUPR as 0.86. Compared with other five methods, GCNLMF performed best. We conducted a case study of the top three predicted genes. Although the association of these three genes with esophageal cancer has not been reported in the database, studies by other researchers have shown that these three genes are significantly associated with esophageal cancer, which illustrates the accuracy of the prediction results of GCNLMF.

KEYWORDS

esophageal cancer, gene, graph convolutional network, logical matrix factorization, gene interaction

Introduction

Esophageal cancer is a common gastrointestinal malignancy, and its common clinical symptoms include retrosternal pain and progressive dysphagia (1). Judging from its prevalence, the incidence of esophageal cancer in China is relatively high globally. The pathological type of esophageal squamous cell carcinoma is more common. The typical symptoms of esophageal cancer patients are not obvious in the early stage, and the disease progresses slowly, so it is difficult to detect early. However, when esophageal cancer develops to the middle and advanced stage, the treatment difficulty increases and the prognosis is poor (2). At present, the treatment of patients with esophageal cancer is mainly surgery, radiotherapy, and chemotherapy. The patients with advanced stage have poor curative effect and high mortality (3).

The occurrence of esophageal squamous cell carcinoma usually goes through a long-term and multi-stage development process. In the original efficient and orderly epithelial renewal cycle, carcinogenic factors are continuously exposed. The basal cells first show morphological changes, atypical hyperplasia and invasion to the surface. The squamous epithelial cells show nuclear atypia and abnormal differentiation. In the early stage of carcinogenesis, this pathological change is limited to the inner part of the mucosal layer and does not break through the basement membrane to infiltrate and invade downward. It is called squamous epithelial dysplasia and is the only recognized form of precancerous lesions of esophageal squamous cell carcinoma (4). A 13 year prospective cohort study (5) conducted a long-term follow-up of normal and precancerous people in Linzhou, Henan Province. It was found that compared with normal people, the relative risk of esophageal squamous cell carcinoma in patients with precancerous lesions (regardless of the degree of specific lesions) was 12.7 (5.5-29.6) times higher than that in normal people. Moreover, the cumulative incidence rate of esophageal squamous cell carcinoma in patients initially diagnosed with precancerous lesions at the end of the study was 58%, which was 8% in the population initially diagnosed with no abnormality. Therefore, atypical hyperplasia of squamous epithelium is a high-risk factor and predictor of esophageal squamous cell carcinoma. Timely early diagnosis of patients with precancerous lesions is an important means to reduce the incidence rate of esophageal squamous cell carcinoma. At present, regular gastroscopy screening for high-risk groups is an effective method for early diagnosis of esophageal cancer. However, due to the heterogeneity between patients, different patients with the same diagnosis still have different outcomes and outcomes. Therefore, an in-depth understanding of the causes of esophageal epithelial progression from normal to precancerous lesions to tumors and a comprehensive analysis of the molecular mechanism of tumor occurrence are of indispensable value for us to evaluate the risk of progression of patients with precancerous lesions, improve the diagnosis and

cure rate of patients, and increase the means and opportunities for early diagnosis and treatment.

With the development and progress of next-generation sequencing technology, multi-omics research on tumors has become an indispensable means to explore the mechanism of tumor occurrence and development. In recent years, a number of esophageal cancer genomic studies, including the Cancer Genome Atlas (TCGA) project, have identified a large number of genomic variants in esophageal squamous cell carcinoma by performing whole-exome or whole-genome sequencing of clinically collected tumor tissue samples (6). Although these studies reveal the important role of the identified genomic alterations in ESCC, the question of how normal epithelial cells are transformed into invasive carcinomas through mutations in precancerous lesions remains unanswered due to the cross-sectional design of previous studies. Compared with studies on esophageal squamous cell carcinoma, there are still few studies on esophageal precancerous lesions. Some researchers used microdissection experimental technology to collect tumor lesions and precancerous lesions adjacent to the tumor on paraffin sections of 45 cases of esophageal squamous cell carcinoma, as well as lesions on paraffin sections of 13 precancerous lesions for full penetrance. Subgroup sequencing analysis showed that epithelial cells in the precancerous stage already have mutations similar to those of tumors, including high-frequency mutations in esophageal cancer driver genes such as TP53, NFE2L2, NOTCH1, FAT1, indicating that the precancerous stage Epithelial cells have undergone the effects of genomic variation (7). Coincidentally, in another report, the researchers performed whole-exome sequencing on 227 different pathological stages of 70 patients with esophageal squamous cell carcinoma, and also found that dysplasia and esophageal squamous cell carcinoma have similar driver genes. Moreover, they also found that there were no genomic alterations of the same type of cancer foci in the tissues of simple non-dysplasia, indicating that most of the genomic events related to canceration started from the stage of precancerous lesions (8). Researchers have performed genomic mutation studies on pathologically normal esophagus (9) and their results have shown that although the exon mutation burden of normal esophageal epithelial cells (derived from human individuals without esophageal squamous cell carcinoma) increases with age, but no cancer-related morphological changes occurred from a histopathological point of view. The results showed that although the exon mutation load of normal esophageal epithelial cells (derived from human individuals without esophageal squamous cell carcinoma) increased with age, there were no cancer-related morphological changes from the perspective of histopathology. The above studies suggest that in the overall organizational environment, the genomic changes of epithelial cells are not enough to fully explain the occurrence of esophageal cancer. Other factors such as immunosuppression in the microenvironment (TME) and cell-cell interaction may also play an important role in the occurrence of esophageal squamous cell carcinoma. A number of experimental studies and clinical

analyses have also revealed the impact of TME on tumorigenesis and development in esophageal squamous cell carcinoma. Kashima et al. (10) found that the positive intensity of cancer associated fibroblasts (CAFs) was significantly positively correlated with lymph node metastasis by staining FFPE tissue sections of patients with esophageal squamous cell carcinoma, so they verified this hypothesis through in vitro experiments and in situ metastasis mouse models, CAFs can promote the metastatic ability of cancer cells and can be used as a marker of patient prognosis. Another experimental study on the microenvironment cells of esophageal squamous cell carcinoma found that the up regulation of transcription factor FOXO1 can promote the polarization of macrophages from M0 to M2 by regulating the expression of CCL20 and csf1, while M2 cells play the regulatory functions of anti-inflammatory and immunosuppression, and promote the occurrence of tumors (11). Similarly, Yang et al. found that blocking the recruitment of tumor associated macrophages (TAMs) can significantly reduce the incidence of tumors in the mouse tumorigenesis model and enhance the anti-tumor effect of CD8 + T cells in the tumor microenvironment. More importantly, M2 polarization increases the expression of PD-L2 in TAMs, leading to immune evasion and tumor promotion through PD-1 signaling pathway (12).

A large number of biological experiments have only found a small number of genes related to esophageal cancer. In recent years, some scholars have identified esophageal cancer-related genes through computational methods such as machine learning.

Liu et al. (13) identified genetic biomarkers of esophageal cancer by SALP-seq and machine learning methods. Wang et al. (14) identified the survival risk of esophageal cancer through the Kohonen network clustering algorithm and kernel extreme learning machine. Li et al. (15) used five conventional machine learning methods to identify key prognostic molecules in esophageal squamous cell carcinoma. Most of these previous studies performed gene differential expression analysis through data from a small number of patients to obtain genes related to esophageal cancer. Its sample size is insufficient and there is a sample-specific bias. It has become a trend to predict disease-related features through associations between biomolecules (16, 17). Therefore, we intend to identify esophageal cancer-related genes by their associations and correlation signatures. Through the known gene signatures associated with esophageal cancer, a computational model was constructed to explore the association of other genes with esophageal cancer.

Materials and methods

41 genes (Supplementary Table 1) are found to be related to esophageal cancer by DisGeNet (18). We constructed a gene interaction network by String (19), which shows as Figure 1.

We implemented Graph Convolutional Network (GCN) to extract feature of each gene from gene interaction network. A graph network requires the input of the node feature matrix and

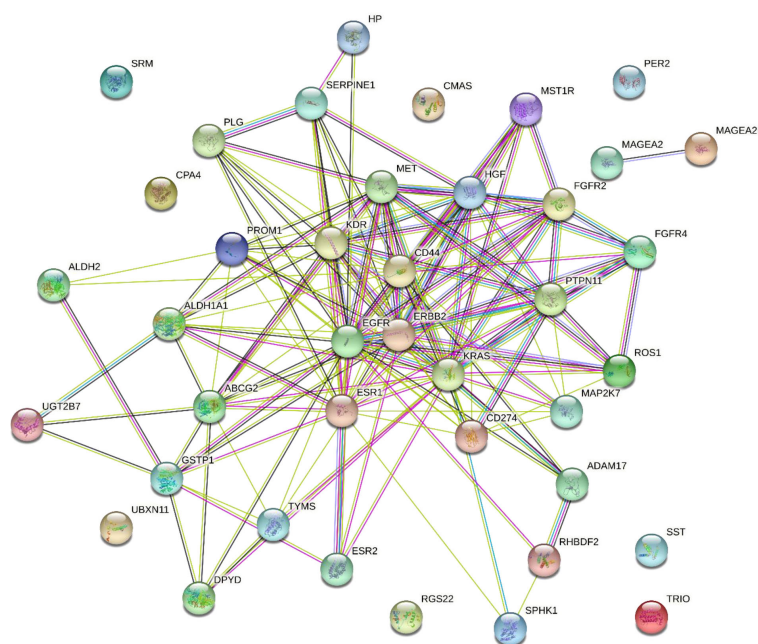


FIGURE 1
Gene interaction network of 41 esophageal cancer-related genes.

the adjacency matrix, so that the aggregation operation of the nodes can be performed. The input of GCN is a feature matrix A and its dimension is $N \times F^0$, where N is the number of nodes in the graph network and F^0 is the number of input features per node. The adjacency matrix A matrix representation of a graph structure whose dimension is $N \times N$.

The aggregated representation of a node does not contain its own features, the representation is the feature aggregation of neighboring nodes, so only nodes with self-loops will contain their own features in this aggregation. Therefore,

$$\bar{A} = A + I \quad (1)$$

The propagation rules for this network are as follows:

$$f(H^i, \bar{A}) = \sigma(\bar{A}H^iW^i) \quad (2)$$

where H^i is the weight matrix of the i -th layer, $\sigma()$ is a nonlinear activation function, and the weights are shared among different nodes.

A node with a large degree will have a large value in its feature representation, and a node with a small degree will have a small value, which may cause the gradient to disappear or explode, and also affect the stochastic gradient descent algorithm. Therefore, the feature table needs to be normalized, the matrix A is multiplied by the inverse of the matrix D , and it is transformed.

$$f(X, \bar{A}) = D\bar{A}X \quad (3)$$

We implemented Logistic Matrix Factorization (LogisticMF) to identify esophageal cancer-related genes. Unlike most previous matrix factorization models, LogisticMF does not use RMSE as its loss function, but a probabilistic approach. Specifically, given an observation matrix R , it is approximated by the inner product of two low-dimensional matrices X_{mf} and X_{mf} , where f is the dimension of the latent factor. Definition l_{ui} means that esophageal cancer (u) is related to gene i , and its conditional probability is given as follows:

$$p(l_{ui} | x_u, y_i, \beta_u, \beta_i) = \frac{\exp(x_u y_i^T + \beta_u + \beta_i)}{1 + \exp(x_u y_i^T + \beta_u + \beta_i)} \quad (4)$$

where β_u, β_i represent the bias.

Similar to Collaborative Filtering for Implicit Feedback Datasets, LogisticMF also uses confidence to represent its frequency. The confidence mapping function can take:

$$c = 1 + a \log(1 + r_{ui}/\epsilon) \quad (5)$$

where a is a smoothing parameter that adjusts the weight of positive and negative examples.

Combining the above formula, we can get:

$$\begin{aligned} L(R|X, Y, \beta) \\ = \prod_{u,i} p(l_{ui} | x_u, y_i, \beta_u, \beta_i)^{ar_{ui}} (1 - p(l_{ui} | x_u, y_i, \beta_u, \beta_i)) \end{aligned} \quad (6)$$

Furthermore, the underlying association matrix of esophageal cancer and genes is assumed to follow a Gaussian distribution:

$$\begin{aligned} p(X|\sigma^2) &= \prod_u N(x_u | 0, \sigma_u^2 I) \\ p(Y|\sigma^2) &= \prod_u N(y_i | 0, \sigma_i^2 I) \end{aligned} \quad (7)$$

Then its posterior probability is:

$$\begin{aligned} \log p(X, Y, \beta | R) \\ = \sum ar_{ui} (x_u y_i^T + \beta_u + \beta_i) \\ - (1 + ar_{ui}) \log(1 + \exp(x_u y_i^T + \beta_u + \beta_i)) \\ - \frac{\lambda}{2} |x_u|^2 - \frac{\lambda}{2} |y_i|^2 \end{aligned} \quad (8)$$

we should maximize the posterior probability, so use alternating gradient descent to optimize:

$$\frac{\partial}{\partial x_u} = \sum ar_{ui} y_i - \frac{y_i(1 + ar_{ui}) \exp(x_u y_i^T + \beta_u + \beta_i)}{1 + \exp(x_u y_i^T + \beta_u + \beta_i)} - \lambda x_u \quad (9)$$

$$\frac{\partial}{\partial \beta_u} = \sum ar_{ui} - \frac{(1 + ar_{ui}) \exp(x_u y_i^T + \beta_u + \beta_i)}{1 + \exp(x_u y_i^T + \beta_u + \beta_i)} - \lambda x_u \quad (10)$$

Results

Experiment workflow

We have obtained 41 genes which are related to esophageal cancer and we also need negative samples to build our model. Therefore, we randomly selected 200 genes as the negative samples. We used 10-cross validation to verify the accuracy of our model. We divided our samples into 10 groups. We used nine groups of datasets to build the model and the rest one to test the model.

Performance of GCNLMF

We apply two evaluation metrics, AUC and AUPR, to evaluate our method. The experimental results of ten tests are shown in Figures 2, 3.

The average of AUC is 0.927 and the standard deviation is 0.035. The average of AUPR is 0.86 and the standard deviation is 0.021. Through the cross-validation experiment, we can see that the prediction accuracy of GCNLMF is very high and stable.

Comparison experiments

To highlight the superiority of GCNLMF, we compare it with five methods. The AUC for each method is the average value obtained by 10-fold cross-validation. The five methods include random forest (RF), gradient boosting decision tree (GBDT), GCN, LMF and Support Vector Machine (SVM). In RF, the number of decision trees was set as 100.

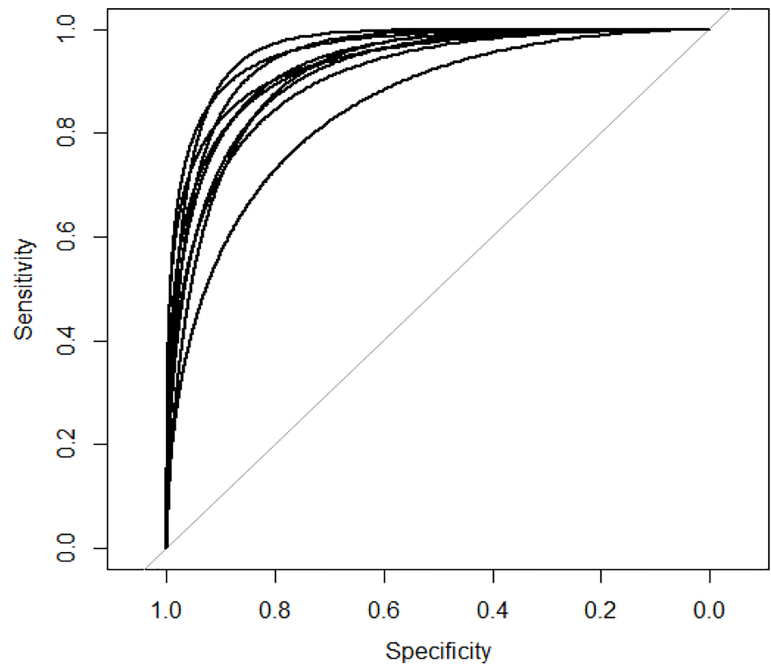


FIGURE 2
AUC curves of GCNLMF in 10-cross validation.

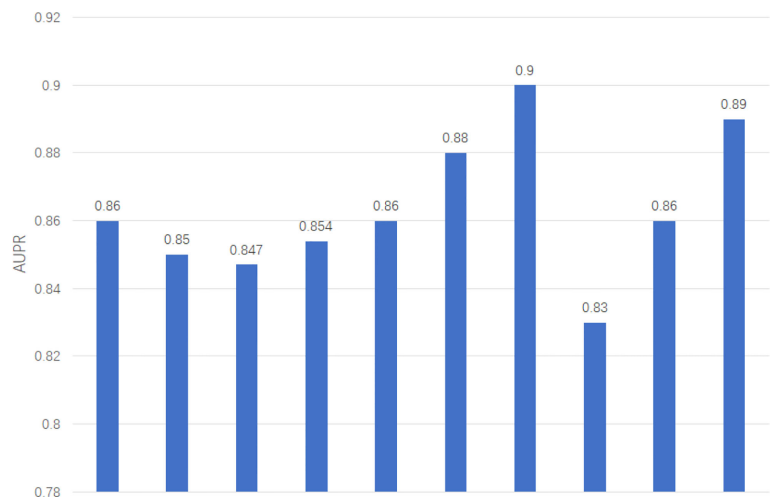


FIGURE 3
AUPR of GCNLMF in 10-cross validation.

The results are shown in Figure 4. The experiment showed that GCNLMF had the highest performance among all methods according to AUC and AUPR scores. Compared with GBDT, RF, GCN, LMF SVM, the AUC of GCNLMF increased by 14%, 9.6%, 1.4%, 3% and 7.4%, respectively. The AUPR scores increased by 15%, 9.7%, 1.4%, 2.3% and 9.6%, respectively.

Case study

After building GCNLMF model, we used it to predict novel esophageal cancer-related genes. IL-10 is not reported to be related to esophageal cancer in the public database and GCNLMF predicted it as an esophageal cancer-related gene. Yang et al. (20) found that the

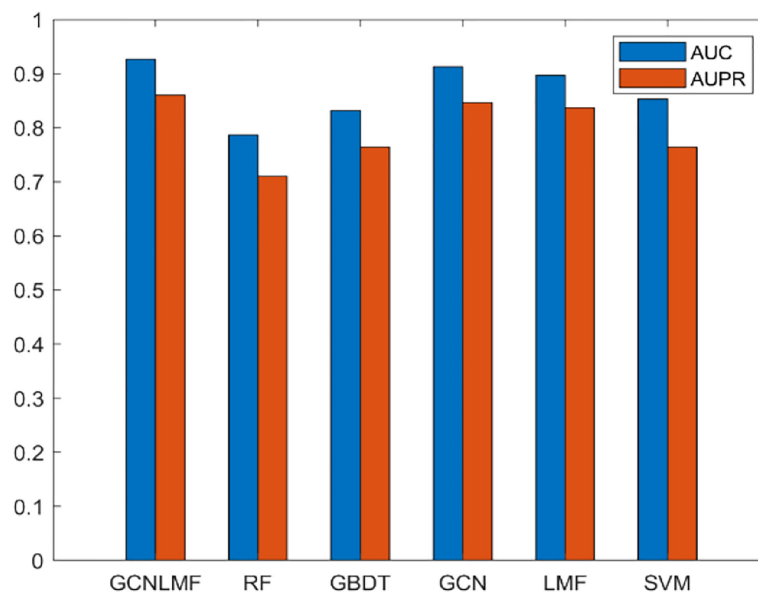


FIGURE 4
Results of GCNLMF compared to the other five methods.

-1082g/a rs1800896 genetic variation can be used as a candidate biomarker to predict the susceptibility of esophageal cancer by comparing the IL10 genotypes of 246 pathologically confirmed esophageal cancer patients and 492 healthy control subjects. Sun et al. (21) found that ETV5 was upregulated in Esophageal squamous cell carcinoma and was associated with tumor staging and prognosis. Knockdown of ETV5 or its downstream genes SKA1 and TRPV2 significantly suppress Esophageal squamous cell carcinoma cells migration and invasion, respectively. Kuerbanjiang et al. (22) detected the expression of BRAF in esophageal cancer samples by tissue microarray, and the results showed that BRAF plays an important role in the proliferation, invasion and metastasis of esophageal cancer, and overexpression of BRAF leads to shortened overall survival.

Conclusions

The incidence of esophageal cancer has obvious familial aggregation phenomenon, which is related to the susceptibility of the population and environmental conditions. In areas with high incidence of esophageal cancer, it is not uncommon for families to have esophageal cancer patients for 3 or more consecutive generations. Therefore, it is important to discover the genetic factors of esophageal cancer.

Most previous studies have compared esophageal cancer patients with healthy people by means of DNA sequencing and

RNA sequencing, so as to find gene mutations and abnormal gene expression associated with esophageal cancer. However, the time and money costs of such methods are high. At the same time, the sample size is limited and there are differences between samples. As a result, the numbers of genes associated with esophageal cancer were both small and inaccurate. Our previous studies have also confirmed the critical role of key genes and signaling pathways in the progression of esophageal cancer (23–25). This paper proposes a method GCNLMF for large-scale identification of esophageal cancer-related genes, which can effectively identify the characteristics of esophageal cancer-related genes. Through the correlation and characteristics between genes, more genes related to esophageal cancer can be predicted.

In order to verify the accuracy of GCNLMF, we used 10-cross validation. The AUC of GCNLMF was 0.927 and the auapr was 0.86. And in ten experiments, the standard deviation of these two indicators is very small, which shows that the method is robust. We also compare GCNLMF with five other commonly used methods, and we find that the accuracy of GCNLMF is significantly higher than other methods. In order to verify the accuracy of the esophageal cancer-related genes predicted by GCNLMF, we selected the top 3 genes in the prediction results to conduct a case study. Although the association of these three genes with esophageal cancer has not been reported in the database, studies by other researchers have shown that these three genes are significantly associated with esophageal cancer, which illustrates the accuracy of the prediction results of GCNLMF.

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

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

Author contributions

XH and FM participated in its design. W-SL, Z-GQ, LZ, H-ML, Y-WH and Y-MZ interpreted and analyzed the data. XH, W-SL and G-sZ wrote the paper. All authors contributed to the article and approved the submitted version.

Funding

Financial support comes from the National Natural Science Foundation of China (No. 81960438, 82103067), Natural Science

Foundation of Jiangxi Province (No. 20212BAB206078) and Jiangxi Provincial Education Foundation (No. GJJ190790, GJJ190792).

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/fonc.2022.982641/full#supplementary-material>

References

- Bhat AA, Nisar S, Maacha S, Carneiro-Lobo TC, Akhtar S, Siveen KS, et al. Cytokine-chemokine network driven metastasis in esophageal cancer; promising avenue for targeted therapy. *J Mol Canc* (2021) 20:1–20.
- Watanabe A, Oshikiri T, Sawada R, Harada H, Urakawa N, Goto H, et al. Actual sarcopenia reflects poor prognosis in patients with esophageal cancer. *J Ann Surg Oncol* (2022) 29:3670–81.
- Visaggi P, Barberio B, Ghisa M, Ribolsi M, Savarino V, Fassan M, et al. Modern diagnosis of early esophageal cancer: from blood biomarkers to advanced endoscopy and artificial intelligence. *J Cancer (Basel)* (2021) 13:3162.
- Taylor PR, Abnet CC, Dawsey SM. Squamous dysplasia—the precursor lesion for esophageal squamous cell carcinoma. *J Cancer Epidemiol Biomarkers Prev* (2013) 22:540–52.
- Wang G, Abnet C, Shen Q, Lewin K, Sun X, Roth M, et al. Histological precursors of esophageal squamous cell carcinoma: results from a 13 year prospective follow up study in a high risk population. *J Gut* (2005) 54:187–92.
- Gao Y-B, Chen Z-L, Li J-G, Hu X-D, Shi X-J, Sun Z-M, et al. Genetic landscape of esophageal squamous cell carcinoma. *Nat Genet* (2014) 46:1097–102.
- Chen X-X, Zhong Q, Liu Y, Yan S-M, Chen Z-H, Jin S-Z, et al. Genomic comparison of esophageal squamous cell carcinoma and its precursor lesions by multi-region whole-exome sequencing. *Nat Commun* (2017) 8:1–12.
- Liu X, Zhang M, Ying S, Zhang C, Lin R, Zheng J, et al. Genetic alterations in esophageal tissues from squamous dysplasia to carcinoma. *Gastroenterology* (2017) 153:166–77.
- Martincorena I, Fowler JC, Wabik A, Lawson AR, Abascal F, Hall MW, et al. Somatic mutant clones colonize the human esophagus with age. *Science* (2018) 362:911–7.
- Kashima H, Noma K, Ohara T, Kato T, Katsura Y, Komoto S, et al. Cancer-associated fibroblasts (CAFs) promote the lymph node metastasis of esophageal squamous cell carcinoma. *Int J Canc* (2019) 144:828–40.
- Wang Y, Lyu Z, Qin Y, Wang X, Sun L, Zhang Y, et al. FOXO1 promotes tumor progression by increased M2 macrophage infiltration in esophageal squamous cell carcinoma. *J Theranost* (2020) 10:11535.
- Yang H, Zhang Q, Xu M, Wang L, Chen X, Feng Y, et al. CCL2-CCR2 axis recruits tumor associated macrophages to induce immune evasion through PD-1 signaling in esophageal carcinogenesis. *Mol Cancer* (2020) 19:1–14.
- Liu S, Wu J, Xia Q, Liu H, Li W, Xia X, et al. Finding new cancer epigenetic and genetic biomarkers from cell-free DNA by combining SALP-seq and machine learning. *Comput Struct Biotechnol J* (2020) 18:1891–903.
- Wang Y, Wang H, Li S, Wang L. Survival risk prediction of esophageal cancer based on the kohonen network clustering algorithm and kernel extreme learning machine. *Mathematics* (2022) 10:1367.
- Li M-X, Sun X-M, Cheng W-G, Ruan H-J, Liu K, Chen P, et al. Using a machine learning approach to identify key prognostic molecules for esophageal squamous cell carcinoma. *BMC Canc* (2021) 21:1–11.
- Cheng N, Chen C, Li C, Huang J. Inferring cell-type-specific genes of lung cancer based on deep learning. *J Curr Gene Ther*. doi: 10.2174/1566523222666220324110914
- Zhao T, Liu J, Zeng X, Wang W, Li S, Zang T, et al. Prediction and collection of protein–metabolite interactions. *Brief Bioinform* (2021) 22:bbab014.
- Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *J Nucleic Acid Res* (2020) 48:D845–55.

19. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *J Nucleic Acid Res* (2021) 49:D605–12.
20. Yang Y, Fa X, Pathology E. Role of IL-10 gene polymorphisms on the susceptibility for esophageal cancer and its association with environmental factors. *Int J Clin Exp Pathol* (2015) 8:9580.
21. Sun M-C, Fang K, Li Z-X, Chu Y, Xu A-P, Zhao Z-Y, et al. ETV5 overexpression promotes progression of esophageal squamous cell carcinoma by upregulating SKA1 and TRPV2. *Int J Med Sci* (2022) 19:1072–81.
22. Kuerbanjiang A, Maimaituerxun M, Zhang Y, Li Y, Cui G, Abuduhabaier A, et al. V-Raf murine sarcoma viral oncogene homolog B1 (BRAF) as a prognostic biomarker of poor outcomes in esophageal cancer patients. *J BMC Gastroenterol* (2021) 21:1–10.
23. He X, Meng F, Qin L, Liu Z, Zhu X, Yu Z, et al. KLK11 suppresses cellular proliferation via inhibition of wnt/ β -catenin signaling pathway in esophageal squamous cell carcinoma. *Am J Cancer Res* (2019) 9:2264–77.
24. Meng F, Li R, Ma L, Liu L, Lai X, Yang D, et al. Porphyromonas gingivalis promotes the motility of esophageal squamous cell carcinoma by activating NF- κ B signaling pathway. *Microbes Infect* (2019) 21:296–304.
25. He X, Meng F, Yu ZJ, Zhu XJ, Qin LY, Wu XR, et al. PLCD1 suppressed cellular proliferation, invasion, and migration via inhibition of wnt/ β -catenin signaling pathway in esophageal squamous cell carcinoma. *Dig Dis Sci* (2021) 66:442–51.



OPEN ACCESS

EDITED BY

Liang Cheng,
Harbin Medical University, China

REVIEWED BY

Kebo LV,
Ocean University of China, China
Taigang Liu,
Shanghai Ocean University, China

*CORRESPONDENCE

Yi Shi
Shiyi.veals@qq.com
Xiaoli Shi
shixl@geneis.cn

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 30 June 2022

ACCEPTED 03 August 2022

PUBLISHED 19 August 2022

CITATION

Li L, Qiu W, Lin L, Liu J, Shi X and Shi Y
(2022) Predicting recurrence and
metastasis risk of endometrial
carcinoma *via* prognostic signatures
identified from multi-omics data.
Front. Oncol. 12:982452.
doi: 10.3389/fonc.2022.982452

COPYRIGHT

© 2022 Li, Qiu, Lin, Liu, Shi and Shi. This
is an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction
in other forums is permitted, provided
the original author(s) and the
copyright owner(s) are credited and
that the original publication in this
journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is
permitted which does not comply with
these terms.

Predicting recurrence and metastasis risk of endometrial carcinoma *via* prognostic signatures identified from multi-omics data

Ling Li¹, Wenjing Qiu², Liang Lin¹, Jinyang Liu^{2,3},
Xiaoli Shi^{2,3*} and Yi Shi^{4*}

¹Department of Gynecological Oncology Surgery, Fujian Cancer Hospital, Fujian Medical University Cancer Hospital, Fuzhou, China, ²Science System Department, Geneis Beijing Co., Ltd., Beijing, China, ³Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, ⁴Department of Molecular Pathology, Fujian Cancer Hospital, Fujian Medical University Cancer Hospital, Fuzhou, China

Objectives: Endometrial carcinoma (EC) is one of the three major gynecological malignancies, in which 15% - 20% patients will have recurrence and metastasis. Though there are many studies on the prognosis on this cancer, the performances of existing models evaluating the risk of its recurrence and metastasis are yet to be improved. In addition, a comprehensive multi-omics analyses on the prognostic signatures of EC are on demand. In this study, we aimed to construct a relatively stable and reliable model for predicting recurrence and metastasis of EC. This will help determine the risk level of patients and choose appropriate adjuvant therapy, thereby avoiding improper treatment, and improving the prognosis of patients.

Methods: The mRNA, microRNA (miRNA), long non-coding RNA (lncRNA), copy number variation (CNV) data and clinical information of patients with EC were downloaded from The Cancer Genome Atlas (TCGA). Differential expression analyses were performed between the recurrence or metastasis group and the non-recurrence/metastasis group. Then, we screened potential prognostic markers from the four kinds of omics data respectively and established prediction models using three classifiers.

Results: We achieved differential expressed mRNAs, lncRNAs, miRNAs and CNVs between the two groups. According to feature selection scores by the random forest algorithm, 275 CNV features, 50 lncRNA features, 150 miRNA features and 150 mRNA features were selected, respectively. And the prediction model constructed by the features of lncRNA data using random forest method showed the best performance, with an area under the curve of 0.763, and an accuracy of 0.819 under 10-fold cross-validation.

Conclusion: We developed a computational model using omics information, which is able to predicting recurrence and metastasis risk of EC accurately.

KEYWORDS

endometrial carcinoma, recurrence and metastasis, lncRNA, CNV, mRNA, miRNA, prediction model

Introduction

Endometrial carcinoma (EC) is a kind of epithelial malignant tumor occurring in the endometrium and is one of the three major gynecological malignancies (1–3). In North America and Europe, it is the fourth leading cancer following breast cancer, lung cancer, colorectal tumor in terms of incidence (4). In China, the incidence of the disease is also increasing year by year and is second only to cervical cancer (5). Obesity, hormonal and metabolic disorders are particularly closely related to the occurrence of EC (6). Its clinical treatment is mainly surgical resection, supplemented by radiotherapy and drug treatment. Although most patients are at the early stage when diagnosed and have a good prognosis, 15% - 20% of patients will have recurrence and metastasis (7–9). The presence of poor prognosis of recurrence or metastasis is the main cause leading to the death of EC patients (10, 11). Therefore, accurate prediction of the recurrence and metastasis of endometrial cancer as early as possible and performed targeted adjuvant therapy are essential to improve the survival rate of EC patients. In fact, it is difficult to identify patients with a high risk of recurrence and metastasis in the early stage. Traditionally, clinicians usually predict the risk of recurrence and metastasis by pathological type, histological grade, depth of myometrial invasion, lymphatic metastasis and extrauterine lesions, and monitor the development of the disease through patients' regular radiologic examination and laboratory examinations (12–15).

Nowadays, with the development of liquid biopsy technology and the popularization of artificial intelligence in the field of medical images, there are many new explorations and novel methods in predicting tumor recurrence and metastasis (16–22). For example, Wu et al. developed a deep convolutional neural network (CNN) model to predict the risk of recurrence

and metastasis from hematoxylin and eosin (H&E) stained sections of lung cancer (23). For estimating the risk of recurrence and metastasis in patients with HER2-positive breast cancer, Yang et al. constructed a novel multimodal fusion model integrating H&E images and clinical characteristics (20), with an area under the curve (AUC) of 0.72 in the independent testing data. Feng et al. identified that detection of somatic mutations of ctDNA could predict recurrence of EC effectively and stably (16). Ye et al. developed a deep convolution network to predict cervical cancer metastasis and recurrence risk (24). Based on the study results of The Cancer Genome Atlas (TCGA), endometrial cancer was classified into four categories according to the mutation spectrum, somatic copy number alterations (SCNAs) and microsatellite instability (MSI): DNA polymerase epsilon (POLE) ultramutated, high microsatellite instability (MSI-H), copy-number low, and copy-number high (25). TCGA molecular typing has initially shown good application prospects in predicting the prognosis of endometrial cancer patients and has been listed in the national comprehensive cancer network (NCCN), which may affect post-surgical adjuvant treatment. However, no applicable prognostic prediction models only based on genomics have been found by retrieving concerned literatures (8, 10).

In this study, all sequencing data and clinical information of patients with EC from TCGA (<http://cancergenome.nih.gov/>) were downloaded and organized to study the association between gene mutation/expression and recurrence or metastasis of EC. Specifically, we first compared the differential expressions of mRNA, long non-coding RNA (lncRNA) and microRNA (miRNA) between patients with recurrence or metastasis and patients without recurrence or metastasis using the DESeq2 package, and then analyzed differences of copy number variations (CNVs) between the two groups by rank-sum test. Furthermore, we analyzed the function of these differential genes, and discussed the molecular mechanism of recurrence and metastasis of EC. After that, characteristic variables were selected by a random forest (RF) algorithm with feature selection and were used to establish prognostic prediction models using three different classifiers. Finally, the RF model based on lncRNA showed the best performance among the twelve models.

Abbreviations: EC, endometrial carcinoma; lncRNAs, long non-coding RNAs; CNV, copy number variation; NCCN, national comprehensive cancer network; TCGA, The Cancer Genome Atlas; AUC, area under the curve; ROC, receiver operating characteristic; RF, random forest; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

Materials and methods

Study participants

TCGA is a great cancer genome project which has produced genomic, epigenomic, transcriptomic and proteomic data of more than 20,000 cancer patients covering multiple cancer types. These data can help researchers have a more comprehensive understanding of cancer and improve the level of cancer screening, diagnosis and treatment. Clinical information of 548 patients with EC was downloaded from the TCGA data portal, including 204 patients without recurrence or metastasis, 43 patients with recurrence or metastasis and 301 patients without information of recurrence and metastasis. In addition, mRNA sequencing data of 543 EC patients, miRNA data of 538 EC patients, lncRNA data of 537 EC patients and CNV data of 534 EC patients were downloaded. Then, we matched the data according to the patient ID, and selected patients with complete omics data and prognostic information into the study.

Difference analysis

The expression data of mRNAs, lncRNAs and miRNAs were displayed as reads per million (RPM) and the expression levels were normalized by DESeq2 (26) package of R language for difference analysis. Then the differentially expressed mRNAs, lncRNAs and miRNAs were calculated by DESeq2 with $\text{Padj} < 0.05$ and the absolute $\log_2\text{FC} > 1$ as the cutoff value, respectively. The CNVs of two groups (recurrence and metastasis group and non-recurrence/metastasis group) were analyzed by SPSS statistical software and significant differences were screened by rank-sum test with $P < 0.005$ as the threshold.

To explore the potential biological functions of these differential genes and the signal pathways they may participate in, we performed Gene Ontology (GO) (27, 28) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses by employing clusterProfiler R package (29) with $p\text{-value} < 0.05$ and $q\text{-value} > 1$ as the threshold.

Feature selection for modeling

The patients in the recurrence and metastasis group and in the group without recurrence or metastasis were divided into the training set and the test set with the ratio of 7:3, respectively. After the division, the patients were fixed in the training set or the test set, that is, in different omics analyses, the same patient was always in the training set or test set.

For each omics data of the training set, a feature selection algorithm based on RF was applied to screen important features

(30–33). Specifically, we screened the characteristic variables with scores according to Gini index. Then the features were grouped in steps of 25 and performed 10 fold cross-validation and scored to confirm the final number of features.

Model construction and comparison

Based on the selected characteristic parameters, RF, logistic regression and support vector machine classifiers were chosen for model construction to select the best model to predict the recurrence or metastasis of patients with endometrial cancer. Specifically, omics data in the training set were grid searched in each classifier to select the best super parameter, and then the final super parameter was determined through 10 fold cross-validation. Finally, we obtained 12 prediction models and compared the prediction performance mainly using the AUC of receiver operating characteristic (ROC) curves, precision and accuracy.

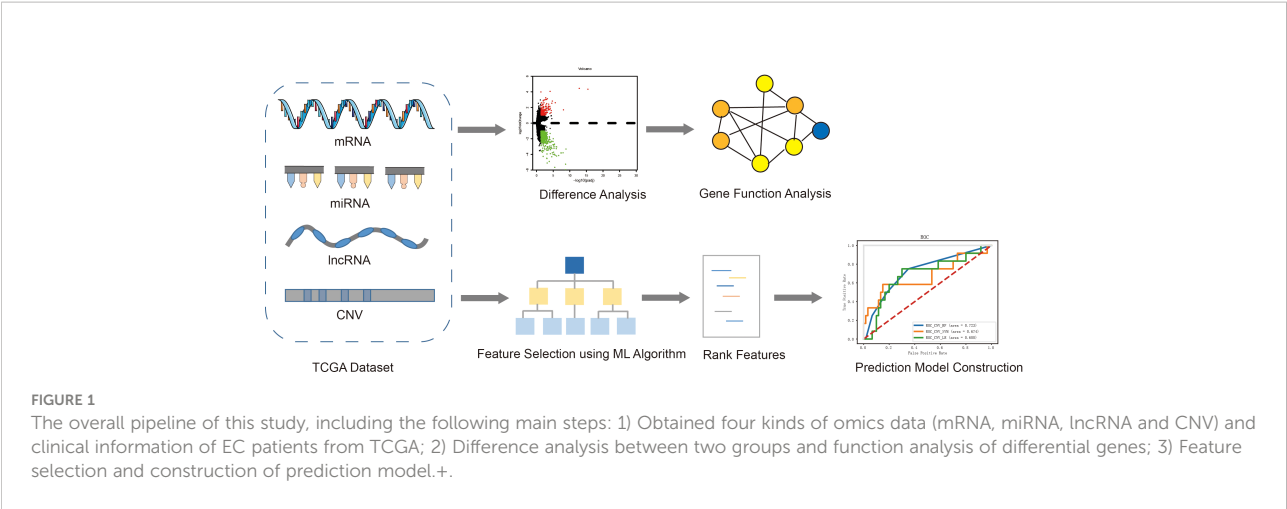
Results

A brief study design of exploring molecular mechanism and establishing prediction model

The overall process of exploring the molecular mechanism of recurrence and metastasis, and establishing risk prediction models using a machine learning algorithm was described in Figure 1. Firstly, four kinds of omics data and clinical information of EC patients were downloaded from TCGA, and then the patients were divided into two groups according to the prognosis status. Secondly, differential expression analysis was performed between the recurrent and metastatic group and non-recurrent metastatic group. Furthermore, we analyzed the function of differential genes using GO and KEGG. At the same time, characteristic variables were selected by a RF algorithm with feature selection and were used to establish prognostic prediction models using three different classifiers.

Clinicopathological features of patients with EC

After matching data according to the patient ID, 238 EC patients with both prognostic information and four kinds of omics data were obtained. Of these patients, 39 patients (16.39%) had cancer recurrence or metastasis whereas 199 patients (83.61%) had no recurrence or metastasis. Clinical and pathological information of these two groups of patients in this study was shown in Table 1. Because some information is absent, such as lymph node, progesterone receptor and estrogen



receptor status, only some factors that may be related to the prognosis of patients (7, 34) were selected for statistical analysis. As can be seen, clinical stage was significantly associated with recurrence or metastasis in this set of data, with a *P*-value of 0.000656. Whereas, there were no significant differences between the recurrent or metastatic group and non-recurrent or metastatic group in ages (median), body mass index (BMI) and pathological type.

Results of differential expression analysis

Among the expression data of 17,958 mRNAs, 592 mRNA genes were expressed significantly different between the recurrent and metastatic group and non-recurrent or metastatic group, with 169 up-regulated genes and 423 down-regulated genes in the recurrent or metastatic group compared to the non-recurrent or metastatic group (Figure 2A). And 3,352 differentially lncRNAs were achieved, in which 87 lncRNAs were significantly different, with 51 down-regulated and 36 up-regulated (Figure 2B). In addition, there were 687 differentially expressed miRNAs, in which 39 miRNAs were significantly

different, with 23 down-regulated and 16 up-regulated (Figure 2C). Heatmaps of the top 50 differentially expressed mRNA, lncRNA and miRNA were shown in Supplementary Figures 1–3. As CNVs had been reported to affect the recurrence of EC (34), it was selected separately for analysis. Finally, 939 significantly different CNVs were got after analyzing by SPSS statistical software with *P* < 0.005 as the threshold.

Then GO and KEGG enrichment analyses were performed to explore the function and involved signal pathways for further investigating the prognostic value and molecular mechanisms. For significantly differentially expressed mRNAs, the top twelve molecular functions with the highest proportion of genes were displayed in Figure 3A, and first twelve enriched signaling pathways were shown in Figure 3B. We found that the molecular functions of these differentially expressed mRNAs mainly enriched in signaling receptor activator activity, receptor ligand activity, growth factor activity, sodium ion transmembrane transporter activity, peptidase inhibitor activity, serine-type endopeptidase inhibitor activity, and so on. And the results of KEGG pathway analysis indicated that the recurrence or metastasis of EC may be correlated to the regulation of cytokine-cytokine receptor interaction, Calcium

TABLE 1 Summary of clinical information of patients with EC.

Clinicopathologic variable	Category	Recurrence or metastasis	Without recurrence or metastasis	<i>P</i> -value
Age (years)	Median	62	61	1.000
Clinical stage	I	18	124	0.000656
	II	5	28	
	III	11	45	
	IV	5	2	
Pathological type	Endometrioid carcinoma	23	142	0.179
	Non-endometrioid carcinoma	16	57	
BMI	≤28	7	62	0.845
	>28	32	137	

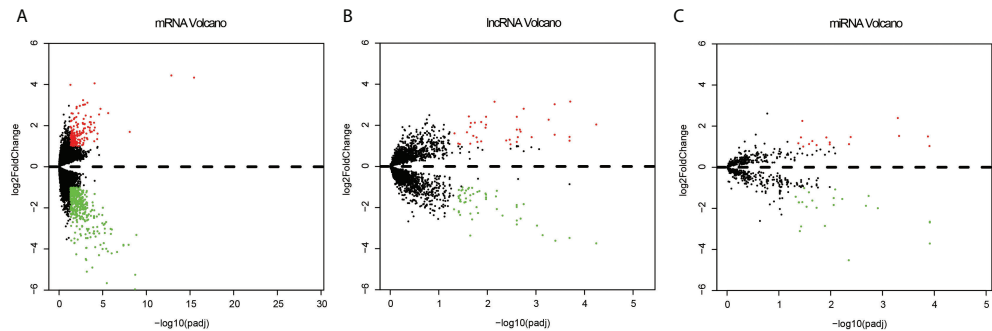


FIGURE 2
Volcano plots of the differentially expressed mRNAs (A), lncRNAs (B) and miRNA (C) between the recurrent or metastatic group and non-recurrent or metastatic group. Red represents up expression and green represents down expression. Black indicates the expression with both the absolute log2FC > 1 and Padj < 0.05. The X axis shows an adjusted P value and the Y axis shows a log2FC.

signaling pathway, Ras signaling pathway, viral protein interaction with cytokine and cytokine receptor. GO enrichment analysis results and KEGG pathways of the different CNVs were displayed in Figures 3C, D, respectively. From the perspective of molecular function, these mainly focus

on glutathione binding, oligopeptide binding, hydrolase activity, hydrolyzing O-glycosyl compounds, hydrolase activity, acting on glycosyl bonds, anion channel activity, transferase activity, transferring alkyl or aryl (other than methyl) groups. The genes with significant CNV differences are mainly involved in the

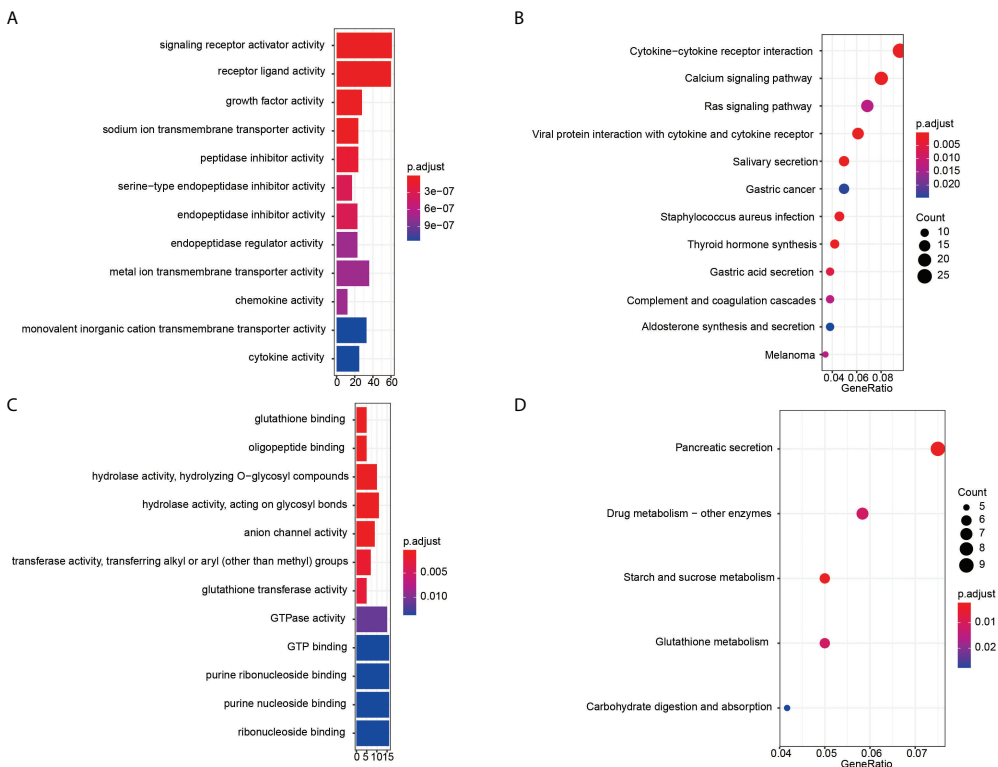


FIGURE 3
Enrichment analysis results. GO enrichment analysis results of significantly differentially expressed mRNAs (A) and CNVs (C). The x-axis is gene counts, the y-axis is GO terms of molecular function. KEGG pathways of the differentially expressed mRNAs (B) and CNVs (D). The x-axis is the ratio of genes in the corresponding pathway, and the y-axis is the name of the pathway.

following five pathways: pancreatic secretion, drug metabolism - other enzymes, starch and sucrose metabolism, glutathione metabolism and carbohydrate digestion and absorption.

Modeling using lncRNA showed the best prediction performance

Among the data of EC downloaded from TCGA, there are 17958 mRNA expression information, 7315 lncRNA expression information, 1881 miRNA expression information and 16383 copy number variation information. Variable selection for these biological data was performed using the RF to determine variable importance measures. The scoring of different number of features screened by RF is shown in Figure 4. The features with the highest 10-CV score were selected for model construction. Specifically, 275 features were chosen from CNV data because the score of 275 features was 0.826, significantly higher than other feature combinations (Figure 4A). And 50 lncRNA features were selected as the score was 0.851, which was higher than others (Figure 4B). For the other two kinds of genomic data, 150 features of miRNA and mRNA were selected, with the highest score of 0.862 and 0.856, respectively (Figures 4C, D).

For each kind of omics data, three classifiers (RF, LR and SVM) were used to construct the prediction model. The ROC curves of three models based on lncRNA data were displayed in Figure 5A, because the model based on the characteristics of lncRNA data represented the best prediction performance. And accuracy, precision, recall and F1-score of the three models were shown in Figure 5B. The RF model constructed by the features of lncRNA data was able to predict recurrence or metastasis of EC with an AUC of 0.763, an accuracy of 0.819. The ROC curves of other models using omics variables were shown in Supplementary Figures 4. The ROC curves of models with the best prediction performance constructed by four omics data (lncRNA, mRNA, miRNA and CNV) were represented in Figure 5C, and the accuracy, precision, recall and F1-score of the prediction models were revealed in Figure 5D.

Discussion

The Oncotype Dx (21 genes) and MammaPrint (70 genes) are two products for predicting recurrence and metastasis of breast cancer which have been internationally recognized (35). However, for patients with EC, there is no effective model based on molecular variations to evaluate the risk of recurrence and

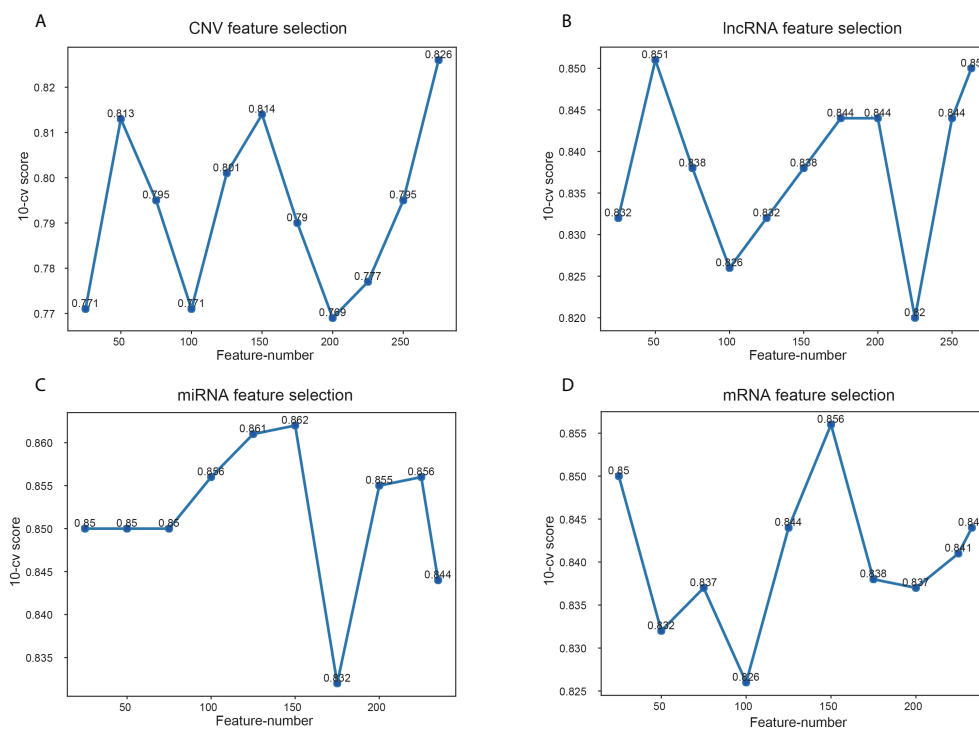


FIGURE 4
Scores of different feature number selection based on omics data. (A) Feature selection of CNV, (B) feature selection of lncRNA, (C) feature selection of miRNA, (D) feature selection of mRNA. The x-axis is feature numbers, the y-axis is the 10-CV score.

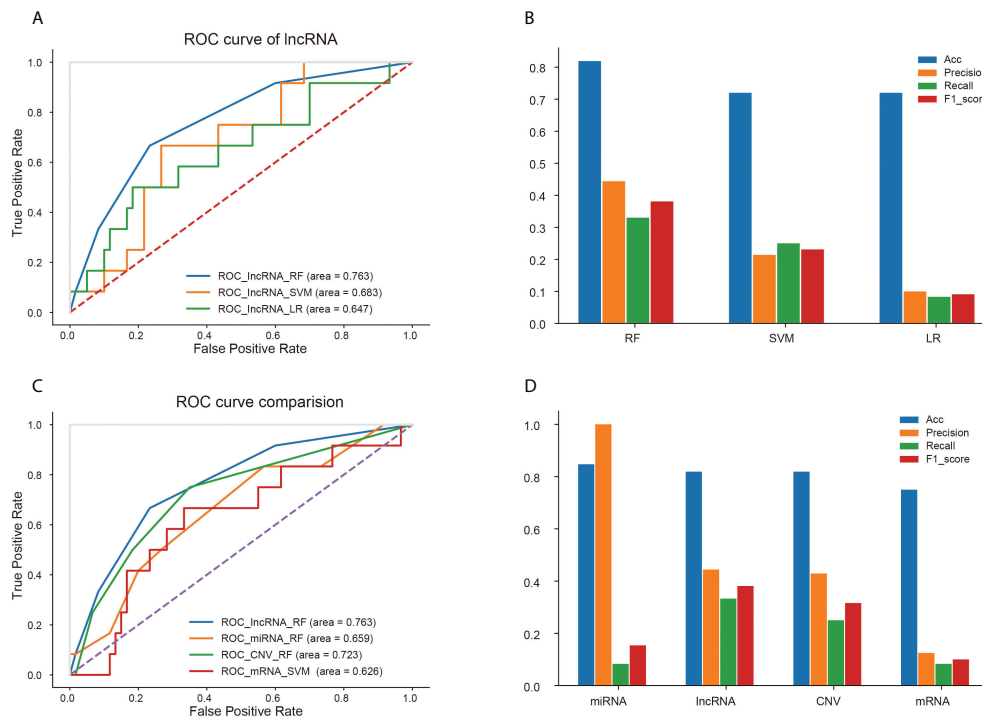


FIGURE 5

Prediction performance of different models based on four kinds of omics data. (A) ROC curves of three models based on lncRNA data. (B) Accuracy, precision, recall and F1-score of three models for lncRNA signatures. Comparison of ROC curves (C) and four properties (D) of optimal models based on four kinds of omics data.

metastasis. EC is a malignant tumor, which often occurs in perimenopausal and postmenopausal women. It usually has a good prognosis if diagnosed early and treated appropriately. So, patients will benefit greatly when a product like Oncotype DX appears, that can help clinicians assess the recurrence risk of patients and adopt adjuvant treatment strategies according to different risk stratification. Previous studies have established prediction models based on clinical characteristics (14, 15) and combined clinical characteristics with molecular data (34). AUC value of the model using clinical features only was about 0.7, whereas M. D. Miller et al. used different kinds of molecular data, up to 5 categories, it may be difficult and expensive to apply clinically.

Here, starting from the data of TCGA, we analyzed the differences of mRNA expression, miRNA expression, lncRNA expression as well as CNVs between patients with recurrence and metastasis and non-recurrence or metastasis, and further analyzed their molecular biological functions and involved signal pathways, trying to explore the molecular biological mechanism of these differences and recurrence and metastasis. Although these molecules are related to recurrence and metastasis, it does not mean that these differential molecules can accurately and reliably predict recurrence and metastasis. To build the

prediction model, it is still necessary to select the most appropriate feature combination by statistical methods (17, 34). Therefore, we used the feature selection algorithm of RF to filter features and selected different classifiers to establish models. Finally, the model using lncRNA data showed the best performance, with an AUC of 0.763, an accuracy of 0.819.

There are still several limitations in this study. Firstly, the sample size was limited and the survival time of tracking was not long enough, which may lead to inaccurate results. However, we have downloaded all samples information from TCGA, a relatively large-scale cancer genome database. Looking for more samples from other open databases or hospitals may be a solution. Secondly, the research on molecular mechanism was not deep enough (36–39). Maybe we should explore and discuss the mechanism of recurrence and metastasis of endometrial cancer in another study (40–42). Thirdly, a more advanced computational model can further improve the prediction accuracy as used elsewhere (43–46). Finally, this study lacked an independent validation set and did not develop a clinically scoring system and thresholds to discriminate risk of recurrence and metastasis. At present, we have not collected enough clinical samples for verification. We will continue to collect data to improve this study.

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.

Ethics statement

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

Author contributions

YS and XS designed the project. LLi, WQ, LLin, and JL collected and analyzed the data of patients with endometrial cancer. LLi and XS searched literatures and wrote the manuscript. All authors have approved the final version of the manuscript.

Conflict of interest

WQ, JL, and XS were employed by the company Geneis Beijing Co., Ltd.

References

- Amant F, Moerman P, Neven P, Timmerman D, Van Limbergen E, Vergote I. Endometrial cancer. *Lancet* (2005) 366(9484):491–505. doi: 10.1016/S0140-6736(05)67063-8
- Liu J, Zhou S, Li S, Jiang Y, Wan Y, Ma X, et al. Eleven genes associated with progression and prognosis of endometrial cancer (EC) identified by comprehensive bioinformatics analysis. *Cancer Cell Int* (2019) 19:136. doi: 10.1186/s12935-019-0859-1
- Bascuas T, Zedira H, Kropp M, Harmening N, Asrih M, Prat-Souteyrand C, et al. Human retinal pigment epithelial cells overexpressing the neuroprotective proteins PEDF and GM-CSF to treat degeneration of the neural retina. *Curr Gene Ther* (2021) 22(2):168–83. doi: 10.2174/1566523221666210707123809
- Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* (2020) 70(1):7–30. doi: 10.3322/caac.21590
- Jiang X, Tang H, Chen T. Epidemiology of gynecologic cancers in China. *J gynecologic Oncol* (2018) 29(1):e7. doi: 10.3802/jgo.2018.29.e7
- Felix AS, Yang HP, Bell DW, Sherman ME. Epidemiology of endometrial carcinoma: Etiologic importance of hormonal and metabolic influences. *Adv Exp Med Biol* (2017) 943:3–46. doi: 10.1007/978-3-319-43139-0_1
- Takahashi A, Matsuura M, Matoda M, Nomura H, Okamoto S, Kanao H, et al. Clinicopathological features of early and late recurrence of endometrial carcinoma after surgical resection. *Int J Gynecologic Cancer* (2017) 27(5):967–72. doi: 10.1097/IGC.0000000000000984
- Del Carmen MG, Boruta DM2nd, Schorge JO. Recurrent endometrial cancer. *Clin Obstet Gynecol* (2011) 54(2):266–77. doi: 10.1097/GRF.0b013e318218c6d1
- Liu H, Qiu C, Wang B, Bing P, Tian G, Zhang X, et al. Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-Origin. *Front Cell Dev Biol* (2021) 9:619330. doi: 10.3389/fcell.2021.619330
- Coll-de la Rubia E, Martinez-Garcia E, Dittmar G, Gil-Moreno A, Cabrera S, Colas E. Prognostic biomarkers in endometrial cancer: A systematic review and meta-analysis. *J Clin Med* (2020) 9(6):1900. doi: 10.3390/jcm9061900
- Zhao T, Cheng L. Mutations in TREM2 change the expression levels of AD-related genes. *Ann Of Neurol* (2020) 88:S98–8. doi: 10.1016/j.jbneur.2022.01.004
- Kang SY, Cheon GJ, Lee M, Kim HS, Kim JW, Park NH, et al. Prediction of recurrence by preoperative intratumoral FDG uptake heterogeneity in endometrioid endometrial cancer. *Transl Oncol* (2017) 10(2):178–83. doi: 10.1016/j.tranon.2017.01.002
- Lee KR, Vacek PM, Belinson JL. Traditional and nontraditional histopathologic predictors of recurrence in uterine endometrioid adenocarcinoma. *Gynecol Oncol* (1994) 54(1):10–8. doi: 10.1006/gyno.1994.1158
- Senol T, Polat M, Ozkaya E, Karateke A. Tumor diameter for prediction of recurrence, disease free and overall survival in endometrial cancer cases. *Asian Pac J Cancer Prev* (2015) 16(17):7463–6. doi: 10.7314/APJCP.2015.16.17.7463

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.2022.982452/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Heatmap of the top 50 differentially expressed mRNA.

SUPPLEMENTARY FIGURE 2

Heatmap of the top 50 differentially expressed lncRNA.

SUPPLEMENTARY FIGURE 3

Heatmap of the top 50 differentially expressed miRNA.

SUPPLEMENTARY FIGURE 4

Prediction performance of three kinds of models based on CNV data (A), miRNA data (B) and mRNA data (C).

15. Versluis MA, de Jong RA, Plat A, Bosse T, Smit VT, Mackay H, et al. Prediction model for regional or distant recurrence in endometrial cancer based on classical pathological and immunological parameters. *Br J Cancer* (2015) 113(5):786–93. doi: 10.1038/bjc.2015.268
16. Feng W, Jia N, Jiao H, Chen J, Chen Y, Zhang Y, et al. Circulating tumor DNA as a prognostic marker in high-risk endometrial cancer. *J Transl Med* (2021) 19(1):51. doi: 10.1186/s12967-021-02722-8
17. Kuhn M. Building predictive models in R using the caret package. *J Stat software* (2008) 28:1–26. doi: 10.18637/jss.v028.i05
18. Muinelo-Romay L, Casas-Arozamena C, Abal M. Liquid biopsy in endometrial cancer: New opportunities for personalized oncology. *Int J Mol Sci* (2018) 19(8):2311. doi: 10.3390/ijms19082311
19. Yang J, Hui Y, Zhang Y, Zhang M, Ji B, Tian G, et al. Application of circulating tumor DNA as a biomarker for non-small cell lung cancer. *Front Oncol* (2021) 11:725938. doi: 10.3389/fonc.2021.725938
20. Yang J, Ju J, Guo L, Ji B, Shi S, Yang Z, et al. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput Struct Biotechnol J* (2022) 20:333–42. doi: 10.1016/j.csbj.2021.12.028
21. He B, Dai C, Lang J, Bing P, Tian G, Wang B, et al. 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* (2020) 1866(11):165916. doi: 10.1016/j.bbdis.2020.165916
22. Cheng L. Omics data and artificial intelligence: New challenges for gene therapy preface. *Curr Gene Ther* (2020) 20(1):1–1. doi: 10.2174/156652322001200604150041
23. Wu Z, Wang L, Li C, Cai Y, Liang Y, Mo X, et al. DeepLRHE: A deep convolutional neural network framework to evaluate the risk of lung cancer recurrence and metastasis from histopathology images. *Front Genet* (2020) 11:768. doi: 10.3389/fgene.2020.00768
24. Ye Z, Zhang Y, Liang Y, Lang J, Zhang X, Zang G, et al. Cervical cancer metastasis and recurrence risk prediction based on deep convolutional neural network. *Curr Bioinf* (2022) 17(2):164–73. doi: 10.2174/1574893616666210708143556
25. Kandath C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, et al. Integrated genomic characterization of endometrial carcinoma. *Nature* (2013) 497(7447):67–73. doi: 10.1038/nature12113
26. 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
27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* (2000) 25(1):25–9. doi: 10.1038/75556
28. The Gene Ontology C. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* (2019) 47(D1):D330–8. doi: 10.1093/nar/gky1055
29. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J Integr Biol* (2012) 16(5):284–7. doi: 10.1089/omi.2011.0118
30. Deviaene M, Testelmans D, Borzee P, Buyse B, Huffel SV, Varon C. Feature selection algorithm based on random forest applied to sleep apnea detection. *Annu Int Conf IEEE Eng Med Biol Soc* (2019) 2019:2580–3. doi: 10.1109/EMBC.2019.8856582
31. Speiser JL. A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. *J BioMed Inform* (2021) 117:103763. doi: 10.1016/j.jbi.2021.103763
32. Hunt C, Montgomery S, Berkenpas JW, Sigafos N, Oakley JC, Espinosa J, et al. Recent progress of machine learning in gene therapy. *Curr Gene Ther* (2021) 22(2):132–43. doi: 10.2174/1566523221666210622164133
33. Zhao T, Hu Y, Peng J, Cheng L. DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* (2020) 36(16):4466–72. doi: 10.1093/bioinformatics/btaa428
34. Miller MD, Salinas EA, Newton AM, Sharma D, Keeney ME, Warrier A, et al. An integrated prediction model of recurrence in endometrial endometrioid cancers. *Cancer Manag Res* (2019) 11:5301–15. doi: 10.2147/CMARS202628
35. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660
36. Birkeland E, Wik E, Mjos S, Hoivik EA, Trovik J, Werner HM, et al. KRAS gene amplification and overexpression but not mutation associates with aggressive and metastatic endometrial cancer. *Br J Cancer* (2012) 107(12):1997–2004. doi: 10.1038/bjc.2012.477
37. Caley DP, Pink RC, Trujillano D, Carter DR. Long noncoding RNAs, chromatin, and development. *ScientificWorldJournal* (2010) 10:90–102. doi: 10.1100/tsw.2010.7
38. Hou M, Tang X, Tian F, Shi F, Liu F, Gao G. AnnoLnc: a web server for systematically annotating novel human lncRNAs. *BMC Genomics* (2016) 17(1):931. doi: 10.1186/s12864-016-3287-9
39. Cheng L, Hu Y, Sun J, Zhou M, Jiang Q. DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* (2018) 34(11):1953–6. doi: 10.1093/bioinformatics/bty002
40. Park SA, Kim LK, Kim YT, Heo TH, Kim HJ. Long non-coding RNA steroid receptor activator promotes the progression of endometrial cancer via wnt/ β -catenin signaling pathway. *Int J Biol Sci* (2020) 16(1):99–115. doi: 10.7150/ijbs.35643
41. Peng L, Yuan XQ, Liu ZY, Li WL, Zhang CY, Zhang YQ, et al. High lncRNA H19 expression as prognostic indicator: data mining in female cancers and polling analysis in non-female cancers. *Oncotarget* (2017) 8(1):1655–67. doi: 10.18632/oncotarget.13768
42. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* (2012) 81:145–66. doi: 10.1146/annurev-biochem-051410-092902
43. Meng Y, Lu C, Jin M, Xu J, Zeng X, Yang J. A weighted bilinear neural collaborative filtering approach for drug repositioning. *Brief Bioinform* (2022) 23(2):bbab581. doi: 10.1093/bib/bbab581
44. Xu J, Cai L, Liao B, Zhu W, Yang J. CMF-impute: an accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* (2020) 36(10):3139–47. doi: 10.1093/bioinformatics/btaa109
45. Liu C, Wei D, Xiang J, Ren F, Huang L, Lang J, et al. An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol Ther Nucleic Acids* (2020) 21:676–86. doi: 10.1016/j.omtn.2020.07.003
46. Tang X, Cai L, Meng Y, Xu J, Lu C, Yang J. Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. *Front Immunol* (2020) 11:603615. doi: 10.3389/fimmu.2020.603615



OPEN ACCESS

EDITED BY

Liang Cheng,
Harbin Medical University, China

REVIEWED BY

Yaoxin Gao,
East China Normal University, China
Pengshuo Yang,
Shandong First Medical University,
China

*CORRESPONDENCE

Jun Zhou
15304690053@163.com
Jia-Sheng Yang
jsyang.mcc@gmail.com

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 07 July 2022

ACCEPTED 15 August 2022

PUBLISHED 20 September 2022

CITATION

Jiang Z-R, Yang L-H, Jin L-Z, Yi L-M,
Bing P-P, Zhou J and Yang J-S (2022)
Identification of novel cuproptosis-
related lncRNA signatures
to predict the prognosis and
immune microenvironment
of breast cancer patients.
Front. Oncol. 12:988680.
doi: 10.3389/fonc.2022.988680

COPYRIGHT

© 2022 Jiang, Yang, Jin, Yi, Bing, Zhou
and Yang. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Identification of novel cuproptosis-related lncRNA signatures to predict the prognosis and immune microenvironment of breast cancer patients

Zi-Rong Jiang¹, Lin-Hui Yang¹, Liang-Zi Jin², Li-Mu Yi³,
Ping-Ping Bing⁴, Jun Zhou^{4*} and Jia-Sheng Yang^{5*}

¹Department of Surgical Oncology, Ningde Municipal Hospital of Ningde Normal University,
Teaching Hospital of Fujian Medical University, Ningde, China, ²Institute of Medical Biology,
Chinese Academy of Medical Sciences and Peking Union Medical College, Kunming, China,

³Department of Pathology, The First Affiliated Hospital of Guangdong University of Pharmacy,
Guangzhou, China, ⁴Academician Workstation, Changsha Medical University, Changsha, China,

⁵School of Electrical & Information Engineering, Anhui University of Technology, Ma'anshan, China

Background: Cuproptosis is a new modality of cell death regulation that is currently considered as a new cancer treatment strategy. Nevertheless, the prognostic predictive value of cuproptosis-related lncRNAs in breast cancer (BC) remains unknown. Using cuproptosis-related lncRNAs, this study aims to predict the immune microenvironment and prognosis of BC patients. and develop new therapeutic strategies that target the disease.

Methods: The Cancer Genome Atlas (TCGA) database provided the RNA-seq data along with the corresponding clinical and prognostic information. Univariate and multivariate Cox regression analyses were performed to acquire lncRNAs associated with cuproptosis to establish predictive features. The Kaplan-Meier method was used to calculate the overall survival rate (OS) in the high-risk and low-risk groups. High risk and low risk gene sets were enriched to explore functional discrepancies among risk teams. The mutation data were analyzed using the "MAFTools" r-package. The ties of predictive characteristics and immune status had been explored by single sample gene set enrichment analysis (ssGSEA). Last, the correlation between predictive features and treatment condition in patients with BC was analyzed. Based on prognostic risk models, we assessed associations between risk subgroups and immune scores and immune checkpoints. In addition, drug responses in at-risk populations were predicted.

Results: We identified a set of 11 Cuproptosis-Related lncRNAs (GORAB-AS1, AC 079922.2, AL 589765.4, AC 005696.4, Cytos, ZNF 197-AS1, AC 002398.1, AL 451085.3, YTH DF 3-AS1, AC 008771.1, LINC 02446), based on which to

construct the risk model. In comparison to the high-risk group, the low-risk patients lived longer ($p < 0.001$). Moreover, cuproptosis-related lncRNA profiles can independently predict prognosis in BC patients. The AUC values for receiver operating characteristics (ROC) of 1-, 3-, and 5-year risk were 0.849, 0.779, and 0.794, respectively. Patients in the high-risk group had lower OS than those in the low-risk group when they were divided into groups based on various clinicopathological variables. The tumor burden mutations (TMB) correlation analysis showed that high TMB had a worse prognosis than low-TMB, and gene mutations were found to be different in high and low TMB groups, such as PIK3CA (36% versus 32%), SYNE1 (4% versus 6%). Gene enrichment analysis indicated that the differential genes were significantly concentrated in immune-related pathways. The predictive traits were significantly correlated with the immune status of BC patients, according to ssGSEA results. Finally, high-risk patients showed high sensitivity in anti-CD276 immunotherapy and conventional chemotherapeutic drugs such as imatinib, lapatinib, and pazopanib.

Conclusion: We successfully constructed of a cuproptosis-related lncRNA signature, which can independently predict the prognosis of BC patients and can be used to estimate OS and clinical treatment outcomes in BRCA patients. It will serve as a foundation for further research into the mechanism of cuproptosis-related lncRNAs in breast cancer, as well as for the development of new markers and therapeutic targets for the disease.

KEYWORDS

cuproptosis, breast cancer, lncRNA, tumor microenvironment, tumor mutation burden

Introduction

Breast cancer (BC) is a cancer that affects mainly women and is a major factor in mortality worldwide (1). BC has overtaken lung cancer as the most usual cancers, with an estimated 2.26 million new cases in the world (11.7%) (2–4). Developing countries account for nearly 60% of mortality (4). Despite early screening and development of anticancer strategies, the prognosis of BC patients has improved significantly (5), but the recurrence rate of BC remains high (6, 7). The prognosis of breast cancer depends not only on the pathological stage at the time of detection but also mainly on the category of breast cancer (8). Breast cancer can be classified into different subgroups, which are primarily on the basis of human epidermal growth factor receptor 2 (HER2), progesterone receptor expression (PR), Ki-67 value, and estrogen receptor (ER). By exploring global gene expression profiles, breast cancer can be also divided into molecular subgroups, HER2-enriched, including Luminal B, Luminal A, Basal-like, and Normal-like (9). BC is a highly heterogeneous tumor, and the search for biomarkers which helps to the diagnosis, prognosis and prediction of breast

cancer has important meaning for monitoring breast cancer recurrence and disclosing new therapeutic target spots throughout the treatment process (10, 11).

Cuproptosis is a new regulatory cell death pattern that mainly relies on the direct binding of fatty acyl components of the tricarboxylic acid (TCA) cycle of mitochondrial respiration, distinguished from other regulatory cell death features like pyroptosis, ferroptosis and apoptosis. With the accumulation of acylated protein and the subsequent decrease in iron-sulfur cluster protein, the cells died due to protein toxic stress. Copper is a vital cofactor for all organisms, but it can become poisonous if density exceed the threshold value maintained by evolutionarily conservative steady-state mechanisms. Current studies have found significant changes in copper content in serum and tumor tissues in some tumor patients (12–14). In addition, mechanisms regarding copper-dependent tumor growth and progression have been recently discovered and summarized in other studies (15, 16). Copper is also capable of promoting angiogenesis, which is essential for tumor metastasis and progression. Overload of copper can also lead to cell death. Because copper plays a crucial part in the

occurrence, severity and development of cancer, it may be an important hub for halting cancer development (17).

Long non-coding RNA (lncRNA) means a non-coding RNA that is more than 200 nucleotides in length (18, 19). Increasing evidence show that lncRNA play a crucial role in regulating tumor occurrence and metastasis (20). For example, it has been demonstrated that the M2 polarization of macrophages and the aggressiveness of BC cells are both influenced by the LINC01140/miR-140-5p/FGF9 axis (21). lncRNA PVT1 can accelerate malignant changes in the phenotype of BC cells by controlling the MIR1-194-5P/BCLAF1 axis as a competing endogenous RNA (22). lncRNA were the key factors for BC chemical resistance (23). Although there are more and more studies on the role of lncRNA in cancer, our understanding of the role of lncRNA in the occurrence is less. Currently, there is a small number of studies on the lncRNA related to cuproptosis, and no studies on the lncRNA related to cuproptosis in BC.

In this study, we established a predictive features and internal validation on the basis of cuproptosis-related lncRNAs. We further analyzed its underlying value in predicting the diagnosis, prognosis, chemotherapy response, and tumor immune infiltration of patients with BC.

Materials and methods

Patients and datasets

We do this from the TCGA official website (<https://portal.gdc.cancer.gov/>); Data were obtained for 1089 alive patients with lncRNA expression. Fifty-two genes related to cuproptosis were downloaded from the genecard database and previous literature (24) (GeneCard filtering condition: Relevance score >20).

Construction of prediction features of cuproptosis-associated lncRNA

Associations between cuproptosis-related genes and lncRNAs were calculated using the “limma” package in R. Using correlation coefficients $|R2| > 0.35$ and $p < 0.05$ as screening standard, in sum, 1,136 cuproptosis-related lncRNA expressions were obtained. We made use of univariate Cox regression analysis to gain cuproptosis-related lncRNA associated with the prognosis of BC patients, then constructed training and test cohorts according to the 1:1 random assignment principle. The training cohort was subjected to lasso-Cox regression analysis to acquire cuproptosis-related lncRNA, and the prediction features of cuproptosis-related lncRNA were constructed. The equation used for this analysis was as follows: Risk score = $(\text{Exp}_i \times \beta_i)$. (Exp: expression level of model gene; β : model gene coefficient).

Construction of nomogram

By combining the risk score with the clinicopathological characteristics of patients' age, phase, and TNM stage, a nomogram that can forecast patients with BRCA's 1-, 3-, and 5-year survival was developed. We verified that predicted survival is consistent with actual survival by using a calibration curve.

Collection and pretreatment of epigenetic mutation data

Somatic changes in the BRCA cohort were obtained from the TCGA database. TMB was defined as the number of somatic, coding, base substitution, and indel mutations per megabase in the genome tested using non-synonymous and transcribing indels at the 5% limit of detection. The “MAF Tools” R package (25) has been used to test the amount of somatic non-sense point mutations in per sample. Reveals how BRCA drives somatic changes in genes for samples with low and high-risk scores.

Gene ontology and Kyoto Encyclopedia of genes and genome pathway analysis

In accordance with the median risk score, patients with BC were distinguish between high-risk team and low-risk team. The error detection rate (FDR) < 0.05 and $|\log_2\text{fold change (FC)}| > 1$ were used as screening standard for obtaining cuproptosis-related genes (DEGs) with divergence in expression. We utilize the “ggplot2” package for GO and KEGG analysis. The difference was regarded as statistically significant when the P value was less than 0.05.

Correlation between risk score and immune cell infiltration

The relative ratio of infiltrating immune cells was calculated using the ssGSEA and CIBERSORT R scripts. In exploring the correlation during risk rating values and immunologically infiltrating cells, we used a Spearman rank correlation analysis.

Role of predictive features in predicting clinical treatment outcome

To assess the role of predictive features in forecasting response to BRCA therapy, we computed the maximum inhibitory concentration (IC50) half of the common

chemotherapeutic agents used to the clinical treatment of BC. In order to compare the IC50 values between high-risk and low-risk teams, the Wilcoxon signed rank test was used.

Data analysis

All data analyses were executed using the R software (version 4.0.2). The expression of cuproptosis-related DEGs in both normal and cancer tissues was examined using Wilcoxon’s test. Using univariate Cox regression analysis, the association between cuproptosis-related lncRNA and overall survival (OS) was examined, and lasso- Cox analysis was applied to sift cuproptosis-related lncRNA to set up predictive features. The Kaplan-Meier approach and the logarithmic rank examine were used to compare the OS of patients in the high-risk team and the low-risk team. The ROC curve was plotted using the “survivalROC” package, and the area under the curve (AUC) value was calculated.

Results

Construction of prediction features of cuproptosis-associated lncRNA

We identified 1136 lncRNA associated with cuproptosis (Supplementary Table S1). Univariate Cox regression analysis showed there were 51 lncRNA related to the recovery results of BC patients. We made use of Lasso multivariate Cox regression analysis to select for efficient prognostic-related symbols, and a prognostic model was established from the training set (Figures 1A, B). It showed that 11 lncRNA (GORA B-AS1, AC 079922.2, AL 589765.4, AC 005696.4, Cytor, ZNF 197-AS1, AC 002398.1, AL 451085.3, YTHDF 3-AS1, AC 008771.1, LINC 02446) associated with cuproptosis were identified to construct predictive signatures. GORAB-AS1, AL589765.4, AC005696.4, CYTOR, YTHDF3-AS1, AC008771.1 were protective while AC079922.2, ZNF197-AS1, AC002398.1, AL451085.3, LINC02446 were risk factors (Figure 1C). The risk grade was

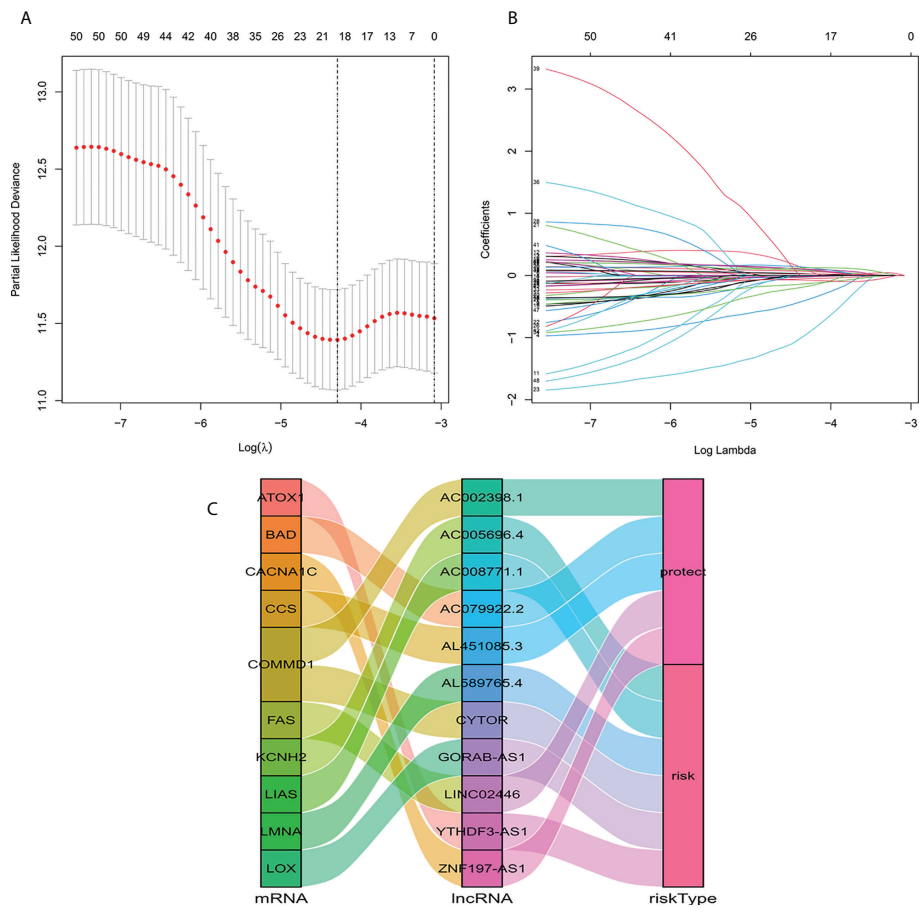


FIGURE 1
Construction of a signature for predicting features of cuproptosis-related lncRNA using the Lasso method. (A) Cross validation diagram. (B) LASSO coefficients of prognostic genes. (C) Sankey diagram of prognostic cuproptosis-related lncRNAs.

calculated as follows: Risk grade = $(0.643 \times \text{GORAB-AS1 expression}) + (-0.892 \times \text{AC079922.2 expression}) + (0.160 \times \text{AL589765.4 expression}) + (0.160 \times \text{AC005696.4 expression}) + (0.126 \times \text{CYTOR expression}) + (-1.927 \times \text{ZNF197-AS1 expression}) + (-0.278 \times \text{AC002398.1 expression}) + (-0.603 \times \text{AL451085.3 expression}) + (0.206 \times \text{YTHDF3-AS1 expression}) + (0.055 \times \text{AC008771.1 expression}) + (-0.147 \times \text{LINC02446 expression})$.

Relationship between predictive features and prognosis of BC patients

According to the formula, each patient's risk grade was determined, and the patients were then divided into high-risk

and low-risk teams based on the median risk score. To ascertain the value of the risk score in predicting the prognosis of BC patients, the OS time in the high-risk and low-risk groups was analyzed using the Kaplan-Meier method. The OS time in the high-risk team was significantly reduced when compared to the low-risk team (Figure 2A, $p < 0.001$). To ascertain if predictive features were an independent prognostic element in patients with BC, Cox regression analysis was executed. Age, T stage, N stage, M stage, and risk score were found to be significantly correlated with OS in univariate Cox regression analysis of BC patients (Figure 2B). In patients with BC, multivariate Cox regression analysis revealed that risk scores and age were independent predictors of OS (Figure 2C). Risk scores for high

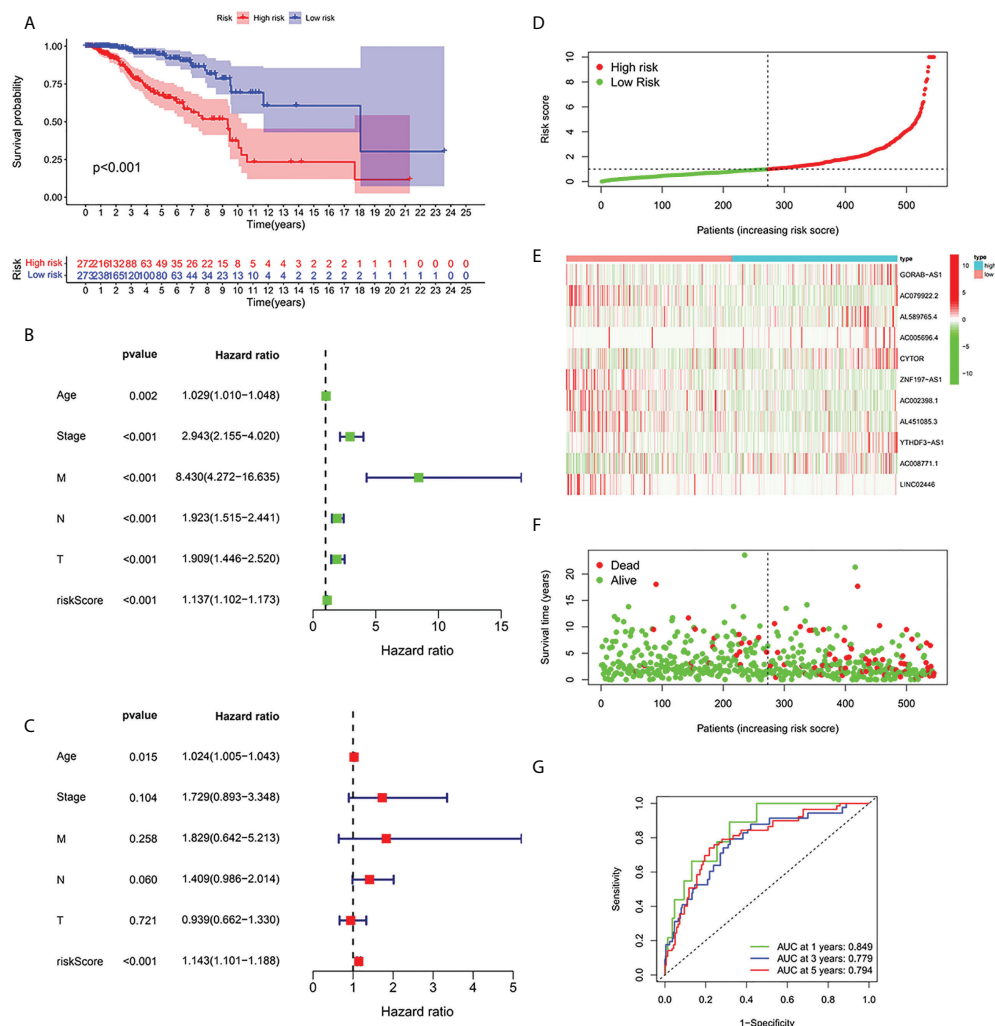


FIGURE 2

Relationship between prognosis and predictive features of patients with BC. (A) OS rates for BC patients in the high-risk and low-risk teams according to a Kaplan-Meier analysis. (B) Forest map of univariate Cox regression analysis. (C) Forest map of multivariate Cox regression analysis. (D) Risk score of BRCA patients calculated according to the model and division of high- and low-risk teams. (E) Gene expression heat maps. (F) Survival status. (G) ROC curve of the predictive characteristic and AUC of 1-, 3-, and 5-year survival. OS, overall survival; ROC, receiver operating characteristics; AUC, area under curve; T, tumor; N, lymph nodes; M, Metastasis.

and low risk teams as displayed in Figures 2D–F. As shown in the figure, the number of patients who died increased as the risk increased, and the expression of Cuproptosis-related lncRNA in the two risk groups was also shown. Indicating good predictive performance, the 1-, 3-, and 5-year survival AUCs were 0.849, 0.779, and 0.794, respectively (Figure 2G).

Identification and verification of nomograms

To further forecast the prognosis of patients with BC, we set up an nomogram with clinical pathology variables and risk scores that predicts the 1, 3, and 5 years prognosis of patients with BC (Figure 3A). The calibration curve demonstrates good agreement between the predicted survival at 1, 3, and 5 years and the actual OS rates (Figure 3B). All results suggest that histograms created by cuproptosis-related lncRNAs have good prognostic potential for BC patients.

The correlation between predictive features and prognosis of BC patients with different clinical pathological variables

In order to investigate the correlation between prognosis and predictive features of patients with BC classified in accordance with different clinical pathological variables, patients with BC were grouped according to age, T stage, N stage and M stage. Significantly fewer patients in the high-risk group survived longer than those in the low-risk group. These findings imply that regardless of clinicopathological factors, the predictive

function can forecast the prognosis of BC patients (Figures 4A–H).

Validation of predictive signature

To verify of the applicability of OS prediction features on the basis of the entire TCGA dataset, validation in both the test cohort and the total cohort, the test cohort showed that patients in the high-risk group had lower OS ratios than those in the low-risk group (Figure 5A, $p < 0.001$). The high-risk team's prognosis was worse than the low-risk team's prognosis for the entire cohort (Figure 5B, $p < 0.001$). The ROC curves of both cohorts showed cracking predictive performance. In the test cohort, the AUC was 0.686, 0.687, and 0.686 for the 1-, 3-, and 5-year survival rates, separately (Figure 5C). In the total cohort, the 1-, 3-, and 5-year survival AUCs were 0.766, 0.734, and 0.736, separately (Figure 5D).

Association between risk signature and TMB

Current studies had emphasized that high TMB were significantly related with abundant CD8+ T cells, which can recognize cancer cells and then lead to an anti-tumor immune outcome (26–29). Therefore, we inferred that TMB might serve as a non-negligible prognostic element in the anti-tumor immunotherapy response, aiming to study interactions between risk scoring and TMB to reveal the genetic variation of subtypes of risk scoring (30). Patients were distributed to different subtypes on the TMB immune set point line. The survival curve indicated that a high TMB value markedly

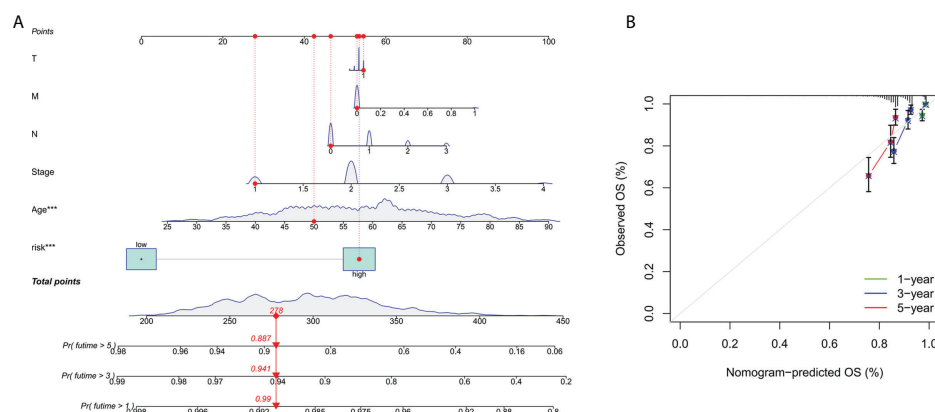


FIGURE 3

Construction and verification of nomograms. (A) The 1,3and 5 years OS of patients with BC is predicted by nomograms that combine clinical pathology variables with risk scores. (B) The calibration curve tested the agreement between the actual OS rate and the predicted 1-,3-, and 5-year survival rates.

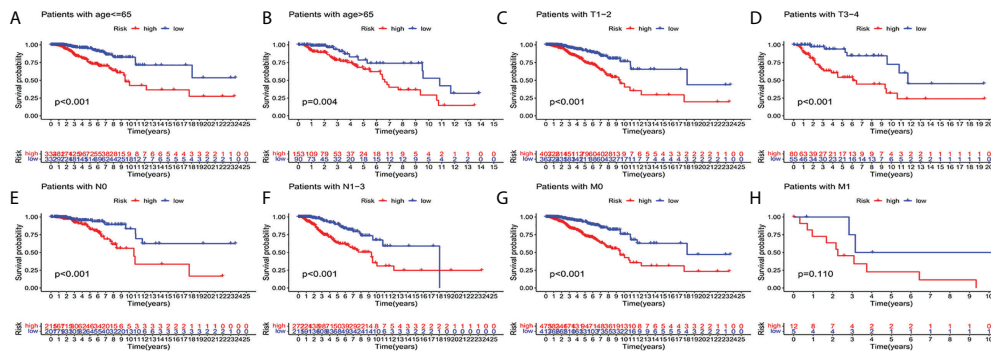


FIGURE 4

Kaplan-Meier survival curves of high-risk and low-risk teams in the patients sorted by different clinical pathological variables. (A, B) age. (C, D) T staging. (E, F) N staging. (G, H) M staging. T, tumor; N, lymph node; M, distant transfer.

indicated a short total survival time ($p = 0.018$, Figure 6A). To further quest for validity of risk scores and consistent prognostic significance of TMB, we tested and verified the synergistic influence of the two markers in predicting the prognosis of BRCA. As shown by the layered survival curve, the TMB status did not interfere with the prognostic performance of the risk score. The risk score subgroup showed significant difference in

prognosis between the low and high TMB status hypotypes ($p < 0.001$, Figure 6B).

In addition, we probed and visualized the distribution of gene mutations among subtypes with different risk scores. The integrated scenery of somatic variations showed mutational models and clinical features of the top 15 most frequently changed driving genes (Figures 6C, D). Significant mutation

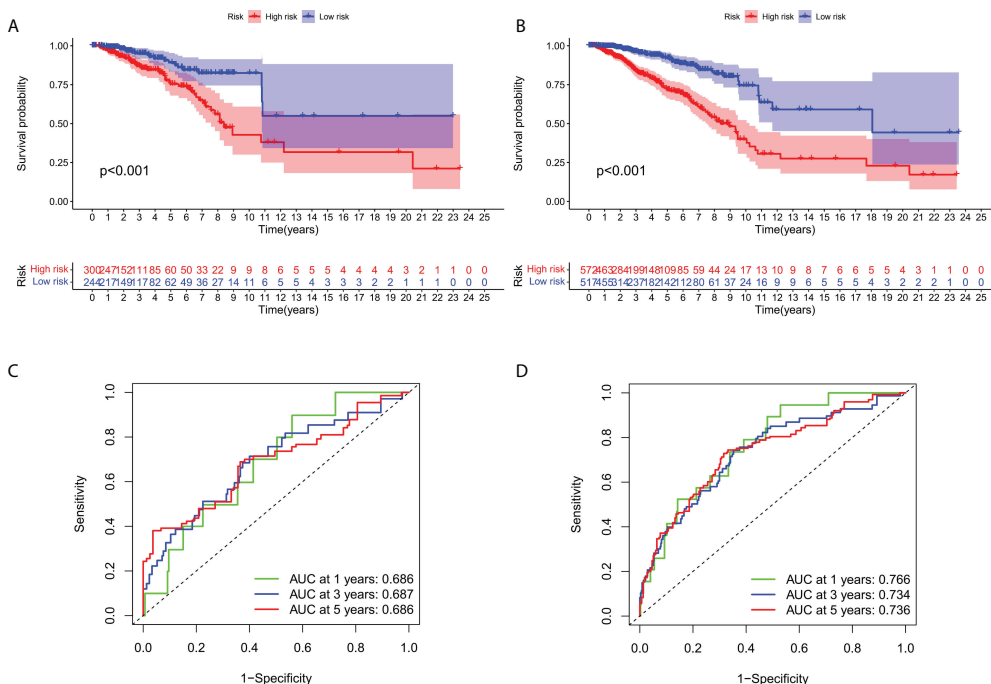


FIGURE 5

Internal verification of OS prediction signatures based on the TCGA dataset. (A) Kaplan-Meier survival curves in the test cohort. (B) Kaplan-Meier survival curves in the total cohort. (C) ROC curves and AUC for 1-, 3-, and 5-year survival in the test cohort. (D) ROC curves and AUC for 1-, 3-, and 5-year survival in the total cohort.

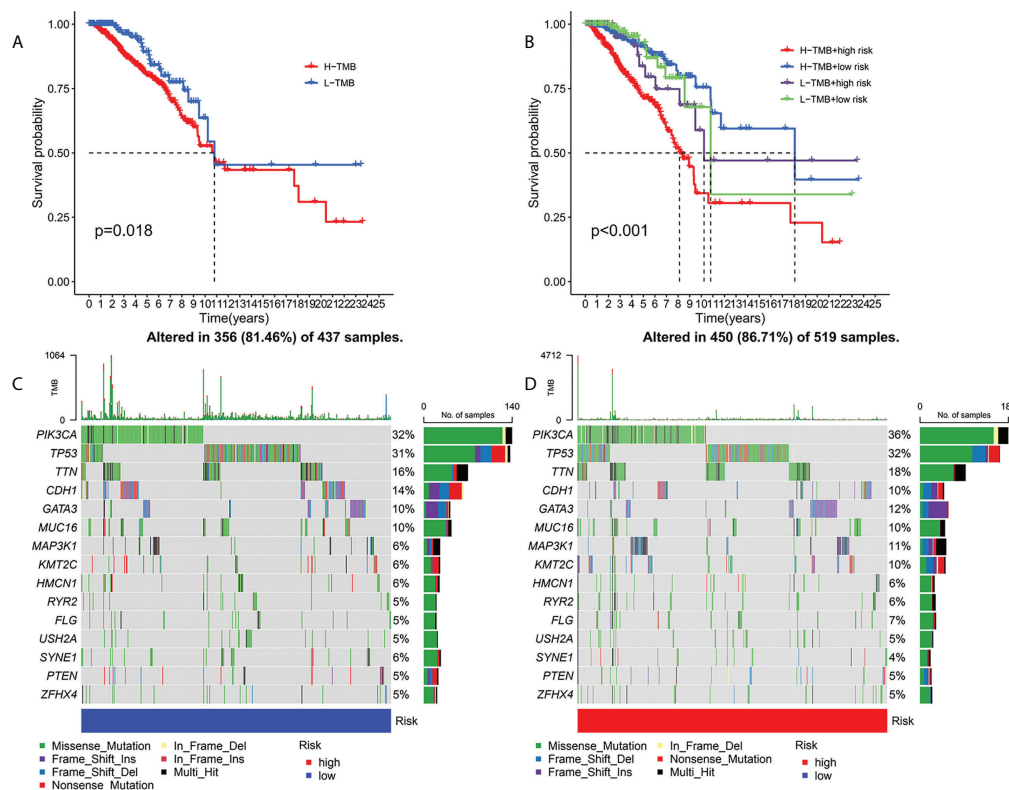


FIGURE 6

Association of risk score with TMB and gene mutations. Kaplan-Meier curves for high and low TMB teams (A). Kaplan-Meier curve for patients layered by TMB and risk score (B). OncoPrint was built with a low-risk score (C) and a high-risk score (D).

(SMG) profiles indicated that PIK3CA (36% versus 32%) had a higher rate of somatic mutations in the high-risk core subtypes, while SYNE1 (4% versus 6%) had a higher rate of somatic mutations in the low-risk core subtypes—a subset of risk scores. These findings may shed new light on the intrinsic link between high and low risks and somatic variation in BRCA immunotherapy. All in all, these consequences suggest that risk scores may serve as independent prognostic predictors and have the potential to access clinical outcomes of anti-tumor immunotherapy.

Heat map and gene enrichment analysis

To investigate the expression of prognostic model genes in clinical characteristics, we established expression heatmaps based on clinical features, correlations between high-risk and low-risk groups, patient age, tumor stage, and lymphoid. Prognostic model genes between nodes were investigated. Lymph node metastasis and Immune Score (Figure 7A). Since the prognosis of patients in the high-risk team and the low-risk team was different, enrichment analysis was executed to study possible discrepancies between the high-risk team and the low-

risk team (Figure 7B). Moreover, we discovered that in biological process (BP), enriched in humoral immune response, protein activation cascade, completion activation, classical pathway. Human immune response mediated by circulating immune globulin, the cellular component is enriched in the immune globulin complex, circulating, blood microparticle, external side of plasma membrane, collagen-containing extracellular matrix. Molecular-rich antigen binding, rage receptor binding, immunoglobulin receptor binding, extra-granular matrix structural constitution. More importantly, KEGG was enriched in IL-17 signaling pathway- κ B signaling pathway, B cell receptor signaling pathway, completion and colonization cascades.

Immune cell infiltration and immune-related function

To investigate further the relationship between risk scores and immune cells and function, we used ssGSEA to calculate enrichment scores for various immune cell subsets, associated functions, or pathways. Results showed that sensitized Dendritic cells (aDC), B cells, CD8+T cells, T follicle helper cells (Tfh) cells, T helper type 1 (Th1) cells, Tumor Infiltrating

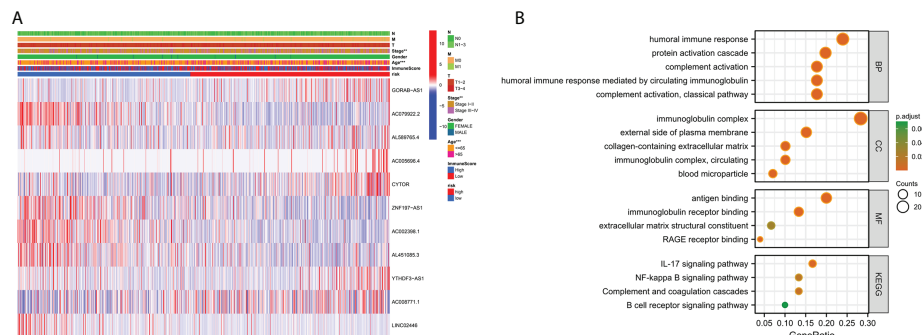


FIGURE 7

Clinically relevant heat map and GO/KEGG pathway enrichment analysis. (A) On the basis of risk characteristics associated with prognosis, a heat map of distribution of cuproptosis-related lncRNA and clinical pathology variables was plotted. The more intense the red, the more intense the expression. The more intense the blue, the more subdued the expression. $**p < 0.01$, $***p < 0.001$. (B) The graph depicts the GO and KEGG analysis of differential genes with high and low risk..

Lymphocytes (TIL) were strikingly different between the high- and low-risk teams (Figure 8A). The IFN response was lower in the high-risk team than in the low-risk team for cytolytic activity, Human Leukocyte Antigen (HLA), T-cell-co-stimulation, inflammation promotion, and Type II immune function scoring (Figure 8B). These findings suggest that immune function is more inactive in high-risk groups.

Correlation between predictive characteristics and BC treatment

Compared with the low-risk team, CD276 expression was significantly increased in the high-risk group, suggesting that

high-risk patients may respond to anti-CD276 immunotherapy. The expression levels of CD274, PDCD1, and CTLA4 in the low-risk team were significantly increased, suggesting that low-risk patients might have a potential response to immunotherapy with PD-1, PD-L1, and CTLA4 (Figure 9). In addition to immunotherapy, we also studied the relationship between predictive features and the general chemotherapeutic response of BC. The IC50 values of imatinib, lapatinib, and pazopanib in the high-risk team were found to be lower, while those of bosutinib, cisplatin, cytarabine, gefitinib, gemcitabine, lenalidomide, nilotinib, paclitaxel, and sunitinib in the high-risk team were found to be higher (Figures 10A–L), which helped explore treatment options for high-risk and low-risk populations.

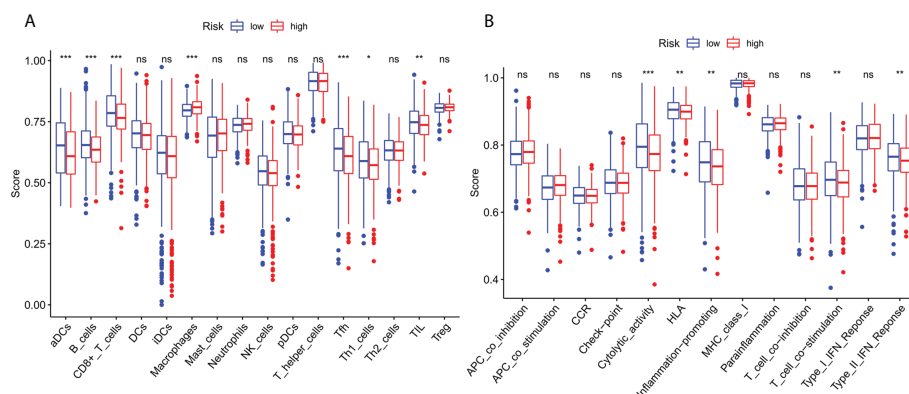


FIGURE 8

Immune infiltration cell score and immune-related function in high-risk and low-risk population. (A) The ssGSEA algorithm was used to calculate the levels of infiltration of 16 immune cells in high-risk and low-risk populations. (B) Relationship between predictive features and 13 immune-related functions. $*p < 0.05$; $**p < 0.01$; $***p < 0.001$; ns, not significant.

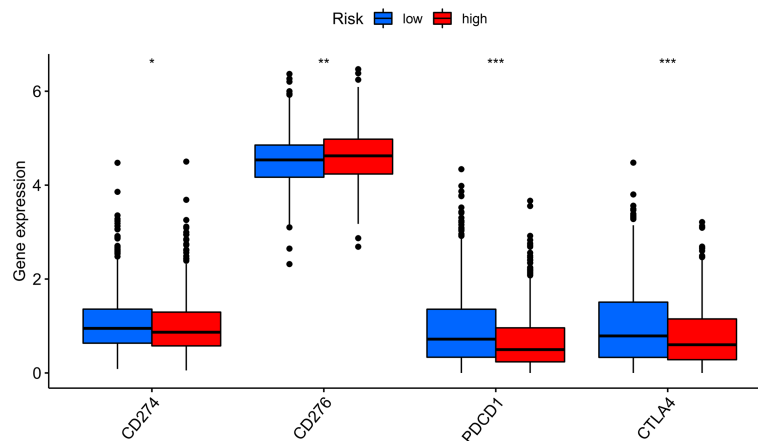


FIGURE 9

Immune checkpoint expression in BRCA patients from two different risk groups. Expression of two immune checkpoints (CD274, CD276, PDCD1 and CTLA4) in the TCGA cohort. ANOVA was applied as significance test, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Discussion

Although the mortality rate of BC has decreased due to the early detection and advanced treatment, the morbidity rate continues to increase annually (31–33). The role of cuproptosis in cancer was complex. Significant changes in copper content in serum and tumor tissues of patients with different cancers (such as breast cancer, cervical cancer, thyroid cancer, lung cancer, pancreatic cancer, prostate cancer, ovarian cancer, oral cancer, breast cancer, and bladder cancer) had been found (12–14). More and more studies had discovered that cuproptosis played a crucial role in the occurrence and development of cancer. However, recent research mainly paid attention to the role of cuproptosis in cancer mechanisms (12–14) and few studies to its role in cancer prognosis.

Researches had shown that lncRNA do not represent transcriptional noise, which played an essential role in tumors (34–37). For example, autophagy-related lncRNA characteristics could accurately forecast the prognosis of patients with Bladder Cancer (38). Cuproptosis-associated lncRNA was a good predictor of prognosis in patients with colon cancer (39). Prognosis of patients with BC had not been studied by constructing predictive features of lncRNA associated with cuproptosis (40, 41). Therefore, it was important to identify predictive features of lncRNA associated with cuproptosis in BC patients. In this research, we applied univariate Cox regression analysis to analyze the relationship between cuproptosis-related lncRNA and the prognosis of patients with BC patients and discovered that 51 lncRNA were related to the prognosis of patients with BC. Through lasso multivariate Cox regression analysis, we determined that 11 lncRNAs associated with cuproptosis were included in the predictive signature.

We also found significant co-expression of mRNA (LOX, BAD, LMNA, KCNH2, COMMD1, CACNA1C, COMMD1, CCS, ATOX1, LIAS, FAS) with these lncRNA. Among them, LIAS was a vital driving factor for the death of copper (24). At present, some clinical data indicate the correlation between genetic changes and immunotherapy response (42, 43). We computed and identified TMB, which was a predictor of immunotherapeutic sensitivity and increases significantly with increased risk scores. Succeeded stratified survival curves indicated that the risk scores had prognostic capabilities independent of TMB, indicating that TMB and risk scores represented various aspects of immunobiology. Furthermore, the risk score, together with the mutation data, revealed a prominent discrepancy in the frequency of gene variation at the transcriptome level between different groupings. In this work, the mutation rate of PIK3CA was significantly increased in subtypes with high-risk scores, while the mutation rate of SYNE1 was increased in patients with low-risk scores.

Since the prognosis of patients in the high-risk team and the low-risk team was different, enrichment analysis was enforced to study possible divergences between the high-risk team and the low-risk team. Besides, we discovered that in biological process (BP), enriched in humoral immune response, protein activation cascade. The cellular component was enriched in the immune globulin complex and immunoglobulin receptor binding. More importantly, KEGG was enriched in IL-17 signaling pathway, completion and colonization cascades, NF- κ B signaling pathway. Enrichment results showed that differential genes with high risk were closely linked to tumor and immune-related pathways.

Subsequently, ssGSEA results indicated that activated dendritic cells (ADCs), CD8+T cells, B cells, and tumor-

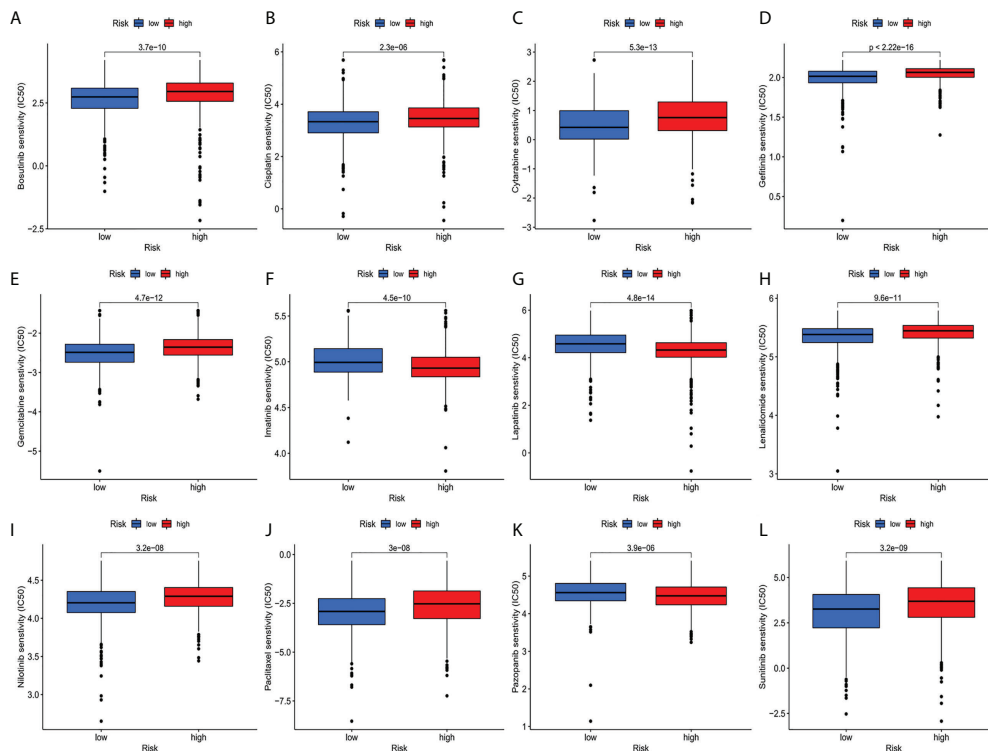


FIGURE 10

(A–L) IC50 for small molecule drugs in high and low risk populations. (A): Bosutinib, (B): Cisplatin, (C): Cytarabine, (D): Gefitinib, (E): Gemcitabine, (F): Imatinib, (G): Lapatinib, (H): Lenalidomide, (I): Nilotinib, (J): Paclitaxel, (K): Pazopanib, (L): Sunitinib. IC50, half maximum inhibitory concentration.

infiltrating lymphocytes (TILs) were markedly different between the high-risk and low-risk teams. The IFN response was lower in the high-risk team than in the low-risk team for cytolytic activity, human leukocyte antigen (HLA), and Type II immune function score, inflammation promotion. These results of the survey suggested that immune function is more inactive in high-risk teams. Aside from increased tumor immune cell infiltration, the high-risk group was associated with decreased antitumor immunity, and HLA and type I IFN responses scored lower in the high-risk teams. Therefore, in breast cancer, reduced antitumor immunity in high-risk teams could explain the poor prognosis. Our study also indicated that high-risk patients may be susceptible to anti-CD276 immunotherapy and conventional chemotherapy drugs imatinib, lapatinib, and pazopanib, but resistant to bosutinib, cisplatin, cytarabine, gefitinib, gemcitabine, lenalidomide, nilotinib, paclitaxel, and sunitinib. This proved that high-risk individuals can obtain a benefit from the combined immunotherapy and chemotherapy, providing a basis for precise and individualized treatment of BC patients.

However, there were some limitations in our study. Firstly, we only used data from the TCGA database for internal

validation, and we needed data from other databases for external validation to test the predictive signature's applicability further. Secondly, the mechanism of action of cuproptosis-related lncRNA in BC needed to be further verified through experiments.

Conclusion

The cuproptosis-related lncRNA signature, in conclusion, can independently predict a patient's prognosis for BC and provides evidence for a potential cuproptosis-related lncRNA mechanism in BC and its response to clinical therapy.

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.

Ethics statement

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

Author contributions

J-SY and JZ conceived and designed the study. Z-RJ, P-PB, and L-HY performed the experiments. L-ZJ and L-MY analyzed the data. Z-RJ wrote the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors thank reviewers for helpful comments on the manuscript.

References

- Bodine PV, Litwack G. Purification and structural analysis of the modulator of the glucocorticoid-receptor complex: evidence that modulator is a novel phosphoglyceride. *J Biol Chem* (1988) 263:3501–12. doi: 10.1016/S0021-9258(18)69099-4
- Ahmad A. Breast cancer statistics: Recent trends. *Adv Exp Med Biol* (2019) 1152:1–7. doi: 10.1007/978-3-030-20301-6_1
- Liu H, Qiu C, Wang B, Bing P, Tian G, Zhang X, et al. Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-Origin. *Front Cell Dev Biol* (2021) 9:619330. doi: 10.3389/fcell.2021.619330
- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2021) 71:209–49. doi: 10.3322/caac.21660
- Davis S, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. *Bioinformatics* (2007) 23:1846–7. doi: 10.1093/bioinformatics/btm254
- Huang Z, Xiao C, Zhang F, Zhou Z, Yu L, Ye C, et al. A novel framework to predict breast cancer prognosis using immune-associated lncRNAs. *Front Genet* (2020) 11:634195. doi: 10.3389/fgene.2020.634195
- Yang J, Ju J, Guo L, Ji B, Shi S, Yang Z, et al. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput Struct Biotechnol J* (2022) 20:333–42. doi: 10.1016/j.csbj.2021.12.028
- Liu X, Yuan P, Li R, Zhang D, An J, Ju J, et al. Predicting breast cancer recurrence and metastasis risk by integrating color and texture features of histopathological images and machine learning technologies. *Comput Biol Med* (2022) 146:105569. doi: 10.1016/j.compbiomed.2022.105569
- Allaoui R, Bergenfelz C, Mohlin S, Hagerling C, Salari K, Werb Z, et al. Cancer-associated fibroblast-secreted CXCL16 attracts monocytes to promote stroma activation in triple-negative breast cancers. *Nat Commun* (2016) 7:13050. doi: 10.1038/ncomms13050
- Weigel MT, Dowsett M. Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocr Relat Cancer* (2010) 17:R245–62. doi: 10.1677/ERC-10-0136
- Shi X, Young S, Cai K, Yang J, Morahan G. Cancer susceptibility genes: update and systematic perspectives. *Innovation* (2022) 3(5):100277. doi: 10.1016/j.xinn.2022.100277

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/fonc.2022.988680/full#supplementary-material>

- Baltaci AK, Dundar TK, Aksoy F, Mogulkoc R. Changes in the serum levels of trace elements before and after the operation in thyroid cancer patients. *Biol Trace Elem. Res* (2017) 175:57–64. doi: 10.1007/s12011-016-0768-2
- Stepien M, Jenab M, Freisling H, Becker NP, Czuban M, Tjonneland A, et al. Pre-diagnostic copper and zinc biomarkers and colorectal cancer risk in the European prospective investigation into cancer and nutrition cohort. *Carcinogenesis* (2017) 38:699–707. doi: 10.1093/carcin/bgx051
- Zhang X, Yang Q. Association between serum copper levels and lung cancer risk: A meta-analysis. *J Int Med Res* (2018) 46:4863–73. doi: 10.1177/0300060518798507
- Ruiz LM, Libedinsky A, Elorza AA. Role of copper on mitochondrial function and metabolism. *Front Mol Biosci* (2021) 8:711227. doi: 10.3389/fmolb.2021.711227
- Ge EJ, Bush AI, Casini A, Cobine PA, Cross JR, Denicola GM, et al. Connecting copper and cancer: from transition metal signalling to metalloplasia. *Nat Rev Cancer* (2022) 22:102–13. doi: 10.1038/s41568-021-00417-2
- Shanbhag VC, Gudekar N, Jasmer K, Papageorgiou C, Singh K, Petris MJ. Copper metabolism as a unique vulnerability in cancer. *Biochim Biophys Acta Mol Cell Res* (2021) 1868:118893. doi: 10.1016/j.bbamcr.2020.118893
- Xiao X, Zhu W, Liao B, Xu J, Gu C, Ji B, et al. BPLDPA: Predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network. *Front Genet* (2018) 9:411. doi: 10.3389/fgene.2018.00411
- Wang L, Xiao Y, Li J, Feng X, Li Q, Yang J. IIRWR: internal inclined random walk with restart for lncRNA-disease association prediction. *IEEE Access* (2019) 7:54034–41. doi: 10.1109/ACCESS.2019.2912945
- Schmitt AM, Chang HY. Long noncoding RNAs in cancer pathways. *Cancer Cell* (2016) 29:452–63. doi: 10.1016/j.ccell.2016.03.010
- Wu S, Xu R, Zhu X, He H, Zhang J, Zeng Q, et al. The long noncoding RNA LINC01140/miR-140-5p/FGF9 axis modulates bladder cancer cell aggressiveness and macrophage M2 polarization. *Aging (Albany NY)* (2020) 12:25845–64. doi: 10.18632/aging.202147
- Chen PH, Wu J, Ding CC, Lin CC, Pan S, Bossa N, et al. Kinome screen of ferroptosis reveals a novel role of ATM in regulating iron metabolism. *Cell Death Differ.* (2020) 27:1008–22. doi: 10.1038/s41418-019-0393-7
- Zangouei AS, Rahimi HR, Mojarad M, Moghbeli M. Non coding RNAs as the critical factors in chemo resistance of bladder tumor cells. *Diagn Pathol* (2020) 15:136. doi: 10.1186/s13000-020-01054-3

24. Tsvetkov P, Coy S, Petrova B, Dreishpoon M, Verma A, Abdusamad M, et al. Copper induces cell death by targeting lipoylated TCA cycle proteins. *Science* (2022) 375:1254–61. doi: 10.1126/science.abf0529
25. Ewing D, Jones SR. Superoxide removal and radiation protection in bacteria. *Arch Biochem Biophys* (1987) 254:53–62. doi: 10.1016/0003-9861(87)90080-4
26. Coscia L, Causa P, Giuliani E, Nunziata A. Pharmacological properties of new neuroleptic compounds. *Arzneimittelforschung* (1975) 25:1436–42.
27. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* (2015) 348:124–8. doi: 10.1126/science.aaa1348
28. Aiken LH, Sloane DM, Barnes H, Cimiotti JP, Jarrin OF, Mchugh MD. Nurses' and patients' appraisals show patient safety in hospitals remains a concern. *Health Aff. (Millwood)* (2018) 37:1744–51. doi: 10.1377/hlthaff.2018.0711
29. Liu J, Lan Y, Tian G, Yang J. A systematic framework for identifying prognostic genes in the tumor microenvironment of colon cancer. *Front Oncol* (2022) 2:899156. doi: 10.3389/fonc.2022.899156
30. Conway CM. Editorial: "Old lamps for new". *Br J Anaesth.* (1975) 47:811–2. doi: 10.1093/bja/47.8.813
31. He B, Lang J, Wang B, Liu X, Lu Q, He J, et al. TOOMe: A novel computational framework to infer cancer tissue-of-Origin by integrating both gene mutation and expression. *Front Bioeng. Biotechnol* (2020) 8:394. doi: 10.3389/fbioe.2020.00394
32. Wang H, Zhao L, Liu H, Luo S, Akinyemiju T, Hwang S, et al. Variants in SNAI1, AMDHD1 and CUBN in vitamin d pathway genes are associated with breast cancer risk: a large-scale analysis of 14 GWASs in the DRIVE study. *Am J Cancer Res* (2020) 10:2160–73. doi: 10.1002/mc.23331
33. Wang B, Yang H, Zhang Y, Tian G, Yang J. A computational framework to trace tumor tissue-of-origin of 19 cancer types based on RNA sequencing. *Res Square* (2022) 1866(11):165916. doi: 10.21203/rs.3.rs-1457167/v1
34. Du R, Bai Y, Li L. Biological networks in gestational diabetes mellitus: insights into the mechanism of crosstalk between long non-coding RNA and N(6)-methyladenine modification. *BMC Pregnancy Childbirth* (2022) 22:384. doi: 10.1186/s12884-022-04716-w
35. Li L, Xie W. LncRNA HDAC11-AS1 suppresses atherosclerosis by inhibiting HDAC11-mediated adropin histone deacetylation. *J Cardiovasc Transl Res* (2022) 1–14. doi: 10.1007/s12265-022-10248-7
36. Zhao C, Xie W, Zhu H, Zhao M, Liu W, Wu Z, et al. LncRNAs and their RBPs: How to influence the fate of stem cells? *Stem Cell Res Ther* (2022) 13:175. doi: 10.1186/s13287-022-02851-x
37. Zheng YJ, Liang TS, Wang J, Zhao JY, Zhai SN, Yang DK, et al. Long non-coding RNA ZNF667-AS1 retards the development of esophageal squamous cell carcinoma via modulation of microRNA-1290-mediated PRUNE2. *Transl Oncol* (2022) 21:101371. doi: 10.1016/j.tranon.2022.101371
38. Sun Z, Jing C, Xiao C, Li T. An autophagy-related long non-coding RNA prognostic signature accurately predicts survival outcomes in bladder urothelial carcinoma patients. *Aging (Albany NY)* (2020) 12:15624–37. doi: 10.18632/aging.103718
39. Cai HJ, Zhuang ZC, Wu Y, Zhang YY, Liu X, Zhuang JF, et al. Development and validation of a ferroptosis-related lncRNAs prognosis signature in colon cancer. *Bosn. J Basic Med Sci* (2021) 21:569–76. doi: 10.17305/bjbm.2020.5617
40. Fu XZ, Zhang XY, Qiu JY, Zhou X, Yuan M, He YZ, et al. Whole-transcriptome RNA sequencing reveals the global molecular responses and ceRNA regulatory network of mRNAs, lncRNAs, miRNAs and circRNAs in response to copper toxicity in ziyang xiangcheng (Citrus junos sieb. ex Tanaka). *BMC Plant Biol* (2019) 19:509. doi: 10.1186/s12870-019-2087-1
41. Du Z, Chai X, Li X, Ren G, Yang X, Yang Z. Nano-CuO causes cell damage through activation of dose-dependent autophagy and mitochondrial lncCyt b-AS/ND5-AS/ND6-AS in SH-SY5Y cells. *Toxicol Mech Methods* (2022) 32:37–48. doi: 10.1080/15376516.2021.1964665
42. Weingarten MA, Feuchtwanger D. Group meetings with new mothers in a family practice: report of a pilot project. *Isr Ann Psychiatr Relat Discip.* (1978) 16:232–42.
43. Burr ML, Sparbier CE, Chan YC, Williamson JC, Woods K, Beavis PA, et al. CMTM6 maintains the expression of PD-L1 and regulates anti-tumour immunity. *Nature* (2017) 549:101–5. doi: 10.1038/nature23643



OPEN ACCESS

EDITED BY

Tianyi Zhao,
Harbin Institute of Technology, China

REVIEWED BY

Ningyi Zhang,
Harbin Institute of Technology, China
Yuansong Zhao,
University of Texas Health Science
Center at Houston, United States

*CORRESPONDENCE

Yu Liu
liuyusph@sina.com
Jiying Wang
wangjiyingsph@sina.com

†These authors share first authorship

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 29 June 2022

ACCEPTED 25 July 2022

PUBLISHED 20 September 2022

CITATION

Cai Y, Wu Q, Chen Y, Liu Y
and Wang J (2022) Predicting
non-small cell lung cancer-related
genes by a new network-based
machine learning method.
Front. Oncol. 12:981154.
doi: 10.3389/fonc.2022.981154

COPYRIGHT

© 2022 Cai, Wu, Chen, Liu and Wang.
This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original author
(s) and the copyright owner(s) are
credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Predicting non-small cell lung cancer-related genes by a new network-based machine learning method

Yong Cai^{1†}, Qiongya Wu^{1†}, Yun Chen^{1†}, Yu Liu^{1*}
and Jiying Wang^{2*}

¹Department of Radiation Oncology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China, ²Department of Oncology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China

Lung cancer is the leading cause of cancer death globally, killing 1.8 million people yearly. Over 85% of lung cancer cases are non-small cell lung cancer (NSCLC). Lung cancer running in families has shown that some genes are linked to lung cancer. Genes associated with NSCLC have been found by next-generation sequencing (NGS) and genome-wide association studies (GWAS). Many papers, however, neglected the complex information about interactions between gene pairs. Along with its high cost, GWAS analysis has an obvious drawback of false-positive results. Based on the above problem, computational techniques are used to offer researchers alternative and complementary low-cost disease–gene association findings. To help find NSCLC-related genes, we proposed a new network-based machine learning method, named deepRW, to predict genes linked to NSCLC. We first constructed a gene interaction network consisting of genes that are related and irrelevant to NSCLC disease and used deep walk and graph convolutional network (GCN) method to learn gene–disease interactions. Finally, deep neural network (DNN) was utilized as the prediction module to decide which genes are related to NSCLC. To evaluate the performance of deepRW, we ran tests with 10-fold cross-validation. The experimental results showed that our method greatly exceeded the existing methods. In addition, the effectiveness of each module in deepRW was demonstrated in comparative experiments.

KEYWORDS

lung cancer, computational techniques, deep walk, graph convolutional network, deep neural network

1 Introduction

Lung cancer continues to be the primary cause of cancer deaths worldwide, causing 1.8 million fatalities annually (1). The two primary kinds of lung cancer are small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). Additionally, nearly 85% of all cases of lung cancer are related to NSCLC (2). More and more researchers found that lung cancer is highly inherited and is associated with certain genes that increase the risk (3).

Genome-wide association studies (GWAS) are a common method to mine disease-related genes. Hung et al. (4) firstly used GWAS and found a locus in chromosome region 15q25 that related to lung cancer. Hu et al. (5) reported that 5p15 locus is related to lung cancer *via* GWAS, and 6p21 was found by Wang et al. (6). With the development of next-generation sequencing (NGS), whole-exome sequencing (WES), whole-genome sequencing (WGS), and other technologies are applied to find disease-related genes. Sun et al. (7) applied WES on 73 advanced NSCLC tumor samples and demonstrated Protein tyrosine phosphatase receptor type D (PTPRD) might be both a prognostic and a predictive biomarker predicting clinical outcomes in non-squamous (ns)-NSCLC patients. Liu et al. (8) found infrequent detrimental mutations in GWAS-nominated sites in dopamine β -hydroxylase (DBH) and coiled-coil domain containing 147 (CDC147) *via* WES.

With the explosive growth of relevant information and data in recent years, GWAS and other methods become more and more time-consuming and laborious. Many studies have focused on drug–disease association tasks and other bioinformatics tasks through machine learning and deep learning methods (9–13). Graph neural network methods that can integrate multiple types of knowledge bases are suitable for this task. (14) used graph convolutional network (GCN) to capture structural information from the network integrating gene and disease. GCN (15) is one type of neural network architecture to learn nodes and edges of graphs. It has been proven that GCN enhances algorithms of

abilities to mine information and make decisions in the bioinformatics field like Deep-DRM (16).

Graph embedding methods are popular in this task. Xiong et al. (17) built a heterogeneous network that incorporates different type datasets and obtained network representation by random walk (RW) to predict gene–disease associations. RW is a common graph embedding approach. This approach has been used to research microRNAs (miRNAs) (18), gene expression (19), and drug repositioning (20). Deep walk (21) is a graph structure data-mining algorithm that combines RW and work2vec. Zhu et al. (22) integrated graph embedding representation and GCN to learn the gene–disease associations. They connected the two methods in series as the encoder to learn features and predicted associations by a decoder.

In the paper, we focused on the problem of mining NSCLC-causing genes. We treated it as a binary classification and proposed a new network-based method. We integrated two types of graph embedding method, deep walk and GCN, to represent the gene interaction network and learn the features and used DNN to predict which genes are related to NSCLC.

2 Methods

We proposed a new method named deepRW based on the gene interaction network to predict NSCLC-related genes. The structure of our method is shown in Figure 1. First, we built a graph network that represented the interactions between genes. Then, we utilized two types of graph embedding method, deep walk and GCN, to learn network information and extract features. Last, we constructed a DNN module to predict disease-related genes.

2.1 Construction of the gene network

The network of gene interactions is represented as a graph network. The graph network we built can be expressed as $G = (V, E)$.

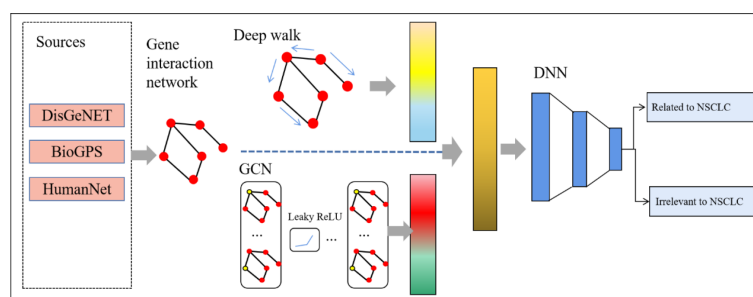


FIGURE 1

The structure of deepRW. GCN, graph convolutional network; DNN, deep neural network.

E) V represents the genes that we selected related to NSCLC; E represents the interactions between genes. It should be emphasized that outliers that did not interact with other genes were eliminated.

2.2 Network representation by deep walk and graph convolutional network

After we obtained the gene interaction network $G = (V, E)$, we used two graph embedding methods to learn the representations of vertices.

2.2.1 Network representation by deep walk

Deep walk uses randomness to produce the sequences of vertices $[v_1, v_2, \dots, v_n]$, where v_{i+1} is a vertex picked at random from the neighbors of vertex v_i and the likelihood of choosing each neighbor is proportional to the weight of the edge in the adjacency matrix that corresponds to it. In the paper, we were able to build sequences at each vertex by using deep walk.

Skip-gram (23) was used to train on the sequences of the vertices by sliding window sampling. Deep walk is actually a combination of RW and skip-gram. RW is responsible for sampling to obtain the co-occurrence relationship between nodes in the graph. Skip-gram trains the embedding vectors of nodes from the relationship. After training, we can get the embedding representation vectors and the probability distribution of the vertices. A representation vector optimizes the conditional probability $P(v_c/v_i)$, where v_c is the vertex that is in the context window of v_i . The loss function of training is:

$$L_{vi} = -\log P(v_{c1}, v_{c2}, \dots, v_{cW} | v_i) = -\log \prod_{j=1}^W P(v_{cj} | v_i) \quad (1)$$

where W represents the window size.

2.2.2 Network representation by graph convolutional network

The other graph embedding method we used is GCN. GCN used the graph network to learn node and edge information of the graph. Compared with deep walk, GCN can not only learn the structure of each node and its neighborhood but also integrate the characteristics of each node into it. If A is the adjacency matrix, the Laplacian matrix is:

$$L = D - A \quad (2)$$

where D means the degree matrix of the network. Since the features of genes should contain not only connections between nodes but also the information itself. So we can get:

$$A' = A + I \quad (3)$$

where I is the identity matrix. Then, the inverse degree matrix D' can be obtained.

$$D' = \sum A' \quad (4)$$

Last, we can get the features as follows:

$$X' = \sigma(D'^{\frac{1}{2}} A' D'^{\frac{1}{2}} X) \quad (5)$$

where X is the feature vector of each vertex, and σ is the activation function. In the study, we used Leaky Rectified Linear Unit (Leaky ReLU) function (24) as the activation function. This activation function may reduce the likelihood of vanishing gradients and boost feature sparsity when compared to other activations. The formula is as follows:

$$\text{LeakyReLU}(x) = \max(0, x) + 0.2 \min(0, x) \quad (6)$$

Two feature vectors of each vertex were generated *via* deep walk and GCN. Then, two feature vectors were fused and delivered to the prediction module.

2.3 Network prediction by deep neural network

To increase the quality of features and determine whether or not the gene is related to NSCLC, we employed a DNN module after network representation by deep walk. Whether there is a linear or non-linear connection between the input and the output, DNN can determine the appropriate mathematical operation to convert the input into the output. Now, most classification methods are shallow structure algorithms, which have the disadvantages of limited representation ability of complex functions in the case of limited samples and calculation suits, and the generalization ability for complex classification problems is limited. Deep learning can realize complex function approximation by learning a deep non-linear network structure and represent the distributed representation of input data. DNN has stronger ability to abstract problems and can also simulate more complex models. The following formula may be used to determine the feature map that advances to the next layer:

$$\text{Output}^l = W^l \text{Input}^l + \text{Bias}^l$$

where Input is the input of the forward propagation, Output is the output, Bias is the bias of layer l , and W is the weight of the neurons. The output of each layer is then sent *via* an activation function, which boosts positive vectors and suppresses negative vectors from the previous layer. We still used Leaky ReLU as the activation function in the predicting module.

Figure 2 depicts the number of layers of the DNN module and the specific parameters of each layer. There are three layers in the DNN module. Identifying NSCLC-related genes is a binary classification task, so we applied softmax as the activation function of the output layer. We used binary cross-entropy as the loss function as follows:

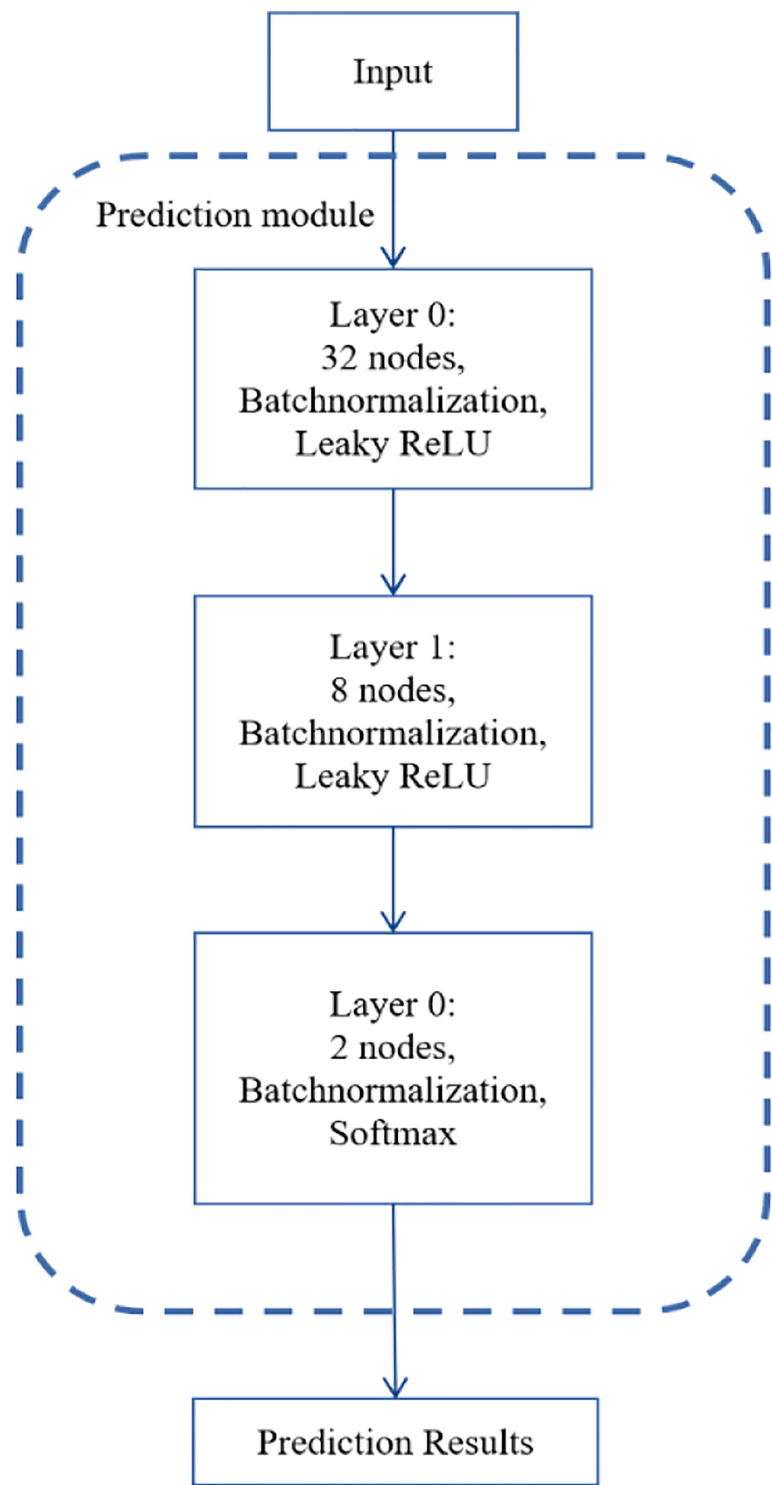


FIGURE 2
The structure of the DNN module.

$$\text{Loss} = -y_i \log(p_i) - (1 - y_i) \log(1 - p_i) \quad (7)$$

where y_i means the true value, p_i means the predicted value. Also, we used batch normalization (25) and early stop in the training process, which can end training if no improvement is shown after 50 epochs.

3 Materials

3.1 Dataset

DisGeNET (26) is used to obtain the gene–disease associations. DisGeNET is a database that contains information on the links between genes and disease. It is one of the biggest collections of genes and variants linked with human diseases. The data in DisGeNET come from a variety of sources, including expert-curated archives, catalogs of GWAS, animal models, and published scientific articles.

HumanNet (27), a probabilistic functional gene database, is used to generate the gene–gene associations; each gene–gene association has a score that represents the probability of the association. The gene network can be expressed as $G^{\text{gg}} = (N^{\text{gg}}, E^{\text{gg}})$, where N^{gg} is the set of genes and E^{gg} is the association of genes. The adjacent matrix of G^{gg} is $W^{\text{gg}} \in R^{N \times N}$, where $W_{ij}^{\text{gg}} = W_{ji}^{\text{gg}} = w$, w is the weight of each association of genes provided by HumanNet (Supplementary Table 1).

In the paper, we found 142 genes linked to NSCLC from DisGeNET, containing stage I, II, III, IIIA, and IIIB types (Supplementary table 1). These 142 genes were positive samples, and another 142 genes were randomly chosen that were reported to be irrelevant to the NSCLC disease. We used gene expression of tissues as the gene features from BioGPS (28).

3.2 Experimental setup

To demonstrate the performance of deepRW, we utilized 10-fold cross-validation to repeat experiments 10 times. The dataset is separated into 10 subsets, in every time experiment, we randomly choose one subset as the test samples, and the others as the train samples. The precision-recall curve (AUPR) and area under the ROC curve (AUROC) is used to evaluate the effectiveness of the methods.

In the training set, the main hyper parameters were set as follows: the window size of the Skip-gram was set to 10, and Skip-gram was trained for 10 iterations. The GCN of three layers and DNN module was trained 50 epochs, and early stopping and Adam with default parameters were used.

To demonstrate the effectiveness of our method, we tested the performance of our method by comparing the models listed as follows.

RWR: Random walk with restart (29) is used to capture relationships between two nodes and the overall structure information of the network by calculating the proximity between two nodes.

KBMF: Kernelized Bayesian matrix factorization (30), which always is used in recommender systems, can take advantage of many side information sources.

RF: Random forest (31) is a classifier that contains multiple decision trees, and its output is determined by the votes on individual trees.

4. Result

4.1 Performance of deep walk and graph convolutional network

First, we discussed the influence of the number of GCN layers. It is known that stacking too many layers into a GCN causes the vanishing gradient problem. This means that back-propagating through these networks leads to over-smoothing, eventually leading to features of graph vertices converging to the same value (32). We constructed the GCN module with two layers, three layers, and four layers. The results are shown in Table 1. From the results, GCN with three layers obtained the highest scores. Stacking four layers slightly reduced performance because of over-smoothing. In the paper, we built GCN with three layers as an encoder.

Then, we demonstrated the effectiveness of each module through the comparison trial on our method missing specific module. In “Without GCN,” we only used deep walk as the network representation module. In “Without deep walk,” we only used GCN. In “DNN,” we directly used DNN as the encoder and the decoder. Table 2 shows the results. Without GCN or deep walk, our method obtained worse scores. We can conclude that GCN and deep walk are important parts in deepRW, and integrating deep walk and GCN can improve the ability of learning the graph network.

4.2 Performance of different methods

To decrease the errors, we repeated the experiment 10 times and calculated the average scores as the final results. From the results in Table 3, we can find that deepRW outperformed all other methods in terms of AUROC and AUPR scores of 0.763 and 0.795. The RWR obtained the worst scores with AUROC and AUPR of 0.636 and 0.659, which are lower than deepRW by 16.64% and 17.11%. The results demonstrated that deepRW works better than a number of machine learning methods for locating NSCLC-related genes. GCN and deepRW are the

TABLE 1 The effectiveness of deep walk and GCN in deepRW.

Number of layers	AUROC	AUPR
Two layers	0.702	0.723
Three layers	0.763	0.795
Four layers	0.741	0.769

AUROC, area under the ROC curve; AUPR, The area under the precision recall curve.

TABLE 2 The AUROC and AUPR scores of different methods.

Method	AUROC	AUPR
DeepRW	0.763	0.795
KBMF	0.701	0.748
RF	0.647	0.697
RWR	0.636	0.659

DeepRW, Deep random walk; KBMF, Kernelized Bayesian matrix factorization; RF, Random forest; RWR, Random walk with restart.

methods that can extract feature information of nodes and edges. The results show that interactions between genes are helpful for enriching the characteristic information of genes. Compared with RWR, deep walk had better performance because deep walk combines RW and word2vec, which makes the algorithm easier to converge. Although deepRW obtained the best performance in the task, this method needs a long time to train and needs more train data to get better results.

5 Conclusion

Lung cancer is the leading cause of cancer death globally, and NSCLC is the main pathological subtype of lung cancer, accounting for about 85%. As the cost of sequencing continues to decrease and the amount of data continues to grow, GWAS and NGS as the main techniques to find disease-causing genes are time-consuming and laborious, and machine learning methods are getting more and more attention. In the paper, we proposed a new network-based method that is integrated with two different graph embedding methods to identify genes related to NSCLC. In order to learn about the relationships between genes and diseases, we first built a gene interaction network made up of both relevant and unrelated genes to the NSCLC disease. Then, we utilized deep walk and GCN to learn gene-disease interactions. Finally, DNN was constructed as the prediction module. This method concerns the gene network topology relationship and is conducive to mining genetic characteristics. We compared our method with several other methods and demonstrated better performance of our method.

We did case studies on new samples to verify the effectiveness of deepRW. We found that tumor protein p63 (TP63) is related to NSCLC. Gürgeç et al. (33) found that TP63 expression values were higher than the predefined cutoff of 12 in

23 NSCLC tumors with squamous cell carcinoma histology. general transcription factor IIH subunit 4(GTF2H4) was also found and supported by Wang et al. (34) who reported that GTF2H4 is associated with lung cancer risk.

Compared with machine learning methods, deepRW as a deep learning method needs more time and more samples to train to obtain better performance. In the future, we will study the ability of deepRW to identify other pathogenic genes.

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

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

Author contributions

YCa, YL, and JW conceived and designed study, collected and analyzed data. QW and YCh statistical analyses. YCa, YL, and YCh drafted and edited manuscript. All authors contributed to the article and approved the submitted version.

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/fonc.2022.981154/full#supplementary-material>

TABLE 3 The AUROC and AUPR scores of different methods.

Method	AUROC	AUPR
DeepRW	0.763	0.795
KBMF	0.701	0.748
RF	0.647	0.697
RWR	0.636	0.659

DeepRW, Deep random walk; KBMF, Kernelized Bayesian matrix factorization; RF, Random forest; RWR, Random walk with restart.

References

- Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* (2018) 34:i457–66. doi: 10.1093/bioinformatics/bty294
- Navada S, Lai P, Schwartz A, Kalemkerian G. Temporal trends in small cell lung cancer: Analysis of the national surveillance, epidemiology, and end-results (SEER) database. *J Clin Oncol* (2006) 24:7082–2. doi: 10.1200/jco.2006.24.18_suppl.7082
- Matakidou A, Eisen T, Houlston R. Systematic review of the relationship between family history and lung cancer risk. *Br J Cancer* (2005) 93:825–33. doi: 10.1038/sj.bjc.6602769
- Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* (2008) 452:633–7. doi: 10.1038/nature06885
- Hu Z, Wu C, Shi Y, Guo H, Zhao X, Yin Z, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat Genet* (2011) 43:792–6. doi: 10.1038/ng.875
- Wang Y, Broderick P, Webb E, Wu X, Vijaykrishnan J, Matakidou A, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* (2008) 40:1407–9. doi: 10.1038/ng.273
- Sun Y, Duan J, Fang W, Wang Z, Du X, Wang X, et al. Identification and validation of tissue or ctDNA PTPRD phosphatase domain deleterious mutations as prognostic and predictive biomarkers for immune checkpoint inhibitors in non-squamous NSCLC. *BMC Med* (2021) 19:1–19. doi: 10.1186/s12916-021-02075-5
- Liu Y, Kheradmand F, Davis CF, Scheurer ME, Wheeler D, Tsavachidis S, et al. Focused analysis of exome sequencing data for rare germline mutations in familial and sporadic lung cancer. *J Thorac Oncol* (2016) 11:52–61. doi: 10.1016/j.jtho.2015.09.015
- Rao A, Vg S, Joseph T, Kotte S, Sivadasan N, Srinivasan R. Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. *BMC Med Genomics* (2018) 11:1–12. doi: 10.1186/s12920-018-0372-8
- Han P, Yang P, Zhao P, Shang S, Liu Y, Zhou J, et al. GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization, in: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 705–13. Available at: <https://dblp.org/rec/conf/kdd/HanYZSLZ0K19.html>
- Wang X, Gong Y, Yi J, Zhang W. (2019). Predicting gene-disease associations from the heterogeneous network using graph embedding, in: *2019 IEEE international conference on bioinformatics and biomedicine (BIBM): IEEE*, pp. 504–11. Available at: <http://ieeexplore.ieee.org/BIBM2019/AcceptedPapers.html>
- Zhao T, Liu J, Zeng X, Wang W, Li S, Zang T, et al. Prediction and collection of protein-metabolite interactions. *Briefings Bioinf* (2021) 22:bbab014. doi: 10.1093/bib/bbab014
- Cheng N, Chen C, Li C, Huang J. Inferring cell-type-specific genes of lung cancer based on deep learning. *Curr Gene Ther* (2022), 1–6. doi: 10.2174/156652322666220324110914
- Li Y, Kuwahara H, Yang P, Song L, Gao XJB. PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks. *bioRxiv* (2019) 532226. doi: 10.1101/532226
- Kipf TN, Welling M. (2016)., *Semi-supervised classification with graph convolutional networks*, Published as a conference paper at ICLR 2016. Available at: <https://arxiv.org/abs/1609.02907>
- Zhao T, Hu Y, Cheng L. Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief Bioinform* (2020) bbab212. doi: 10.1093/bib/bbaa212
- Xiong Y, Guo M, Ruan L, Kong X, Tang C, Zhu Y, et al. Heterogeneous network embedding enabling accurate disease association predictions. *BMC Med Genomics* (2019) 12:1–17. doi: 10.1186/s12920-019-0623-3
- Yu L, Shen X, Zhong D, Yang J. Three-layer heterogeneous network combined with unbalanced random walk for miRNA-disease association prediction. *Front Genet* (2019) 10:1316. doi: 10.3389/fgene.2019.01316
- Zhao T, Lyu S, Lu G, Juan L, Zeng X, Wei Z, et al. SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res* (2021) 49:D1413–9. doi: 10.1093/nar/gkaa838
- Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F. deepDR: A network-based deep learning approach to *in silico* drug repositioning. *Bioinformatics* (2019) 35:5191–8. doi: 10.1093/bioinformatics/btz418
- Perozzi B, Al-Rfou R, Skiena S. (2014). Deepwalk: Online learning of social representations, in: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* pp. 701–10. doi: 10.1145/2623330.2623732
- Zhu L, Hong Z, Zheng H. (2019). Predicting gene-disease associations via graph embedding and graph convolutional networks, in: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): IEEE*, pp. 382–9. Available at: <http://ieeexplore.ieee.org/BIBM2019/AcceptedPapers.html>
- Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. (2013). arXiv:1301.3781v3.
- Xu B, Wang N, Chen T, Li M. Empirical evaluation of rectified activations in convolutional network. (2015). <https://arxiv.org/abs/1505.00853>.
- Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *OALib Journal* (2015) 3: 448–456.
- Piñero J, Bravo A, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* (2016), 45(D1): D833–D839.
- Hwang S, Kim CY, Yang S, Kim E, Hart T, Marcotte EM, et al. HumanNet v2: human gene networks for disease research. *Nucleic Acids Res* (2019) 47:D573–80. doi: 10.1093/nar/gky1126
- Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* (2009) 10:1–8. doi: 10.1186/gb-2009-10-11-r130
- Tong H, Faloutsos C, Pan J-Y. (2006). Fast random walk with restart and its applications, in: *Sixth international conference on data mining (ICDM'06): IEEE*, pp. 613–22. Available at: <https://ieeexplore.ieee.org/document/4053087>
- Gönen M, Kaski S. Kernelized Bayesian Matrix Factorization. *IEEE Trans Pattern Anal Mach Intell* (2014) 36(10):2047–60. doi: 10.1109/TPAMI.2014.2313125
- Breiman L. Random forests. *Mach Learn* (2001) 45:5–32. doi: 10.1023/A:1010933404324
- Li G, Muller M, Thabet A, Ghanem B. (2019). Deepgcns: Can gcns go as deep as cnns?, in: *Proceedings of the IEEE/CVF international conference on computer vision* pp. 9267–76. Available at: <https://dblp.uni-trier.de/rec/conf/iccv/Li0TG19.html>
- Gürgeç D, Conrad T, Becker M, Sebens S, Röcken C, Hoffmann J, et al. breaking the crosstalk of the cellular tumorigenic network by low-dose combination therapy in lung cancer patient-derived xenografts. *Commun Biol* (2022) 5:1–10.
- Wang M, Liu H, Liu Z, Yi X, Heike B, Hung RJ, et al. Genetic variant in DNA repair gene GTF2H4 is associated with lung cancer risk: A large-scale analysis of six published GWAS datasets in the TRICL consortium. *Carcinogenesis* (2016), 37(9):888–896.



OPEN ACCESS

EDITED BY
Tianyi Zhao,
Harbin Institute of Technology, China

REVIEWED BY
Quan Zou,
University of Electronic Science and
Technology of China, China
Junwei Han,
Harbin Medical University, China

*CORRESPONDENCE
Zhufang Kuang
zfkuan@163.com

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

RECEIVED 06 July 2022
ACCEPTED 14 September 2022
PUBLISHED 29 September 2022

CITATION
Duan T, Kuang Z and Deng L (2022)
SVMMMDR: Prediction of miRNAs-drug
resistance using support vector
machines based on
heterogeneous network.
Front. Oncol. 12:987609.
doi: 10.3389/fonc.2022.987609

COPYRIGHT
© 2022 Duan, Kuang and Deng. This is
an open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use,
distribution or reproduction is
permitted which does not comply
with these terms.

SVMMMDR: Prediction of miRNAs-drug resistance using support vector machines based on heterogeneous network

Tao Duan, Zhufang Kuang* and Lei Deng

School of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha, China

In recent years, the miRNA is considered as a potential high-value therapeutic target because of its complex and delicate mechanism of gene regulation. The abnormal expression of miRNA can cause drug resistance, affecting the therapeutic effect of the disease. Revealing the associations between miRNAs-drug resistance can help in the design of effective drugs or possible drug combinations. However, current conventional experiments for identification of miRNAs-drug resistance are time-consuming and high-cost. Therefore, it's of pretty realistic value to develop an accurate and efficient computational method to predicting miRNAs-drug resistance. In this paper, a method based on the Support Vector Machines (SVM) to predict the association between MiRNA and Drug Resistance (SVMMMDR) is proposed. The SVMMMDR integrates miRNAs-drug resistance association, miRNAs sequence similarity, drug chemical structure similarity and other similarities, extracts path-based Hetesim features, and obtains inclined diffusion feature through restart random walk. By combining the multiple feature, the prediction score between miRNAs and drug resistance is obtained based on the SVM. The innovation of the SVMMMDR is that the inclined diffusion feature is obtained by inclined restart random walk, the node information and path information in heterogeneous network are integrated, and the SVM is used to predict potential miRNAs-drug resistance associations. The average AUC of SVMMMDR obtained is 0.978 in 10-fold cross-validation.

KEYWORDS

miRNA, drug resistance, support vector machines, hetesim score, random walk with restart

1 Introduction

In recent years, the difficulty of drug target selection has led to the increase of drug development cost and the low efficiency of pharmaceutical industry. So far, it has been discovered that the human genome can encode up to 25,000 genes. But only 600 of the disease-causing proteins have targeted drugs, meaning a significant number are

“undrugable”. Therefore, the focus of target selection has now shifted to other macromolecules, such as non-coding RNA.

According to the genetic central dogma, the DNA is a carrier that carries genetic information. During growth and development, the genetic information in DNA is transcribed into RNA. Then the RNA is translated into various proteins to perform specific biological functions. There are many types of RNAs with complex functions. Research shows that only 2% of the RNA could code for proteins, and 98% couldn't. In biology, RNAs with non-coding are called non-coding RNAs (ncRNAs). Among ncRNAs, miRNA is a group of non-coding RNAs encoded by the genome with a length of about 20–23 nucleotides. The miRNAs play a major role in gene expression regulation. They have a significant meaning in many biological processes such as cell differentiation, development and cellular signaling pathways.

The miRNAs play an important role in the understanding of life sciences. The miRNAs are significant in many aspects such as cellular biological processes, gene expression regulation at transcriptional and post-transcriptional levels, and others.

There are many studies on the biological mechanisms and interactions between genes, miRNAs, lncRNAs, diseases and drugs, such as the relationship between genes and diseases, miRNAs and diseases, lncRNAs and diseases, miRNAs and lncRNAs, etc.

For the association between genes and diseases, a network impulse dynamics framework based on multiple biological networks NIDM is proposed by Xiang et al. (1) to predict potential disease-gene associations. The HyMM is proposed by Xiang et al. (2) to more effectively predict disease-related genes by integrating information from the structure of multi-scale modules. The PrGeFNE is proposed by Xiang et al. (3) based on fast network embedding. An understanding of the association between genetics and disease can help understand the pathogenesis of disease.

For the association between miRNAs and diseases, a meta-path-based MDPBMP is proposed by Yu et al. (4). The information carried by the nodes is extracted and integrated through MDPBMP, and the miRNAs-disease association is predicted using embedded feature vectors. The VGAE-MDA, a deep learning framework with variational graph autoencoder, is proposed by Ding et al. (5). The MLPMDA, a miRNAs-disease association prediction method using multilayer linear projection, is proposed by Guo et al. (6). The prediction method GRPAMDA is proposed by Zhong et al. (7). The GRPAMDA combines the graph random propagation network based on DropFeature and attention network. The NIMGSA is proposed by Jin et al. (8) to predict miRNAs-disease association based on neural induction matrix completion. An ensemble learning framework with resampling method for miRNA-disease association ERMDA prediction is proposed by Dai et al. (9). A double random walk model is proposed by Zhu et al. (10). The end-to-end deep learning method PDMDA is proposed by Yan et al. (11). A

computational framework SMALF based on XGBoost is proposed by Liu et al. (12). The algorithm MSCDE is proposed by Han et al. (13) based on a variety of biological source information. The method based on tensor factorization and label propagation (TFLP) is proposed by Yu et al. (14) for multi-type miRNA-disease association prediction.

For the association between lncRNAs and diseases, a non-negative matrix factorization based on graph regularization LDGRNMF is proposed by Wang et al. (15) to predict the lncRNAs-disease association. The internal inclined random walk with restart (IIRWR) is used by Wang et al. (16) to infer potential lncRNA-disease associations. A lncRNAs-disease association prediction method GBDTLRL2D based on Gradient Boosting Decision Tree and Logistic Regression is proposed by Duan et al. (17). The GCRFLDA, a prediction method based on graph convolution matrix completion, is proposed by Fan et al. (18). An end-to-end computational method based on graph attention network GANLDA is proposed by Lan et al. (19). A method called LRWRHLDA is proposed by Wang et al. (20). The LRWRHLDA designs a multi-layer network using six known heterogeneous networks, and uses Laplace normalized random walk and restart algorithm to predict. A dual attention network is proposed by Liu et al. (21).

For the association between miRNAs and lncRNAs, the LMI-INGI, based on interactome network and graphlet interaction, is proposed by Zhang et al. (22) to predict the lncRNAs-miRNAs associations. The NALMA is proposed by Zhang et al. (23) to use network distance analysis. The DWLMI proposed by Yang et al. (24) utilizes lncRNAs-miRNAs-disease-protein-drug diagram. The structural perturbation method SPMLMI is proposed by Xu et al. (25). A logical matrix factorization with neighborhood regularized, LMFNRLMI, is proposed by Liu et al. (26).

Advances in genomics and bioinformatics have facilitated the identification of miRNAs. The miRNAs have also been found to interact with a variety of drugs. It is possible to develop resistance or sensitivity during drug treatment because of the regulation of genes by miRNAs (27). For example, scientists have found that miRNA let-7b is resistant to the drug cisplatin (28). Cisplatin is an important drug in the treatment of many diseases, such as sarcoma. Cisplatin has also been reported to down-regulate miRNA let-7b expression, lead to up-regulation of Cyclin D1, and induce resistance to cisplatin. Similarly, miRNA Mir-106a is found to enhance the sensitivity of OVCAR3/CIS cells to cisplatin (29). Since both the increase and decrease of miRNA expression level can cause diseases, miRNA-targeted therapy drugs can be divided into miRNA mimics and miRNA inhibitors. Their aim is to induce gene silencing and selective up-regulation of proteins.

There are several public databases that collate miRNAs-drug relationships. For instance, the database of miRNAs-drug interactions, pharmaco-miR is developed by Rukov et al. (30) according to the interaction between miRNA target genes and

drug proteins. The database mTD is developed by Chen et al. (31) to collect information about the impact of miRNAs during drug treatment. The ncDR is developed by Dai et al. (32) to provide information of noncoding RNAs related to drug resistance. However, the known link between miRNAs and drug resistance is limited. Because biological experiments are time-consuming and expensive, it is necessary to develop computation-based methods to predict the potential association between miRNAs and drug resistance.

Different computational methods have been developed to identify and predict miRNAs-drug resistance. For example, an algorithm for predicting potential miRNAs-drug resistance associations through Bi-Random Walk (BiRW) is proposed by Xu et al. (33). The method SNMFSSMA based on symmetric non-negative matrix factorization and kronecker regularized least squares is developed by Zhao et al. (34) for prediction of small molecular-miRNAs association. The GCMR proposed by Huang et al. (35) uses graph convolution to built potential factor model, learns the graph embedding feature of miRNAs and drugs, and expresses the problem of predicting miRNA-drug association as a link prediction problem involving two-miRNA-drug sensitivity associations, named LGCMDS, is proposed by Yu et al. (36). The MDIPA, a matrix factor-based method, is proposed by Jamali et al. (37) to predict the unknown interactions between miRNAs and drug resistance. Predicting associations between small molecular and microRNAs using functional similarity of miRNAs and multiple similarity measures of small molecular is proposed by Qu et al. (38). In addition, combined with clinical, chemical, and biological information, a method based on non-negative matrix factorization to predict miRNAs-small molecule relationships is developed by Luo et al. (39).

Although there are some studies on predictive tools for miRNAs-drug resistance associations, these methods cannot fully utilize the structure and semantics in heterogeneous networks to extract higher-quality information. At the same time, the accuracy and performance obtained by these methods need to be improved. To address these issue, a method for predicting miRNAs-drug resistance based on support vector machines SVMMDR is proposed in this paper. The SVMMDR considers the path information between nodes in heterogeneous networks. The hetesim measures the correlation between nodes of the same type or different types within a unified framework. At the same time, based on the search path between two nodes, the measure between node pairs is defined by following a sequence. The node information and path information in heterogeneous networks are integrated. The SVM is used to predict potential miRNAs-drug resistance associations. The contribution of our method mainly consists of the following parts:

- The SVMMDR introduces the concept of miRNA and drug groups. On this basis, a roaming network is

established. Walker is more inclined to choose the next node of the walk. The inclined diffusion feature is obtained by inclined restart random walk.

- The SVMMDR integrates node information and path information in heterogeneous networks. The data feature is obtained by combining the inclined diffusion feature and hetesim score.
- The SVMMDR algorithm improves prediction accuracy and has the highest AUC values when compared to existing algorithms.

2 Materials and methods

The miRNAs-drug resistance association data required in this paper are downloaded from ncDR database (32). The ncDR collected 5864 validated relationships between 145 compounds and 1039 ncRNAs (877 miRNAs and 162 lncRNAs) from approximately 900 published papers. We only need the correlation between miRNAs-drug resistance among them. After removing duplicate data, the 625 miRNAs, 85 drugs and 2301 miRNAs-drug resistance associations are obtained.

In this paper, an SVM-based method SVMMDR is proposed to predict the association between miRNAs-drug resistance. The SVMMDR integrates miRNAs-drug resistance association, miRNAs sequence similarity, drugs chemical structure similarity and other similarities. The path-based hetesim feature is obtained, and the concepts of miRNAs group and drug group are introduced to obtain the inclined diffusion feature through inclined random walk. Finally, the SVM algorithm is used to predict the association between miRNAs and drug resistance. This mainly includes the following steps:

- (1) The miRNAs-drug resistance association data set is downloaded from the ncDR, and the list of miRNAs and drugs, the matrix *A* of miRNAs-drug resistance association are obtained by de-duplication of the downloaded data. Then the gaussian interaction profile kernel similarity matrix of miRNAs GSM and of drug GSD are calculated.
- (2) The sequence of miRNA list is downloaded from miRBase database, and the miRNAs sequence similarity matrix SSM between miRNAs is calculated. The drug chemical structure similarity matrix ESD is obtained by using the published tool SimComp.
- (3) The miRNAs similarity network is obtained based on the GSM and SSM, and drugs similarity network is obtained based on the GSD and ESD.
- (4) The miRNA-resistance association network, miRNA similarity network, and drug similarity network are integrated to construct a heterogeneous network. In

the heterogeneous networks, the inclined diffusion feature are obtained based on the inclined random walk with restart. Then the low-dimensional inclined diffusion feature are obtained by using Singular Value Decomposition (SVD).

- (5) The hetesim scores for miRNA-drug pairs are calculated based on paths in the heterogeneous network.
- (5) The inclined diffusion feature and the hetesim score are combined to obtain the feature data set. The combined features are used in the SVM classifier to obtain the predicted scores for miRNAs-drug resistance. The flow of SVMMDR is shown in Figure 1.

2.1 Calculate Gaussian interaction profile kernel similarity

The matrix A of miRNAs-drug resistance association network is obtained. The number of rows of A is the number of miRNAs, and the number of columns of A is the number of drugs, as shown in the formula (1):

$$A(m_i, d_j) = \begin{cases} 1 & m_i \text{ is associated with } d_j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Where $A(m_i, d_j) = 1$ indicates that there is a resistance between miRNA m_i and drug d_j .

For any given miRNA m_i and m_j , the gaussian interaction profile kernel similarity $GSM(m_i, m_j)$ can be obtained based on A , as shown in the formula (2) and (3):

$$GSM(m_i, m_j) = \exp(-\delta_m \|A(i, :) - A(j, :)\|^2) \quad (2)$$

$$\delta_m = \delta'_m / \left(\frac{1}{nm} \sum_{i=1}^{nm} \|A(i, :)\|^2 \right) \quad (3)$$

where nm is the number of miRNAs and $A(i, :)$ is the i_{th} row of the adjacency matrix A . The δ_m is used to control the frequency band, it represents the normalized frequency band of Gaussian interaction profile kernel similarity based on the new frequency band parameter δ'_m . The gaussian interaction profile kernel similarity between drugs can be obtained in the same way, represented by GSD, which is given by (4) and (5):

$$GSD(d_x, d_y) = \exp(-\delta_d \|A(:, x) - A(:, y)\|^2) \quad (4)$$

$$\delta_d = \delta'_d / \left(\frac{1}{nd} \sum_{x=1}^{nd} \|A(:, x)\|^2 \right) \quad (5)$$

where nd is the number of miRNAs and $A(:, x)$ is the x_{th} col of the A .

2.2 Calculate miRNA sequence similarity and drug chemical structure similarity

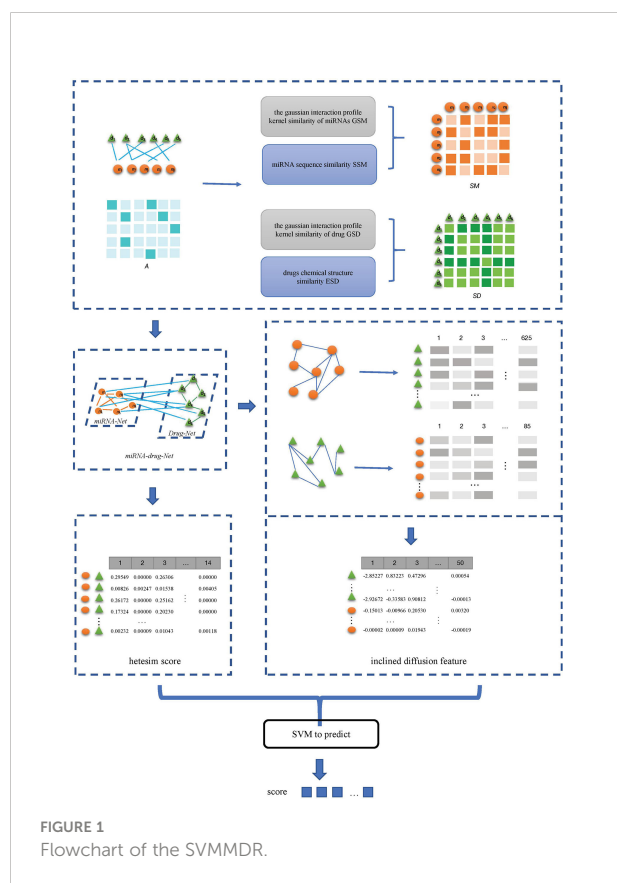
The sequences of relevant miRNAs are downloaded from the public database miRBase (<https://mirbase.org/>) (40). The miRBase database provides information including miRNAs sequence data, annotations, and predicted gene targets. The sequence similarity SSM between miRNAs is calculated as shown in the formula (6):

$$SSM(m_i, m_j) = 1 - \frac{\text{Levenshtein}(m_i, m_j)}{\text{len}(m_i) + \text{len}(m_j)} \quad (6)$$

$$0 \leq \text{Levenshtein}(m_i, m_j) \leq \text{len}(m_i) + \text{len}(m_j) \quad (7)$$

Where $\text{len}(m_i)$ represents the length of miRNA m_i sequence, $\text{len}(m_j)$ represents the length of miRNA m_j sequence, $\text{Levenshtein}(m_i, m_j)$ is defined as the class editing distance of the transformation from m_i sequence to m_j sequence.

With the Kyoto Encyclopedia of Genes and Genomes (KEGG) database entry number corresponding to drugs in the DLREFD database as the parameter, the chemical structural



similarity matrix ESD between drugs is calculated using the SimComp tool (41).

2.3 Integer similarity

In this section, miRNAs similarity network and drugs similarity network are constructed. The miRNAs similarity network is expressed as SM. The SM is fused by SSM and GSM, which is given by (8):

$$SM(m_i, m_j) = \begin{cases} GSM(m_i, m_j) & \text{if } SSM(m_i, m_j) = 0 \\ SSM(m_i, m_j) & \text{otherwise} \end{cases} \quad (8)$$

Similarly, denote SD as the drug similarity network, which is fused by ESD and GSD, as follows:

$$SD(d_i, d_j) = \begin{cases} GSD(d_i, d_j) & \text{if } SSM(d_i, d_j) = 0 \\ SSM(d_i, d_j) & \text{otherwise} \end{cases} \quad (9)$$

2.4 Obtain low-dimensional network inclined diffusion features

A global heterogeneous network is constructed by integrating the association matrix A of miRNAs-drug resistance network, the similarity matrix SM of miRNA and the similarity matrix SD of drugs. The concepts of miRNAs group and drugs group are introduced to obtain miRNA weight matrix and drug weight matrix to construct roaming network. The restart random walk is used to calculate the inclined diffusion feature on the roaming network, and the high dimensional inclined diffusion feature are obtained. Then, the SVD is used to reduce the dimension of the high-dimensional inclined diffusion feature, and the low-dimensional inclined diffusion feature is obtained. The specific sub-steps are as follows:

2.4.1 Building a heterogeneous network

The heterogeneous network $G = (V, E)$ is constructed. The dimension of the matrix G is $(nm + nd) * (nm + nd)$, where nm and nd is the number of miRNAs and drugs, as shown in formula (10):

$$G = \begin{bmatrix} SM & A \\ A^T & SD \end{bmatrix} \quad (10)$$

where A^T is the transpose of A.

2.4.2 Obtain the weight matrix

The drugs associated with the same miRNA are regarded as a drug group. If one miRNA with high similarity to this miRNA are associated with a drug in this drug group, this miRNA is

considered to have a potential association with other drugs in the drug group.

For example, for drug d_i , miRNAs associated with d_i are regarded as a miRNA group. If d_j with high similarity to d_i is associated with miRNA in this miRNA group, then it is assumed that d_j may be associated with other miRNAs in the miRNA group. Based on the above assumptions, miRNAs weight matrix W_{MM} of $nd * nm$ dimension and drugs weight matrix W_{DD} of $nm * nd$ dimension are obtained, as shown in the formula (11) and (12):

$$W_{MM}(d_i, m_j) = \frac{SS_M(d_i, m_j)}{\max_{1 \leq i \leq nd} \{SS_M(d_i, m_j)\}} \quad (11)$$

$$SS_M(d_i, m_j) = \sum_{m_k \in DM(d_i)} SM(m_k, m_j) \quad (12)$$

where $DM(d_i) = \{m_k \mid \forall m_k \in \{if(A(m_k, d_i) = 1)\}, 1 \leq k \leq nm\}$ represents the miRNA group associated with the drug d_i . $if(A(m_k, d_i) = 1, 1 \leq k \leq nm)$ represents that miRNA m_k is associated with drug d_i . $SM(m_k, m_j)$ is the similarity between miRNA m_k and m_j .

The drugs weight matrix W_{DD} of $nm * nd$ dimension can also be obtained:

$$W_{DD}(m_j, d_i) = \frac{SS_D(m_j, d_i)}{\max_{1 \leq j \leq nm} \{SS_D(m_j, d_i)\}} \quad (13)$$

$$SS_D(m_j, d_i) = \sum_{d_z \in DD(m_j)} SD(d_z, d_i) \quad (14)$$

where $DD(m_j) = \{d_z \mid \forall d_z \in \{if(A(d_z, m_j) = 1)\}, 1 \leq z \leq nd\}$ represents the drug group associated with the drug m_j . $if(A(d_z, m_j) = 1, 1 \leq z \leq nd)$ represents that drug d_z is associated with drug m_j . $SD(d_z, m_j)$ is the similarity between drug d_z and m_j .

2.4.3 Construct roaming network

When drug d_x is a walker, it walks on the miRNAs node network. T_D is the transition probability matrix of the roaming network. For any given miRNA m_i and m_j , denote T_D as the probability of m_i transferring to m_j during the walking process.

$$T_D(m_i, m_j) = \frac{ST_D(m_i, m_j)}{\sum_{n=1}^{nm} ST_D(m_n, m_j)} \quad (15)$$

$$ST_D(m_i, m_j) = \begin{cases} W_{DD}(m_j, d_x) & \text{If } W_{DD}(m_j, d_x) > 0 \\ SM(m_i, m_j) & \text{otherwise} \end{cases} \quad (16)$$

where, m_i is the current node of migration, and m_j is the next node. If the value of m_j and d_x in the weight matrix is not 0, it means that m_j and d_x have potential correlation, namely $ST_D(m_i, m_j) = W_{DD}(m_j, d_x)$. Otherwise, the probability of m_i transferring to m_j is related to miRNA similar matrix, $ST_D(m_i, m_j) = SM(m_i, m_j)$.

When miRNA m_y is a walker, it walks on the miRNAs node network. Denote T_M as the transition probability matrix of the roaming network. For any given drug d_i and d_j , T_M represents the probability of d_i transferring to d_j during the walking process.

$$T_M(d_i, d_j) = \frac{ST_M(d_i, d_j)}{\sum_{n=1}^{nd} ST_M(d_n, d_j)} \quad (17)$$

$$ST_M(d_i, d_j) = \begin{cases} W_{MM}(d_j, d_x) & \text{If } W_{MM}(d_j, m_y) > 0 \\ SD(d_i, d_j) & \text{otherwise} \end{cases} \quad (18)$$

2.4.4 Obtain inclined diffusion feature by IIRWR

Based on the transfer probability matrix T_D and T_M obtained, the drugs inclined diffusion feature $P_D = [P^1, P^2, P^3, \dots, P^x, \dots, P^{nd}]$ can be obtained by random walk, where P^x represents the nm -dimensional inclined diffusion feature of drug node d_x . Meanwhile, the miRNAs inclined diffusion feature $P_M = [P^1, P^2, P^3, \dots, P^y, \dots, P^{nm}]$, where P^y denotes the nd -dimensional inclined diffusion feature of miRNA node m_y . The nm and nd denote the number of miRNA nodes and drug nodes.

When the inclined diffusion feature P^x of drug node d_x is calculated, each step of the walking is faced with two choices: randomly selecting adjacent miRNA node or returning to the starting node. The walking process is shown in the equation (19):

$$P_{t+1}^x = (1 - r) \times T_D \times P_t^x + r \times P_0^x \quad (19)$$

$$P_0^x(m_i) = \frac{A(m_i, d_x)}{\sum_{n=1}^{nm} A(m_n, d_x)} \quad (20)$$

When the inclined diffusion feature P^y of miRNA node m_y is calculated, the walking process is shown as follows:

$$P_{t+1}^y = (1 - r) \times T_M \times P_t^y + r \times P_0^y \quad (21)$$

$$P_0^y(d_j) = \frac{A(m_y, d_j)}{\sum_{n=1}^{nd} A(m_y, d_n)} \quad (22)$$

Where r is the restart probability, P_t^x is a nd -dimensional transition probability vector of node m_y , and its k -th element represents the probability of accessing node k at t step, $k \in \{1, 2, \dots, nd\}$. P_0^y represents the initial migration probability vector of node m_y , and $P_0^y(d_j)$ represents the initial migration probability of m_y visiting node d_j .

After several iterations, the difference between the two iterations of p^x and p^y is less than 10^{-10} . The miRNA inclined diffusion feature P_M and the drug inclined diffusion feature P_D reach a stable state, and the final inclined diffusion feature is obtained.

2.4.5 Calculate the low-dimensional inclined diffusion feature

The more nodes in the heterogeneous network, the higher the feature dimension obtained by the inclined restart random walk. However, when the feature dimension is high, there will be data redundancy. The sample distribution of the high-dimension space is sparse. The SVD is used to reduce the dimension of the inclined diffusion feature.

Suppose that the $m \times n$ dimensional matrix P can be decomposed by $P = U\Sigma V^T$, where U is a $m \times m$ -dimensional matrix and V is a $n \times n$ -dimensional matrix. The U and V are left singular vectors and the right singular vectors, both unitary matrices, that is, $UU^T = I$, $VV^T = I$. The $m \times n$ dimensional matrix Σ has values only on the main diagonal, and all other elements are zero. Every element along the main diagonal is called singular value. The singular values are arranged from largest to smallest, and the decrease is extremely fast. In many cases, the sum of the first 10% or even 1% of the singular values accounts for more than 99% of the total singular values. In other words, we can also use the largest d singular values and corresponding left and right singular vectors to approximate the matrix, as follows:

$$P_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T \approx U_{m \times d} \Sigma_{d \times d} V_{d \times n}^T \quad (23)$$

where d is far less than n , and the low-dimensional feature vector X can be obtained by formula (24):

$$X = U_{m \times d} (\Sigma_{d \times d})^{1/2} \quad (24)$$

The SVD is performed on P_D and P_M respectively to obtain low-dimensional node feature matrix X_D and X_M .

2.5 Calculate the hetesim score

In heterogeneous networks, the types of nodes are different, and the relationship between nodes has various meanings. In order to obtain the correlation between different nodes, the hetesim scores are calculated (42). The hetesim is a path-based measurement method used to measure the correlation of objects (including objects of the same type or different types) in heterogeneous networks.

(1) The transition probability matrix I_{MD} from miRNA to drug, I_{DD} from drug to drug, I_{MM} from miRNA to miRNA are obtained as follows:

$$I_{MD}(m_i, d_j) = \frac{A(m_i, d_j)}{\sum_{k=1}^{nd} A(m_i, d_k)} \quad (25)$$

$$I_{DD}(d_i, d_j) = \frac{SD(d_i, d_j)}{\sum_{k=1}^{nd} SD(d_i, d_k)} \quad (26)$$

$$I_{MM}(m_i, m_j) = \frac{SM(m_i, m_j)}{\sum_{k=1}^{nm} SM(m_i, m_k)} \quad (27)$$

(2) The N is the node, and there are only miRNA and drug node. The path with length l between any two nodes is represented by $\rho = N_1 N_2 \dots N_{l+1}$, and the reachable probability matrix $PM = I_{N_1} N_2 I_{N_2} N_3 \dots I_{N_l} N_{l+1}$. Divide the path in half, get the PM_{ρ_L} and PM_{ρ_R} .

- when l is even, $\rho_L = N_1 N_2 \dots N_{\frac{l}{2}+1}$, $\rho_R = N_{\frac{l}{2}+1} N_{\frac{l}{2}+2} \dots N_{l+1}$. The PM_{ρ_L} and PM_{ρ_R} is calculated.
 - When l is odd, $\rho_{L_1} = N_1 N_2 \dots N_{\frac{l+1}{2}}$, $\rho_{L_2} = N_1 N_2 \dots N_{\frac{l+1}{2}}$, $\rho_{R_1} = N_{\frac{l+1}{2}} N_{\frac{l+1}{2}+1} \dots N_{l+1}$, $\rho_{R_2} = N_{\frac{l+1}{2}} N_{\frac{l+1}{2}+1} \dots N_{l+1}$.
- Then $PM_{\rho_L} = \frac{PM_{\rho_{L_1}} + PM_{\rho_{L_2}}}{2}$, $PM_{\rho_R} = \frac{PM_{\rho_{R_1}} + PM_{\rho_{R_2}}}{2}$.

(3) The PM_{ρ_L} and $PM_{\rho_R^{-1}}$ are calculated, where ρ_R^{-1} represents the reverse of ρ_R , for example, if $\rho_R = \text{MMDD}$, the $\rho_R^{-1} = \text{DDMM}$.

$$\text{Hetesim}(a, b | \rho) = \frac{PM_{\rho_L} (PM_{\rho_R^{-1}})^T}{\|PM_{\rho_L}\|_2 * \|PM_{\rho_R^{-1}}\|_2} \quad (28)$$

where $\text{Hetesim}(a, b | \rho)$ represents the hetesim score of the node a reaching the node b through path ρ . As shown in Table 1, there are 14 different paths from a miRNA to a drug when the $l < 5$. So, the 14-dimensional hetesim feature between each miRNA-drug node pair in the heterogeneous network is obtained.

TABLE 1 The paths from a miRNA to a drug in the heterogeneous network when $l < 5$.

id	path	meaning
1	MMD	miRNA-miRNA-drug
2	MDD	miRNA-drug-drug
3	MMMD	miRNA-miRNA-miRNA-drug
4	MDMD	miRNA-drug-miRNA-drug
5	MMDD	miRNA-miRNA-drug-drug
6	MDDD	miRNA-drug-drug-drug
7	MMMD	miRNA-miRNA-miRNA-miRNA-drug
8	MMDD	miRNA-miRNA-miRNA-drug-drug
9	MDMD	miRNA-miRNA-drug-miRNA-drug
10	MDDDD	miRNA-miRNA-drug-drug-drug
11	MDMD	miRNA-drug-miRNA-miRNA-drug
12	MDMD	miRNA-drug-miRNA-drug-drug-drug
13	MDDMD	miRNA-drug-drug-miRNA-drug
14	MDDDD	miRNA-drug-drug-drug-drug

2.6 Training the support vector machine classifier

Feature data are obtained by combining inclined diffusion feature and hetesim score. For each pair of drug and miRNA sample in the calculated Hetesim score matrix, the 50-dimensional inclined diffusion feature of the corresponding miRNA and drug are obtained respectively, and 114-dimensional feature is obtained. For example, sample drug d_i and miRNA m_j , the 14-dimensional HeteSim score of $d_i - m_j$ pair is combined with the 50-dimensional inclined diffusion feature of the corresponding drug d_i and the 50-dimensional inclined diffusion feature of miRNA m_j , namely, the i -th row of the X_D and the j -th row of the matrix X_M , to obtain the 114-dimensional feature of drug d_i and miRNA m_j . The 114-dimension feature data of all sample pair are obtained by a similar method. The obtained feature data are used for SVM classifier to predict the miRNAs-drug resistance relationship.

The SVM is an effective classification method and has been widely used in the classification of biological data (43–45). The SVM can transform sample space into high-dimensional or even infinite-dimensional feature space (46). The goal of SVM is to find a hyperplane so that the sample points close to the hyperplane can have a larger distance. The steps of SVM for the algorithm are as follows:

(1) The kernel function $K(x_i, x_j)$ and punish parameter C need to be selected first. The optimization problem is constructed and solved.

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^N \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, i = 1, 2, \dots, N \end{aligned} \quad (29)$$

where $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$. The punish function $C = 64$. The optimal solution is obtained as $\alpha^* = \{\alpha_1^*, \alpha_2^*, \dots, \alpha_N^*\}^T$ (2). A positive component of α^* is selected, $0 \leq \alpha_i^* \leq C$:

$$b^* = y_j - \sum_{i=1}^N \alpha_i^* y_i K(x_i, x_j) \quad (30)$$

(3) The decision function is constructed.

$$f(x) = \text{sign} \left(\sum_{i=1}^N \alpha_i^* y_i K(x_i, x_j) + b^* \right) \quad (31)$$

2.7 The SVMMDR algorithm

In this section, Algorithm 1 describes the implementation details of SVMMDR. In lines 2 to 15 of Algorithm 1, the low-dimensional inclined diffusion feature matrix X_{Ma} and X_{Da} are

obtained by using the inclined random walk with restart and SVD. The hetesim score between any two nodes in a heterogeneous network is obtained from lines 16 to 41. In lines 42 to 45, the combined features is obtained and used to train the SVM classifier. Then the final prediction score is obtained.

Input: miRNAs set, drugs set, The association matrix of the miRNA-drug resistance, A ;

Output: The gaussian interaction profile kernel similarity matrixs, GSM and GSD . The miRNAs sequence similarity matrix, SSM . The chemical structural similarity matrix, ESD . The similarity matrix SM and SD . Prediction score.

Construct the adjacency matrix G ;

Obtain the weight matrix W_{MM} and W_{DD} ;

Initialize the global transition probability matrix T_D and T_M ;

Initialize the transition probability vector for each node

$$P_0^x(m_i) = \frac{A(m_i, d_x)}{\sum_{n=1}^{nm} A(m_n, d_x)}, P_0^y(d_j) = \frac{A(m_y, d_j)}{\sum_{n=1}^{nd} A(m_y, d_j)}$$

while $P_{t+1}^x - P_t^x > 10^{-10}$ **do**:

Obtain the updated probability vector:

$$P_{t+1}^y = (1-r) * T_D * P_t^x + r * P_0^y$$

end while

$$P_{m*n} = U_{m*m} \Sigma_{m*n} V_{n*n}^T \approx U_{m*d} \Sigma_{d*d} V_{d*n}^T X / U_{n*d} \Sigma_{d*d}^{1/2}$$

Get low-dimensional inclined diffusion feature X_M and X_D

Calculate $I_{MD}(m_i, d_j)$, $I_{DD}(d_i, d_j)$, $I_{MM}(m_i, m_j)$

$$I_{MD}(m_i, d_j) = \frac{A(m_i, d_j)}{\sum_{k=1}^{nd} A(m_i, d_k)}$$

$$I_{DD}(d_i, d_j) = \frac{SD(d_i, d_j)}{\sum_{k=1}^{nd} SD(d_i, d_k)}$$

$$I_{MM}(m_i, m_j) = \frac{SM(m_i, m_j)}{\sum_{k=1}^{nm} SM(m_i, m_k)}$$

for $l=1 \rightarrow 5$ **do**

Divide the path into two parts.

if $l \% 2 == 0$ **then**

$$\rho_L = N_1 N_2 \dots N_{\frac{l}{2}+1}$$

$$\rho_R = N_{\frac{l}{2}+1} N_{\frac{l}{2}+2} \dots N_{l+1}$$

$$PM_{\rho_L} = I_{N_1 N_2} I_{N_2 N_3} \dots I_{N_{\frac{l}{2}} N_{\frac{l}{2}+1}}$$

$$PM_{\rho_R} = I_{N_{\frac{l}{2}+1} N_{\frac{l}{2}+2}} I_{N_{\frac{l}{2}+2} N_{\frac{l}{2}+3}} \dots I_{N_l N_{l+1}}$$

end if

if $l \% 2 != 0$ **then**

$$\rho_{L_1} = N_1 N_2 \dots N_{\frac{l+1}{2}}$$

$$\rho_{L_2} = N_1 N_2 \dots N_{\frac{l+3}{2}}$$

$$\rho_{R_1} = N_{\frac{l+1}{2}} N_{\frac{l+1}{2}+1} \dots N_{l+1}$$

$$\rho_{R_2} = N_{\frac{l+3}{2}} N_{\frac{l+3}{2}+1} \dots N_{l+1}$$

$$PM_{\rho_{L_1}} = I_{N_1 N_2} I_{N_2 N_3} \dots I_{N_{\frac{l+1}{2}} N_{\frac{l+1}{2}+1}}$$

$$PM_{\rho_{L_2}} = I_{N_1 N_2} I_{N_2 N_3} \dots I_{N_{\frac{l+1}{2}} N_{\frac{l+3}{2}}}$$

$$PM_{\rho_{R_1}} = I_{N_{\frac{l+1}{2}} N_{\frac{l+3}{2}}} I_{N_{\frac{l+3}{2}} N_{\frac{l+5}{2}}} \dots I_{N_l N_{l+1}}$$

$$PM_{\rho_{R_2}} = I_{N_{\frac{l+3}{2}} N_{\frac{l+5}{2}}} I_{N_{\frac{l+5}{2}} N_{\frac{l+7}{2}}} \dots I_{N_l N_{l+1}}$$

$$PM_{\rho_L} = \frac{PM_{\rho_{L_1}} + PM_{\rho_{L_2}}}{2}$$

$$PM_{\rho_R} = \frac{PM_{\rho_{R_1}} + PM_{\rho_{R_2}}}{2}$$

end if

$$Hetesim(a, b | \rho) = \frac{PM_{\rho_L} (PM_{\rho_R})^T}{\|PM_{\rho_L}\|_2 * \|PM_{\rho_R}\|_2}$$

end for

Combined with the inclined diffusion feature and HeteSim score to get the data set

$$D_{train} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

$$D_{test} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

Use D_{train} to train the Support Vector Machines (SVM) as classifier

ALGORITHM 1

SVMMDR algorithm.

3 Result and discussion

3.1 Data sets

The miRNAs-drug resistance association data are downloaded from ncDR database. After deduplication, 85 drugs and 625 miRNAs are obtained, and 2301 miRNAs-drug resistance known association are obtained, all as positive samples. Negative samples are randomly selected from unknown associations with three times the number of positive samples. The final sample dataset is constructed from 2301 positive samples and 6903 negative samples.

3.2 Performance measures

The 10-fold Cross-Validation(10-CV) is performed to evaluate the performance of SVMMDR. The process of 10-CV is as follows: the sample data is equally divided into 10 groups. The 9 group of data is used as the training set, and the remaining group is used as the validation set. After ten times of the above process, each of the 10 groups in turn is used as a validation data to obtain 10 performance results. The final performance evaluation is obtained by averaging the 10 performance results. Multiple measures are used to evaluate performance, such as the area under the receiver operating characteristic curves (AUC), recall (REC), accuracy (ACC), F1-score and Matthews Correlation Coefficient (MCC). They can be presented as below:

$$Recall = \frac{TP}{TP + FN}, \quad (32)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (33)$$

$$F1 - Score = \frac{2 \times TP}{2TP + FP + FN}, \quad (34)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (35)$$

where the *TP* is the number of samples that are correctly classified as positive, the *FP* is the number of samples that are misclassified as positive, the *TN* represents the number of samples that are correctly classified as negative, and the *FN* is the number of samples that are misclassified as negative.

3.3 Performance comparison with existing machine learning methods

In order to reflect the performance of SVM, the proposed SVMMDR methods will be compared with the following solution, including using logistic regression (LR) as a classifier,

the use of random forests (RF) used as a classifier, K nearest neighbor (KNN) as a classifier. The same features of the same training sample are used to train the corresponding classifiers. To get performance, the 10-fold cross-validation is applied. For KNN classifier, the 10 nearest neighbors and leaf size of 20 point is used. The RF builds a number of decision tree classifiers trained on a set of randomly selected samples of the benchmark to improve the performance. For LR, the maxiter and tol parameters are optimized to 500 and 0.001, respectively.

Figure 2 indicates the ROC curves of SVMMDR using other classifiers. The AUC of SVMMDR, KNN, RF and LR are 0.978, 0.939, 0.892 and 0.948. Furthermore, Table 2 shows the values of performance measures such as ACC, Pre, Recall, F1-score, MCC. The results show that the AUC value obtained by SVMMDR is the highest. The value of performance measure is also better than other classifiers. The SVM classifier can achieve effective classification by mapping features to higher dimensional space through kernel function changes. At the same time, the optimal solution is obtained with constraints, which can make the classification more accurate.

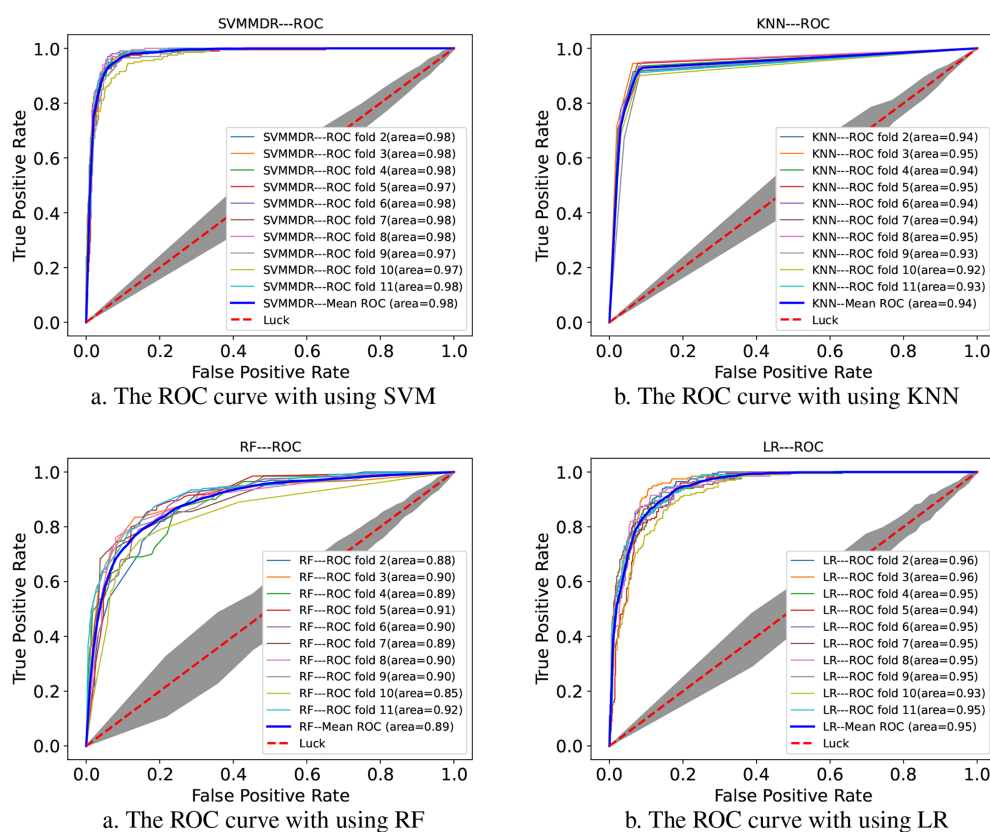


FIGURE 2

The ROC curve comparison with existing machine learning methods.

TABLE 2 Performance comparison of existing machine learning methods.

Method	ACC	Pre	RECALL	F1-score	MCC
SVMMMDR	0.9393	0.8705	0.8905	0.8800	0.8399
RF	0.8381	0.81733	0.4625	0.5704	0.5236
LR	0.8925	0.8020	0.7565	0.7785	0.7082
KNN	0.9080	0.8936	0.7175	0.7957	0.7448

3.4 Performance comparison with different topological features

To demonstrate the advantages of combining features in SVMMMDR, different feature groups (hetesim+ inclined diffusion feature, hetesim feature, and inclined diffusion feature) are used for comparison experiment. The comparison results are shown in Figures 3 and 4. Denote “SVMMMDR”, “Hetesim” and “in-Diff” as the combination feature, hetesim feature and inclined diffusion feature. Figure 3 shows the ROC curves of different feature groups. It can be seen that the combination of hetesim and inclined diffusion obtained a higher AUC than the two separate feature, and the AUC obtained by inclined diffusion feature alone is higher than that obtained by hetesim alone. Figure 4 represents the performance achieved for the different feature groups. It can also see that the combination of the two features has best performance. Although the AUC of inclined diffusion feature reaches 0.96, the Pre, F1 and MCC are all relatively low. The combination of inclined diffusion feature and hetesim feature can solve this problem and improve performance.

3.5 Performance comparison with existing methods

To further illustrate the superiority of the proposed method, the SVMMMDR is compared with existing miRNA-resistance prediction algorithms, such as GCMDR, MDIPA and BiRW-MD, all of which use the sample set in this paper. Performance measures are obtained by performing 10-fold cross-validation.

GCMDR (35): Data from multiple data sources are fused. The latent factor method are constructed using graph convolution to learn the graph embedding feature of miRNAs and drugs, and end-to-end prediction method are built.

MDIPA (37): The identification of potential miRNAs-drug interactions is seen as a matrix completion problem, the unknown associations are predicted based on weighted non-negative matrix factorization. The path-based miRNAs similarity matrix and drugs similarity matrix based on drugs structure information are obtained, which are combined with extracted drugs and miRNAs neighbor information for prediction.

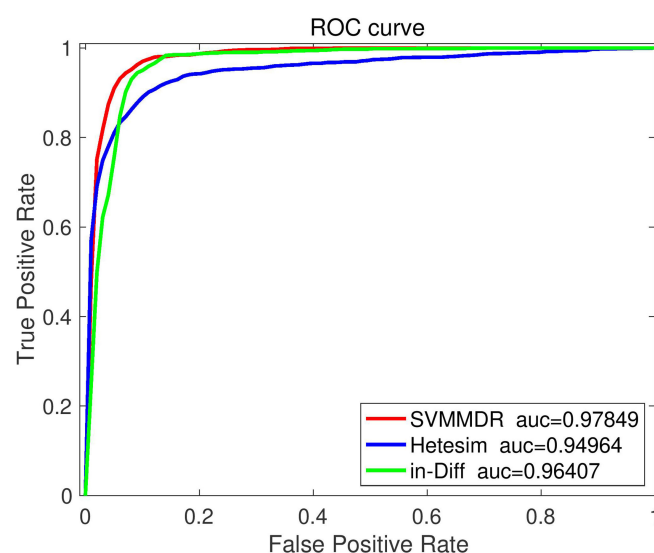


FIGURE 3
The ROC curve comparison with different feature.

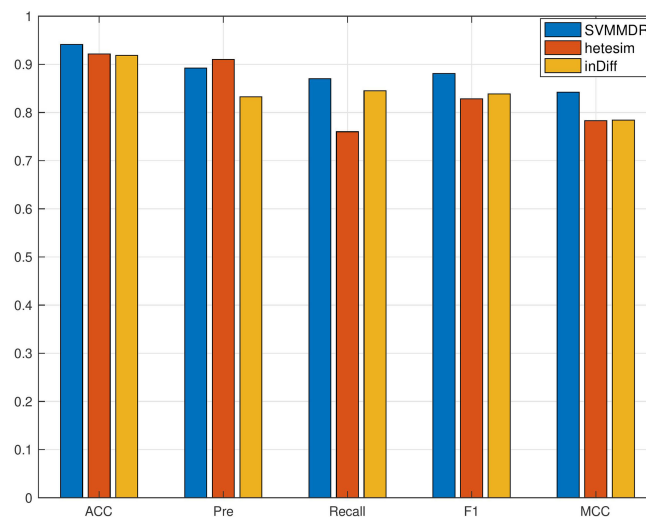


FIGURE 4
The performance comparison with different feature.

BiRW-WD (33): The multiple similarity networks and miRNAs-drugs association network are integrated to construct a heterogeneous network. In the heterogeneous networks, the bi-directional random walk (BiRW) are used to predict potential miRNAs-drug effect associations.

Figure 5 illustrates the comparison results. It can be seen that the proposed SVMMDR method achieves the best performance. The reasons are as follows: (1) The drug group and miRNA group are introduced. When restart random walk is used to obtain

diffusion feature, the walker is more inclined to select the node of the next walk. The inclined diffusion feature contribute to the prediction accuracy. (2) The hetesim score is obtained from the path information of two nodes in the heterogeneous network. Regardless of the same or different node types, the hetesim measures their correlation within a unified framework. At the same time, according to the search path between two nodes, the measure between node pairs is defined by following a sequence. (3) The SVM with high accuracy is used as the classifier.

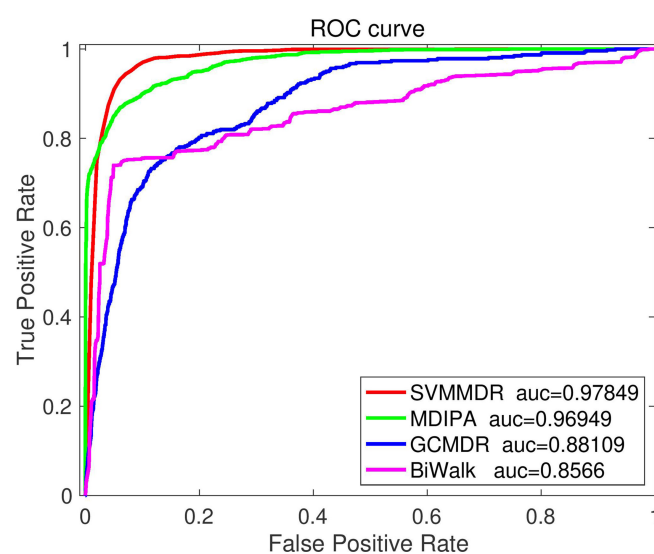


FIGURE 5
The ROC curve comparison with existing methods.

TABLE 3 The top 20 predicted miRNA related to 5-FU.

miRNA	PMID	miRNA	PMID
miR-3665	25117811	let-7b	22382630
miR-3619	25117811	miR-99a	24304648
miR-3141	25117811	miR-4465	25117811
miR-1234	25117811	miR-4484	25117811
miR-1275	25356050	miR-887	25117811
miR-29b-2	23167930	miR-4750	25117811
miR-423	25117811	miR-221	25117811
miR-107	22382630	miR-191	24304648
miR-296	20485139	miR-23a	24249161
miR-483	24304648	miR-21	24275137

3.6 Case study

In order to illustrate the effectiveness of the proposed method, we present a case study of the drug 5-fluorouracil(5-FU). The 5-FU is an antimetabolite drug widely used in cancer treatment, especially colorectal cancer (CRC) Longley et al. (47). There are 244 miRNAs related to 5-FU in ncDR database. We remove these associations in the association matrix A and use the rest as test data. The SVMMDR algorithm proposed in this paper is used for prediction, and 174 miRNAs with prediction scores greater than 0.95 are obtained. For the top 20 predicted miRNAs, we verify whether predicted miRNAs-drug resistance associations are confirmed by searching the PubMed literature. Table 3 indicates the miRNAs and the PMIDs of publications mentioning the association between miRNAs and 5-FU. For example, miR-21 expression levels are confirmed to lead to 5-Fluorouracil resistance Tomimaru et al. (48). The miR-23a enhances 5-FU resistance in microsatellite instability (MSI) CRC cells through targeting ABCF1 Li et al. (49).

4 Conclusion

More and more evidence indicates that miRNA expression level is related to drug resistance, affecting the therapeutic effect of disease. Predicting the association between miRNA-drug resistance can help to select more appropriate drugs for clinical treatment and promote the cure of disease. However, there are also very few computation-based predictive tools for miRNA- drug resistance.

Therefore, in this paper, a method based on the Support Vector Machines to predict the relationship between MiRNA and Drug Resistance (SVMMDR) is proposed. The SVMMDR integrates miRNAs-drug resistance association, miRNAs

sequence similarity, drug chemical structure similarity and other similarities, extracts path-based hetesim features, and obtains inclined diffusion features through inclined restart random walk. The machine learning algorithm SVM is used to predict the association between miRNAs and drug resistance.

The 10-fold cross-validation is used to assess the performance of SVMMDR. The area under the ROC curve AUC is used as a measure of performance. The AUC of SVMMDR reaches 0.978. The results show that SVMMDR has a significant performance advantage.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <http://www.jianglab.cn/ncDR/>.

Author contributions

TD, ZFK, and LD conceived this work and designed the experiments. TD and ZFK carried out the experiments. TD and ZFK collected the data and analyzed the results. TD, ZFK and LD wrote, revised, and approved the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grants Nos. 62072477, 61309027, 61702562 and 61702561, the Hunan Provincial Natural Science Foundation of China under Grants No.2018JJ3888, the Scientific Research Fund of Hunan Provincial Education Department under Grant No.18B197, the

National Key R&D Program of China under Grant No.2018YFB1700200, the Open Research Project of Key Laboratory of Intelligent Information Perception and Processing Technology (Hunan Province) under Grant No.2017KF01, the Hunan Key Laboratory of Intelligent Logistics Technology 2019TP1015.

Acknowledgments

We would like to thank the Experimental Center of School of Computer and Information Engineering, Central South University of Forestry and Technology, for providing computing resources.

References

- Xiang J, Zhang J, Zheng R, Li X, Li M. NIDM: network impulsive dynamics on multiplex biological network for disease-gene prediction. *Briefings Bioinf* (2021) 22. doi: 10.1093/bib/bbab080
- Xiang J, Meng X, Wu F-X, Li M. HyMM: Hybrid method for disease-gene prediction by integrating multiscale module structures. *bioRxiv* (2021) 23. doi: 10.1101/2021.04.30.442111
- Xiang J, Zhang N-R, Zhang J-S, Lv X-Y, Li M. PrGeFNE: Predicting disease-related genes by fast network embedding. *Methods* (2021) 192:3–12. doi: 10.1016/j.ymeth.2020.06.015
- Yu L, Zheng Y, Gao L. miRNA–disease association prediction based on meta-paths. *Briefings Bioinf* (2022) 23. doi: 10.1093/bib/bbab571
- Ding Y, Tian L-P, Lei X, Liao B, Wu F-X. Variational graph auto-encoders for miRNA-disease association prediction. *Methods* (2021) 192:25–34. doi: 10.1016/j.ymeth.2020.08.004
- Guo L, Shi K, Wang L. MLPMDA: Multi-layer linear projection for predicting miRNA- disease association. *Knowledge-Based Syst* (2021) 214:106718. doi: 10.1016/j.knsys.2020
- Zhong T, Li Z, You Z-H, Nie R, Zhao H. Predicting miRNA–disease associations based on graph random propagation network and attention network. *Briefings Bioinf* (2022). doi: 10.1093/bib/bbab589
- Jin C, Shi Z, Lin K, Zhang H. Predicting miRNA-disease association based on neural inductive matrix completion with graph autoencoders and self-attention mechanism. *Biomolecules* (2022) 12. doi: 10.3390/biom12010064
- Dai Q, Wang Z, Liu Z, Duan X, Song J, Guo M. Predicting miRNA-disease associations using an ensemble learning framework with resampling method. *Briefings Bioinf* (2022) 23:bbab543. doi: 10.1093/bib/bbab543
- Zhu Q, Fan Y, Pan X. Fusing multiple biological networks to effectively predict miRNA-disease associations. *Curr Bioinf* (2021) 16:371–84. doi: 10.2174/1574893615999200715165335
- Yan C, Duan G, Li N, Zhang L, Wu F-X, Wang J. PDMDA: predicting deep-level miRNA–disease associations with graph neural networks and sequence features. *Bioinformatics* (2022) 508:2226–34. doi: 10.1093/bioinformatics/btac077
- Liu D, Huang Y, Nie W, Zhang J, Deng L. SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. *BMC Bioinf* (2021) 22:219. doi: 10.1186/s12859-021-04135-2
- Han G, Kuang Z, Deng L. MSCNE:predict miRNA-disease associations using neural network based on multi-source biological information. *IEEE/ACM Trans Comput Biol Bioinf* (2021) 1–1. doi: 10.1109/TCBB.2021.3106006
- Yu N, Liu Z-P, Gao R. Predicting multiple types of MicroRNA-disease associations based on tensor factorization and label propagation. *Comput Biol Med* (2022) 146:105558. doi: 10.1016/j.compbiomed.2022.105558
- Wang M-N, You Z-H, Wang L, Li L-P, Zheng K. LDGRNMF: LncRNA-disease associations prediction based on graph regularized non-negative matrix factorization. *Neurocomputing* (2021) 424:236–45. doi: 10.1016/j.neucom.2020.02.062
- Wang L, Xiao Y, Li J, Feng X, Li Q, Yang J. IIRWR: Internal inclined random walk with restart for LncRNA-disease association prediction. *IEEE Access* (2019) 7:54034–41. doi: 10.1109/ACCESS.2019.2912945
- Duan T, Kuang Z, Wang J, Ma Z. GBDTLRL2D: Predicts LncRNA-disease associations using MetaGraph2Vec and K-means based on heterogeneous network. *Front Cell Dev Biol* (2021) 9:753027. doi: 10.3389/fcell.2021.753027
- Fan Y, Chen M, Pan X. GCRFLDA: scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field. *Briefings Bioinf* (2021) 23. doi: 10.4231093/bib/bbab361.Bbab361
- Lan W, Wu X, Chen Q, Peng W, Wang J, Chen YP. GANLDA: Graph attention network for lncRNA-disease associations prediction. *Neurocomputing* (2022) 469:384–93. doi: 10.1016/j.neucom.2020.09.094
- Wang L, Shang M, Dai Q, He P-A. Prediction of lncRNA-disease association based on a laplace normalized random walk with restart algorithm on heterogeneous networks. *BMC Bioinf* (2022) 23:5. doi: 10.1186/s12859-021-04538-1
- Liu Y, Yu Y, Zhao S. Dual attention mechanisms and feature fusion networks based method for predicting LncRNA-disease associations. *Interdiscip Sciences: Comput Life Sci* (2022) 14:358–71. doi: 10.1007/s12539-021-00492-x
- Zhang L, Liu T, Chen H, Zhao Q, Liu H. Predicting lncRNA–miRNA interactions based on interactome network and graphlet interaction. *Genomics* (2021) 113:874–80. doi: 10.1016/j.ygeno.2021.02.002
- Zhang L, Yang P, Feng H, Zhao Q, Liu H. Using network distance analysis to predict lncRNA–miRNA interactions. *Interdiscip Sciences: Comput Life Sci* (2021) 13:535–45. doi: 10.1007/s12539-021-00458-z
- Yang L, Li L-P, Yi H-C. DeepWalk based method to predict lncRNA-miRNA associations via lncRNA-miRNA-disease-protein-drug graph. *BMC Bioinf* (2022) 22:621. doi: 10.1186/s12859-022-04579-0
- Xu M, Chen Y, Lu W, Kong L, Fang J, Li Z, et al. SPMLMI: predicting lncRNA–miRNA interactions in humans using a structural perturbation method. *PeerJ* (2021) 9:e11426. doi: 10.7717/peerj.11426
- Liu H, Ren G, Chen H, Liu Q, Yang Y, Zhao Q. Predicting lncRNA-miRNA interactions based on logistic matrix factorization with neighborhood regularized. *Knowledge-Based Syst* (2020) 191:105261. doi: 10.1016/j.knsys.2019.105261
- Li Z, Rana TM. Therapeutic targeting of microRNAs: current status and future challenges. *Nat Rev Drug Discovery* (2014) 13:622–38. doi: 10.1038/nrd4359
- Guo Y, Yan K, Fang J, Qu Q, Zhou M, Chen F. Let-7b expression determines response to chemotherapy through the regulation of cyclin D1 in glioblastoma. *J Exp Clin Cancer Res* (2013) 32:1–10. doi: 10.1186/1756-9966-32-41
- Li H, Xu H, Shen H. microRNA-106a modulates cisplatin sensitivity by targeting PDCCD4 in human ovarian cancer cells. *Oncol Lett* (2014) 7:183–8. doi: 10.3892/ol.2013.1644
- Rukov JL, Wilentzik R, Jaffe I, Vinther J, Shomron N. PharmacomiR: linking microRNAs and drug effects. *Briefings Bioinform* (2014) 15:648–59. doi: 10.1093/bib/bbs082

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.

31. Chen X, Xie W-B, Xiao P-P, Zhao X-M, Yan H. mTD: a database of microRNAs affecting therapeutic effects of drugs. *J Genet Genomics* (2017) 44:269–71. doi: 10.1016/j.jgg.2017.04.003
32. Dai E, Yang F, Wang J, Zhou X, Song Q, An W, et al. ncDR: a comprehensive resource of non-coding RNAs involved in drug resistance. *Bioinformatics* (2017) 33:4010–1. doi: 10.1093/bioinformatics/btx523
33. Xu P, Wu Q, Rao Y, Kou Z, Fang G, Liu W, et al. Predicting the influence of MicroRNAs on drug therapeutic effects by random walking. *IEEE Access* (2020) 8:117347–53. doi: 10.1109/ACCESS.2020.3004512
34. Zhao Y, Chen X, Yin J, Qu J. SNMFSSMA: using symmetric nonnegative matrix factorization and kronecker regularized least squares to predict potential small molecule-microRNA association. *RNA Biol* (2020) 17:281–91. doi: 10.1080/15476286.2019.1694732
35. Huang Y-a, Hu P, Chan KC, You Z-H. Graph convolution for predicting associations between miRNA and drug resistance. *Bioinformatics* (2020) 36:851–8. doi: 10.1093/bioinformatics/btz621
36. Yu S, Xu H, Li Y, Liu D, Deng L. (2021). LGCMDS: Predicting miRNA-drug sensitivity based on light graph convolution network, in: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. New York: IEE. pp. 217–22. doi: 10.1109/BIBM52615.2021.9669566
37. Jamali AA, Kusalik A, Wu F-X. MDIPA: a microRNA–drug interaction prediction approach based on non-negative matrix factorization. *Bioinformatics* (2020) 36:5061–7. doi: 10.1093/bioinformatics/btaa577
38. Qu J, Chen X, Sun Y-Z, Zhao Y, Cai S-B, Ming Z, et al. In silico prediction of small molecule-miRNA associations based on the HeteSim algorithm. *Mol Therapy-Nucleic Acids* (2019) 14:274–86. doi: 10.1016/j.omtn.2018.12.002
39. Luo J, Shen C, Lai Z, Cai J, Ding P. Incorporating clinical, chemical and biological information for predicting small molecule-microRNA associations based on non-negative matrix factorization. *IEEE/ACM Trans Comput Biol Bioinf* (2021) 18:2535–45. doi: 10.1109/TCBB.2020.2975780
40. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* (2019) 47:D155–62. doi: 10.1093/nar/gky1141
41. Hattori M, Tanaka N, Kanehisa M, Goto S. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res* (2010) 38:W652–6. doi: 10.1093/nar/gkq367
42. Shi C, Kong X, Huang Y, Philip SY, Wu B. Hetsim: A general framework for relevance measure in heterogeneous networks. *IEEE Trans Knowledge Data Eng* (2014) 26:2479–92. doi: 10.1109/TKDE.2013.2297920
43. Zou Y, Wu H, Guo X, Peng L, Ding Y, Tang J, et al. MK-FSVM-SVDD: a multiple kernel-based fuzzy SVM model for predicting DNA-binding proteins via support vector data description. *Curr Bioinf* (2021) 16:274–83. doi: 10.2174/1574893615999200607173829
44. Wang M, Liang Y, Hu Z, Chen S, Shi B, Heidari AA, et al. Lupus nephritis diagnosis using enhanced moth flame algorithm with support vector machines. *Comput Biol Med* (2022) 145:105435. doi: 10.1016/j.combiomed.2022.105435
45. Ao C, Yu L, Zou Q. Prediction of bio-sequence modifications and the associations with diseases. *Briefings Funct Genomics* (2020) 20:1–18. doi: 10.1093/bfpg/elaa023
46. Hearst M, Dumais S, Osuna E, Platt J, Scholkopf B. "Support vector machines," in *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, July–Aug. 1998, doi: 10.1109/5254.708428
47. Longley DB, Harkin DP, Johnston PG. 5-fluorouracil: mechanisms of action and clinical strategies. *Nat Rev Cancer* (2003) 3:330–8. doi: 10.1038/nrc1074
48. Tomimaru Y, Eguchi H, Nagano H, Wada H, Tomokuni A, Kobayashi S, et al. MicroRNA-21 induces resistance to the anti-tumour effect of interferon- α /5-fluorouracil in hepatocellular carcinoma cells. *Br J Cancer* (2010) 103:1617–26. doi: 10.1038/sj.bjc.6605958
49. Li X, Li X, Liao D, Wang X, Wu Z, Nie J, et al. Elevated microRNA-23a expression enhances the chemoresistance of colorectal cancer cells with microsatellite instability to 5-fluorouracil by directly targeting ABCF1. *Curr Protein Pept Sci* (2015) 16:301–9. doi: 10.2174/138920371604150429153309



OPEN ACCESS

EDITED BY

Tianyi Zhao,
Harbin Institute of Technology, China

REVIEWED BY

Mario Perez-Medina,
Instituto Politécnico Nacional (IPN),
Mexico
Chi-Chang Chang,
Chung Shan Medical University,
Taiwan
Zhang Shuyao,
Guangzhou Red,
Cross Hospital, China

*CORRESPONDENCE

Tongcun Zhang
zhangtongcun@wust.edu.cn
Fan Sun
sunfan@wust.edu.cn
Weidong Hu
huwd@whu.edu.cn

[†]These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 29 July 2022

ACCEPTED 14 September 2022

PUBLISHED 29 September 2022

CITATION

Wu H, Chen C, Gu L, Li J, Yue Y,
Lyu M, Cui Y, Zhang X, Liu Y, Zhu H,
Liao X, Zhang T, Sun F and Hu W
(2022) B cell deficiency promotes the
initiation and progression of
lung cancer.
Front. Oncol. 12:1006477.
doi: 10.3389/fonc.2022.1006477

COPYRIGHT

© 2022 Wu, Chen, Gu, Li, Yue, Lyu, Cui,
Zhang, Liu, Zhu, Liao, Zhang, Sun and
Hu. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

B cell deficiency promotes the initiation and progression of lung cancer

Han Wu^{1,2†}, Chen Chen^{3†}, Lixing Gu^{1,2,4}, Jiapeng Li^{1,2,4},
Yunqiang Yue^{1,2}, Mengqing Lyu^{1,2}, Yeting Cui^{1,2},
Xiaoyu Zhang^{1,2}, Yu Liu^{1,2}, Haichuan Zhu^{1,2}, Xinghua Liao^{1,2},
Tongcun Zhang^{1,2*}, Fan Sun^{1,2*} and Weidong Hu^{5*}

¹College of Life Sciences and Health, Wuhan University of Science and Technology, Wuhan, China,

²Institute of Biology and Medicine, Wuhan University of Science and Technology, Wuhan, China,

³Department of Biological Repositories, Zhongnan Hospital of Wuhan University, Wuhan, China,

⁴College of Science, Wuhan University of Science and Technology, Wuhan, China, ⁵Department of
Thoracic Surgery, Zhongnan Hospital of Wuhan University, Wuhan, China

Currently commercialized CAR-T cell therapies targeting CD19 and BCMA show great efficacy to cure B cell malignancies. However, intravenous infusion of these CAR-T cells severely destroys both transformed and normal B cells in most tissues and organs, in particular lung, leading to a critical question that what the impact of normal B cell depletion on pulmonary diseases and lung cancer is. Herein, we find that B cell frequency is remarkably reduced in both smoking carcinogen-treated lung tissues and lung tumors, which is associated with advanced cancer progression and worse patient survival. B cell depletion by anti-CD20 antibody significantly accelerates the initiation and progression of lung tumors, which is mediated by repressed tumor infiltration of T cells and macrophage elimination of tumor cells. These findings unveil the overall antitumor activity of B cells in lung cancer, providing novel insights into both mechanisms underlying lung cancer pathogenesis and clinical prevention post CAR-T cell therapy.

KEYWORDS

B cell, CAR-T, lung cancer, initiation, progression

Introduction

CAR-T cell therapy represents the most revolutionary breakthrough in treatment of hematopoietic cancers in the new century (1). At present, there are eight products of CAR-T cells marketed all over the world with six targeting CD19 and two BCMA to cure B-cell leukemia/lymphoma and multiple myeloma, respectively (2, 3). In addition to strongly expressed in B cell malignancies, both molecules are constitutively and exclusively expressed on naïve and memory B cells. Consequently, current CAR-T cell

therapy results in severe loss of both malignant and normal B cells in most tissues and organs, in particular lung, the major accumulation site of CAR-T cells by intravenous infusion (1–3). However, the long-term effect of normal B cell depletion from lung frequently exposed to environmental risks is unknown.

B cell is one of the most abundant immune cells in lung tissues under both physiological and pathological conditions, including lung cancer (4, 5). The main function of B cell comprises antibody production for humoral immunity, antigen presentation for T cell immunity, and immune regulation. However, different from T lymphocytes, the role of B lymphocytes in cancers is in debate (6–8). In murine pre-malignancy models, failure in immune cell recruitment and tumorigenesis in K14-HPV16 mice (T and B cell-deficient) can be overcome by adoptive transfer of B cells or serum from HPV16 mice to induce strong infiltration of innate immune cells and malignant progression, emphasizing the indispensable role for B cell in establishment of chronic inflammatory state to promote carcinogenesis (9). In line with these findings, B cell depletion by administration of anti-CD20 antibody is capable of preventing premalignant dysplasia in K14-HPV16 mice with resultant reduced levels of serum IgG and immune cells in neoplastic site (10). Similarly, compared to B cells with selective TNF α deletion, adoptive transfer of B cells from wild type mice into TNF α knockouts notably stimulates papilloma development in DMBA/TPA murine model of skin tumorigenesis, highlighting the tumor-promoting function of Breg, the known source of TNF α (11). Meanwhile, IFN γ secretion and T cell infiltration are elevated in TNF α knockout mice (11). Consistently, B cells not only induce a non-protective humoral immune response, but also inhibit CD4⁺ T cells to mount CTL-mediated tumor immunity (12).

The pro-tumor role of B cells is also observed in human cancers. An increase of tumor-infiltrating B cells is associated with poor prognosis and survival in patients with metastatic ovarian cancer (13, 14). In tumor tissues, STAT3 activation in B cells is positively associated with tumor angiogenesis (15). Moreover, partial B-cell deletion by rituximab leads to repressed tumor burden in half patients with advanced colorectal cancer (16). These studies support the positive role for B cells in fostering malignant transition and progression in both murine and human cancers.

On the other hand, the antitumor role of B cells is found in murine established tumor models. Anti-CD20-mediated B cell depletion exacerbates primary tumor burden and pulmonary metastasis in B16 melanoma model (17), whereas adoptive transfer of CpG-primed B cells inhibits tumor progression of melanoma in B-cell deficient mice (18). Accordingly, increased tumor growth and metastasis is observed in mice bearing 4T1 breast tumor by administration of anti-CD20 antibody, with enriched Breg abundance and impaired T cell activity, which can be reverted by adoptive transfer of CpG-activated B cells (19). The tumor-inhibiting role of B cells involves both direct killing

of tumor cells through FasL/Fas and granzyme B/perforin pathways and indirect eradication of tumor by promotion of tumor infiltration of T cells, complement-dependent tumor cell lysis, and antibody-mediated ADCC and phagocytosis of tumor cells (20).

The antitumor role of B cells is also reported in human cancers in literature. B cell frequency in tumor is positively correlated with patient survival, representing a novel prognostic cellular biomarker (21). Moreover, the presence of B cells in tertiary lymphoid structure (TLS) is positively associated with protective tumor immunity, including high levels of naïve, memory, and activated T cells, and low levels of exhausted T cells and Tregs (22–24).

In summary, the role of B cells in tumor initiation and progression is controversial in both human and murine cancers, in particular lung cancer. Given malignant and normal B cell depletion by current CAR-T cell therapy, to explore its consequence is urgently needed to improve clinical treatment and/or prevention post therapy. In the present study, the prognostic role of B cells in lung cancer was extensively analyzed with public databases. Then, the dynamic changes of B cell frequency in lung tissues and lung tumors were examined during lung tumorigenesis. The effect of B cell deficiency on tumor infiltration of immune cells, T cell activation, tumor initiation and progression was investigated in endogenous lung tumor model. These findings about B cell function will provide novel insights into both lung cancer pathogenesis and clinical prevention post CAR-T cell therapy.

Materials and methods

Public data mining

The gene expression data of B cell specific markers (CD19, CD79A, CD79B, BLK, and CD20/MS4A1) and clinical characteristics in human lung cancer were retrieved from The Cancer Genome Atlas (TCGA) database in Xena website (<http://xena.ucsc.edu>). The mean value of these five B cell specific markers was utilized as B cell set for correlation analysis. Individual B cell specific marker was also analyzed. The association of immune infiltrates with clinical outcomes was determined by Kaplan-Meier Plotter (<http://kmplot.com/analysis/index.php?p=service&cancer=lung>) and TIMER (<http://timer.cistrome.org/>) as described (25). All information in detail was indicated in Figure Legends.

Mouse and tumor cell

FVB/N and BALB/c mice originally from Beijing HFK Bioscience Co., Ltd were housed in pathogen-free conditions and used according to protocols approved by the Animal Ethics

Committee of Wuhan University of Science and Technology. To induce endogenous lung tumor, 1 g/kg body weight of urethane was intraperitoneally injected once a week for six consecutive weeks in 6-week-old female wildtype mice. After 6-week tumor initiation and 6-week tumor progression, the urethane-treated mice were sacrificed and/or kept for lung tumor analysis and immune profiling at the indicated time points (26). For B cell depletion, anti-CD20 antibody (200 µg per mouse) was administrated through intravenous tail vein injection every three weeks, starting two days before urethane treatment. An isotype control antibody (Ctrl) was included for comparison. For spontaneous lung tumor, female FVB/N wildtype mice were maintained for 24 months and sacrificed for tumor examination and immune cell analysis. For syngeneic xenograft, murine lung tumor cells (LAP0297 and MAD109, 10^6 cells per mouse) were subcutaneously inoculated (sc) in the right flank or intravenously injected (iv) through tail vein. Then, tumor tissues or lung tissues were collected for immune cell analysis at the indicated time as described (27).

FACS analysis

The single cell suspensions from fresh lung tissues and lung tumors were blocked with α CD16/CD32 and stained with the antibody against cell surface antigen. If needed, the cells were then fixed with paraformaldehyde (2%), permeabilized and incubated with antibodies against intracellular antigens. For nuclear staining, Permeabilization/Fixation buffer was utilized. For IFN γ staining, cells were treated with phorbol 12-myristate 13-acetate (PMA, 50 ng/ml), ionomycin (1 µM), brefeldin A (BFA, 3 µg/ml) and monensin (2 µM) for 4 hours before they were collected for staining. Data were acquired by Accuri C6 or LSRFortessa I and analyzed by FlowJo software as described (28, 29). In brief, CD45 was included as a marker to distinguish immune cells from other cells in cell suspensions both *in vitro* and *in vivo*. Immune cells (CD45 $^+$) were gated with the indicated markers as follows. Lymphocytes were gated for CD4 $^+$ T cells (CD3 $^+$ CD4 $^+$ CD8 $^-$), Treg cells (CD3 $^+$ CD4 $^+$ Foxp3 $^+$ CD25 $^+$), CD8 $^+$ T cells (CD3 $^+$ CD4 $^-$ CD8 $^+$), NK cells (NKp46 $^+$ CD3 $^-$), and B cells (B220 $^+$ CD3 $^-$). B lymphocytes were further classified into Breg cells (B220 $^+$ CD1d $^+$ CD5 $^+$), memory B cells (B220 $^+$ IgD $^+$ IgM $^+$ CD38 $^+$), and plasma cells (B220 $^-$ CD138 $^+$). Myeloid cells were gated sequentially as follows: neutrophils (CD11b $^+$ Ly6G $^+$), macrophages (Mac: MerTK $^+$ CD64 $^+$; AM: CD11c $^+$ CD11b $^+$; IM: CD11c $^-$ CD11b $^+$), dendritic cells (MHC-II $^+$ CD11c $^+$), monocytes (CD11b $^+$ SSC lo). Tumor cells were gated as CD45 $^+$ EpCAM $^+$ cells. All antibodies used for FACS analysis were shown in [Supplementary Table S1](#).

Immunohistochemistry assay

Mouse lung tissues and lung tumors were excised, fixed in formalin, embedded in paraffin, and cut into 4-µm-thick

sections. Sections were subjected to sequential incubations with the indicated primary antibodies, biotinylated secondary antibodies and streptavidin-horseradish peroxidase (HRP) as described (29). Images of the staining were analyzed using the image J software. The data represented were from five mice per group, with over 500 cells counted in each mouse. Antibodies used for IHC assay were listed in [Supplementary Table S1](#).

Statistical analysis

Two tailed, unpaired Student's *t* test was employed to assess significance of differences between two groups. Ordinary one-way ANOVA was performed to analyze the significance of differences among multiple groups. Chi-square test was carried out to determine the association between clinical characteristics. Log-rank test was used to compare survival between groups. All data are represented as bars (means \pm SEM) with sample dots. The experimental replication and sample numbers for tumor analysis, FACS, and immunohistochemistry were indicated in [FIGURE](#) or [FIGURE LEGENDS](#). The *p* values were indicated as **p* < 0.05, ***p* < 0.01, ns, not statistically significant. The *p* values < 0.05 and 0.01 were considered statistically significant and highly statistically significant, respectively.

Results

Tumor-infiltrating B cells are positively associated with patient survival in lung cancer

B cells are key lymphocytes to mediate humoral immunity against extracellular pathogens, however, the overall function of these antibody-secreting immune cells in lung cancer remains elusive. To determine the role of B cell in lung cancer, public data from TCGA LUNG cohort were retrieved to examine the association between B cell and patient survival. Initially, five specific B cell markers (CD19, CD79A, CD79B, BLK, and CD20/MS4A1) were chosen as a gene set for B cell characterization in RNA sequencing data of lung tumors. Intriguing, both B cell set and individual specific B cell markers were positively correlated with patient survival, including overall survival (OS), disease specific survival (DSS), disease free interval (DFI), and progression free interval (PFI) ([Figure 1A](#); [Supplementary Figure S1](#)). Data from Kaplan-Meier Plotter also exhibited better overall survival (OS), first progression (FP), and post-progression survival (PPS) in patients with high CD20 expression, compared to that in patients with low CD20 expression ([Supplementary Figure S2A](#)). Taken together, these results indicate the antitumor role of B cell in lung cancer.

To confirm this opinion, TIMER database was employed to investigate the association between tumor-infiltrating B cells and

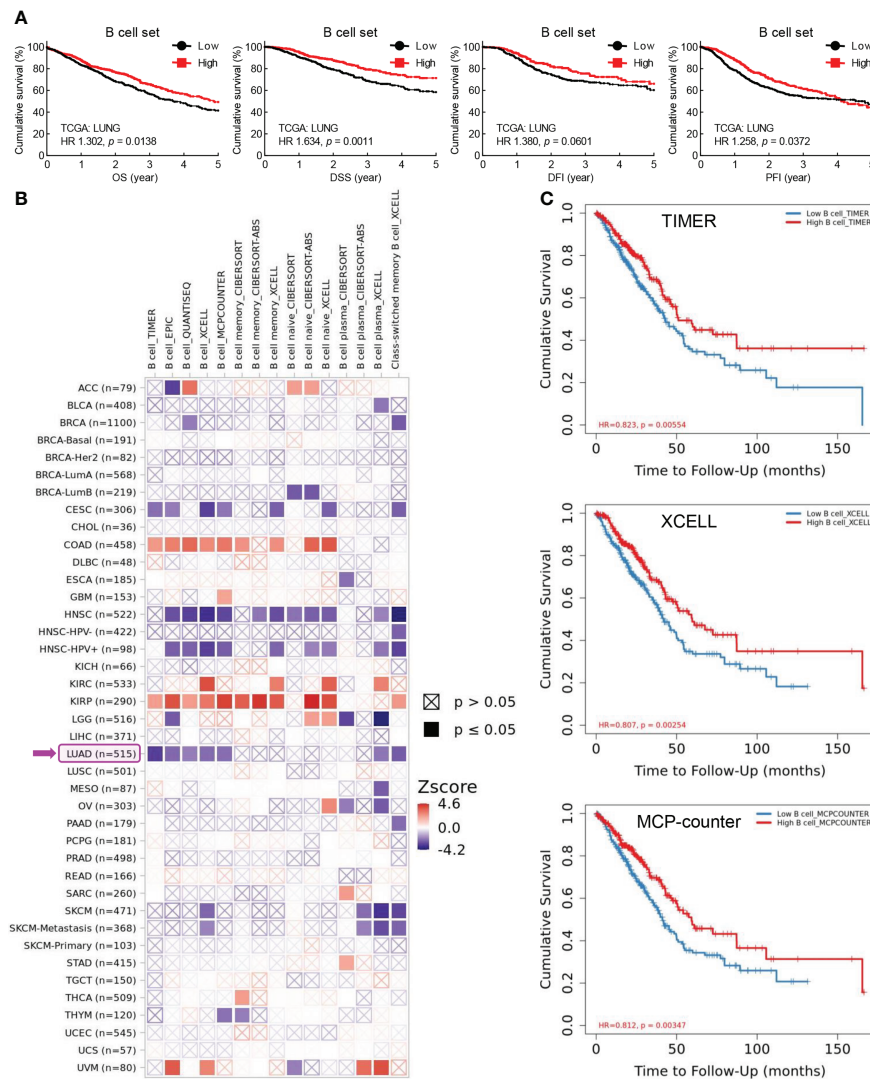


FIGURE 1

Tumor-infiltrating B cells are positively associated with patient survival in lung cancer. **(A)** TCGA LUNG data showing positive correlation between B cell frequency in lung tumor and overall survival (OS), disease specific survival (DSS), disease free interval (DFI), and progression free interval (PFI). B cell set comprises five specific B cell markers (CD19, CD79A, CD79B, BLK, and CD20). **(B)** TIMER data showing association between B cell infiltrates and clinical outcome in human cancers. Z-score > 0 : increased risk, Z-score < 0 : decreased risk. **(C)** Kaplan-Meier curves displaying positive association of B cell infiltrates with patient survival in lung adenocarcinoma (LUAD) by TIMER, XCELL, and MCP-counter algorithms. The infiltration level is equally divided into low and high levels. The hazard ratio (HR) and the log-rank p value for Kaplan-Meier curve were shown in plots **(A, C)**.

clinical outcomes. As shown in Figure 1B, B cell abundance was a favorable biomarker of decreased risk in lung adenocarcinoma (LUAD), which was further evidenced by Kaplan-Meier curves displaying better survival of patients with high tumor infiltration level of B cells estimated by TIMER, XCELL, MCP-counter, EPIC, and QUANTISEQ algorithms (Figure 1C; Supplementary Figure S2B). Strikingly, the prognostic role of tumor-infiltrating B cells was only observed in LUAD, but not in lung squamous cell carcinoma (LUSC) (Figure 1B). This unexpected finding was substantiated by the fact that high level of B cell set was

associated with better survival (OS, DSS, DFI, and PFI) in patients with LUAD but not LUSC (Supplementary Figure S2C).

In addition, high level of CD20, the specific marker for B cell, and B cell set in lung cancer was correlated with better overall survival in patients received chemotherapy, improved clinical outcomes of primary and follow-up treatments, and delayed tumor progression, respectively (Supplementary Figure S3). Taken together, these data suggest that B cells probably exert antitumor function to inhibit tumor progression and promote patient survival in lung cancer.

B cell is reduced during lung tumorigenesis

To clarify the role of B cell in lung cancer, smoking carcinogen urethane-induced endogenous murine lung tumor model was employed (Figure 2A), which faithfully recapitulates the molecular characteristics and histologic patterns of human lung cancer, and in particular adenocarcinoma associated with tobacco smoking, the most common type of lung cancer that makes up about 40% of all lung cancers. Interestingly, the frequency of B cell was significantly and gradually decreased in lung tissues exposed to urethane (Figure 2B). Moreover, urethane-induced lung tumors possessed much fewer tumor-infiltrating B cells in comparison to untreated lung tissues. Consistently, the tumor infiltration of B cells were severely reduced in multiple lung tumor models, including spontaneous lung tumor, subcutaneous lung tumor, and oncogenic *Kras*^{G12D}-induced lung tumor (Figure 2C). To further confirm these data found in FVB/N mice, similar experiments were carried out in BALB/C mice, revealing B cell inhibition in both urethane-treated lung tissues and lung tumors (Figure 2D). These data suggest that B cell is repressed in both lung tumors and surrounding tissues.

B cell depletion leads to increased lung tumorigenesis

What is the function of B cell repression in lung tumorigenesis? To address this question, anti-CD20 antibody was utilized to deplete B cells in smoking carcinogen urethane-induced endogenous lung tumor model, which is widely used to study the mechanisms underlying lung tumorigenesis (Figure 3A). Firstly, B cells were successfully depleted by anti-CD20 antibody as evidenced by remarkably fewer B lymphocytes in multiple tissues and organs from mice received anti-CD20 antibody, including blood, lung, TDLN (mediastinum lymph node), and spleen, compared to that in tissues from control mice (Figure 3B, C). Surprisingly, compared to mice from control group, both lung tumor number and tumor burden were significantly increased in mice underwent B cell depletion (Figure 3D). In detail, more small lung tumors and larger average tumor size and burden were identified in mice administrated with anti-CD20 antibody, suggesting enhanced tumor initiation and progression (Figures 3E, F). In line with this finding, decreased apoptosis rate was detected in lung tumors from mice treated with anti-CD20 antibody (Figures 3G, Supplementary Figure S4A). Meanwhile, α CD20-mediated B cell depletion had minimal effect on TDLN weight and body weight (Supplementary

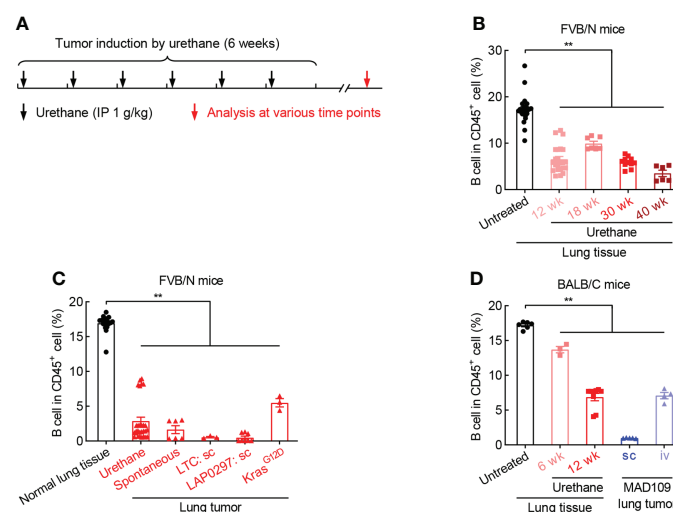


FIGURE 2

B cell is reduced during lung tumorigenesis. (A) Schematic of smoking carcinogen urethane-induced endogenous lung tumorigenesis. (B) FACS data showing decrease of B cell in lung tissues exposed to urethane in FVB/N mice. Untreated (n = 24), Urethane 12 wk (n = 23), Urethane 18 wk (n = 7), Urethane 30 wk (n = 11), Urethane 40 wk (n = 6). (C) FACS data showing decrease of B cell in multiple lung tumor models in FVB/N mice. Normal lung tissue (n = 17), Urethane-induced lung tumor (n = 28), Spontaneous lung tumor (n = 6), LTC: sc (n = 3), LAP0297 sc (n = 8), *Kras*^{G12D}-induced lung tumor (n = 3). (D) FACS data showing loss of B cell in both urethane-induced endogenous lung tumor and xenograft models in BALB/C mice. Untreated (n = 6), Urethane 6 wk (n = 3), Urethane 12 wk (n = 9), MAD109 sc (n = 5), MAD109 iv (n = 4). LTC: Lung Tumor Cell which was established from spontaneous FVB/N lung tumor in our laboratory. LAP0297 and MAD109 are lung cancer cell lines originally derived from spontaneous lung tumors developed in FVB/N and BALB/C mice, respectively. sc, subcutaneous injection; iv, intravenous injection (tail vein). Ordinary one-way ANOVA was performed. Data represented means \pm SEM (B-D). ***p* < 0.01.

Figures S4B, C). These data suggest that B cell deficiency promotes the initiation and progression of lung cancer.

B cell deficiency impairs T cell killing of lung tumor cells

How B cell depletion boosts lung tumorigenesis? Initially, the activity of T cell, the well-known direct killer of tumor cells, was analyzed in the context of anti-CD20 antibody treatment. Unexpectedly, in lung tissues and lung tumors, both CD4⁺ and CD8⁺ T cells expressed comparable levels of IFN γ between B cell-depleted mice and control mice (Figures 4A, B). Similar results were obtained with respect to granzyme B (Granz B), another T cell activation marker (Figures 4C, D). These data

suggest that B cell deficiency has negligible effect on T cell activity in lung cancer.

Although T cell activity unchanged in both lung tissues and lung tumors, the tumor infiltration of immune cells was severely inhibited by B cell depletion (Figure 5A). In detail, lymphocytes, including B cell, CD4⁺ T cell, CD8⁺ T cell, and NK, were significantly decreased in tumors from mice treated with anti-CD20 antibody, compared to that in control mice (Figure 5B), and so were myeloid cells, including macrophage, dendritic cell, monocyte, but not neutrophil (Figure 5C). In parallel, only NK and monocyte were suppressed by α CD20-mediated B cell depletion in lung tissues, whereas CD4⁺ T cell and neutrophil increased (Figures 5D, E), indicating differential roles of anti-CD20 antibody in modulating immune cell populations in lung tumors and lung tissues. It was worth noting that Treg cells were

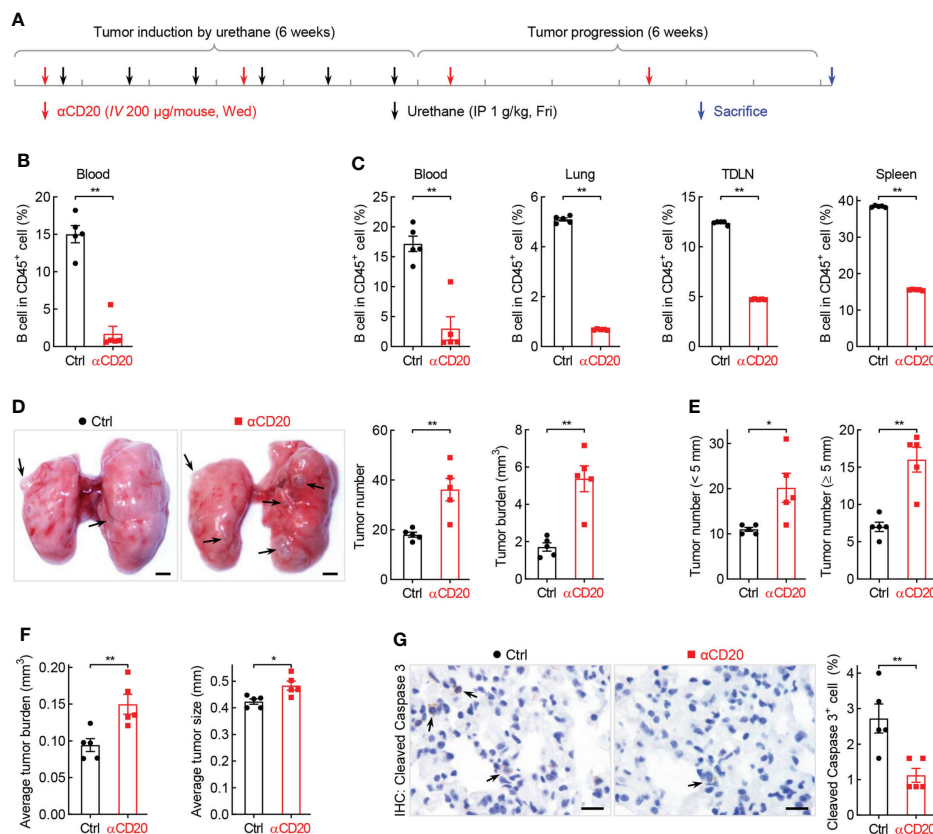


FIGURE 3

B cell depletion leads to increased lung tumorigenesis. (A) Schematic of B cell depletion by α CD20 in smoking carcinogen urethane-induced endogenous lung tumor model. (B) FACS data showing B cell depletion by α CD20 in blood before first dose of urethane ($n = 5$). (C) FACS data showing B cell depletion by α CD20 in blood, lung, tumor-draining lymph node (TDLN), and spleen at the endpoint ($n = 5$). (D) Tumor examination showing increased lung tumor number and tumor burden by α CD20-mediated B cell depletion ($n = 5$). (E) Tumor examination showing elevated both small and large lung tumor numbers by α CD20-mediated B cell depletion ($n = 5$). (F) Tumor examination showing enlarged individual lung tumor by α CD20-mediated B cell depletion ($n = 5$). (G) IHC analysis showing decreased tumor cell apoptosis in mice with α CD20-mediated B cell depletion ($n = 5$). TDLN: mediastinum lymph node. Data shown were representative of two independent experiments with similar results. Scale bar: 1 mm (D) and 20 μ m (G). Student's t test (two tailed, unpaired) was performed (B–G). Data represented means \pm SEM (B–G). * $p < 0.05$; ** $p < 0.01$.

remarkably elevated by B cell depletion in all tissues/organs tested, including tumor, lung, TDLN, and spleen (Figure 5F; Supplementary Figure S5). Taken together, these data suggest that B cell deprivation dampens tumor infiltration of most immune cells, but enhances Treg cells in lung cancer.

B cell deficiency impedes macrophage elimination of lung tumor cells

Besides T cell-mediated tumor eradication, macrophage-guided phagocytosis is critical for tumor cell elimination as well. Surprisingly, both CD24 and CD47, two classic ligands of “don’t eat me” signal, were notably upregulated on tumor cells from mice received anti-CD20 antibody (Figure 6A), whereas comparable expression levels of their receptors were observed on macrophages, Siglec-10 and SIRP α , respectively (Data not shown). These data indicate that B cell deficiency shifts macrophage’s function towards “don’t eat me” signal in lung cancer.

Given B cells are the essential source of antibodies, signaling molecules responding for antibody-dependent cellular phagocytosis

were then examined. Intriguingly, CD64, also known as Fc γ RI which is a high-affinity receptor for IgG to initiate specific phagocytosis of pathogens and tumor cells, was significantly downregulated on tumor-infiltrating macrophages in anti-CD20 antibody-treated mice, compared to that in control mice (Figure 6B). Similar results were achieved in respect of PD-L1 (Figure 6B), which was recently identified as a novel stimulator of phagocytosis in alveolar macrophages (28). Moreover, the expression of PD-L1, but not CD64, was abated on macrophages in lung tissues from B cell-depleted mice as well (Figure 6C). Taken together, these data suggest that B cell deficiency restrains macrophage-mediated elimination of tumor cells in lung cancer.

B cell subsets are diminished by CD20 blockade in lung cancer

Given many B cell subpopulations in lung tissues and lung tumors (30), the effect of anti-CD20 antibody on some primary B cell subsets was surveyed. As expected, memory B cells and Breg cells, both expressing CD20 molecule, were significantly

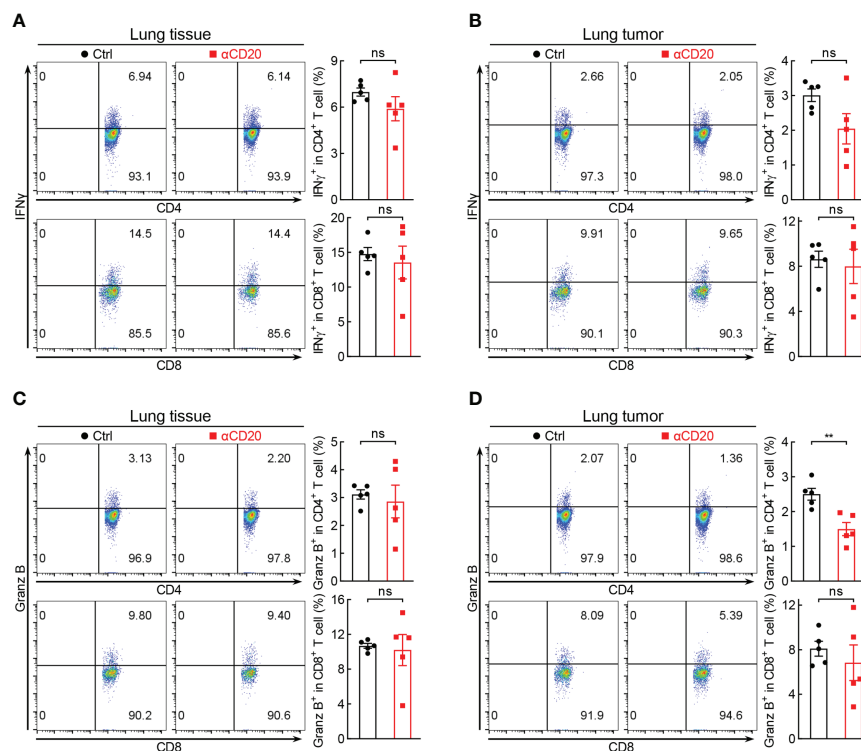


FIGURE 4

B cell deficiency has negligible effect on T cell activity in lung cancer. (A, B) FACS data showing comparable expression levels of IFN γ in both lung tissues and lung tumors between Ctrl and α CD20 groups (*n* = 5). (C, D) FACS data showing comparable expression levels of Granzyme B (granzyme B) in both lung tissues and lung tumors between Ctrl and α CD20 groups (*n* = 5). Data shown were representative of two independent experiments with similar results. Student's *t* test (two tailed, unpaired) was performed. Data represented means \pm SEM. ***p* < 0.01; ns, not statistically significant.

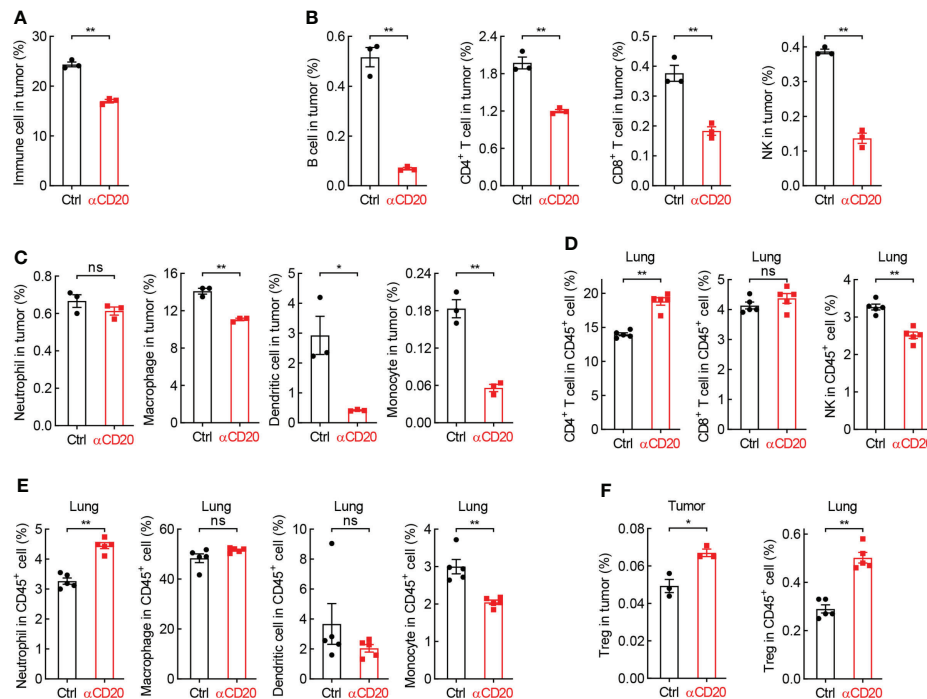


FIGURE 5

B cell deprivation dampens tumor infiltration of immune cells in lung cancer. (A) FACS data showing reduced abundance of immune cells in tumor from mice with αCD20-mediated B cell depletion ($n = 3$). (B, C) FACS data showing decreased tumor-infiltrating lymphocytes (B) and myeloid cells (C) by αCD20-mediated B cell depletion ($n = 3$). (D, E) FACS data showing differential changes of lymphocytes (D) and myeloid cells (E) in lung tissue by αCD20-mediated B cell depletion ($n = 5$). (F) FACS data showing augmented Treg cell frequency in both lung tumor ($n = 3$) and lung tissue ($n = 5$) by αCD20-mediated B cell depletion. Student's t test (two tailed, unpaired) was performed. Data represented means \pm SEM. * $p < 0.05$; ** $p < 0.01$; ns, not statistically significant.

decreased in various tissues/organs from mice treated with anti-CD20 antibody, including lung, TDLN, spleen, and tumor; however, plasma cells only diminished in TDLN and tumor (Figure 7; Supplementary Figure S6). Considering the exacerbated lung tumor initiation and progression by B cell depletion, these data suggest that the pro-tumor function of Breg cells is dominated by the antitumor activity of memory B cells and plasma cells in tumor microenvironment, rendering an overall immune suppression by CD20 blockade in lung cancer.

Discussion

Given the contradictory conclusions of B cells' function in cancer in literature (6–8), to determine the role of B cell in lung tumorigenesis is of great importance for clinical treatment and prevention post CAR-T cell therapy, which eliminates both transformed and normal B cells in most tissues and organs, including lung, the primary targeting site of intravenous infusion. In the present study, tumor-infiltrating B cells are characterized as a positive predictor for delayed cancer progression, improved therapeutic response, and extended survival in patients with lung

cancer, in particular lung adenocarcinoma. In murine lung tumor models, B cell abundance is severely reduced during carcinogenesis, whereas B cell depletion by anti-CD20 antibody significantly promotes the initiation and progression of lung tumor, accompanied with impaired tumor infiltration of immune cells, inhibited phagocytic signaling, and diminished memory B cells and plasma cells. These findings unmask the overall antitumor role of B cells in lung cancer, shedding light on lung cancer pathogenesis and clinical prevention after CAR-T cell therapy.

B cell abundance in tumor was found increased in several papers (21–24), but decreased in others (13–16). In the present study, B cell frequency is severely reduced in both carcinogen-treated lung tissues and lung tumors from various lung cancer models, compared to that in normal lung tissues (Figure 2). This discrepancy is probably caused by cancers with different stages used for analysis. The immune cell composition and B cell subpopulations are evolved with tumor progression and therapy, suggesting a switch from tumor-inhibiting to tumor-promoting function (31, 32). For example, a recently identified novel proangiogenic B cell subset, Breg, and Treg are enriched along with cancer progression, while CTL decreased (33). This hypothesis fits well with the conflicting observations of both positive and negative prognostic roles of B cells in human

cancers (6–8). In addition, carcinogen-induced oncogenic mutation status also has an impact on B cell infiltration into tumor, as evidenced by lower B cell frequency in lung cancer patients with Kras mutation (34, 35). These data indicate that both infiltration level and overall role of B cells in cancers are dependent on tumor stages, mutations, and therapeutic treatments.

Besides antibody production, B cells indirectly restrict lung cancer initiation and progression through regulation of other immune cells. Although comparable activation status of CD4⁺ and CD8⁺ T cells in lung tissues and lung tumors in both groups, there is a remarkable reduction of tumor infiltration of T cells by B cell depletion (Figures 4–5B), resulting in a net inhibition of T cell immunity in tumor microenvironment. In addition, B cells limit Treg function in lung cancer, substantiated by higher Treg frequency by B cell depletion in tumor, lung, TDLN, and spleen (Figure 5F; Supplementary Figure S5; Ref 36). Other immune cells responsible for ADCC and phagocytosis, including NK, macrophage, dendritic cell, and monocyte, are repressed in lung tumors by anti-CD20 antibody treatment (Figures 5B, C). Moreover, the classic and non-classic “don’t eat me” signals are boosted by B cell depletion, whereas “eat me” signals suppressed (Figure 6). These data indicate B cells restrain lung tumor initiation

and progression probably through activating T cell response, NK-mediated ADCC, and APC-dependent phagocytosis.

Tumor-infiltrating B cell frequency not only predicts better survival in patients with lung cancer, but also therapeutic response, including PD-1/PD-L1 immune checkpoint blockade (37–39), which is in part explained by higher PD-L1 expression level in B cell-enriched tumors (40). Another reason is that B cell can reprogram the tumor microenvironment by recruitment of immune cells, turning “cold” lung tumor to “hot” to improve the efficacy of immunotherapy (Figure 5). Furthermore, positive predictable roles of B cells in lung cancer are also found in the context of chemotherapy, primary therapy, and follow-up treatment (Supplementary Figure S3). These data demonstrate B cell as a novel prognostic biomarker for better therapeutic response and survival in patients with lung cancer.

In summary, we find tumor-infiltrating B cell abundance is positively associated delayed cancer progression, improved therapeutic response, and extended survival in patients with lung cancer. In murine lung tumor models, B cell frequency is severely reduced, whereas B cell depletion by anti-CD20 antibody significantly accelerates the initiation and progression of lung tumors, which is mediated by repressed tumor infiltration of

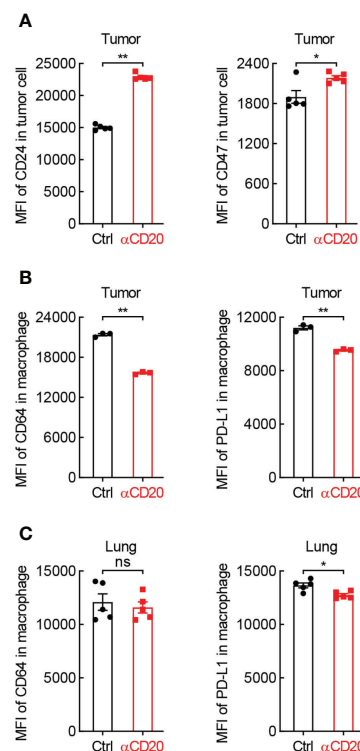


FIGURE 6

B cell shortage impedes the expression of phagocytosis-related genes in lung cancer. (A) FACS data showing elevated expression of CD24 and CD47 on tumor cells from mice with αCD20-mediated B cell depletion (n = 5). (B) FACS data showing impaired expression of PD-L1 and CD64 on tumor-infiltrating alveolar macrophages from mice with αCD20-mediated B cell depletion (n = 3). (C) FACS data showing reduced expression of PD-L1 but not CD64 on alveolar macrophages in lung tissues from mice with αCD20-mediated B cell depletion (n = 5). Student's *t* test (two tailed, unpaired) was performed. Data represented means ± SEM. **p* < 0.05; ***p* < 0.01; ns, not statistically significant.

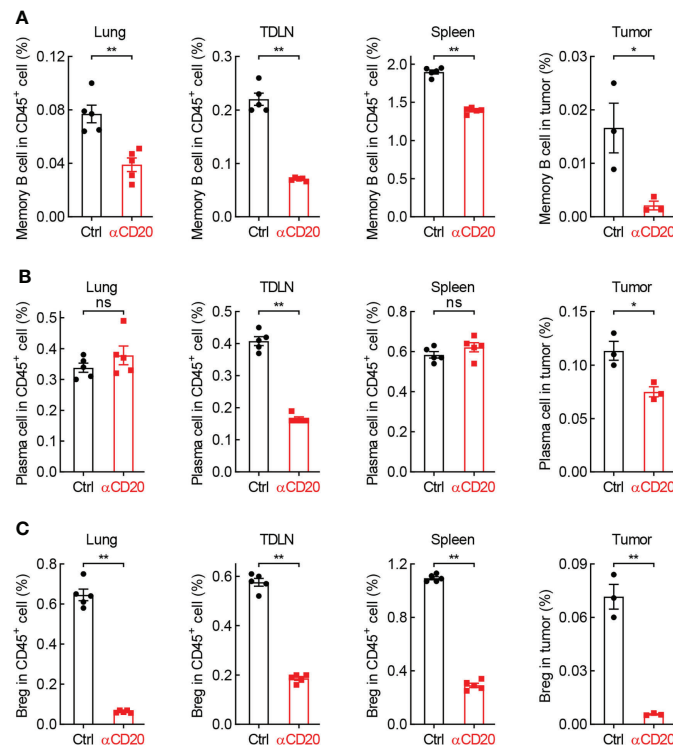


FIGURE 7

B cell subsets are diminished by CD20 blockade in lung cancer. (A) FACS data showing fewer memory B cells in multiple tissues and organs from mice with α CD20-mediated B cell depletion. (B) FACS data showing reduced plasma cells in TDLN and tumor, but not lung tissue and spleen from mice with α CD20-mediated B cell depletion. (C) FACS data showing decreased Breg cells in multiple tissues and organs from mice with α CD20-mediated B cell depletion. TDLN: mediastinum lymph node. Lung tissue: $n = 5$; TDLN: $n = 5$; Spleen: $n = 5$; Lung tumor: $n = 3$. Student's t test (two tailed, unpaired) was performed. Data represented means \pm SEM. * $p < 0.05$; ** $p < 0.01$; ns, not statistically significant.

immune cells, inhibited phagocytic signaling, and diminished memory B cells and plasma cells. Taken together, these findings discover the overall antitumor role of B cells in lung cancer, providing novel insights into cellular mechanisms underlying lung cancer pathogenesis and clinical prevention post CAR-T cell therapy.

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 animal study was reviewed and approved by Animal Ethics Committee of Wuhan University of Science and Technology.

Author contributions

FS, TZ, and WH conceived and designed the study. HW, CC, LG, JL, YY, ML, and YC performed the experiments. HW, CC, XZ, and YL analyzed the data. FS, HZ, and XL wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by grants from the National Natural and Science Foundation of China (82103621, 2017YFE0129100), China Postdoctoral Science Foundation (2021M702538), Department of Education of Hubei Province (B2021023), Department of Science and Technology of Hubei Province (2019ACA168), and Zhongnan Hospital of Wuhan University (LCYF202208, ZNJC202015, PTXM2021019, and cxy2019088).

Acknowledgments

The authors thank Baiyin Yuan and Jianhong Sun for their technical assistance and support in animal experiments, Xiang Zhou and Yao Xu for their critical and constructive advice and feedback.

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.

References

- Zhang X, Zhu L, Zhang H, Chen S, Xiao Y. CAR-T cell therapy in hematological malignancies: Current opportunities and challenges. *Front Immunol* (2022) 13:927153. doi: 10.3389/fimmu.2022.927153
- Wang N, Hu X, Cao W, Li C, Xiao Y, Cao Y, et al. Efficacy and safety of CAR19/22 T-cell cocktail therapy in patients with refractory/relapsed b-cell malignancies. *Blood* (2020) 135:17–27. doi: 10.1182/blood.2019000017
- Li C, Cao W, Que Y, Wang Q, Xiao Y, Gu C, et al. A phase I study of anti-BCMA CAR T cell therapy in relapsed/refractory multiple myeloma and plasma cell leukemia. *Clin Transl Med* (2021) 11:e346. doi: 10.1002/ctm2.346
- Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat Med* (2018) 24:1277–89. doi: 10.1038/s41591-018-0096-5
- Stankovic B, Bjorhovde HAK, Skarshaug R, Aamodt H, Frafjord A, Muller E, et al. Immune cell composition in human non-small cell lung cancer. *Front Immunol* (2018) 9:3101. doi: 10.3389/fimmu.2018.03101
- Fridman WH, Petitprez F, Meylan M, Chen TW, Sun CM, Roumenina LT, et al. B cells and cancer. to b or not to b? *J Exp Med* (2021) 218:e20200851. doi: 10.1084/jem.20200851
- Patel AJ, Richter A, Drayson MT, Middleton GW. The role of b lymphocytes in the immuno-biology of non-small-cell lung cancer. *Cancer Immunol Immunother.* (2020) 69:325–42. doi: 10.1007/s00262-019-02461-2
- Guy TV, Terry AM, Bolton HA, Hancock DG, Shklovskaya E, Fazekas de St. Groth B. Pro- and anti-tumour effects of b cells and antibodies in cancer. *A comparison Clin Stud preclinical models Cancer Immunol Immunother.* (2016) 65:885–96. doi: 10.1007/s00262-016-1848-z
- de Visser KE, Korets LV, Coussens LM. De novo carcinogenesis promoted by chronic inflammation is b lymphocyte dependent. *Cancer Cell* (2005) 7:411–23. doi: 10.1016/j.ccr.2005.04.014
- Affara NI, Ruffell B, Medler TR, Gunderson AJ, Johansson M, Bornstein S, et al. B cells regulate macrophage phenotype and response to chemotherapy in squamous carcinomas. *Cancer Cell* (2014) 25:809–21. doi: 10.1016/j.ccr.2014.04.026
- Schioppa T, Moore R, Thompson RG, Rosser EC, Kulbe H, Nedospasov S, et al. B regulatory cells and the tumor-promoting actions of TNF-alpha during squamous carcinogenesis. *Proc Natl Acad Sci* (2011) 108:10662–7. doi: 10.1073/pnas.1100994108
- Qin Z, Richter G, Schuler T, Ibe S, Cao X, Blankenstein T. B cells inhibit induction of T cell-dependent tumor immunity. *Nat Med* (1998) 4:627–30. doi: 10.1038/nm0598-627
- Dong HP, Elstrand MB, Holth A, Silins I, Berner A, Trope CG, et al. NK- and b-cell infiltration correlates with worse outcome in metastatic ovarian carcinoma. *Am J Clin Pathol* (2006) 125:451–8. doi: 10.1309/15B66DQMFY78CJ
- Lundgren S, Berntsson J, Nodin B, Micke P, Jirstrom K. Prognostic impact of tumour-associated b cells and plasma cells in epithelial ovarian cancer. *J Ovarian Res* (2016) 9:21. doi: 10.1186/s13048-016-0232-0
- Yang C, Lee H, Pal S, Jove V, Deng J, Zhang W, et al. B cells promote tumor progression via STAT3 regulated-angiogenesis. *PLoS One* (2013) 8:e64159. doi: 10.1371/journal.pone.0064159
- Barbera-Guillem E, Nelson MB, Barr B, Nyhus JK, May KF Jr., Feng L, et al. B lymphocyte pathology in human colorectal cancer. experimental and clinical therapeutic effects of partial b cell depletion. *Cancer Immunol Immunother.* (2000) 48:541–9. doi: 10.1007/PL00006672
- DiLillo DJ, Yanaba K, Tedder TF. B cells are required for optimal CD4+ and CD8+ T cell tumor immunity: therapeutic b cell depletion enhances B16 melanoma growth in mice. *J Immunol* (2010) 184:4006–16. doi: 10.4049/jimmunol.0903009
- Li Q, Lao X, Pan Q, Ning N, Yet J, Xu Y, et al. Adoptive transfer of tumor reactive b cells confers host T-cell immunity and tumor regression. *Clin Cancer Res* (2011) 17:4987–95. doi: 10.1158/1078-0432.CCR-11-0207
- Bodogai M, Moritoh K, Lee-Chang C, Hollander CM, Sherman-Baust CA, Wersto RP, et al. Immunosuppressive and prometastatic functions of myeloid-derived suppressive cells rely upon education from tumor-associated b cells. *Cancer Res* (2015) 75:3456–65. doi: 10.1158/0008-5472.CAN-14-3077
- Tao H, Lu L, Xia Y, Dai F, Wang Y, Bao Y, et al. Antitumor effector b cells directly kill tumor cells via the Fas/FasL pathway and are regulated by IL-10. *Eur J Immunol* (2015) 45:999–1009. doi: 10.1002/eji.201444625
- Zhang Y, Yin X, Wang Q, Song X, Xia W, Mao Q, et al. A novel gene expression signature-based on b-cell proportion to predict prognosis of patients with lung adenocarcinoma. *BMC cancer* (2021) 21:1098. doi: 10.1186/s12885-021-08805-5
- Germain C, Gnjatich S, Tamzalit F, Knockaert S, Remark R, Goc J, et al. Presence of b cells in tertiary lymphoid structures is associated with a protective immunity in patients with lung cancer. *Am J Respir Crit Care Med* (2014) 189:832–44. doi: 10.1164/rccm.201309-1611OC
- Cui C, Wang J, Fagerberg E, Chen PM, Connolly KA, Damo M, et al. Neoantigen-driven b cell and CD4 T follicular helper cell collaboration promotes anti-tumor CD8 T cell responses. *Cell* (2021) 184:6101–18. e6113. doi: 10.1016/j.cell.2021.11.007
- Bruno TC, Ebner PJ, Moore BL, Squalls OG, Waugh KA, Eruslanov EB, et al. Antigen-presenting intratumoral b cells affect CD4(+) TIL phenotypes in non-small cell lung cancer patients. *Cancer Immunol Res* (2017) 5:898–907. doi: 10.1158/2326-6066.CIR-17-0075
- Sun F, Guo ZS, Gregory AD, Shapiro SD, Xiao G, Qu Z. Dual but not single PD-1 or TIM-3 blockade enhances oncolytic virotherapy in refractory lung cancer. *J Immunother Cancer* (2020) 8:e000294. doi: 10.1136/jitc-2019-000294
- Sun F, Qu Z, Xiao Y, Zhou J, Burns TF, Stabile LP, et al. NF-kappaB1 p105 suppresses lung tumorigenesis through the Tpl2 kinase but independently of its NF-kappaB function. *Oncogene* (2016) 35:2299–310. doi: 10.1038/onc.2015.299
- Sun F, Li L, Yan P, Zhou J, Shapiro SD, Xiao G, et al. Causative role of PDLIM2 epigenetic repression in lung cancer and therapeutic resistance. *Nat Commun* (2019) 10:5324. doi: 10.1038/s41467-019-13331-x
- Sun F, Li L, Xiao Y, Gregory AD, Shapiro SD, Xiao G, et al. Alveolar macrophages inherently express programmed death-1 ligand 1 for optimal protective immunity and tolerance. *J Immunol* (2021) 207:110–4. doi: 10.4049/jimmunol.2100046
- Li L, Sun F, Han L, Liu X, Xiao Y, Gregory AD, et al. PDLIM2 repression by ROS in alveolar macrophages promotes lung tumorigenesis. *JCI Insight* (2021) 6:e144394. doi: 10.1172/jci.insight.144394

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.2022.1006477/full#supplementary-material>

30. Chen J, Tan Y, Sun F, Hou L, Zhang C, Ge T, et al. Single-cell transcriptome and antigen-immunoglobulin analysis reveals the diversity of b cells in non-small cell lung cancer. *Genome Biol* (2020) 21:152. doi: 10.1186/s13059-020-02064-6
31. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* (2020) 11:2285. doi: 10.1038/s41467-020-16164-1
32. Maynard A, McCoach CE, Rotow JK, Harris L, Haderk F, Kerr DL, et al. Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing. *Cell* (2020) 182:1232–51.e1222. doi: 10.1016/j.cell.2020.07.017
33. van de Veen W, Globinska A, Jansen K, Straumann A, Kubo T, Verschoor D, et al. A novel proangiogenic b cell subset is increased in cancer and chronic inflammation. *Sci Adv* (2020) 6:eaz3559. doi: 10.1126/sciadv.aaz3559
34. Pinto R, Petriella D, Lacalamita R, Montrone M, Catino A, Pizzutilo P, et al. KRAS-driven lung adenocarcinoma and b cell infiltration: Novel insights for immunotherapy. *Cancers* (2019) 11:1145. doi: 10.3390/cancers11081145
35. Li S, MacAlpine DM, Counter CM. Capturing the primordial kras mutation initiating urethane carcinogenesis. *Nat Commun* (2020) 11:1800. doi: 10.1038/s41467-020-15660-8
36. Germain C, Devi-Marulkar P, Knockaert S, Biton J, Kaplon H, Letaief L, et al. Tertiary lymphoid structure-b cells narrow regulatory T cells impact in lung cancer patients. *Front Immunol* (2021) 12:626776. doi: 10.3389/fimmu.2021.626776
37. Budczies J, Kirchner M, Kluck K, Kazdal D, Glade J, Allgauer M, et al. A gene expression signature associated with b cells predicts benefit from immune checkpoint blockade in lung adenocarcinoma. *Oncoimmunology* (2021) 10:1860586. doi: 10.1080/2162402X.2020.1860586
38. Xia L, Guo L, Kang J, Yang Y, Yao Y, Xia W, et al. Predictable roles of peripheral IgM memory b cells for the responses to anti-PD-1 monotherapy against advanced non-small cell lung cancer. *Front Immunol* (2021) 12:759217. doi: 10.3389/fimmu.2021.759217
39. Patil NS, Nabat BY, Muller S, Koeppen H, Zou W, Giltane J, et al. Intratumoral plasma cells predict outcomes to PD-L1 blockade in non-small cell lung cancer. *Cancer Cell* (2022) 40:289–300.e284. doi: 10.1016/j.ccell.2022.02.002
40. Ho KH, Chang CJ, Huang TW, Shih CM, Liu AJ, Chen PH, et al. Gene landscape and correlation between b-cell infiltration and programmed death ligand 1 expression in lung adenocarcinoma patients from the cancer genome atlas data set. *PloS One* (2018) 13:e0208459. doi: 10.1371/journal.pone.0208459



OPEN ACCESS

EDITED BY

Liang Cheng,
Harbin Medical University, China

REVIEWED BY

Nizhuan Wang,
ShanghaiTech University, China
Shuaiqun Wang,
Shanghai Maritime University, China

*CORRESPONDENCE

Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

[†]These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 19 July 2022

ACCEPTED 14 September 2022

PUBLISHED 29 September 2022

CITATION

Jian F, Huang F, Zhang Y-H, Huang T
and Cai Y-D (2022) Identifying anal
and cervical tumorigenesis-associated
methylation signaling with machine
learning methods.
Front. Oncol. 12:998032.
doi: 10.3389/fonc.2022.998032

COPYRIGHT

© 2022 Jian, Huang, Zhang, Huang and
Cai. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Identifying anal and cervical tumorigenesis-associated methylation signaling with machine learning methods

Fangfang Jian^{1†}, FeiMing Huang^{2†}, Yu-Hang Zhang³,
Tao Huang^{4,5*} and Yu-Dong Cai^{2*}

¹Department of Obstetrics & Gynecology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, ²School of Life Sciences, Shanghai University, Shanghai, China,

³Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, ⁴Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, ⁵CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

Cervical and anal carcinoma are neoplastic diseases with various intraepithelial neoplasia stages. The underlying mechanisms for cancer initiation and progression have not been fully revealed. DNA methylation has been shown to be aberrantly regulated during tumorigenesis in anal and cervical carcinoma, revealing the important roles of DNA methylation signaling as a biomarker to distinguish cancer stages in clinics. In this research, several machine learning methods were used to analyze the methylation profiles on anal and cervical carcinoma samples, which were divided into three classes representing various stages of tumor progression. Advanced feature selection methods, including Boruta, LASSO, LightGBM, and MCFS, were used to select methylation features that are highly correlated with cancer progression. Some methylation probes including cg01550828 and its corresponding gene RNF168 have been reported to be associated with human papilloma virus-related anal cancer. As for biomarkers for cervical carcinoma, cg27012396 and its functional gene HDAC4 were confirmed to regulate the glycolysis and survival of hypoxic tumor cells in cervical carcinoma. Furthermore, we developed effective classifiers for identifying various tumor stages and derived classification rules that reflect the quantitative impact of methylation on tumorigenesis. The current study identified methylation signals associated with the development of cervical and anal carcinoma at qualitative and quantitative levels using advanced machine learning methods.

KEYWORDS

cervical carcinoma, anal carcinoma, DNA methylation, machine learning, classification rule

1 Introduction

Anal carcinoma is a malignant proliferative disease associated with anal abnormalities (1). With strong sex bias (females have higher mortality), more than 1000 people die of anal carcinoma per year in the United States (2, 3). The risk factors for anal carcinoma include aging, sex (more than two-thirds of patients are women), smoking, and most importantly human papilloma virus (HPV) infection (4, 5). Cervical carcinoma occurs in the cervix, which is located beneath the uterus and connects to the vagina (6). Smoking, immune suppression caused by human immunodeficiency virus (HIV) infection, and HPV infection are the major risk factors for cervical carcinoma (7). Both anal carcinoma and cervical carcinoma are malignant diseases associated with HPV infection. However, HPV cannot directly trigger the initiation and progression of such malignant diseases. The underlying mechanisms for HPV-mediated cancer initiation and progression have not been fully revealed. Therefore, for a long time, finding potential carcinogenic mechanisms associated with HPV infection and related biomarkers in anal and cervical carcinoma have been one of the major challenges in this field.

DNA methylation is a common biological process that regulates the activity of a DNA segment without changing the sequence (8, 9). It has been shown to be abnormally regulated during tumorigenesis in multiple cancer subtypes including anal and cervical carcinoma (10, 11). The demethylation of oncogenic genes and the methylation of tumor suppressors have been widely observed in cancers (8). In anal and cervical carcinoma, a genome-wide host methylation profiling under HIV infection revealed the potential associations between abnormal methylation status and anal and cervical carcinogenesis by monitoring the methylation alteration from normal to intraepithelial neoplasm and malignant tumorigenesis (12). Potential epigenetic markers to predict cancer risk and drive carcinogenesis around genes such as ASCL1, ATP10A, and CCDC81 have been identified. However, the quantitative association between biomarkers and disease risk has not been fully established.

In the present study, the methylation data, retrieved from Gene Expression Omnibus (GEO) database, on anal and cervical carcinoma samples was investigated. Three stages: normal control, intraepithelial neoplasia (also known as stage 0 of tumorigenesis reflecting the intermediate stage), and tumor, were included. To reveal the underlying biomarkers for distinguishing different stages, we applied multiple machine learning algorithms on the methylation data, which treated methylation as features. Some essential methylation sites were extracted, which can be latent biomarkers to distinguish different stages. Furthermore, some quantitative rules were also discovered for carcinogenesis monitoring, also indicating the different methylation patterns on various stages. Finally, some perfect classifiers were built to identify the stage of samples. All

in all, this study provided a novel effective computational analysis for cancer biomarker recognition and progression monitoring on anal and cervical carcinoma.

2 Materials and methods

2.1 Data

The methylation profiling of 143 anal carcinoma samples and 28 cervical carcinoma samples was accessed from the GEO database under the accession number GSE186859 (12). The three different stages of cancer were involved in the 143 anal carcinoma samples: 9 normal samples, 13 anal intraepithelial neoplasia-3 (AIN3) samples, and 121 tumor samples. Similarly, the 28 cervical carcinoma samples contained 10 normal samples, 9 cervical intraepithelial neoplasia-3 (CIN3) samples, and 9 tumor samples. AIN3 or CIN3 is an intermediate state between normal and tumor. The 485,512 methylation probes were extracted for each anal and cervical carcinoma sample.

2.2 Boruta feature filtering

Because of the enormous number of original methylation features and limited methylations related to anal or cervical carcinomas, Boruta was employed for initial filtering (13–16).

Boruta is a random forest (RF)-based feature selection method for confirming whether variables in the classification are statistically superior to random variables. In a nutshell, Boruta analysis compares all variables to random variables, which are duplicates of the original variables by shuffling. RF is used to evaluate the importance of all variables, including actual and random variables. Actual variables that outperform the best random variables are labeled as confirmed, whereas those that do not outperform the best of the random variables are labeled as rejected. The above procedures are repeated numerous times, resulting in a binomial distribution for the binary outcome (confirmed or denied) of a series of n trials. The variables in the rejection region of the distribution were removed, whereas those in the acceptance region are preserved. To obtain the best classification accuracy, Boruta selects features that are strongly and weakly important, unlike wrapper techniques, which strive to discover a few powerfully relevant features.

The Boruta program used in this study was obtained at https://github.com/scikit-learn-contrib/boruta_py. It was performed on anal and cervical carcinoma samples using default parameters, respectively. Key methylation features for anal and cervical carcinomas were selected for further analysis, respectively.

2.3 Feature ranking algorithms

With Boruta, key methylation features for anal and cervical carcinomas can be obtained, respectively. However, they evidently provided different roles to depict anal or cervical carcinomas. Thus, further investigation was necessary. Here, three feature ranking algorithms: Monte Carlo feature selection (MCFS) (17), light gradient boosting machine (LightGBM) (18) and least absolute shrinkage and selection operator (LASSO) (19), followed to uncover the importance of features selected by Boruta. Their brief descriptions were as follows.

2.3.1 Monte Carlo feature selection

In MCFS, the importance of features are determined according to their roles in multiple decision trees (DTs) (17). This method has been commonly used to process biological data (20–22). As part of the current study, t classification trees are built based on m randomly chosen methylation features and random division of training and test samples. Such procedures are executed s times. Consequently, $s \times t$ DTs are built, based on which a measurement, relative importance (RI), is computed for each feature. Such measurement is determined by how many times it has been selected in these $s \times t$ trees and how much it contributes to predicting the class of the $s \times t$ trees. It can be estimated as follows:

$$RI_g = \sum_{\tau=1}^s (wAcc)^u \sum_{n_g(\tau)} IG(n_g(\tau)) \left(\frac{no.in\ n_g(\tau)}{no.in\ \tau} \right)^v, \quad (1)$$

where $wAcc$ is the weighted accuracy, $IG(n_g(\tau))$ is the information gain (IG) of node $n_g(\tau)$, $(no.in\ n_g(\tau))$ is the number of samples in node $n_g(\tau)$, and $(no.in\ \tau)$ is the sample sizes in the tree root. u and v are two settled positive integers.

The MCFS program was downloaded at <https://home.ipipan.waw.pl/m.draminski/mcfs.html>, which was executed using default parameters. According to the decreasing order of RI values, features were ranked in a list, named MCFS feature list.

2.3.2 Light gradient boosting machine

The LightGBM algorithm uses a gradient boosting framework, and it is an improved version of the gradient boosting DT with the advantages of high efficiency, support for parallelism, and large-scale data processing (18). The total number of times each feature participated in the trees is used by LightGBM to evaluate the importance of features. The higher the frequency with which features are selected, the more important they are. Based on this criterion, features can be ranked in a list with the decreasing order of times.

In this study, we used the LightGBM program (<https://lightgbm.readthedocs.io/en/latest/>), implemented by Python. It was run under default parameters. The feature list generated by LightGBM was called LightGBM feature list.

2.3.3 Least absolute shrinkage and selection operator

LASSO is a classic feature selection method (19). In this method, L1 paradigm is used to create a penalty function that selectively eliminates features by imposing a higher penalty on features with higher coefficients and more prediction errors, resulting in a model with fewer features and less overfitting. The coefficients of input features that do not contribute favorably to the prediction of a machine learning model are scaled down. As a result, the coefficients of the features are used to rank features in a list.

Here, the LASSO package collected in Scikit-learn (23) was used. Likewise, default parameters were used. For clear descriptions, the list yielded by LASSO was termed as LASSO feature list.

2.4 Incremental feature selection

Based on one feature ranking algorithm, a feature list can be obtained. However, which features are optimal for classification is still a problem. This study adopted incremental feature selection (IFS) method (24–28) to analyze the list and extract optimal features for a given classification algorithm. In this method, the feature list with n features is first divided into n feature subsets, with the number of features differing by 1 in turn. Subsequently, the feature subsets and target variables are fed into one classification algorithm to construct classifiers. Their classification performance is evaluated through 10-fold cross-validation (29). The optimal feature subset for one classification algorithm is defined as the subset of features with the highest classification performance and the classifier with the optimal feature subset is defined as the optimal classifier.

2.5 Synthetic minority oversampling technique

Two datasets for anal and cervical carcinomas, respectively, were investigated in this study. As mentioned in Section 2.1, the anal carcinoma dataset was imbalanced. In such dataset, tumor samples were about 13 times as many as normal samples. The classifiers directly built on such dataset would create bias. It was necessary to tackle such problem. In this study, synthetic minority oversampling technique (SMOTE) was adopted (30–32).

SMOTE is an oversampling method for dealing with imbalanced problems. It generates new samples for each minority class until the sizes of all classes are same. The samples of the minority class are synthesized by first selecting one sample to serve as a seed sample and then randomly selecting one of the k -nearest neighbors for linear combination. The synthesis formula is as follows:

$$s = x + \beta(x - y), \quad (2)$$

where x represents the feature vector of the seed sample, y represents the feature vector of its randomly selected neighbor, and β is a random value between 0 and 1. In this study, the SMOTE program downloaded from <https://github.com/scikitlearn-contrib/imbalanced-learn> was used. Default parameters were adopted to execute this program.

2.6 Classification algorithm

To execute IFS method, one classification algorithm was necessary. Two classic classification algorithms: RF (33) and DT (34), were attempted in this study as they are widely used in dealing with biological and medical problems (35–40). The below text gave the brief descriptions on these two algorithms.

RF is one of the most classic and powerful classification algorithms in machine learning. In fact, it is an ensemble algorithm containing multiple DTs. To construct each DT, samples are randomly selected, with replacement, from the original dataset and the selected sample number is equal to the number of samples in the original dataset. Furthermore, features are also randomly chosen from all features. RF integrates constructed DTs with majority voting. It is quite interesting that although DT is quite weak, RF is much more powerful and can avoid overfitting. Thus, it was adopted in this study to construct efficient classifiers.

Above-mentioned RF is generally much stronger than DT. However, it also loses the merits of its component DT. It is widely accepted that DT is a white-box algorithm, which means that its decision-making process is completely open. This makes it possible for us to understand the principle of DT. For the problems investigated in this study, DT can help us uncover essential methylation differences on three stages of anal and cervical carcinomas, thereby improving our comprehension on these two carcinomas. Generally, DT is a tree-like structure. There are two node types in this structure. One is branch node, which is in charge of determining which branch a test sample goes through down. The other is leaf node, which determines the class of the test sample reaching the leaf node. Besides, DT can also be represented by a set of classification rules. Each rule is generated by a path from the root to one leaf node. The investigation of these rules can uncover the different patterns of various stages of anal and cervical carcinomas.

To quickly implement DT and RF, related packages in Scikit-learn (23) were employed. These packages were executed using default parameters.

2.7 Performance evaluation

The weighted F1 was adopted to assess the overall performance of classifiers. To calculate such measurement, the

F1 score on each class should be calculated first, which is defined as

$$F1 \text{ score}_i = \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i}, \quad (3)$$

where TP_i is the true positive for the i -th class, FP_i and FN_i stand for the false positive and false negative for the i -th class. The weighted F1 is defined as the weighted mean of F1 scores on all classes. On the other hand, the direct mean of F1 scores on all classes defines another measurement, macro F1, which was also provided in this study.

Moreover, the accuracy (ACC) and Matthew correlation coefficients (MCC) (41) were also used in this study. ACC is the most classic measurement, which is defined as the proportion of correctly predicted samples. MCC is much more perfect than ACC when the class sizes are quite different. To compute the MCC, two binary matrices X and Y should be constructed in advance, where X and Y stores the true and predicted class of each sample, respectively. Then, MCC can be calculated by

$$MCC = \frac{\text{cov}(X, Y)}{\sqrt{\text{cov}(X, X)\text{cov}(Y, Y)}} \quad (4)$$

3 Results

In the present study, we developed a robust computational pipeline, which combined several machine learning algorithms. The entire procedures are illustrated in Figure 1. The detailed results were listed as below.

3.1 Results of Boruta and feature ranking algorithms

As lots of methylation features were used to represent each sample. Boruta was adopted for preliminary feature filtering. On anal carcinoma dataset, 571 methylation features were selected by Boruta, whereas 26 features were selected on the cervical carcinoma dataset. The selected features on two datasets are provided in Supplementary Table S1.

Subsequently, three feature ranking algorithms were used on both datasets to rank the filtered features by their importance. On each dataset, three feature lists were obtained, which are available in Supplementary Table S1. On the anal carcinoma dataset, three features were assigned RI values 0 by MCFS method. Thus, they were removed from the MCFS feature list. Furthermore, a biological analysis of how the top-ranked features affected the development of anal or cervical carcinomas would be given in Section 4.1.

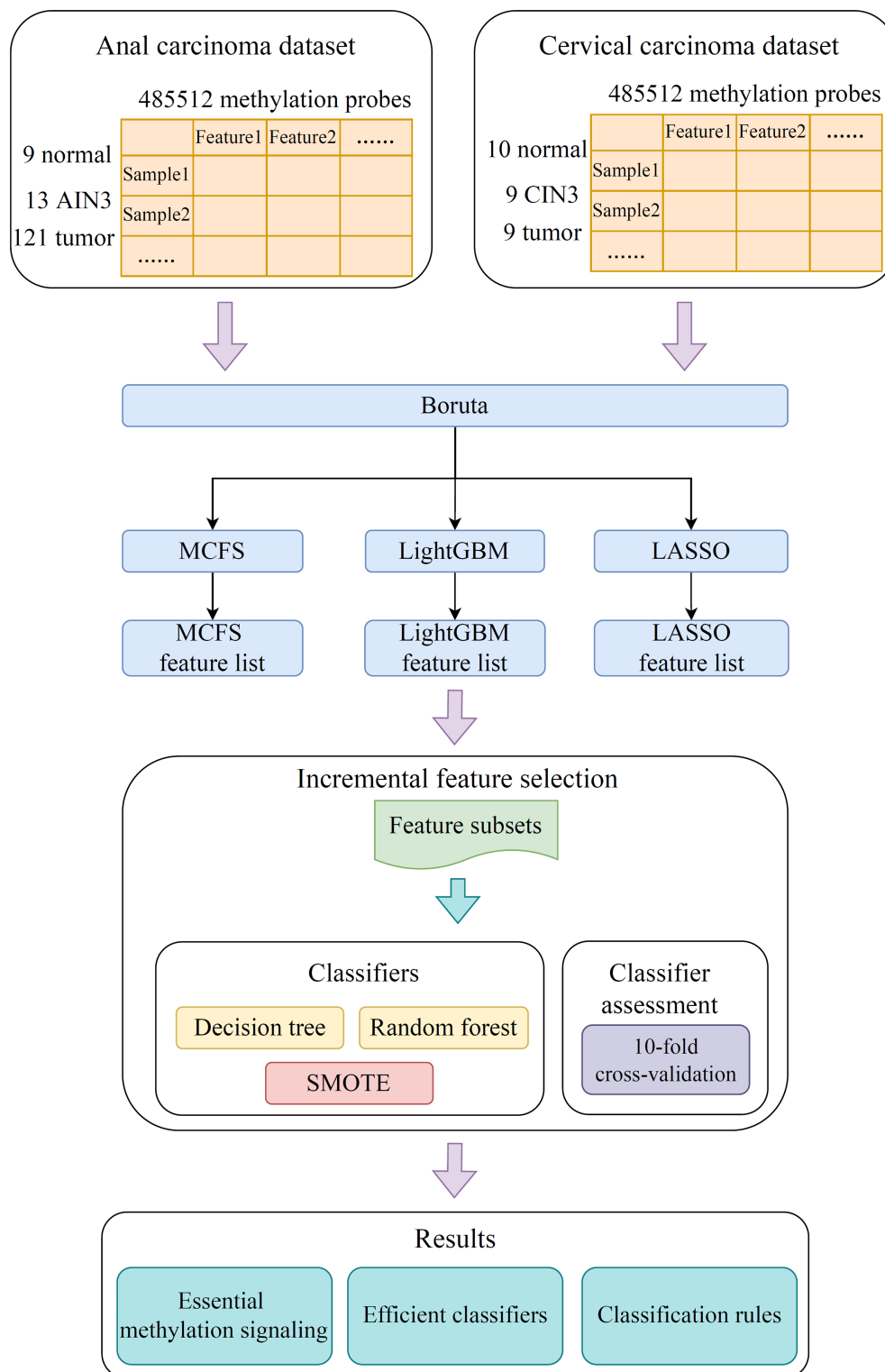


FIGURE 1

Flow chart of the entire analysis process. The 485,512 methylation probes in the anal or cervical carcinoma dataset are filtered by Boruta and ranked according to feature importance by using three feature ranking algorithms, namely, MCFS, LightGBM, and LASSO. Afterward, each of three feature lists is fed into the incremental feature selection (IFS) computational framework containing two efficient classification algorithms (decision tree, random forest) to extract essential methylations, construct efficient classifiers and classification rules.

3.2 Results of IFS method

Based on the three feature lists on each dataset, IFS method was executed using RF or DT as the classification algorithm. All possible feature subsets were constructed and RF or DT classifier was built on each of them, which was evaluated by 10-fold cross-validation. The cross-validation results are provided in [Supplementary Table S2](#).

On the anal carcinoma dataset, the performance of classifiers, measured by weighted F1, was illustrated by six IFS curves, as shown in [Figure 2A](#), in which weighted F1 was set as Y-axis and number of features was defined as X-axis. When DT was used in the IFS method, the highest weighted F1 values on the LASSO, MCFS and LightGBM feature lists, were all 0.993. Such performance was obtained by using top 215, 17 and 6 features in three feature lists, respectively. These features also comprised the optimal feature subsets for DT, on which three optimal DT classifiers were constructed. Their overall performance, measured by ACC, MCC and Macro F1, is listed in [Table 1](#). Interestingly, their performance was same with ACC of 0.993, MCC of 0.975 and macro F1 of 0.981. Furthermore, the performance (F1 score) of these three DT optimal classifiers on three stages (normal, AIN3 and tumor) is shown in [Figure 3A](#). The three classifiers also provided equal performance on three stages (0.947 on normal, 1.000 on AIN3 and 0.996 on tumor). Above results indicated the good performance of three optimal DT classifiers. As for the IFS results with RF, three curves were also plotted, as shown in [Figure 2A](#). RF provided the perfect performance (weighted F1 = 1) on all three feature lists when top 13, 15 and 5 features in the LASSO, MCFS and LightGBM lists, respectively, were used. These features constituted the optimal feature subsets for RF on different lists. Accordingly, three optimal RF classifiers were set up with the optimal feature subsets. The ACC, MCC and macro F1 values of these classifiers are listed in [Table 1](#) and their performance on three classes is shown in [Figure 3A](#). All measurements were equal to 1.000, also suggesting the perfect performance of three optimal RF classifiers.

On the cervical carcinoma dataset, the same IFS procedure was conducted. Three curves for DT and RF, respective, are plotted, as shown in [Figure 2B](#). For IFS results with DT, the highest weighted F1 on the MCFS feature list was 1.000 and it was 0.964 on other two lists. The optimal feature subsets were constructed by picking up top 19, 4 and 18 features in the LASSO, MCFS and LightGBM lists, respectively. On these feature subsets, three optimal DT classifiers were set up. Their ACC, MCC and macro F1 values are listed in [Table 2](#). Clearly, the optimal DT classifier on MCFS feature list provided perfect values on three measurements and the other two classifiers gave lower performance with ACC of 0.964, MCC of 0.948 and macro F1 of 0.965. Their performance (F1 score) on three stages (normal, CIN3 and tumor) is illustrated in [Figure 3B](#). Again,

the optimal DT classifier on MCFS feature list provided the perfect performance on all three stages and the other two classifiers yielded the same performance on three stages (0.952 on normal, 0.941 on CIN3 and 1.000 on tumor). Evidently, all three optimal DT classifiers generated perfect or nearly perfect performance. As for IFS results with RF, when top 5, 4, and 19 features in the LASSO, LightGBM and MCFS lists were adopted, RF produced perfect performance. The optimal feature subsets were constructed using these features and three optimal RF classifiers were built with them. These classifiers also provided perfect performance measured by other measurements ([Table 2](#) and [Figure 3B](#)).

With the above IFS results, the optimal RF classifiers generally provided better performance than the optimal DT classifiers. All optimal RF classifiers yielded perfect performance, suggesting that they can be efficient tools to classify anal or cervical carcinoma samples.

3.3 Classification rules

One of the main purposes of this study was to depict the methylation patterns for two carcinomas on different stages. On anal carcinoma dataset, the top features in the LightGBM feature list were selected as they yielded the highest performance and they were least. The DT was applied on all anal carcinoma samples represented by these five features, yielding four rules, as listed in [Table 3](#). Two rules were for identifying tumor samples and one rule was for predicting AIN3 and normal samples, respectively. Similarly, on the cervical carcinoma dataset, we selected the top four features in the MCFS feature list to construct the classification rules. Three rules were generated, as shown in [Table 4](#). Each stage was assigned one rule. These rules would be discussed in detail in Section 4.2.

4 Discussion

By employing multiple machine learning algorithms, methylation datasets on anal and cervical carcinomas were deeply analyzed. Three feature lists, generated by three feature ranking algorithms, were obtained for each dataset. The methylation features with high ranks in three lists may be essential for two carcinomas, which can be novel methylation biomarkers associated with carcinoma progression from normal to precancerous lesions and from precancerous lesions to malignant cancer in anal and cervical carcinomas. Some of them were discussed in this section. Furthermore, some rules were set up for anal and cervical carcinomas, respectively. They were also discussed in the section.

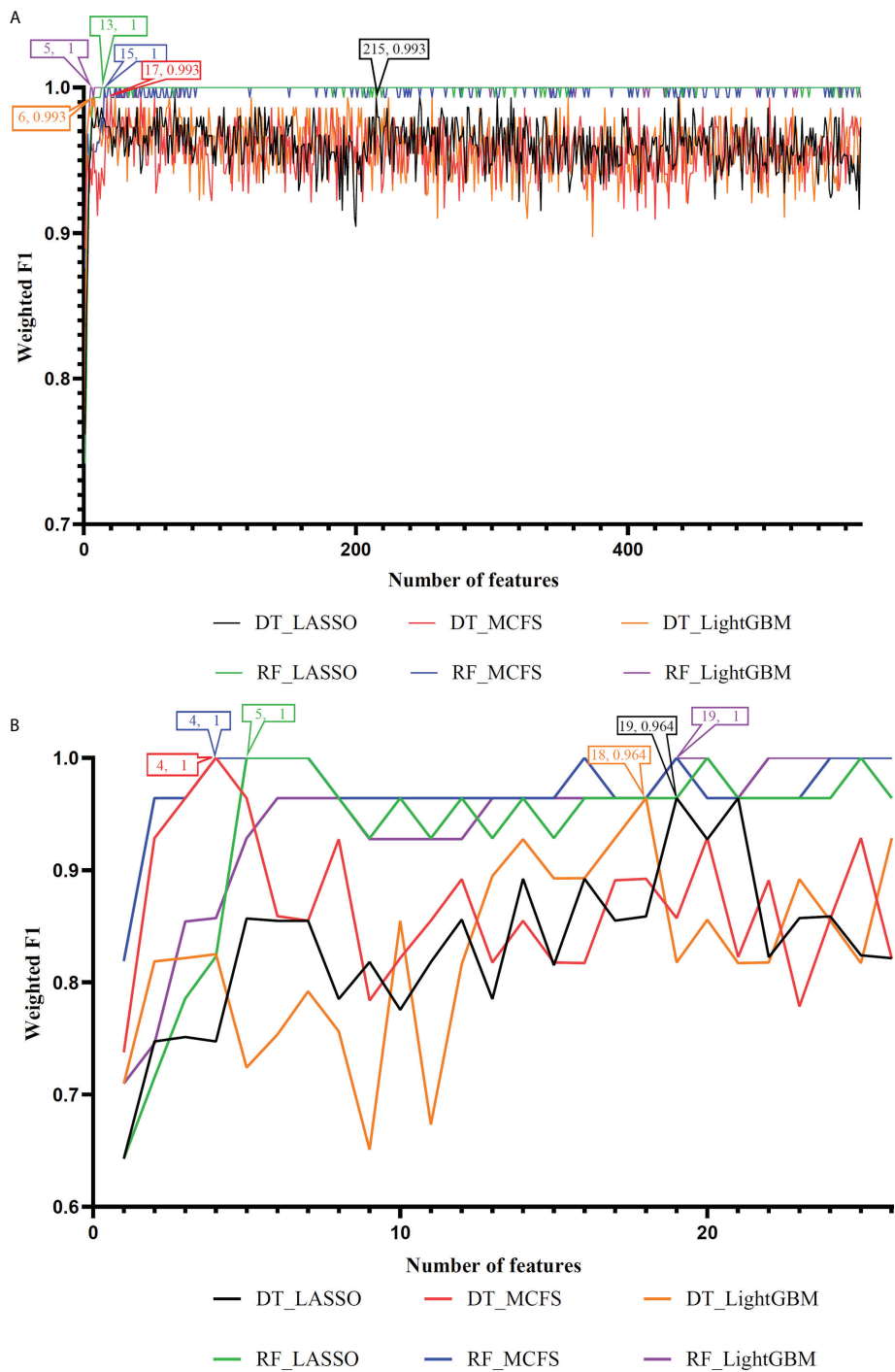


FIGURE 2 IFS curves to show the performance (weighted F1) of decision tree (DT) and random forest (RF) under different feature subsets in the anal and cervical carcinoma datasets. (A) IFS curves for the anal carcinoma dataset. (B) IFS curves for the cervical carcinoma dataset.

TABLE 1 Performance of the optimal classifiers on anal carcinoma dataset.

Feature ranking algorithm	Classification algorithm	Number of features	ACC	MCC	Macro F1	Weighted F1
MCFS	DT	17	0.993	0.975	0.981	0.993
	RF	15	1.000	1.000	1.000	1.000
LightGBM	DT	6	0.993	0.975	0.981	0.993
	RF	5	1.000	1.000	1.000	1.000
LASSO	DT	215	0.993	0.975	0.981	0.993
	RF	13	1.000	1.000	1.000	1.000

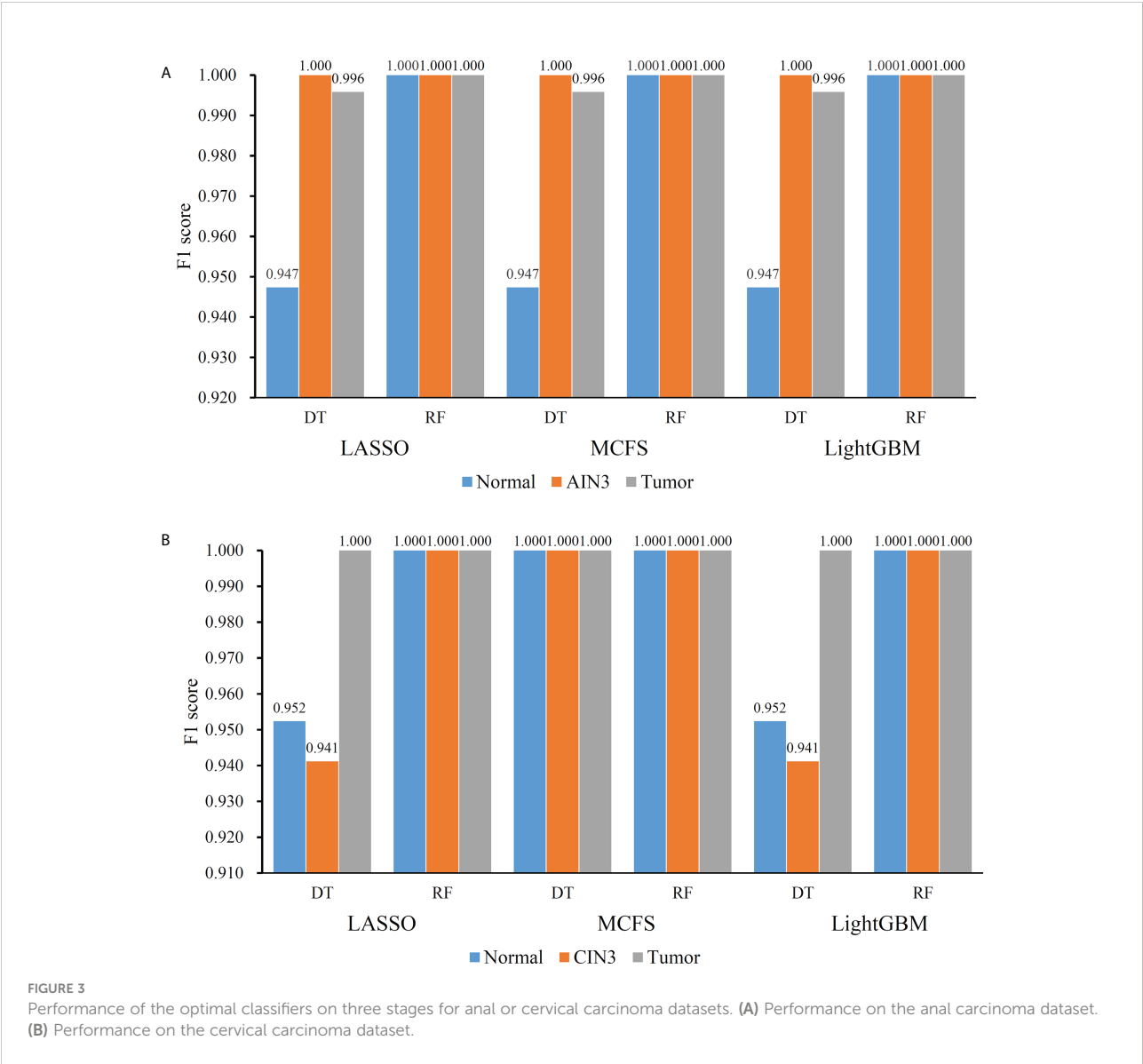


TABLE 2 Performance of the optimal classifiers on cervical carcinoma dataset.

Feature ranking algorithm	Classification algorithm	Number of features	ACC	MCC	Macro F1	Weighted F1
MCFS	DT	4	1.000	1.000	1.000	1.000
	RF	4	1.000	1.000	1.000	1.000
LightGBM	DT	18	0.964	0.948	0.965	0.964
	RF	19	1.000	1.000	1.000	1.000
LASSO	DT	19	0.964	0.948	0.965	0.964
	RF	5	1.000	1.000	1.000	1.000

4.1 Methylation biomarkers for anal and cervical carcinoma

The first methylation marker for anal carcinoma is cg23197559, near functional gene PTMA. According to recent publications, no direct reports confirm the association between PTMA and anal carcinoma. However, a recent study confirmed that in anal cancer and esophageal carcinoma, the methylation of a calcium-binding protein, S100A7, and PTMA has been shown to have specific methylation-mediated protein overexpression, validating the specific role of PTMA-related methylation alteration during anal carcinoma (42). The next probe cg07713411 is near gene MGA. MGA has been widely reported to be associated with tumor invasion (43) and progression (44). MGA has also been reported to contribute to MYC-mediated pathway in colorectal cancer cell lines. Considering the similarities between colorectal cancer and anal carcinoma, such gene regulated by our predicted methylation marker may also participate in the regulation of anal carcinoma, validating our prediction (44). The next probe cg25578064 regulates gene SFRS6, which is also a key driver gene for multiple gastrointestinal cancer subtypes (45), although it has no direct link with anal cancer. No functional genes were annotated around probe cg18954144 but the CpG site has been reported to be a typical signature for cancer overall survival (46), indicating that such probe may also be valuable for anal carcinoma monitoring. The final biomarker probe cg01550828 regulates a functional gene RNF168, encoding a ring finger protein. RNF168 has been selected as a candidate associated with HPV-related anal cancer (47), validating the efficacy and accuracy of our prediction.

As for biomarkers for cervical carcinoma, the first probe cg10417457 has been listed as a functional probe for cancer status monitoring according to a recent patent describing a systematic method to monitor cancer status established on 126 tumors (48). Therefore, such probe may also be functional to monitor cervical

carcinoma. No reports associated with cg02871554 have been found. As for the probe cg27012396 near a functional gene HDAC4, various publications have confirmed that HDAC4 regulates the glycolysis and survival of hypoxic tumor cells in cervical carcinoma (49–51). Therefore, it is reasonable for us to predict that such probe may be a biomarker for cervical carcinoma. The next biomarker is cg05713971 near an effective gene called HERPUD1. Antineoplastic activity has been shown to be associated with gene HERPUD1 and further related to human cervical carcinoma according to a recent *in vitro* experiment (52). Such gene has also been detected to be regulated by functional microRNA miR-375 and further contributes to HPV-positive cervical cancer, validating our prediction (53).

4.2 Quantitative rules for anal and cervical carcinoma

For monitoring the status of anal carcinoma, three rules for recognizing three different clusters separately include the functional probes cg01550828 and cg18954144, both of which are associated with anal tumorigenesis as we have discussed above. According to our rules, a higher methylation level of cg01550828 and a lower methylation level of cg18954144 indicate a pathogenic status of anal carcinoma, consistent with previous studies (46, 47). Interestingly, we also identified lower methylation of cg01550828, associated with gene RNF168 as a biomarker for pathogenesis of intermediate status (precancerous lesions/intraepithelial neoplasia), providing a novel approach for predicting the precancerous lesion stage. As we have discussed above, all the top rules are established based on our qualitative biomarkers, indicating the consistency between different machine learning models and validating the efficacy and accuracy of our prediction.

TABLE 3 Classification rules on anal carcinoma.

Index	Condition	Result
Rule 1	(cg01550828>0.0817) and (cg18954144>0.8291)	Tumor
Rule 2	cg01550828 ≤ 0.0817	AIN3
Rule 3	(cg01550828>0.0817) and (cg18954144 ≤ 0.8291) and (cg01550828>0.4363)	Normal
Rule 4	(cg01550828>0.0817) and (cg18954144 ≤ 0.8291) and (cg01550828 ≤ 0.4363)	Tumor

TABLE 4 Classification rules on cervical carcinoma.

Index	Condition	Result
Rule 1	$\text{cg10417457} \leq 0.4173$	Normal
Rule 2	$(\text{cg10417457} > 0.4173)$ and $(\text{cg02871554} > 0.6087)$	CIN3
Rule 3	$(\text{cg10417457} > 0.4173)$ and $(\text{cg02871554} \leq 0.6087)$	Tumor

For monitoring the status of cervical carcinoma, the top rules for normal control, precancerous lesion, and tumorigenesis prediction include the same group of features: cg10417457 and cg02871554. Although no direct association between cg02871554 and tumors has been recognized, cg10417457 has been validated to be an effective cancer-associated biomarker. Therefore, it is reasonable for our rules to summarize that a higher methylation of such probe may indicate a malignant change of cervical tissues. Further annotation on cg02871554 may be needed to explain its capacity for distinguishing precancerous lesions from malignant cancers.

5 Conclusion

In the present study, efficient feature selection algorithms, namely, Boruta, MCFS, LightGBM, and LASSO, were used to identify methylation signals associated with anal and cervical tumorigenesis. Subsequently, advanced machine learning algorithms were used to evaluate the performance of the filtered features for distinguishing different stages of anal or cervical carcinomas. Moreover, a DT was built to mine the classification rules for anal and cervical tumorigenesis. Taken together, this study provided a novel analysis to recognize key methylations for anal and cervical tumorigenesis qualitatively and quantitatively. The identified biomarkers and rules not only established an accurate and effective guideline for cancer differential diagnosis and progression stage monitoring, but also revealed potential mechanisms for the initiation and progression of anal and cervical tumorigenesis, indicating the specific roles of some methylations during the pathogenesis of these two diseases.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE186859>.

Author contributions

TH and Y-DC designed the study. FJ and FH performed the experiments. FJ, FH and Y-HZ analyzed the results. FJ and FH wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences [XDA26040304, XDB38050200], the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences [202002].

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/fonc.2022.998032/full#supplementary-material>

SUPPLEMENTARY TABLE 1

Feature lists yielded by MCFS, LightGBM, and LASSO on cervical and anal carcinoma datasets

SUPPLEMENTARY TABLE 2

Performance of DT and RF classifiers by applying IFS methods on feature lists yielded by MCFS, LightGBM, and LASSO on the cervical and anal carcinoma datasets.

References

- Engstrom PF, Arnoletti JP, Benson AB, Berlin JD, Berry JM, Chen Y-J, et al. Anal carcinoma. *J Natl Compr Cancer Network* (2010) 8:106–20. doi: 10.6004/jccn.2010.0007
- Lee GC, Kunitake H, Milch H, Savitt LR, Stafford C, Bordeianou LG, et al. What is the risk of anal carcinoma in patients with anal intraepithelial neoplasia III? *Dis Colon Rectum* (2018) 61:1350. doi: 10.1097/DCR.0000000000001219
- Donà MG, Giuliani M, Rollo F, Vescio MF, Benevolo M, Giglio A, et al. Incidence and clearance of anal high-risk human papillomavirus infection and their risk factors in men who have sex with men living with HIV. *Sci Rep* (2022) 12:184. doi: 10.1038/s41598-021-03913-5
- Somia IKA, Teeratakulpisarn N, Joo WS, Yee IA, Pankam T, Nonenoy S, et al. Prevalence of and risk factors for anal high-risk HPV among HIV-negative and HIV-positive MSM and transgender women in three countries at south-East Asia. *Medicine* (2018) 97:e9898. doi: 10.1097/MD.00000000000009898
- Lerman J, Hennequin C, Etienney I, Abramowitz L, Goujon G, Gornet J-M, et al. Impact of tobacco smoking on the patient's outcome after (chemo) radiotherapy for anal cancer. *Eur J Cancer* (2020) 141:143–51. doi: 10.1016/j.ejca.2020.09.039
- Devine C, Viswanathan C, Faria S, Marcal L, Sagebiel TL. Imaging and staging of cervical cancer. *Seminars in ultrasound, CT and MRI*. (2019), 40: 280–6. doi: 10.1053/j.sult.2019.03.001
- Barukčić I. Human papillomavirus—the cause of human cervical cancer. *J Biosci Medicines* (2018) 6:106. doi: 10.4236/jbm.2018.64009
- Koch A, Joosten SC, Feng Z, De Ruijter TC, Draht MX, Melotte V, et al. Analysis of DNA methylation in cancer: location revisited. *Nat Rev Clin Oncol* (2018) 15:459–66. doi: 10.1038/s41571-018-0004-4
- Pfeifer GP. Defining driver DNA methylation changes in human cancer. *Int J Mol Sci* (2018) 19:1166. doi: 10.3390/ijms19041166
- Zhu H, Zhu H, Tian M, Wang D, He J, Xu T. DNA Methylation and hydroxymethylation in cervical cancer: diagnosis, prognosis and treatment. *Front Genet* (2020) 11:347. doi: 10.3389/fgene.2020.00347
- Van Der Zee RP, Van Noesel CJ, Martin I, Ter Braak TJ, Heideman DA, De Vries HJ, et al. DNA Methylation markers have universal prognostic value for anal cancer risk in HIV-negative and HIV-positive individuals. *Mol Oncol* (2021) 15:3024–36. doi: 10.1002/1878-0261.12926
- Siegel EM, Ajidahun A, Berglund A, Guerrero W, Eschrich S, Putney RM, et al. Genome-wide host methylation profiling of anal and cervical carcinoma. *PLoS One* (2021) 16:e0260857. doi: 10.1371/journal.pone.0260857
- Kursa MB, Rudnicki WR. Feature selection with the boruta package. *J Stat Softw* (2010) 36:1–13. doi: 10.18637/jss.v036.i11
- Huang F, Chen L, Guo W, Zhou X, Feng K, Huang T, et al. Identifying COVID-19 severity-related SARS-CoV-2 mutation using a machine learning method. *Life* (2022) 12:806. doi: 10.3390/life12060806
- Li Z, Huang F, Chen L, Huang T, Cai Y-D. Identifying *In vitro* cultured human hepatocytes markers with machine learning methods based on single-cell RNA-seq data. *Front Bioeng Biotechnol* (2022) 10:916309. doi: 10.3389/fbioe.2022.916309
- Zhou X, Ding S, Wang D, Chen L, Feng K, Huang T, et al. Identification of cell markers and their expression patterns in skin based on single-cell RNA-sequencing profiles. *Life* (2022) 12:550. doi: 10.3390/life12040550
- Micha D, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski, et al. Monte Carlo Feature selection for supervised classification. *Bioinformatics* (2008) 24:110–7. doi: 10.1093/bioinformatics/btm486
- Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* (2017) 30:3146–54.
- Tibshirani RJ. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B: Methodological* (1996) 73:273–82. doi: 10.1111/j.1467-9868.2011.00771.x
- Chen L, Li J, Zhang YH, Feng K, Wang S, Zhang Y, et al. Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J Cell Biochem* (2018) 119:3394–403. doi: 10.1002/jcb.26507
- Chen X, Jin Y, Feng Y. Evaluation of plasma extracellular vesicle MicroRNA signatures for lung adenocarcinoma and granuloma with Monte-Carlo feature selection method. *Front Genet* (2019) 10:367. doi: 10.3389/fgene.2019.00367
- Li J, Lu L, Zhang YH, Xu Y, Liu M, Feng K, et al. Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine. *Cancer Gene Ther* (2020) 27:56–69. doi: 10.1038/s41417-019-0105-y
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* (2011) 12:2825–30.
- Liu H, Setiono R. Incremental feature selection. *Appl Intell* (1998) 9:217–30. doi: 10.1023/A:1008363719778
- Zhao X, Chen L, Lu J. A similarity-based method for prediction of drug side effects with heterogeneous information. *Math Biosci* (2018) 306:136–44. doi: 10.1016/j.mbs.2018.09.010
- Zhang YH, Li Z, Zeng T, Pan X, Chen L, Liu D, et al. Distinguishing glioblastoma subtypes by methylation signatures. *Front Genet* (2020) 11:604336. doi: 10.3389/fgene.2020.604336
- Chen L, Li Z, Zhang S, Zhang Y-H, Huang T, Cai Y-D. Predicting RNA 5-methylcytosine sites by using essential sequence features and distributions. *BioMed Res Int* (2022) 2022:4035462. doi: 10.1155/2022/4035462
- Ding S, Wang D, Zhou X, Chen L, Feng K, Xu X, et al. Predicting heart cell types by using transcriptome profiles and a machine learning method. *Life* (2022) 12:228. doi: 10.3390/life12020228
- Kohavi R. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *International joint conference on artificial intelligence*. (San Francisco, CA, United States: Morgan Kaufmann Publishers Inc.) 1995 p. 1137–45.
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* (2002) 16:321–57. doi: 10.1613/jair.953
- Zhang X, Chen L, Guo Z-H, Liang H. Identification of human membrane protein types by incorporating network embedding methods. *IEEE Access* (2019) 7:140794–805. doi: 10.1109/ACCESS.2019.2944177
- Pan X, Chen L, Liu I, Niu Z, Huang T, Cai YD. Identifying protein subcellular locations with embeddings-based node2loc. *IEEE/ACM Trans Comput Biol Bioinform* (2021) 19:666–75. doi: 10.1109/TCBB.2021.3080386
- Breiman L. Random forests. *Mach Learn* (2001) 45:5–32. doi: 10.1023/A:1010933404324
- Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Trans syst man cybern* (1991) 21:660–74. doi: 10.1109/21.97458
- Chen L, Li Z, Zeng T, Zhang YH, Feng K, Huang T, et al. Identifying COVID-19-specific transcriptomic biomarkers with machine learning methods. *BioMed Res Int* (2021) 2021:9939134. doi: 10.1155/2021/9939134
- Onesime M, Yang Z, Dai Q. Genomic island prediction via chi-square test and random forest algorithm. *Comput Math Methods Med* (2021) 2021:9969751. doi: 10.1155/2021/9969751
- Zhang Y-H, Zeng T, Chen L, Huang T, Cai Y-D. Determining protein-protein functional associations by functional rules based on gene ontology and KEGG pathway. *Biochim Biophys Acta (BBA) - Proteins Proteomics* (2021) 1869:140621. doi: 10.1016/j.bbapap.2021.140621
- Ran B, Chen L, Li M, Han Y, Dai Q. Drug-drug interactions prediction using fingerprint only. *Comput Math Methods Med* (2022) 2022:7818480. doi: 10.1155/2022/7818480
- Tang S, Chen L. iATC-NFMLP: Identifying classes of anatomical therapeutic chemicals based on drug networks, fingerprints and multilayer perceptron. *Curr Bioinf* (2022). doi: 10.2174/1574893617666220318093000
- Yang Y, Chen L. Identification of drug-disease associations by using multiple drug and disease networks. *Curr Bioinf* (2022) 17:48–59. doi: 10.2174/1574893616666210825115406
- Gorodkin J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput Biol Chem* (2004) 28:367–74. doi: 10.1016/j.compbiolchem.2004.09.006
- Su Y-F, Chen Y-J, Tsai F-T, Li W-C, Hsu M-L, Wang D-H, et al. Current insights into oral cancer diagnostics. *Diagnostics* (2021) 11:1287. doi: 10.3390/diagnostics11071287
- Mathsyaraja H, Catchpole J, Eastwood E, Babaeva E, Geuenich M, Cheng PF, et al. Loss of MGA mediated polycomb repression promotes tumor progression and invasiveness. *bioRxiv* (2020). doi: 10.1101/2020.10.16.334714
- Demma MJ. *Targeting the oncogenic MYC pathway by miniproteins: Understanding mechanism of action*. University of Miami (2019).
- Brim H, Abu-Asab MS, Nouria M, Salazar J, Deleo J, Razjouyan H, et al. An integrative CGH, MSI and candidate genes methylation analysis of colorectal tumors. *PLoS One* (2014) 9:e82185. doi: 10.1371/journal.pone.0082185
- Wang Y, Chen L, Ju L, Qian K, Wang X, Xiao Y, et al. Epigenetic signature predicts overall survival clear cell renal cell carcinoma. *Cancer Cell Int* (2020) 20:564–4. doi: 10.1186/s12935-020-01640-x

47. Szymonowicz KA, Chen J. Biological and clinical aspects of HPV-related cancers. *Cancer Biol Med* (2020) 17:864. doi: 10.20892/j.issn.2095-3941.2020.0370
48. Zhang K, Hou R, Zheng L. Method and system for determining cancer status. In: Google Patents, Patent No 9,984,201 (2018). Zhang K, Hou R, Zheng L. *Method and system for determining cancer status*. In: Google Patents, Patent No 9,984,201 (2018).
49. Yeasmin S, Nakayama K, Rahman MT, Rahman M, Ishikawa M, Katagiri A, et al. Biological and clinical significance of NAC1 expression in cervical carcinomas: a comparative study between squamous cell carcinomas and adenocarcinomas/adenosquamous carcinomas. *Hum Pathol* (2012) 43:506–19. doi: 10.1016/j.humpath.2011.05.021
50. Liu Y, Lu Z, Xu R, Ke Y. Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology. *Oncotarget* (2016) 7:5852. doi: 10.18632/oncotarget.6809
51. Zhang Y, Ren Y, Guo L, Ji C, Hu J, Zhang H, et al. Nucleus accumbens-associated protein-1 promotes glycolysis and survival of hypoxic tumor cells via the HDAC4-HIF-1 α axis. *Oncogene* (2017) 36:4171–81. doi: 10.1038/onc.2017.51
52. De Souza CEA, Pires ADRA, Cardoso CR, Carlos RM, Cadena SMSC, Acco A. Antineoplastic activity of a novel ruthenium complex against human hepatocellular carcinoma (HepG2) and human cervical adenocarcinoma (HeLa) cells. *Heliyon* (2020) 6:e03862. doi: 10.1016/j.heliyon.2020.e03862
53. Zeng J-H, Liang X-Z, Lan H-H, Zhu X, Liang X-Y. The biological functions of target genes in pan-cancers and cell lines were predicted by miR-375 microarray data from GEO database and bioinformatics. *PloS One* (2018) 13:e0206689. doi: 10.1371/journal.pone.0206689



OPEN ACCESS

EDITED BY

Liang Cheng,
Harbin Medical University, China

REVIEWED BY

Jing Yang,
ShanghaiTech University, China
Wei Kong,
Shanghai Maritime University, China

*CORRESPONDENCE

Zhenbing Zeng
zbzeng@shu.edu.cn
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

[†]These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 27 June 2022

ACCEPTED 14 September 2022

PUBLISHED 29 September 2022

CITATION

Lu J, Li J, Ren J, Ding S, Zeng Z,
Huang T and Cai Y-D (2022)
Functional and embedding feature
analysis for pan-cancer classification.
Front. Oncol. 12:979336.
doi: 10.3389/fonc.2022.979336

COPYRIGHT

© 2022 Lu, Li, Ren, Ding, Zeng, Huang
and Cai. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Functional and embedding feature analysis for pan-cancer classification

Jian Lu^{1,2†}, JiaRui Li^{3†}, Jingxin Ren⁴, Shijian Ding⁴,
Zhenbing Zeng^{1*}, Tao Huang^{2,5*} and Yu-Dong Cai^{4*}

¹Department of Mathematics, School of Sciences, Shanghai University, Shanghai, China, ²CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Science, Shanghai, China, ³Advanced Research Computing, University of British Columbia, Vancouver, BC, Canada, ⁴School of Life Sciences, Shanghai University, Shanghai, China, ⁵CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

With the increasing number of people suffering from cancer, this illness has become a major health problem worldwide. Exploring the biological functions and signaling pathways of carcinogenesis is essential for cancer detection and research. In this study, a mutation dataset for eleven cancer types was first obtained from a web-based resource called cBioPortal for Cancer Genomics, followed by extracting 21,049 features from three aspects: relationship to GO and KEGG (enrichment features), mutated genes learned by word2vec (text features), and protein-protein interaction network analyzed by node2vec (network features). Irrelevant features were then excluded using the Boruta feature filtering method, and the retained relevant features were ranked by four feature selection methods (least absolute shrinkage and selection operator, minimum redundancy maximum relevance, Monte Carlo feature selection and light gradient boosting machine) to generate four feature-ranked lists. Incremental feature selection was used to determine the optimal number of features based on these feature lists to build the optimal classifiers and derive interpretable classification rules. The results of four feature-ranking methods were integrated to identify key functional pathways, such as olfactory transduction (hsa04740) and colorectal cancer (hsa05210), and the roles of these functional pathways in cancers were discussed in reference to literature. Overall, this machine learning-based study revealed the altered biological functions of cancers and provided a reference for the mechanisms of different cancers.

KEYWORDS

pan-cancer, cancer mutation, enrichment, embedding, feature selection, rule learning

1 Introduction

Cancer is one of the most common causes of death in human beings. According to World Health Organization (WHO), about 10 million patients died because of cancer in 2020. Early cancer diagnosis significantly improves the survival, but more than half of patients with cancer have been diagnosed in advanced stages (1). The average 5-year survival rate after surgery in the early stage is 91%, which is higher than the 26% survival rate in the late stage (2).

The identification of tumor type and tissue origin is of paramount importance for cancer treatment. Most cancer types are diagnosed *via* invasive biopsy; however, non-invasive early detection is lacking (3). Circulating tumor DNA (ctDNA) could be a potential biomarker for early cancer diagnosis (4). Despite the multiple challenges in developing non-invasive liquid biopsy based on ctDNA in blood plasma, such as the limited materials of cancer DNA in blood plasma to achieve a high sensitivity (5), enormous efforts and progresses have been made in the past decades. Studies on identification methods for tumor tissue of origin mainly focused on characterizing and utilizing tumor-specific DNA methylation, gene expression profiling, and genomic alteration (6–8). Machine learning methods, especially deep learning models, have been developed and widely used to identify tumor tissue of origin (9). In our previous study, we developed a bioinformatics pipeline based on machine learning algorithms to identify the tissue of origin in five tumors according to the enrichment of gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) using the mutated genes (10); the approach was proven to be of high efficacy and robustness. However, the limitation of previous methods in analyzing small datasets restricted our previous analysis with only five cancer types.

In this study, we applied machine learning algorithms to investigate a large mutation data, which involved eleven cancer types. Each sample was represented by three feature types: (1) relationship to GO terms and KEGG pathways; (2) word embeddings of mutated genes; (3) network embeddings of mutated genes. Several machine learning algorithms were applied to such dataset. First, the irrelevant features were excluded by Boruta feature selection. Then, remaining features were deeply analyzed by four different feature selection methods, resulting in four feature-ranked lists. In the next step, each feature list is subjected to incremental feature selection (IFS) (11) combined with the different classification algorithms to determine the optimal number of features and build the optimal classifiers. Some essential features were identified by each feature selection method and those identified by multiple methods were deemed to be more important. Features related to GO terms and KEGG pathways were analyzed. Furthermore, this study also reported several classification rules, indicating different patterns on various cancer types. From the results

yielded by four feature selection methods, they were quite different, suggesting that the four methods are complement with each other. Incorporating multiple methods in the pipeline can help us achieve a more comprehensive result.

2 Materials and methods

2.1 Data sources

Mutation data with eleven cancer types were acquired from the cBioPortal for Cancer Genomics (http://cbio.mskcc.org/cancergenomics/pancan_tcga/) (12, 13). This dataset mainly includes bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma/rectum adenocarcinoma esophageal carcinoma (COADREAD), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), acute myeloid leukemia (LAML), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), and uterine corpus Endometrial Carcinoma (UCEC). A total of 3478 samples were obtained, and the sample size for each cancer type is listed in Table 1. This cancer mutation dataset was then used in the next step of the analysis.

2.2 Feature representation

In this work, three approaches were utilized to encode the feature vectors to extract relevant information from each cancer sample in the mutant dataset: GO and KEGG enrichment theory, word2vec, and node2vec. Accordingly, three feature types were generated from each sample, namely, enrichment, text and network features, respectively. A total of 21,049 features were created, with 20,293 enrichment features derived from GO

TABLE 1 Number of samples under different cancer types.

Cancer type	Sample size
BLCA	100
BRCA	513
COADREAD	499
GBM	276
HNSC	306
KIRC	473
LAML	201
LUAD	230
LUSC	177
OV	456
UCEC	247

and KEGG, 256 text features yielded by word2vec, and 500 network features generated by node2vec. A detailed description of these features is presented below.

2.2.1 Enrichment features derived from GO and KEGG

GO terms and KEGG pathways give crucial functional information for gene characterization in biology study and the discovery of underlying biological mechanisms. The data obtained could be helpful for further research when GO terms and KEGG pathways are used for feature encoding. As a commonly used approach in quantifying the overlap between the gene set and GO terms or KEGG pathways, the GO and KEGG enrichment theory (14) were used to measure the impact of alterations in biological functions among patients with cancer.

For a specific cancer individual p and a GO term GO_j , G_{GO} represents the gene set that is annotated by GO_j , and G_p represents the variant gene set for individual p . The relationship between p and GO_j is defined as the hypergeometric test p-values of G_p and G_{GO} , called GO enrichment score, which can be computed by

$$Score_{GO}(p, GO_j) = -\log_{10} \left(\sum_{k=m}^n \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} \right), \quad (1)$$

where N and M indicate the total number of human genes and the number of genes in G_{GO} , respectively; n denotes the number of mutant genes in G_p , and m represents the number of genes both in G_p and G_{GO} . According to the high enrichment score, the mutation in patient p has a deep functional impact on the GO term GO_j .

Similarly, for the KEGG pathway, the enrichment score for a cancer individual p and a KEGG pathway K_j can be calculated as follows, called KEGG enrichment score,

$$Score_{KEGG}(p, K_j) = -\log_{10} \left(\sum_{k=m}^n \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} \right), \quad (2)$$

where N and n are defined as shown in Eq. 1, and M and m indicate the number of genes in pathway K_j and the number of genes both in G_p and K_j . A total of 20,293 GO terms and KEGG pathways were adopted in this study, with the enrichment scores between patients with cancer and these functional terms serving as the feature values. Each patient with cancer is represented by 20,293 enrichment scores, which can be used for subsequent feature analysis. For convenience, such features were called enrichment features. These features were calculated by our in-house program.

2.2.2 Text features generated by Word2vec

Word2vec is a natural language processing model that uses unsupervised learning to learn word associations from a text corpus (15). It obtains the word embedding vectors by training two-layer neural networks to reconstruct linguistic contexts of words, making the semantic and syntactic similar words close in distance in a specific space. Words are embedded in a continuous vector space, with close vectors for similar words. The training algorithms of word2vec are mainly CBOW or Skip-gram. Here, the word2vec algorithm in Gensim (<https://github.com/RaRe-Technologies/gensim>) was adopted. It took the name of each gene as a word and the genes presenting in each sample as sentences. The second class of features was the average of the vectors corresponding to the genes under each sample. In summary, word2vec program with default parameters was used to produce a 256-dimensional feature vector for each sample based on gene names. For convenience, these features were called text features.

2.2.3 Network features generated by Node2vec

The gene interaction network provides information on the features of gene interactions. In this study, gene names were inputted into a gene network based on the PPI network in STRING (16). Each node in this network represents a gene, each edge denotes the interaction between two genes. Evidently, each edge indicates a PPI. To reflect different strengths of PPIs, edges are assigned the confidence scores of their corresponding PPIs. Thus, this gene interaction network is a weighted version. The feature vector of each gene is obtained using node2vec (17).

The node2vec algorithm can be regarded as a generalized version of Skip-gram, which can process network data. It first generates several paths starting from each node in the network. Each path is extended from the current endpoint to one of its neighbors in a well-defined way. After a predefined number of paths have been generated, they are fed into the word2vec with Skip-gram, where nodes in paths are termed as words and paths are considered as sentences, to yield the feature vector of each node. The node2vec program was retrieved from <https://snap.stanford.edu/node2vec/>. Default parameters were used.

The gene interaction network mentioned above was fed into the node2vec program, assigning a vector feature for each node (gene). The feature vector of each sample was further constructed from the feature vectors of genes and was defined as the mean vector of the feature vectors of genes related to the sample. In this study, a 500-dimensional feature vector was generated for each of the samples from the gene interaction network. For convenience, these features were called network features.

2.3 Feature selection methods

2.3.1 Boruta feature filtering

A total of 21,049-dimensional feature vectors were obtained after feature encoding. Directly employing these features for analysis would require massive computation. Therefore, non-essential features were eliminated from the dataset using Boruta feature selection (18). In each iteration round, Boruta compares the importance of the original feature to that of the shadow feature with random forest (RF) classifier. If the original feature is statistically more important than the shadow features, then the original feature is deemed important. If the original feature is statistically less essential than the shadow feature, then the original feature is considered unimportant. After Boruta analysis, the important features were retained for the next step of feature ranking, and the computational efficiency was improved.

In this work, the Boruta program from https://github.com/scikit-learn-contrib/boruta_py was used and executed with default parameters.

2.3.2 Least absolute shrinkage and selection operator

Lasso (19) is a regression model that uses L1 regularization technology. The overfitting problem is reduced by adding a high penalty to parameters with high coefficients and great prediction errors, thus reducing the number of parameters and lowering the feature dimension because some feature coefficients are reduced to 0 and eliminated from the model. As a result, Lasso is frequently used for the selection of features that are prioritized by importance according to their coefficients. The Lasso package, obtained from Scikit-learn (20), was applied on the features selected by Boruta. Its default parameters were adopted. The obtained feature list was called Lasso feature list.

2.3.3 Minimum redundancy maximum relevance

mRMR (21) is a feature selection method that has been widely applied in biology. Its main goal is to maximize the correlation between features and categorical variables while minimizing feature-to-feature redundancy. Mutual information between individual features and category variables is used to determine the correlation between features and categories, and mutual information between features and features is used to calculate the redundancy. A ranked feature list can be obtained after feature selection using mRMR. The mRMR program was derived from <http://home.penglab.com/proj/mRMR/>, and it was executed with default parameters. The list yielded by mRMR was called mRMR feature list.

2.3.4 Monte Carlo feature selection

MCFS (22) is used to identify the essential features in the dataset for a particular classification problem. The method

resamples the original dataset c times, separates it into c pairs of training and test sets, randomly selects m features from all features for s times to build a decision tree (DT), and generates s DTs each time. Finally, the entire procedure yields $c \times s$ DTs.

The relative importance (RI) is computed for each feature based on these DTs. Features are sorted in descending order of RI values to produce a ranked feature list. Here, MCFS was implemented using the dmLab software provided by Draminski (22) with parameters u and v set to 1, which can be obtained at <http://www.ipi-pan.eu/staff/m.draminski/mcfs.html>. Features with RI scores equal to 0 from the calculation results were deleted in the next analysis. The list yielded by MCFS was termed as MCFS feature list.

2.3.5 Light gradient boosting machine

LightGBM (23) is a fast gradient boosting DT implementation that recurrently fits a new DT by using the negative gradient of the loss function of the current DT as the approximate value of the residual. This approach saves computer resources by employing two strategies called gradient-based one-side sampling and exclusive feature bundling. Given that LightGBM is based on a tree model, the importance of a feature can be quantified by the number of times the feature is involved in building the DTs. In this study, a python version of the LightGBM program with default parameters, which can be downloaded from <https://lightgbm.readthedocs.io/en/latest/>, was used to rank features selected by Boruta. This list was called LightGBM feature list.

2.4 Incremental feature selection

Features selected by Boruta were sorted in descending order of importance using the Lasso, mRMR, MCFS, and lightGBM algorithms. However, the features in each feature list that were critical to the classification of cancer types were not determined. Therefore, IFS (11) was used to detect the optimal number of features in each ranked list and build the optimal classifiers.

Given a feature list, IFS produces a series of feature subsets depending on a specified interval step initially. For example, when the interval is 5, the first feature subset includes the top 5 features in the list, and the second feature subset includes the top 10 features. All possible feature subsets can be generated when the interval was set as 1. The sample data including each of these feature subsets are then applied to train a classifier with a given classification algorithm (e.g., DTs (24), random forest (RF) (25), and support vector machine (SVM) (26)). Such classifier is tested by 10-fold cross-validation (Kohavi, 1995). When training the classifier, we adopted Synthetic Minority Oversampling Technique (SMOTE) (27) to balance the sample sizes of different cancer types in this study. Ultimately, all classifiers

built by the succession of feature subsets were compared using a performance metric to determine the optimal number of features and the consequent optimal classifier.

2.5 Synthetic minority oversampling technique

In this study, the sample sizes for the eleven cancer types were markedly unequal as indicated in Table 1. The obtained results are frequently unsatisfactory when the classifier is built by directly utilizing an unbalanced sample dataset. Hence, overcoming the categorization difficulty provided by uneven data has become a machine learning challenge. SMOTE is a synthetic sampling strategy in which new samples for a minority class are generated using any randomly selected sample and its nearest neighbors (27). In this study, SMOTE was utilized in the imblearn module (with default parameters) to synthesize new samples for minority cancer types and generate an equal number of cancer samples in each type in the training set.

2.6 Classification algorithms

In IFS method, one classification algorithm was necessary. To fully test each feature subset, three classic classification algorithms: DT (24), RF (25), and SVM (26), were employed in this study. These classification algorithms have been applied to tackle various medical or biological problems (28–36).

2.6.1 Support vector machine

SVM is one of the most classic classification algorithms. The main idea is to determine a hyperplane by learning the distribution of samples in different classes. Generally, such hyperplane in the original feature space is difficult to obtain. SVM adopts the kernel trick to translate samples to a high-dimensional feature space. In this case, such hyperplane is easy to discover. The class of a test sample is determined according to the side of the hyperplane it belongs to.

2.6.2 Random forest

RF is also a classic classification algorithm, which is quite different from SVM. In fact, it is an ensemble algorithm consisting of several DTs. Each DT is built on a new dataset, in which samples were randomly selected, with replacement, from the original dataset. And such new dataset has same number of samples in the original dataset. Furthermore, each DT is constructed based on randomly selected features. The predicted results of RF are determined by the majority voting on the results yielded by all DTs.

2.6.3 Decision tree

Above two classification algorithms are generally deemed to be powerful. However, their decision principles are quite

complicated, which is impossible for us to understand. This is a great block for us to learn new knowledge from a large dataset. For this study, we cannot extract mutation patterns on different cancer types only based on RF and SVM. In view of this, DT was also used in this study, which is deemed to be a type of white-box algorithm. It employs a tree structure and contains leaf nodes and branch nodes. The branch nodes are in charge of classifying samples, whereas the leaf nodes are responsible for determining classes. Besides the tree representation, a DT can also be represented by a set of IF-THEN rules. Each rule is obtained by a path from root node to one leaf node. These rules make the classification procedures completely open, providing opportunities for us to understand different patterns on various cancer types.

In this work, the corresponding packages that implement above SVM, RF and DT, in scikit-learn (20) were employed. Each package was performed with default parameters.

2.7 Performance measurement

In the IFS, the classifiers were trained using training samples consisting of the feature subsets. The performance of the classifiers was then evaluated using 10-fold cross-validation (37). The commonly used main model metrics for each class are accuracy (recall), precision and F_1 score (38–41). Here, F_1 score was used as the main metric to measure the performance of the classifier on one class, which can be calculated as follows:

$$F_1 \text{ score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

As above F_1 score only measures the performance of the classifier on one class. F_1 scores on all classes can be integrated to give an overall evaluation on the classifier. There are two ways to integrate these scores. The first way is to calculate the mean of all F_1 scores. Such obtained measurement is called macro F_1 . The second way further considers the class sizes, the weighted mean of all F_1 scores is computed, which is termed as weighted F_1 . As the sizes of different cancers are quite different, weighted F_1 was more proper than macro F_1 to fully evaluate the overall performance of classifiers. Thus, it was selected as the key measurement in this study.

Besides, the overall accuracy (ACC) and Matthew correlation coefficients (MCC) (42, 43) were also employed. ACC is a generally measurement, which indicates the proportion of correctly predicted samples. MCC is much more complex. However, it is deemed as a balanced measurement even if the sizes of classes are quite different. To compute MCC, two matrices X and Y should be constructed in advance, where X stores the true class of each sample and Y includes the predicted class of each sample. Then, the MCC can be computed by

$$MCC = \frac{\text{cov}(X, Y)}{\sqrt{\text{cov}(X, X)\text{cov}(Y, Y)}}, \quad (4)$$

where $\text{cov}(Y, X)$ stands for the covariance of two matrices.

3 Results

In this study, we first downloaded a mutation dataset containing 3478 cancer samples from the cBioPortal for Cancer Genomics database, which included eleven cancer types. Three feature types (enrichment, text and network features) were generated to represent each cancer sample. The Boruta feature filtering method was used to remove irrelevant features and selected features were further analyzed by Lasso, mRMR, MCFS, and LightGBM methods, respectively, to produce four feature-ranked lists. Each feature list was subjected to IFS combined with classification algorithms and model evaluation measurements to determine the optimal number of features, build the optimal classifiers, and extract the important classification rules. The entire analysis pipeline is shown in Figure 1. This section details the obtained results.

3.1 Results of feature selection methods

First, a large dataset containing 3,478 samples and 21,049 features was generated. To filter key informative features from these features, the Boruta feature filtering method was applied to such dataset. 18,835 features were excluded and 2,214 important features were retained, which are provided in Supplementary

Table S1. Among the selected 2,214 features, enrichment features were most, followed by network and text features. The numbers of selected features on three types are shown in Figure 2. Enrichment features were important to classify samples into different cancer types. However, considering the fact that the original enrichment features were much more than other two feature types, such result was reasonable. Furthermore, the selected enrichment features only occupied 8.33% of all enrichment features, and such proportions for text and network features were 60.16% and 74.00%, respectively. It was indicated that text and network features also provided key contributions on the classification of cancer samples.

In the next step, a refined dataset with 3,478 samples and 2,214 selected features was produced. Four feature selection methods (Lasso, mRMR, MCFS, and LightGBM) were executed on such dataset to analyze the importance of the 2,214 features. Four feature lists: Lasso, mRMR, MCFS and LightGBM feature lists, were obtained. These lists are also provided in Supplementary Table S1.

3.2 Results of IFS method on different feature lists

We obtained four feature-ranked lists but were still unable to determine the features in each list that could effectively distinguish cancer types. Therefore, we employed the IFS combined with classification algorithms to determine the optimal results. For each list, IFS first generated a series of feature subsets with interval 5, on which the DT, RF, and SVM

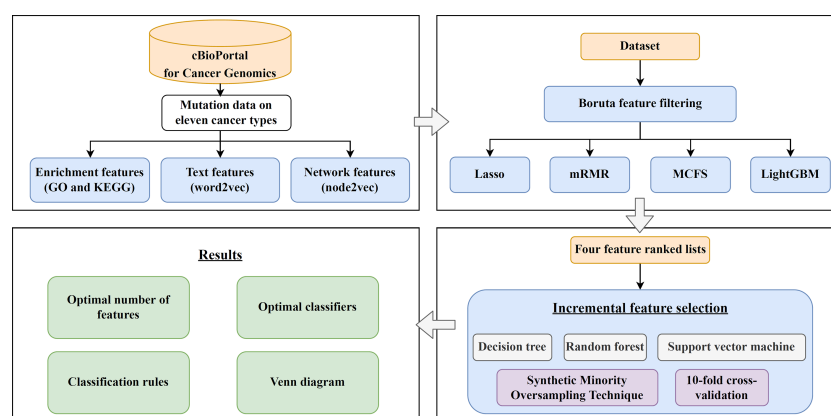


FIGURE 1

Computational framework of this study. First, the cancer samples obtained from the cBioPortal for Cancer Genomics database are represented by three feature types, derived by GO and KEGG enrichment, word2vec, and node2vec. Then, the Boruta feature filtering is adopted to exclude irrelevant features and retained features are ranked by Lasso, mRMR, MCFS, and LightGBM methods in four feature-ranked lists. These feature lists are subjected to incremental feature selection combined with classification algorithms to determine the optimal number of features, build optimal classifiers, and extract important classification rules. Furthermore, the Venn diagram analysis is conducted on the key features identified by different feature selection methods.

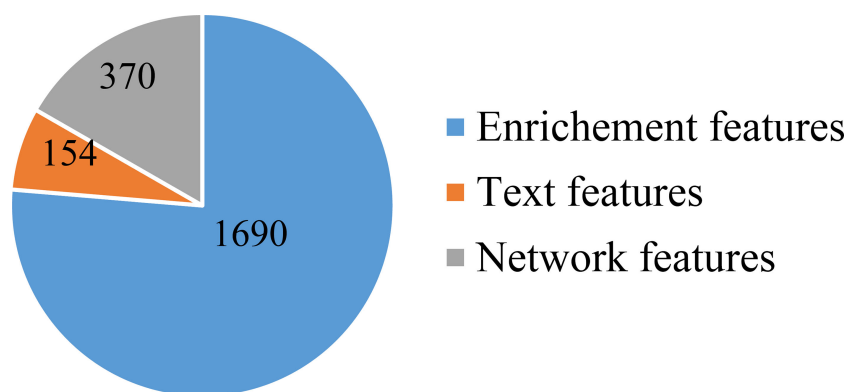


FIGURE 2

Pie chart to show the distribution of features selected by Boruta on three feature types. Enrichment features are most, followed by network and text features.

classifiers were constructed. We then used 10-fold cross-validation for evaluation with weighted F1 as the key performance metric. The results of the IFS on different feature selection methods are provided in [Supplementary Table S2](#).

For the IFS results on the Lasso feature list, an IFS curve was plotted for each classification algorithm, as shown in [Figure 3](#), where weighted F1 was set as Y-axis and number of features was set as X-axis. It can be observed that the highest weighted F1 values for DT, RF and SVM were 0.4215, 0.6134 and 0.6772, respectively. These values were obtained by using top 1770, 2055 and 1905, respectively, features in the list, which constituted the optimal feature subsets for three classification algorithms, respectively. Furthermore, the optimal DT, RF and SVM classifiers were built using the corresponding optimal feature

subsets. The values of ACC, MCC and Macro F1 yielded by these optimal classifiers are listed in [Table 2](#). Evidently, the optimal SVM classifier provided the highest performance. The performance of the optimal classifiers on eleven cancer types are illustrated in [Figure 4A](#), from which we can see that the optimal SVM classifier provided the best performance on all cancer types. This further confirmed the superiority of the optimal SVM classifier.

With regard to the IFS results on the mRMR feature list, three IFS curves were also plotted, as illustrated in [Figure 5](#). DT, RF and SVM provided the highest weighted F1 of 0.4347, 0.6170 and 0.6200, respectively. Top 490, 1505 and 1810, respectively, features in the mRMR feature list were used to generate such performance. On these features, the optimal DT, RF and SVM classifiers were built. Their additional performance

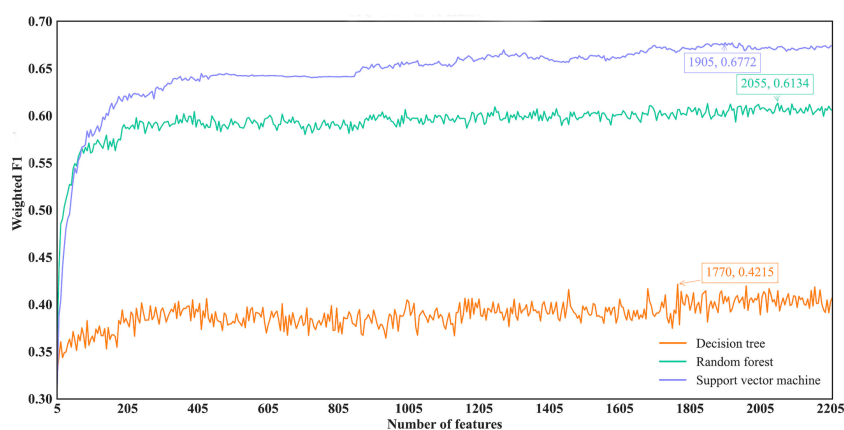


FIGURE 3

IFS curves of different classification algorithms on the Lasso feature list. Three classification algorithms provided highest weighted F1 values of 0.4215, 0.6134 and 0.6772, respectively, based on top 1770, 2055 and 1905, respectively, features in the list.

TABLE 2 Detailed performance of the optimal classifiers for different feature selection methods and classification algorithms.

Feature selection method + classification algorithm	Number of features	ACC	MCC	Macro F1	Weighted F_1
Lasso + DT	1770	0.4218	0.3574	0.4270	0.4215
Lasso + RF	2055	0.6236	0.5844	0.6503	0.6134
Lasso + SVM	1905	0.6811	0.6443	0.7275	0.6772
mRMR + DT	490	0.4373	0.3748	0.4454	0.4347
mRMR + RF	1505	0.6268	0.5880	0.6547	0.6170
mRMR + SVM	1810	0.6271	0.5873	0.6611	0.6200
MCFS + DT	460	0.3982	0.3325	0.4088	0.3966
MCFS + RF	385	0.6024	0.5615	0.6200	0.5917
MCFS + SVM	550	0.5871	0.5401	0.6254	0.5823
LightGBM + DT	1880	0.4278	0.3645	0.4281	0.4273
LightGBM + RF	315	0.6288	0.5893	0.6529	0.6218
LightGBM + SVM	2015	0.6803	0.6430	0.7275	0.6771

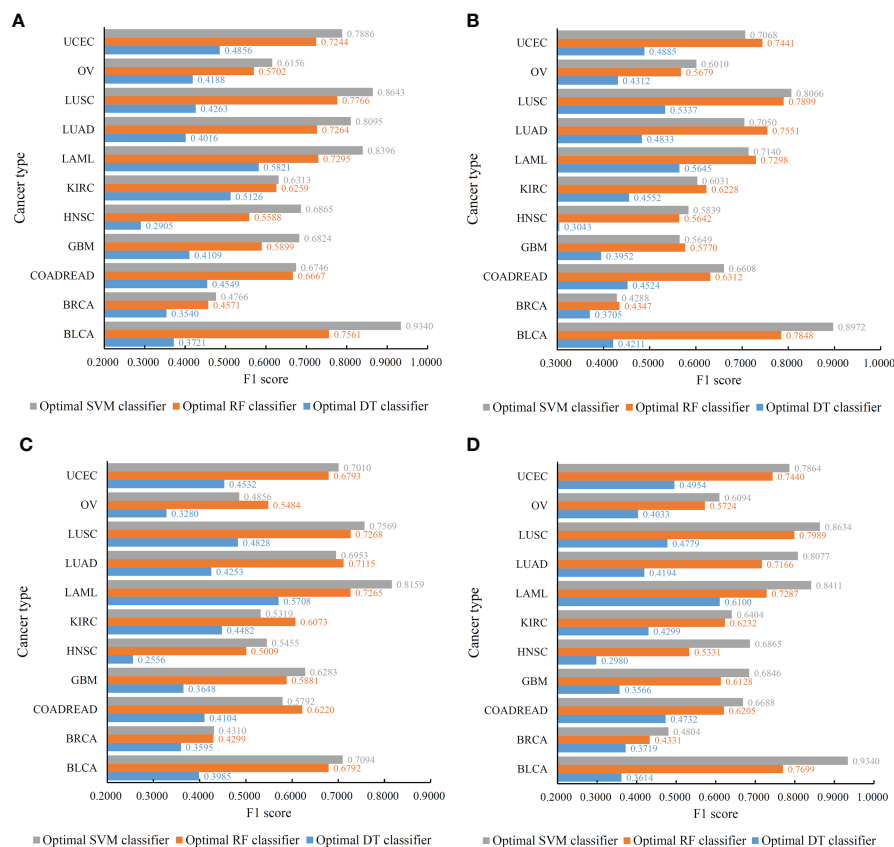


FIGURE 4 Performance of the optimal classifiers on eleven cancer types for different feature lists. (A) Lasso feature list; (B) mRMR feature list; (C) MCFS feature list; (D) LightGBM feature list.

measurements are listed in Table 2 and Figure 4B. The optimal SVM classifier still gave the highest performance. However, its superiority to the optimal RF classifier was not very evident. The optimal DT/RF classifier gave almost the equal performance of

the optimal DT/RF classifier on Lasso feature list, but the performance of the optimal SVM classifier was evidently declined compared with that of the optimal SVM classifier on the Lasso feature list.

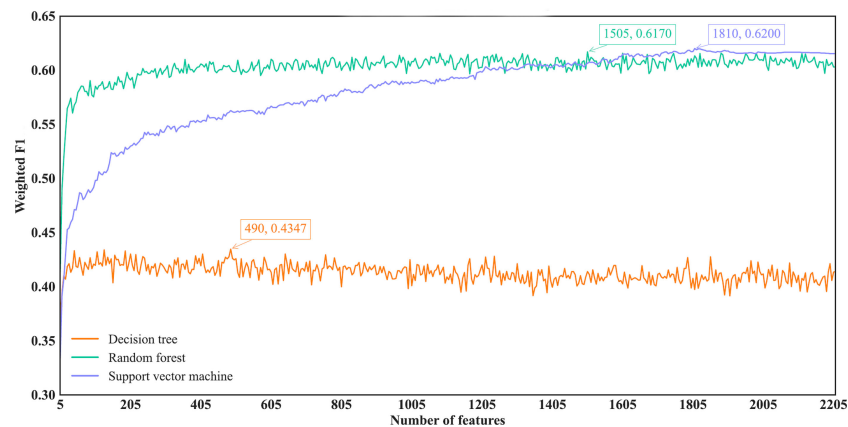


FIGURE 5

IFS curves of different classification algorithms on the mRMR feature list. Three classification algorithms provided highest weighted F1 values of 0.4347, 0.6170 and 0.6200, respectively, based on top 490, 1505 and 1810, respectively, features in the list.

For the IFS results on the MCFS feature list, we also plotted an IFS curve for each classification algorithm to clearly show its performance on different feature subsets, as illustrated in Figure 6. When top 460, 385 and 550, respectively, features in the list were used, DT, RF and SVM provided the highest weighted F1 values of 0.3966, 0.5917 and 0.5823, respectively. Accordingly, the optimal DT, RF and SVM classifiers were built with these optimal features. Their ACC, MCC and Macro F1 are listed in Table 2 and their performance on eleven cancer types is shown in Figure 4C. The optimal RF and SVM classifiers were almost at the same level. Relatively speaking, the optimal RF classifier was slightly better than the optimal SVM classifier. These three optimal classifiers gave lower performance than above optimal classifiers using the same classification algorithm.

As for the IFS results on the LightGBM feature list, similar investigation was conducted. Figure 7 shows the three IFS curves of three classification algorithms. It can be observed that DT, RF and SVM provided the highest weighted F1 values of 0.4273, 0.6218 and 0.6771, respectively, when top 1880, 315 and 2015, respectively, features in the list were adopted. These features were used to build the optimal DT, RF and SVM classifiers. Other overall measurements of these optimal classifiers are listed in Table 2. And their performance on all cancer types is shown in Figure 4D. Evidently, the optimal SVM classifier was better than other two classifiers. The performance of these classifiers is quite similar to that of the optimal classifiers on the Lasso feature list.

Among above optimal classifiers, the optimal SVM classifiers on the Lasso and LightGBM feature lists were evidently better

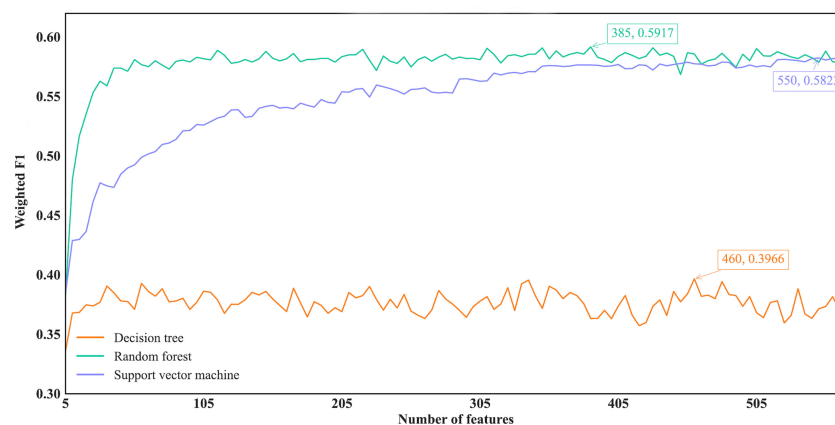


FIGURE 6

IFS curves of different classification algorithms on the MCFS feature list. Three classification algorithms provided highest weighted F1 values of 0.3966, 0.5917 and 0.5823, respectively, based on top 460, 385 and 550, respectively, features in the list.

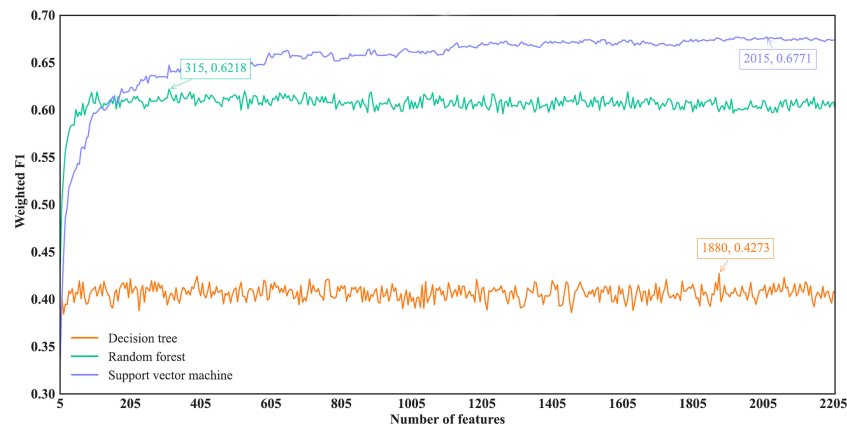


FIGURE 7

IFS curves of different classification algorithms on the LightGBM feature list. Three classification algorithms provided highest weighted F1 values of 0.4273, 0.6218 and 0.6771, respectively, based on top 1880, 315 and 2015, respectively, features in the list.

than other classifiers. They can be efficient tools to classify cancer samples into different types.

3.3 Investigation on key features

Several optimal classifiers were built in Section 3.2. Features used in these classifiers were deemed to be related to cancer classification. Here, we investigated the distribution of these features on three feature types. The distribution is shown in Figure 8. Except the results on MCFS feature list, enrichment features always occupied most. Network features were most for the results on MCFS feature list. Based on different feature

selection methods, some common features can be extracted, whereas some exclusive features can also be discovered by a certain feature selection method. Integrating the results derived from different feature selection methods can give a full overview on cancer classification.

As mentioned in Section 3.2, the optimal SVM classifier was generally the best among all optimal classifiers on a certain feature list. However, these classifiers were of low efficiency due to the large number of features used. In view of this, we carefully checked the IFS results with SVM, trying to finding out a SVM classifier with high performance but with less features. Finally, the SVM classifiers using top 350 features in the Lasso feature list, top 375 features in the mRMR feature list,

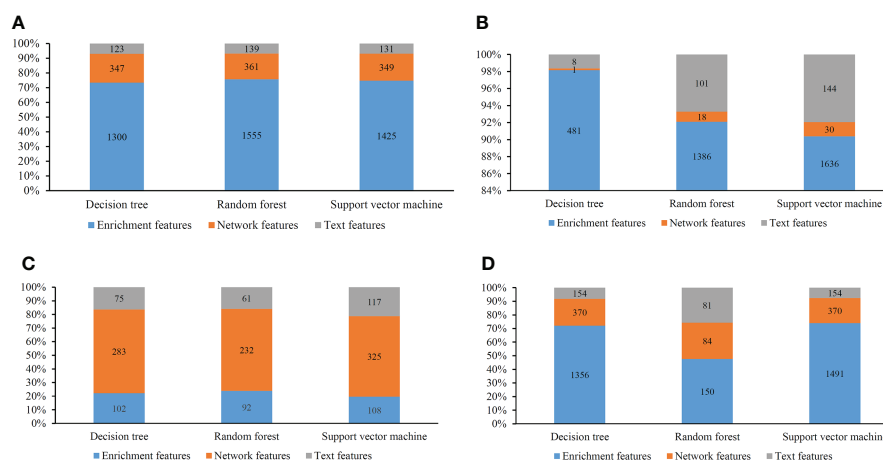


FIGURE 8

Distribution of the optimal features for different classification algorithms and feature lists. (A) Lasso feature list; (B) mRMR feature list; (C) MCFS feature list; (D) LightGBM feature list.

top 150 features in the MCFS feature list and top 315 features in the LightGBM feature list were picked up. For convenience, these classifiers were called feasible SVM classifiers. Their performance is listed in Table 3. It can be observed that their performance is slightly lower than the corresponding optimal SVM classifiers. However, they were more efficient than the optimal classifiers as much less features were involved. Since the optimal features except those used in the feasible classifiers can provide limited improvement, features used in the feasible classifiers were evidently more important than the rest optimal features. Further investigation on these features was helpful to uncover the essential differences of various cancer types. Thus, four feature sets consisting of features in four feasible SVM classifiers were set up and a Venn diagram was plotted, as shown in Figure 9. Detailed intersections are listed in Supplementary Table S3. The results showed that the four feature sets intersected in one important feature (hsa04740), with 26 intersections in three feature sets. These features were deemed to be highly related to cancer classification. Among these 27 (1 + 26) features, 20 were enrichment features (occupied 74%), three were network features and four were text features. The biological implications of these features for cancer classification would be presented in Section 4.1.

3.4 Classification rules derived from the optimal DT classifiers

Although SVM generally achieved the best performance in the above tasks, it is a black-box algorithm that is difficult to interpret in a biological sense. Meanwhile, DT has a low predictive power but can provide easily understandable decision rules because of its IF-THEN rule architecture, which simplifies the discussion on the biological implications of the features. In view of this, DT was used to conduct some additional investigations.

On each feature list, the number of the optimal features for DT had been determined *via* IFS method. Based on these features, DT was applied to all cancer samples and a tree was learned. A set of classification rules was extracted from such tree, which is provided in Supplementary Table S4. The number of rules for each cancer type under different feature selection methods is shown in Figure 10. The results showed that BRCA and HNSC were given plenty of classification rules. The

biological importance of these rules would be discussed in Section 4.2.

4 Discussion

We compared the optimal feature sets identified by the four feature selection methods and found that these methods generated different results (Figure 9). A total of 195 optimal features were identified by two or more feature selection methods, and 772 features were method-specific. The features shared by multiple methods may play key roles in cancer-type-specific development. For example, the feature hsa04550, which is the KEGG pathway of “Signaling pathways regulating pluripotency of stem cells - Homo sapiens (human)” was shared by three methods. Tumorigenesis and the generation of induced pluripotent stem cells (iPSCs) are highly similar processes, and iPSCs from different cell types are led by different reprogramming processes (44). Only one feature was shared by all four methods, which is hsa04740, the KEGG pathway “Olfactory transduction - Homo sapiens (human)”. Previous studies found that the 301 olfactory receptor genes showed different expression patterns in 968 cancer cell lines derived from different cancer types (45); this finding indicated the specific roles of this pathway in different cancers.

4.1 Clustering of optimal GO terms features indicated the functional groups in categorizing the cancer types

Given the abundance of algorithm-specific features, we applied Revigo to cluster the GO terms to assess the relevance of these features in cancer categorization (46). This relevance infers the distance between two GO terms according to the pair-wise semantic similarity. We highlighted the GO terms representing the clusters and ranked the top by four algorithms by displaying their descriptions.

4.1.1 Analysis of biological process

Our literature review confirmed the relevance of these GO terms and clusters to cancer type classification (Figure 11A). For example, a cluster was enriched with GO terms involved in T cell responses. T-cell apoptosis could be triggered by up-regulating

TABLE 3 Performance of the feasible SVM classifiers for different feature selection methods.

Feature selection method	Number of features	ACC	MCC	Macro F1	Weighted F1
Lasso	350	0.6449	0.6038	0.6903	0.6401
mRMR	375	0.5604	0.5150	0.5756	0.5527
MCFS	150	0.5472	0.4957	0.5874	0.5416
LightGBM	315	0.6521	0.6124	0.6887	0.6472

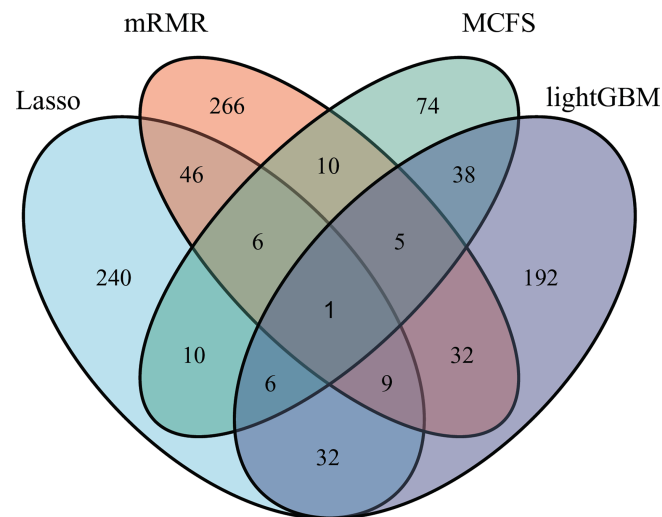


FIGURE 9

Venn diagram to show the intersection of key features identified by different feature selection methods. Lasso, mRMR, MCFS and LightGBM indicate the feature subsets identified by the feature selection methods with the same names.

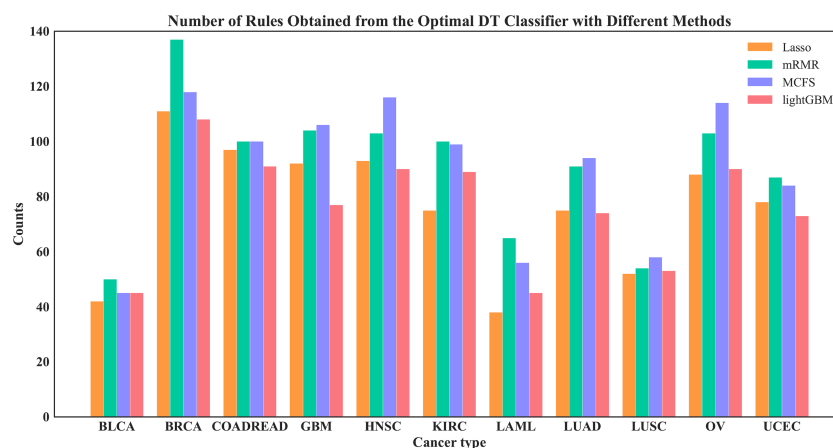


FIGURE 10

Number of rules under each cancer type obtained by the four optimal DT classifiers on four feature lists.

FAS/FASL system in cancer cells. The mutations in FAS and FASL genes reduce the risks of certain types of cancer but not the others, indicating that T cell apoptosis behaves differently in different cancers. IN the present study, cellular responses, especially immune responses involved in T cells, were one of the most critical function groups in distinguishing cancer types. The polymorphisms of the fundamental immunosuppressive cytokine, cytotoxic T-lymphocyte antigen-4 (CTLA4, CD152), which terminates the T-cell response and prohibits T-cell activation, are associated with the risk of breast and cervical cancers (47). This finding proved again the relevance of these

GO term clusters to cancer types. In addition, the GO terms involved in fibroblast growth factor receptor (FGFR) signaling pathway are clustered because FGFRs are recurrently altered in many human cancers. The prevalence of the mutations in this gene depends on the cancer type (48). The other three GO terms clusters include chromosome damage or rearrangement, cellular or tissue development, and regulations of biological processes, including epigenetic modifications. These biological processes have specific signatures in different cancers: different tumors with different origin of cell types are underlined by cancer-type-specific tumorigenesis processes because of the diverse

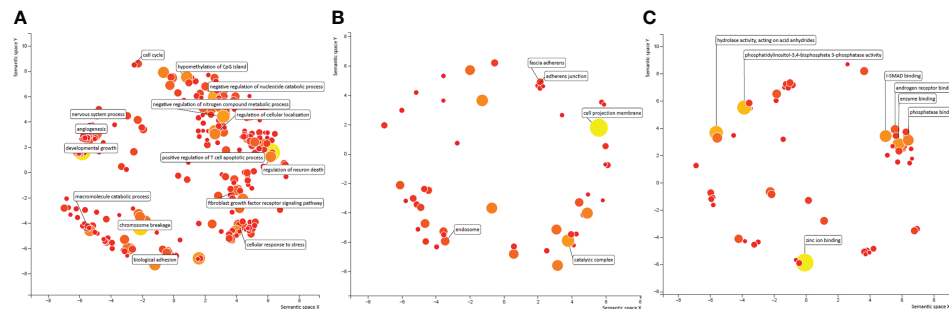


FIGURE 11

(A) GO term clustering of Biological Process identified by only one of the four algorithms. The distance between two GO terms were inferred based on pair-wise semantic similarities. (B) GO term clustering of Cellular Component identified by only one of the four algorithms. The distance between two GO terms were inferred based on pair-wise semantic similarities. (C) GO term clustering of Molecular Function identified by only one of the four algorithms. The distance between two GO terms were inferred based on pair-wise semantic similarities.

characteristics of different cell types. Studies using the Pan-Cancer Analysis of Whole Genome (PCAWG) and The Cancer Genome Atlas (TCGA) data identified chromoanagenesis landscape in different cancers (49), implying the different distributions of mutation types. Moreover, different tumor development mechanisms are caused by aneuploidy, a context-dependent, cancer-type-specific oncogenic event (50).

4.1.2 Analysis of molecular function

In contrast to biological process, we saw only 1 cluster enriched in the other two categories of GO terms, molecular function and cellular component. This cluster was found in the GO term category of cellular function and was enriched by several protein-protein binding functions, such as I-SMAD binding. The nuclear accumulation of active Smad complexes transduces the transforming growth factor beta-superfamily signals from transmembrane receptors into the nucleus. Genetic and epigenetic changes, such as DNA mutations, methylation, and miRNA expression, contribute to the transcriptional activity of TGF- β signaling in all cancer types (51). Previous studies identified the mutation hotspots in SMAD and inhibitors, indicating that the different alterations of Smad and the binding proteins play an important role in different cancer types by regulating TGF- β signaling through various ways.

4.1.3 Analysis of cellular component

In addition to the only cluster enriched by GO terms involved in protein-protein binding, the other GO terms in Cellular Component and Molecular Function showed less similarity with each other (Figures 11B, 11C). However, the top ranked GO terms were well recognized as highly cancer type specific. For example, the top GO term in Cellular Component was cell projection membrane, which is a cell protrusion that is involved in many biological functions, such as cancer cell invasion, cell motility, and cytokinesis. Glypicans play a role in

cellular and tissue development, morphogenesis, and cell motility and show differential expression in different cancer types by behaving as tumor promoters and suppressors in a cancer type-specific manner (52).

In summary, we confirmed that these algorithm-specific features are extensively relevant to cancer types. Each method has a unique strength in a different aspect; therefore, all four methods must be incorporated for the comprehensive inference of cancer type classification.

4.2 Biological relevance of identified rules to cancer type classification

Besides essential features, several interesting classification rules were also obtained in this study (Supplementary Table S4). Here, some rules were examined. We found some features that can distinguish multiple cancer types with high impacts (passed counts ≥ 100). For example, the feature GO:0019002 (GMP binding) can classify BRCA, COADREAD, KIRC and OV, which is expected because GMP is the pharmacological target for treating multiple types of cancers (53–55). Another GO term group that can be used to classify multiple cancer type contains two GO terms, namely, GO:0031049 (programmed DNA elimination) and GO:0031052 (programmed DNA elimination by chromosome breakage), which are also involved in oncogenesis. Activation-induced deaminase is crucial in tumorigenesis because it is implicated in B cell lymphomas. DNA deaminases show preferred targeting, which provides solutions to identify their mutation foot-print in tumors. This finding also indicated their roles in genetic mutation in various cancer types (56). In addition to GO terms, some KEGG pathways can classify multiple cancer types, such as hsa00562 (Inositol phosphate metabolism - Homo sapiens (human)). We found it could be used to distinguish BRCA and COADREAD in

this study. The main influential pathway contributing to CRC was inositol phosphate metabolism (57), which also had the most impact on the metabolic pathway in breast cancer. All these previous findings support our results and suggest the robustness of the methods in the present study.

5 Conclusion

This study was conducted on a cancer mutation dataset. After feature coding, irrelevant features were excluded using Boruta feature selection. Different feature ranking and IFS methods were then employed to identify the optimal number of features, construct efficient classifiers and extract interpretable classification rules. The results of the four methods were combined to identify the most important functional pathways and features, which were further discussed and validated with academic literature, providing a new understanding of the altered biological functions of different cancer types.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://cbio.mskcc.org/cancergenomics/pancan_tcg/.

Author contributions

ZZ, TH and Y-DC designed the study. JL and SD performed the experiments. JRL and JR analyzed the results. JL and JRL wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), Strategic

Priority Research Program of Chinese Academy of Sciences [XDB38050200, XDA26040304], the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences [202002].

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/fonc.2022.979336/full#supplementary-material>

SUPPLEMENTARY TABLE 1

Feature lists obtained by using four feature selection methods after Boruta feature filtering.

SUPPLEMENTARY TABLE 2

Detailed results for incremental feature selection under different feature lists and different classification algorithms.

SUPPLEMENTARY TABLE 3

Venn results for the feature sets used in the feasible SVM classifiers on different feature lists yielded by four feature selection methods.

SUPPLEMENTARY TABLE 4

Classification rules extracted by the optimal DT classifier under each feature list.

References

1. Crosby D, Bhatia S, Brindle KM, Coussens LM, Dive C, Emberton M, et al. Early detection of cancer. *Sci (New York N.Y.)* (2022) 375:eaay9040. doi: 10.1126/science.aay9040
2. Huang AC, Postow MA, Orlowski RJ, Mick R, Bengsch B, Manne S, et al. T-Cell invigoration to tumour burden ratio associated with anti-PD-1 response. *Nature* (2017) 545:60–5. doi: 10.1038/nature22079
3. Screening PDQ, Prevention Editorial B. "Cancer screening overview (PDQ®): Health professional version,". In: *PDQ Cancer information summaries*. Bethesda (MD: National Cancer Institute (US (2002).
4. Donaldson J, Park BH. Circulating tumor DNA: Measurement and clinical utility. *Annu Rev Med* (2018) 69:223–34. doi: 10.1146/annurev-med-041316-085721
5. Aravanis AM, Lee M, Klausner RD. Next-generation sequencing of circulating tumor DNA for early cancer detection. *Cell* (2017) 168:571–4. doi: 10.1016/j.cell.2017.01.030
6. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, Liu MC, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol* (2020) 31:745–59. doi: 10.1016/j.annonc.2020.02.011
7. Ye Q, Wang Q, Qi P, Chen J, Sun Y, Jin S, et al. Development and clinical validation of a 90-gene expression assay for identifying tumor tissue origin. *J Mol Diagnostics* (2020) 22:1139–50. doi: 10.1016/j.jmoldx.2020.06.005
8. Zhang S, Sun K, Zheng R, Zeng H, Wang S, Chen R, et al. Cancer incidence and mortality in chin. *J Natl Cancer Center* (2021) 1:2–11. doi: 10.1016/j.jncc.2020.12.001

9. Divate M, Tyagi A, Richard DJ, Prasad PA, Gowda H, Nagaraj SH. Deep learning-based pan-cancer classification model reveals tissue-of-Origin specific gene expression signatures. *Cancers* (2022) 14:1185. doi: 10.3390/cancers14051185
10. Wang S, Cai Y. Identification of the functional alteration signatures across different cancer types with support vector machine and feature analysis. *Biochim Biophys Acta (BBA) - Mol Basis Dis* (2018) 1864:2218–27. doi: 10.1016/j.bbadis.2017.12.026
11. Liu HA, Setiono R. Incremental feature selection. *Appl Intell* (1998) 9:217–30. doi: 10.1023/A:1008363719778
12. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* (2012) 2:401–4. doi: 10.1158/2159-8290.CD-12-0095
13. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signaling* (2013) 6:pl1. doi: 10.1126/scisignal.2004088
14. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: A web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol* (2007) 8:R3. doi: 10.1186/gb-2007-8-1-r3
15. Mikolov T, Chen K, Corrado G, Dean J. "Efficient estimation of word representations in vector space". In: *International conference on learning representations*. (Scottsdale, Arizona, USA: arXiv) (2013).
16. Mering CV, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: A database of predicted functional associations between proteins. *Nucleic Acids Res* (2003) 31:258–61. doi: 10.1093/nar/gkg034
17. Grover A, Leskovec J. "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. (San Francisco, California, USA: ACM) (2016).
18. Kursa MB, Rudnicki WR. Feature selection with the boruta package. *J Stat Softw* (2010) 36:1–13. doi: 10.18637/jss.v036.i11
19. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* (2006) 101:1418–29. doi: 10.1198/016214506000000735
20. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* (2011) 12:2825–30.
21. Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* (2005) 27:1226–38. doi: 10.1109/TPAMI.2005.159
22. Draminski M, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J. Monte Carlo Feature selection for supervised classification. *Bioinformatics* (2008) 24:110–7. doi: 10.1093/bioinformatics/btm486
23. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. "LightGBM: A highly efficient gradient boosting decision tree". (Redmond, WA, USA: NIPS) (2017).
24. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Trans systems man cybernetics* (1991) 21:660–74. doi: 10.1109/21.97458
25. Breiman L. Random forests. *Mach Learn* (2001) 45:5–32. doi: 10.1023/A:1010933404324
26. Suthaharan S. "Support vector machine". In: *Machine learning models and algorithms for big data classification*. (Boston, MA, USA: Springer) (2016). p. 207–35.
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* (2002) 16:321–57. doi: 10.1613/jair.953
28. Zhou J-P, Chen L, Guo Z-H. iATC-NRAKEL: An efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. *Bioinformatics* (2020) 36:1391–6. doi: 10.1093/bioinformatics/btaa166
29. Chen L, Li Z, Zhang S, Zhang Y-H, Huang T, Cai Y-D. Predicting RNA 5-methylcytosine sites by using essential sequence features and distributions. *BioMed Res Int* (2022) 2022:4035462. doi: 10.1155/2022/4035462
30. Ding S, Wang D, Zhou X, Chen L, Feng K, Xu X, et al. Predicting heart cell types by using transcriptome profiles and a machine learning method. *Life* (2022) 12:228. doi: 10.3390/life12020228
31. Li X, Lu L, Chen L. Identification of protein functions in mouse with a label space partition method. *Math Biosci Eng* (2022) 19:3820–42. doi: 10.3934/mbe.2022176
32. Ran B, Chen L, Li M, Han Y, Dai Q. Drug-drug interactions prediction using fingerprint only. *Comput Math Methods Med* (2022) 2022:7818480. doi: 10.1155/2022/7818480
33. Wang R, Chen L. Identification of human protein subcellular location with multiple networks. *Curr Proteomics*. (2022) 19:344–56. doi: 10.2174/1570164619666220531113704
34. Wu Z, Chen L. Similarity-based method with multiple-feature sampling for predicting drug side effects. *Comput Math Methods Med* (2022) 2022:9547317. doi: 10.1155/2022/9547317
35. Yang Y, Chen L. Identification of drug-disease associations by using multiple drug and disease networks. *Curr Bioinf* (2022) 17:48–59. doi: 10.2174/1574893616666210825115406
36. Zhou X, Ding S, Wang D, Chen L, Feng K, Huang T, et al. Identification of cell markers and their expression patterns in skin based on single-cell RNA-sequencing profiles. *Life* (2022) 12:550. doi: 10.3390/life12040550
37. Kohavi R. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *International joint conference on artificial intelligence*. (Montreal, QC, Canada: Lawrence Erlbaum Associates Ltd). p. 1137–45.
38. Zhao X, Chen L, Lu J. A similarity-based method for prediction of drug side effects with heterogeneous information. *Math Biosci* (2018) 306:136–44. doi: 10.1016/j.mbs.2018.09.010
39. Zhao X, Chen L, Guo Z-H, Liu T. Predicting drug side effects with compact integration of heterogeneous networks. *Curr Bioinf* (2019) 14:709–20. doi: 10.2174/1574893614666190220114644
40. Liang H, Chen L, Zhao X, Zhang X. Prediction of drug side effects with a refined negative sample selection strategy. *Comput Math Methods Med* (2020) 2020:1573543. doi: 10.1155/2020/1573543
41. Tang S, Chen L. iATC-NFMLP: Identifying classes of anatomical therapeutic chemicals based on drug networks, fingerprints and multilayer perceptron. *Curr Bioinf* (2022). doi: 10.2174/1574893617666220318093000
42. Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta (BBA)-Protein Structure* (1975) 405:442–51. doi: 10.1016/0005-2795(75)90109-9
43. Gorodkin J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput Biol Chem* (2004) 28:367–74. doi: 10.1016/j.compbiolchem.2004.09.006
44. Izgi K, Canatan H, Iskender B. Current status in cancer cell reprogramming and its clinical implications. *J Cancer Res Clin Oncol* (2017) 143:371–83. doi: 10.1007/s00432-016-2258-5
45. Ranzani M, Iyer V, Ibarra-Soria X, Del Castillo Velasco-Herrera M, Garnett M, Logan D, et al. Revisiting olfactory receptors as putative drivers of cancer. *Wellcome Open Res* (2017) 2:9. doi: 10.12688/wellcomeopenres.10646.1
46. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One* (2011) 6:e21800. doi: 10.1371/journal.pone.0021800
47. Zhao HY, Duan HX, Gu Y. Meta-analysis of the cytotoxic T-lymphocyte antigen 4 gene +6230G/A polymorphism and cancer risk. *Clin Trans Oncol Off Publ Fed Spanish Oncol Societies Natl Cancer Institute Mexico* (2014) 16:879–85. doi: 10.1007/s12094-014-1159-9
48. Grillo E, Ravelli C, Corsini M, Gaudenzi C, Zammataro L, Mitola S. Novel potential oncogenic and druggable mutations of FGFRs recur in the kinase domain across cancer types. *Biochim Et Biophys Acta Mol Basis Dis* (2022) 1868:166313. doi: 10.1016/j.bbadis.2021.166313
49. Rasnic R, Linial M. Chromoanagenesis landscape in 10,000 TCGA patients. *Cancers* (2021) 13:4197. doi: 10.3390/cancers13164197
50. Ben-David U, Amon A. Context is everything: aneuploidy in cancer. *Nat Rev Genet* (2020) 21:44–62. doi: 10.1038/s41576-019-0171-x
51. Korkut A, Zaidi S, Kanchi RS, Rao S, Gough NR, Schultz A, et al. A pan-cancer analysis reveals high-frequency genetic alterations in mediators of signaling by the TGF- β superfamily. *Cell Syst* (2018) 7:422–437.e427. doi: 10.1016/j.cels.2018.08.010
52. Kaur SP, Cummings BS. Role of glypicans in regulation of the tumor microenvironment and cancer progression. *Biochem Pharmacol* (2019) 168:108–18. doi: 10.1016/j.bcp.2019.06.020
53. Bianchi-Smiraglia A, Wawrzyniak JA, Bagati A, Marvin EK, Ackroyd J, Moparthy S, et al. Pharmacological targeting of guanosine monophosphate synthase suppresses melanoma cell invasion and tumorigenicity. *Cell Death Differentiation* (2015) 22:1858–64. doi: 10.1038/cdd.2015.47
54. Lv Y, Wang X, Li X, Xu G, Bai Y, Wu J, et al. Nucleotide *de novo* synthesis increases breast cancer stemness and metastasis via cGMP-PKG-MAPK signaling pathway. *PloS Biol* (2020) 18:e3000872. doi: 10.1371/journal.pbio.3000872
55. Wang Q, Guan YF, Hancock SE, Wahi K, Van Geldermalsen M, Zhang BK, et al. Inhibition of guanosine monophosphate synthetase (GMPS) blocks glutamine metabolism and prostate cancer growth. *J Pathol* (2021) 254:135–46. doi: 10.1002/path.5665
56. Schmitz K-M, Petersen-Mahrt SK. AIDing the immune system-DIAbolic in cancer. *Semin In Immunol* (2012) 24:241–5. doi: 10.1016/j.smim.2012.07.001
57. Zhu G, Wang Y, Wang W, Shang F, Pei B, Zhao Y, et al. Untargeted GC-MS-Based metabolomics for early detection of colorectal cancer. *Front In Oncol* (2021) 11:729512. doi: 10.3389/fonc.2021.729512



OPEN ACCESS

EDITED BY

Xin Zhang,
Jiangmen Central Hospital, China

REVIEWED BY

Christopher Asquith,
University of Eastern Finland, Finland
Sander Piersma,
Amsterdam University Medical Center,
Netherlands

*CORRESPONDENCE

Zhongjun Liu
zjliu@bjmu.edu.cn
Liang Jiang
jiangliang@bjmu.edu.cn
Wanqiong Yuan
yuanwanqiong@bjmu.edu.cn

[†]These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 11 May 2022

ACCEPTED 14 September 2022

PUBLISHED 30 September 2022

CITATION

Hang J, Ouyang H, Wei F, Zhong Q,
Yuan W, Jiang L and Liu Z (2022)
Proteomics and phosphoproteomics
of chordoma biopsies reveal
alterations in multiple pathways and
aberrant kinases activities.
Front. Oncol. 12:941046.
doi: 10.3389/fonc.2022.941046

COPYRIGHT

© 2022 Hang, Ouyang, Wei, Zhong,
Yuan, Jiang and Liu. This is an open-
access article distributed under the
terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use,
distribution or reproduction is
permitted which does not comply with
these terms.

Proteomics and phosphoproteomics of chordoma biopsies reveal alterations in multiple pathways and aberrant kinases activities

Jing Hang^{1,2,3,4†}, Hanqiang Ouyang^{5,6,7†}, Feng Wei^{5,6,7†},
Qihang Zhong¹, Wanqiong Yuan^{5,6,7*}, Liang Jiang^{5,6,7*}
and Zhongjun Liu^{5,6,7*}

¹Center for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China, ²Key Laboratory of Assisted Reproduction, Ministry of Education, Beijing, China, ³Beijing Key Laboratory of Reproductive Endocrinology and Assisted Reproduction, Beijing, China, ⁴National Clinical Research Center for Obstetrics and Gynecology, Beijing, China, ⁵Department of Orthopedics, Peking University Third Hospital, Beijing, China, ⁶Beijing Key Laboratory of Spinal Disease, Beijing, China, ⁷Engineering Research Center of Bone and Joint Precision Medicine, Beijing, China

Background: Chordoma is a slow-growing but malignant subtype of bone sarcoma with relatively high recurrence rates and high resistance to chemotherapy. It is urgent to understand the underlying regulatory networks to determine more effective potential targets. Phosphorylative regulation is currently regarded as playing a significant role in tumorigenesis, and the use of tyrosine kinase inhibitors in clinical practice has yielded new promise for the treatment of a variety of sarcoma types.

Materials and methods: We performed comprehensive proteomic and phosphoproteomic analyses of chordoma using four-dimensional label-free liquid chromatography–tandem mass spectrometry (LC-MS/MS) and bioinformatics analysis. The potential aberrantly expressed kinases and their functions were validated using western blotting and CCK-8 assays.

Results: Compared with paired normal muscle tissues, 1,139 differentially expressed proteins (DEPs) and 776 differentially phosphorylated proteins (DPPs) were identified in chordoma tumor tissues. The developmentally significant Wnt-signaling pathway and oxidative phosphorylation were aberrant in chordoma. Moreover, we predicted three kinases (AURA, CDK9, and MOK) with elevated activity by kinase-pathway network analysis (KiPNA) and verified their increased expression levels. The knockdown of these kinases markedly suppressed chordoma cell growth, and this was also the case for cells treated with the CDK9 inhibitor AZD4573. We additionally examined 208 proteins whose expression and phosphorylation levels were synergetically altered.

Conclusions: We herein depicted the collective protein profiles of chordomas, providing insight into chordomagenesis and the potential development of new therapeutic targets.

KEYWORDS

chordoma, proteomics, phosphoproteomics, integrated kinase network, kinase inhibitor

Introduction

Chordoma is a rare but malignant primary bone tumor that typically occurs along with the axial skeleton, with an incidence rate of approximately two cases per million people that represents 1%–4% of all bone cancers (1, 2). Chordoma is hypothesized to arise from the neoplastic transformation of embryonic notochordal remnant tissues (3). The rod-like notochord produces several signaling proteins (e.g., sonic hedgehog and Wnt) and secretes signaling factors (e.g., BMP and FGF) to direct organogenesis within the early embryo. The T-box-family transcription factor brachyury (*TBXT*), a crucial regulator of notochord development, and is characteristically overexpressed in chordomas and usually regarded as a hallmark of chordoma (4, 5). Despite its manifestation as a low-to-intermediate grade tumor that usually occurs sporadically, chordoma exhibits a strong tendency toward local recurrence, and over 30% of patients develop metastases (6).

As chordoma shows low sensitivity to conventional cytotoxic chemotherapies and radiation, radical surgery remains the primary and most effective treatment; however, their proximity to neurovascular structures hampers complete resection, and some patients have already developed metastatic diseases upon initial diagnosis. Several recent studies (including our own) have revealed that EGFR-signaling dysbiosis modulated by CMTM3 and cMET exacerbated the chordoma process (7, 8), and that DEPDC1B regulated ubiquitination of BIRC5 also promoted chordoma progression (9). However, more information should be depicted to overcome its lengthy latency and poor response to treatment (10).

Phosphorylation constitutes one of the most common post-translational modifications (PTMs) of proteins, with kinases exhibiting tumorigenic actions in various malignancies, and relevant small-molecule inhibitors have been widely and successfully employed for the clinical treatment of diverse types of cancer (11). Indeed, a series of tyrosine kinase inhibitors (TKIs) have been used in the pre-clinical trials of chordoma (12), including the inhibitors against the epidermal growth factor receptor (EGFR) like afatinib, erlotinib, and lapatinib (13), the inhibitors against tyrosine-protein kinase Met (c-MET) like crizotinib, the inhibitors against Platelet-derived growth factor receptor (PDGFR) such as

imatinib (14), and the inhibitors of the type I IGF receptor/insulin receptor (IGF1R/INSR) inhibitor linsitinib together with erlotinib (15). However, their clinical applications all exhibited varied degrees of limitations. The responses of lapatinib were not as good as it could have been expected (16). Erlotinib incompletely suppressed chordoma growth *in vivo* (17). Not all PDGFR-positive chordomas patients are sensitive to imatinib and it did not achieve dimensional and long-lasting responses as well (18, 19). Although the European multi-center clinical trial involving afatinib is still ongoing, it is well established that single-afatinib therapies have not yielded lasting effects and the tumoral evolutions lead to the drug resistance (20, 21). Besides, machine learning together with synergistic drug combinations have also been proposed in chordoma (22).

A majority of recent studies on chordoma have focused on histologic and genetic analysis (23–28). Moreover, although numerous investigators have assessed the roles of several individual proteins in chordoma (29, 30), we only uncovered a few reports on chordoma-related proteins that exploited proteomics (31–33) and little is known with respect to the PTMs of proteins. Therefore, a more comprehensive understanding of chordoma is urgently needed to elucidate phosphorylation status and the aberrant activation of signaling pathways that contribute to chordomagenesis.

In the present study, we profiled and integrated the proteomic and phosphoproteomic landscapes with label-free mass spectrometry, identified several kinases as potentially druggable molecular targets, and verified the effects of these kinases on chordoma cell proliferation. Our findings will enable us to better understand the underlying pathogenesis of chordoma, and thus yield novel targets and enhance the effectiveness of clinical treatments.

Materials and methods

Patients and tissue samples

Patients that had been diagnosed and treated as chordoma according to the preoperative images (including plain and enhanced head Magnetic Resonance Imaging (MRI), thin layer skull base Computed Tomography (CT) scanning, and 3-D

reconstruction) at the Peking University Third Hospital were included in this study. All the cases were diagnosed by two independent pathologists using immunohistochemistry and histopathological examination. The detailed clinical characteristics of the enrolled fifteen patients are listed in Table 1. Chordoma tumor samples were obtained at the time of surgery, washed efficiently using ice-cold phosphate-buffered solution (PBS), snap-frozen in liquid nitrogen for 3 min, and stored at -80°C for further use. The distal normal muscle tissues were obtained as control and processed in the same way as chordoma samples. These paired tissues were used for the following integrated proteomics and phosphoproteomics analyses.

Sample preparation

Protein extraction and digestion

We weighed and ground samples into cell powder using a mortar and liquid nitrogen. Then we added four volumes of lysis buffer (8M Urea, 100 mM NaCl, 100 mM Tris-HCl, pH 8.0, 1% Protease Inhibitor, 1% Phosphatase Inhibitor) and treated the samples with sonication. The cell debris was removed and the supernatant was collected. The total protein concentration was determined with the BCA kit (Thermo Fisher Scientific, USA) following the manufacturer's instructions.

The protein quantity and volume from each sample were adjusted to the same, and TCA were added to a final concentration of 20% and the proteins were precipitated at 4°C for 2 hours. The protein pellets were collected by centrifugation at 4,500 g for 5 min and washed with pre-cold acetone 3 times. 100 mM TEAB was added and the precipitation was dispersed by ultrasound. A 1:50 mass ratio of trypsin was used for the overnight digestion for 800 μg proteins. Then the solutions were reduced with 5 mM dithiothreitol (DTT) for 30 min at 56°C and alkylated with 11 mM iodoacetamide (IAA) for 15 min at room temperature in darkness. The digested peptides were desalted with C18 SPE (3M company) according to the manufacturer's instructions.

Phosphopeptides enrichment

0.8 mg peptides per sample was used for phosphoproteomic analysis. Peptide mixtures were reconstituted in loading buffer (50% acetonitrile/6% trifluoroacetic acid) and enriched by immobilized metal-chelated affinity chromatography (IMAC) microsphere (High-SelectTM Fe-NTA Phosphopeptide Enrichment Kit). After gentle incubation by rotation and successive washing with 50% acetonitrile/6% trifluoroacetic acid and 30% acetonitrile/0.1% trifluoroacetic acid, the bound phosphopeptides were eluted with 10% NH_4OH . The eluted phosphopeptides were vacuum-lyophilized and desalted by C18 ZipTips (Millipore).

Label-free liquid chromatography-tandem mass spectrometry

LC-MS/MS analysis

The tryptic peptides of 0.5 μg were dissolved in solvent A (0.1% formic acid in 2% acetonitrile) and then separated on a homemade reversed-phase analytical column of 25-cm length, 100 μm i.d. (packed with 1.9 μm /120 Å ReproSil-PurC18 resins) using NanoElute Ultra-Performance Liquid Chromatography (UPLC) system. The gradient was comprised of an increase from 7% to 24% over 72 min, 24% to 32% over 12 min, and climbing to 80% in 3 min, then holding at 80% for the last 3 min. All at a constant flow rate of 450 nL/min. The peptides separated by the UPLC system were then subjected to ionization by both electron spray (the Capillary) and captivespray source, followed by tandem mass spectrometry (MS/MS) in timsTOF Pro (Bruker Daltonics, USA) for analysis. The electrospray voltage applied was 1.6 kV. 1/k0 was set as 0.75–1.40 $\text{V}\cdot\text{s}/\text{cm}^2$; resolution was set as custom; ramp time was set as 100.0 ms; spectra rate was set as 9.43 Hz, and lock duty cycle was set to 100%. The m/z scan range was 100 to 1,700 for the MS2 spectrum. The data acquisition used a parallel accumulation serial fragmentation (PASEF) procedure. We selected precursors with charge states 0 to 5 for fragmentation, and 10 PASEF-MS/MS scans were acquired per cycle and set the dynamic exclusion time as 30 s. The overall procedures for phosphoproteomics were similar, except that the peptides were dissolved in solvent C (0.1% formic acid in water) and separated by an IonOpticks C18 reversed-phase analytical column (100 μm i.d. \times 25 cm). The elution gradient was set as: 0–50 min, 2%~22%; 50–52 min, 22%~35%; 52–55 min, 35%~90%; 55–60 min, 90%. The flow rate was set constantly at 300 nL/min.

Database search

The resulting MS/MS data were processed using the Maxquant search engine (v.1.6.6.0) using “match between runs”, “second peptide search” and LFQ. Tandem mass spectra were searched against the human SwissProt database concatenated with reverse decoy database. Trypsin was specified as cleavage enzyme allowing up to 2 missing cleavages. The minimum length of peptide segment was set as 7 amino acid residues and the maximum modification number of the peptide segment was set to 5. The mass tolerance for precursor ions was set as 20 ppm in First search and 20 ppm in Main search, and the mass tolerance for fragment ions was set as 0.02 Da. Alkylation (cysteine) was specified as fixed modification, while oxidation (methionine), acetylation (protein N-term), desamidization (asparagine and glutamine) were specified as variable modifications. The FDRs of both protein identification and peptide spectrum matching (PSM) were set to 1%. The addition of phosphorylation (serine, threonine, and tyrosine) as modifications was used for phosphoproteomic data.

Bioinformatics analysis

Statistical analysis

We used horizontally normalized LFQ (MaxQuant) as a quantitative protein intensity. This value is divided by the mean quantification of all samples and used as the normalized value for subsequent analysis. To compare the differences between these nine chordoma tumor tissues (CT) and their nine paired normal tissues (CN) that applied to proteomics profiling, we compared the expression of each protein with paired *t*-test and determined differentially expressed proteins (DEPs) and differentially phosphorylation sites (DPSs) with $|\text{fold change (FC)}| \geq 2.0$ and $p\text{-value} < 0.05$. We used statistical analysis methods including principal component analysis (PCA) and Pearson's correlation coefficient to test sample repeatability. Statistical significance ($p < 0.05$) was assessed by using Student's *t*-test. Based on clinical characteristics, we divided the tumor tissues into large (diameter > 5.0 cm, $n=4$) and small (diameter < 5.0 cm, $n=5$) (L/S) subgroups. DEPs and DPSs were determined similarly. The statistics were analyzed using `ttest_ind` function in `scipy` of Python 3.7.6.

Pathway enrichment analyses and hierarchical clustering

For the annotation and enrichment of DEPs, Kyoto Encyclopedia of Genes and Genomes (KEGG) database were classified against all identified proteins by two-sided Fisher's exact probability test using `fisher_exact` function in `scipy.stats` of Python 3.7.6. The KEGG with an adjusted $p\text{-value} < 0.05$ was considered significant.

To find the functional correlation of different groups between DEPs or DPPs, we performed clustering analysis by hierarchical clustering and visualized it by heatmap based on p -values from Fisher's exact test using `pheatmap` function in R package "pheatmap". The horizontal dimension of the heatmap represented Fisher's exact test results of different groups, while the longitudinal one described functional classification.

Protein-Protein Interaction (PPI) Network

The database accession numbers or protein sequences of all DEPs in different groups were searched in the STRING database (version 11) for protein-protein interactions. We fetched all interactions with a confidence score ≥ 0.7 . Interaction network form STRING was visualized *via* `forceNetwork` function in R package "networkD3".

Categorical gene set enrichment analysis (GSEA)

GSEA was performed on total proteomics and phosphoproteomics with GSEA v4.0.3 software (34). Official gene sets were downloaded from the GSEA website (www.broadinstitute.org/gsea/) for enrichment. A permutation number of 1000 was adopted.

Kinase activity prediction

Prediction of kinase-substrate regulations

We adopted iGPS V1.0 software (35) which utilized GPS2.0 algorithm for the prediction of site-specific kinase-substrate relations, and PPI information was used as the major contextual factor to filtrate potentially false-positive hits.

Prediction of kinase activities

We used Gene Set Enrichment Analysis (GSEA) method that adopted from the established PTM signature enrichment analysis (PTM-SEA) (36) to predict kinase activities. Normalized enrichment scores (NESs) of enrichment results were regarded as kinase activity scores. For each kinase, the kinase was predicted as positive if the predominant change of substrates was an increase in phosphorylation and vice versa.

Kinase-pathway network analysis (KiPNA)

We used a network analysis framework that is developed for relating kinase signaling with pathway dysregulation (37). Nodes stand for significantly enriched proteins, leading edge phosphosites, enriched kinases, and enriched pathways. The connected edges include enriched kinases, leading-edge phosphosites connect to target proteins, and proteins connect to enriched pathways. Networks were visualized using `cytoscape`.

Cell line, cell culture, and transfection

Human chordoma cell lines JHC7 and U-CH1 were purchased from American Type Culture Collection (ATCC, Manassas, VA, USA). JHC7 cells were cultured in DMEM: F12 medium (ATCC® 30-2006™, ATCC, USA) and U-CH1 cells were cultured in RPMI-1640 medium (11875093, Gibco, US) supplemented with an additional 1% L-glutamine (25030081, Gibco, US). To culture U-CH1, coating buffer (50 $\mu\text{g/mL}$ rat tail type I collagen (354236, BD Biosciences) was added to the culture flask for one hour at room temperature prior to adding cells. All cells were cultured with 10% characterized fetal bovine serum (FBS) (10099141, Gibco, US), 10 units/mL penicillin and 10 mg/mL streptomycin (10378016, Gibco, US), and maintained in humidified incubators at 37°C with 5% CO₂.

To knock down endogenous AURA, CDK9, and MOK, siRNA constructs were generated with the target sequences shown in Table S4. The siRNAs were purchased commercially (HanBio Inc., Shanghai, China). Chordoma cells were transfected with siRNA at a final concentration of 50 nM using Lipofectamine 3000 transfection reagent (Catalog No. L3000015, Invitrogen, USA). The suppression efficiency of AURA, CDK9, and MOK was analyzed by western blot on the third day post transfection.

Western blotting

Proteins were lysed and extracted by RIPA (Thermo Fisher Scientific) cell lysis buffer supplemented with protease inhibitor and phosphatase inhibitor cocktails. Protein quantification was measured by the Pierce BCA protein assay kit (Thermo Fisher Scientific). 10 μ g of total proteins were applied to SDS-PAGE electrophoresis, transferred to PVDF membrane, and detected by conventional protocols for western blotting. Primary antibodies against Aurora kinase (AURA) (66757-1-Ig, Proteintech, US), Cyclin-dependent kinase 9 (CDK9) (2316, CST, US) and MAPK/MAK/MRK overlapping kinase (MOK) (23926-1-AP, Proteintech, US), Brachyury-T (81694, CST, US) and β -actin (4970, CST, US), and subsequently with the appropriate horseradish peroxidase-conjugated secondary antibodies (7074, CST, US).

Cell proliferation analysis

Cell proliferation was measured by a Counting Kit-8 (CCK-8) detection kit (Catalog No. CK04, Dojindo Molecular Technologies, Japan). The cells were seeded in a 96-well plate with 3000 cells per well and treated with siRNA transfection. At the indicated time points, 10 μ L of CCK-8 solution was added, followed by incubation at 37°C for two hours, and absorbance at 450 nm was determined.

For *in vitro* inhibitor assays, AZD4573 (S8719, Selleck) was dissolved to 10 mM in DMSO and then dosed to yield a final DMSO concentration of \leq 0.3%. Chordoma cells was treated with different concentration of AZD4573 from 0.4 nM to 4000

nM (ten-folds dilution) for nine days and the cell proliferation was tested by CCK-8 as well.

Immunohistochemical and hematoxylin-eosin staining

IHC and HE staining were performed on formalin-fixed, paraffin-embedded tissues as previously described (38). Briefly, samples were incubated with primary anti-Brachyury antibody (81694, CST, USA, 1:200 dilution) at 37°C for 1 hour and second antibody at 37°C for 30 min. Then, the samples were treated with the DAB Substrate kit (PV-8000, ZSGB-BIO, China).

Results

Overview of the global proteomic and phosphoproteomic landscape with respect to chordoma

We enrolled 15 patients who were diagnosed with chordoma according to both preoperative images and IHC examination (Table 1; Figures S1A–D): samples from nine of these were subjected to proteomic study and the other six were used for western blot analysis. Since it is impracticable to sample the embryonic notochord (the ideal chordoma control), we compared CT to their CN as performed by many other investigative groups (32, 39, 40) (our entire protocol is illustrated in Figure 1A). We then chose to further validate

TABLE 1 Summary of clinical information of chordoma tumor samples and classification.

Patient Code	Age	Gender	Tumor size (cm)	Size classification [#]	Site	primary or recurrent	usage in this study
C1	34	Male	3.8 \times 4.8 \times 5.4	Large	C	primary	Proteomics
C2	64	Male	2.8 \times 4.4 \times 3.4	Small	C	primary	Proteomics
C3	62	Male	3.4 \times 3.7 \times 5.8	Large	C	primary	Proteomics
C4	21	Female	5.9 \times 4.4 \times 7.5	Large	C	recurrent	Proteomics
C5	50	Female	2.5 \times 2.0 \times 2.8	Small	C	recurrent	Proteomics
C6	69	Female	2.4 \times 1.5 \times 2.4	Small	C	primary	Validation
C7	71	Female	4.3 \times 4.2 \times 3.8	Small	T	primary	Validation
C8	66	Female	2.4 \times 2.7 \times 3.1	Small	T	primary	Proteomics
C9	48	Male	9.7 \times 2.3 \times 2.6	Large	C	primary	Proteomics
C10	56	Female	3.7 \times 2.6 \times 2.3	Small	T	recurrent	Validation
C11	70	Male	14.6 \times 9.8 \times 14.3	Large	C	recurrent	Proteomics
C12	58	Male	3.1 \times 2.1 \times 4.8	Small	C	recurrent	Proteomics
C13	65	Female	9.8 \times 3.4 \times 4.8	Large	T	recurrent	Validation
C14	58	Female	3.9 \times 4.2 \times 5.1	Large	C	primary	Validation
C15	63	Male	3.1 \times 2.2 \times 2.7	Small	T	primary	Validation

[#]Large is identified as at least one edge is over 5.0 cm. C, cervical spine; T, thoracic vertebra.

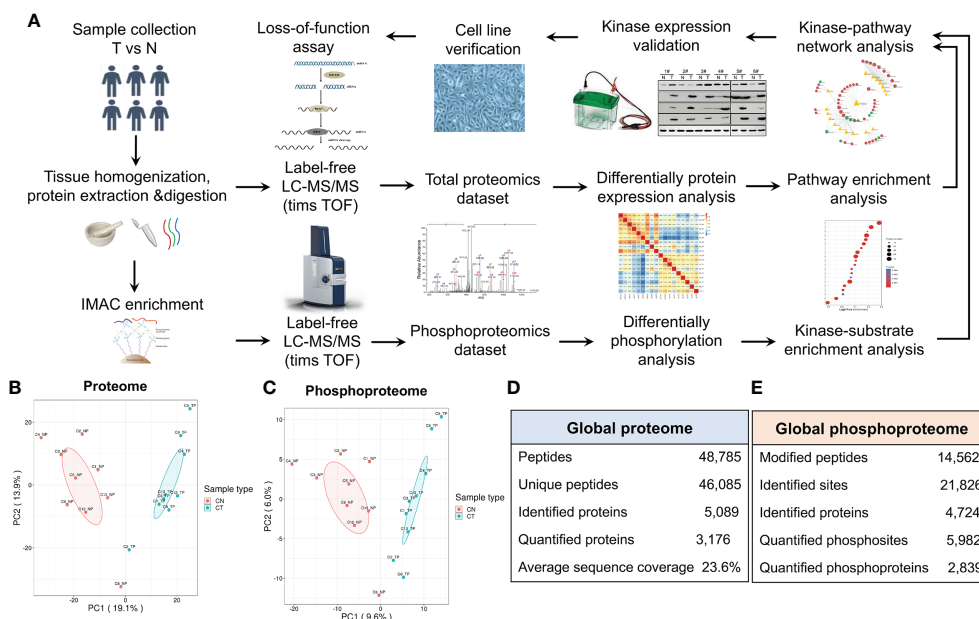


FIGURE 1

Characterization of the proteome and phosphoproteome in chordoma tumor and normal tissues. (A) General pipeline of (phospho)peptide enrichment and the quantitative mass spectrometric protocol followed by pathway analyses and biochemical validation. (B, C) Principal component analysis (PCA) of the quantified proteome (B) and quantified phosphoproteome (C). (D, E) Statistical analysis of the 4D label-free proteomic (D) and phosphoproteomic (E) datasets.

some intriguing and significant targets retrieved from the high-throughput LC-MS/MS results.

PCA of both the proteome and phosphoproteome revealed similarities among the sample groups (Figures 1B, C), confirming the validity of our dataset and indicating its reliability for further assessment. We then added mobility (4-D label-free) to our LC-MS/MS analysis to increase the peptide-identification depth. In total, we identified 46,085 unique peptides from 5,089 proteins and 21,826 phosphosites from 4,724 proteins (Figures 1D, E).

Identification of differentially expressed proteins and alterations in the Wnt-signaling pathway

We divided the DEPs into four quantified groups (Q1–Q4) according to their differentially expressed magnification and determined that 364 proteins in Q1 ($0 < \text{ratio} \leq 0.5$) and 775 proteins in Q4 ($\text{ratio} > 2$) were down- and up-regulated proteins, respectively (Figure 2A). These 1,139 proteins were regarded as significantly DEPs, and visualization of their distribution patterns indicated that a majority were upregulated (Figure 2B). Hierarchical clustering of DEPs indicated that the protein expression status of CT was different from that of CN (Figure 2C). Then we adopted functional KEGG-pathway

analysis on all the DEPs to evaluate their biological processes, and noted that oxidative phosphorylation (the metabolic pathway linking electron transport and phosphorylation) was dramatically altered in chordoma tissues (Figure 2D). Since most of the proteins were over-expressed in chordoma, we then analyzed the KEGG pathway for the upregulated proteins and ascertained that non-homologous end-joining was dramatically enriched (Figure S2A). We also executed GSEA to better elucidate the functions of the DEPs (Figure S2B) and found oxidative phosphorylation and lysosome to be enriched using both KEGG and GSEA (Figures S3A–D), suggesting their relatively robust functions in chordomagenesis.

As has been stated previously, the Wnt-signaling pathway plays a crucial role in embryonic and notochord development. We therefore assessed whether this signaling pathway was aberrantly regulated in chordoma, we determined that at least 10 proteins of the ~100 known proteins in the Wnt-signaling pathway exhibited anomalies (six were upregulated and four were downregulated; Figures 2E–N). The pivotal protein β -catenin (gene symbol: *CTNNB1*) showed an elevated (~two-fold) expression (Figure 2E), and as a subunit of the cadherin protein complex that acts as an intracellular signal transducer, mutations and increased expression of β -catenin are associated with many cancers (41). Our findings signified that chordomas shared similar behaviors. For example, calcium/calmodulin-dependent protein kinase type II (CaM-kinase II) is a ubiquitous Ser/Thr-

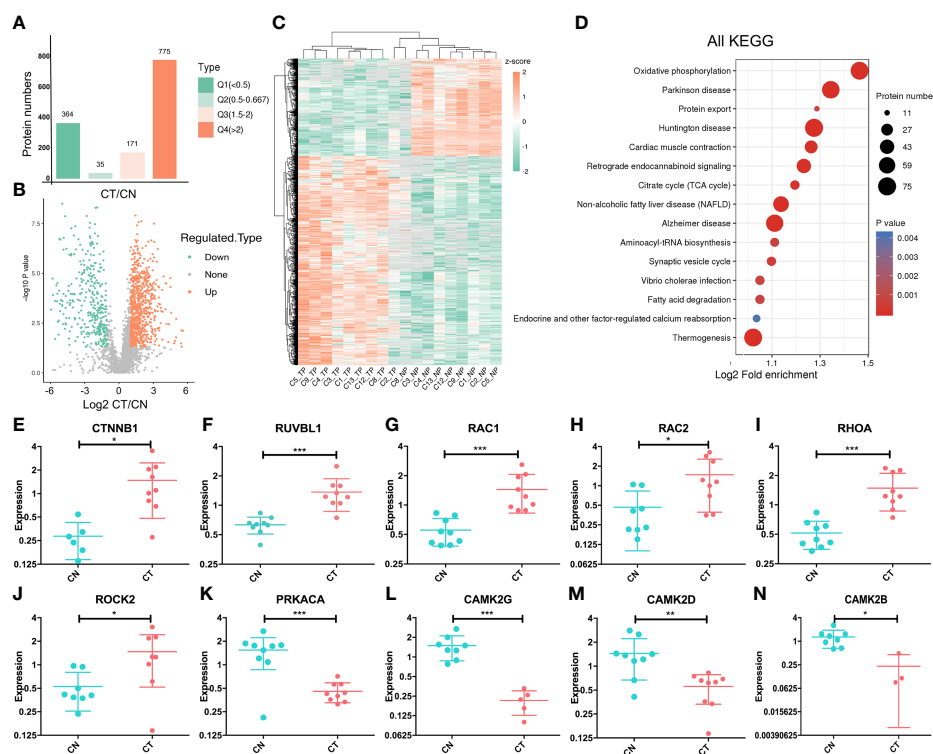


FIGURE 2

Targeted proteomic study and altered expression of Wnt-signaling-related proteins. **(A)** Statistical comparison of regulated proteins between chordoma tissues (CT) and their paired normal adjacent muscle tissues (CN) by Q category. Q1, 0 < ratio ≤ 0.5, corresponding to downregulated proteins; Q2, 0.5 < ratio ≤ 0.667; Q3, 1.5 < ratio ≤ 2; Q4, ratio > 2, corresponding to upregulated proteins. Protein numbers for Q1–Q4 are 364/35/171/775, respectively. **(B)** Volcano plot of the distributional patterns of statistical significance (-log P value) and magnitude of the changes (log₂FC) for all DEPs. **(C)** Unsupervised hierarchical clustering heatmap of the significantly regulated proteins identified from chordoma tissues. Unique proteins (n=1,139; rows) were significantly identified from nine paired samples (columns). TP, tumor proteins; NP, normal proteins. Unsupervised hierarchical clustering was performed using the Cluster program with Pearson correlation and pairwise complete-linkage analyses. **(D)** KEGG pathway-enrichment analysis was executed to identify important pathways that depended upon all DEPs. The colored blocks that correspond to functional classification indicate the magnitude of enrichment, and are displayed by colors ranging from blue (weak enrichment) to red (strong enrichment). **(E–N)** Relative normalized expression of CTNNB1 (E), RUVBL1 (F), RAC1 (G), RAC2 (H), RHOA (I), ROCK2 (J), PRKACA (K), CAMK2G (L), CAMK2D (M), and CAMK2B (N). The vertical axis signifies log (2). CN, blue; CT, red. Student's paired t-test was applied to distinguish the expression differences; *, p < 0.05, **, p < 0.01, ***, p < 0.001.

directed protein kinase comprising a family of four genes (alpha, beta, gamma, and delta) that regulates the Wnt pathway (42). In our dataset, three of these four proteins (CAMK2G, CAMK2B, and CAMK2D) were downregulated (Figures 2L–N), suggesting a potential for aberrant regulation of CaM-kinase II and its corresponding phosphorylation.

Integration of the proteome and phosphoproteome reveals multiple regulatory levels

In addition to determining protein expression levels, we quantified the DPPs to elucidate their potential impacts, and identified 823 upregulated phosphosites that corresponded to 593 proteins and 403 downregulated phosphosites that corresponded to

183 proteins (Figure 3A). We subsequently implemented KEGG pathway enrichment analysis for up- and down-regulated DPPs, demonstrating that 34 pathways were enriched, and we delineated the top 10 upregulated pathways in Figure 3B. Notably, spliceosome (involving 23 DPPs) was the most significantly different between CT and CN tissues (Table S1), and five proteins that contained multiple phosphorylation sites were dually phosphorylated and functioned in both up- and down-regulation (Table 2), thereby indicating a complicated regulatory function for protein PTM. The Notch-signaling pathway was also markedly enriched, indicating its aberrant regulation in chordoma formation (Figure 3B).

We next analyzed the proteins that were regulated at both the expression and post-translational modification (PTM) levels (Figures 3C, D). Using the intersection of the proteomic and phosphoproteomic datasets, we uncovered 208 proteins of which 55.8% were upregulated at both levels (Figure 3D, Table S2).

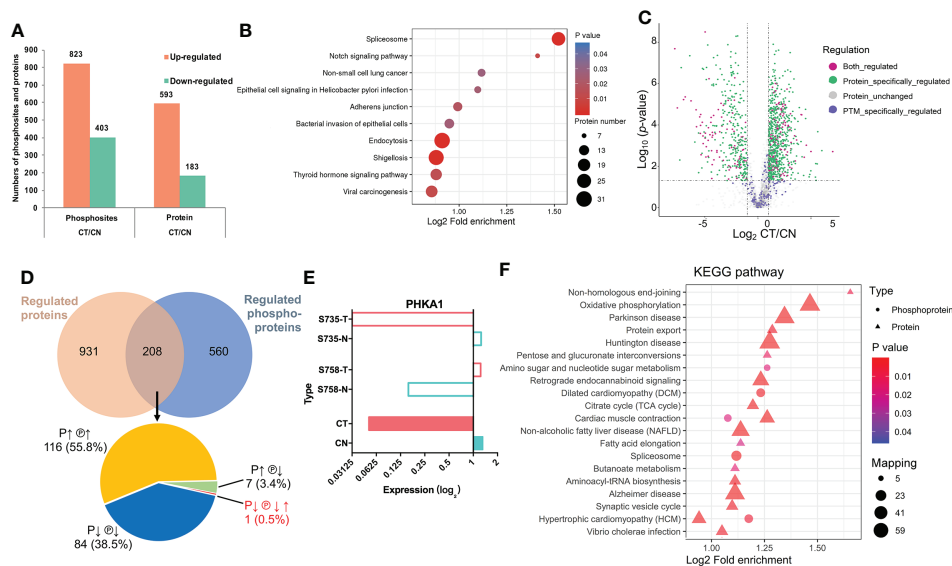


FIGURE 3

Phosphoproteomic analysis and its combinative interpretation in relation to the proteome. **(A)** Statistical comparison of differentially phosphorylated sites and proteins between CT and CN. **(B)** KEGG pathway-enrichment analysis was implemented separately for the upregulated DPPs. **(C)** Volcano plot of the distribution patterns of protein-specific (green), phosphorylation-specific (purple), and combined (magenta) protein regulations. **(D)** Venn diagram of proteomics and phosphoproteomics analyses. Two-hundred eight proteins were found to be dually regulated: of these, 116 (55.8%) were both over-expressed (P↑) and hyper-phosphorylated (P↑@↑); 84 (38.5%) exhibited the opposite conditions (P↓@↓); seven (3.4%) proteins were over-expressed but hypo-phosphorylated (P↑@↓); and while the expression of one (0.5%) protein was downregulated (P↓), it possessed two phosphosites that showed opposing patterns (P↓@↑). **(E)** Detailed information with respect to PHKA1 regulation. Although serine735 (S735) and serine758 (S758) were phosphorylated in opposing fashions, their expression was dramatically downregulated. Colors are the same as in Figures 2E–N (CN, blue; CT, red). **(F)** Combined KEGG pathway-enrichment analysis for DEPs and DPPs. Circles and triangles designate DEPs and DPPs respectively, with different colors denoting their differential expression (red represents upregulation and blue represents downregulation).

TABLE 2 Proteins with both up- and down- regulated phosphorylated sites.

Protein accession	Name	Position	Ratio	P value	Regulated Type
O14639	ABLIM1	490S	0.333	0.049	Down
		655S	7.651	0.021	Up
		706S	4.732	0.009	Up
P14618	PKM	77S	0.166	0.000	Down
		127S	0.226	0.006	Down
		519S	2.400	0.008	Up
P27816	MAP4	696S	0.361	0.032	Down
		928S	7.811	0.027	Up
Q15149	PLEC	201S	0.225	0.000	Down
		1435S	0.266	0.008	Down
		2782S	7.928	0.045	Up
		4618S	10.853	0.014	Up
Q8WWI1	LMO7	4385S	13.973	0.007	Up
		704S	2.541	0.015	Up
		805S	0.357	0.010	Down
		991S	0.424	0.022	Down
		1510S	0.312	0.004	Down
		1586S	0.294	0.008	Down

Intriguingly, the phosphorylase b kinase regulatory subunit alpha (PHKA1) displayed a unique pattern, with diminished protein abundance but containing both hyper- and hypo-phosphorylation sites (Figure 3E). We also noted seven proteins that showed downregulated phosphorylation but upregulated expression, indicating that phosphorylation influenced their protein abundance and consequent functions. KEGG pathway-enrichment analysis was ultimately performed on the 208 proteins to characterize their possible biologic functions more intuitively (Figure 3F).

Cell proliferation-related kinases promote chordomagenesis

To investigate our datasets more systematically and to uncover some potential targets, we executed KiPNA tailored

for phosphoproteomic profiles (Figure 4A). KiPNA is an integrated computational approach that enables us to predict the activities of kinases based on the differences in their substrate phosphorylation. We identified 28 possible kinases with altered activity (Figure 4B; Table S3), and initially chose the top 10 for validation (data not shown). Three of these kinases (Aurora kinase A, AURA; Cyclin-dependent kinase 9, CDK9; and MAPK/MAK/MRK overlapping kinase, MOK) appeared to show significant differences between CT and CN samples. Aurora kinases are essential enzymes in the control of the cell cycle. For instance, AURA is critical to the regulation of cancer progression, and its mutations and deregulations are associated with several cancers (43). CDK9 reactivates epigenetically silenced genes in cancer, and inhibition of CDK9 by drugs such as flavopiridol, dinaciclib, seliciclib, SNS-032, and RGB-286638 is exploited in cancer therapy (44, 45). MOK is closely

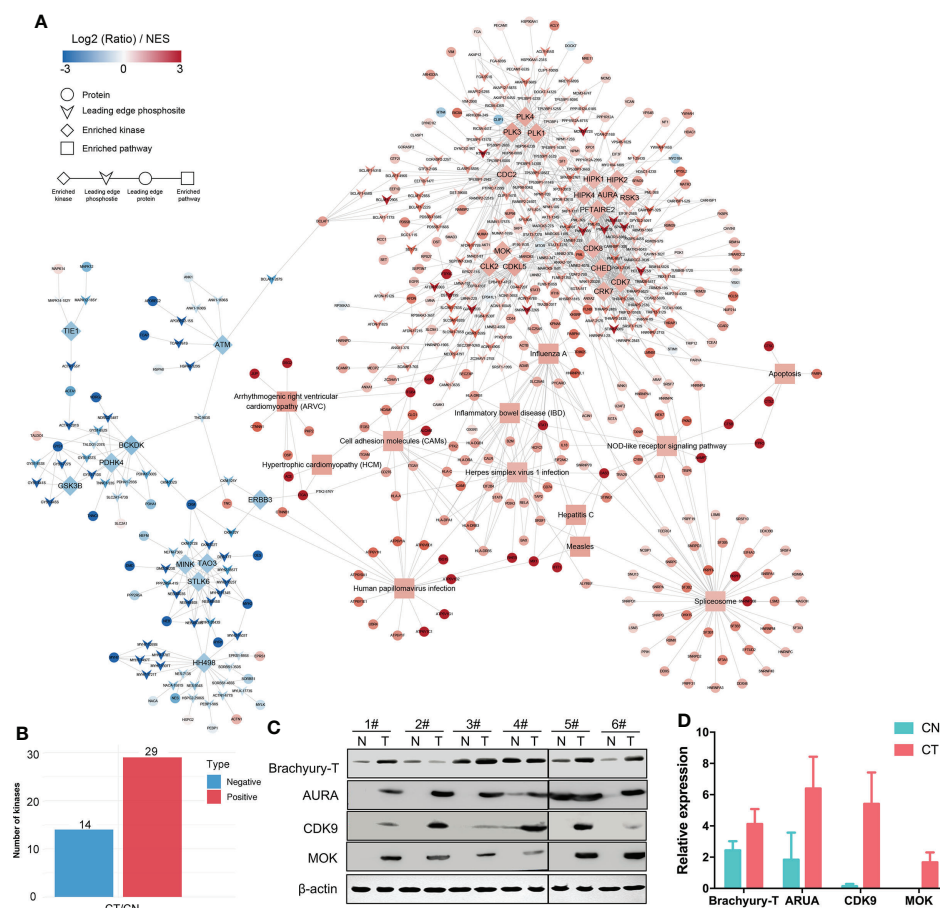


FIGURE 4

KiPNA analysis and the expression validation of specific proteins. (A) Kinase-pathway network analysis (KiPNA) indicating the proteomic and phosphoproteomic signaling network in chordoma. (B) Fourteen and 29 kinases were conjectured to be negatively and positively regulated, respectively. (C) The elevated expression of brachyury-T, AURA, MOK, and CDK9 were verified by western immunoblotting analysis. The samples were derived from six independent patients who were distinct from the patients whose samples were used for mass spectrometry. T, tumor tissues; N, normal tissues. (D) Densitometric quantification of western blot results from (B) and presented relative to β-actin expression. The data represent the mean ± SD of each experiment carried out in triplicate.

related to serine/threonine protein kinases in the protein kinome, and the expression of MOK is augmented in various tumors (46). Considering these results, we analyzed the roles of these kinases in chordoma.

We first confirmed that the specific marker of chordoma, brachyury (gene symbol: *TBXT*), was over-expressed in chordoma tissues when compared with normal tissues in six different cases (Figure 4C, Table 1). We then compared and validated the elevated expression of these kinases in CT to CN [although a few special cases were inconsistent, most likely due to individual heterogeneity (Figures 4C, D)]. Second, to evaluate the roles of AURA, CDK9, and MOK in chordoma cell proliferation, we silenced AURA, CDK9, and MOK protein expression in chordoma cell lines using siRNAs, and demonstrated that knockdown efficacy to be over 70% in JHC7 cells (Figures 5A–F). Despite the slow growth of chordoma cells, JHC7 proliferation rate after knockdown of AURA, CDK9, and MOK was nevertheless significantly suppressed relative to control cells as measured by CCK8 assay

(Figures 5G–I). We observed similar results with U-CH1, another chordoma cell line (Figures 5J–L). These results collectively indicate that AURA, CDK9, and MOK promote chordoma oncogenesis and could therefore be promoted as potential therapeutic targets in the treatment of chordoma.

Inhibition of CDK9 reduces tumor growth in chordoma

Since numerous kinase inhibitors have been discovered and developed, we assessed whether their inhibition would also lead to consequences analogous to those with siRNA transfection. Aurora family members are crucial to faithful mitotic transition, and their specific inhibitors include Barasertib (AZD1152) and Alisertib (MLN8237) (47). There also exists an abundance of CDK9 inhibitors, from the first-generation inhibitors Alvocidib (flavopiridol) and Seliciclib (roscovitine/CYC202) to the more specific and more potent

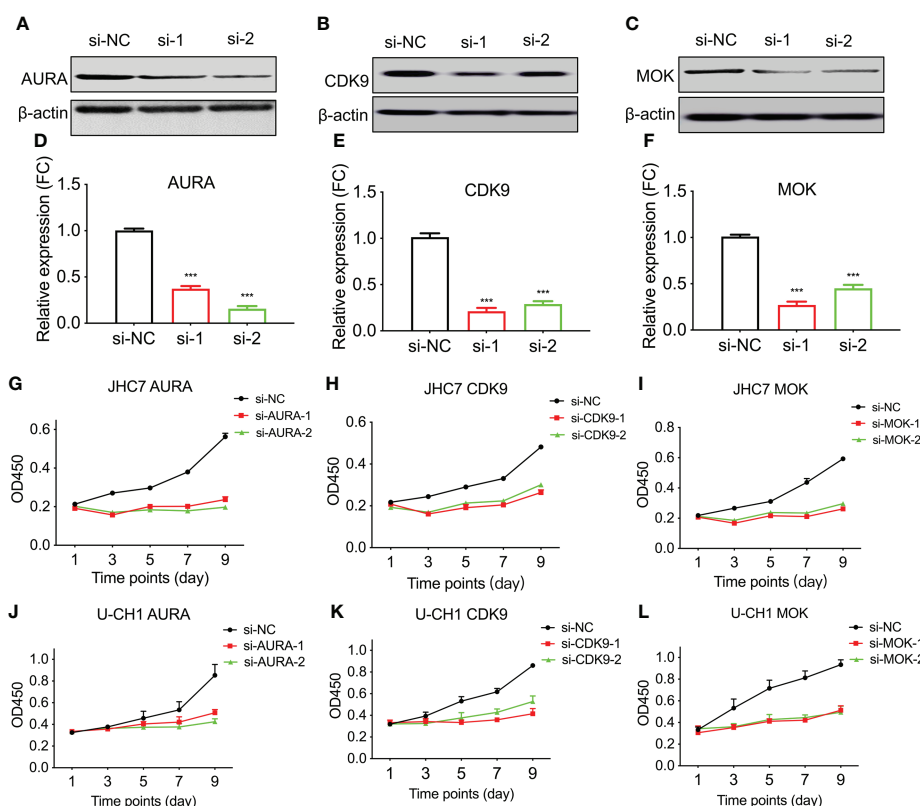


FIGURE 5

Knockdown of AURA, MOK, and CDK9 compromises chordoma cell line proliferation. The human chordoma cell lines JHC7 and U-CH1 were used to evaluate the proliferative effects of kinases. (A–C) The efficiency of siRNA-knockdown transfection was confirmed by western blotting of whole-cell extracts from JHC7 cells for AURA (A), CDK9 (B), and MOK (C). si-NC, negative control; si-1 and si-2, two independent siRNAs. (D–F) Densitometric quantification of western blot results from (A–C). Student's t-test was used. (G–I) Cellular proliferation after siRNA transfection was quantified by CCK8 assays. Cell growth of JHC7 (G–I) and U-CH1 cell lines (J–L) was significantly attenuated by knockdown of AURA, CDK9, and MOK. The data were acquired from three biologically and independently repeated experiments ($n=3$, data are mean \pm s.e.m.) ***: $P < 0.001$.

second-generation inhibitors as represented by BAY1143572 (atavaciclib) and AZD4573 (45). There are, however, far fewer data with respect to inhibitors of MOK. Considering their inhibitory potency and selectivity, we chose CDK9 and AZD4573 for further validation.

CDK9 is a functional subunit of P-TEFb (positive transcription elongation factor) and regulates transcriptional pause-release during gene transcriptional elongation (48). AZD4573 was optimized as a highly selective CDK9 inhibitor with an IC₅₀ of less than 4 nM and exhibited a greater than 10-fold selectivity for CDK9 over CDKs 1–7 (Figures 6A, B); it is currently regarded as a clinical candidate for the treatment of hematologic malignancies (49, 50). We treated chordoma cells with various concentrations (using 10-fold dilutions) of AZD4573 and found that, commensurate with increasing concentrations of AZD4573, the inhibitory effects on cellular proliferation were also gradually augmented (Figures 6C, D). These results indicate that functional inhibition of a key kinase suppresses chordoma cell proliferation and may thus provide a therapeutic option for the treatment of chordoma.

Chordoma tumor size is correlated with cytoskeletal anomalies

Due to the slow proliferation rate of chordomas and their adjacency to the spine, chordoma tumor size is directly related to patient survival rate (51). We regarded tumor size to be large if its maximal diameter was greater than 5.0 cm, and small if under 5.0 cm, and thereby noted four large (L, n=4) and five small (S, n=5) chordomas (Table 1). Unsupervised hierarchical clustering of these samples indicated differential protein expression patterns (Figure 7A), while Toll- and Imd-signaling pathways and ubiquitin-mediated proteolysis were significant functional KEGG enrichments with respect to both DEPs and DPPs (Figure 7B). Only two proteins were regulated at both the expression and phosphorylation levels (Figure 7C). To ascertain which proteins were uniformly altered in the CT/CN and L/S groups, we compared these two datasets and uncovered seven proteins in which phosphorylation levels were augmented in both (Figure 7D). We hypothesize that these dually changed proteins are plausible in promoting chordoma cell proliferation and tumor growth.

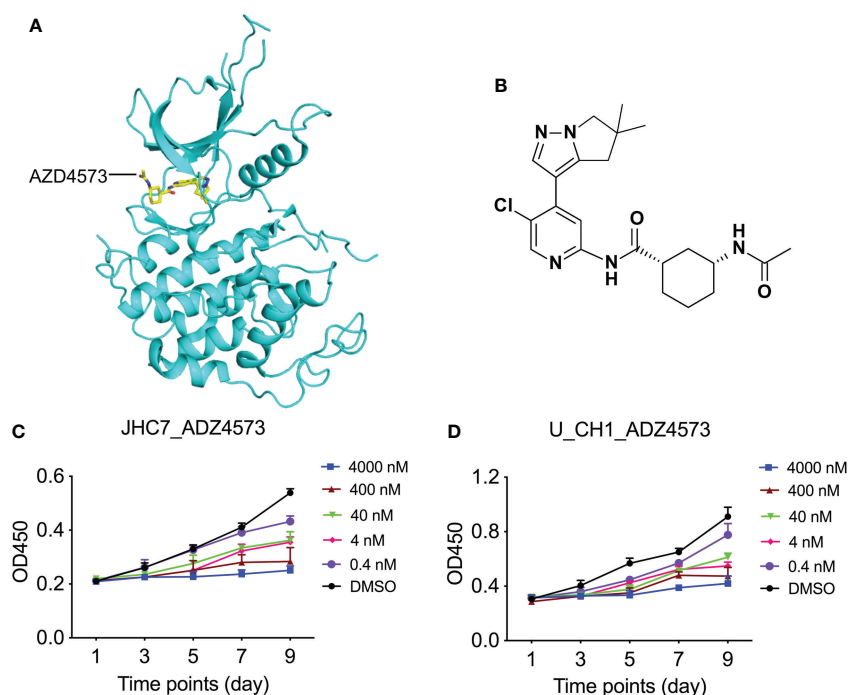


FIGURE 6
Dose-dependent treatment of AZD4573 results in cellular proliferation in chordoma cell lines. (A) Structural cartoon representation of CDK9 protein (pdb ID: 6Z45); AZD4573 is depicted as yellow sticks. (B) Chemical structure of AZD4573. (C, D) AZD4573 drug-sensitivity assays. JHC7 (C) and U-CH1 (D) cells were treated with graded concentrations of AZD4573 (from 0.4 nM to 4000 nM) and proliferation was subsequently determined with the CCK-8 cell-proliferation assay.

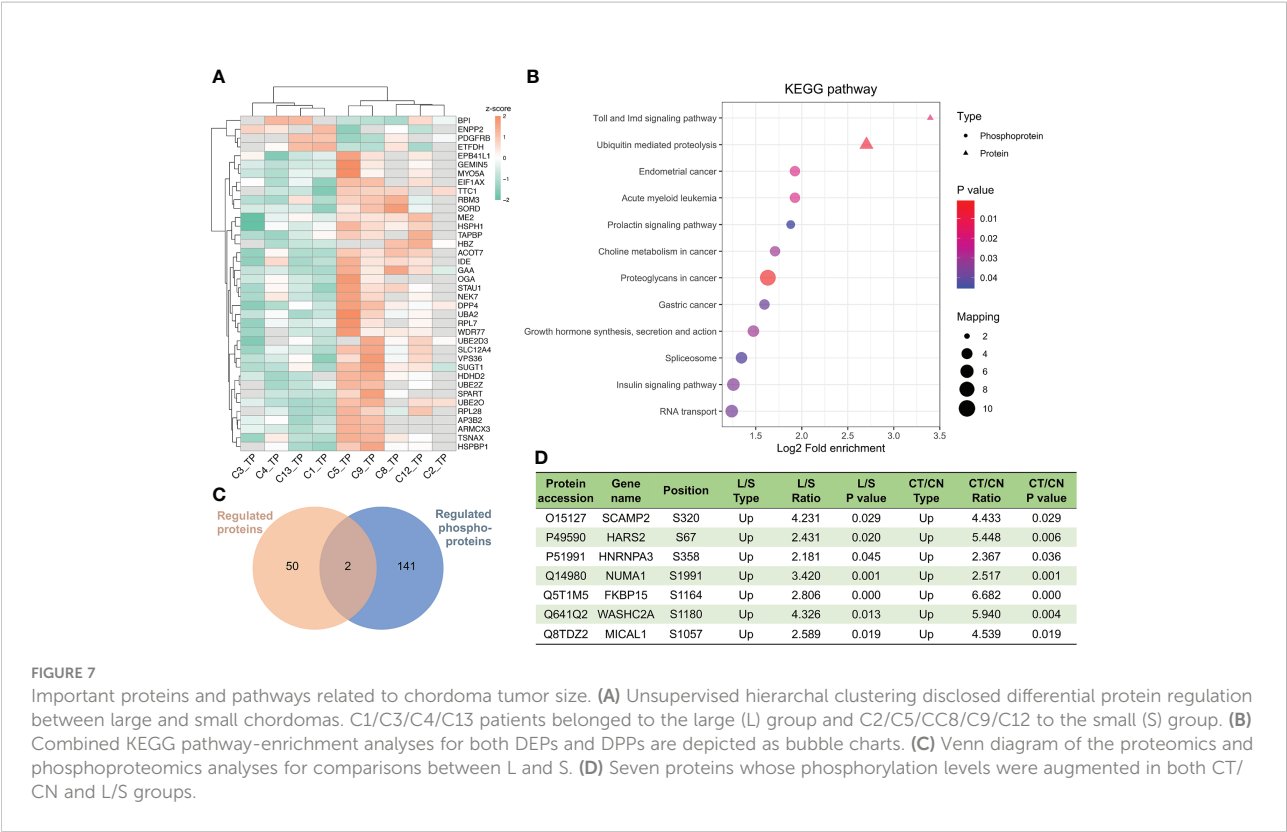


FIGURE 7 Important proteins and pathways related to chordoma tumor size. **(A)** Unsupervised hierarchal clustering disclosed differential protein regulation between large and small chordomas. C1/C3/C4/C13 patients belonged to the large (L) group and C2/C5/CC8/C9/C12 to the small (S) group. **(B)** Combined KEGG pathway-enrichment analyses for both DEPs and DPPs are depicted as bubble charts. **(C)** Venn diagram of the proteomics and phosphoproteomics analyses for comparisons between L and S. **(D)** Seven proteins whose phosphorylation levels were augmented in both CT/ CN and L/S groups.

Discussion

Chordomas are dual epithelial-mesenchymal tissue tumors that show slow progression, but despite their low malignancy, chordomas always manifest a high recurrence rate and frequently lead to local invasion and distant metastasis in advanced stages (52, 53). Chordoma is resistant to conventional chemotherapy and radiotherapy and surgical eradication remains challenging due to the complicated anatomical location of the tumors; patients are also vulnerable to relapse. It is therefore of the utmost necessity to discern the detailed molecular underpinnings of chordoma progression, and the exploration of novel therapeutic targets would also be of great value for chordoma patients.

Several investigators have assessed the genomic stability, epigenetic aberrations, genetic variations, gene transcription, microRNA expression profiles (23–28), and proteomic profiles of chordomas (32, 54). However, ours is the first-ever phosphoproteomics analysis of chordoma. We herein illuminated the PTM landscape together with protein expression by executing high-throughput, 4D, label-free proteomic and phosphoproteomic analyses, and then combined the bioinformatic results with cellular and biochemical verification so as to elucidate potential phosphorylative regulatory mechanisms in chordomas. The ideal control would be normal non-tumoral notochordal tissues, but as normal notochord is small and limited to the embryonic stage, sufficient material needed for experimentation is too difficult to

procure; thus, as for numerous other studies (32, 39, 40), we utilized normal adjacent muscle tissues as a substitute control as both tissues types originate from mesenchymal cells.

From nine paired tissues, we identified 5,089 proteins and 4,724 phosphorylated proteins. The key regulators of Wnt signaling are critical to embryonic development, and we noted marked alterations in CT tissues relative to CN tissues (Figure 2). Using intersection analysis, we identified 208 proteins that were dually regulated, and observed alterations in multiple pathways at both proteomic and phosphoproteomic levels (Figure 3). Moreover, we uncovered 28 kinases as possible therapeutic targets through KiPNA, of which AURA, CDK9, and MOK were validated both *in vivo* and *in vitro* (Figures 4 and 5), and their inhibition by small molecules proved their potential as drug targets (Figure 6). Considered the tumor size, we determined that secretory carrier-associated membrane protein 2 (SCAMP2) and the other seven proteins were upregulated in both the CT/ CN and Large/Small groups (Figure 7). These data provided a basis for the aforementioned phosphorylated proteins being related to malignancy *via* the promotion of tumor-cell amplification.

Small molecules that specifically target brachyury-T have been screened (55), and several kinase inhibitors have also been evaluated as to their suppression of chordoma tumor cell growth (29, 30). However, these few relevant publications were limited to assessments of phosphorylation and specific kinases in

chordomas, while our datasets provided more suitable and more widely applicable kinase candidates for future research.

Although we acknowledge limitations to the present study, such as the relatively small sample size ($n=9$ for our proteomics analysis) due to the rarity of chordomas and the lack of *in vivo* validation models such as xenografts, our findings still reveal useful information regarding chordomas. Our proteomic and phosphoproteomic data provide much-needed options for additional studies with respect to their validation and the application and establishment of more efficient therapies for chordoma treatment.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/[Supplementary Material](#).

Ethics statement

The studies involving human participants were reviewed and approved by the ethical review board of the Peking University Third Hospital [M2019233]. The patients/participants provided their written informed consent to participate in this study.

Author contributions

This work was performed by all authors. JH, HO and FW performed the majority of experiments. HO and QZ executed the experiments. JH and WY performed the majority of data and statistical analysis. ZL and LJ conceived and designed experiments. JH, HO and WY wrote and edited the manuscript. ZL, LJ and WY directed the study. FW collected chordoma tissues and clinical patient information. All authors contributed to the article and approved the submitted version.

References

1. Pan Y, Lu L, Chen J, Zhong Y, Dai Z. Analysis of prognostic factors for survival in patients with primary spinal chordoma using the SEER registry from 1973 to 2014. *J Orthop Surg Res* (2018) 13:76. doi: 10.1186/s13018-018-0784-3
2. Pennington Z, Ehresman J, McCarthy EF, Ahmed AK, Pittman PD, Lubelski D, et al. Chordoma of the sacrum and mobile spine: a narrative review. *Spine J* (2021) 21:500–17. doi: 10.1016/j.spinee.2020.10.009
3. Salisbury JR. The pathology of the human notochord. *J Pathol* (1993) 171:253–5. doi: 10.1002/path.1711710404
4. Vujovic S, Henderson S, Presneau N, Odell E, Jacques TS, Tirabosco R, et al. Brachyury, a crucial regulator of notochordal development, is a novel biomarker for chordomas. *J Pathol* (2006) 209:157–65. doi: 10.1002/path.1969
5. Pillay N, Plagnol V, Tarpey PS, Lobo SB, Presneau N, Suzhai K, et al. A common single-nucleotide variant in T is strongly associated with chordoma. *Nat Genet* (2012) 44:1185–7. doi: 10.1038/ng.2419
6. Walcott BP, Nahed BV, Mohyeldin A, Coumans JV, Kahle KT, Ferreira MJ. Chordoma: current concepts, management, and future directions. *Lancet Oncol* (2012) 13:e69–76. doi: 10.1016/S1470-2045(11)70337-0
7. Yuan W, Wei F, Ouyang H, Ren X, Hang J, Mo X, et al. CMTM3 suppresses chordoma progress through EGFR/STAT3 regulated EMT and TP53 signaling pathway. *Cancer Cell Int* (2021) 21:510. doi: 10.1186/s12935-021-02159-5
8. Lohberger B, Scheipl S, Heitzer E, Quehenberger F, de Jong D, Suzhai K, et al. Higher cMET dependence of sacral compared to clival chordoma cells: contributing to a better understanding of cMET in chordoma. *Sci Rep* (2021) 11:12466. doi: 10.1038/s41598-021-92018-0

Funding

This work was supported by the National Natural Science Foundation of the P. R. of China (82071658), Beijing Natural Science Foundation (7204327), the Beijing Nova Program (Z201100006820010), Capital's Funds for Health Improvement and Research (2020-4-40916), the Key Clinical Program of Peking University Third Hospital (BYSY2017002), and Clinical Medicine Plus X - Young Scholars Project, Peking University, the Fundamental Research Funds for the Central Universities (PKU2021LCXQ005).

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/fonc.2022.941046/full#supplementary-material>

9. Wang L, Tang L, Xu R, Ma J, Tian K, Liu Y, et al. DEPDC1B regulates the progression of human chordoma through UBE2T-mediated ubiquitination of BIRC5. *Cell Death Dis* (2021) 12:753. doi: 10.1038/s41419-021-04026-7
10. Barber SM, Sadrameli SS, Lee JJ, Fridley JS, Teh BS, Oyelese AA, et al. Chordoma-current understanding and modern treatment paradigms. *J Clin Med* (2021) 10:1054. doi: 10.3390/jcm10051054
11. Bhullar KS, Lagaron NO, McGowan EM, Parmar I, Jha A, Hubbard BP, et al. Kinase-targeted cancer therapies: progress, challenges and future directions. *Mol Cancer* (2018) 17:48. doi: 10.1186/s12943-018-0804-2
12. Frezza AM, Botta L, Trama A, Dei Tos AP, Stacchiotti S. Chordoma: update on disease, epidemiology, biology and medical therapies. *Curr Opin Oncol* (2019) 31:114–20. doi: 10.1097/CCO.0000000000000502
13. Scheipl S, Barnard M, Cottone L, Jorgensen M, Drewry DH, Zuercher WJ, et al. EGFR inhibitors identified as a potential treatment for chordoma in a focused compound screen. *J Pathol* (2016) 239:320–34. doi: 10.1002/path.4729
14. Hindi N, Casali PG, Morosi C, Messina A, Palassini E, Pilotti S, et al. Imatinib in advanced chordoma: A retrospective case series analysis. *Eur J Cancer* (2015) 51:2609–14. doi: 10.1016/j.ejca.2015.07.038
15. Macaulay VM, Middleton MR, Eckhardt SG, Rudin CM, Juergens RA, Gedrich R, et al. Phase I dose-escalation study of linsitinib (OSI-906) and erlotinib in patients with advanced solid tumors. *Clin Cancer Res* (2016) 22:2897–907. doi: 10.1158/1078-0432.CCR-15-2218
16. Stacchiotti S, Tamborini E, Lo Vullo S, Bozzi F, Messina A, Morosi C, et al. Phase II study on lapatinib in advanced EGFR-positive chordoma. *Ann Oncol* (2013) 24:1931–6. doi: 10.1093/annonc/mdt117
17. Siu IM, Ruzevick J, Zhao Q, Connis N, Jiao Y, Bettgeowda C, et al. Erlotinib inhibits growth of a patient-derived chordoma xenograft. *PLoS One* (2013) 8:e78895. doi: 10.1371/journal.pone.0078895
18. Hindi N, Casali PG, Morosi C, Messina A, Palassini E, Pilotti S, et al. Imatinib in advanced chordoma: A retrospective case series analysis. *Eur J Cancer* (2015) 51:2609–14. doi: 10.1016/j.ejca.2015.07.038
19. Stacchiotti S, Longhi A, Ferraresi V, Grignani G, Comandone A, Stupp R, et al. Phase II study of imatinib in advanced chordoma. *J Clin Oncol* (2012) 30:914–20. doi: 10.1200/JCO.2011.35.3656
20. Magnaghi P, Salom B, Cozzi L, Amboldi N, Ballinari D, Tamborini E, et al. Afatinib is a new therapeutic approach in chordoma with a unique ability to target EGFR and brachyury. *Mol Cancer Ther* (2018) 17:603–13. doi: 10.1158/1535-7163.MCT-17-0324
21. Zhao T, Siu IM, Williamson T, Zhang H, Ji C, Burger PC, et al. AZD8055 enhances *in vivo* efficacy of afatinib in chordomas. *J Pathol* (2021) 255:72–83. doi: 10.1002/path.5739
22. Anderson E, Havener TM, Zorn KM, Foil DH, Lane TR, Capuzzi SJ, et al. Synergistic drug combinations and machine learning for drug repurposing in chordoman. *Sci Rep* (2020) 10:12982. doi: 10.1038/s41598-020-70026-w
23. Scheil-Bertram S, Kappler R, von Baer A, Hartwig E, Sarkar M, Serra M, et al. Molecular profiling of chordoma. *Int J Oncol* (2014) 44:1041–55. doi: 10.3892/ijo.2014.2268
24. Alholle A, Brini AT, Bauer J, Gharanei S, Niada S, Slater A, et al. Genome-wide DNA methylation profiling of recurrent and non-recurrent chordomas. *Epigenetics* (2015) 10:213–20. doi: 10.1080/15592294.2015.1006497
25. Duan Z, Choy E, Nielsen GP, Rosenberg A, Iafraite J, Yang C, et al. Differential expression of microRNA (miRNA) in chordoma reveals a role for miRNA-1 in met expression. *J Orthop Res* (2010) 28:746–52. doi: 10.1002/jor.21055
26. Tarpey PS, Behjati S, Young MD, Martincorena I, Alexandrov LB, Farndon SJ, et al. The driver landscape of sporadic chordoma. *Nat Commun* (2017) 8:890. doi: 10.1038/s41467-017-01026-0
27. Liang WS, Dardis C, Helland A, Sekar S, Adkins J, Cuyugan L, et al. Identification of therapeutic targets in chordoma through comprehensive genomic and transcriptomic analyses. *Cold Spring Harb Mol Case Stud* (2018) 4:a003418. doi: 10.1101/mcs.a003418
28. Bai J, Shi J, Li C, Wang S, Zhang T, Hua X, et al. Whole genome sequencing of skull-base chordoma reveals genomic alterations associated with recurrence and chordoma-specific survival. *Nat Commun* (2021) 12:757. doi: 10.1038/s41467-021-21026-5
29. Thanindratar P, Dean DC, Nelson SD, Hornicek FJ, Duan Z. T-LAK cell-originated protein kinase (TOPK) is a novel prognostic and therapeutic target in chordoma. *Cell Prolif* (2020) 53:e12901. doi: 10.1111/cpr.12901
30. Hao S, Song H, Zhang W, Seldomridge A, Jung J, Giles AJ, et al. Protein phosphatase 2A inhibition enhances radiation sensitivity and reduces tumor growth in chordoma. *Neuro Oncol* (2018) 20:799–809. doi: 10.1093/neuonc/nox241
31. Chen S, Xu W, Jiao J, Jiang D, Liu J, Chen T, et al. Differential proteomic profiling of primary and recurrent chordomas. *Oncol Rep* (2015) 33:2207–18. doi: 10.3892/or.2015.3818
32. Wu Z, Wang L, Guo Z, Wang K, Zhang Y, Tian K, et al. Experimental study on differences in clivus chordoma bone invasion: an iTRAQ-based quantitative proteomic analysis. *PLoS One* (2015) 10:e0119523. doi: 10.1371/journal.pone.0119523
33. Shen Y, Li M, Xiong Y, Gui S, Bai J, Zhang Y, et al. Proteomics analysis identified ASNS as a novel biomarker for predicting recurrence of skull base chordoma. *Front Oncol* (2021) 11:698497. doi: 10.3389/fonc.2021.698497
34. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* (2005) 102:15545–50. doi: 10.1073/pnas.0506580102
35. Song C, Ye M, Liu Z, Cheng H, Jiang X, Han G, et al. Systematic analysis of protein phosphorylation networks from phosphoproteomic data. *Mol Cell Proteomics* (2012) 11:1070–83. doi: 10.1074/mcp.M111.012625
36. Krug K, Mertins P, Zhang B, Hornbeck P, Raju R, Ahmad R, et al. A curated resource for phosphosite-specific signature analysis. *Mol Cell Proteomics* (2019) 18:576–93. doi: 10.1074/mcp.TIR118.000943
37. Brubaker DK, Paulo JA, Sheth S, Poulin EJ, Popow O, Joughin BA, et al. Proteogenomic network analysis of context-specific KRAS signaling in mouse-to-Human cross-species translation. *Cell Syst* (2019) 9:258–270 e6. doi: 10.1016/j.cels.2019.07.006
38. Zhang C, Qu L, Lian S, Meng L, Min L, Liu J, et al. PRL-3 promotes ubiquitination and degradation of AURKA and colorectal cancer progression via dephosphorylation of FZR1. *Cancer Res* (2019) 79:928–40. doi: 10.1158/0008-5472.CAN-18-0520
39. Garofalo F, di Summa PG, Christoforidis D, Pracht M, Laudato P, Cherix S, et al. Multidisciplinary approach of lumbo-sacral chordoma: From oncological treatment to reconstructive surgery. *J Surg Oncol* (2015) 112:544–54. doi: 10.1002/jso.24026
40. Zhou H, Chen CB, Lan J, Liu C, Liu XG, Jiang L, et al. Differential proteomic profiling of chordomas and analysis of prognostic factors. *J Surg Oncol* (2010) 102:720–7. doi: 10.1002/jso.21674
41. Morin PJ. Beta-catenin signaling and cancer. *Bioessays* (1999) 21:1021–30. doi: 10.1002/(SICI)1521-1878(199912)22:1<1021::AID-BIES6>3.0.CO;2-P
42. Gaertner TR, Kolodziej SJ, Wang D, Kobayashi R, Koomen JM, Stoops JK, et al. Comparative analyses of the three-dimensional structures and enzymatic properties of alpha, beta, gamma and delta isoforms of Ca²⁺-calmodulin-dependent protein kinase II. *J Biol Chem* (2004) 279:12484–94. doi: 10.1074/jbc.M313597200
43. Shah KN, Bhatt R, Rotow J, Rohrberg J, Olivas V, Wang VE, et al. Aurora kinase A drives the evolution of resistance to third-generation EGFR inhibitors in lung cancer. *Nat Med* (2019) 25:111–8. doi: 10.1038/s41591-018-0264-7
44. Zhang H, Pandey S, Travers M, Sun H, Morton G, Madzo J, et al. Targeting CDK9 reactivates epigenetically silenced genes in cancer. *Cell* (2018) 175:1244–1258 e26. doi: 10.1016/j.cell.2018.09.051
45. Anshabo AT, Milne R, Wang S, Albrecht H. CDK9: A comprehensive review of its biology, and its role as a potential target for anti-cancer agents. *Front Oncol* (2021) 11:678559. doi: 10.3389/fonc.2021.678559
46. Chen T, Wu D, Moskaluk CA, Fu Z. Distinct expression patterns of ICK/MAK/MOK protein kinases in the intestine implicate functional diversity. *PLoS One* (2013) 8:e79359. doi: 10.1371/journal.pone.0079359
47. Bavetsias V, Linardopoulos S. Aurora kinase inhibitors: Current status and outlook. *Front Oncol* (2015) 5:278. doi: 10.3389/fonc.2015.00278
48. O'Brien T, Hardin S, Greenleaf A, Lis JT. Phosphorylation of RNA polymerase II c-terminal domain and transcriptional elongation. *Nature* (1994) 370:75–7. doi: 10.1038/370075a0
49. Barlaam B, Casella R, Cidado J, Cook C, De Savi C, Dishington A, et al. Discovery of AZD4573, a potent and selective inhibitor of CDK9 that enables short duration of target engagement for the treatment of hematological malignancies. *J Med Chem* (2020) 63:15564–90. doi: 10.1021/acs.jmedchem.0c01754
50. Cidado J, Boiko S, Proia T, Ferguson D, Criscione SW, San Martin M, et al. AZD4573 is a highly selective CDK9 inhibitor that suppresses MCL-1 and induces apoptosis in hematologic cancer cells. *Clin Cancer Res* (2020) 26:922–34. doi: 10.1158/1078-0432.CCR-19-1853
51. Xu JC, Lechrich BM, Yasaka TM, Fong BM, Hsu FPK, Kuan EC. Characteristics and overall survival in pediatric versus adult skull base chordoma: a population-based study. *Childs Nerv Syst* (2021) 37:1901–8. doi: 10.1007/s00381-021-05046-6
52. Catton C, O'Sullivan B, Bell R, Laperriere N, Cummings B, Fornasier V, et al. Chordoma: long-term follow-up after radical photon irradiation. *Radiother Oncol* (1996) 41:67–72. doi: 10.1016/S0167-8140(96)91805-8
53. Cannizzaro D, Tropeano MP, Milani D, Spaggiari R, Zaed I, Mancarella C, et al. Microsurgical versus endoscopic trans-sphenoidal approaches for clivus chordoma: a pooled and meta-analysis. *Neurosurg Rev* (2021) 44:1217–25. doi: 10.1007/s10143-020-01318-y

54. Davies JM, Robinson AE, Cowdrey C, Mummaneni PV, Ducker GS, Shokat KM, et al. Generation of a patient-derived chordoma xenograft and characterization of the phosphoproteome in a recurrent chordoma. *J Neurosurg* (2014) 120:331–6. doi: 10.3171/2013.10.JNS13598

55. Sharifnia T, Wawer MJ, Chen T, Huang QY, Weir BA, Sizemore A, et al. Small-molecule targeting of brachyury transcription factor addition in chordoma. *Nat Med* (2019) 25:292–300. doi: 10.1038/s41591-018-0312-3



OPEN ACCESS

EDITED BY

Liang Cheng,
Harbin Medical University, China

REVIEWED BY

Ping Dong,
Shanghai Jiao Tong University, China
Jun Ding,
Zhejiang University, China

*CORRESPONDENCE

Mengjun Hu
hu-menjun@163.com
Wandong Hong
xhmk-hwd@163.com
Xiangjian Chen
wz1370@126.com

[†]These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 30 June 2022

ACCEPTED 14 September 2022

PUBLISHED 03 October 2022

CITATION

Zhang Y, Wang Z, Basharat Z, Hu M,
Hong W and Chen X (2022)
Nomogram of intra-abdominal
infection after surgery in patients with
gastric cancer: A retrospective study.
Front. Oncol. 12:982807.
doi: 10.3389/fonc.2022.982807

COPYRIGHT

© 2022 Zhang, Wang, Basharat, Hu,
Hong and Chen. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Nomogram of intra-abdominal infection after surgery in patients with gastric cancer: A retrospective study

Yue Zhang^{1,2†}, Zhengfei Wang^{3†}, Zarrin Basharat⁴,
Mengjun Hu^{5*}, Wandong Hong^{6*} and Xiangjian Chen^{2*}

¹Department of Otolaryngology, Wenzhou People's Hospital, Wenzhou, China, ²Department of Gastrointestinal Surgery, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China, ³Department of Hepato-biliary Surgery, The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou, China, ⁴Jamil-ur-Rahman Center for Genome Research, Dr. Panjwani Center for Molecular Medicine and Drug Research, International Center for Chemical and Biological Sciences, University of Karachi, Karachi, Pakistan, ⁵Department of Pathology, Zhuji Affiliated Hospital of Wenzhou Medical University, Shaoxing, China, ⁶Department of Gastroenterology and Hepatology, the First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China

Background: Surgical resection is still the primary way to treat gastric cancer. Therefore, postoperative complications such as IAI (intra-abdominal infection) are major problems that front-line clinical workers should pay special attention to. This article was to build and validate IAI's RF (regression function) model. Furthermore, it analyzed the prognosis in patients with IAI after surgery for stomach cancer. The above two points are our advantages, which were not involved in previous studies.

Methods: The data of this study was divided into two parts, the training data set and the validation data set. The training data for this article were from the patients treated surgically with gastric cancer in our center from December 2015 to February 2017. We examined IAI's morbidity, etiological characteristics, and prognosis in the training data set. Univariate and multivariate logistic regression analyses were used to screen risk factors, establish an RF model and create a nomogram. Data from January to March 2021 were used to validate the accuracy of the RF model.

Results: The incidence of IAI was 7.2%. The independent risk factors for IAI were hypertension (Odds Ratio [OR] = 3.408, P = 0.001), history of abdominal surgery (OR = 2.609, P = 0.041), combined organ excision (OR = 4.123, P = 0.010), and operation time ≥ 240 min (OR = 3.091, P = 0.005). In the training data set and validation data set, the area under the ROC curve of IAI predicted by the RF model was 0.745 ± 0.048 (P < 0.001) and 0.736 ± 0.069 (P = 0.003), respectively. In addition, IAI significantly extended the length of hospital stay but had little impact on survival.

Conclusions: Patients with hypertension, combined organ excision, a history of abdominal surgery, and a surgical duration of 240 min or more are prone to IAI, and the RF model may help to identify them.

KEYWORDS

gastric cancer, surgery, postoperative complications, intra-abdominal infection, receiver operating characteristic curve, nomogram

Introduction

Gastric cancer is one of the most prevalent malignancies worldwide. According to the 2020 global cancer data (1), gastric cancer ranks fifth and fourth in morbidity and mortality, respectively. With diagnostic techniques such as endoscopy, the detection ratio of non-advanced gastric cancer is increasing, especially in Japan and South Korea. However, in China, there is no nationwide screening for gastric cancer (2). Only a small percentage of patients with early stomach cancer could receive treatment with ESD (Endoscopic Submucosal Dissection) or EMR (Endoscopic mucosal resection) (3). The remaining patients with advanced stage were treated with subtotal gastrectomy/total gastrectomy and lymph node dissection. Despite significant advances in surgical and postoperative care techniques for gastric cancer, severe postoperative complications can still occur at a high rate and affect the prognosis of patients (4–6). Therefore, determining how to reduce the occurrence of IAI is critical. The analysis of risk factors and the establishment of prediction models have been widely used in clinical disease research. Eun Hye Kim et al. developed a valid predictive model that can be used to determine the patients who will receive non-curative ESD resection (7). Screening the risk factors of IAI after gastric cancer surgery and establishing a prediction model can help clinicians take targeted measures to prevent the occurrence of IAI. Many scholars have studied surgical site infections, including their incidence, risk factors, prognosis, etc. However, most were superficial incision infections. Research on deep infections is not comprehensive enough, such as modeling and validation. Our innovation lies in the inclusion of more risk factors that were not included in previous studies, such as PNI (Prognostic nutritional index), neoadjuvant chemotherapy, lesion location, surgical method, and pathological type, to establish an RF (regression function) model and verify its predictive value by internal validation. In addition to the risk factor analysis of IAI occurrence and the establishment of the RF model, this paper also verified the RF model, which was never seen in previous studies. Through multivariate logistic regression analysis, we can obtain this RF

model, which can help us predict the probability of IAI for each patient. Finally, we also studied the impact of IAI on prognosis. This article added risk factors such as preoperative adjuvant chemotherapy. Since most previous authors have studied its relationship with overall postoperative complications (8, 9), this article aims to explore its relationship with a single complication, intraperitoneal infection, so this factor is included.

Materials and methods

Patients

The paper collected the data from 520 gastric cancer patients who were admitted to the gastrointestinal surgery department of our hospital for surgery from December 2015 to February 2017. The inclusion criteria of this study were patients who were surgically treated with gastric cancer in our department, aged > 18 years old, and without organ dysfunction. The exclusion criteria had emergency surgery, postoperative pathologic indication of non-primary gastric cancer, extensive peritoneal metastasis without surgical treatment, and preoperative intra-abdominal infection. According to the pathological report, 27 out of 520 cases were classified as pathological inconsistencies, including 5 having a neuroendocrine tumor, 2 suffering from lymphoma, 8 with chronic inflammatory changes such as chronic ulcers, 11 having intra-epithelial neoplasia, 1 with remnant gastric cancer, and 21 cases excluded due to surgical inconsistencies. As is displayed in Figure 1, 472, patients who met the criteria were finally included in this study. Among the 472 patients, 413 underwent radical gastrectomy with D2 lymph node dissection, and 59 underwent palliative resection. In addition, 101 patients underwent laparoscopic-assisted radical gastrectomy. Based on the same inclusion and exclusion criteria, 135 patients were selected from January to March 2021 to validate the prediction model. Each tumor was pathologically diagnosed and staged according to the 8th edition of the AJCC (American Joint Committee on Cancer) TNM classification system of Gastric Cancer (10).

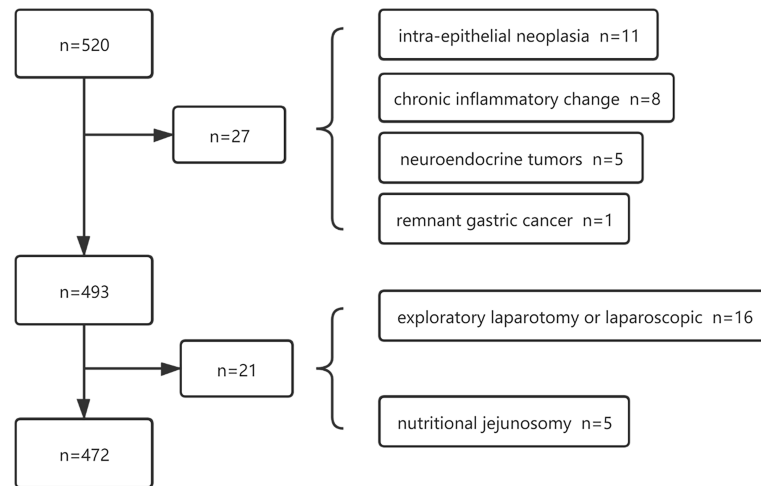


FIGURE 1

A total of 520 patients were initially screened and 472 patients were identified of the training data.

Surgical procedures and perioperative management

The patients included in this article were treated by our experienced gastrointestinal surgery team. The scope of lymph node dissection and the mode of gastrointestinal reconstruction was determined according to the fourth edition of the Japanese gastric cancer treatment guidelines (11). According to the guidelines, patients in cT1a or cT1b without lymph nodes or distant metastases should undergo D1 or D1+ lymph node dissection. Standard D2 or D2+ lymphadenectomy is feasible for patients with the following requirements: T2-4 or N+. When a patient was preoperatively diagnosed as M1, it was decided whether to perform combined organ resection and enlarged lymph node radical resection to achieve the R0 resection standard. If radical surgery was impossible, palliative resection or gastrointestinal short-circuit surgery was performed to relieve the suffering of patients and improve their future quality of life. Preoperative patients were given routine fasting for 8h and intestinal cleaning. A postoperative drainage tube was routinely placed in the sub-hepatic and splenic fossa. Perioperative treatment with cephalosporins was routinely used to prevent infection until 3–5 days after surgery. If no drainage fluid is found or the drainage fluid is relatively clear for 2–3 days, and the patient has no discomfort such as abdominal pain/fever, the drainage tube may be removed. The antibiotic was changed based on bacterial susceptibility testing or clinical experience if a patient was diagnosed with IAI. The treatment of IAI included routine surgical monitoring and nursing, simple rubber tube drainage, double cannula flushing and drainage, analgesic, antipyretic, anti-infection treatment, maintenance of

water & electrolyte balance, nutritional support, and surgery (12).

Clinical and surgical outcomes

The following variables were obtained from the patient's medical records at our hospital: Sex, age, BMI (Body Mass Index), chronic diseases (diabetes, hypertension), ASA (American Society of Anesthesiologists) score, preoperative chemotherapy history, earlier abdominal surgery, the existence of anemia/hypo-albuminemia, presence of hyperlipemia, site of primary carcinoma (clinical stage), time of operation, operation method, combined organ excision, PNI (Prognostic nutritional index), and BTF (perioperative blood transfusion history). The following formula counted PNI: $10 \times \text{serum albumin value (g/dL)} + 0.005 \times \text{total lymphocyte count in the peripheral blood (per mm}^3\text{)}$ (13). BTF is a transfusion of red blood cells during hospitalization (14). The percentage of deaths occurring within 30 days of surgery is known as postoperative mortality. This study considered complications in IAI patients diagnosed with Clavien-Dindo grade II.

Definition of IAI

From postoperative hospitalization to the post-discharge outpatient follow-up period, physicians closely monitored the occurrence and progression of IAI in patients. According to the findings during the second operation, clinical symptoms, temperature $\geq 38^\circ\text{C}$ (15), abdominal signs such as tenderness

and rebound pain, laboratory tests such as leukocyte, CRP (C-reactive protein), and PCT (Procalcitonin) (15), culture results of drainage fluid, and abdominal CT (computerized tomography) were performed to check whether the patient had an intra-abdominal infection (16). IAI can be divided into two categories according to whether it is caused by intestinal leakage. The first category includes anastomotic and duodenal stump leakage, and the second category is abdominal effusion accompanied by infection without intestinal leakage. The anastomotic fistula was confirmed by endoscopy, laboratory examination, radiological examination (17), or secondary surgical exploration (18).

Statistical analysis

Statistical analyses were performed using the IBM SPSS Statistics 25, R statistical software, and the Graphpad Prism 8 for Windows OS. The results of continuous variables were presented as the mean and standard deviation, and the categorical variables were presented as frequencies. The differences between groups for continuous variables were compared using an independent sample T-test, while the differences between groups for categorical variables were compared by the Chi-square test or Fisher's exact test. The ROC (receiver operating characteristic) curve of each variable was drawn by SPSS data processing software, and the Jorden index was calculated to determine the critical value of each variable. The training data set used univariate and multivariate logistic regression analyses to screen risk factors and establish an RF model. In the univariate analysis, variables with P -value < 0.1 were included in the multivariate logistic regression analysis (Forward: LR). $P < 0.05$ were considered statistically significant in multivariate logistic regression analysis, and a nomogram was created using the R statistical software (R Foundation for Statistical Computing, Vienna, Austria). The score obtained can be converted into the probability of IAI occurrence prediction by substituting the data from the validation set into the equation obtained by binary logistic regression analysis. The ROC curve was applied to calculate the accuracy of the nomogram to predict the diagnosis of IAI. The model's validity was measured using the AUROC (area under the receiver operating characteristic curve). A model with an AUROC above 0.7 was considered useful in diagnostic accuracy (19). The GraphPad Prism (Version 8) was used to describe the survival curves of the two groups.

Results

Incidence and clinical outcomes

The baseline characteristics of the training data set and validation data set were displayed in Tables 1 and 2. In 42

patients of the training data set who underwent surgical treatment (1. Radical gastrectomy with D2 lymph node dissection, 2. Palliative gastric cancer surgery) with primary gastric cancer, 34 (7.2%) patients suffered from intra-abdominal infection, including 15 (44.1%) cases of anastomotic leakage or duodenal stump leakage, 19 (55.8%) cases of peritoneal effusion with infection. As demonstrated in Tables 1 and 2, the length of hospital stay in IAI patients was significantly longer in terms of short-term prognosis. Studies have revealed that intra-abdominal abscess is one of the common causes of readmission (20), severely affecting patients' prognosis and quality of life. According to the Extended Clavien-Dindo classification of intra-abdominal infection (21), 1 case (2.9%) reached II, 30 cases (88.2%) reached IIIa, 3 cases (8.9%) reached IIIb stage. Fortunately, none of the patients had multiple organ failures or died from an intra-abdominal infection. Under the careful management of doctors and nurses in the treatment group, all the patients were improved and discharged after sufficient drainage, antibiotics, and other symptomatic support treatment (22).

The enrolled patients were contacted by phone to obtain and analyze their prognosis, with the most recent follow-up in May 2019. However, 25% of the patients were lost to follow-up and were excluded from the analysis. Finally, survival analysis was described in 353 patients with radical D1 or D2 lymphadenectomy. None of these patients died within 30 days after surgery. As is displayed in Figure 2, a significant difference did not appear in OS (overall survival) in the two groups ($P = 0.64$). A study by Ru-Hong Tu et al. also demonstrated that intra-abdominal infection after therapeutic gastrectomy did not lead to reduced long-term survival in patients (23). Furthermore, neither overall nor major surgical complications were risk factors for decreased survival in patients who did not die from early postoperative complications within 30 days of surgery (24). That was also consistent with our research results.

Pathogens

The abdominal drainage fluid of 34 patients diagnosed with IAI was cultured, and 19 (55.8%) were positive. The collection of abdominal drainage fluid follows the Sterile principle. Among the 19 patients with positive culture results, 4 had mixed growth of more than three strains (the possibility of specimen contamination could not be ruled out), 5 had mixed growth of two strains, and 10 had single strains. There were 6 gram-negative strains (46.2%), 6 gram-positive strains (46.2%), and 1 *Candida* spp. (7.6%). The most common microorganism was *Streptococcus anginosus*, *Enterococcus faecalis*, and *Klebsiella pneumoniae*. *Streptococcus anginosus* is one of the common colonized bacteria of the oropharynx, which can migrate to the digestive tract and become a pathogenic bacteria of postoperative intra-abdominal infection (25, 26). Previously, Xiao et al. (27) reported the presence of gram-negative bacilli

TABLE 1 Clinicopathological characteristics of the patients in training data (n = 472).

Variable	IAI (n = 34)	Non-IAI (n = 438)	χ^2 or <i>t</i> -value	<i>P</i> -value
Sex (Male: Female)	25:9	325:113	0.007	0.931
Age (years) #	66.26 ± 9.665	64.32 ± 11.109	0.991	0.322
BMI (kg/m ²)#	23.19 ± 3.07	22.29 ± 3.05	1.648	0.100
Preoperative white blood cell count (×10 ⁹ /L) #	6.50 ± 1.95	6.14 ± 2.23	-0.891	0.373
Preoperative lymphocyte count (×10 ⁹ /L) #	1.55 ± 0.56	1.61 ± 0.59	-0.60	0.551
Preoperative hemoglobin (g/L) #	111.38 ± 21.68	117.39 ± 25.15	-1.354	0.177
Preoperative albumin (g/L) #	36.20 ± 4.84	38.00 ± 4.80	-2.104	0.036
PNI	43.95 ± 6.20	46.12 ± 6.29	-1.933	0.054
ASA (1 + 2/3+4)	30:4	425:13	4.726	0.03
Diabetes mellitus (yes/no)	3:31	47:391	0.003	0.953
Hypertension (yes/no)	19:15	119:319	12.573	< 0.001
History of abdominal surgery (yes/no)	8:26	45:393	4.311	0.038
Neoadjuvant chemotherapy (yes/no)	2:32	17:421	0.014	0.905
Lesion location (limited/diffuse)	33:1	435:3	0.000	1.000
Upper	5	64		
Middle	7	62		
lower	21	309		
other	1	3		
Time of operation (min)	205.91 ± 58.12	188.99 ± 46.81	1.993	0.047
Operation type				
Radical surgery:	28:6	385:53	0.453	0.501
non-radical surgery				
Operation method				
Open : Laparoscopic-assisted	25:9	347:91	0.613	0.434
Combined organ excision (yes/no)	6:28	20:418	8.011	0.005
BTF (yes/no)	9:25	73:365	2.113	0.146
Pathological type				
Signet-ring cell carcinoma:	3:31	89:349	2.657	0.103
Non-signet ring cell carcinoma				
Tumor stage (I+II/III+IV)	11:23	215:223	3.540	0.060
I	5	119		
II	6	96		
III	15	149		
IV	8	74		
Post-operative hospital stays (days) #	27.06 ± 14.043	14.30 ± 6.392	5.257	< 0.001

ASA, American Society of Anesthesiologist; BMI, Body Mass Index; #mean ± SD, IAI, intra-abdominal infection.

in 73/1835, i.e., 4% of patients undergoing gastrectomy for gastric cancer. *Klebsiella pneumonia* was commonly linked to body mass index >25 kg/m².

Due to the significant difference in pH (Pondus Hydrogenii) between the oral cavity and the stomach, it is generally believed that the bacteria do not remain active during the migration process in the digestive tract. Therefore, there are two possibilities: first is that for gastric cancer patients, a measure of perioperative management is using proton pump inhibitors, which can reduce the pH of the gastric mucosa. The second is that the most common site of gastric cancer is the antrum, so the site secreting more gastric acid was just removed during the operation.

Risk factors

According to the univariate analysis of this data (Table 3), the IAI would occur easier in patients with a BMI ≥ 25 kg/m², ASA score ≥ 3, history of abdominal surgery, hypertension, combined organ excision, and operative time ≥ 240 min. The tumor stage (III + IV) was a potential risk factor. Diabetes, radical surgery, or laparoscopic-assisted surgery execution, in addition to the pathological type and tumor stage, were not considered risk factors for the occurrence of IAI. The multivariate analysis demonstrated that hypertension (Odds Ratio [OR] = 3.408, 95% confidence interval [CI]: 1.632–7.117, *P* = 0.001), operation time ≥

TABLE 2 Clinicopathological characteristics of the patients in internal validation data (n = 135).

Variable	IAI (n = 15)	Non-IAI (n = 120)	χ^2 or <i>t</i> -value	<i>P</i> -value
Sex (Male: Female)	11:4	76:44	0.582	0.446
Age (years) #	69.40 ± 8.175	64.76 ± 10.971	-1.582	0.116
BMI (kg/m2)#	22.42 ± 3.08	23.06 ± 2.83	0.810	0.419
Preoperative white blood cell count (×10 ⁹ /L) #	6.46 ± 2.12	5.98 ± 1.85	-0.917	0.361
Preoperative lymphocyte count (×10 ⁹ /L) #	1.41 ± 0.55	1.57 ± 0.53	1.089	0.278
Preoperative hemoglobin (g/L) #	117.40 ± 15.33	116.73 ± 21.93	-0.114	0.909
Preoperative albumin (g/L) #	36.11 ± 2.92	36.74 ± 3.98	0.588	0.557
PNI	43.19 ± 4.34	44.60 ± 5.17	1.016	0.311
ASA (1 + 2/3+4)	14:1	103:17	0.162	0.687
Diabetes mellitus (yes/no)	1:14	23:97	0.698	0.403
Hypertension (yes/no)	5:10	26:94	0.472	0.492
History of abdominal surgery (yes/no)	6:9	14:106	6.385	0.012
Neoadjuvant chemotherapy (yes/no)	2:13	6:113	0.489	0.485
Lesion location (limited/diffuse)	12:3	113:7	2.109	0.146
upper	5	14		
middle	3	31		
lower	4	68		
other	3	7		
Time of operation (min)	235.73 ± 47.35	218.36 ± 55.18	-1.166	0.246
Operation type				
Radical surgery:	12:3	109:11	0.720	0.396
non-radical surgery				
Operation method				
Open : Laparoscopic-assisted	8:7	30:90	3.984	0.046
Combined organ excision (yes/no)	2:13	3:117	1.876	0.171
BTF (yes/no)	4:11	16:104	0.970	0.325
Pathological type				
Signet-ring cell carcinoma:	1:14	25:95	0.930	0.335
Non-signet ring cell carcinoma				
Tumor stage (I+II/III+IV)	6:9	80:40	4.101	0.043
I	3	54		
II	3	26		
III	6	36		
IV	3	4		
Post-operative hospital stays (days) #	16.07 ± 7.94	11.79 ± 4.53	-3.124	0.002

ASA, American Society of Anesthesiologist; BMI, Body Mass Index; #mean ± SD, IAI, intra-abdominal infection.

240 min (OR = 3.091, 95% CI: 1.408–6.783, *P* = 0.005), the history of abdominal surgery (OR = 2.609, 95% CI: 1.042–6.530, *P* = 0.041), and combined organ resection (OR = 4.123, 95% CI: 1.403–12.121, *P* = 0.01) were independent risk factors for IAI (Table 4).

Regression function model and validation

According to the analysis results in Table 4, the RF model for IAI could be obtained as follows: *estimated probability* =

$\frac{1}{1+EXP(-X)}$, $X = -3.63 + (1.226 \cdot hypertension) + (0.959 \cdot history\ of\ abdominal\ surgery) + (1.128 \cdot operation\ time \geq 240mins) + (1.417 \cdot combined\ organ\ excision)$. The ROC curve for the RF model based on the training data set for the prediction of IAI is demonstrated in Figure 3 (AUROC = 0.745 ± 0.048, *P* < 0.001, 95% CI: 0.650–0.840). Intuitively, the RF model was presented as a nomogram that could visualize the RF model (Figure 4). The ROC curve of the nomogram based on the validation data set for the prediction of IAI is displayed in Figure 5 (AUROC = 0.736 ± 0.069, *P* = 0.003, 95% CI: 0.602–0.871).

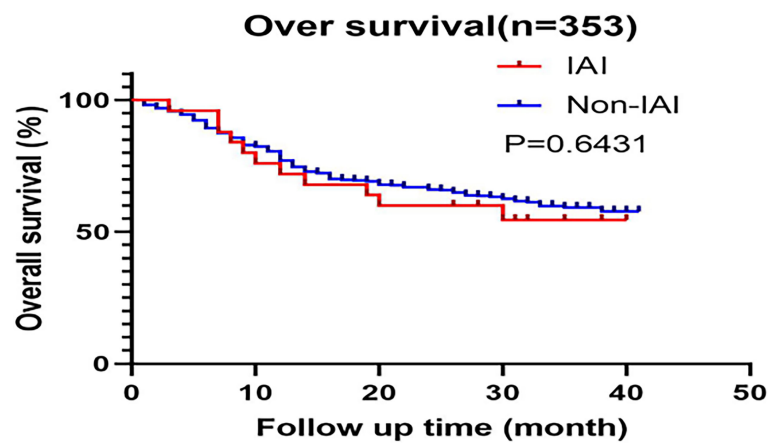


FIGURE 2

Kaplan-Meier curves of overall survival based on the training data.

TABLE 3 Univariate analysis of possible predictors of risk of IAI based on training data.

Variable	IAI (n = 34)	Non-IAI (n = 438)	Univariate analysis		
			OR	95%CI	P-value
Sex (Male: Female)	25:9	325:113	0.968	0.465-2.016	0.931
Age (years) ≥ 65 / <65	22:12	236:202	1.521	0.771-3.001	0.222
BMI (kg/m ²) ≥ 25 / <25	12:22	87:351	2.055	1.054-4.006	0.033
Preoperative white blood cell count ($\times 10^9$ /L) ≥ 4 / <4	33:1	398:40	3.139	0.441-22.364	0.358
Preoperative lymphocyte count ($\times 10^9$ /L) <0.8 / ≥ 0.8	33:1	21:417	1.806	0.594-5.490	0.532
Preoperative hemoglobin (g/L) <100 / ≥ 100	10:24	102:336	1.339	0.661-2.715	0.419
Preoperative albumin (g/L) <35 / ≥ 35	10:24	97:341	1.421	0.702-2.878	0.330
PNI <47 / ≥ 47	24:10	238:200	1.924	0.941-3.932	0.066
ASA (3 + 4/1+2)	4:30	13:425	3.569	1.416-8.992	0.03
Diabetes mellitus (yes/no)	3:31	47:391	0.817	0.259-2.575	0.953
Hypertension (yes/no)	19:15	119:319	3.066	1.605-5.856	0.000
History of abdominal surgery (yes/no)	8:26	45:393	2.433	1.162-5.094	0.038
Neoadjuvant chemotherapy (yes/no)	2:32	17:421	1.490	0.385-5.764	0.905
Time of operation (min) ≥ 240 / <240	13:21	64:374	3.176	1.663-6.066	<0.001
Operation type	28:6	385:53	0.667	0.288-1.542	0.501
Radical surgery: non-radical surgery					
Operation method	25:9	347:91	0.747	0.360-1.548	0.434
Open : Laparoscopic-assisted					
Combined organ excision (yes/no)	6:28	20:418	3.676	1.671-8.084	0.005
BTF (yes/no)	9:25	73:365	1.712	0.830-3.531	0.146
Pathological type	3:31	89:349	0.400	0.125-1.279	0.103
Signet-ring cell carcinoma: Non-signet ring cell carcinoma					
Tumor stage (I+II/III+IV)	11:23	215:223	1.921	0.958-3.851	0.060

TABLE 4 Multivariate analysis of risk factors of IAI based on internal validation data.

Risk factors	β coefficients	Standard error (SE)	Odds Ratio [OR]	95% Confidence Interval [CI]	P value
Intercept	-3.63	0.327			< 0.001
Hypertension	1.226	0.376	3.408	1.632-7.117	0.001
History of abdominal surgery	0.959	0.468	2.609	1.042-6.53	0.041
Operation time (min): ≥ 240	1.128	0.401	3.091	1.408-6.783	0.005
Combined organ excision	1.417	0.550	4.123	1.403-12.121	0.010

Discussion

Gastric cancer still has a relatively high incidence, and surgical resection is the primary treatment method. Therefore, postoperative complications are important problems that front-line clinical workers should pay special attention to. Abdominal infection is one of gastric cancer's most severe postoperative complications, resulting in significantly longer hospital stays, septic shock, multiple organ failures, and even death. In this single-center retrospective study, the incidence of postoperative abdominal infection for gastric cancer was 7.2%. A study by Felipe J.F.Coimbra MD (28) revealed that the overall incidence of postoperative complications of gastric cancer was 33.5%, among which the most common surgical complication was intra-abdominal abscess with an incidence of 7.9%, which was close to the data obtained in this retrospective study.

Chen Ke et al. have demonstrated that total laparoscopic gastrectomy has less bleeding, shorter hospitalization, and fewer

postoperative complications than open gastrectomy (29). However, Inokuchi, M et al.'s meta-analysis demonstrated an insignificant difference in the intra-abdominal abscesses between the laparoscopic-assisted distal gastrectomy group and the open distal gastrectomy group (30). The results of this study also indicated that laparoscopic-assisted gastrectomy did not reduce the incidence of intra-abdominal infection.

Studies on high BMI (≥ 25 kg/m²) as a potential risk factor have drawn different conclusions. The meta-analysis by Zhao et al. demonstrated that high BMI patients had a higher risk of wound infection and IAI in both open and laparoscopic-assisted gastrectomy (31). However, the analysis by Sun et al. revealed that although high BMI patients had a higher risk of wound infection than those with low BMI (< 25 kg/m²), there was an insignificant difference in the incidence of anastomotic fistula among them (32). Previous studies have concluded that low PNI (< 47) is an independent risk factor for postoperative complications in patients with gastric cancer and will affect

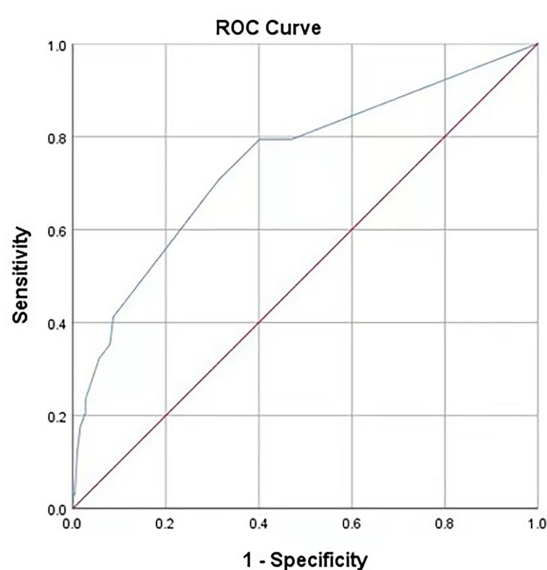


FIGURE 3

Receiver operating characteristic curves of the predictive model on the training data. AUC (95%CI) = 0.745 (0.650-0.840). The areas under receiver operating characteristic curves were 0.745 ± 0.048 ($P < 0.001$). The ideal area under the curve was 1.00. The reference line represents that based on chance alone (area under the curve 0.50).

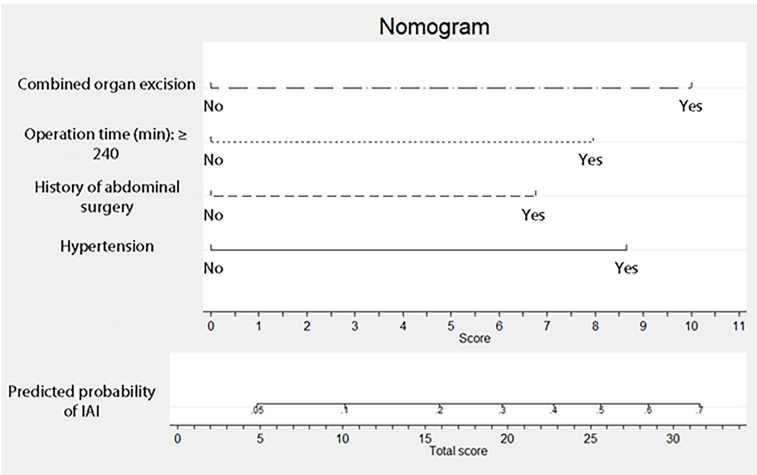


FIGURE 4
Nomogram for intra-abdominal infection after surgery for gastric cancer. To estimate the probability of intra-abdominal infection, mark patient values at each axis, draw a straight line perpendicular to the point axis, and sum the points for all variables. Next, mark the sum on the total point axis and draw a straight line perpendicular to the probability axis.

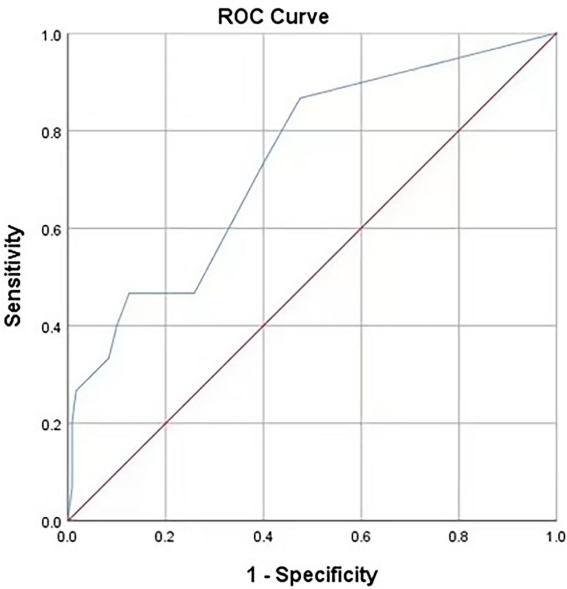


FIGURE 5
Receiver operating characteristic curves of the predictive model based on the internal validation data set. AUC (95%CI) = 0.736 (0.602-0.871), P=0.003.

long-term prognosis (33). Therefore, it may be more useful than BMI in predicting postoperative IAI for gastric cancer patients. Our data analysis suggests that the low PNI (P=0.066) group may be at greater risk of developing IAI as a postoperative complication. The differences in the results of these studies may be due to sampling error. In addition, regional climate and

dietary habits may make the BMI or PNI of a certain group generally higher or lower.

According to previous literature reports and clinicians' experience, diabetes patients are more likely to develop infectious complications. It may be because of the physiological mechanisms, including lipid metabolism

disorders, endothelial cell damage and dysfunction, abnormal platelet function, and blood vessel atherosclerosis, resulting in poor blood supply at the anastomotic and residual ends, thus increasing the risk of fistula (34). In addition, high blood pressure and diabetes often co-exist, causing damage to blood vessels together (35). Jönsson et al. (36) indicated that collagen synthesis depends on tissue oxygenation, thus demonstrating disturbed anastomotic healing in insufficient blood supply. This study found insignificant differences between the two groups, whether or not they had diabetes. Patients with hypertension, however, were at greater risk of developing IAI. It may be due to sampling error or bias.

Postoperative adhesions form in 50% to 100% of all abdominopelvic interventions (37). Due to the presence of more or less tissue adhesion in the abdominal cavity, patients with previous abdominal surgery must have separated adhesion next time. Then, the operation time will be prolonged.

Splenectomy and partial pancreas resection accounted for most of the combined organ resection. Spleen is the largest immune organ in the body, and its removal may affect the immune function of the human body. For example, splenectomy increases the risk of developing sepsis in response to *Streptococcus pneumoniae*, *Neisseria meningitidis*, and *Hemophilus influenza* type B infections (38–40). There is also an increased risk of pancreatic fistula associated with infection in patients with partial pancreatectomy.

Patients with combined organ resection and a history of abdominal surgery generally have longer surgery times and longer gastrointestinal opening times, which increases the risk of surgical site infection.

The area under the receiver operating characteristic (ROC) curve for the RF model based on the training data set was 0.745 ± 0.048 , and that of the nomogram based on the validation data set was 0.736 ± 0.069 . It revealed that this nomogram had good predictive power.

IAI is one of the common complications of abdominal surgery, which can be life-threatening to a certain extent and cannot be ignored. Therefore, it is urgent to thoroughly study the risk factors of abdominal infection and its influence on the prognosis to better guide clinical work. This retrospective analysis demonstrated that hypertension, combined organ resection, history of abdominal surgery, and operation time ≥ 240 min were independent risk factors that could increase the risk of postoperative intra-abdominal infection. Therefore, we should minimize unnecessary tissue damage to reduce the wound surface and the operation time. Stijn Blot et al. summarized the etiological characteristics of 1,982 patients with intra-abdominal infection. They found that most patients were infected with gram-negative bacteria, among which *Escherichia coli* in Enterobacteria was the most common, and *Enterococcus* sp. accounted for the most in gram-positive bacteria (41). Our data demonstrated that the top three pathogens were *Streptococcus anginosus*, *Klebsiella pneumoniae*, and *Enterococcus faecalis*. *Streptococcus anginosus* is a commonly colonized bacteria in the oral cavity. *Klebsiella pneumoniae* and *Enterococcus faecalis*

are common colonized bacteria in the intestinal tract. From the standpoint of pathogens, we must conduct well in perioperative oral management. Patients with severe periodontitis need to be treated by a stomatologist before surgery. Moreover, it is necessary to make good intestinal preparation before operation (42), strictly follow the principle of sterility during operation, apply a sufficient course of antibiotics after the operation, and ensure a good drainage effect.

Conclusions

IAI is allied with gastric cancer surgery complications, including pathogenic growth. Species such as *Klebsiella*, *Streptococcus*, and *Enterococcus* dominated the variety in this study. Independent risk factors impacting IAI included hypertension, combined organ resection, history of abdominal surgery, and operation time of more than 240 mins. Diabetes did not increase the chance of infection. Compared to conventional electrosurgery, the extent of operative time may be reduced with energy devices, such as ultrasonically activated coagulating shears. Since this study is a single-center retrospective study, there is a possibility that the samples taken do not conform to the general population. Besides, the selective and observational bias in the retrospective study are also limitations of this type of study. A larger sample size and patients from diverse areas could help reduce these limitations. In conclusion, gastric cancer patients with the risk factors above require more attention. This is the first study to establish an RF model of IAI and verify it, and the verified result shows that the RF model has a significant predictive ability for the occurrence of IAI after gastric cancer surgery.

Data availability statement

The data supporting the findings are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Author contributions

YZ, ZW, XC, and WH designed this study. YZ collected and analyzed the data, made tables and figures, and wrote the manuscript. ZB, MH, XC, and WH revised the manuscript. All authors contributed to the article and approved the submitted version.

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

References

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2021) 71(3):209–49. doi: 10.3322/caac.21660
- Yuan Y. A survey and evaluation of population-based screening for gastric cancer. *Cancer Biol Med* (2013) 10(2):72–80. doi: 10.7497/j.issn.2095-3941.2013.02.002
- Ono H, Yao K, Fujishiro M, Oda I, Nimura S, Yahagi N, et al. Guidelines for endoscopic submucosal dissection and endoscopic mucosal resection for early gastric cancer. *Dig Endosc*. (2016) 28(1):3–15. doi: 10.1111/den.12518
- Tokunaga M, Tanizawa Y, Bando E, Kawamura T, Terashima M. Poor survival rate in patients with postoperative intra-abdominal infectious complications following curative gastrectomy for gastric cancer. *Ann Surg Oncol* (2013) 20(5):1575–83. doi: 10.1245/s10434-012-2720-9
- Sano T, Sasako M, Yamamoto S, Nashimoto A, Kurita A, Hiratsuka M, et al. Gastric cancer surgery: morbidity and mortality results from a prospective randomized controlled trial comparing D2 and extended para-aortic lymphadenectomy–Japan clinical oncology group study 9501. *J Clin Oncol* (2004) 22(14):2767–73. doi: 10.1200/JCO.2004.10.184
- Papenfuss WA, Kukar M, Oxenberg J, Attwood K, Nurkin S, Malhotra U, et al. Morbidity and mortality associated with gastrectomy for gastric cancer. *Ann Surg Oncol* (2014) 21(9):3008–14. doi: 10.1245/s10434-014-3664-z
- Kim EH, Park JC, Song IJ, Kim YJ, Joh DH, Hahn KY, et al. Prediction model for non-curative resection of endoscopic submucosal dissection in patients with early gastric cancer. *Gastrointest Endosc* (2017) 85(5):976–83. doi: 10.1016/j.gie.2016.10.018
- Wu L, Ge L, Qin Y, Huang M, Chen J, Yang Y, et al. Postoperative morbidity and mortality after neoadjuvant chemotherapy versus upfront surgery for locally advanced gastric cancer: a propensity score matching analysis. *Cancer Manag Res* (2019) 11:6011–8. doi: 10.2147/CMAR.S203880
- Zhou Y, Tian Z, Zeng J, Zhou W, Wu K, Shen W. Effect of neoadjuvant treatment combined with radical gastrectomy on postoperative complications and prognosis of gastric cancer patients. *Scand J Gastroenterol* (2021) 56(11):1343–8. doi: 10.1080/00365521.2021.1966092
- Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The eighth edition AJCC cancer staging manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin* (2017) 67(2):93–9. doi: 10.3322/caac.21388
- Japanese Gastric Cancer Association/Japanese gastric cancer treatment guidelines 2014 (ver. 4). *Gastric Cancer* (2017) 20(1):1–19. doi: 10.1007/s10120-016-0622-4
- Sartelli M, Chichom-Mefire A, Labricciosa FM, Hardcastle T, Abu-Zidan FM, Adesunkanmi AK, et al. The management of intra-abdominal infections from a global perspective: 2017 WSES guidelines for management of intra-abdominal infections. *World J Emerg Surg* (2017) 12:29. doi: 10.1186/s13017-017-0141-6
- Onodera T, Goseki N, Kosaki G. [Prognostic nutritional index in gastrointestinal surgery of malnourished cancer patients]. *Nihon Geka Gakkai Zasshi* (1984) 85(9):1001–5.
- Xiao H, Quan H, Pan S, Yin B, Luo W, Huang G, et al. Impact of peri-operative blood transfusion on post-operative infections after radical gastrectomy for gastric cancer: a propensity score matching analysis focusing on the timing, amount of transfusion and role of leukocyte depletion. *J Cancer Res Clin Oncol* (2018) 144(6):1143–54. doi: 10.1007/s00432-018-2630-8
- Vincenzi B, Fioroni I, Pantano F, Angeletti S, Dicuonzo G, Zoccoli A, et al. Procalcitonin as diagnostic marker of infection in solid tumors patients with fever. *Sci Rep* (2016) 6:28090. doi: 10.1038/srep28090
- Emmi V, Sganga G. Clinical diagnosis of intra-abdominal infections. *J Chemother* (2009) 21 Suppl 1:12–8. doi: 10.1179/joc.2009.21.Supplement-1.12
- Haaga JR. Imaging intraabdominal abscesses and nonoperative drainage procedures. *World J Surg* (1990) 14(2):204–9. doi: 10.1007/BF01664874
- Facy O, Paquette B, Orry D, Binquet C, Masson D, Bouvier A, et al. Diagnostic accuracy of inflammatory markers as early predictors of infection after elective colorectal surgery: Results from the IMACORS study. *Ann Surg* (2016) 263(5):961–6. doi: 10.1097/SLA.0000000000001303
- Giannini EG, Zaman A, Kreil A, Floreani A, Dulbecco P, Testa E, et al. Platelet count/spleen diameter ratio for the noninvasive diagnosis of esophageal varices: results of a multicenter, prospective, validation study. *Am J Gastroenterol* (2006) 101(11):2511–9. doi: 10.1111/j.1572-0241.2006.00874.x
- Asaoka R, Kawamura T, Makuuchi R, Irino T, Tanizawa Y, Bando E, et al. Risk factors for 30-day hospital readmission after radical gastrectomy: a single-center retrospective study. *Gastric Cancer* (2019) 22(2):413–20. doi: 10.1007/s10120-018-0856-4
- Katayama H, Kurokawa Y, Nakamura K, Ito H, Kanemitsu Y, Masuda N, et al. Extended clavian-dindo classification of surgical complications: Japan clinical oncology group postoperative complications criteria. *Surg Today* (2016) 46(6):668–85. doi: 10.1007/s00595-015-1236-x
- Solomkin JS, Mazuski JE, Bradley JS, Rodvold KA, Goldstein EJ, Baron EJ, et al. Diagnosis and management of complicated intra-abdominal infection in adults and children: guidelines by the surgical infection society and the infectious diseases society of America. *Clin Infect Dis* (2010) 50(2):133–64. doi: 10.1086/649554
- Tu RH, Lin JX, Desiderio J, Li P, Xie JW, Wang JB, et al. Does intra-abdominal infection after curative gastrectomy affect patients' long-term prognosis? a multi-center study based on a Large sample size. *Surg Infect (Larchmt)* (2019) 20(4):271–7. doi: 10.1089/sur.2018.246
- Galata C, Blank S, Weiss C, Ronellenfitsch U, Reissfelder C, Hardt J. Role of postoperative complications in overall survival after radical resection for gastric cancer: A retrospective single-center analysis of 1107 patients. *Cancers (Basel)* (2019) 11(12):1980. doi: 10.3390/cancers11121890
- Sasaki M, Kodama Y, Shimoyama Y, Ishikawa T, Kimura S. Acidity and acid tolerance mechanisms of streptococcus anginosus. *J Gen Appl Microbiol* (2018) 64(4):174–9. doi: 10.2323/jgam.2017.11.005
- Nishikawa M, Honda M, Kimura R, Kobayashi A, Yamaguchi Y, Hori S, et al. The bacterial association with oral cavity and intra-abdominal abscess after gastrectomy. *PLoS One* (2020) 15(11):e0242091. doi: 10.1371/journal.pone.0242091
- Xiao H, Xiao Y, Quan H, Liu W, Pan S, Ouyang Y. Intra-abdominal infection after radical gastrectomy for gastric cancer: Incidence, pathogens, risk factors and outcomes. *Int J Surg* (2017) 48:195–200. doi: 10.1016/j.ijsu.2017.07.081
- Coimbra FJF, de Jesus VHF, Franco CP, Calsavara VF, Ribeiro HSC, Diniz AL, et al. Predicting overall and major postoperative morbidity in gastric cancer patients. *J Surg Oncol* (2019) 120(8):1371–8. doi: 10.1002/jso.25743
- Chen K, Pan Y, Cai JQ, Xu XW, Wu D, Mou YP. Totally laparoscopic gastrectomy for gastric cancer: a systematic review and meta-analysis of outcomes compared with open surgery. *World J Gastroenterol* (2014) 20(42):15867–78. doi: 10.3748/wjg.v20.i42.15867
- Inokuchi M, Sugita H, Otsuki S, Sato Y, Nakagawa M, Kojima K. Laparoscopic distal gastrectomy reduced surgical site infection as compared with open distal gastrectomy for gastric cancer in a meta-analysis of both randomized controlled and case-controlled studies. *Int J Surg* (2015) 15:61–7. doi: 10.1016/j.ijsu.2015.01.030
- Zhao B, Zhang J, Mei D, Luo R, Lu H, Xu H, et al. Does high body mass index negatively affect the surgical outcome and long-term survival of gastric cancer patients who underwent gastrectomy: A systematic review and meta-analysis. *Eur J Surg Oncol* (2018) 44(12):1971–81. doi: 10.1016/j.ejso.2018.09.007
- Sun L, Zhao B, Huang Y, Lu H, Luo R, Huang B. Feasibility of laparoscopy gastrectomy for gastric cancer in the patients with high body mass index: A systematic review and meta-analysis. *Asian J Surg* (2020) 43(1):69–77. doi: 10.1016/j.asjsur.2019.03.017
- Jiang N, Deng JY, Ding XW, Ke B, Liu N, Zhang RP, et al. Prognostic nutritional index predicts postoperative complications and long-term outcomes of gastric cancer. *World J Gastroenterol* (2014) 20(30):10537–44. doi: 10.3748/wjg.v20.i30.10537
- Zawada AE, Moszak M, Skrzypczak D, Grzymislawski M. Gastrointestinal complications in patients with diabetes mellitus. *Adv Clin Exp Med* (2018) 27(4):567–72. doi: 10.17219/acem/67961

35. Yamazaki D, Hitomi H, Nishiyama A. Hypertension with diabetes mellitus complications. *Hypertens Res* (2018) 41(3):147–56. doi: 10.1038/s41440-017-0008-y
36. Jönsson K, Jiborn H, Zederfeldt B. Breaking strength of small intestinal anastomoses. *Am J Surg* (1983) 145(6):800–3. doi: 10.1016/0002-9610(83)90144-7
37. diZerega GS. Contemporary adhesion prevention. *Fertil Steril* (1994) 61(2):219–35. doi: 10.1016/S0015-0282(16)56507-8
38. Amlot PL, Hayes AE. Impaired human antibody response to the thymus-independent antigen, DNP-ficoll, after splenectomy. implications for post-splenectomy infections. *Lancet* (1985) 1(8436):1008–11. doi: 10.1016/S0140-6736(85)91613-7
39. Ram S, Lewis LA, Rice PA. Infections of people with complement deficiencies and patients who have undergone splenectomy. *Clin Microbiol Rev* (2010) 23(4):740–80. doi: 10.1128/CMR.00048-09
40. Robinette CD, Fraumeni JF Jr. Splenectomy and subsequent mortality in veterans of the 1939–45 war. *Lancet* (1977) 2(8029):127–9. doi: 10.1016/S0140-6736(77)90132-5
41. Blot S, Antonelli M, Arvaniti K, Blot K, Creagh-Brown B, de Lange D, et al. Epidemiology of intra-abdominal infection and sepsis in critically ill patients: "AbSeS", a multinational observational cohort study and ESICM trials group project. *Intensive Care Med* (2019) 45(12):1703–17. doi: 10.1007/s00134-019-05819-3
42. Toh JWT, Phan K, Hitos K, Pathma-Nathan N, El-Khoury T, Richardson AJ, et al. Association of mechanical bowel preparation and oral antibiotics before elective colorectal surgery with surgical site infection: A network meta-analysis. *JAMA Netw Open* (2018) 1(6):e183226. doi: 10.1001/jamanetworkopen.2018.3226



OPEN ACCESS

EDITED BY

Liang Cheng,
Harbin Medical University, China

REVIEWED BY

Leila Mostafavi,
Massachusetts General Hospital and
Harvard Medical School, United States
Lei Cheng,
Nanjing Medical University, China
Liwen Wu,
Hunan Children's Hospital, China

*CORRESPONDENCE

Xiangrong Zheng
zxr_168@126.com
Yantong Zhu
zhytong1020@163.com

SPECIALTY SECTION

This article was submitted to
Cancer Genetics,
a section of the journal
Frontiers in Oncology

RECEIVED 06 July 2022

ACCEPTED 24 October 2022

PUBLISHED 06 December 2022

CITATION

Zhu Y and Zheng X (2022)
Microscopic polyangiitis presenting
with persistent cough and
hemoptysis in pediatrics: A case
report and review of the literature.
Front. Oncol. 12:987507.
doi: 10.3389/fonc.2022.987507

COPYRIGHT

© 2022 Zhu and Zheng. This is an
open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use,
distribution or reproduction is
permitted which does not comply with
these terms.

Microscopic polyangiitis presenting with persistent cough and hemoptysis in pediatrics: A case report and review of the literature

Yantong Zhu* and Xiangrong Zheng*

Department of Pediatrics, Xiangya Hospital, Central South University, Changsha, China

Background: Microscopic polyangiitis (MPA) is a necrotizing vasculitis that involves small- and medium-sized vessels and is associated with the presence of antineutrophil cytoplasmic antibodies with a perinuclear staining pattern (p-ANCA). The kidney and lungs are the organs primarily affected. MPA is rare in children and is easily misdiagnosed. Below is a complete case history of the course of the disease.

Case presentation: An 11-year-old girl with a 1-month history of cough and hemoptysis showed no improvement after imipenem-cilastatin treatment. p-ANCA and microscopic hematuria and proteinuria were positive, and a chest CT revealed an area of shadow in the bilateral lower lobe of the lungs. Renal biopsies showed crescentic glomerulonephritis, and MPA was diagnosed based on these criteria. The patient exhibited dramatic clinical and imaging improvements after immunosuppressive treatment.

Conclusion: The organs most commonly involved in MPA in children are the lungs, kidneys, skin, nervous system organs, and organs of the gastrointestinal tract. Careful examination should be carried out in these patients while biopsies of the kidney or any other organs remain the gold standard for diagnostic purposes. Pulmonary involvement may be the initial symptom of the disease and should not be confused with pneumonia. A urinalysis should be performed in patients with hemoptysis. Antibiotics should be used with caution.

KEYWORDS

microscopic polyangiitis, cough, hemoptysis, children, renal

Abbreviations: ANA, antinuclear antibodies; CRP, C-reactive protein; CT, computerized tomography; DAH, diffuse alveolar hemorrhage; ESR, erythrocyte sedimentation rate; ECMO, extra-corporeal mechanical oxygenation; p-ANCA, perinuclear staining pattern; PPD, purified protein derivative; MPA, microscopic polyangiitis.

Introduction

Antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis (AAV) is a multisystem autoimmune disease that primarily involves small- and medium-sized blood vessels throughout the body. Clinically, it is classified into three types: microscopic polyangiitis (MPA), granulomatosis with polyangiitis (GPA), and eosinophilic GPA (EGPA) (1). Microscopic polyangiitis (MPA) is a necrotizing vasculitis that involves small- and medium-sized vessels, primarily affecting the lungs and kidneys. An analysis of several retrospective adult cases showed that most patients typically present with renal involvement (RI, 80%–100%), cutaneous involvement (CI, 50%), pulmonary involvement (PI, 25%–55%), gastrointestinal involvement (GI, 30%–50%), and nervous system involvement (NSI, 28%) (2). Currently, there are no unified diagnostic criteria for AAV. ANCA positivity is a specific serological marker for AAV, and accurate ANCA testing is important for diagnosis. For example, MPA is associated with the presence of antineutrophil cytoplasmic antibodies with a perinuclear staining pattern (p-ANCA); however, detection of ANCAs is not in itself diagnostic of AAV, and a biopsy remains the gold standard for diagnosis, especially in cases with negative serology or unusual clinical presentation. However, the estimated collective incidence of pediatric vasculitis is about 0.05% (3), and it is even rarer in MPA. We herein describe a complete course of the disease and analyze the clinical features and lab examinations reported for pediatric MPA patients in the past 10 years.

Case description

A previously healthy 11-year-old girl (weight, 38.4 kg; height, 155 cm) presented to our department with a 1-month history of cough and hemoptysis. She had an intermittent fever but was without dyspnea, joint pain, headache, or other symptoms. She was initially treated in the local hospital where they suspected infection, and the symptoms improved after 7 days of anti-infection treatment administered on 10 April 2017. The symptoms appeared again 1 month later, and she visited our outpatient department on 10 May 2017. Her parents were healthy and nonconsanguineous. A hemogram reported a leucocyte count of $8.1 \times 10^9/L$ (neutrophils 52%), platelet count of $407 \times 10^9/L$, and hemoglobin count of 106 g/L. The chest X-ray revealed areas of abnormal density (symmetric distribution) in the bilateral lower lobes of the lungs (Figure 1A). The patient was diagnosed with lobar pneumonia because of the respiratory symptoms, fever, and X-ray evidence and was treated with parenteral imipenem-cilastatin but had no clinical improvement. Three days later, she was admitted to our hospital to undergo a detailed work-up for the cough and hemoptysis. Upon admission, she had a blood pressure of 96/

54 mmHg and a body temperature of 38.5°C. The physical examination on admission was not remarkable. Further examination revealed the following notable laboratory test findings, in chronological order: urinalysis revealed microscopic hematuria (+++) and proteinuria (++) on 14 May; renal function revealed BUN 4.85 mmol/L, Cr 81.6 $\mu\text{mol/L}$, UA 213.7 $\mu\text{mol/L}$, and C3 1,300 mg/L; inflammatory markers were slightly increased (ESR 29 mm/h and CRP 19 mg/L); ANA was present in a relatively low titer (1:80), without any antigen-specific antibodies; p-ANCA was positive and MPO-ANCA revealed 50.9 U/ml (normal <20 U/ml); PR3 and GBM were normal on 16 May; the chest CT performed on May 16 revealed an area of dense shadow in the bilateral lower lobes of the lungs, without ground-glass opacity and reticulation (Figure 1B), but pulmonary artery CT and PPD were normal. To ultimately establish a proper diagnosis, renal biopsies were performed as a result of the proteinuria and positive MPO and p-ANCA identified on 18 May; finally, the biopsies revealed glomerular cellular and microcellular crescents (Figures 1E–H). Immunofluorescence microscopy showed IgA (+) and IgG (++) deposits in the glomerular capillary loop but no C3 and C1q deposits. Based on the respiratory symptoms (cough and hemoptysis), positive MPO and p-ANCA, and the renal biopsies, our patient was finally diagnosed with microscopic polyangiitis (MPA), and so we started immunosuppressive treatment: methylprednisolone (800 mg/day*3 days, two cycles), plasma exchange (1,500 ml/day*4 days), and IV-cyclophosphamide (CTX 800 mg/day once), followed by oral formulation of methylprednisolone (48 mg/day) and tacrolimus (4 mg daily). Methylprednisolone pulse therapy and plasmapheresis resulted in dramatic clinical and imaging improvements (Figures 1C, D). The patient stopped the drugs after 2 months with proteinuria (++) and returned to our hospital.

We prescribed methylprednisolone (500 mg/day for 3 days) again, following which she received methylprednisolone (48 mg/day for a month, 24 mg/day for the next month, followed by a gradually reduced dose of 6 mg/day per month, stopping in February 2018) and tacrolimus (4 mg/day, which was gradually reduced (1 mg/day for 3 months) and halted in July 2018). Proteinuria turned negative in September 2017, and the patient continues to have no clinical symptoms, with proteinuria remaining negative. The clinical timeline is presented in Figure 1I.

Discussion and conclusions

AAV is a group of diseases caused by inflammation of the blood vessels. It has a complex pathogenesis. GPA and MPA are associated with a loss of immunological tolerance to PR3 or MPO, whereas EGPA pathogenesis involves two distinct

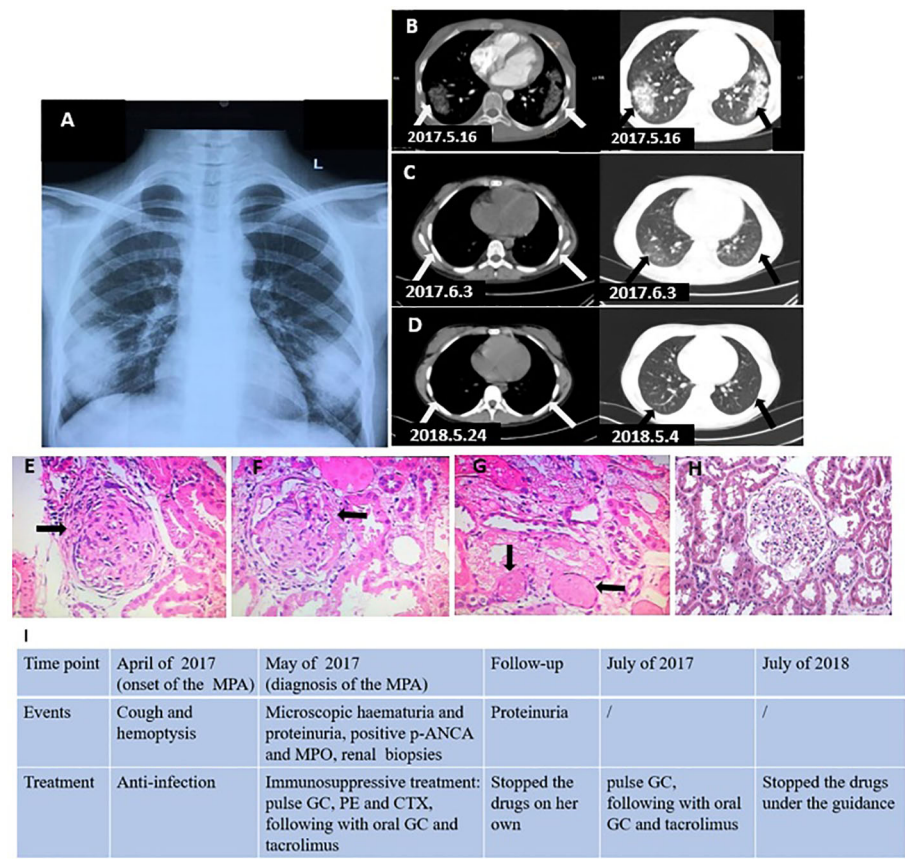


FIGURE 1 (A) X-ray revealed abnormal density areas (arrows) in the bilateral lower lobe of the lungs (symmetric distribution). (B) Chest CT revealed a large area of shadow (arrows) in the bilateral lower lobe of the lungs (before the treatment). (C) Chest CT revealed the shadows (arrows) in bilateral lungs improved (during the treatment). (D) Chest CT revealed the lesions disappeared (after the treatment). (E, F) Renal biopsies (HE x40) revealed glomerulonephritic fibrocellular crescents (arrows). (G) Renal biopsies revealed vascular occlusion (arrows). (H) Normal control glomerulus. (I) Clinical timeline. GC, glucocorticoid; CTX, cyclophosphamide; PE, plasma exchange.

mechanisms that are associated with ANCA-positive and ANCA-negative (4). ANCAs are considered to be central to vasculitis. MPO or PR3 on neutrophil plasma membranes combines with ANCAs that are produced by B cells, which promote the release of destructive proteases, oxygen radicals, as well as autoantigens. Subsequent tissue damage may result in organ dysfunction (4, 5).

Since MPA is rare in children, we initially diagnosed this case with cough and hemoptysis, intermittent fever, and areas of abnormal density in the lungs, such as lobar pneumonia, supported by the fact that the symptoms improved after the first anti-infection treatment. However, further treatment resulted in no improvement. The first anti-infection treatment may have been effective because the infection can cause or worsen MPA when the MPA is not severe. Since there was no further improvement, we then performed urinalysis (hematuria +++, proteinuria ++), MPO-ANCA (50.9 U/ml), and renal

biopsies (crescentic glomerulonephritis). Based on the results, MPA was considered. After immunosuppressive treatment, for remission-induction of new-onset organ-threatening or life-threatening MPA, we recommend treatment with a combination of glucocorticoids and cyclophosphamide. In addition, plasma exchange could be used for the treatment of severe diffuse alveolar hemorrhage. The patient's condition obviously improved, and proteinuria has remained negative. Given the paucity of clinical trials in pediatric ANCA-associated vasculitis, pediatric rheumatologists have relied on adult AAV evidence for management. The European League Against Rheumatism (EULAR) (6) suggests remission-induction of new-onset organ-threatening AAV and recommends a combination of glucocorticoids and either cyclophosphamide or rituximab, and in the case of severe diffuse alveolar hemorrhage, a plasma exchange could be considered for treatment.

Clinical discussion

Patients with MPA always manifest with multiple system symptoms, such as respiratory symptoms, gastrointestinal symptoms, renal involvement, and nerve system involvement. We recently published a summary of children's case reports in PubMed ([Supplementary Table](#)) that showed that the organs involved in MPA in children are the lungs (manifest as dyspnea, cough, and hemoptysis, about 80%), kidneys (manifest as hematuria, proteinuria, and even anuria, about 80%), skin (manifest as purpuric lesions, bullae/hemorrhagic bullae, about 20%), nervous system (manifest as abnormal eye movements, episodes of seizure, and even coma, about 16%), and gastrointestinal tract (manifest as abdominal pain, digestive tract hemorrhage, about 16%). These present as the initial symptoms and affect the prognosis and survival rate. In addition, anemia can also be an initial symptom. Pulmonary involvement, as the initial syndrome, should be carefully distinguished from pneumonia ([Supplementary Table](#), the overall usage rate of antibiotics is about 43%, in patients manifesting with pulmonary involvement as the initial syndrome, usage is about 71%), especially in patients who manifest with pulmonary symptoms only. Physical examination always relates to the affected organs, and we should perform further examination in the following order: urinalysis, CRP, and ESR; ANA, ANCA, and chest CT or X-ray; and renal biopsies. We could start immunosuppressive treatment after an MPA diagnosis. There are no unified criteria for the diagnosis of MPA. Granulomatosis polyangiitis (GPA), which is always present with granulomatous inflammation, should be excluded (7, 8). In addition, systemic lupus erythematosus (9), Sjögren syndrome (10), and other connective tissue diseases also involve multiple organs, such as the lungs and kidneys, and we need to distinguish these other systemic diseases from MPA.

Radiological discussion

The lung is the most commonly affected organ in MPA. Chest CTs show diffuse areas of ground-glass opacity (11), reticulation (12), patchy shadows (13), nodular thickening (14), interstitial pneumonia (15), emphysema (16), and pleural effusion (17). The lesions are always of a symmetric distribution, and all these changes could be alone or coexist in one patient. X-ray and CT findings are inconclusive; MPA can be limited to the lungs. Chest X-rays revealed lesions with an alveolar-filling pattern, which are most often bilateral (2). The imaging manifestations of MPA in the lungs can be easily misdiagnosed as infections or another disease, especially in patients with only lung involvement. Normal X-rays do not

mean the lungs are unaffected (18), and most patients will be required to have a chest CT. MPA can also involve the nervous system. Magnetic resonance imaging (MRI) of the brain revealed non-hemorrhagic multiple lesions (13), reversible posterior leukoencephalopathy syndrome (16), and even hemorrhagic stroke (19). We recommend that brain MRI be performed on patients diagnosed with, or suspected of having, MPA.

Pathological discussion

Biopsies of the kidney or any other organ remain the gold standard for diagnostic purposes (6). Generally, renal biopsies can reveal pathological changes such as necrotizing glomerulonephritis (18), cellular and microcellular crescent glomerulonephritis, and sclerosed glomerulonephritis (12). Immunofluorescence microscopy demonstrates pauci-immune glomerulonephritis, and slight IgA deposits have been reported (20). In addition to the confirmation of MPA diagnosis, a renal biopsy also can help the clinician to provide a renal prognosis of the disease.

In summary, MPA is an unusual disease in children involving multiple systems. The clinical symptoms vary, but the organs most involved in MPA in children are the lungs, kidneys, skin, nervous system, and gastrointestinal tract. A thorough examination should be carried out in these patients, with biopsies of the kidney or any other organs remaining the gold standard for diagnostic purposes. Pulmonary involvement should be carefully distinguished from pneumonia; urinalysis should be taken in those patients with hemoptysis; and finally, antibiotics should be used with caution.

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.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics committee of Xiangya Hospital, Central South University Approval Documents for Scientific Research Projects (No. 2018121104). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

Author contributions

All authors read and approved the final manuscript. The first draft of the manuscript was produced by YZ and XZ. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors thank the parents for approval for publication.

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/fonc.2022.987507/full#supplementary-material>

References

- Zhao WM, Wang ZJ, Shi R, Zhu YY, Zhang S, Wang RF, et al. Environmental factors influencing the risk of ANCA-associated vasculitis. *Front Immunol* (2022) 13:991256. doi: 10.3389/fimmu.2022.991256
- Villiger PM, Guillevin L. Microscopic polyangiitis: Clinical presentation. *Autoimmun Rev* (2010) 9(12):812–9. doi: 10.1016/j.autrev.2010.07.009
- Eleftheriou D, Batu ED, Ozen S, Brogan PA. Vasculitis in children. *Nephrol Dial Transplant* (2015) 30 Suppl 1:i94–103. doi: 10.1093/ndt/gfu393
- Kitching AR, Anders HJ, Basu N, Brouwer E, Gordon J, Jayne DR, et al. ANCA-associated vasculitis. *Nat Rev Dis Primers* (2020) 6(1):71. doi: 10.1038/s41572-020-0204-y
- Trivioli G, Marquez A, Martorana D, Tesi M, Kronbichler A, Lyons PA, et al. Genetics of ANCA-associated vasculitis: role in pathogenesis, classification and management. *Nat Rev Rheumatol* (2022) 18(10):559–74. doi: 10.1038/s41584-022-00819-y
- Yates M, Watts RA, Bajema IM, Cid MC, Crestani B, Hauser T, et al. EULAR/ERA-EDTA recommendations for the management of ANCA-associated vasculitis. *Ann Rheum Dis* (2016) 75(9):1583–94. doi: 10.1136/annrheumdis-2016-209133
- Jiang B, Zhao YY, Wei SH. Granulomatosis with polyangiitis: the relationship between ocular and nasal disease. *Ocul Immunol Inflammation* (2013) 21(2):115–8. doi: 10.3109/09273948.2012.747618
- Jennette JC, Falk RJ, Bacon PA, Basu N, Cid MC, Ferrario F, et al. 2012 revised international chapel hill consensus conference nomenclature of vasculitides. *Arthritis Rheum* (2013) 65(1):1–11. doi: 10.1002/art.37715
- Zhang R, Dang X, Shuai L, He Q, He X, Yi Z. Lupus erythematosus panniculitis in a 10-year-old female child with severe systemic lupus erythematosus: A case report. *Med (Baltimore)* (2018) 97(3):e9571. doi: 10.1097/MD.00000000000009571
- Li H, Xiong Z, Liu J, Li Y, Zhou B. [Manifestations of the connective tissue associated interstitial lung disease under high resolution computed tomography]. *Zhong Nan Da Xue Xue Bao Yi Xue Ban* (2017) 42(8):934–9. doi: 10.11817/j.issn.1672-7347.2017.08.010
- Brunner J, Freund M, Prelog M, Binder E, Sailer-Hoeck M, Jungtraithmayr T, et al. Successful treatment of severe juvenile microscopic polyangiitis with rituximab. *Clin Rheumatol* (2009) 28(8):997–9. doi: 10.1007/s10067-009-1177-0
- Jindal G, Cruz SD, Punia RP, Kaur R. Refractory anemia as a presenting feature of microscopic polyangiitis: a rare vasculitis in children. *Indian J Pediatr* (2011) 78(10):1287–9. doi: 10.1007/s12098-011-0459-0
- Wang S, Habib S, Umer S, Reisman L, Raman V. Recurrent posterior reversible encephalopathy syndrome in a child with microscopic polyangiitis. *J Clin Rheumatol* (2015) 21(2):113–4. doi: 10.1097/RHU.0000000000000222
- Alpigiani MG, Calcagno A, Salvati P, Rossi GA, Barbano G, Ghiggeri G, et al. Late onset of pANCA renal and pulmonary vasculitis in a girl affected by undifferentiated connective tissue disease. *Lupus* (2010) 19(5):655–7. doi: 10.1177/0961203309349740
- Roszkiewicz J, Smolewska E. From fibrosis to diagnosis: a paediatric case of microscopic polyangiitis and review of the literature. *Rheumatol Int* (2018) 38(4):683–7. doi: 10.1007/s00296-017-3923-y
- Bhadu D, Kumar P, Malhotra KP, Sharma A, Sharma M, Srivastava D. Central nervous system vasculitis in pediatric microscopic polyangiitis. *Acta Reumatol Port* (2016) 41(4):372–5.
- Wang H, Sun L, Tan W. Clinical features of children with pulmonary microscopic polyangiitis: report of 9 cases. *PloS One* (2015) 10(4):e0124352. doi: 10.1371/journal.pone.0124352
- Dziuban EJ, Castle VP, Haftel HM. Microscopic polyangiitis in an adolescent presenting as severe anemia and syncope. *Rheumatol Int* (2011) 31(11):1507–10. doi: 10.1007/s00296-009-1270-3
- Iglesias E, Eleftheriou D, Mankad K, Prabhakar P, Brogan PA. Microscopic polyangiitis presenting with hemorrhagic stroke. *J Child Neurol* (2014) 29(8):NP1–4. doi: 10.1177/0883073813488661
- Kaseda K, Marui Y, Suwabe T, Hoshino J, Sumida K, Hayami N, et al. Kidney transplantation for a patient with refractory childhood-onset ANCA-associated vasculitis. *Mod Rheumatol* (2016) 26(2):307–9. doi: 10.3109/14397595.2013.877327

Frontiers in Oncology

Advances knowledge of carcinogenesis and tumor progression for better treatment and management

The third most-cited oncology journal, which highlights research in carcinogenesis and tumor progression, bridging the gap between basic research and applications to improve diagnosis, therapeutics and management strategies.

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

