PREDICTING HIGH-RISK INDIVIDUALS FOR COMMON DISEASES USING MULTI-OMICS AND EPIDEMIOLOGICAL DATA

EDITED BY: Lu Zhang, Bailiang Li, Xin Zhou and Yuanwei Zhang

PUBLISHED IN: Frontiers in Genetics and Frontiers in Bioengineering and Biotechnology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-459-9
DOI 10.3389/978-2-88971-459-9

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

PREDICTING HIGH-RISK INDIVIDUALS FOR COMMON DISEASES USING MULTI-OMICS AND EPIDEMIOLOGICAL DATA

Topic Editors:
Lu Zhang, Hong Kong Baptist University, SAR China
Bailiang Li, Stanford University, United States
Xin Zhou, Vanderbilt University, United States
Yuanwei Zhang, University of Science and Technology of China, China

Citation: Zhang, L., Li, B., Zhou, X., Zhang, Y., eds. (2021). Predicting High-Risk Individuals for Common Diseases Using Multi-Omics and Epidemiological Data. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-459-9

Table of Contents

frontiers in Genetics

EDITORIAL
published: 18 August 2021
doi: 10.3389/fgene.2021.737598

Editorial: Predicting High-Risk Individuals for Common Diseases Using Multi-Omics and Epidemiological Data

*Debajyoti Chowdhury[1,2], Xin Zhou[3], Bailiang Li[4], Yuanwei Zhang[5], William K. Cheung[6], Aiping Lu[1,2] and Lu Zhang[1,6]**

[1] Computational Medicine Lab, Hong Kong Baptist University, Kowloon Tong, Hong Kong, [2] School of Chinese Medicine, Institute of Integrated Bioinformedicine and Translational Sciences, Hong Kong Baptist University, Kowloon Tong, Hong Kong, [3] Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, United States, [4] Department of Radiation Oncology, Stanford University School of Medicine, Stanford, CA, United States, [5] The Chinese Academy of Sciences Key Laboratory of Innate Immunity and Chronic Diseases, Hefei National Laboratory for Physical Sciences at the Microscale, Chinese Academy of Sciences Center for Excellence in Molecular Cell Science, Collaborative Innovation Center of Genetics and Development, School of Life Sciences, The First Affiliated Hospital of University of Science and Technology of China, Hefei, China, [6] Department of Computer Science, Faculty of Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Keywords: multi-omics analyses, disease prediction, machine learning, complex diseases, data integration analysis

Editorial on the Research Topic

Predicting High-Risk Individuals for Common Diseases Using Multi-Omics and Epidemiological Data

Physiological data are the reflections of the physiological status of living systems (Terranova et al., 2021). It is precious and preserves meticulous information. Capturing, interpreting, and rationalizing them is imperative for next-generation medicine. Obtaining real-time, patient-centric data have been progressively positioned at the core of digital disruption in healthcare. It promises to deliver an accurate yet early diagnosis, and personalized precision therapy (Esteva et al., 2019). The advent of multi-omics technologies and proficiency in utilizing complex, multi-dimensional biological, epidemiological, and clinical data from bench-side to real-world have significantly steered biomedical research and healthcare practices. With the mounting resources of multi-omics data including transcriptomics, genomics, proteomics, metabolomics, and epigenomics, it becomes challenging to integrate and infer them to insights. However, it is essential in reimagining the scopes of discoveries in predictive healthcare (Boniolo et al., 2021; Ding et al., 2021).

This special issue congregated 15 different studies demonstrating different computational frameworks, algorithms, and methods for inferring multi-omics, high-throughput data for predictive health and early diagnosis of many common diseases. This issue covered different conditions including sleep, gynecological, and oral health, common viral infections, and different cancers including breast cancers (BC), multiple myeloma (MM), stomach adenocarcinoma (SA), esophageal cancer (OC), gastric cancer (GC), and hepatocellular carcinoma (HC).

The majority of the studies published in this topic have introduced diverse methods to predict risks for different cancers (Guo et al.; He et al.; Liu et al.; Pang et al.; Song et al.; Sun J. R. et al.; Sun Z. et al.; Zhao et al.; Zhang et al.; Zhou et al.). Zhou et al. introduced a novel long non-coding RNAs (lncRNAs) based screening method that can indicate risk score for MM. They obtained the raw transcriptome data from Gene Expression Omnibus by performing weighted gene co-expression

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes, University of Nottingham, United Kingdom

***Correspondence:**
Lu Zhang
ericluzhang@hkbu.edu.hk

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *07 July 2021*
Accepted: *26 July 2021*
Published: *18 August 2021*

Citation:
Chowdhury D, Zhou X, Li B, Zhang Y, Cheung WK, Lu A and Zhang L (2021) Editorial: Predicting High-Risk Individuals for Common Diseases Using Multi-Omics and Epidemiological Data. Front. Genet. 12:737598. doi: 10.3389/fgene.2021.737598

network analysis (WGCNA) and principal component analysis to identify several risk lncRNAs. Successively, they employed univariate, least absolute shrinkage, and selection operator (LASSO) Cox regression and multivariate Cox hazard regression analysis to identify the reliable targets of the lncRNAs, LINC00996 and LINC00525 to devise a predictive risk score system. These lncRNAs were associated with survival and involved in the occurrence and progression of MM. Similarly, Zhao et al. identified the six-lncRNA signature as a potential prognostic marker to predict disease-free survival of BC patients. Liu et al. introduced an effective multi-gene modeling framework to predict the overall prognosis of heterogenous SA including their signature mutations. They collected two independent SA cohorts with both genetic profiling and clinical follow-up data to investigate the association between the somatic mutations and prognosis. Guo et al. identified a practical and robust nine-gene prognostic model based on an immune gene dataset. Immune-related genes (IRGs) are crucial contributors to the development of EC. The authors studied the transcriptome data and matched it with the clinical data of OC patients from The Cancer Genome Atlas (TCGA) database. GEPIA2.0 was employed to analyze 4,094 differentially expressed prognostic genes among the 286 normal from Genotype-Tissue Expressions (GTEx) and 182 TCGA samples. Then, they used Clusterprofiler for Gene Ontology annotations and Kyoto Encyclopedia of Genes and Genomes enrichment analysis and performed joint Cox regression analysis to study candidate prognostic biomarkers for OC. Relying on this, they estimated the risk scores of each patient from the expressions of differentially expressed IRGs and the regression coefficient from the regression model.

Sun J. R. et al. focused on alternative splicing (AS) and flagged the AS events as a reliable biomarker for the prognosis of OC. They constructed the splicing factors-AS correlation networks to offer new insights in identifying the potential regulatory mechanisms associated with OC development. In the second study by this team, genomic scores (GS) were calculated based on Genome-Wide Network Analysis to predict the survival in GC (Sun Z. et al.). Their multivariate analysis revealed a GS strategy as a novel prognostic factor that comprises 7 miRNAs, 8 mRNA, and 19 DNA methylation sites.

The power of machine learning models have emerged in the study by He et al. Sequencing-based identification of tumor tissue-of-origin (TOO) is critical for patients with cancers of unknown primary lesions. There has always been a probability of misdiagnosis. To avoid those issues, He et al., developed a machine learning model using the expression of a 150-gene panel to infer the tumor TOO for 15 common solid tumor cancer types, including lung, breast, liver, colorectal, gastroesophageal, ovarian, cervical, endometrial, pancreatic, bladder, head and neck, thyroid, prostate, kidney, and brain cancers. They studied 7,460 primary tumor samples across those 15 cancer types and employed the Support vector machines based recursive feature elimination algorithm to perform the feature selection and classification modeling on gene expression data. It designated 154 out of the 11,925 genes with distinct biological significance. Thus, they elucidated a robust classifier on gene expression data to predict TOO-based accurate reclassifications of cancer types which were supplemented with clinical examination.

Zhang et al. introduced an interesting method relying on miRNA-based nomogram to predict distal lung metastasis of BC. They acquired miRNA and clinicopathological data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and screened out 8 miRNAs as highly relevant to lung metastasis of BC patients. They used the limma package to distinguish miRNAs annotated within the METABRIC dataset and differentially expressed miRNAs (DEMs). They employed LASSO regression to select the most suitable predictive miRNAs from the 16-lung metastasis-related DEMs and formulated a risk-score prediction tool relying on 8-miRNAs for predicting lung metastasis status of BC patients in the training set. Then, they used univariate and multivariate logistic regression analysis to determine the proficiency of those 8 miRNAs as predictors and employed decision-curve analysis to test its clinical applicability. Song et al. investigated a vital direction to identify the hub genes associated with HC. Using a Robust Rank Aggregation method combined with WGCNA, they constructed a clinically relevant prediction model to uncover the complex biological mechanisms of HC.

Sleep is one of the most neglected public health concerns. Sambou et al. instituted a large study comprising the big data obtained from 328,850 participants to endorse a data-driven decision on the associations of the quality of sleep and the healthier life span.

Implantation failure (IF) is one of the recurring issues in assisted pregnancy (Busnelli et al., 2021). Thin endometrium (TE) is a critical factor in IF. mRNA-miRNA cross-talks have been repeatedly flagged as one of the essential etiologies for IF. Xu, B et al., reconstructed integrative transcriptional regulatory networks based on the miRNA-mRNA expression profiles in the TE and normal endometrium tissue obtained from 8 patients (Zong et al.). It involved the miRNA sequence analysis using the DeAnnIso tool (Zhang et al., 2016). They employed Solexa CHASTITY and Cutadapt pipeline to process mRNA sequence data and identified multiple hub genes by constructing the miRNA–mRNA regulatory networks that illuminate new insights underpinning the TE formation (Zong et al.). Huang et al. studied single-cell transcriptional profiles to identify the impact of sex and age on the gene expression of endothelial cells. The transcriptomes of endothelial cells from 5 organs, heart-aorta, fat, lungs, limb, muscle, kidney of the mouse were analyzed. It discovered that older mice had increased expressions of genes involved in inflammation in endothelial cells, which may contribute to the development of chronic, non-communicable diseases like atherosclerosis, hypertension, and Alzheimer's disease with age.

Another study focused on host-pathogen interactions and devised oligoadenylate synthetases-like (OASL) as a potential biomarker for early detection of flu-mediated acute respiratory infection (ARI) cases (Li et al.). This study was aimed to distinguish a strong single-gene biomarker with a superior diagnostic accuracy by using integrated bioinformatics analysis with XGBoost, a feature selection method relying on recursive feature elimination with cross-validation (Li et al.). They

analyzed transcriptome profiles to reconstruct a co-expression network by employing WGCNA to identify the OASL as a hub gene for ARI. Pang et al. applied random forest to predict dental caries risks among teenagers. They constructed the caries risk prediction model that serves as an easy, accessible community-level tool to identify individuals with high caries risk.

All of the research articles published under this topic introduced the state-of-the-art technologies employed on multiplexed physiological data. It offers a newer perspective on the early diagnosis of different diseases using data-driven approaches. We anticipate it will be impactful in accelerating the scopes in predictive healthcare research and applications.

AUTHOR CONTRIBUTIONS

This editorial was designed by DC and LZ, written by DC, edited and revised by LZ, XZ, BL, and YZ, and supported by WC and AL. All authors made a direct and intellectual contribution to this topic and approved the article for publication.

FUNDING

This work was supported by Research Grant Council Early Career Scheme (HKBU 22201419), IRCMS HKBU (Grant No. IRCMS/19-20/D02), Guangdong Basic and Applied Basic Research Foundation (Grant Nos. 2019A1515011046 and 2021A1515012226).

ACKNOWLEDGMENTS

We would thank Research Grants Council of Hong Kong, Hong Kong Baptist University and HKBU Research Committee for their kind support of this project. We also thank all the authors who contributed to this topic.

REFERENCES

Boniolo, F., Dorigatti, E., Ohnmacht, A. J., Saur, D., Schubert, B., and Menden, M. P. (2021). Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opin. Drug Discov.* doi: 10.1080/17460441.2021.1918096. [Epub ahead of print].

Busnelli, A., Somigliana, E., Cirillo, F., Baggiani, A., and Levi-Setti, P. E. (2021). Efficacy of therapies and interventions for repeated embryo implantation failure: a systematic review and meta-analysis. *Sci. Rep.* 11:1747. doi: 10.1038/s41598-021-81439-6

Ding, J., Blencowe, M., Nghiem, T., Ha, S. M., Chen, Y. W., Li, G., et al. (2021). Mergeomics 2.0: a web server for multi-omics data integration to elucidate disease networks and predict therapeutics. *Nucleic Acids Res.* 49, W375–W387. doi: 10.1093/nar/gkab405

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. doi: 10.1038/s41591-018-0316-z

Terranova, N., Venkatakrishnan, K., and Benincosa, L. J. (2021). Application of machine learning in translational medicine: current status and future opportunities. *AAPS J.* 23, 1–10. doi: 10.1208/s12248-021-00593-x

Zhang, Y., Zang, Q., Zhang, H., Ban, R., Yang, Y., Iqbal, F., et al. (2016). DeAnnIso: a tool for online detection and annotation of isomiRs from small RNA sequencing data. *Nucleic Acids Res.* 44, W166–W175. doi: 10.1093/nar/gkw427

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.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 02 July 2020
doi: 10.3389/fbioe.2020.00729

OASL as a Diagnostic Marker for Influenza Infection Revealed by Integrative Bioinformatics Analysis With XGBoost

Yang Li[1†], Hongjie Liu[2†], Quan Xu[3], Rui Wu[4], Yi Zhang[1], Naizhe Li[1], Xiaozhou He[1], Mengjie Yang[1], Mifang Liang[1*] and Xuejun Ma[1,5*]*

[1] NHC Key Laboratory of Medical Virology and Viral Diseases, National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China, [2] BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China, [3] ChosenMed Technology (Beijing) Co., Ltd., Beijing, China, [4] Department of Pathology, Peking University Third Hospital, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, China, [5] Center for Biosafety Mega-Science, Chinese Academy of Sciences, Wuhan, China

OPEN ACCESS

Edited by:
Xin Maizie Zhou,
Stanford University, United States

Reviewed by:
Aline Silva Mello Cesar,
University of São Paulo, Brazil
Yan Gong,
Wuhan University, China

****Correspondence:***
Yang Li
yeli7068@outlook.com
Mifang Liang
mifangl@163.com
Xuejun Ma
maxj@ivdc.chinacdc.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *22 February 2020*
Accepted: *09 June 2020*
Published: *02 July 2020*

Citation:
Li Y, Liu H, Xu Q, Wu R, Zhang Y, Li N, He X, Yang M, Liang M and Ma X (2020) OASL as a Diagnostic Marker for Influenza Infection Revealed by Integrative Bioinformatics Analysis With XGBoost. Front. Bioeng. Biotechnol. 8:729. doi: 10.3389/fbioe.2020.00729

Host response biomarkers offer a promising alternative diagnostic solution for identifying acute respiratory infection (ARI) cases involving influenza infection. However, most of the published panels involve multiple genes, which is problematic in clinical settings because polymerase chain reaction (PCR)-based technology is the most widely used genomic technology in these settings, and it can only be used to measure a small number of targets. This study aimed to identify a single-gene biomarker with a high diagnostic accuracy by using integrated bioinformatics analysis with XGBoost. The gene expression profiles in dataset GSE68310 were used to construct a co-expression network using weighted correlation network analysis (WGCNA). Fourteen hub genes related to influenza infection (blue module) that were common to both the co-expression network and the protein–protein interaction network were identified. Thereafter, a single hub gene was selected using XGBoost, with feature selection conducted using recursive feature elimination with cross-validation (RFECV). The identified biomarker was oligoadenylate synthetases-like (OASL). The robustness of this biomarker was further examined using three external datasets. OASL expression profiling triggered by various infections was different enough to discriminate between influenza and non-influenza ARI infections. Thus, this study presented a workflow to identify a single-gene classifier across multiple datasets. Moreover, OASL was revealed as a biomarker that could identify influenza patients from among those with flu-like ARI. OASL has great potential for improving influenza diagnosis accuracy in ARI patients in the clinical setting.

Keywords: influenza infection, host response, OASL, XGBoost, WGCNA

INTRODUCTION

Acute respiratory infection (ARI) is responsible for significant levels of morbidity and mortality worldwide related to infectious diseases. Viruses and bacteria are the main causes of ARI. Among the viruses, influenza virus kills more people than other viruses. It has been estimated that there were 250,000–500,000 additional deaths during the first 12 months of the global circulation of the 2009 pandemic H1N1 influenza A virus (Dawood et al., 2012). Better diagnostics for ARI (with or without influenza virus) are urgently needed in both inpatient and outpatient settings. However,

discriminating between influenza and non-influenza flu-like illnesses on clinical grounds is often difficult, because these ARIs share similar clinical features (e.g., cough and fever).

Diagnostic methods for viral pathogens, such as culture, serodiagnosis, nucleic acid-based methods, and high-throughput sequencing, are important to guide disease management. When the presence of a viral pathogen is confirmed by these methods, this does not exclude a possible coinfection with bacteria, leading to antimicrobial prescriptions "just in case" (Tsalik et al., 2016). Moreover, as for most respiratory pathogens, the presence of influenza virus is sometimes unrelated to the presenting illness (Jansen et al., 2011). There is currently widespread interest in tests for virus detection in general and tests for "active" virus detection.

The host response to infection provides an alternative target for "active" virus detection. It has been reported that biomarkers based on host gene expression have great potential for distinguishing ARI patients infected with viruses versus bacteria (Herberg et al., 2016; Sweeney et al., 2016b; Tsalik et al., 2016; Yu et al., 2019). In addition to ARI, other infectious diseases such as tuberculosis (Sweeney et al., 2016a), systemic inflammation (Sampson et al., 2017) and hemorrhagic fevers (Robinson et al., 2019) have been studied using this approach. Most published panels for detecting the host response to infections contained multiple genes, making it difficult to apply them in clinical settings, as polymerase chain reaction (PCR)-based technologies could only measure a small number of targets. Recently, interferon alpha-inducible protein 27 (IFI27) was found to be able to distinguish influenza and non-influenza flu-like illnesses in a large cohort, with an area under the curve (AUC) value of 0.87 (Tang et al., 2017). However, IFI27 was the most upregulated gene during influenza virus, respiratory syncytial virus (RSV), and human rhinovirus (HRV) infections (Ioannidis et al., 2012; Zhai et al., 2015). Here, we aimed to follow the single-gene strategy to improve the discrimination between influenza and non-influenza flu-like illnesses based on an integrated bioinformatics analysis with XGBoost (**Figure 1**).

MATERIALS AND METHODS

Study Design

The purpose of this study was to use an integrated bioinformatics analysis to analyze multiple gene expression datasets in order to identify a biomarker that can accurately classify patients with influenza or non-influenza flu-like illnesses, including bacterial infections and other viral infections. The general study workflow was shown in **Figure 1**.

Data Collection

In brief, data were obtained from the Gene Expression Omnibus (GEO) database[1] in December 2019 using the keyword "influenza cohort." The following exclusion criteria were applied to the microarray data: (1) only involved influenza infection; (2) no or insufficient clinical data; (3) concerned influenza vaccine responses; and (4) used non-baseline ("healthy") controls. After review, GSE68310, which contains 880 samples from 133 subjects with influenza infection or other viral ARIs, was selected for biomarker discovery (Zhai et al., 2015).

For the validation stage, three external independent microarray datasets were selected. GSE6269 (Ramilo et al., 2007) was used to evaluate the diagnostic performance between influenza and bacterial infections. Both GSE42026 (Herberg et al., 2013) and GSE38900 (Mejias et al., 2013) were used to estimate the discriminatory power to differentiate the influenza against other viral infections. In addition to controls, the three datasets contained cases with common bacterial and viral respiratory infections, i.e., *Streptococcus pneumoniae*, *Staphylococcus aureus*, influenza virus, HRV, and RSV etc. Before further analysis, the expression matrices were normalized and log2-transformed.

Differentially Expressed Genes Screening

The limma R package was used to screen the influenza infection associated differential expressed genes (DEGs). DEGs analyses contrasting the Day 0 influenza A virus infected individual data with the baseline samples were performed by function for linear model fitting in the R package limma (Ritchie et al., 2015). Correction for multiple testing was addressed by controlling the false discovery rate (FDR) using the Benjamini–Hochberg (B.H.) method. Criteria for DEGs were an absolute $\log_2$ fold change (Log_2FC) of 0 and the FDR-adjusted P-value of <0.05.

Co-expression Network Construction

A co-expression network was constructed using the normalized GSE68310 data by the weighted correlation network analysis (WGCNA) in R (Langfelder and Horvath, 2008). Briefly, quality assessment of GSE68310 samples was conducted using the cluster method. The soft-thresholding power was then calculated, with the type of network set to signed. The correlation coefficient threshold was 0.90. Network construction was then performed based on the calculated power. In addition, the minimum number of genes in each module was 30 and the threshold for cut height was set to 0.25 to merge possible similar modules.

Identification of Modules Related to Influenza Infection

For a given module, the expression profile was summarized into a single characteristic expression profile, designated module eigengenes (MEs). MEs were considered as the first principal component in the principal component analysis (PCA). Thereafter, a Pearson correlation analysis, calculating the Student asymptotic P-values for the correlations, between MEs and clinical traits (Progression, Baseline, Day0 of viral infection and gender) was conducted.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Analyses

To understand the functions of enriched genes in interesting modules, Gene Ontology (GO) (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2017) analyses were performed using clusterProfiler

[1] http://www.ncbi.nlm.nih.gov/geo/

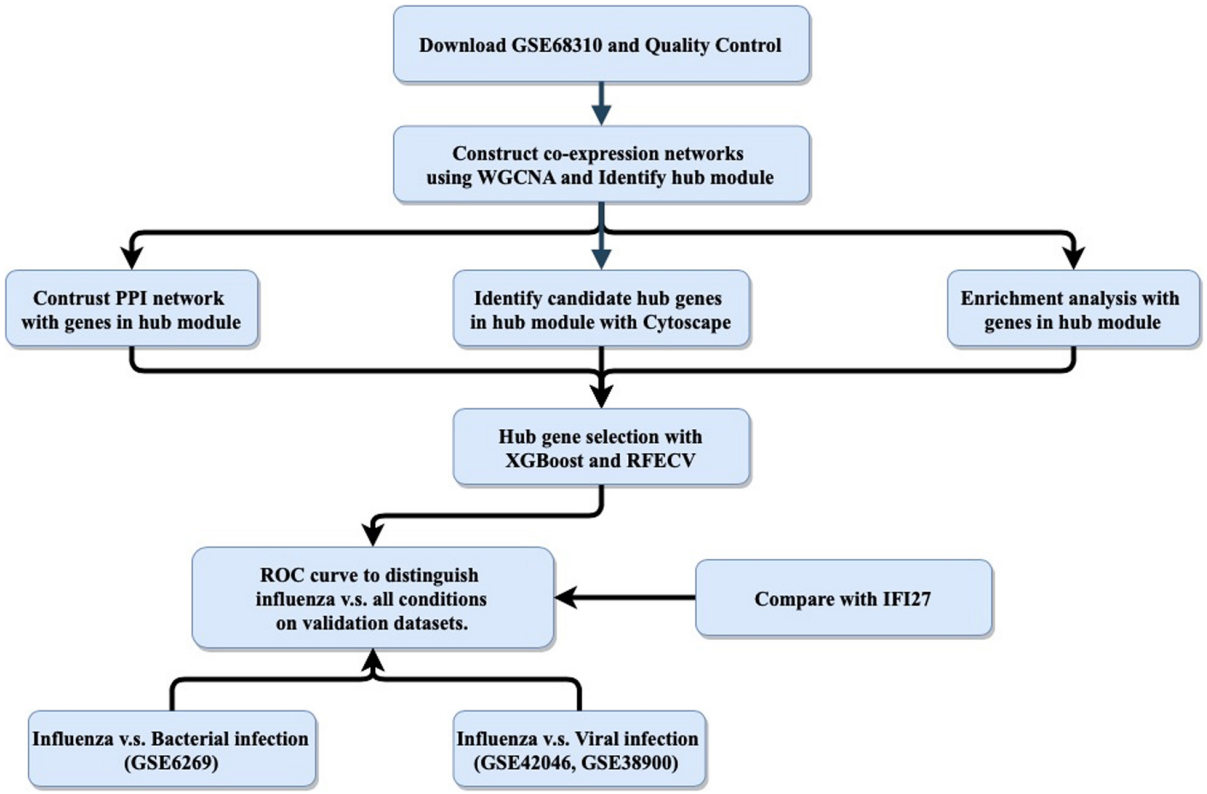


FIGURE 1 | General study workflow: data collection, *in silico* analysis, and external validation. PPI, protein–protein interaction; RFECV, recursive feature elimination with cross-validation.

(Yu et al., 2012), identifying significant results based on a Benjamini–Hochberg FDR-adjusted *P*-value ≤0.05.

Candidate Hub Gene Selection

Three bioinformatics approaches were combined to select the hub genes. First, the module that was most highly correlated with influenza infection was selected. Hub genes in the module were determined by both gene significance and module membership. Second, all the interesting genes were uploaded to the Search Tool for the Retrieval of Interacting Genes (STRING) database[2] to create a protein–protein interaction network (PPIN) (Szklarczyk et al., 2019). Hub genes in PPIN were selected by maximum neighborhood component (MNC), degree and maximal clique centrality (MCC) using cytoHubba with Cytoscape (Shannon et al., 2003; Chin et al., 2014). Thereafter, hub genes common to both networks were chosen. Finally, a single hub gene was selected using XGBoost with recursive feature elimination with cross-validation (RFECV) (Pedregosa et al., 2011; Chen and Guestrin, 2016).

External Dataset Validation of the Hub Gene

We validated the hub gene-based classification performance related to distinguishing influenza and non-influenza acute respiratory illness using the external datasets GSE6269, GSE42026, and GSE38900. We also compared the performance of the selected hub gene to the performance of IFI27, which is a biomarker that discriminates influenza from all other conditions, with an AUC value of 0.87 (Tang et al., 2017). Additionally, a receiver operating characteristic (ROC) curve was plotted, and AUC was calculated using "pROC" (Robin et al., 2011) to evaluate the performance of the selected hub gene regarding distinguishing influenza infection from all other conditions.

Statistical Analysis

R (version 3.5.1) was used for most analyses, with hub gene selection being performed using XGBoost in Python (version 3.6). The statistical significance of pairwise differences between groups was analyzed using a two-tailed *t*-test. *P*-value ≤0.05 was considered statistically significant.

RESULTS

Quality Control and Sample Selection

Raw data in dataset GSE68310 was subjected to background adjustment, variance stabilization after log2 transformation, rank invariant normalization, and quality control evaluation with a detection *P*-value less than 0.05 by using corresponding functions

[2]https://string-db.org

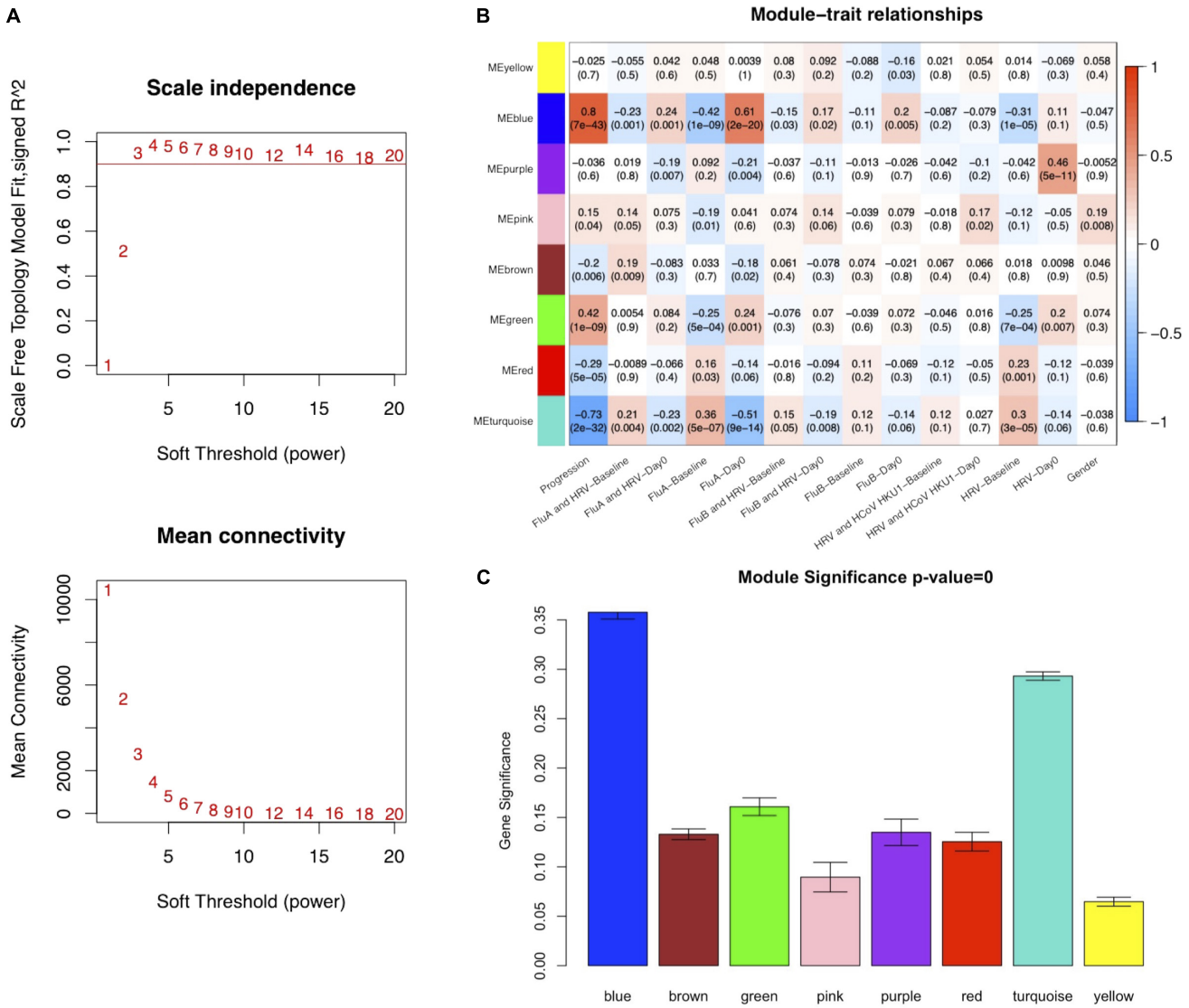


FIGURE 2 | Co-expression network constructed using weighted correlation network analysis (WGCNA). **(A)** Analysis of the scale-free fit index with a threshold of 0.90 (top) and mean connectivity (bottom) for various soft-thresholding power values. **(B)** Distribution of average gene significance and errors in the modules associated with Influenza infections (FluA-Day0). **(C)** Heatmap of the correlation between module eigengenes and the clinical traits recorded in GSE68310. FluA, influenza A virus; FluB, influenza B virus; HRV, human rhinovirus; HCoV, human coronavirus.

in the R package lumi (Du et al., 2008). The preprocessed expression matrix was then normalized by quantile method in R package limma. Thereafter, the probe sets with known gene symbol were kept, with 20,914 probes out of 47,254 remaining. No samples were removed after cluster analysis (**Supplementary Figure S2**).

Influenza Associated DEGs

After quality control, we obtained the normalized expression matrices from GSE68310. Under the threshold of FDR < 0.05 and $|\log_2 FC| \geq 0$, a total of 6142 DEGs (2465 up-regulated and 3677 down-regulated) were achieved. The volcano plot of DEGs were shown in **Supplementary Figure S2**.

Weighted Co-expression Network and Identification of the Influenza Infection-Related Module

To ensure that a scale-free network was constructed, a soft-thresholding power of 3 was selected while 0.90 was used as the correlation coefficient threshold (**Figure 2A**). After removing the gray module which contained unassigned genes (n = 10,047), a total of eight modules were identified and constructed in the WGCNA analysis (**Figure 2B**). The module with the most genes was the turquoise (n = 3127) module, followed by the blue (n = 1930), and brown (n = 1155) modules (**Supplementary Figure S3**). Modules with a greater MS were considered to have more connection with the influenza infections, and we

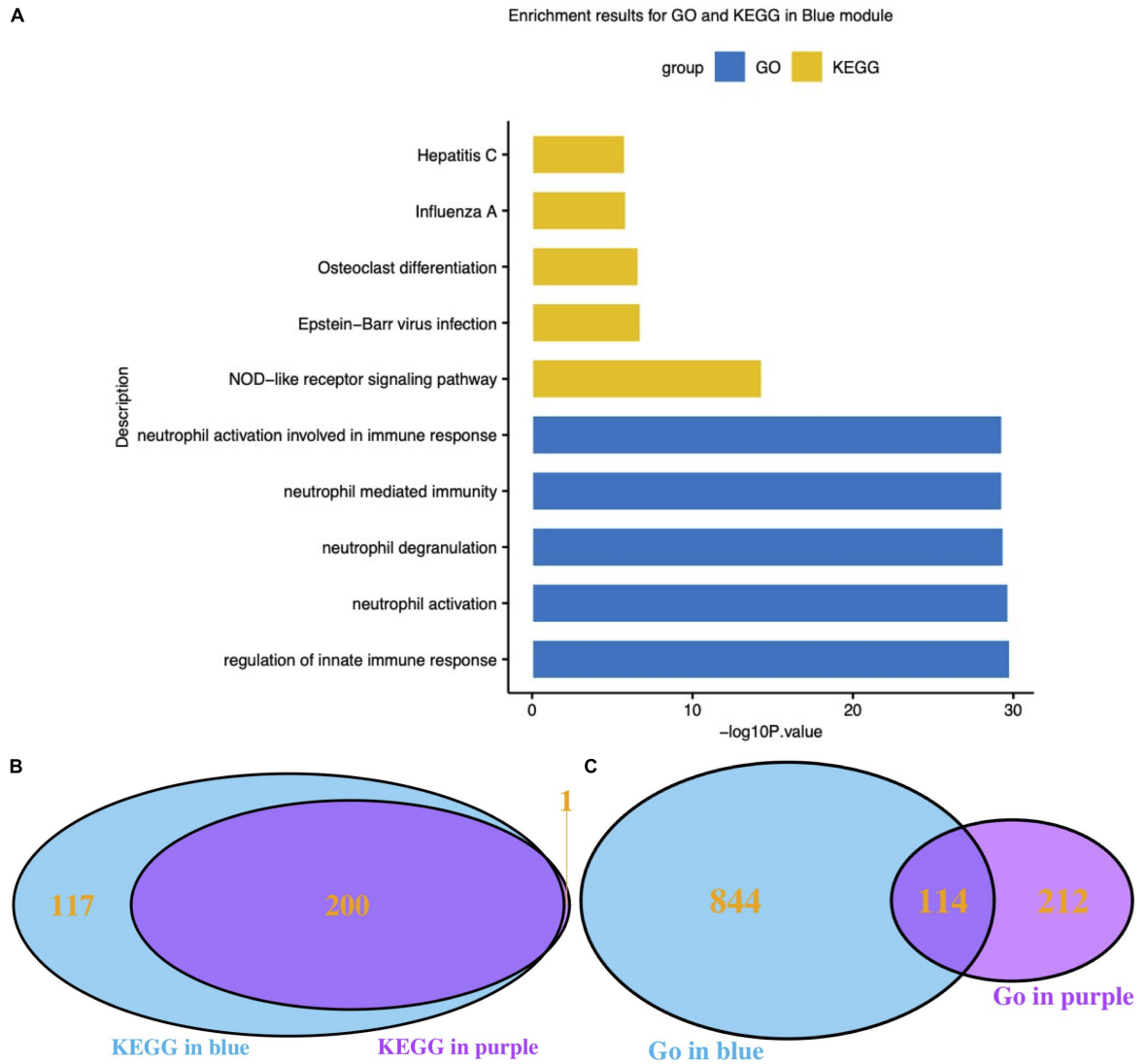


FIGURE 3 | Functional analysis of interesting modules. **(A)** GO and KEGG enrichment results for the blue module; **(B)** Venn diagram of KEGG results for the blue and purple modules; **(C)** Venn diagram of GO results for the blue and purple modules. GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes.

found that the MS of the blue module was higher than those of any other modules (**Figure 2C**). In addition, module–trait correlation analyses showed that multiple modules were related to influenza infection. The Pearson correlation analysis, which involved calculating the Student asymptotic *P*-values for the correlations, between the MEs of each module and clinical traits is shown in **Figure 2B**. The blue module was the module most relevant to influenza infection, while the purple module was related to HRV infection.

Quality Control of Modules Using Functional Analysis

Functional enrichment results of genes in the blue module, which was highly related to influenza infection, should hypothetically be related to the immune response to viruses. The GO and KEGG functional enrichment results were both used to examine this hypothesis (**Figure 3A**). The most highly enriched GO terms included regulation of innate immune response, neutrophil activation, neutrophil degranulation, neutrophil mediated immunity, and neutrophil activation involved in immune response. The KEGG results directly included the influenza A pathway (**Figure 3**).

It has been reported that different respiratory viruses can cause similar symptoms via different mechanisms. As the purple module was associated with HRV infection, GO and KEGG analyses were also performed on the genes in the purple module. The KEGG pathway results clearly suggested that the blue module (influenza-related) and the purple module (HRV-related) shared highly similar KEGG pathways (**Figure 3B**). Conversely, the GO Biological Process results were very dissimilar (**Figure 3C**). Thereafter, the correlation between

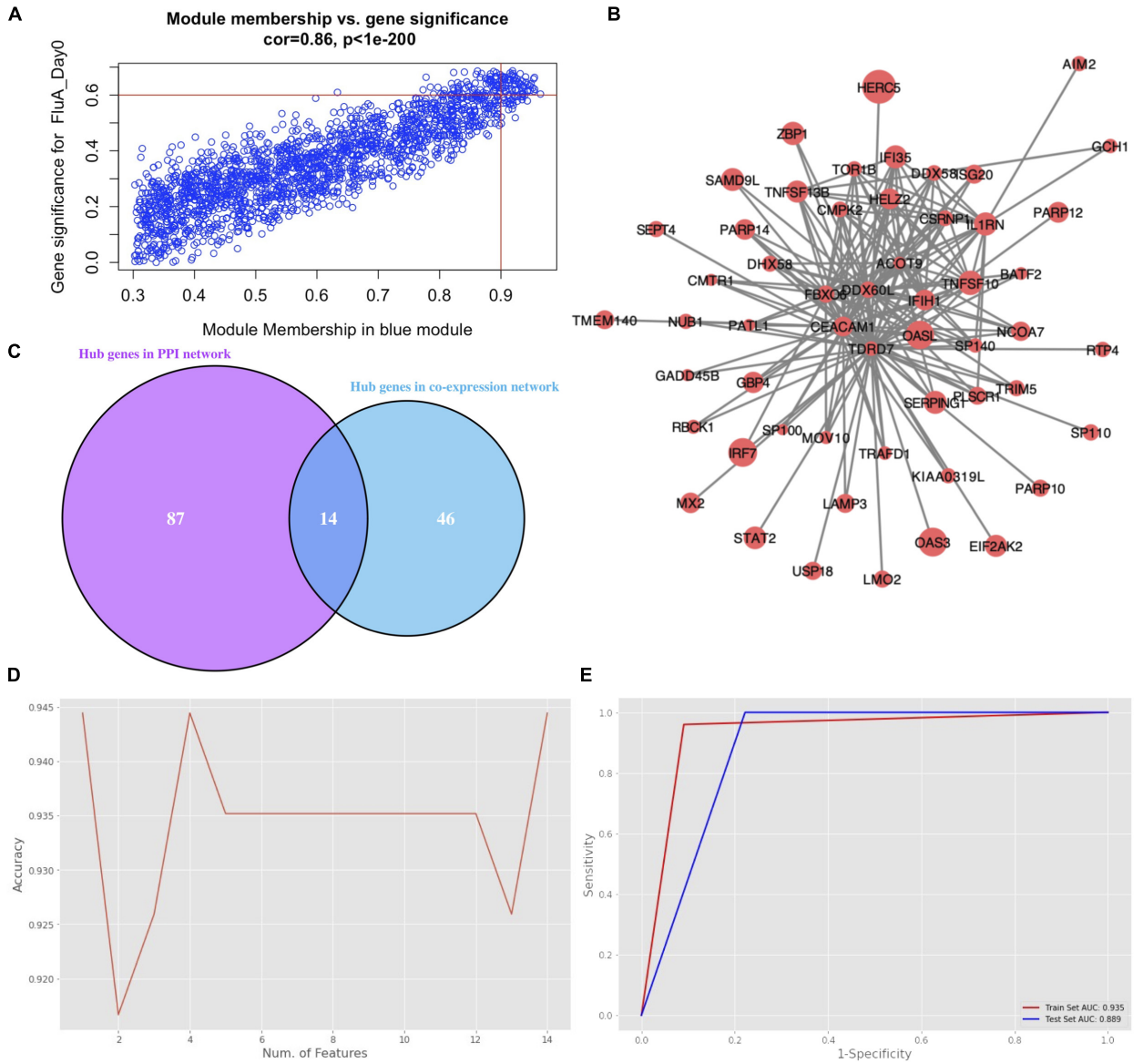


FIGURE 4 | Hub gene selection. **(A)** Scatter plot of module eigengenes in the blue module with selection thresholds. **(B)** Visualization of the network connections among the most connected genes in the blue module. The size of circles was equal to the $\log_2$ fold change. **(C)** Common hub genes in both the PPI and co-expression networks. **(D)** Classification accuracy versus number of genes, based on the combination of XGBoost and recursive feature elimination with cross-validation. **(E)** Evaluation of classification performance of the selected hub gene, oligoadenylate synthetases-like (OASL), using dataset GSE68310.

module membership regarding the blue module and gene significance for HRV was assessed. No correlation was found, as shown in **Supplementary Figure S4** ($r = -0.11$, $P = 1.3\text{e-}6$). Therefore, the presence of a unique set of genes in the blue module was correlated with influenza infections.

Hub Gene Selection

The genes in the blue module were identified as candidate hub genes by the co-expression network approach. A total of 106 genes were selected using a gene significance threshold of 0.9 and a module membership significance of 0.6 (**Figure 4A** and **Supplementary Table S1**). In addition, the network connections among the most connected genes in the blue module was displayed through Cytoscape (**Figure 4B**). Next, a PPIN of all the genes in the blue module was constructed using Cytoscape based on the STRING database. The top 101 genes shared by MNC, degree and MCC through cytoHubba were considered as hub genes (**Supplementary Table S1**). Thereafter, 14 genes that were common to both networks were selected as the candidates to be further analyzed (**Figure 4C** and **Supplementary Figure S5**).

Hub gene selection based on XGBoost and RFECV was carried out using the 14 candidate genes. The samples labeled "Day0" (meaning that samples were collected within 48 h of ARI onset, i.e., in the acute phase) with data on the 14 genes were firstly

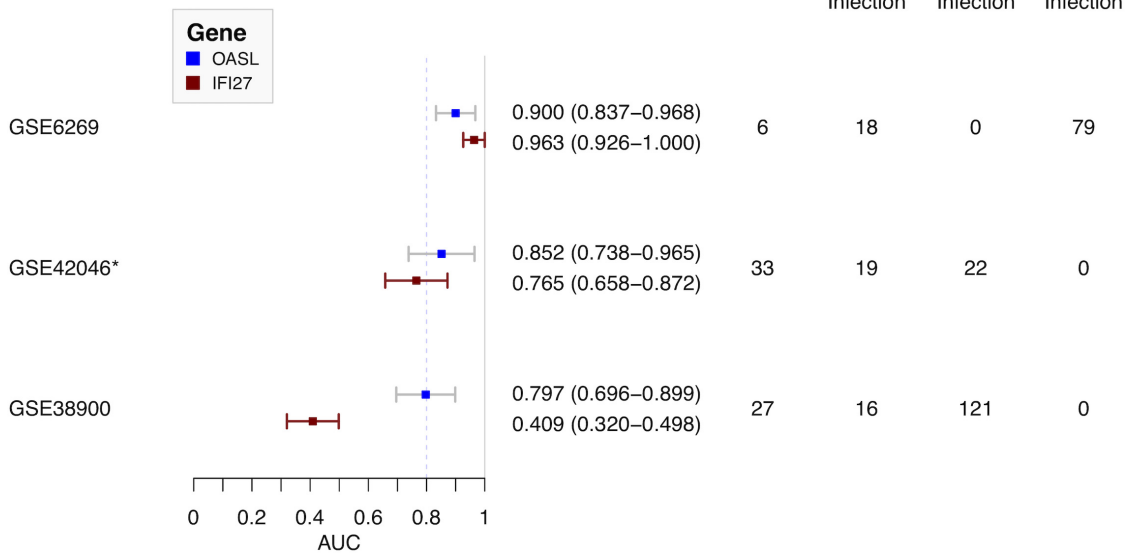


FIGURE 5 | Forest plot of diagnostic performance of OASL and IFI27 on external cohorts. AUC, Area under curve. *Cases with bacterial infections were removed.

standardized. They were then randomly assigned at a 7:3 ratio to a training set (93 samples) and a test set (40 samples). The "XGBoost" package in Python was used for data classification. Parameter max_depth was defined as 3; learning_rate was defined as 0.01; gamma was defined as 0.05; n_estimators was defined as 100. To obtain the best XGBoost model parameter combination (learning_rate, max_depth, gamma, and n_estimators) with the highest classification accuracy, fivefold cross-validation and grid search were applied to the training set. RFECV was then applied for feature selection based on the feature importance scores calculated by XGBoost. Parameter step was defined as 1; cv was defined as 5. The highest accuracy of classification was 0.944 which could be achieved through a single gene, oligoadenylate synthetases-like (OASL) (**Figure 4D**). Moreover, the AUC score in the training and test sets for this single gene was 0.935 and 0.889, respectively (**Figure 4E**).

External Validation Cohorts

Three external cohorts were chosen to evaluate the diagnostic performance of the single gene-based classifier (**Figure 5**). First of all, GSE6269 was used to evaluate the diagnostic performance between influenza and bacterial infections. Both OASL and IFI27 showed high diagnostic accuracy (0.900 and 0.963, respectively). Next, GSE42026 and GSE38900 were used to estimate the discriminatory power to differentiate the influenza virus against other respiratory viruses. To meet this aim, cases with bacterial infection (n = 18) were firstly removed in GSE42026. After that, the AUC of OASL was 0.852 (95% CI: 0.738–0.965) while the AUC of IFI27 was 0.765 (95% CI: 0.658–0.872). For GSE38900, the AUC of OASL was 0.797 (95% CI: 0.696–0.899) while the AUC of IFI27 was 0.409 (95% CI: 0.320–0.498). AUC values were calculated using bootstrapping validation (Robin et al., 2011). Based on these findings, OASL achieved overall accurate results.

DISCUSSION

Over the last decade, considerable achievements have been made regarding the discovery of gene expression biomarkers of infections, especially respiratory illnesses (Herberg et al., 2016; Sweeney et al., 2016b; Tang et al., 2017; Robinson et al., 2019; Yu et al., 2019). In clinical settings, panels with multiple genes are problematic for infection diagnostics, as the most widely used genomic technology in clinical settings is PCR-based technologies, which can only be used to assess a handful of targets. To overcome this barrier, a single gene-based diagnostic strategy will be highly beneficial. IFI27 has recently been reported to be able to distinguish between influenza and bacterial infections (with an AUC of 0.91) and between influenza and non-influenza but flu-like illness (with an AUC of 0.87) (Tang et al., 2017). However, IFI27 has been found to be the highest upregulated gene during both influenza and RSV infections (Ioannidis et al., 2012). Therefore, an integrated bioinformatics analysis with machine learning was performed in this study to identify a hub gene that was specific to influenza infection.

As ARIs share similar clinical features and various respiratory viruses trigger a variety of interferon-stimulated genes (ISGs), an ideal dataset for biomarker discovery should include not only influenza infections, but also other respiratory infections. GSE68310 was finally selected (Zhai et al., 2015). To discriminate influenza infections from other viral infections, WGCNA, an unsupervised analysis method that clusters genes based on their expression profiles, was the first step to identify the hub module associated with influenza infection. Moreover, quality control involving enrichment analysis was performed on both the blue (influenza-related) module and the purple (HRV-related) module. Although diverse GO results were observed, similar KEGG pathways were enriched, which provides insights as to why

the clinical features are similar among various viral infections (**Figures 3B,C**). The ISGs related to different viral infections were unique, which was consistent with previous research (Ioannidis et al., 2012; Andres-Terre et al., 2015). Therefore, the presence of a distinctive set of genes in the blue module was as expected.

To obtain a single hub gene for influenza infection, XGBoost was applied to the high-dimensional gene expression matrix. Compared with other ensemble machine learning algorithms, XGBoost extends simple classification and regression trees (CARTs) instead of building a single tree. Building many trees and then aggregating them to form a single consensus prediction model can improve the prediction accuracy (Chen and Guestrin, 2016). In addition, as a tree-based algorithm, XGBoost provided an importance score for each gene in each tree model. The importance score revealed how informative the gene was. RFECV showed good performance regarding feature reduction. Finally, the hub gene OASL was selected and tested in the discovery dataset GSE68310 (**Figure 4**).

To evaluate the diagnostic performance of OASL, three external datasets were selected (**Figure 1**). Firstly, both OASL and IFI27 shared similar highly accurate performance in discriminating between influenza and bacterial infections on GSE6269. To classify influenza and viral infections, OASL outperformed IFI27 slightly on GSE42026 with an AUC of 0.852 (95% CI 0.738–0.965) versus 0.765 (95% CI 0.658–0.872). In addition, we investigated another external cohort GSE38900 as a challenge dataset which contained 121 cases with non-influenza viral infections. Although both OASL and IFI27 showed reduced AUC on GSE38900, it was worth of noting that the AUC of OASL still remained close to 0.8. To avoid poor reproducibility across external patient populations, more studies with larger sample sizes were needed to verify the diagnostic performance of OASL.

Oligoadenylate synthetases-like, a member of the OAS family, mediates antiviral activities via promoting retinoic acid-inducible gene I (RIG-I)-mediated signaling by mimicking polyubiquitin (pUb) (Zhu et al., 2014). Notably, to evade host innate immunity, a number of viruses (especially influenza virus) target ubiquitin ligases or encode deubiquitinases (DUBs) and DUB-like molecules (Gack et al., 2009). Thus, in the absence of pUb (which is caused by influenza viruses), the activation of RIG-I triggered by OASL plays central roles in host antiviral activities. Recently, OASL has been considered as a new player in controlling antiviral innate immunity (Zhu et al., 2015). In addition, OASL was included by previous panels for discriminating viral and bacterial infections (Andres-Terre et al., 2015; Sampson et al., 2017). It was consistent with present results. OASL has considerable discriminatory power in differentiating between viral and bacterial infections (**Figure 5**). It was worthy of noting the expressions of OASL triggered by various viruses were different enough to tell influenza infection apart from other viral infections (**Figure 5** and **Supplementary Figure S8**). The role of expressions of OASL triggered by different viruses in the pathogenesis of ARI need to be studied in the future.

Compared with other genomic technologies, influenza-targeted quantitative reverse transcription polymerase chain reaction (qRT-PCR) was widespread in clinical practice. The performance of PCR was limited because samples tend to be collected prior to ARI onset (and, sometimes, late in the illness), there is often a limited specimen quantity, and the nucleic acid (typically RNA) is often degraded. However, OASL was found to be upregulated during the progression of influenza infection (**Supplementary Figure S9**). To our surprise, OASL remained upregulated at 21 days after ARI onset which was the timepoint the subject had clinically recovered. The same trend was observed for IFI27 (**Supplementary Figure S9**). This might be caused by the influenza virus load was reduced but not eliminated. Therefore, identification of OASL expression might indicate the presence of an influenza infection when PCR indicated a negative result. As the OASL expression value was important and influenza is an RNA virus, we suggested using qRT-PCR to detect both OASL expression and influenza virus to distinguish between influenza and non-influenza flu-like cases in clinical settings.

Nevertheless, our study had certain limitations. First of all, the performances of OASL in the external datasets were moderate (AUC < 0.9). Secondly, limited types of viral infections were validated in the datasets. ARI is not caused by one or two viruses but a diverse viral community in the respiratory tract. We previously found that RSV, human coronaviruses (HCoV), human bocavirus (HBoV), influenza virus, human adenoviruses (HAdV), and human parainfluenza virus (HPIV) may be the main causes of severe ARI in Beijing, China (Wang et al., 2016). Thirdly, although it is accepted that the current study provides useful baseline data for future study, an ideal approach should be to perform a prospective study to verify the usefulness of OASL as an influenza ARI biomarker. Yet, it will be challenging to collect ARI specimens currently during the COVID-19 pandemic. Moreover, qRT-PCR is a commonly used validation tool for confirming gene expression results obtained from microarray. Therefore, we shall apply qRT-PCR to test the OASL assay's accuracy with various ARI in the future work.

On the whole, this study addressed a major challenge related to translating genomic science into clinical practice. It has recently been reported that transcriptomes in nasal and blood samples from ARI patients exhibit similar patterns of type I interferon response (Yu et al., 2019). Thereafter, we suggested that a combination of both OASL and universal influenza detection, as measured by qRT-PCR using nasal samples, could be utilized to identify influenza infection in individuals with flu-like illness. Ultimately, before the OASL and influenza assay is used in clinical practice, there will be a need for prospective studies to establish its clinical utility as well as cost-effectiveness analyses.

DATA AVAILABILITY STATEMENT

The microarray datasets GSE68310, GSE6269, GSE42026, and GSE38900 for this study can be found in the Gene Expression Omnibus (GEO) database hosted by the National Center for Biotechnology Information of the US National Institutes of Health (https://www.ncbi.nlm.nih.gov/geo/).

AUTHOR CONTRIBUTIONS

YL, ML, and XM conceived of the study. YL, HL, and QX collected and analyzed the data. RW, YZ, NL, XH, and MY analyzed the data partially. YL and HL drafted the manuscript. ML and XM revised the manuscript. All authors read and approved the manuscript.

FUNDING

This study was supported by grants from the China Mega-Projects for Infectious Disease (2018ZX10711-001, 2017ZX10104-001, and 2018ZX10713-002), National Natural Science Foundation of China (82041023 and 81601997), and Beijing Natural Science Foundation (7164308). The funders had no role in the design, execution, or analysis of the study, nor in the preparation or approval of the manuscript.

ACKNOWLEDGMENTS

We thank the researchers who obtained the microarray data used in this study. We thank Yiming Zhou from Beijing Neoantigen Biotechnology Co., Ltd., for his efforts to improve the performance of codes. We thank Xianbing Yu from the Chemistry Department at the University of Chicago for his helpful discussion. We also thank Xiaoxian Cui from Shanghai Municipal Center for Disease Control & Prevention for her valuable suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00729/full#supplementary-material

REFERENCES

Andres-Terre, M., McGuire, H. M., Pouliot, Y., Bongen, E., Sweeney, T. E., Tato, C. M., et al. (2015). Integrated, multi-cohort analysis identifies conserved transcriptional signatures across multiple respiratory viruses. *Immunity* 43, 1199–1211. doi: 10.1016/j.immuni.2015.11.003

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining*, New York, NY.

Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., and Lin, C.-Y. (2014). CytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8:S11. doi: 10.1186/1752-0509-8-S4-S11

Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P.-Y., et al. (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *Lancet Infect. Dis.* 12, 687–695.

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 24, 1547–1548. doi: 10.1093/bioinformatics/btn224

Gack, M. U., Albrecht, R. A., Urano, T., Inn, K.-S., Huang, I.-C., Carnero, E., et al. (2009). Influenza A virus NS1 targets the ubiquitin ligase TRIM25 to evade recognition by the host viral RNA sensor RIG-I. *Cell Host Microb.* 5, 439–449. doi: 10.1016/j.chom.2009.04.006

Herberg, J. A., Kaforou, M., Gormley, S., Sumner, E. R., Patel, S., Jones, K. D., et al. (2013). Transcriptomic profiling in childhood H1N1/09 influenza reveals reduced expression of protein synthesis genes. *J. Infect. Dis.* 208, 1664–1668. doi: 10.1093/infdis/jit348

Herberg, J. A., Kaforou, M., Wright, V. J., Shailes, H., Eleftherohorinou, H., Hoggart, C. J., et al. (2016). Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial vs viral infection in febrile children. *JAMA* 316, 835–845. doi: 10.1001/jama.2016.11236

Ioannidis, I., McNally, B., Willette, M., Peeples, M. E., Chaussabel, D., Durbin, J. E., et al. (2012). Plasticity and virus specificity of the airway epithelial cell immune response during respiratory virus infection. *J. Virol.* 86, 5422–5436. doi: 10.1128/jvi.06757-11

Jansen, R. R., Wieringa, J., Koekkoek, S. M., Visser, C. E., Pajkrt, D., Molenkamp, R., et al. (2011). Frequent detection of respiratory viruses without symptoms: toward defining clinically relevant cutoff values. *J. Clin. Microbiol.* 49, 2631–2636. doi: 10.1128/jcm.02094-10

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 9:559. doi: 10.1186/1752-0509-8-S4-559

Mejias, A., Dimo, B., Suarez, N. M., Garcia, C., Suarez-Arrabal, M. C., Jartti, T., et al. (2013). Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection. *PLoS Med.* 10:e1001549. doi: 10.1371/journal.pmed.1001549

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Ramilo, O., Allman, W., Chung, W., Mejias, A., Ardura, M., Glaser, C., et al. (2007). Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood* 109, 2066–2077. doi: 10.1182/blood-2006-02-002477

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* 12:77. doi: 10.1186/1752-0509-8-S4-77

Robinson, M., Sweeney, T. E., Barouch-Bentov, R., Sahoo, M. K., Kalesinskas, L., Vallania, F., et al. (2019). A 20-gene set predictive of progression to severe dengue. *Cell Rep.* 26, 1104–1111. doi: 10.1016/j.celrep.2019.01.033

Sampson, D. L., Fox, B. A., Yager, T. D., Bhide, S., Cermelli, S., McHugh, L. C., et al. (2017). A Four-Biomarker blood signature discriminates systemic inflammation due to viral infection versus other etiologies. *Sci. Rep.* 7:e02325-28. doi: 10.1038/s41598-017-02325-8

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sweeney, T. E., Braviak, L., Tato, C. M., and Khatri, P. (2016a). Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. *Lancet Respirat. Med.* 4, 213–224. doi: 10.1016/s2213-2600(16)00048-5

Sweeney, T. E., Wong, H. R., and Khatri, P. (2016b). Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Sci. Transl. Med.* 8:346ra391.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613.

Tang, B. M., Shojaei, M., Parnell, G. P., Huang, S., Nalos, M., Teoh, S., et al. (2017). A novel immune biomarker IFI27 discriminates between influenza and bacteria in patients with suspected respiratory infection. *Eur. Respir. J.* 49:1602098. doi: 10.1183/13993003.02098-2016

Tsalik, E. L., Henao, R., Nichols, M., Burke, T., Ko, E. R., McClain, M. T., et al. (2016). Host gene expression classifiers diagnose acute respiratory illness etiology. *Sci. Transl. Med.* 8:322ra311.

Wang, Y., Zhu, N., Li, Y., Lu, R., Wang, H., Liu, G., et al. (2016). Metagenomic analysis of viral genetic diversity in respiratory samples from children with severe acute respiratory infection in China. *Clin. Microbiol. Infect.* 22, e451–e459. doi: 10.1016/j.cmi.2016.01.006

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, J., Peterson, D. R., Baran, A. M., Bhattacharya, S., Wylie, T. N., Falsey, A. R., et al. (2019). Host gene expression in nose and blood for the diagnosis of viral respiratory infection. *J. Infect. Dis.* 219, 1151–1161. doi: 10.1093/infdis/jiy608

Zhai, Y., Franco, L. M., Atmar, R. L., Quarles, J. M., Arden, N., Bucasas, K. L., et al. (2015). Host transcriptional response to influenza and other acute respiratory viral infections - a prospective cohort study. *PLoS Pathog.* 11:e1004869. doi: 10.1371/journal.pmed.1004869

Zhu, J., Ghosh, A., and Sarkar, S. N. (2015). OASL—a new player in controlling antiviral innate immunity. *Curr. Opin. Virol.* 12, 15–19. doi: 10.1016/j.coviro.2015.01.010

Zhu, J., Zhang, Y., Ghosh, A., Cuevas, R. A., Forero, A., Dhar, J., et al. (2014). Antiviral activity of human OASL protein is mediated by enhancing signaling of the RIG-I RNA sensor. *Immunity* 40, 936–948. doi: 10.1016/j.immuni.2014.05.007

Conflict of Interest: QX was employed by ChosenMed Technology (Beijing) Co. Ltd.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 July 2020
doi: 10.3389/fgene.2020.00827

Replication of the Association Between Keratoconus and Polymorphisms in *PNPLA2* and *MAML2* in a Han Chinese Population

Jing Zhang[1,2,3], *Yue Li*[1,2,3], *Yiqin Dai*[1,2,3] *and Jianjiang Xu*[1,2,3]*

[1] Eye Institute and Department of Ophthalmology, Eye & ENT Hospital, Fudan University, Shanghai, China, [2] NHC Key Laboratory of Myopia, Fudan University, Shanghai, China, [3] Shanghai Key Laboratory of Visual Impairment and Restoration, Shanghai, China

Keratoconus (KC) is a complex ocular disease that is affected by both genetic and non-genetic triggers. A recent genome-wide association study (GWAS) identified a genome-wide significant locus for KC in the region of *PNPLA2* (rs61876744), as well as a suggestive signal in the *MAML2* (rs10831500) locus. In order to validate their findings, here we performed a replication study of the Han Chinese population, with 120 sporadic KC cases and 206 gender and age matched control subjects, utilizing the TaqMan SNP genotyping assays. SNP rs10831500, as well as two proxy SNPs for rs61876744, named rs7942159 and rs28633403, were subjected to genotyping. However, we did not find a significant difference ($P > 0.05$) in all the three genotyped SNPs between KC cases and the controls. A further meta-analysis on four previous cohorts of white patients and this Han Chinese cohort showed a significant genetic heterogeneity within the replicated loci. Thus, the current study suggests that SNP rs61876744 (or its proxy SNPs) and rs10831500 might not be associated with KC susceptibility in this Han Chinese cohort, and a large-scale association analysis focusing on the loci is therefore warranted in further investigations.

Keywords: keratoconus, association study, Han Chinese population, SNP, replication

OPEN ACCESS

Edited by:
Yuanwei Zhang,
University of Science and Technology of China, China

Reviewed by:
Weichen Zhou,
University of Michigan, United States
Milind B. Ratnaparkhe,
ICAR Indian Institute of Soybean Research, India

***Correspondence:**
Jianjiang Xu
jianjiangxu@126.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *22 April 2020*
Accepted: *09 July 2020*
Published: *22 July 2020*

Citation:
Zhang J, Li Y, Dai Y and Xu J (2020) Replication of the Association Between Keratoconus and Polymorphisms in PNPLA2 and MAML2 in a Han Chinese Population. Front. Genet. 11:827. doi: 10.3389/fgene.2020.00827

INTRODUCTION

Keratoconus (KC) is a degenerative ocular disorder that is characterized by continuous corneal thinning and steepening, which finally causes moderate to severe visual impairment (Rabinowitz, 1998). Most of these diagnosed cases are sporadic, while a familial form of KC is also observed. The prevalence of KC has been estimated to be 1:2,000 in the general population. However, a strikingly higher incidence among Asians has been reported, and Asians are younger at presentation and require corneal grafting at an earlier age. This is suggestive of substantial influences of ethnic differences underlying this disease (Kok et al., 2012). The therapeutic intervention of KC varies heavily on the clinical stage. Contact lenses and corneal collagen UV cross-linking are major effective approaches for the management of KC at early stages, achieving biomechanical stabilization of the cornea and reducing the disease progression rate (Karolak and Gajecka, 2017).

Unfortunately, not all KC cases are recognized at early stages, and as the disease progresses, corneal transplantation is necessitated for up to 20% of KC patients. KC is therefore one of the major indications for corneal transplantation in western countries (Faria-Correia et al., 2015). This makes finding specific biomarkers that can target KC at its early stage of particular importance.

KC has a complicated etiology, with UV exposure (Arnal et al., 2011), atopy (Bawazeer et al., 2000), contact lens wear (Steahly, 1978), and constant eye rubbing (McMonnies, 2009) considered as the main behavioral and environmental risk factors for the disease. Biologically, down-regulation of collagens and structural proteins like lumican, keratocan, and decorin, as well as increased expression of catabolic enzymes were observed in KC patients, indicating the dramatic rearrangement of the corneal architecture (Sharif et al., 2018; Ferrari and Rama, 2020). Altered TGF-β signaling, which is a key regulator of extracellular matrix (ECM) secretion and assembly, was found to be involved in KC progression (Engler et al., 2011). In addition, increased oxidative stress and classic pro-inflammatory proteins including IL1, IL6, MMP9, and TNF-α were also found in KC corneas (Mas Tur et al., 2017; Vallabh et al., 2017). More importantly, an increasing body of evidence suggests a substantial genetic basis underlying KC, such as the increased probability for siblings of KC to develop the same disease (Naderan et al., 2016), the higher concordance rate in monozygotic twins compared to dizygotic twins (Tuft et al., 2012), and the observation of multi-generation pedigrees with KC (Burdon and Vincent, 2013). Many efforts have therefore been made to identify the genetic risks for KC, mainly based on approaches including linkage analyses and genome-wide association studies (GWAS). To date, single nucleotide polymorphism (SNPs) in these genes have been identified, including *CAST, RAB3GAP1, DOCK9, LOX, HGF, ZNF469, VSX1, IL1A, IL1B, WNT10A, SOD1* (De Bonis et al., 2011; Bykhovskaya et al., 2012; Czugala et al., 2012; Li et al., 2012, 2013a,b; Wang et al., 2013; Cuellar-Partida et al., 2015), and some central corneal thickness (CCT) related loci including *MDPZ-NF1B, FOXO1, FND3B, COL4A3, COL4A4,* and *COL5A* (Lu et al., 2013; Iglesias et al., 2018). Several of them were independently investigated in other ethnicities, including the Han Chinese population, whilst substantial heterogeneity remains across various ethnicities (Wang et al., 2013, 2016, 2018; Hao et al., 2015; Zhang et al., 2018).

Recently, McComish et al. performed a GWAS study of four independent cohorts of white patients with KC. Two novel loci showed genome-wide significance, rs61876744 in the *PNPLA2* gene on chr11, and rs138380 in the *CSNK1E* gene on chr22. They also reported a suggestive association signal from rs10831500, which was close to the *MAML2* gene on chr11 (McComish et al., 2019). However, given the potential genetic heterogeneity underlying KC etiology, it still remains unclear whether these newly identified SNPs are still in association with KC risk in other populations. An intensive investigation on the loci of interest, is therefore in demand. We thus conducted a replication study here to examine their roles in KC susceptibility in an independent Han Chinese cohort.

MATERIALS AND METHODS

Subjects

A total of 120 sporadic Han Chinese keratoconus cases, as well as 206 age and gender matched controls were recruited. KC cases were collected from the Department of Ophthalmology at the EENT Hospital of Fudan University from October 2015 to March 2018. They all lived in East China and were of Han Chinese ethnicity. KC cases were diagnosed based on both clinical examination and videokeratography pattern analysis, according to the following criteria: (1) at least one KC sign by slit-lamp examination (stromal thinning, Fleischer's ring, Munson's sign, and Vogt's striae); (2) an asymmetric bowtie pattern in corneal topography; refractive errors; signs of videokeratography; (3) KISA index >100; central K reading >47D. The control subjects had no ocular disease and attended the same hospital due to accidental injury. Written informed consent forms were signed by all participants. This study was performed in accordance with the declaration of Helsinki and was approved by the Ethics Committee of the EENT Hospital of Fudan University.

DNA Extraction

Genomic DNA was extracted from the monocytes in peripheral blood, with the QIAGEN FlexiGene DNA kit (Qiagen, Germany) following the standard protocol. DNA concentration was tested by a NanoDrop spectrophotometer. DNA samples were stored at −20°C before use.

SNP Genotyping

SNP rs10831500, as well as two proxy SNPs for rs61876744, named rs7942159 and rs28633403 were subjected to genotyping. The probes were designed by ThermoFisher TaqMan™ SNP genotyping Assay (Catalog nos. C__30938976_10 for rs10831500, C__11279798_10 for rs7942159, C__64236579_10 for rs28633403). The probe for SNP rs138380 failed to be designed by the custom TaqMan™ SNP genotyping Assay, and it was not further investigated here. Real-time PCR (Applied Biosystems VII, USA) was applied to complete the genotyping assay. Each reaction for the samples was prepared as 5 μL 2×SuperMix for SNP Genotyping (ThermoFisher, USA), 0.25 μL 40×probe, 2.5 μL ddH_2O, and 2 μL DNA. PCR cycling conditions were 95°C for 10 min, 45 cycles of 95°C for 15 s and 60°C for 1 min. Fluorescence data were automatically analyzed by QuantStudio™ Real-Time PCR Software (Applied

TABLE 1 | Characteristics of KC cases and controls included in this study.

Feature	Cases (n = 120)	Controls (n = 206)
Gender (female/male)	29/91	79/127
Average age (years)*	22.77 ± 5.69	26.23 ± 4.17
Age range (years)	13–45	15–33
Disease onset age (years)*	20.96 ± 5.08	NA
Visual activity*	OS: 0.61 ± 0.25	NA
	OD: 0.35 ± 0.26	

*OS, left eye; OD, right eye. *Data is shown as mean ± S.D.*

Biosystems, USA). Genotypes were classified by the ratio of the two fluorescence signals (FAM and VIC).

Data Analysis

The statistical analyses were mainly carried out by PLINK (Purcell et al., 2007). The validation of SNP frequency in cases and controls was calculated for departure from the Hardy-Weinberg equilibrium through an exact test. The allele frequency of each SNP between the cases and controls was calculated with a χ^2-test. The logistic regression model, with adjustment for gender and age, was applied to evaluate odds ratios (ORs) and their 95% confidence intervals (CIs). The linkage disequilibrium (LD) among SNPs was calculated using the LDlink package (Machiela and Chanock, 2015). A meta-analysis was performed by weighting effect size estimates using the inverse of the corresponding standard errors. The between-study heterogeneity was evaluated by the I^2-value. OR and 95% CI for the minor allele were calculated with the random effects model when $I^2 > 50\%$.

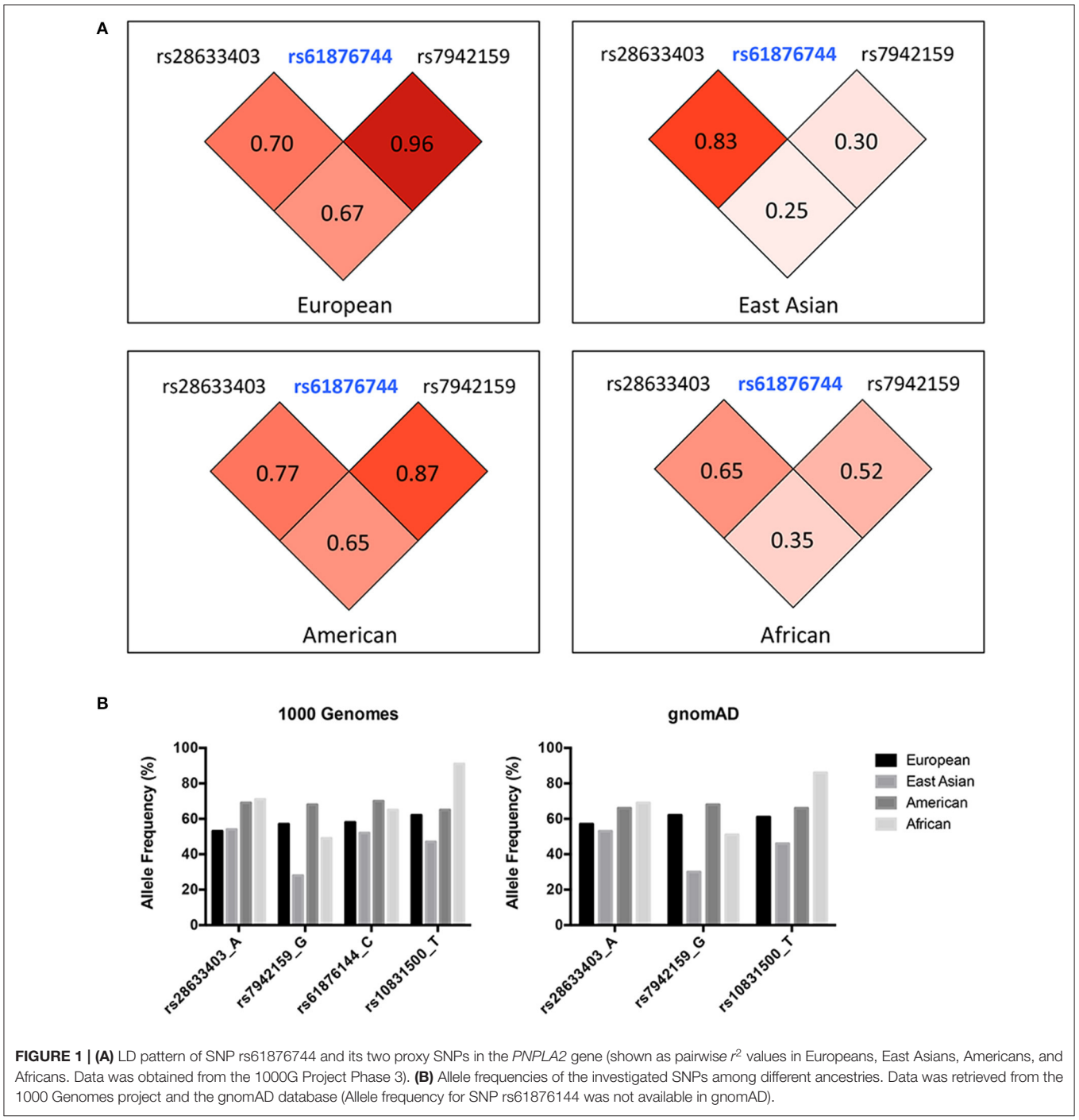


FIGURE 1 | (A) LD pattern of SNP rs61876744 and its two proxy SNPs in the *PNPLA2* gene (shown as pairwise r^2 values in Europeans, East Asians, Americans, and Africans. Data was obtained from the 1000G Project Phase 3). **(B)** Allele frequencies of the investigated SNPs among different ancestries. Data was retrieved from the 1000 Genomes project and the gnomAD database (Allele frequency for SNP rs61876144 was not available in gnomAD).

The statistical significance of SNP association was calculated by the *Z*-test. The *P*-values were transformed from the *Z*-scores and a pooled $P < 0.05$ was considered as statistically significant.

RESULTS

A total of 120 sporadic Han Chinese KC cases and 206 controls were recruited for this study. As presented in **Table 1**, KC cases showed an average age of 22.77 ± 5.69 yrs, and 75.8% of them were male. The control subjects showed an average age of 26.23 ± 4.17 yrs, and the percentage of males was 61.6%, similar to that of the case group.

Three SNPs were subjected to genotyping in our cohort. SNP rs10831500 in the *MAML2* gene was directly replicated here to investigate its association in this Han Chinese cohort. SNP rs61876744 in the *PNPLA2* gene showed the most significant association signal in the original GWAS, however, the TaqMan probe for this SNP failed to be designed, probably due to the features of flanking sequences around this SNP, and thereby its two proxy SNPs, rs28633403 (the most correlated SNP in Asians, $r^2 = 0.83$) and rs7942159 (the most correlated SNP in Europeans, $r^2 = 0.95$) were selected for further replication. The LD pattern among the three SNPs in the *PNPLA2* region, and their allele frequencies varied a lot in different ancestries (shown in **Figure 1**). Another suggestive signal in the *CSNK1E* gene, SNP rs138378, was not further replicated due to the failure of designing its custom probe for genotyping, as well as the lack of suitable proxy SNP ($r^2 > 0.8$).

We achieved an averaged genotyping call rate of 92.9% for the investigated SNPs. The two proxy SNPs for rs61876744 were in Hardy-Weinberg equilibrium in the controls, whilst SNP rs10831500 showed a slight deviation ($P = 0.02905$). Allelic association analyzed by PLINK showed that none of the SNPs were significantly in association with KC susceptibility in this Han Chinese cohort (**Table 2**). The minor allele frequency (MAF) of rs28633403 in the case group was almost comparable to that in the control group (49.4 vs. 50.0%). SNP rs7942159, the other proxy SNP for rs61876744 showed a 5% MAF difference between the cases and the controls, but did not reach nominal significance. Interestingly, its risk allele "G" had much lower frequency in Asians (Asians: 30%, Europeans: 61.5%; gnomAD data). For SNP rs10832500, its protective allele "T" in the original GWAS, presented a risk role in this Han Chinese cohort. A following genotypic association analysis was performed. However, only the genotype distribution of rs7942159 presented a borderline difference ($P = 0.06726$). The frequencies of the GG, GA, and AA genotypes of rs7942159 were found to be 8.2, 53.4, and 38.4% in the KC case group, compared to 11.1, 37.7, and 51.2% in the control group. A higher OR of 1.69 was shown when the dominant model was applied (**Table 3**).

Of note, in addition to rs10832500, SNP rs28633403, and rs7942159 were also genotyped in the original GWAS project, and the raw summary data was obtained (**Supplementary Table 1**). A meta-analysis of association results from previous four cohorts of white patients and this Han Chinese cohort was then further performed (**Figure 2**). It was found that these SNPs presented opposite trends among the included five cohorts, and substantial between-study heterogeneity was found. Therefore, the random-effects model was used here. SNP rs28633403 and rs7942159 were found to be in association with KC by meta-analysis ($P\_meta = 0.004$ and 0.04, respectively). However, their contributions to KC susceptibility remain questionable, as substantial heterogeneity existed ($I^2 > 50\%$) and their association *P*-values in 3 out of 5 cohorts were bigger than the 0.05 cutoff. SNP rs10832500

TABLE 2 | Basic association result of the genotyped SNPs in this study.

SNP	Allele	MAF_ Case %	MAF_ Control %	χ^2	*P*-value	OR (95% CI)
rs28633403	A/G	49.4	50.0	0.01639	0.8981	0.98 (0.67–1.42)
rs7942159	G/A	34.9	29.9	1.163	0.2808	1.26 (0.83–1.91)
rs10831500	T/G	51.4	45.8	1.716	0.1901	1.25 (0.89–1.74)

MAF, minor allele frequency, the minor allele of each SNP was underlined; OR, odds ratio, with respect to the minor allele; 95% CI: Lower/Upper bound of 95% confidence interval for OR; P-values and ORs were calculated after adjustment for age and gender.

TABLE 3 | Genotype frequencies of the genotyped SNPs and their association with susceptibility to KC.

SNP/group	Group frequency			*P*-value	Dominant model		Recessive model	
					OR (95% CI)	*P*-value	OR (95% CI)	*P*-value
rs28633403	AA	AG	GG	0.5483	AA&AG vs. GG		AA vs. GG&AG	
Cases	20.7%	57.3%	22.0%		1.18 (0.63–2.22)	0.5962	0.78 (0.41–1.49)	0.4556
Controls	25.0%	50.0%	25.0%					
rs7942159	GG	GA	AA	0.07729	GG&GA vs. AA		GG vs. AA&GA	
Cases	8.2%	53.4%	38.4%		1.69 (0.96–2.97)	0.06726	0.72 (0.27–1.89)	0.4981
Controls	11.1%	37.7%	51.2%					
rs10831500	TT	TG	GG	0.3785	TT&TG vs. GG		TT vs. GG&TG	
Cases	32.4%	37.8%	29.7%		1.18 (0.71–1.96)	0.5171	1.18 (0.71–1.96)	0.5171
Controls	25.0%	41.7%	33.3%					

OR, odds ratio; CI, confidence interval; P-values and ORs were calculated after adjustment for age and gender.

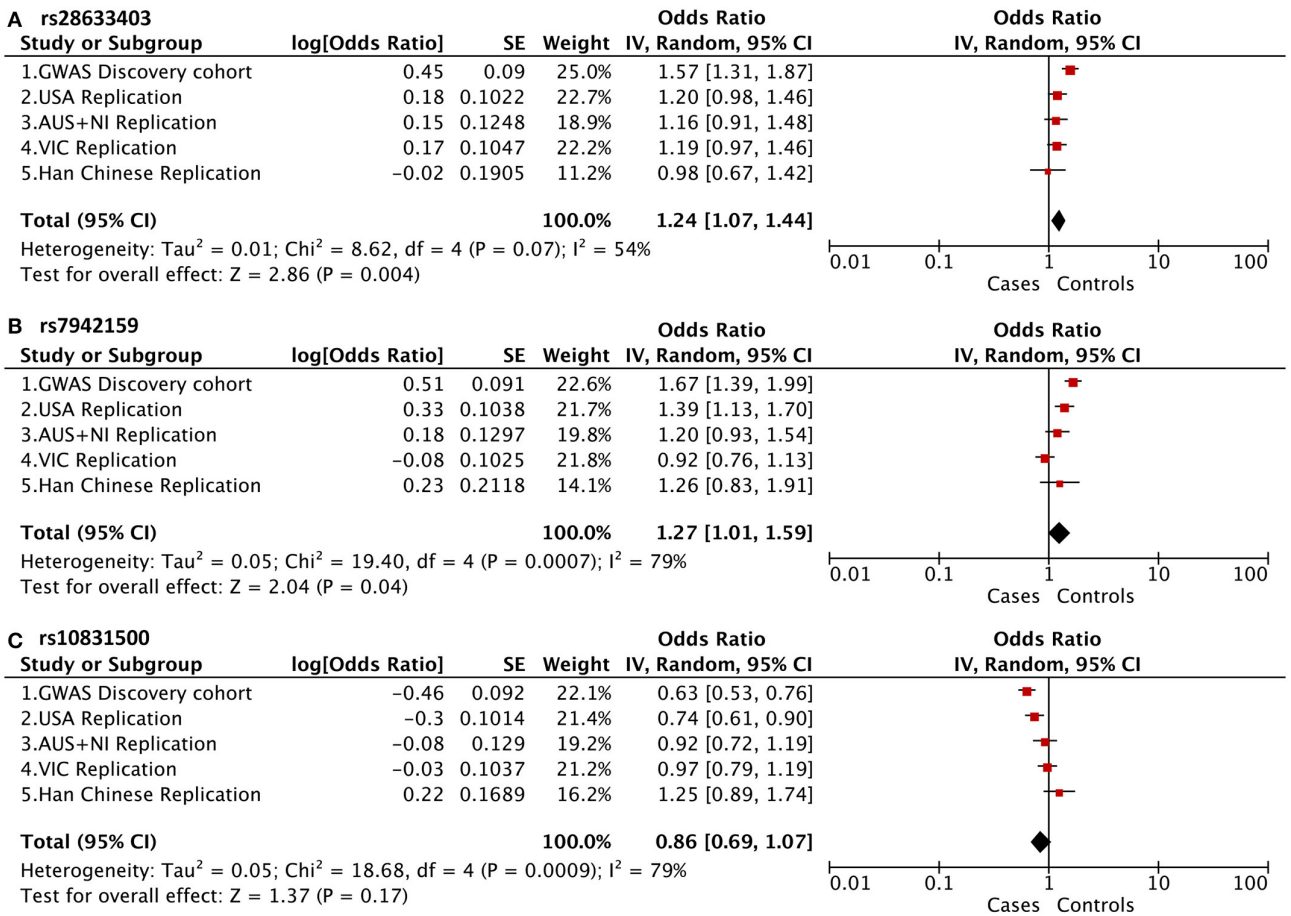


FIGURE 2 | Meta-analysis of association results from previous four cohorts of white patients and this Han Chinese cohort, shown as a forest plot of the three genotyped SNPs. **(A)** *PNPLA2* rs28633403; **(B)** *PNPLA2* rs7942159; **(C)** *MAML2* rs10831500. The size of the box is proportional to the weight of the study. Squares indicate the study-specific odds ratio (OR). Horizontal lines indicate 95% confidence interval (CI). A diamond shows the summary OR with its corresponding 95% CI. USA, United States; AUS+NI, Australia and Northern Ireland; VIC, Victoria, Australia.

did not show significant association with KC by meta-analysis. Taken together, due to the substantial heterogeneity within the replicated loci, the current study did not support the association between KC and SNPs in *PNPLA2* and *MAML2* in this Han Chinese cohort.

DISCUSSION

The etiology of KC is not well-understood, with genetic, environmental, and behavioral risk factors all contributing to the disease. Identifying the genetic risk factors for KC has proved challenging. Recently, well-powered GWAS for keratoconus and central corneal thickness have uncovered many risk loci, but most of them were performed in western populations (Burdon et al., 2011; Li et al., 2012; Lu et al., 2013; Cuellar-Partida et al., 2015; Khawaja et al., 2019; McComish et al., 2019). Some of those reported KC susceptibility loci have been further investigated in a Han Chinese cohort, including by our group (Wang et al., 2013, 2018; Hao et al., 2015; Zhang et al., 2018). However, not all these established KC-associated loci could be successfully validated, highlighting the great genetic heterogeneity underlying this complicated disease between Asians and Europeans.

Here we replicated the association of SNPs in the *PNPLA2* and *MAML2* gene with KC susceptibility in a Han Chinese cohort. We were unable to discover a remarkable difference ($P > 0.05$) in all the three genotyped SNPs between KC cases and the controls. Further meta-analysis on previous four cohorts of white patients and this Han Chinese cohort showed a significant genetic heterogeneity within the replicated loci. Thus, the current study suggested that SNP rs61876744 (or its proxy SNPs) and rs10831500 might not link with KC susceptibility in this Han Chinese cohort. Actually, based on the original GWAS, only rs61876744 was selected to represent the association signal of this locus due to its qualified *P*-value ($P < 5 \times 10^{-8}$) and the same direction of association among the four examined white cohorts. It is also possible that other SNPs within the *PNPLA2* locus may confer the risk to KC susceptibility, and thereby a large-scale association analysis on other candidate SNPs is required in further investigations.

The current study indicated great heterogeneity within the *PNPLA2* and *MAML2* region, as the I^2-values calculated by the meta-analysis for all these investigated SNPs were larger than 50%. The discrepancy between original GWAS and the meta-analysis results might come from the existence of false positive signals from GWAS, and more likely, could be explained by their substantial population differences across various ancestries. Indeed, the allele frequency (AF) of these SNPs, as well as the LD patterns within, varied a lot among different populations (**Figure 1B**). For SNP rs7942159, which was in high LD (r^2 = 0.96) with rs61876144, the lead SNP in previous GWAS in Europeans, showed a markedly reduced AF in East Asians (57 vs. 28%). Consistently, its LD (shown as r^2) with rs61876144 reduced to 0.30 in East Asians. The heterogeneity *P*-value for rs7942159 in the meta-analysis on four white cohorts and this Chinese cohort was 0.0007. The "G" allele of rs7942159 was the risk allele in both Europeans and East Asians, although the "G" allele is the minor allele in East Asians, but major allele in Europeans. Similarly, the "A" allele of rs28633403 was the risk allele for both populations, while its AF differed a lot. For SNP rs10831500 (*MAML2* locus), replication in the Han Chinese cohort and the subsequent meta-analysis did not support its association to KC susceptibility. Actually, in the original GWAS, the signal from rs10831500 was supported by the US replication cohort only. Its association *P*-values in another two white cohorts were both larger than 0.5. More interestingly, its risk allele was even contradictory in the Han Chinese cohort, making the causative role of rs10831500 to KC susceptibility questionable.

This study had several limitations that need to be noted. The primary limitation came from the relatively small sample size here, which might cause lower power and negative findings. We suggested that SNP rs28633403 and rs10831500 should not be associated to KC in Han Chinese, due to their similar allele frequencies in KC cases and controls, or the contrasting risk allele among different cohorts. However, the contribution of rs7942159 to KC risk is worth further exploration with an increased sample size, although the dominant allele differed among ethnicities. The association of other outstanding SNPs in the *PNPLA2* also needs attention. Secondly, due to the failure to design suitable probes for direct genotyping on the lead SNP in previous GWAS, two proxy SNPs for rs61876144 were genotyped instead. We speculated that the failure of designing suitable probes might be due to the features of the flanking sequences around rs61876144, as they may affect the efficiency or specificity of PCR amplification reactions. Although we have already selected the most correlated proxy SNPs for replication instead, they were not in absolute LD with the lead SNP, and this might influence the outcomes.

In conclusion, this case-control study of a Han Chinese cohort did not support the association of SNPs in the *PNPLA2* and *MAML2* gene and KC susceptibility, which was suggested by a previous GWAS report. Nevertheless, we could not fully rule out the probability that other SNPs within the loci might contribute to KC risk. Further investigations are required to explore other potential causative variants within the loci.

DATA AVAILABILITY STATEMENT

All data relevant to the study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Eye and ENT Hospital of Fudan University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JX and JZ designed the study. JZ, YL, and YD performed the experiments. JZ analyzed the data. JZ, YL, YD, and JX wrote and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The authors were sponsored by the National Natural Science Foundation of China (81700806, 81870630). The sponsor or funding organization had no role in the design or conduct of this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00827/full#supplementary-material

REFERENCES

Arnal, E., Peris-Martinez, C., Menezo, J. L., Johnsen-Soriano, S., and Romero, F. J. (2011). Oxidative stress in keratoconus? *Invest. Ophthalmol. Visual Sci.* 52, 8592–8597. doi: 10.1167/iovs.11-7732

Bawazeer, A. M., Hodge, W. G., and Lorimer, B. (2000). Atopy and keratoconus: a multivariate analysis. *Br. J. Ophthalmol.* 84, 834–836. doi: 10.1136/bjo.84.8.834

Burdon, K. P., Macgregor, S., Bykhovskaya, Y., Javadiyan, S., Li, X., Laurie, K. J., et al. (2011). Association of polymorphisms in the hepatocyte growth factor gene promoter with keratoconus. *Invest. Ophthal. Vis. Sci.* 52, 8514–8519. doi: 10.1167/iovs.11-8261

Burdon, K. P., and Vincent, A. L. (2013). Insights into keratoconus from a genetic perspective. *Clin. Exp. Optometry* 96, 146–154. doi: 10.1111/cxo.12024

Bykhovskaya, Y., Li, X., Epifantseva, I., Haritunians, T., Siscovick, D., Aldave, A., et al. (2012). Variation in the lysyl oxidase (LOX) gene is associated with keratoconus in family-based and case-control studies. *Investig. Ophthalmol. Visual Sci.* 53, 4152–4157. doi: 10.1167/iovs.11-9268

Cuellar-Partida, G., Springelkamp, H., Lucas, S. E., Yazar, S., Hewitt, A. W., Iglesias, A. I., et al. (2015). WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness. *Hum. Mol. Genet.* 24, 5060–5068. doi: 10.1093/hmg/ddv211

Czugala, M., Karolak, J. A., Nowak, D. M., Polakowski, P., Pitarque, J., Molinari, A., et al. (2012). Novel mutation and three other sequence variants segregating

with phenotype at keratoconus 13q32 susceptibility locus. *Eur. J. Hum. Genet.* 20, 389–397. doi: 10.1038/ejhg.2011.203

De Bonis, P., Laborante, A., Pizzicoli, C., Stallone, R., Barbano, R., Longo, C., et al. (2011). Mutational screening of VSX1, SPARC, SOD1, LOX, and TIMP3 in keratoconus. *Mol. Vision* 17, 2482–2494.

Engler, C., Chakravarti, S., Doyle, J., Eberhart, C. G., Meng, H., Stark, W. J., et al. (2011). Transforming growth factor-beta signaling pathway activation in Keratoconus. *Am. J. Ophthalmol.* 151, 752–9 e2. doi: 10.1016/j.ajo.2010.11.008

Faria-Correia, F., Luz, A., and Ambrosio, R. (2015). Managing corneal ectasia prior to keratoplasty. *Expert. Rev. Ophthalmol.* 10, 33–48. doi: 10.1586/17469899.2015.991390

Ferrari, G., and Rama, P. (2020). The keratoconus enigma: a review with emphasis on pathogenesis. *Ocular Surf.* 18, 363–373. doi: 10.1016/j.jtos.2020.03.006

Hao, X. D., Chen, P., Chen, Z. L., Li, S. X., and Wang, Y. (2015). Evaluating the Association between Keratoconus and Reported Genetic Loci in a Han Chinese Population. *Ophthalmic genetics* 36, 132–136. doi: 10.3109/13816810.2015.1005317

Iglesias, A. I., Mishra, A., Vitart, V., Bykhovskaya, Y., Hohn, R., Springelkamp, H., et al. (2018). Cross-ancestry genome-wide association analysis of corneal thickness strengthens link between complex and Mendelian eye diseases. *Nat. Commun.* 9:1864. doi: 10.1038/s41467-018-03646-6

Karolak, J. A., and Gajecka, M. (2017). Genomic strategies to understand causes of keratoconus. *Mol. Genet. Genom.* 292, 251–269. doi: 10.1007/s00438-016-1283-z

Khawaja, A. P., Rojas Lopez, K. E., Hardcastle, A. J., Hammond, C. J., Liskova, P., Davidson, A. E., et al. (2019). Genetic variants associated with corneal biomechanical properties and potentially conferring susceptibility to keratoconus in a genome-wide association study. *JAMA Ophthal.* 137, 1005–1012. doi: 10.1001/jamaophthalmol.2019.2058

Kok, Y. O., Tan, G. F., and Loon, S. C. (2012). Review: keratoconus in Asia. *Cornea* 31, 581–93. doi: 10.1097/ICO.0b013e31820cd61d

Li, X., Bykhovskaya, Y., Canedo, A. L., Haritunians, T., Siscovick, D., Aldave, A. J., et al. (2013b). Genetic association of COL5A1 variants in keratoconus patients suggests a complex connection between corneal thinning and keratoconus. *Investig. Ophthalmol. Visual Sci.* 54, 2696–2704. doi: 10.1167/iovs.13-11601

Li, X., Bykhovskaya, Y., Haritunians, T., Siscovick, D., Aldave, A., Szczotka-Flynn, L., et al. (2012). A genome-wide association study identifies a potential novel gene locus for keratoconus, one of the commonest causes for corneal transplantation in developed countries. *Hum. Mol. Genet.* 21, 421–429. doi: 10.1093/hmg/ddr460

Li, X., Bykhovskaya, Y., Tang, Y. G., Picornell, Y., Haritunians, T., Aldave, A. J., et al. (2013a). An association between the calpastatin (CAST) gene and keratoconus. *Cornea* 32, 696–701. doi: 10.1097/ICO.0b013e3182821c1c

Lu, Y., Vitart, V., Burdon, K. P., Khor, C. C., Bykhovskaya, Y., Mirshahi, A., et al. (2013). Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nat. Genet.* 45, 155–163. doi: 10.1038/ng.2506

Machiela, M. J., and Chanock, S. J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 31, 3555–3557. doi: 10.1093/bioinformatics/btv402

Mas Tur, V., MacGregor, C., Jayaswal, R., O'Brart, D., and Maycock, N. (2017). A review of keratoconus: diagnosis, pathophysiology, and genetics. *Survey Ophthalmol.* 62, 770–783. doi: 10.1016/j.survophthal.2017.06.009

McComish, B. J., Sahebjada, S., Bykhovskaya, Y., Willoughby, C. E., Richardson, A. J., Tenen, A., et al. (2019). Association of genetic variation with keratoconus. *JAMA Ophthalmol.* 138, 174–181. doi: 10.1001/jamaophthalmol.2019.5293

McMonnies, C. W. (2009). Mechanisms of rubbing-related corneal trauma in keratoconus. *Cornea* 28, 607–615. doi: 10.1097/ICO.0b013e318198384f

Naderan, M., Rajabi, M. T., Zarrinbakhsh, P., Naderan, M., and Bakhshi, A. (2016). Association between family history and keratoconus severity. *Curr. Eye Res.* 41, 1414–1418. doi: 10.3109/02713683.2015.1128553

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Rabinowitz, Y. S. (1998). Survey of ophthalmology. *Keratoconus* 42, 297–319. doi: 10.1016/S0039-6257(97)00119-7

Sharif, R., Bak-Nielsen, S., Hjortdal, J., and Karamichos, D. (2018). Pathogenesis of Keratoconus: the intriguing therapeutic potential of Prolactin-inducible protein. *Progress Retinal Eye Res.* 67, 150–167. doi: 10.1016/j.preteyeres.2018.05.002

Steahly, L. P. (1978). Keratoconus following contact lens wear. *Ann. Ophthalmol.* 10, 1177–1179.

Tuft, S. J., Hassan, H., George, S., Frazer, D. G., Willoughby, C. E., and Liskova, P. (2012). Keratoconus in 18 pairs of twins. *Acta Ophthalmol.* 90, e482–e486. doi: 10.1111/j.1755-3768.2012.02448.x

Vallabh, N. A., Romano, V., and Willoughby, C. E. (2017). Mitochondrial dysfunction and oxidative stress in corneal disease. *Mitochondrion* 36, 103–113. doi: 10.1016/j.mito.2017.05.009

Wang, Y., Jin, T., Zhang, X., Wei, W., Cui, Y., Geng, T., et al. (2013). Common single nucleotide polymorphisms and keratoconus in the Han Chinese population. *Ophthalmic Genet.* 34, 160–166. doi: 10.3109/13816810.2012.743569

Wang, Y., Wei, W., Zhang, C., Zhang, X., Liu, M., Zhu, X., et al. (2016). Association of Interleukin-1 gene single nucleotide polymorphisms with keratoconus in Chinese Han Population. *Curr. Eye Res.* 41, 630–635. doi: 10.3109/02713683.2015.1045083

Wang, Y. M., Ma, L., Lu, S. Y., Chan, T. C. Y., Yam, J. C. S., Tang, S. M., et al. (2018). Analysis of multiple genetic loci reveals MPDZ-NF1B rs1324183 as a putative genetic marker for keratoconus. *Br. J. Ophthalmol.* 102, 1736–1741. doi: 10.1136/bjophthalmol-2018-312218

Zhang, J., Wu, D., Li, Y., Fan, Y., Chen, H., and Xu, J. (2018). Evaluating the association between calpastatin (CAST) gene and keratoconus in the Han Chinese population. *Gene* 653, 10–13. doi: 10.1016/j.gene.2018.02.016

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 July 2020
doi: 10.3389/fgene.2020.00796

Genome-Wide Profiling of Alternative Splicing Signature Reveals Prognostic Predictor for Esophageal Carcinoma

Jian-Rong Sun[1,2†], Chen-Fan Kong[1,3†], Yan-Ni Lou[2], Ran Yu[1,2], Xiang-Ke Qu[1,4] and Li-Qun Jia[2]*

[1] Graduate School, Beijing University of Chinese Medicine, Beijing, China, [2] Oncology Department of Integrated Traditional Chinese and Western Medicine, China-Japan Friendship Hospital, Beijing, China, [3] Gastroenterology Department, Beijing University of Chinese Medicine Affiliated Dongzhimen Hospital, Beijing, China, [4] Rheumatism Department of Traditional Chinese Medicine, China-Japan Friendship Hospital, Beijing, China

Background: Alternative splicing (AS) is a molecular event that drives protein diversity through the generation of multiple mRNA isoforms. Growing evidence demonstrates that dysregulation of AS is associated with tumorigenesis. However, an integrated analysis in identifying the AS biomarkers attributed to esophageal carcinoma (ESCA) is largely unexplored.

Methods: AS percent-splice-in (PSI) data were obtained from the TCGA SpliceSeq database. Univariate and multivariate Cox regression analysis was successively performed to identify the overall survival (OS)-associated AS events, followed by the construction of AS predictor through different splicing patterns. Then, a nomogram that combines the final AS predictor and clinicopathological characteristics was established. Finally, a splicing regulatory network was created according to the correlation between the AS events and the splicing factors (SF).

Results: We identified a total of 2389 AS events with the potential to be used as prognostic markers that are associated with the OS of ESCA patients. Based on splicing patterns, we then built eight AS predictors that are highly capable in distinguishing high- and low-risk patients, and in predicting ESCA prognosis. Notably, the area under curve (AUC) value for the exon skip (ES) prognostic predictor was shown to reach a score of 0.885, indicating that ES has the highest prediction strength in predicting ESCA prognosis. In addition, a nomogram that comprises the pathological stage and risk group was shown to be highly efficient in predicting the survival possibility of ESCA patients. Lastly, the splicing correlation network analysis revealed the opposite roles of splicing factors (SFs) in ESCA.

Conclusion: In this study, the AS events may provide reliable biomarkers for the prognosis of ESCA. The splicing correlation networks could provide new insights in the identification of potential regulatory mechanisms during the ESCA development.

Keywords: esophageal carcinoma, alternative splicing, survival, prognosis, splicing factor

OPEN ACCESS

Edited by:
Xin Maizie Zhou,
Stanford University, United States

Reviewed by:
Yanqing Liu,
Columbia University, United States
Meng Xu,
National Institutes of Health (NIH), United States

****Correspondence:***
Li-Qun Jia
liqun-jia@hotmail.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *11 February 2020*
Accepted: *03 July 2020*
Published: *22 July 2020*

Citation:
Sun J-R, Kong C-F, Lou Y-N, Yu R, Qu X-K and Jia L-Q (2020) Genome-Wide Profiling of Alternative Splicing Signature Reveals Prognostic Predictor for Esophageal Carcinoma. Front. Genet. 11:796. doi: 10.3389/fgene.2020.00796

INTRODUCTION

Being the seventh most frequently occurring tumor in humans, esophageal carcinoma (ESCA) ranks the sixth in causing fatalities worldwide. In year 2018 alone, the number of new ESCA cases and ESCA-related deaths was estimated to be 572,034 and 508,585, respectively (Bray et al., 2018). Although the development of early diagnosis and treatment approaches for ESCA have seen much improvement in recent years, the five-year survival rate of 15–20% is unsatisfactory (Pennathur et al., 2013). Due to the high morbidity and mortality rates of ESCA, there is an urgent call for the development of a highly efficient prognostic method. Over the past few decades, a great deal of effort has been made to identify prognostic biomarkers and therapeutic targets for ESCA. Although the studies showed some promising results, the research only focused on aspects such as mutation-driving factors and transcriptional levels (Zhu J. et al., 2018), thereby neglecting the diversity of RNA isoforms driven by post-translational modifications.

Alternative splicing (AS) is a crucial molecular mechanism by which mRNA is spliced into different RNA transcripts in order to be translated into diverse protein products (Tress et al., 2017). Recent studies showed that AS modifies about 94% of all human genes and plays an important role in the biological process (Matera and Wang, 2014; Oltean and Bates, 2014). Dysregulation of AS is associated with manifold pathological processes, including cancers where it promotes cancer development by causing the loss-of-function in tumor suppressors or the activation of oncogenes and cancer pathways. A recent study has shown multiple AS events participated in carcinogenesis, including proliferation, angiogenesis, invasion and metastasis (Mao et al., 2019). Tumor cells often tend to generate isoform switches where the variants produced are utilized to promote cell growth, drug resistance, invasion, immune escape and metastasis (Chen and Weiss, 2015; Climente-Gonzalez et al., 2017; Kim et al., 2018). For example, ZAK has two isoforms, namely ZAKα and ZAKβ (Lee et al., 2018), that play an opposite role in cancer development. Whilst ZAKα exerts an anti-neoplastic effect, ZAKβ exhibits an anti-proliferation feature. In BRCA2, one of the splicing variants BRCA2-Δ3 (Gelli et al., 2019), has been shown to be associated with a high risk of developing breast or ovarian cancer (Muller et al., 2011; Caputo et al., 2018). CXCR3 is another tumor-related gene in humans with three different splice variants: CXCR3A, CXCR3B, and CXCR3-alt. Recent studies have shown that the CXCR3 protein level is often heightened in tumor tissues than that of adjacent tissues. A high expression of CXCR3 is usually associated with adverse prognosis in cancer patients. Other studies have found that the CXCR3A variant promotes tumor cell growth while the CXCR3B variant induces tumor cell apoptosis (Ruytinx et al., 2018).

In addition, splicing factors have been shown to play a role in regulating tissue- or cell-type-specific AS (Tripathi et al., 2010), Alterations in the expression and activity of critical splicing factors can cause a string of changes to the AS, which then jointly promote tumor cell growth and survival (Ladomery, 2013). Therefore, an integrated analysis of AS events is needed in order to dissect the molecular mechanisms of ESCA and to identify potential prognostic markers for cancer.

With the continuous development of genome-wide sequencing technologies in recent years, it is now possible to identify cancer-specific molecules and prognostic biomarkers for patients (Griffith et al., 2010; Katz et al., 2010). Although systematic analysis of prognostic AS signature in liver cancer, lung cancer, head and neck cancer, and breast cancer has been reported (Suo et al., 2015; Li Y. et al., 2017; Liang et al., 2019; Wu et al., 2019), the AS signature in ESCA is largely unknown.

In the current study, we revealed numerous AS events connected with the overall survival (OS) of ESCA patients through an integrated profiling for the genome-wide AS events in the ESCA cohort from TCGA SpliceSeq. Based on the AS events identified, we constructed prognostic predictors. Then, we presented an AS-clinicopathologic nomogram which is useful in predicting the survival probability for ESCA patients. Finally, we established an SF-AS correlation network to demonstrate the underlying regulation mechanism for ESCA prognosis.

MATERIALS AND METHODS

The flowchart of the current study was presented in **Figure 1A**.

Data Acquisition

The RNA-seq data and clinical information of the TCGA ESCA cohort were obtained from the TCGA data portal[1]; while the Percent-splice-in (PSI) data of AS events for ESCA were obtained from the TCGA SpliceSep[2], a data portal that provides AS profiles across 33 tumors based on the TCGA RNA-seq data. There are seven types of AS events (**Figure 1B**) identified to date, namely Alternate Acceptor site (AA), Alternate Terminator (AT), Mutually Exclusive Exons (ME), Retained Intron (RI), Alternate Donor site (AD), Alternate Promoter (AP), and Exon Skip (ES) (Ryan et al., 2016). PSI values ranging from zero to one were used to quantify the AS events. Thus, to obtain a reliable set of AS events, we set a strict screening filter so that the percentage of samples containing PSI values exceeds 75%.

The AS events were annotated by combining the splicing type, ID number in the SpliceSeq and the corresponding parent gene symbol. For example, in "ERBB2| 99888| ES", ERBB2 denotes

Abbreviations: AA, alternate acceptor site; AD, alternate donor site; AP, alternate promoter; AS, alternative splicing; AT, alternate terminator; AUC, area under curve; DCA, decision curve analysis; ES, exon skip; ESCA, esophageal carcinoma; HR, hazard ratio; ME, mutually exclusive exons; OS, overall survival; PSI, Percent Spliced In; RI, retained intron; RNA-seq, RNA sequencing; ROC, receiver operating characteristic; SFs, splicing factors; TCGA, The Cancer Genome Atlas.

[1]https://portal.gdc.cancer.gov, version 18.0

[2]https://bioinformatics.mdanderson.org/TCGASpliceSeq/

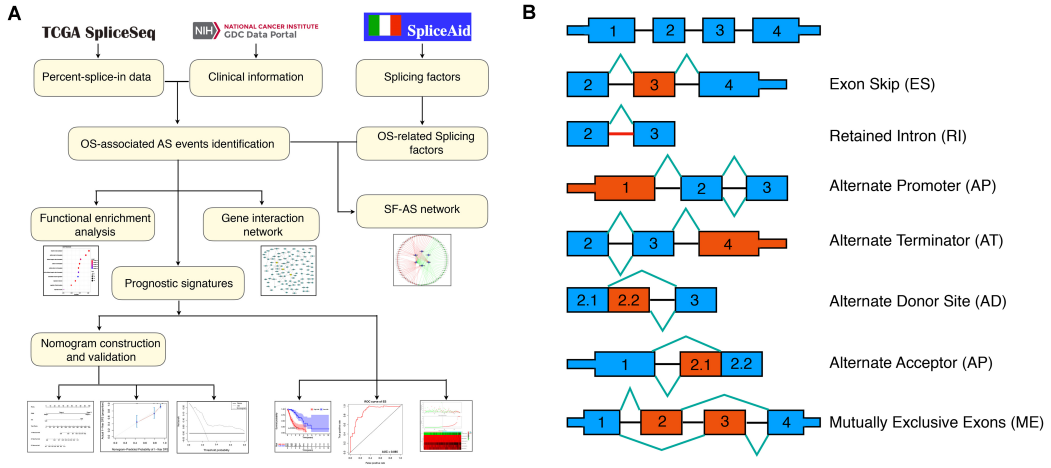


FIGURE 1 | Flowchart of the present study and splicing pattern of AS events. **(A)** Study flowchart. **(B)** Illustrations for seven types of AS events, including exon skip (ES), retained intron (RI), alternate promoter (AP), alternate terminator (AT), alternate donor site (AD), alternate promoter (AP), mutually exclusive exons (ME).

the corresponding parent gene name, 99888 represents the ID of splicing variant and ES indicates the splicing type.

Survival Analysis of AS Events, Gene Interaction Network, Functional, and Pathway Enrichment Analysis

The clinical information of ESCA patients was downloaded from the TCGA database. Based on the median PSI values, the patients were divided into two subgroups (high- and low-PSI). Univariate Cox regression analysis was conducted to detect the association between the alternative splicing (AS) events and the overall survival (OS) of ESCA patients, with $P < 0.05$ being considered significant. UpSetR (version 1.4.0) was used to create Upset plots in order to analyze the intersections of all seven types of OS-associated AS events in ESCA (Lex et al., 2014). Subsequently, the corresponding parent genes of OS-associated AS events were selected to construct a gene interaction network using Reactome FI plugin in Cytoscape (version 3.7.1), and the key genes in the network were identified using CentiScaPe2.2 plugin in Cytoscape (version 3.7.1). Functional enrichment analysis was performed by Database for Annotation, Visualization and Integrated Discovery (DAVID) online functional annotation tool[3] using the parent genes (Dennis et al., 2003). Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways with $P < 0.05$ were considered statistically significant. Then, the significant pathways in KEGG and the top 10 terms in each GO category, namely containing cellular (CC), molecular function (MF), and biological process (BP) were visualized by ggplot2 package in R (version 3.3.0).

Construction of the Prognostic Predictor for ESCA Patients

Firstly, Lasso regression analysis was performed for OS-associated AS events in each splicing type in order to screen for candidates in subsequent analysis and to avoid model over-fitting. Secondly, the screened AS events were used in multivariate Cox regression analysis to construct the prognostic predictor (McNeish, 2015). Meanwhile, considering that all seven AS types have differences in their individual mode of action that is independent from each other in post-transcriptional modification, the screened AS events in each splicing type above were consolidated to construct another prognostic predictor. Then, the risk scores were computed based on each prognostic predictor and the formula used for calculating the risk score for each patient is as follows: $\text{Risk score} = \beta_{\text{AS event1}} \times \text{PSI}_{\text{AS event1}} + \beta_{\text{AS event2}} \times \text{PSI}_{\text{AS event2}} + \cdots + \beta_{\text{AS eventn}} \times \text{PSI}_{\text{AS eventn}}$. The patients were divided into two subgroups (high- and low-risk) according to the median risk score in order to perform Kaplan-Meier test for estimating the predictive accuracy of each prognostic predictor. The predictive accuracy of each prognostic predictor was assessed by computing the area under the curve (AUC) value at 3 years of the Receiver operating characteristic (ROC) curve by the survival ROC package (version 1.0.3). Since fewer events occurred after 5 years (see Kaplan-Meier curves), the dynamic AUC value from 1 to 5 years was calculated by time ROC package (version 0.4) in order to obtain an optimal signature. Besides, the mutations of parent genes in final signature were analyzed using maftools package in R (version 3.10).

Finally, stratified Cox survival analysis was performed to verify the independent prognostic power of the final signature in ESCA cohort such as age, gender, pathological stage and tumor grade.

Development and Validation of an AS-Clinicopathologic Nomogram

In order to detect whether the prognostic predictor along with all clinical variables described above was associated with the OS of ESCA patients, Univariate Cox regression analysis was performed. Subsequently, the OS-related variables were used for multivariate Cox regression analysis to screen for independent

[3]https://david.ncifcrf.gov/, version 6.8

prognostic factors and to develop a nomogram model that can better predict the survival probability of patients. Subsequently, to make sure that the results obtained were reliable, the nomogram model was validated by the Bootstrap method with the resample number set as 1000. The calibration curves were used to assess the predictive ability of the nomogram and the C-statistic were calculated to evaluate the discriminative ability using Hmisc package in R (version 4.1.1). A calibration curve close to 45° is an indication of good prediction ability of the model constructed by this factor. To verify clinical application of the nomogram, the decision curve analysis (DCA) was conducted using stdca package[4].

Construction of Underlying SF-AS Correlation Network

Splicing factors (SFs) were retrieved from the SpliceAid 2 database (Piva et al., 2012). The mRNA expression data of SFs were obtained from the TCGA database and normalized using the trimmed mean method of *M*-values (TMM) from edgeR package in R (version 3.6.0). Univariate Cox regression analysis was performed to screen the OS-associated SFs. Then, the Spearman correlation analysis was performed between the PSI values of OS-associated AS events and the expression level of OS-associated SFs, with $P < 0.05$ being set as a cut-off value. Finally, Cytoscape (version 3.7.1) was used to generate an underlying SF-AS correlation network among the significant result of spearman correlation analysis, with the correlation coefficient greater than 0.5.

RESULTS

Integrated AS Events Profiles in TCGA ESCA Cohort

Within the integrated AS events profiles of 185 ESCA patients from TCGA SpliceSeq, we detected a total of 50342 AS events in 10766 genes, which included 20843 ESs in 7174 genes, 10033 APs in 4046 genes, 8448 ATs in 3690 genes, 4145 AAs in 2871 genes, 3590 ADs in 2463 genes, 3038 RIs in 2001 genes, and 245 MEs in 237 genes (**Figure 2A**). The results showed that, among the seven types of AS events, ES was the main splicing pattern while ME was the least frequent event in ESCA patients.

Detection and Functional Enrichment Analysis of OS-Associated AS Events

The clinical information of ESCA patients was downloaded from the TCGA database. A total of 185 ESCA patients with fully characterized tumors were included in the analysis. The demographic and clinical characteristics of patients are provided in Supplementary (**Supplementary Table S1**).

Using the AS events profiles in the ESCA cohort, we identified 2389 AS events which were significantly associated with the OS of ESCA patients ($P < 0.05$) by univariate Cox regression analysis. In particular, we found one gene with potentially more than one AS events that were significantly connected with patient survival. In order to better visualize intersecting sets, an UpSet plot was created as shown in **Figure 2B**. Interestingly, our analysis revealed that one gene can exhibit up to four types of AS events that were all found to be significantly associated with patient survival. Specifically, ES, AA, AD, and RI of *CIRBP* were all significantly linked to the OS of patients. The distribution of top 20 AS events in different splicing type presented in **Figure 3** clearly showed that, the majority of AS event was related to good prognosis. Furthermore, all parent genes of OS-associated AS events were used in functional and pathway enrichment analysis. A total of 74 Gene Ontology (GO) terms and 15 Kyoto Encyclopedia of Genes and Genomes (KEGG) terms were identified significantly in the analysis ($P < 0.05$). The top pathways of GO and KEGG enrichment were shown in **Figures 4A–D**.

In order to dissect the biological relationships between the corresponding parent genes of OS-associated AS events in ESCA, a gene interaction network was created using Cytoscape. Our results revealed, three vital hub genes in the network, namely *SIN3A, YWHAZ*, and *RPA3* (**Figure 4E**), which may be closely related to the development of ESCA.

Construction of the Prognostic Predictor for ESCA Patients

To avoid model over-fitting, the significant OS-associated AS events ($P < 0.05$) in each AS type were analyzed by lasso regression (**Supplementary Figure S1**), and the results were selected to perform multivariate Cox regression analysis, respectively. Meanwhile, the AS events screened above in each splicing type were amalgamated to fit another multivariate Cox regression. Finally, a total of eight AS models were constructed, namely AA, AT, ME, RI, AD, AP, ES, and ALL models. The specific formulas of each model shown in **Table 1** were used to compute the risk score of each patient, which were then divided into high- and low-risk subgroups according to the median of risk scores. Kaplan-Meier survival analysis of each model was considerably efficient in distinguishing good or poor outcome between the two subgroups (**Figures 5A–H**). To compare the level of efficiency among different AS models, ROC curves were created with the AUC values calculated at 3 years survival, respectively (**Figures 6A–H**). The AUC value of ROC for the ES prognostic predictor was calculated to be 0.885, which remained higher than other AS models over time, suggesting that ES has a higher level of efficiency than other prognostic predictors (**Figure 7A**). The distribution of patients' survival status, risk score and AS events for the ES prognostic predictors as illustrated in **Figure 7B** showed that, the risk score increased as the patient's survival time decreased, which resulted in a significant increase ($P < 0.05$) in the number of deaths (red dots in the upper part of **Figure 7B**). The corresponding parent genes of AS events included in the ES prognostic predictor were shown in **Table 2**. Moreover, among these seven parent genes, ERBB2 and C19orf82 possessed the most frequent genetic mutation and the missense mutation was the most common alteration (**Figure 8A**). The mutant of ERBB2 and C19orf82 also indicated a significantly shorter OS time than the wild type (**Figures 8B,C**).

[4]https://www.mskcc.org/

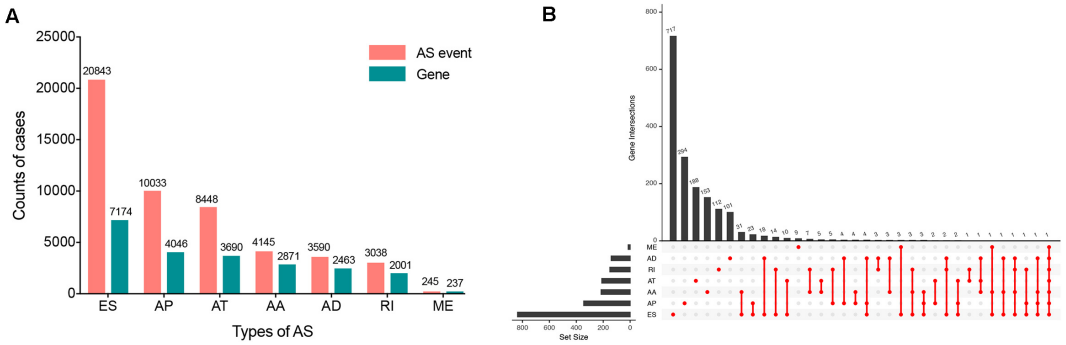


FIGURE 2 | Overview of seven types of AS in this study. **(A)** Number of AS events and related genes in ESCA. **(B)** UpSet plots in ESCA, showing the interactions among the seven types of OS-associated AS events. One gene may have up to seven types of AS events.

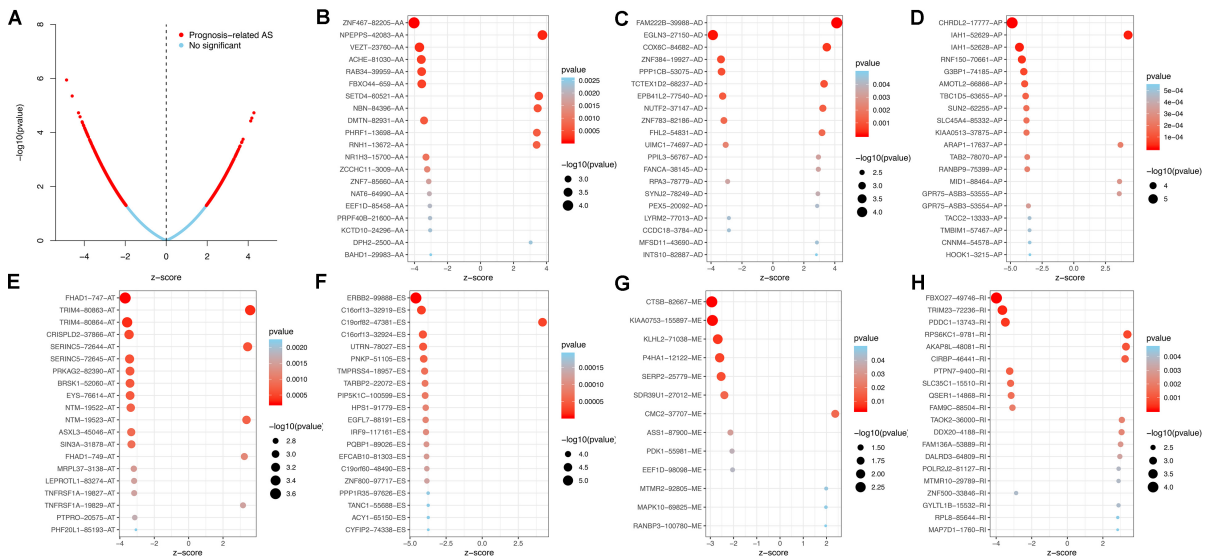


FIGURE 3 | Top 20 most significant AS events in ESCA. **(A)** The volcano plots of prognosis-connected AS events. The top 20 AS events associated with survival outcome for ESCA in different splice patterns, including **(B)** AA, alternate acceptor site. **(C)** AD, alternate donor site. **(D)** AP, alternate promoter. **(E)** AT, alternate terminator. **(F)** ES, exon skip. **(G)** ME, mutually exclusive exons. **(H)** RI, retained intron.

Furthermore, to verify the prognostic value of the final predictor, we performed Cox survival analysis in stratified ESCA cohort where the patients were classified by clinicopathological characteristics, including age, gender, tumor grade and different pathological stages, such as T stage, M stage, and N stage. The results clearly showed that the high-risk group had a worse prognosis than that of the low-risk group in almost all cohorts (**Table 3**). Taken together, our results showed that the final predictor can maintain its efficiency to precisely identify patients with adverse prognosis, regardless of clinical parameters.

Development and Efficiency of AS-Clinicopathologic Nomogram

To screen for potential factors correlated with the OS of ESCA patients, the risk level (high or low) based on the ES prognostic predictor along with clinicopathologic variables mentioned earlier were studied by univariate Cox analysis. The results showed that tumor grade, pathological stage and risk score level were statistically significant ($P < 0.05$) (**Table 4**). Multivariate Cox regression analysis revealed that the risk score level derived from the ES prognostic predictor and the pathological stage were the only independent prognostic factors associated with the OS of ESCA patients (**Table 4**). These independent prognostic factors were used in the construction of subsequent nomograms (**Figure 9A**). The calibration curve of the nomogram for the probability of survival at 1, 3, 5 years showed good uniformity between prediction and actual observation (**Figures 9B–D**). The C-statistic for OS prediction of ESCA patients was 0.78, indicating that the predictive ability of this nomogram model was efficient. The DCA of this nomogram for 1, 3, 5 years as shown in **Figures 9E–G** demonstrated that this nomogram had good clinical usefulness, which meant that if the threshold probability was less than 80%, using this nomogram to predict prognosis in

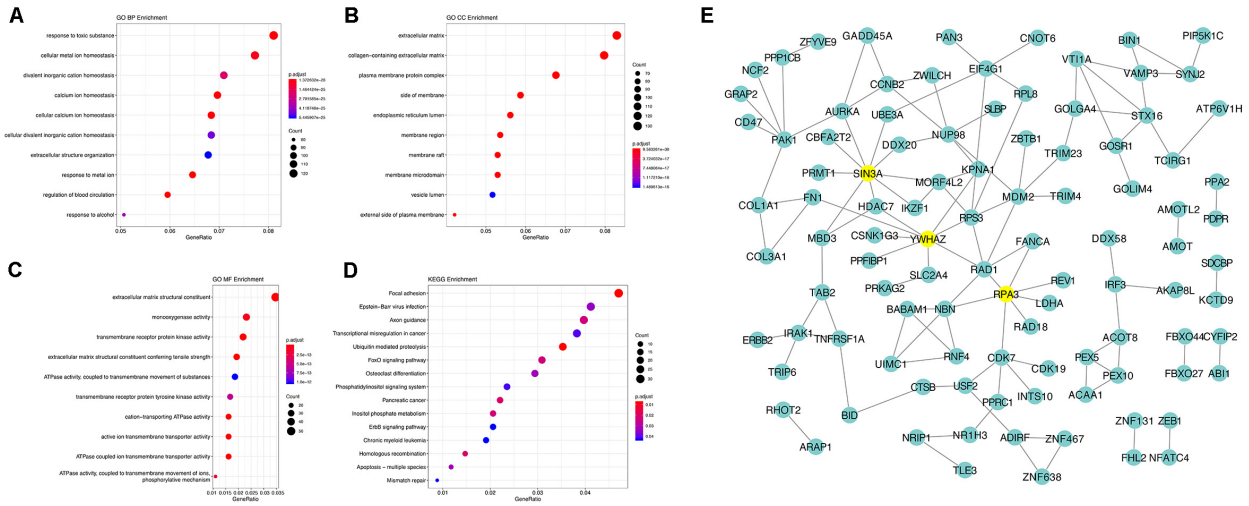


FIGURE 4 | Gene interaction network and functional analysis of OS-associated alternative splicing events in ESCA. **(A)** biological processes (BP). **(B)** cellular component (CC). **(C)** molecular function (MF). **(D)** KEGG pathway analysis. **(E)** Gene interaction network of corresponding parent genes of OS-associated AS events generated by Cytoscape.

TABLE 1 | Formula of prognostic signature for esophageal carcinoma.

Type	Formula
AA	ZNF467 |82205| AA × (−8.77) + NPEPPS |42083| AA × 3.59 + VEZT |23760| AA × (−9.88) + RAB34 |39959| AA × (−35.9) + FBXO44 |659| AA × (−4.92) + SETD4 |60521| AA × 1.69 + NR1H3 |15700| AA × (−2.05) + ZCCHC11 |3009| AA × (−6.17)
AD	FAM222B |39988| AD × 4.77 + EGLN3 |27150| AD × (−11.07) + COX6C |84682| AD × 3.29 + ZNF384 |19927| AD × (−5.16) + ZNF783 |82186| AD × (−5.33) + UIMC1 |74697| AD × (−4.84) +FANCA |38145| AD × (6.74) + RPA3 |78779| AD × (−31.74) + LYRM2 |77013| AD × (−5.92)
AP	CHRDL2 |17777| AP × (−7.74) + IAH1 |52629| AP × 8.73 + RNF150 |70661| AP × (−8.97) + RANBP9 |75399| AP × (−21.7) +GPR75 − ASB3 |53555| AP × (3.44) + TACC2 |13333| AP × (−3.86)+HOOK1 |3215| AP × (−13.94)
AT	FHAD1 |747| AT × (−1.51) + TRIM4 |80863| AT × (3.28) + BRSK1 |52060| AT × (−8.77) + EYS |76614| AT × (−8.45) + NTM |19522| AT × (−3.88) +MRPL37 |3138| AT × (−9.51) + LEPROTL1 |83274| AT × (−6.95) + PTPRO |20575| AT × (−3.16)
ES	ERBB2 |99888| ES × (−25.65) + C19orf82 |47381| ES × 3.35 + C16orf13 |32924| ES × (−4.68) +UTRN |78027| ES × (−2.93) + TMPRSS4 |18957| ES × (−11.68) + HPS1 |91779| ES × (−5.21)+FCAB10 |81303| ES × (−10.77)
ME	CTSB |82667| ME × (−15.83) + KIAA0753 |155897| ME × (−3.57) + KLHL2 |71038| ME × (−1.82) + P4HA1 |12122| ME × (−3.46) + CMC2 |37707| ME × (1.91) + EEF1D |98098| ME × (−1.03)+MTMR2 |92805| ME × 2.43 + MAPK10 |69825| ME × 2.55
RI	FBXO27 |49746| RI × (−18.33) + TRIM23 |72236| RI × (−28.37) + PDDC1 |13743| RI × (−20.12) + AKAP8L |48081| RI × 4.51 + PTPN7 |9400| RI × (−25.60) + SLC35C1 |15510| RI × (−14.10)+POLR2J2 |81127| RI × (2.55)
ALL	CHRDL2 |17777| AP × (−7.19) + ERBB2 |99888| ES × (−34.28) + IAH1 |52629| AP × (6.44) + C16orf13 |32919| ES × (−3.21) +C19orf82 |47381| ES × 2.64 + RNF150 |70661| AP × (−8.56) + PNKP |51105| ES × (−29.80) + ZNF467 |82205| AA × (−8.53) +TMPRSS4 |18957| ES × (−10.62) + HPS1 |91779| ES × (−3.26)

1, 3, or 5 years added more benefit than either the treat-none scheme or treat-all scheme.

Establishment of the SF-AS Correlation Network

To explore the upstream mechanism of AS regulation, we calculated the gene expression levels of SFs from the TCGA ESCA level 3 RNA-seq data and subsequently conducted univariate Cox regression analysis. The results showed that a total of 15 SFs were significantly related to the OS of ESCA patients ($P < 0.05$) (**Supplementary Table S2**). For instance, the expression level of SFs *CLK1* and *SNRPB2* was found to be associated with poor prognosis (**Figures 10A,B**). In addition, the correlations between the PSI values of OS-associated AS events and the gene expression levels of OS-associated SFs were investigated using Spearman's test. Our analysis identified a total of six key SFs that are associated with poor prognosis, including *CLK1, SNRPB2,*

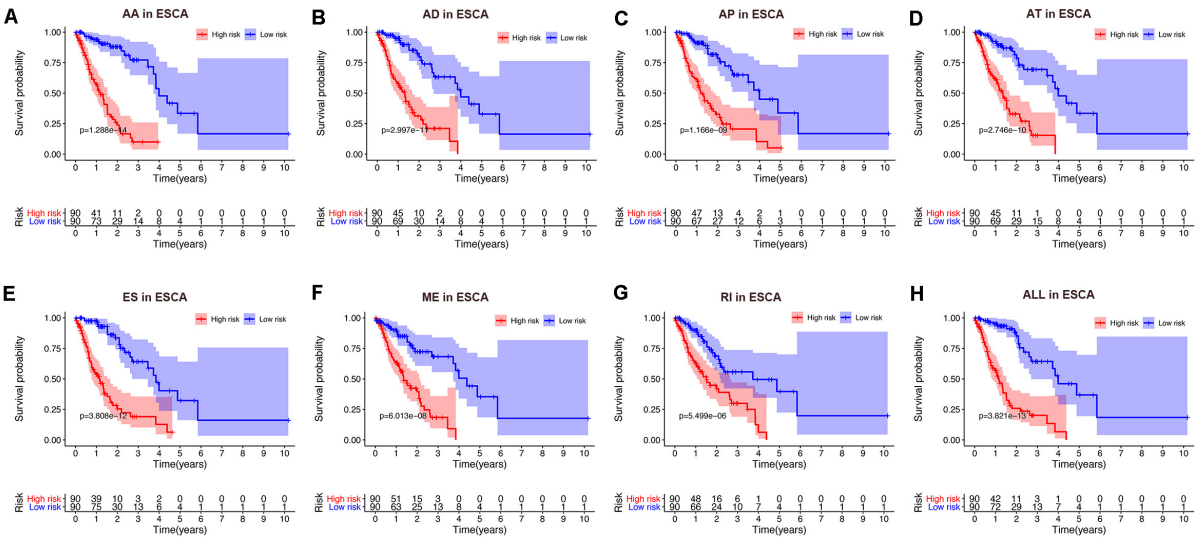


FIGURE 5 | Kaplan-Meier curve of prognostic predictors constructed with either one type or all seven AS types in the ESCA cohort. **(A)** AA: alternate acceptor site. **(B)** AD: alternate donor site. **(C)** AP: alternate promoter. **(D)** AT: alternate terminator. **(E)** ES: exon skip. **(F)** ME: mutually exclusive exons. **(G)** RI: retained intron. **(H)** ALL: all seven AS types combined. Red line indicates high-risk subgroup while blue line indicates low-risk subgroup.

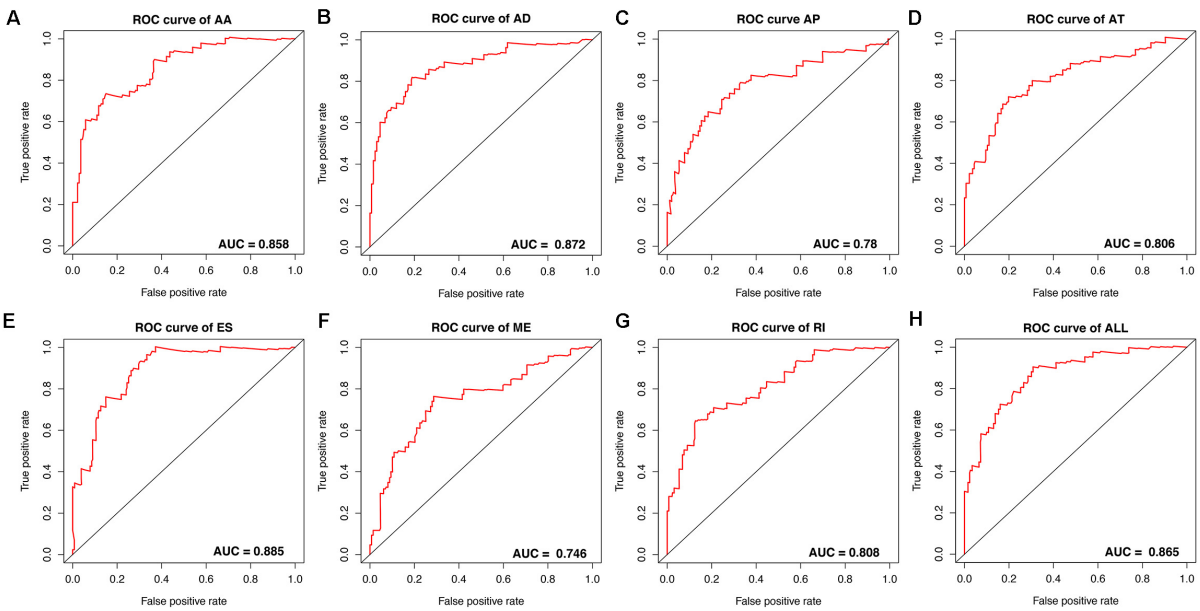


FIGURE 6 | ROC curves with calculated AUC values of prognostic predictors constructed with either one type or all seven AS types in the ESCA cohort. **(A)** AA: alternate acceptor site. **(B)** AD: alternate donor site. **(C)** AP: alternate promoter. **(D)** AT: alternate terminator. **(E)** ES: exon skip. **(F)** ME: mutually exclusive exons. **(G)** RI: retained intron. **(H)** ALL: all seven AS types combined.

TCERG1, *HTATSF1*, *RBMX2*, and *HNRNPH1*, indicating that the abnormal expression of these key SFs may play a role in the dysregulation of the splicing patterns in ESCA. The correlation network as shown in **Figure 10C** revealed a total of 5 OS-associated SFs (blue triangles) that were significantly correlated with 77 OS-associated AS events (red and blue dots). The red dots indicate adverse prognosis (HR > 1) while green dots denote favorable clinical outcomes (HR < 1). Additionally, we found that most adverse survival prognostic AS events (red dots) were positively correlated (red lines) with the expression of SFs (blue triangles); while most favorable prognosis AS events (green dots) were negatively correlated (green lines) with the expression of SFs. The representative dot plots of correlation between the SFs and AS events were shown in **Figures 10D,E**. Based on our observations, we bypothesize that the oncogenic SFs play a key role in

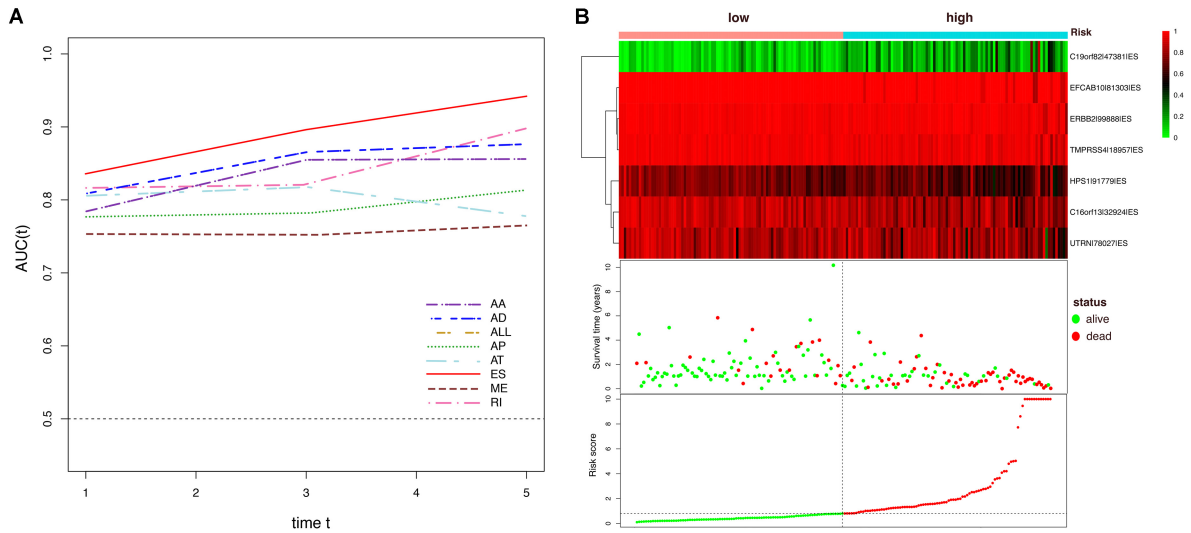


FIGURE 7 | Dynamic AUC values of ROC curves for each AS model and determination of the ES prognostic signature in the ESCA cohort. **(A)** The curves of time-dependent AUCs versus time (1–5 years) of each signature: AUC(t) versus *t* **(B)** Patients were divided into high- and low-risk subgroups based on the median of risk scores based on the ES prognostic predictor. The upper part is the heatmap of AS events involved in the prognostic predictor, color transition from green to red indicates the increasing PSI score of corresponding AS event from low to high. The middle part is the survival status and survival time of each individual. Color of each plot represents the survival status of each patient. The bottom part is risk score of each individual.

TABLE 2 | Prognostic predictors for esophageal carcinoma.

Gene	AS id	Splicing type	Exons	HR	Lower95	Upper95	*P*-value	Index
ERBB2	99888	ES	22	7.24E-12	1.86E-16	2.82E-07	1.97E-06	−25.651926
C19orf82	47381	ES	2:03	28.41024	5.787874	139.4539703	3.74E-05	3.346750
C16orf13	32924	ES	2	0.009271	0.000568	0.151374336	0.00102	−4.680888
UTRN	78027	ES	67	0.053443	0.007712	0.370361388	0.003021	−2.929139
TMPRSS4	18957	ES	11	8.48E-06	6.10E-09	0.011780172	0.001563	−11.678310
HPS1	91779	ES	9	0.005465	0.000377	0.079114333	0.000133	−5.209448
EFCAB10	81303	ES	2	2.11E-05	1.71E-08	0.026086922	0.003038	−10.765213

AS, alternative splicing; ES, Exon Skip; HR, hazard ratio.

meditating the dysregulation of AS in ESCA, which leads to cancer development.

DISCUSSION

AS is a post-translational modification process that generates multiple mRNA isoforms from a single gene. The resulting RNA transcripts can function differently and participate in various physiological processes. Dysregulation of AS in cancer-related genes has been found to participate in many biological processes in tumors, and these abnormally regulated genes can be used as molecular markers for cancer prognosis and treatment. However, an integrated analysis of the AS signature in ESCA remains largely unknown.

In this study, we performed a systematic analysis of OS-associated AS events in 185 of ESCA patients from TCGA SpliceSeq. A total of 2389 AS events were found to be significantly associated with the OS of ESCA patients. Among these OS-associated AS events, some splice variants that have been identified to play an important part in tumor biology were also included in our analysis. For instance, ECM1b, a splice isoform derived from ECM1 (due to an ES event based on our data) can enhance chemosensitivity by suppressing MTORC2/MYC/MTORC1 signaling pathway. One study has demonstrated that ECM1b expression sensitizes ESCA cells to cisplatin, a drug commonly used in ESCA patient treatment (Yu et al., 2019). MUC1, a spliced variant of PUF60 (following an ES event based on our data) can promote carcinogenesis by regulating P53 and β-catenin. An increased expression level of MUC1 is associated with malignant transformation of various malignancies in different tissues, such as breast, colon and pancreas. MUC1 itself has nine main splice variants in which MUC1/C, D and Z are associated with cancer progression (Kahkhaie et al., 2014). Therefore, our comprehensive analysis of AS events nicely complements the AS atlas of ESCA.

The carcinogenesis of ESCA is correlated to multiple pathological processes with a complicated regulatory network. Therefore, predicting tumor prognosis by amalgamating multiple biomarkers and establishing a model is far more effective than

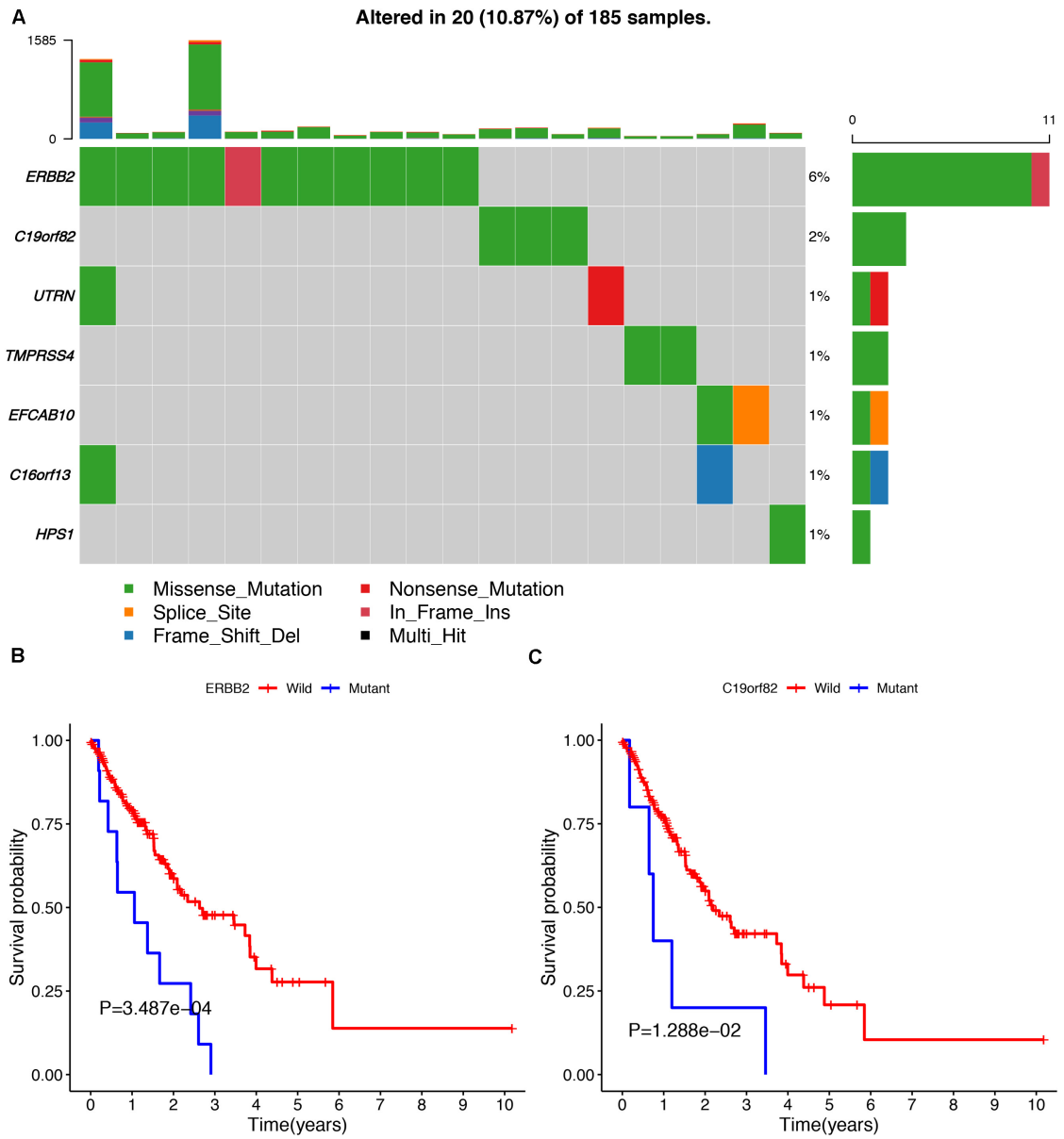


FIGURE 8 | The mutation profiling of parent genes in ESCA samples. **(A)** The waterfall plot of parent genes in ESCA cohort. **(B,C)** Kaplan-Meier survival curves of two different mutated genes (ERBB2 and C19orf82).

that of using a single clinical indicator. Over the past decade, numerous studies have integrated genome-wide prognostic biomarkers to improve the prognosis and diagnosis of ESCA. However, most studies are limited at the transcriptome level, as the focus were given to mRNA, lncRNA or miRNA as the prognostic predictors (Fan and Liu, 2016; Xue et al., 2018). In this study, we focused on AS which belongs to the gene posttranscriptional regulation level. Therefore, we created the prognostic predictors for each type of AS by multivariate Cox regression analysis. Our results showed that the ES model with the best AUC value at 0.885 exhibited a high prediction efficiency than other models. Some parent genes of AS events in the ES model have also been reported to play critical roles in cancer biology. For instance, TMPRSS4, a type-II transmembrane serine protease found to be upregulated in many solid cancers can promote the proliferation, invasion and migration of cancer cells (Jin et al., 2016; Li X.M. et al., 2017; Jianwei et al., 2018). ERBB2, a common oncogene that has been used as one of the key prognostic and treatment indicators in breast cancer, exhibits an overexpressed level in approximately 25–30% of breast cancers and confers a worse biological effect. Besides breast cancer, ERBB2 overexpression is also commonly detected in gastric, esophageal and endometrial cancers (Moasser, 2007). Notably, ES was found to be the most frequent splicing type in our study.

TABLE 3 | Analysis of the final AS signature in stratified ESCA cohorts.

Characteristics	High risk	Low risk	HR (95% CI)	*P*-value
Age (years)				
≤60	47	46	5.29 (2.45-11.4)	<0.001
>60	46	46	6.74 (2.95-15.43)	<0.001
Gender				
Male	83	75	5.49 (3.14-9.60)	<0.001
Female	11	16	7.73 (0.86-69.63)	0.068
Tumor grade				
G1/2	43	49	7.53 (3.09-18.35)	<0.001
G3	31	17	9.87 (2.32-42.09)	0.002
Pathological stage				
Stage I/II	47	56	5.04 (2.21-11.50)	<0.001
Stage III/IV	43	34	6.80 (3.28-14.09)	<0.001
T stage				
T1/2	32	40	4.18 (1.83-9.56)	<0.001
T3/4	56	50	6.39 (3.13-13.08)	<0.001
M stage				
M0	70	75	5.57 (2.99-10.37)	<0.001
M1	8	4	7.32 (0.88-60.59)	0.065
N stage				
N0	34	42	6.63 (2.10-20.93)	0.001
N1	37	34	4.84 (2.41-9.73)	<0.001
N2/3	18	11	5.07 (1.45-17.72)	0.011

HR, hazard ratio; CI, confidence interval.

TABLE 4 | Univariate and multivariate Cox regression analysis for clinical variables.

Variables	Univariate analysis		Multivariate analysis	
	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value
Age	1.01 (0.99–1.03)	0.45	–	–
Gender	2.92 (0.91–9.38)	0.07	–	–
Tumor grade	1.63 (1.07–2.48)	0.02	1.15 (0.66–1.98)	0.63
Pathological stage	2.51 (1.74–3.61)	<0.001	3.03 (1.12–8.18)	0.03
T	1.65 (1.13–2.41)	<0.01	1.05 (0.56–1.62)	0.86
N	1.76 (1.33–2.34)	<0.001	1.03 (0.60–1.75)	0.93
M	2.93 (1.30–6.58)	<0.01	1.25 (0.08–2.55)	0.36
Risk score	1.17 (1.12–1.22)	<0.001	1.13 (1.08–1.19)	<0.001

The "–" indicates that the value is not available; HR, hazard ratio; CI, confidence interval.

In agreement with this, some studies have shown that some splicing variants of genes generated through ES was upregulated in some solid cancers, and can increase the motility of cancer cells (Oltean and Bates, 2014). D16ERBB2, a splice variant of ERBB2 generated through the skipping of exon 16, has been shown to exert high tumorigenecity, and a close association with increased tumor invasive properties and metastasis (Gautrey et al., 2015). Interestingly, our analysis showed that the AS events of ERBB2 is a favorable prognostic predictor, indicating that depending on the exon deletion site, the resulting splicing variant may play an entirely opposite role in tumor development. However, few studies have reported the detailed biological significance of other parent genes in the ES model. Hence, the underlying mechanism of these splicing events involved in final model is largely unclear. Therefore, further research with functional experiments is urgently in need.

Furthermore, to enable the prognostic predictor achieve a more reliable and valuable prediction efficacy in clinical settings, the prognostic nomogram that comprises the pathological stage and the risk level based on the ES prognostic predictor, was developed for assessing individual survival risk of patients with satisfactory discrimination. The calibration curve, C-statistic, and DCA curve demonstrated that the nomogram had great potential to be applied in clinical practice. Moreover, we performed functional enrichment analysis to explore the biological function of AS events in ESCA. Our CC of GO enrichment analysis showed that AS can mediate extracellular matrix-related pathways to promote tumor cell proliferation, invasion and metastasis (Wang et al., 2016). Additionally, KEGG analysis revealed several significant signaling pathways, such as ubiquitin-mediated proteolysis and focal adhesion signaling, which were consistent with the comprehensive analysis of AS in gastrointestinal adenocarcinomas and correlated with the tumorigenesis and prognosis of ESCA (Lin et al., 2018; Zhu R. et al., 2018). Therefore, we hypothesize that the cancer-associated outcomes due to AS alteration may be associated with these common pathways.

As the main regulator of the AS event, SF can affect the choice of splicing sites through recognition and binding of the mRNA precursor. In this study, we identified 6 SFs (*CLK1, SNRPB2, TCERG1, HTATSF1, RBMX2*, and *HNRNPH1*) associated with adverse prognosis of ESCA. Some of these SFs have been reported previously. For example, HNRNPH1, an RNA-binding protein highly expressed in many cancers, was found to alter the splicing of some oncogenes following knockdown, which then inhibits the tumor formation and growth in Rhabdomyosarcoma (Li et al., 2018). CLK1, a member of the CLKs family that phosphorylates SR proteins involved in splicing, was shown to promote the phosphorylation of SPF45 when overexpressed, which ultimately induces cell migration and invasion of ovarian cancer (Liu et al., 2013). Finally, our SF-AS correlation network outlined an obvious trend, showing that whilst most favorable prognostic AS events were negatively associated with the expression level of SFs in ESCA; adverse prognostic AS events were positively associated with the expression level of SFs. Notably, this phenomenon proposed an assumption that the dysregulation of AS in ESCA was related to the up-regulation of SFs. This study provided another approach to understand the splicing patterns and their mechanistic connection to SFs in the ESCA, which will enable us to dissect the potential mechanism of AS events in the development of ESCA.

Although our predictor performed well in ESCA prognosis prediction, there are inevitably several limitations in the current study that can be improved. Firstly, the number of patients included in the ESCA cohorts were limited. Secondly, this study lacks other independent cohort of ESCA patients that can be used to demonstrate the reproducibility of the prognostic predictors constructed in this report. Nevertheless, our comprehensive analysis of the splicing pattern provides some fundamental

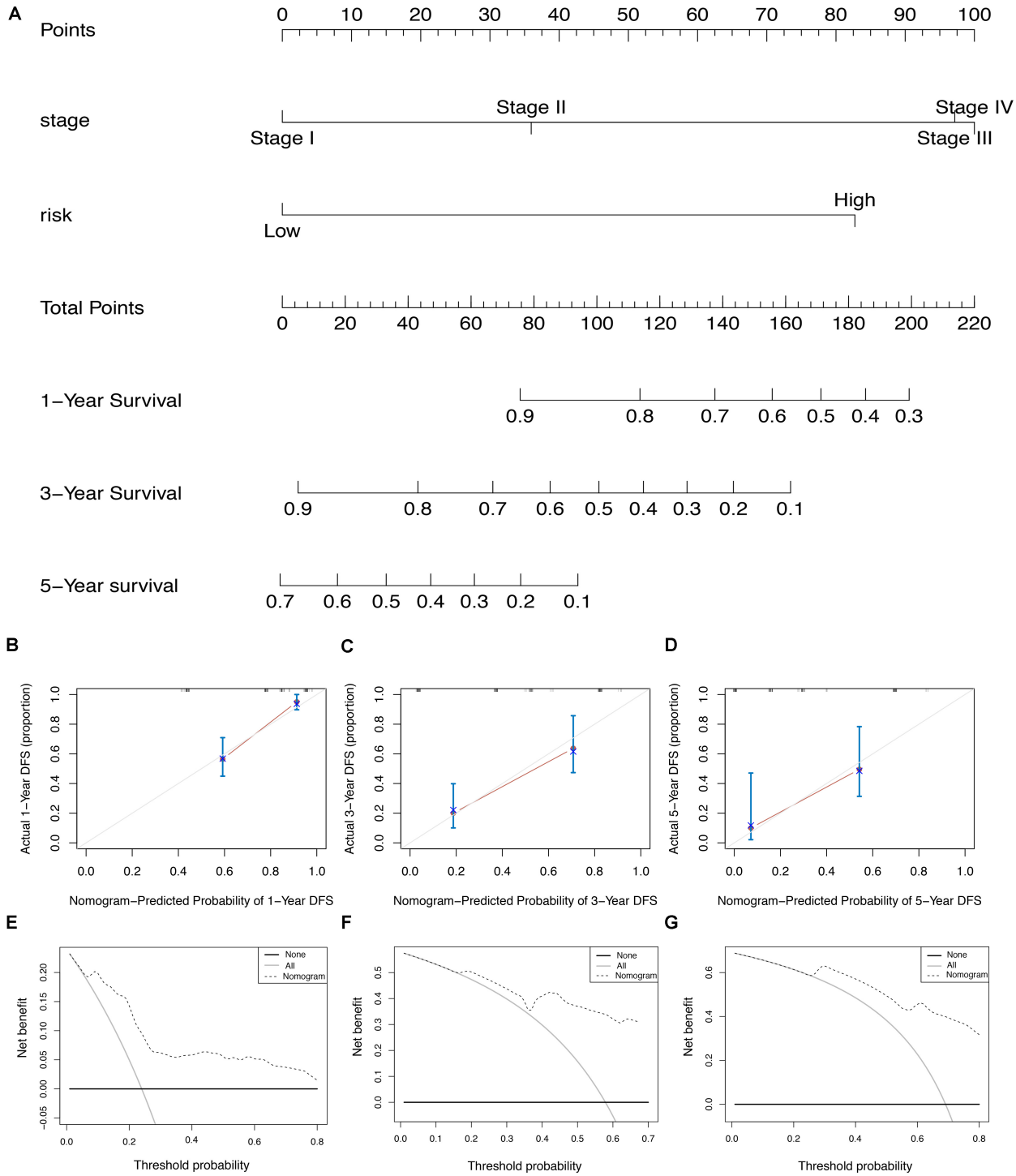


FIGURE 9 | The AS-clinicopathologic nomogram for prediction on survival probability in patients with ESCA. **(A)** Development of AS-clinicopathologic nomogram for predicting 1-, 3-, and 5-years OS for ESCA patients. **(B–D)** Calibration plot of the AS-clinicopathologic nomogram in terms of agreement between nomogram-predicted and observed 1-, 3-, and 5-years outcomes in the ESCA cohort. The actual performances of our model are shown in red lines. And the silver line of 45° represents the ideal performance. **(E–G)** Decision curve analyses of the AS-clinicopathologic nomogram for 1-, 3-, and 5-years risk in ESCA cohort. The gray line represents the net benefit of treat-all scheme varying with threshold probability, while the black line represents the net benefit of treat-no scheme. The net benefits by using our nomogram for predicting 1-, 3-, and 5-years OS are displayed with imaginary line.

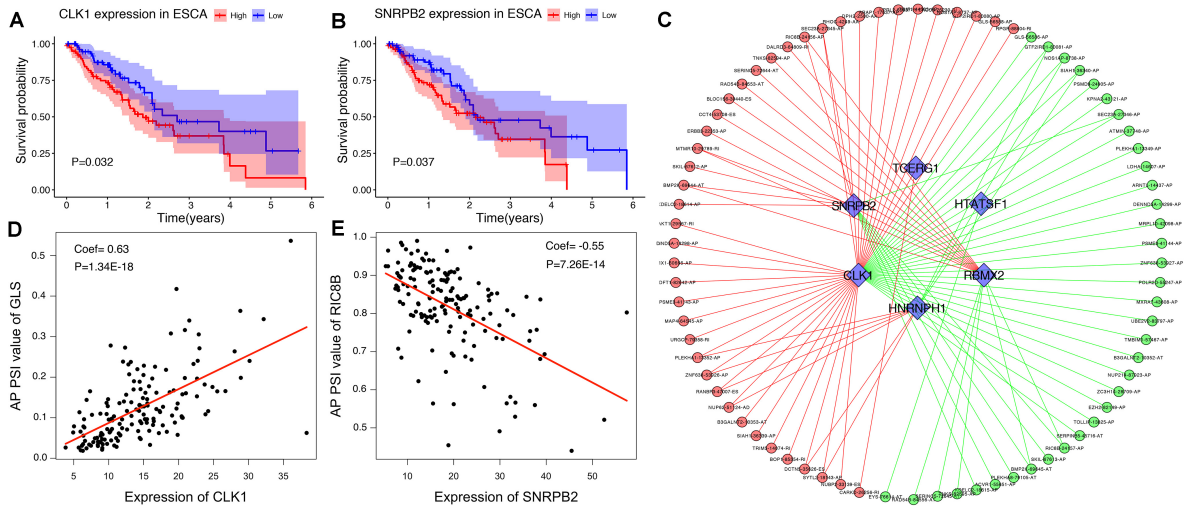


FIGURE 10 | The OS-associated splicing factors in ESCA and SF-AS correlation network. **(A,B)** Kaplan-Meier curves of survival-related splicing factors. **(C)** SF-AS correlation network. Blue triangles were OS-associated splicing factors. Red/green lines represent positive/negative correlations between substances. Red/green dots represent adverse/favorable AS events. **(D,E)** Dot plots of correlations between expression of SFs (CLK1 and SNRPB2) and PSI values of OS-associated AS events.

knowledge to study the molecular mechanism and to identify potential drug targets for ESCA.

CONCLUSION

In conclusion, we performed an integrated analysis for RNA splicing patterns of ESCA and constructed a prognostic predictor that can be used to predict the survival probability of ESCA patients. More importantly, we constructed a well-executed nomogram that combines clinicopathological variables with the final prognostic predictor, which showed a great potential to be applied in clinical settings. The correlation network between prognostic AS events and SFs suggested a potential mechanism of the oncogenic process in ESCA. Additionally, the AS events revealed in our study, particularly those that can be used as a prognostic predictor, exhibited considerable potential for clinical application as prognostic markers as well as therapeutic targets. Our study also provided valuable fundamental knowledge to understand the underlying mechanism of ESCA development.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the TCGA SpliceSeq (https://bioinformatics.mdanderson.org/TCGASpliceSeq/), the TCGA database (https://portal.gdc.cancer.gov, version 18.0).

ETHICS STATEMENT

As our data were downloaded from the TCGA SpliceSeq and TCGA database, there is no requirement for ethics committee approval and consent to participate.

AUTHOR CONTRIBUTIONS

L-QJ contributed to design the study and revised the manuscript. J-RS, C-FK, Y-NL, X-KQ, and RY collected and assembled the data. J-RS, C-FK, and Y-NL conducted the data analysis and interpretation. J-RS drafted the manuscript. All the authors read and approved the final manuscript.

FUNDING

2019 Chinese and Western Medicine Clinical Collaborative Capacity Building Project for Major Difficult Diseases (2019-ZX-005).

ACKNOWLEDGMENTS

We would like to express our sincere appreciation to the TCGA program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00796/full#supplementary-material

FIGURE S1 | Lasso regression analysis for different types of OS-associated AS events. **(A)** AA: alternate acceptor site. **(B)** AD: alternate donor site. **(C)** AP: alternate promoter. **(D)** AT: alternate terminator. **(E)** ES: exon skip. **(F)** ME: mutually exclusive exons. **(G)** RI: retained intron. **(H)** ALL: all types of AS events.

TABLE S1 | Clinical parameters of patients from the TCGA.

TABLE S2 | The survival-associated splicing factors.

REFERENCES

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Caputo, S. M., Leone, M., Damiola, F., Ehlen, A., Carreira, A., Gaidrat, P., et al. (2018). Full in-frame exon 3 skipping of BRCA2 confers high risk of breast and/or ovarian cancer. *Oncotarget* 9, 17334–17348. doi: 10.18632/oncotarget.24671

Chen, J., and Weiss, W. A. (2015). Alternative splicing in cancer: implications for biology and therapy. *Oncogene* 34, 1–14. doi: 10.1038/onc.2013.570

Climente-Gonzalez, H., Porta-Pardo, E., Godzik, A., and Eyras, E. (2017). The functional impact of alternative splicing in cancer. *Cell Rep.* 20, 2215–2226. doi: 10.1016/j.celrep.2017.08.012

Dennis, G. Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* 4:3.

Fan, Q., and Liu, B. (2016). Identification of a RNA-Seq based 8-Long non-coding RNA signature predicting survival in esophageal cancer. *Med. Sci. Monitor Int. Med. J. Exp. Clin. Res.* 22, 5163–5172. doi: 10.12659/msm.902615

Gautrey, H., Jackson, C., Dittrich, A. L., Browell, D., Lennard, T., and Tyson-Capper, A. (2015). SRSF3 and hnRNP H1 regulate a splicing hotspot of HER2 in breast cancer cells. *RNA Biol.* 12, 1139–1151. doi: 10.1080/15476286.2015.1076610

Gelli, E., Colombo, M., Pinto, A. M., De Vecchi, G., Foglia, C., Amitrano, S., et al. (2019). Usefulness and limitations of comprehensive characterization of mRNA splicing profiles in the definition of the clinical relevance of BRCA1/2 variants of uncertain significance. *Cancers* 11:295. doi: 10.3390/cancers11030295

Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., et al. (2010). Alternative expression analysis by RNA sequencing. *Nat. Methods* 7, 843–847. doi: 10.1038/nmeth.1503

Jianwei, Z., Qi, L., Quanquan, X., Tianen, W., and Qingwei, W. (2018). TMPRSS4 upregulates TWIST1 expression through STAT3 activation to induce prostate cancer cell migration. *Patho. Oncol. Res. POR* 24, 251–257. doi: 10.1007/s12253-017-0237-z

Jin, J., Shen, X., Chen, L., Bao, L. W., and Zhu, L. M. (2016). TMPRSS4 promotes invasiveness of human gastric cancer cells through activation of NF-κB/MMP-9 signaling. *Biomed. Pharmacother. Biomed. Pharm.* 77, 30–36. doi: 10.1016/j.biopha.2015.11.002

Kahkhaie, K. R., Moaven, O., Abbaszadegan, M. R., Montazer, M., and Gholamin, M. (2014). Specific MUC1 splice variants are correlated with tumor progression in esophageal cancer. *World J. Surg.* 38, 2052–2057. doi: 10.1007/s00268-014-2523-1

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* 7, 1009–1015. doi: 10.1038/nmeth.1528

Kim, H. K., Pham, M. H. C., Ko, K. S., Rhee, B. D., and Han, J. (2018). Alternative splicing isoforms in health and disease. *Pflugers Arch. Eur. J. Physiol.* 470, 995–1016. doi: 10.1007/s00424-018-2136-x

Ladomery, M. (2013). Aberrant alternative splicing is another hallmark of cancer. *Int. J. Cell Biol.* 2013:463786. doi: 10.1155/2013/463786

Lee, J. S., Lin, Y. Y., Wang, T. S., Liu, J. Y., Lin, W. W., and Yang, J. J. (2018). Antitumorigenic effects of ZAKbeta, an alternative splicing isoform of ZAK. *Chin. J. Physiol.* 61, 25–34. doi: 10.4077/cjp.2018.Bag528

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans. Visual. Comp. Graph.* 20, 1983–1992. doi: 10.1109/tvcg.2014.2346248

Li, X. M., Liu, W. L., Chen, X., Wang, Y. W., Shi, D. B., Zhang, H., et al. (2017). Overexpression of TMPRSS4 promotes tumor proliferation and aggressiveness in breast cancer. *Int. J. Mol. Med.* 39, 927–935. doi: 10.3892/ijmm.2017.2893

Li, Y., Bakke, J., Finkelstein, D., Zeng, H., Wu, J., and Chen, T. (2018). HNRNPH1 is required for rhabdomyosarcoma cell growth and survival. *Oncogenesis* 7:9. doi: 10.1038/s41389-017-0024-4

Li, Y., Sun, N., Lu, Z., Sun, S., Huang, J., Chen, Z., et al. (2017). Prognostic alternative mRNA splicing signature in non-small cell lung cancer. *Cancer Lett.* 393, 40–51. doi: 10.1016/j.canlet.2017.02.016

Liang, Y., Song, J., He, D., Xia, Y., Wu, Y., Yin, X., et al. (2019). Systematic analysis of survival-associated alternative splicing signatures uncovers prognostic predictors for head and neck cancer. *J. Cell. Physiol.* 234, 15836–15846. doi: 10.1002/jcp.28241

Lin, P., He, R. Q., Ma, F. C., Liang, L., He, Y., Yang, H., et al. (2018). Systematic analysis of survival-associated alternative splicing signatures in gastrointestinal pan-adenocarcinomas. *EBioMedicine* 34, 46–60. doi: 10.1016/j.ebiom.2018.07.040

Liu, Y., Conaway, L., Rutherford Bethard, J., Al-Ayoubi, A. M., Thompson Bradley, A., Zheng, H., et al. (2013). Phosphorylation of the alternative mRNA splicing factor 45 (SPF45) by Clk1 regulates its splice site utilization, cell migration and invasion. *Nucleic Acids Res.* 41, 4949–4962. doi: 10.1093/nar/gkt170

Mao, S., Li, Y., Lu, Z., Che, Y., Sun, S., Huang, J., et al. (2019). Survival-associated alternative splicing signatures in esophageal carcinoma. *Carcinogenesis* 40, 121–130. doi: 10.1093/carcin/bgy123

Matera, A. G., and Wang, Z. (2014). A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* 15, 108–121. doi: 10.1038/nrm3742

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: a method long overlooked in behavioral sciences. *Multi. Behav. Res.* 50, 471–484. doi: 10.1080/00273171.2015.1036965

Moasser, M. M. (2007). The oncogene HER2: its signaling and transforming functions and its role in human cancer pathogenesis. *Oncogene* 26, 6469–6487. doi: 10.1038/sj.onc.1210477

Muller, D., Rouleau, E., Schultz, I., Caputo, S., Lefol, C., Bieche, I., et al. (2011). An entire exon 3 germ-line rearrangement in the BRCA2 gene: pathogenic relevance of exon 3 deletion in breast cancer predisposition. *BMC Med. Genet.* 12:121. doi: 10.1186/1471-2350-12-121

Oltean, S., and Bates, D. O. (2014). Hallmarks of alternative splicing in cancer. *Oncogene* 33, 5311–5318. doi: 10.1038/onc.2013.533

Pennathur, A., Gibson, M. K., Jobe, B. A., and Luketich, J. D. (2013). Oesophageal carcinoma. *Lancet* 381, 400–412. doi: 10.1016/s0140-6736(12)60643-6

Piva, F., Giulietti, M., Burini, A. B., and Principato, G. (2012). SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. *Hum. Mutat.* 33, 81–85. doi: 10.1002/humu.21609

Ruytinx, P., Proost, P., and Struyf, S. (2018). CXCL4 and CXCL4L1 in cancer. *Cytokine* 109, 65–71. doi: 10.1016/j.cyto.2018.02.022

Ryan, M., Wong, W. C., Brown, R., Akbani, R., Su, X., Broom, B., et al. (2016). TCGASpliceSeq a compendium of alternative mRNA splicing in cancer. *Nucleic Acids Res.* 44, D1018–D1022. doi: 10.1093/nar/gkv1288

Suo, C., Hrydziuszko, O., Lee, D., Pramana, S., Saputra, D., Joshi, H., et al. (2015). Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. *Bioinformatics* 31, 2607–2613. doi: 10.1093/bioinformatics/btv164

Tress, M. L., Abascal, F., and Valencia, A. (2017). Alternative splicing may not be the key to proteome complexity. *Trends Biochem. Sci.* 42, 98–110. doi: 10.1016/j.tibs.2016.08.008

Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., et al. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* 39, 925–938. doi: 10.1016/j.molcel.2010.08.011

Wang, J., Zhang, G., Wang, J., Wang, L., Huang, X., and Cheng, Y. (2016). The role of cancer-associated fibroblasts in esophageal cancer. *J. Transl. Med.* 14:30. doi: 10.1186/s12967-016-0788-x

Wu, H. Y., Peng, Z. G., He, R. Q., Luo, B., Ma, J., Hu, X. H., et al. (2019). Prognostic index of aberrant mRNA splicing profiling acts as a predictive indicator for hepatocellular carcinoma based on TCGA SpliceSeq data. *Int. J. Oncol.* 55, 425–438. doi: 10.3892/ijo.2019.4834

Xue, W. H., Fan, Z. R., Li, L. F., Lu, J. L., Ma, B. J., Kan, Q. C., et al. (2018). Construction of an oesophageal cancer-specific ceRNA network based on miRNA, lncRNA, and mRNA expression data. *World J. Gastroent.* 24, 23–34. doi: 10.3748/wjg.v24.i1.23

Yu, V. Z., Ko, J. M. Y., Ning, L., Dai, W., Law, S., and Lung, M. L. (2019). Endoplasmic reticulum-localized ECM1b suppresses tumor growth and regulates MYC and MTORC1 through modulating MTORC2 activation in

esophageal squamous cell carcinoma. *Cancer Lett.* 461, 56–64. doi: 10.1016/j.canlet.2019.07.005

Zhu, J., Chen, Z., and Yong, L. (2018). Systematic profiling of alternative splicing signature reveals prognostic predictor for ovarian cancer. *Gynecol. Oncol.* 148, 368–374. doi: 10.1016/j.ygyno.2017.11.028

Zhu, R., Liu, Y., Zhou, H., Li, L., Li, Y., Ding, F., et al. (2018). Deubiquitinating enzyme PSMD14 promotes tumor metastasis through stabilizing SNAIL in human esophageal squamous cell carcinoma. *Cancer Lett.* 418, 125–134. doi: 10.1016/j.canlet.2018.01.025 doi: 10.1016/j.canlet.2018.01.025

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 July 2020
doi: 10.3389/fgene.2020.00673

Identification of a Six-lncRNA Signature With Prognostic Value for Breast Cancer Patients

Erjie Zhao†, Yujia Lan†, Fei Quan†, Xiaojing Zhu, Suru A, Linyun Wan, Jinyuan Xu and Jing Hu**

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Breast cancer (BRCA) is the most common cancer and a major cause of death in women. Long non-coding RNAs (lncRNAs) are emerging as key regulators and have been implicated in carcinogenesis and prognosis. In this study, we aimed to develop a lncRNA signature of BRCA patients to improve risk stratification. In the training cohort (GSE21653, *n* = 232), 17 lncRNAs were identified by univariate Cox proportional hazards regression, which were significantly associated with patients' survival. The least absolute shrinkage and selection operator-penalized Cox proportional hazards regression analysis was used to identify a six-lncRNA signature. According to the median of the signature risk score, patients were divided into a high-risk group and a low-risk group with significant disease-free survival differences in the training cohort. A similar phenomenon was observed in validation cohorts (GSE42568, *n* = 101; GSE20711, *n* = 87). The six-lncRNA signature remained as independent prognostic factors after adjusting for clinical factors in these two cohorts. Furthermore, this signature significantly predicted the survival of grade III patients and estrogen receptor-positive patients. Furthermore, in another cohort (GSE19615, *n* = 115), the low-risk patients that were treated with tamoxifen therapy had longer disease-free survival than those who underwent no therapy. Overall, the six-lncRNA signature can be a potential prognostic tool used to predict disease-free survival of patients and to predict the benefits of tamoxifen treatment in BRCA, which will be helpful in guiding individualized treatments for BRCA patients.

Keywords: long non-coding RNA, signature, prognosis, disease-free survival, breast cancer

OPEN ACCESS

Edited by:
Bailiang Li,
Stanford University, United States

Reviewed by:
Hongyi Zhang,
University of Texas Southwestern Medical Center, United States
Yan Zhang,
Harvard Medical School, United States

****Correspondence:***
Jing Hu
hjistb@gmail.com
Jinyuan Xu
xujinyuan@hrbmu.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *04 March 2020*
Accepted: *02 June 2020*
Published: *24 July 2020*

Citation:
Zhao E, Lan Y, Quan F, Zhu X, A S, Wan L, Xu J and Hu J (2020) Identification of a Six-lncRNA Signature With Prognostic Value for Breast Cancer Patients. Front. Genet. 11:673. doi: 10.3389/fgene.2020.00673

INTRODUCTION

Breast cancer (BRCA) is the second leading cause of cancer death among women. More than 268,000 new patients are diagnosed with BRCA each year and 41,760 patients will die from BRCA (DeSantis et al., 2019; Siegel et al., 2019). The current treatment for BRCA, which can improve survival of BRCA patients, includes mastectomy, hormone therapy (Early Breast Cancer Trialists Collaborative et al., 2011), surgery with adjuvant radiation therapy (Bradley and Mendenhall, 2018; Chargari et al., 2019), and chemotherapy (Oikonomou et al., 2019). Immunotherapy of BRCA patients is a recent emerging area of treatment (Greenlee et al., 2017; Jia et al., 2017; Adams et al., 2019). Although the TNM stage system is a valuable resource for the classification of BRCA patients, it does not predict the prognosis of patients. Therefore, the molecular markers need to be identified so that the survival of BRCA patients can be evaluated (Giuliano et al., 2017; Zhang et al., 2017).

Long non-coding RNAs (lncRNA, >200 nucleotides in length) are a class of non-coding RNAs transcribed from mammalian genomes (Yu et al., 2018). Some lncRNAs are found to be deregulated between cancer and normal tissues, such as BRCA (Liu et al., 2015), lung cancer (Jen et al., 2017), gastric cancer (Liu et al., 2017), and prostate cancer (Xu et al., 2018). Furthermore, lncRNAs have been confirmed to participate in diverse biological processes by acting as key regulators in cancers. Gupta et al. (2010) found that dysregulated *HOTAIR* increased cancer invasiveness and metastasis through dependence on PRC2, and lncRNA *HOXD-AS1* regulated the Rho GTPase activating protein 11A (ARHGAP11A), which resulted in induced metastasis (Lu et al., 2017). In recent years, some lncRNAs have been found to be biomarkers of predicting BRCA patient outcomes, such as lncRNA *BCYRN1* (Booy et al., 2017) and *HOTAIR* (Zhang et al., 2013a), which has attracted increasing attention.

In this study, we developed a six-lncRNA signature based on lncRNA expression, with the ability to predict disease-free survival of patients with BRCA, and we assessed its prognostic value in the training and validation cohorts. This signature had an independent prognostic value after adjusting for clinical factors. Furthermore, the lncRNA signature also significantly predicted survival of grade III and estrogen receptor (ER)-positive BRCA patients. Moreover, the signature predicted survival benefits of tamoxifen therapy in BRCA patients.

MATERIALS AND METHODS

Study Samples

Breast cancer gene expression data generated by the Affymetrix HG-U133 Plus 2.0 microarray platform and corresponding clinical information were obtained from the publicly available GEO database[1]. To analyze the correlation of lncRNA expression with disease-free survival (DFS) for BRCA, we selected those data sets that included patients with survival status information. In total, 232 samples from GSE21653 (Sabatier et al., 2011a,b), 101 samples from GSE42568 (Clarke et al., 2013), and 87 samples from GSE20711 (Dedeurwaerder et al., 2011) were obtained. The GSE21653 data set was defined as the training cohort, and the GSE42568 and GSE20711 data sets were treated as the validation cohort. Another dataset, GSE19615 (n = 115) (Li et al., 2010), which contained 62 patients treated with tamoxifen, was obtained to validate the prognostic value of the signature for patients after hormone treatment. Detailed clinical information of patients with BRCA in this study is shown in **Table 1**.

Microarray Data Processing and lncRNA Re-annotation

All the raw microarray data (CEL files) of BRCA patients were downloaded from the GEO database and background adjusted and normalized using the Robust Multichip Average (RMA) algorithm (Irizarry et al., 2003a,b) and "Affy" package (Gautier et al., 2004). The probe sequences of Affymetrix HG-U133 Plus 2.0 array were downloaded from the Affymetrix website[2] and uniquely mapped to the human genome (hg19). Specific probes of lncRNAs were obtained by matching the chromosomal position of probes to the chromosomal position of lncRNA genes based on the annotations from GENCODE (release 23) according to the previous studies (Du et al., 2013; Zhou et al., 2015). When multiple probes were mapped to the same lncRNA, expression values of these probes were integrated using the median value to represent the expression value of the single lncRNA. As a result, 2,673 lncRNAs were obtained for further analysis.

Identification of a Survival-Related lncRNA Signature Set Associated With Breast Cancer

A univariate Cox proportional hazards regression analysis was carried out to evaluate the association between expression levels of lncRNAs and patients' disease-free survival in the training cohort. Only those lncRNAs with a p-value of <0.01 were considered statistically significant. We then conducted the least absolute shrinkage and selection operator (LASSO) penalized Cox proportional hazards regression analysis to select the prognostic markers of the above lncRNAs (Tibshirani, 1997; Zhang et al., 2013b). We created a risk-score formula by a linear combination of the expressions of these six lncRNAs, weighted by their respective Cox regression coefficients as follows (Zhang et al., 2012, 2013c):

$$Risk\ Score = \sum_{i=1}^{N} \left(Exp_i \times Coef_i\right)$$

where N is the number of prognostic genes, Exp_i is the expression value of the i gene, and $Coef_i$ is the estimated regression coefficient of the i gene in the univariate Cox regression analysis. Using the median signature risk score in each cohort as the cutoff point, BRCA patients in every cohort were divided into low- and high-risk groups.

Statistical Analysis

The association between the lncRNA gene expression and the patient's survival was assessed by univariable Cox regression analysis. LASSO logistic regression analysis was used to identify the lncRNAs comprising the prognostic signature with non-zero coefficients in the training cohort using "glmnet" package (Friedman et al., 2010). Kaplan–Meier survival analysis and the log-rank test were used to compare the difference in disease-free survival between the high-risk group and low-risk group using the R package "survival." Furthermore, we used Cox multivariate analysis to test whether the lncRNA signature was independent of patient age and histological grade. Hazard ratio (HR) and 95% confidence intervals (CI) were estimated by the Cox proportional hazards regression model. The time-dependent receiver operating characteristic (ROC)

[1]https://www.ncbi.nlm.nih.gov/geo/

[2]http://www.affymetrix.com/

TABLE 1 | The Clinical and pathological characteristics of patients in four GEO cohorts.

Characteristic	GSE21653 (*n* = 232)	GSE42568 (*n* = 101)	GSE20711 (*n* = 87)	GSE19615 (*n* = 115)
Age (years)	55.0 (24.0–85.0)	56.9 (31.1–90.0)	53.8 (32.1–82.1)	53.0 (32.0–85.0)
Grade				
Grade I	39	10	13	23
Grade II	76	40	4	28
Grade III	117	51	70	64
ER status				
Positive	128	67	42	70
Negative	104	34	45	45
Median follow up (months)	51.8	66.0	71.4	64.0
Disease-free status				
Relapse	74	45	39	14
No relapse	158	56	48	101
Hormone therapy	–	–	–	
Tamoxifen	–	–	–	62
Arimidex	–	–	–	2
None	–	–	–	47
Unknown	–	–	–	4

curves were used to compare the prognostic accuracy of the six-lncRNA signature for survival. Statistical significance was defined as two-tailed *p*-values being less than 0.05. All of the statistical analyses were performed using R program 3.5.2[3] and Bioconductor.

RESULTS

Identifying a Six-lncRNA Signature in the Training Cohort

As summarized in the workflow (**Figure 1**), we first performed an univariable Cox proportional hazards regression analysis to assess the association between lncRNA expression and disease-free survival of patients with BRCA in the training cohort. A set of 17 lncRNAs that were significantly correlated with patients' survival ($p \leq 0.01$, **Table 2**) was identified. We found six lncRNAs (*LINC00917*, *AL391840.1*, *TRIM52-AS1*, *AL355075.4*, *AC093802.2*, and *AC091544.4*) to comprise a prognostic signature using a LASSO-penalized Cox proportional hazards regression analysis for the above 17 lncRNAs with optimal tuning parameters. All six lncRNAs have positive coefficients, which indicates that their high expressions are associated with shorter survival. Finally, we calculated the signature risk score based on a linear combination of the expression levels of six prognostic lncRNAs, weighted by the coefficients derived from the univariable Cox regression analysis as follows: Risk Score = (1.6348 × expression value of *LINC00917*) + (1.7487 × expression value of *AL391840.1*) + (0.6661 × expression value of *TRIM52-AS1*) + (0.9439 × expression value of *AL355075.4*) + (1.1742 × expression value of *AC093802.2*) + (0.4818 × expression value of *AC091544.4*).

[3]https://www.r-project.org/

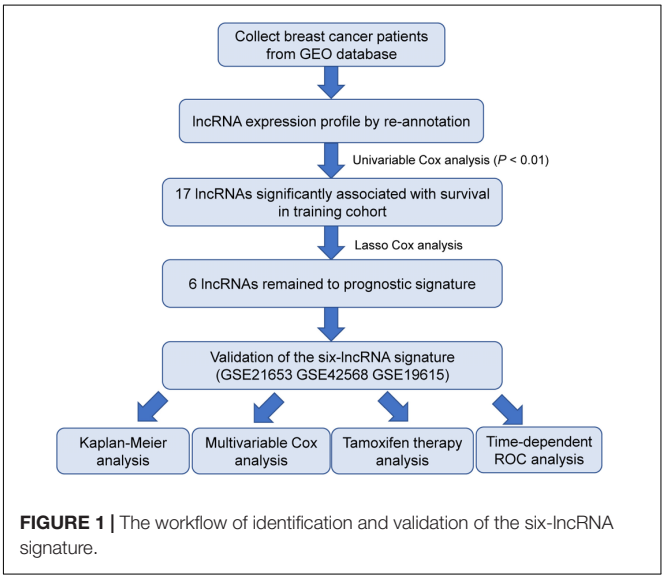


FIGURE 1 | The workflow of identification and validation of the six-lncRNA signature.

The Six-lncRNA Signature Predicts Disease-Free Survival of Patients With Breast Cancer

We calculated the six-lncRNA signature risk score for each patient in the training cohort (GSE21653, $n = 232$). The patients were divided into a high-risk group ($n = 116$) and a low-risk group ($n = 116$) using the median risk score as the cutoff. Compared with the low-risk patients, the high-risk patients had shorter disease-free survival (median survival 62.4 months vs greater than 200 months, HR = 1.67, 95% CI = 1.05–2.66, $p = 0.028$, **Figure 2A**). The prognostic value of the six-lncRNA signature was then evaluated in the validation cohort (GSE42568, $n = 101$). The signature classified patients into two groups, including a high-risk group ($n = 50$) and a low-risk group

TABLE 2 | The 17 lncRNAs that are significantly associated with the disease-free survival in the training cohort (n = 232).

Ensembl ID	lncRNA name	*P*-value	HR (95%CI of HR)
ENSG00000168367	LINC00917	0.0015	5.128 (1.872 – 14.048)
ENSG00000215231	LINC01020	0.0046	10.021 (2.036 – 49.326)
ENSG00000227467	LINC01537	0.0055	8.267 (1.860 – 36.742)
ENSG00000224699	LAMTOR5-AS1	0.0091	1.762 (1.151 – 2.698)
ENSG00000226754	AL606760.1	0.0034	2.184 (1.294 – 3.685)
ENSG00000231533	AL391840.1	0.0054	5.747 (1.678 – 19.686)
ENSG00000259001	AL355075.4	0.0082	2.570 (1.276 – 5.176)
ENSG00000231312	MAP4K3-DT	0.0005	3.687 (1.778 – 7.642)
ENSG00000231528	FAM225A	0.0027	2.220 (1.318 – 3.738)
ENSG00000254887	AC010247.1	0.0054	3.369 (1.432 – 7.923)
ENSG00000259889	AC093802.2	0.0008	3.236 (1.619 – 6.468)
ENSG00000260337	AC091544.4	0.0036	1.619 (1.170 – 2.240)
ENSG00000261292	AC110491.1	0.0036	2.041 (1.262 – 3.301)
ENSG00000260027	HOXB7	0.0094	1.975 (1.181 – 3.302)
ENSG00000248275	TRIM52-AS1	0.0041	1.947 (1.235 – 3.068)
ENSG00000261357	AC099518.2	0.0015	7.756 (2.189 – 27.479)
ENSG00000267317	AC027307.2	0.0022	2.219 (1.331 – 3.699)

(n = 51), based on the median risk score. The disease-free survival of the high-risk group was significantly shorter than that of the low-risk group (median survival 69.7 months vs greater than 100 months, HR = 2, 95% CI = 1.09–3.66, p = 0.022, **Figure 2B**). Similarly, in another validated cohort (GSE20711, n = 87), the high-risk group still had a poorer prognosis than the low-risk group (median survival 77.8 months vs 122.5 months, HR = 1.54, 95% CI = 1.02–2.91, p = 0.040, **Figure 2C**).

Next, we assessed whether the prognostic value of the six-lncRNA signature was independent of other clinical factors. We performed univariate and multivariate Cox proportional hazards regression analysis for factors, including age, ER status, histological grade, and the signature. In the training cohort, the high-risk six-lncRNA signature (HR = 1.789, 95% CI = 1.122–2.852, p = 0.015), grade III (HR = 3.174, 95% CI = 1.314–7.666, p = 0.010) and grade II (HR = 2.881, 95% CI = 1.181–7.028, p = 0.020) were significantly correlated with DFS of patients (**Table 3**). We found that the signature (HR = 2.327, 95% CI = 1.256–4.311, p = 0.007) and ER status (HR = 0.472, 95% CI = 0.234–0.877, p = 0.017) significantly independently predicted patients' disease-free survival in the validation cohort GSE42568 (**Table 3**). Moreover, the six-lncRNA signature was also an independent prognostic factor associated with disease-free survival in the GSE20711 dataset (HR = 1.631, 95% CI = 1.037–3.105, p = 0.043). These results indicate that the six-lncRNA signature is an independent prognostic factor for BRCA patients' disease-free survival.

The Six-lncRNA Signature Predicts Survival of Patients During Diverse BRCA Groups

We explored whether the six-lncRNA signature was effective for patients within different histological grades using a Kaplan–Meier survival analysis. For grade III patients, the signature significantly classified patients into two groups with distinctively different survival times (median survival 55.2 months vs greater than 150 months, HR = 2.39, 95% CI = 1.26–4.51, p = 0.0057, **Figure 3A**), including the high-risk group (n = 56) and the low-risk group (n = 61) in the training cohort. The signature showed a similar prognostic value for grade III patients in the validation cohort (median survival 25.2 months vs greater than 69.3 months, HR = 3.01 95% CI = 0.96–9.46, p = 0.048, **Figure 3B**). In grade I patients, there were no significant survival differences among the high-risk groups and the low-risk groups in two cohorts (**Supplementary Figure S1A,B**). A similar phenomenon was observed in grade II patients from the GSE21653 data set (**Supplementary Figure S1C**). However, in grade II patients from the GSE42568 data set, the high-risk and low-risk groups had significant survival differences (HR = 5.29, 95% CI = 1.17–23.9, p = 0.015, **Supplementary Figure S1D**).

Furthermore, Kaplan–Meier survival analysis was performed after patient stratification according to ER status. The ER-positive patients were divided into high-risk and low-risk groups. The high-risk ER-positive patients had shorter disease-free survival than low-risk ER-positive patients in the training cohort (HR = 1.77, 95% CI = 0.93–3.38, p = 0.078, **Figure 4A**) and the validation cohort (HR = 3.32, 95% CI = 1.31–8.38, p = 0.0072, **Figure 4B**). There were no significant survival differences between the high-risk and low-risk ER-negative patients in these two cohorts when using the same risk formula (**Supplementary Figure S2**).

The Six-lncRNA Signature Predicts Patient Outcome After Tamoxifen Therapy

We further tested whether the six-lncRNA was useful to guide therapy in an independent cohort (GSE19615). In this cohort,

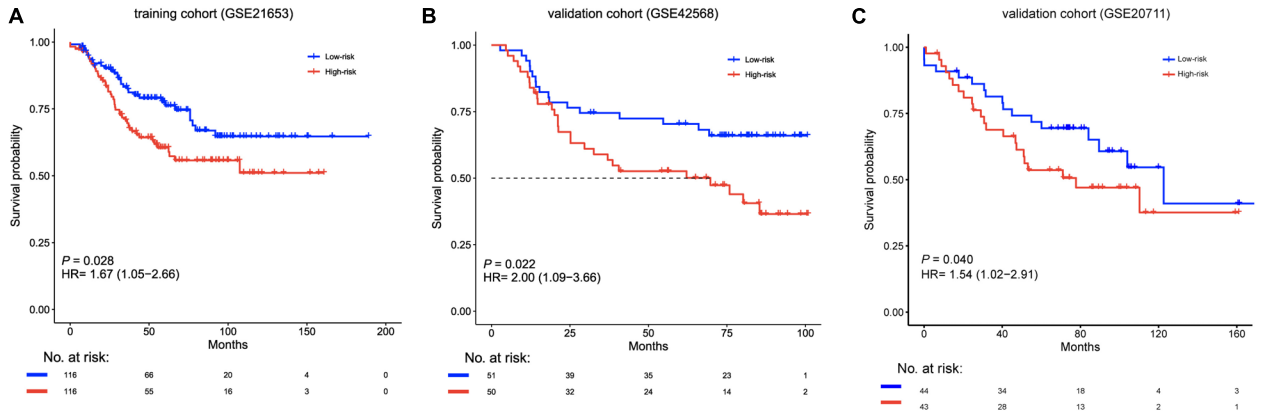


FIGURE 2 | Kaplan–Meier survival curves of disease-free survival between high-risk (red) and low-risk (blue) patients in the **(A)** training cohort (GSE21653, *n* = 232), and the **(B,C)** validation cohort (GSE42568, *n* = 101; GSE20711, *n* = 87). The differences between the two curves were determined by the two-sided log-rank test. The number of patients at risk is listed below the survival curves. HR, hazard ratio.

TABLE 3 | Multivariate analysis for the six-lncRNA signature of disease-free survival in cohorts.

Variables		Univariate analysis			Multivariate analysis		
		HR	95% CI	*P*-value	HR	95% CI	*P*-value
Training set GSE21653 (*n* = 232)							
Age		1.006	0.988−1.024	0.533	1.007	0.989−1.026	0.423
ER status							
	Positive vs Negative	0.670	0.424−1.059	0.087	0.772	0.468−1.272	0.310
Grade							
	Grade II vs Grade I	2.748	1.129−6.691	0.026*	2.881	1.181−7.028	0.020*
	Grade III vs Grade I	3.395	1.437−8.026	0.005*	3.174	1.314−7.666	0.010*
Six-lncRNA signature							
	High-risk vs Low-risk	1.674	1.052−2.664	0.030*	1.789	1.122−2.852	0.015*
Validation set GSE42568 (*n* = 101)							
Age		0.995	0.969−1.021	0.700	1.001	0.975−1.027	0.962
ER status							
	Positive vs Negative	0.439	0.243−0.793	0.006*	0.472	0.254−0.877	0.017*
Grade							
	Grade II vs Grade I	1.497	0.337−6.638	0.596	1.059	0.234−4.788	0.940
	Grade III vs Grade I	3.966	0.943−16.679	0.060	2.880	0.662−12.53	0.158
Six-lncRNA signature							
	High-risk vs Low-risk	1.998	1.092−3.655	0.025*	2.327	1.256−4.311	0.007*
Validation set GSE20711 (*n* = 87)							
Age		1.041	1.010−1.073	0.009*	1.043	1.013−1.075	0.005*
ER status							
	Positive vs Negative	0.554	0.286−1.070	0.079	0.637	0.308−1.316	0.223
Grade							
	Grade II vs Grade I	1.941	0.315−11.947	0.474	2.028	0.275−14.976	0.488
	Grade III vs Grade I	2.564	0.786−8.362	0.118	2.177	0.592−8.013	0.242
Six-lncRNA signature							
	High-risk vs Low-risk	1.539	1.021−2.905	0.040*	1.631	1.037−3.105	0.043*

there were 62 patients who received tamoxifen therapy and 47 who did not. We classified each patient into high- and low-risk groups based on the lncRNA signature risk score. Among the 58 low-risk patients, tamoxifen treatment could prolong the disease-free survival of these patients (HR = 0.08, 95% CI = 0.01–0.62, p = 0.0018, **Figure 5A**), while there were no significant survival differences between patients with and without tamoxifen therapy in the high-risk group (**Figure 5B**). This

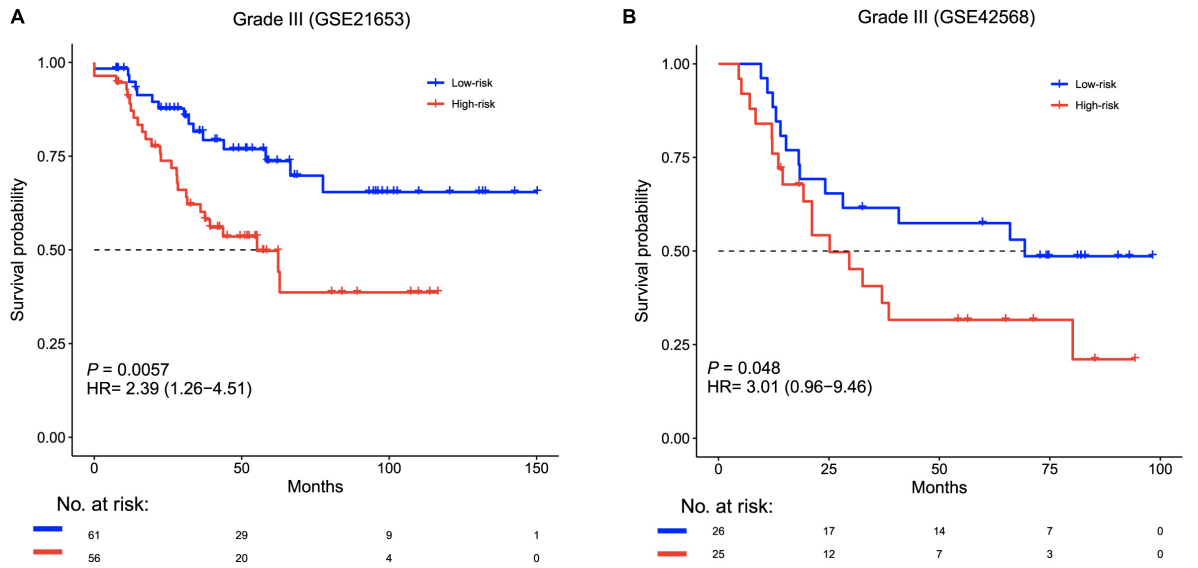


FIGURE 3 | Survival analysis of grade III patients based on the six-lncRNA signature. Kaplan–Meier survival curves for grade III patients in **(A)** GSE21653 (n = 117) and **(B)** GSE42568 (n = 51).

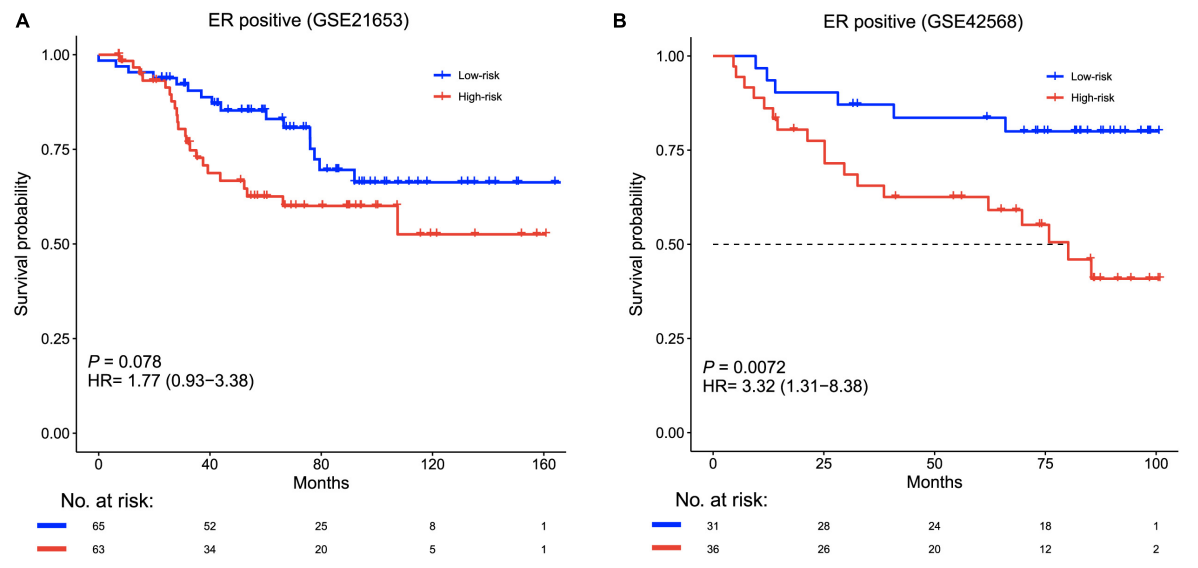


FIGURE 4 | Survival analysis of ER-positive patients based on the six-lncRNA signature. Kaplan–Meier survival curves for ER-positive patients in **(A)** GSE21653 (n = 128) and **(B)** GSE42568 (n = 67).

result revealed that tamoxifen treatment was only beneficial for low-risk BRCA patients.

Comparison of the Survival Prediction Power Between Clinical Factors and the Six-lncRNA Signature

To compare the sensitivity and specificity in survival prediction between clinical factors (histological grade and ER status) and the six-lncRNA signature, we performed a time-dependent ROC analysis in the training cohort. We also constructed a prognostic model by combining our signature with histological grade or ER status. There were no significant differences between histological grade and the lncRNA signature (p = 0.171). A similar result was found between the signature and ER status (p = 0.997). Moreover, for the histological grade, we observed that the histologic grade combined with the six-lncRNA signature (AUC = 0.73) had a higher area under the ROC curve than the histological grade alone (AUC = 0.68, **Figure 6A**). The six-lncRNA signature could also improve the prognostic accuracy of the ER status (0.63 vs 0.59, **Figure 6B**). In addition, for further clinical utility, we constructed a full clinical prognostic model by combining

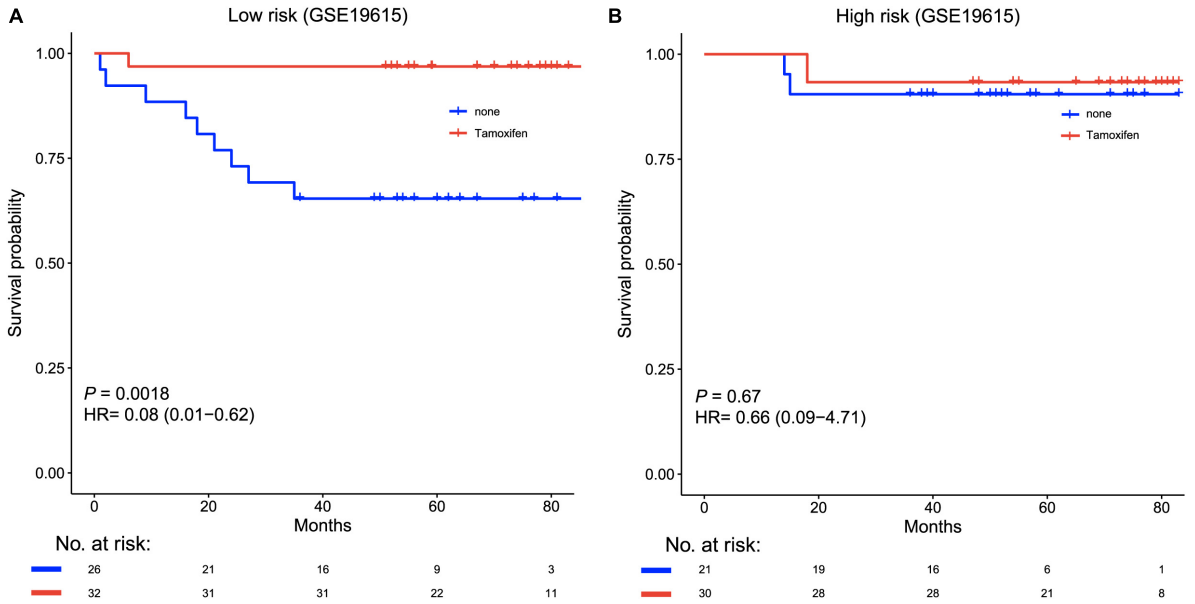


FIGURE 5 | Kaplan–Meier curves of the disease-free survival according to low-risk or high-risk scores, stratified by tamoxifen therapy in an independent cohort (GSE19615, n = 115). **(A)** Patients with tamoxifen therapy (red) had significantly longer disease-free survival than patients without treatment (blue) in the low-risk group. **(B)** Patients who received tamoxifen therapy (red) and those who did not (blue) showed no survival differences in the high-risk group.

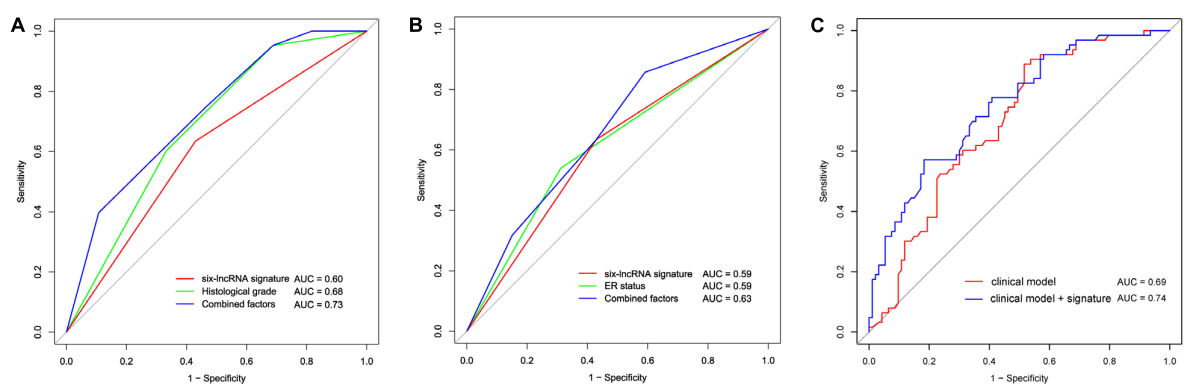


FIGURE 6 | Comparison of sensitivity and specificity for survival prediction by the six-lncRNA signature, histological grade, and ER status. **(A)** The receiver operating characteristics (ROC) curves of the six-lncRNA signature, histological grade, and the combination of the two factors. **(B)** The ROC curves of the six-lncRNA signature, ER status, and the combination of the two factors. **(C)** The ROC curves of the clinical model and the clinical model combined with the six-lncRNA signature. AUC, the area under the curve.

all clinical factors including age, histological grade, and ER status. After adding the six-lncRNA signature into the clinical prognostic model, the prediction accuracy of the model was effectively improved (0.74 vs 0.69, **Figure 6C**). These results suggest that our six-lncRNA signature can add a complementary value to known clinical factors.

DISCUSSION

In the current study, we developed and validated a prognostic six-lncRNA signature based on lncRNA expression, which stratified BRCA patients into two groups (high-risk group and low-risk group) with different disease-free survival. We demonstrated that this signature could predict the survival of grade III BRCA patients. The ER-positive patients who were classified as the low-risk group achieved better survival benefits. Furthermore, by using this signature, we can find a subgroup of patients who are likely to benefit from tamoxifen therapy. In sum, the six-lncRNA signature for BRCA patients may be a prognostic tool that is helpful in guiding individualized treatment of patients.

Histological classification of BRCA into grades I, II, and III, determines the treatment of BRCA patients (Cortes et al., 2012; Harris et al., 2016; Waks and Winer, 2019). The tumor cells of grade III cancer tend to grow more quickly and look different from normal breast cells (Wani et al., 2010). We observed that

the six-lncRNA signature significantly predicted the survival of grade III BRCA patients. This finding suggests that this lncRNA signature predicted survival in patients with invasive cancer. In addition, we found that high-risk ER-positive BRCA patients had shorter disease-free survival than low-risk ER-negative patients. Some studies have confirmed that ER is an essential predictor for responding to therapy, such as tamoxifen therapy, in metastatic BRCA (Fisher et al., 1997).

Given the heterogeneity of cancer, reliable prognostic biomarkers are needed to identify patients who can benefit from therapy (Li et al., 2017; Zhang et al., 2018). There is growing research on several gene signatures to improve decision-making and individualization of BRCA therapy (Cronin et al., 2007; Cardoso et al., 2016; Yu et al., 2019). However, it is difficult to apply all of them for clinical management. Our prognostic signature could identify a group of patients at low risk, where the use of tamoxifen therapy led to significantly extended disease-free survival. This suggests that our signature may hold special clinical value by separating responders to tamoxifen treatment, from non-responders, independent of pathological stage. Such separation could spare non-responders from therapy that is not beneficial and could promote the exploration of more effective therapeutic regimens.

The six-lncRNA (*LINC00917*, *AL391840.1*, *TRIM52-AS1*, *AL355075.4*, *AC093802.2*, and *AC091544.4*) signature in BRCA suggests that lncRNAs can be used as prognostic factors for the survival of patients. To avoid the influence of protein-coding genes, we annotated these probes with protein-coding genes, and found that only one lncRNA overlapped with protein-coding gene RPPH1. This gene had no predictive performance for survival, whether by itself or in combination with other lncRNAs (p = 0.39 and 0.16 respectively, **Supplementary Figure S3**). In addition, among these lncRNAs, *TRIM52-AS1* was dominantly up-regulated in triple-negative breast cancer (TNBC) tissues compared to non-TNBC tissues by a RT-PCR (Lv et al., 2016). Moreover, another study found that the overexpression of TRIM52-AS1 suppressed cell migration and proliferation and induced cell apoptosis in renal cell carcinoma (Liu et al., 2016). However, these six lncRNAs have not been studied in BRCA. Thus, this is a novel study on the association between lncRNA expression and the disease-free survival of patients with BRCA.

Although the signature demonstrated an accurate survival prediction, several limitations should be noted. Because the sample size of our study was limited, large-scale cohort studies should be performed to investigate the prognostic value of this six-lncRNA signature. In addition, we only used a bioinformatics method to predict the six-lncRNA signature in BRCA, thus, further *in vitro* or *in vivo* experiments need to be conducted. Third, we investigated the efficacy of tamoxifen therapy in a low-risk BRCA group, thus more examinations are required to evaluate its efficacy and safety.

In conclusion, the six-lncRNA signature that we identified predicted the disease-free survival of patients with BRCA. This signature also predicted the survival of grade III and ER-positive patients. Furthermore, our findings revealed that the six-lncRNA signature could predict the benefits to patients treated with tamoxifen therapy. Further validation studies are needed to test the prognostic power of this signature before it is used clinically.

DATA AVAILABILITY STATEMENT

All datasets used in this study were publicly available from the GEO database (https://www.ncbi.nlm.nih.gov/geo) at accession numbers GSE21653, GSE42568, GSE20711, and GSE19615.

AUTHOR CONTRIBUTIONS

JH conceived the idea and conceptualized the study. EZ, YL, FQ, and JX conducted the bioinformatics analysis and interpreted the results. EZ and JX collected and pre-processed data. EZ, YL, SA, and LW generated the figures and tables. EZ and YL wrote the manuscript. JH and JX supervised the whole study process and revised the manuscript. All authors have read and approved the final version of manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (Grant No. 31900478), the Heilongjiang Postdoctoral Foundation (LBH-Z17157), and the Fundamental Research Funds for the Provincial Universities (Grant Nos. 2018-KYYWF-0454 and 2018-KYYWF-0455).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00673/full#supplementary-material

REFERENCES

Adams, S., Gatti-Mays, M. E., Kalinsky, K., Korde, L. A., Sharon, E., Amiri-Kordestani, L., et al. (2019). Current landscape of immunotherapy in breast cancer: a review. *JAMA Oncol.* doi: 10.1001/jamaoncol.2018.7147 [Online ahead of print].

Booy, E. P., McRae, E. K., Koul, A., Lin, F., and McKenna, S. A. (2017). The long non-coding RNA BC200 (BCYRN1) is critical for cancer cell survival and proliferation. *Mol. Cancer* 16:109. doi: 10.1186/s12943-017-0679-7

Bradley, J. A., and Mendenhall, N. P. (2018). Novel radiotherapy techniques for breast cancer. *Annu. Rev. Med.* 69, 277–288. doi: 10.1146/annurev-med-042716-103422

Cardoso, F., van't Veer, L. J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., et al. (2016). 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. *N. Engl. J. Med.* 375, 717–729. doi: 10.1056/NEJMoa1602253

Chargari, C., Deutsch, E., Blanchard, P., Gouy, S., Martelli, H., Guerin, F., et al. (2019). Brachytherapy: an overview for clinicians. *CA Cancer J. Clin.* 69, 386–401. doi: 10.3322/caac.21578

Clarke, C., Madden, S. F., Doolan, P., Aherne, S. T., Joyce, H., O'Driscoll, L., et al. (2013). Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis* 34, 2300–2308. doi: 10.1093/carcin/bgt208

Cortes, J., Calvo, V., Ramirez-Merino, N., O'Shaughnessy, J., Brufsky, A., Robert, N., et al. (2012). Adverse events risk associated with bevacizumab addition to breast cancer chemotherapy: a meta-analysis. *Ann. Oncol.* 23, 1130–1137. doi: 10.1093/annonc/mdr432

Cronin, M., Sangli, C., Liu, M. L., Pho, M., Dutta, D., Nguyen, A., et al. (2007). Analytical validation of the Oncotype DX genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. *Clin. Chem.* 53, 1084–1091. doi: 10.1373/clinchem.2006.076497

Dedeurwaerder, S., Desmedt, C., Calonne, E., Singhal, S. K., Haibe-Kains, B., Defrance, M., et al. (2011). DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol. Med.* 3, 726–741. doi: 10.1002/emmm.201100801

DeSantis, C. E., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. D., Goding Sauer, A., et al. (2019). Breast cancer statistics, 2019. *CA Cancer J. Clin.* 69, 438–451. doi: 10.3322/caac.21583

Du, Z., Fei, T., Verhaak, R. G., Su, Z., Zhang, Y., Brown, M., et al. (2013). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* 20, 908–913. doi: 10.1038/nsmb.2591

Early Breast, Cancer Trialists, Collaborative, G., Davies, C., Godwin, J., Gray, R., et al. (2011). Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet* 378, 771–784. doi: 10.1016/S0140-6736(11)60993-8

Fisher, B., Dignam, J., Wolmark, N., DeCillis, A., Emir, B., Wickerham, D. L., et al. (1997). Tamoxifen and chemotherapy for lymph node-negative, estrogen receptor-positive breast cancer. *J. Natl. Cancer Inst.* 89, 1673–1682. doi: 10.1093/jnci/89.22.1673

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Giuliano, A. E., Connolly, J. L., Edge, S. B., Mittendorf, E. A., Rugo, H. S., Solin, L. J., et al. (2017). Breast Cancer-Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J. Clin.* 67, 290–303. doi: 10.3322/caac.21393

Greenlee, H., DuPont-Reyes, M. J., Balneaves, L. G., Carlson, L. E., Cohen, M. R., Deng, G., et al. (2017). Clinical practice guidelines on the evidence-based use of integrative therapies during and after breast cancer treatment. *CA Cancer J. Clin.* 67, 194–232. doi: 10.3322/caac.21397

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076. doi: 10.1038/nature08975

Harris, L. N., Ismaila, N., McShane, L. M., Andre, F., Collyar, D. E., Gonzalez-Angulo, A. M., et al. (2016). Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of clinical oncology clinical practice guideline. *J. Clin. Oncol.* 34, 1134–1150. doi: 10.1200/JCO.2015.65.2289

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31, e15. doi: 10.1093/nar/gng015

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Jen, J., Tang, Y. A., Lu, Y. H., Lin, C. C., Lai, W. W., and Wang, Y. C. (2017). Oct4 transcriptionally regulates the expression of long non-coding RNAs NEAT1 and MALAT1 to promote lung cancer progression. *Mol. Cancer* 16:104. doi: 10.1186/s12943-017-0674-z

Jia, H., Truica, C. I., Wang, B., Wang, Y., Ren, X., Harvey, H. A., et al. (2017). Immunotherapy for triple-negative breast cancer: existing challenges and exciting prospects. *Drug Resist. Updat.* 32, 1–15. doi: 10.1016/j.drup.2017.07.002

Li, B., Cui, Y., Diehn, M., and Li, R. (2017). Development and validation of an individualized immune prognostic signature in early-stage nonsquamous non-small cell lung cancer. *JAMA Oncol.* 3, 1529–1537. doi: 10.1001/jamaoncol.2017.1609

Li, Y., Zou, L., Li, Q., Haibe-Kains, B., Tian, R., Li, Y., et al. (2010). Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. *Nat. Med.* 16, 214–218. doi: 10.1038/nm.2090

Liu, B., Sun, L., Liu, Q., Gong, C., Yao, Y., Lv, X., et al. (2015). A cytoplasmic NF-kappaB interacting long noncoding RNA blocks IkappaB phosphorylation and suppresses breast cancer metastasis. *Cancer Cell* 27, 370–381. doi: 10.1016/j.ccell.2015.02.004

Liu, Z., Chen, Z., Fan, R., Jiang, B., Chen, X., Chen, Q., et al. (2017). Over-expressed long noncoding RNA HOXA11-AS promotes cell cycle progression and metastasis in gastric cancer. *Mol. Cancer* 16:82. doi: 10.1186/s12943-017-0651-6

Liu, Z., Yan, H. Y., Xia, S. Y., Zhang, C., and Xiu, Y. C. (2016). Downregulation of long non-coding RNA TRIM52-AS1 functions as a tumor suppressor in renal cell carcinoma. *Mol. Med. Rep.* 13, 3206–3212. doi: 10.3892/mmr.2016.4908

Lu, S., Zhou, J., Sun, Y., Li, N., Miao, M., Jiao, B., et al. (2017). The noncoding RNA HOXD-AS1 is a critical regulator of the metastasis and apoptosis phenotype in human hepatocellular carcinoma. *Mol. Cancer* 16:125. doi: 10.1186/s12943-017-0676-x

Lv, M., Xu, P., Wu, Y., Huang, L., Li, W., Lv, S., et al. (2016). LncRNAs as new biomarkers to differentiate triple negative breast cancer from non-triple negative breast cancer. *Oncotarget* 7, 13047–13059. doi: 10.18632/oncotarget.7509

Oikonomou, E. K., Kokkinidis, D. G., Kampaktsis, P. N., Amir, E. A., Marwick, T. H., Gupta, D., et al. (2019). Assessment of prognostic value of left ventricular global longitudinal strain for early prediction of chemotherapy-induced cardiotoxicity: a systematic review and meta-analysis. *JAMA Cardiol.* 4, 1007–1018. doi: 10.1001/jamacardio.2019.2952

Sabatier, R., Finetti, P., Adelaide, J., Guille, A., Borg, J. P., Chaffanet, M., et al. (2011a). Down-regulation of ECRG4, a candidate tumor suppressor gene, in human breast cancer. *PLoS One* 6:e27656. doi: 10.1371/journal.pone.0027656

Sabatier, R., Finetti, P., Cervera, N., Lambaudie, E., Esterni, B., Mamessier, E., et al. (2011b). A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res. Treat.* 126, 407–420. doi: 10.1007/s10549-010-0897-9

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA Cancer J. Clin.* 69, 7–34. doi: 10.3322/caac.21551

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16, 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385:aid-sim380<3.0.co;2-3

Waks, A. G., and Winer, E. P. (2019). Breast cancer treatment: a review. *JAMA* 321, 288–300. doi: 10.1001/jama.2018.19323

Wani, F. A., Bhardwaj, S., Kumar, D., and Katoch, P. (2010). Cytological grading of breast cancers and comparative evaluation of two grading systems. *J. Cytol.* 27, 55–58. doi: 10.4103/0970-9371.70738

Xu, J., Lan, Y., Yu, F., Zhu, S., Ran, J., Zhu, J., et al. (2018). Transcriptome analysis reveals a long non-coding RNA signature to improve biochemical recurrence prediction in prostate cancer. *Oncotarget* 9, 24936–24949. doi: 10.18632/oncotarget.25048

Yu, F., Quan, F., Xu, J., Zhang, Y., Xie, Y., Zhang, J., et al. (2019). Breast cancer prognosis signature: linking risk stratification to disease subtypes. *Brief. Bioinform.* 20, 2130–2140. doi: 10.1093/bib/bby073

Yu, F., Zhang, G., Shi, A., Hu, J., Li, F., Zhang, X., et al. (2018). LnChrom: a resource of experimentally validated lncRNA-chromatin interactions in human and mouse. *Database* 2018:bay039. doi: 10.1093/database/bay039

Zhang, H., Deng, Y., Zhang, Y., Ping, Y., Zhao, H., Pang, L., et al. (2017). Cooperative genomic alteration network reveals molecular classification across 12 major cancer types. *Nucleic Acids Res.* 45, 567–582. doi: 10.1093/nar/gkw1087

Zhang, J. X., Han, L., Bao, Z. S., Wang, Y. Y., Chen, L. Y., Yan, W., et al. (2013a). HOTAIR, a cell cycle-associated long noncoding RNA and a strong predictor of

survival, is preferentially expressed in classical and mesenchymal glioma. *Neuro Oncol.* 15, 1595–1603. doi: 10.1093/neuonc/not131

Zhang, J. X., Song, W., Chen, Z. H., Wei, J. H., Liao, Y. J., Lei, J., et al. (2013b). Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. *Lancet Oncol.* 14, 1295–1306. doi: 10.1016/S1470-2045(13)70491-1

Zhang, X. Q., Sun, S., Lam, K. F., Kiang, K. M., Pu, J. K., Ho, A. S., et al. (2013c). A long non-coding RNA signature in glioblastoma multiforme predicts survival. *Neurobiol. Dis.* 58, 123–131. doi: 10.1016/j.nbd.2013.05.011

Zhang, X., Sun, S., Pu, J. K., Tsang, A. C., Lee, D., Man, V. O., et al. (2012). Long non-coding RNA expression profiles predict clinical phenotypes in glioma. *Neurobiol. Dis.* 48, 1–8. doi: 10.1016/j.nbd.2012.06.004

Zhang, Y., Li, X., Zhou, D., Zhi, H., Wang, P., Gao, Y., et al. (2018). Inferences of individual drug responses across diverse cancer types using a novel competing endogenous RNA network. *Mol. Oncol.* 12, 1429–1446. doi: 10.1002/1878-0261.12181

Zhou, M., Guo, M., He, D., Wang, X., Cui, Y., Yang, H., et al. (2015). A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J. Transl. Med.* 13:231. doi: 10.1186/s12967-015-0556-3

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.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 05 August 2020
doi: 10.3389/fbioe.2020.00737

A Neural Network Framework for Predicting the Tissue-of-Origin of 15 Common Cancer Types Based on RNA-Seq Data

Binsheng He[1†], Yanxiang Zhang[2†], Zhen Zhou[3†], Bo Wang[2], Yuebin Liang[2], Jidong Lang[2], Huixin Lin[2], Pingping Bing[1], Lan Yu[4], Dejun Sun[4], Huaiqing Luo[1*], Jialiang Yang[1,2*] and Geng Tian[2*]*

[1] Academician Workstation, Changsha Medical University, Changsha, China, [2] Geneis (Beijing) Co., Ltd., Beijing, China, [3] Department of Radiology, Beijing Chest Hospital, Capital Medical University, Beijing Tuberculosis and Thoracic Tumor Research Institute, Beijing, China, [4] Inner Mongolia People's Hospital, Huhhot, China

Sequencing-based identification of tumor tissue-of-origin (TOO) is critical for patients with cancer of unknown primary lesions. Even if the TOO of a tumor can be diagnosed by clinicopathological observation, reevaluations by computational methods can help avoid misdiagnosis. In this study, we developed a neural network (NN) framework using the expression of a 150-gene panel to infer the tumor TOO for 15 common solid tumor cancer types, including lung, breast, liver, colorectal, gastroesophageal, ovarian, cervical, endometrial, pancreatic, bladder, head and neck, thyroid, prostate, kidney, and brain cancers. To begin with, we downloaded the RNA-Seq data of 7,460 primary tumor samples across the above mentioned 15 cancer types, with each type of cancer having between 142 and 1,052 samples, from the cancer genome atlas. Then, we performed feature selection by the Pearson correlation method and performed a 150-gene panel analysis; the genes were significantly enriched in the GO:2001242 Regulation of intrinsic apoptotic signaling pathway and the GO:0009755 Hormone-mediated signaling pathway and other similar functions. Next, we developed a novel NN model using the 150 genes to predict tumor TOO for the 15 cancer types. The average prediction sensitivity and precision of the framework are 93.36 and 94.07%, respectively, for the 7,460 tumor samples based on the 10-fold cross-validation; however, the prediction sensitivity and precision for a few specific cancers, like prostate cancer, reached 100%. We also tested the trained model on a 20-sample independent dataset with metastatic tumor, and achieved an 80% accuracy. In summary, we present here a highly accurate method to infer tumor TOO, which has potential clinical implementation.

Keywords: cancer of unknown primary, tissue-of-origin, neural network, RNA sequencing, the Pearson correlation

OPEN ACCESS

Edited by:
Lu Zhang,
Hong Kong Baptist University,
Hong Kong

Reviewed by:
Missaoui Nabiha,
University of Sousse, Tunisia
Jianghan Qu,
Veracyte, United States

***Correspondence:**
Binsheng He
hbscsmu@163.com
Huaiqing Luo
luohuaiqing@csu.edu.cn;
linhx2@geneis.cn
Jialiang Yang
yangjl@geneis.cn
Geng Tian
tiang@geneis.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *05 March 2020*
Accepted: *10 June 2020*
Published: *05 August 2020*

Citation:
He B, Zhang Y, Zhou Z, Wang B, Liang Y, Lang J, Lin H, Bing P, Yu L, Sun D, Luo H, Yang J and Tian G (2020) A Neural Network Framework for Predicting the Tissue-of-Origin of 15 Common Cancer Types Based on RNA-Seq Data. Front. Bioeng. Biotechnol. 8:737. doi: 10.3389/fbioe.2020.00737

INTRODUCTION

Worldwide, almost one in three cancer patients is clinically diagnosed with distant metastases. In most cases, primary and metastatic lesions are identified simultaneously; however, some primary tumors cannot be found after systematic clinicopathological diagnosis (Tomuleasa et al., 2017). Cases with cancer of unknown primary (CUP) lesions account for approximately 3–5% of all newly diagnosed cancers (Richardson et al., 2015); due to its poor prognosis, CUP is the fourth-highest cause of cancer-related deaths around the world (Pavlidis and Fizazi, 2005; Kamposioras et al., 2013). Cancer of unknown primary patients are generally treated with non-selective empirical chemotherapy, which leads to a very low short-term survival rate (Kurahashi et al., 2013). Thus, identifying the primary site is critical for improving long-term survival in CUP patients, especially when considering cancer-type specific targeted therapy (Hudis, 2007; Varadhachary et al., 2008; Hyphantis et al., 2013).

To identify the primary lesion of CUP, a systematic assessment is performed which consists of physical examination, patient-history analysis, serum markers, radiological imaging; as well as immunohistochemical analysis. Immunohistochemical markers are very important for determining tissue-of-origin (TOO; MacReady, 2010; Molina et al., 2012; Oien and Dennis, 2012; Pavlidis and Pentheroudakis, 2012); however, the expressed markers may be non-specific sometimes (Handorf et al., 2013; Montezuma et al., 2013; Tothill et al., 2013). Recently, studies have shown that cellular-origin signatures, which are sufficiently retained in primary tissue, persist after primary cancer cells undergo dedifferentiation and colonization in different tissue types (Ma et al., 2005; Tothill et al., 2005). Molecular profiling is a promising technique that can improve primary-site diagnosis in CUP patients (Ma et al., 2005; Lazaridis et al., 2008; Meiri et al., 2012); it is based on expression microarrays and the quantitative real-time polymerase chain reaction (qRT-PCR) experimental platform (Ma et al., 2005; Lazaridis et al., 2008; Greco et al., 2012; Meiri et al., 2012).

In recent years, cancer classification based on gene expression data such as RT-PCR has attracted great interest and has been implemented in different studies (Lapointe et al., 2004; Mramor et al., 2007; Liu et al., 2008). Single studies are prone to laboratory-specific bias; they are usually limited to a relatively small number of samples and fail to yield novel markers for clinical application. However, applying Next Generation Sequencing (NGS) technology helps alleviate the issue of batch effect by providing gene expression data sets from multiple studies; thus, the integrative analysis of such data can be considered a source of cancer classification. In this regard, establishing a robust classification model is a challenging task; bioinformatics feature selection techniques for establishing such models have been introduced in a previous review (Saeys et al., 2007).

Support vector machines (SVMs) based on the recursive feature elimination (RFE) algorithm represent embedded methods used for feature selection and classification modeling based on microarray gene expression data, which mined 11,925 genes to 154 genes with definite biological significance (Xu et al., 2016). More than 20,000 genes were generated from NGS RNA-Seq data in other studies (Bhowmick et al., 2019); this number is almost twice as much as that from microarray gene expression data. Hence, RNA-Seq data from nine cancer types (lung, liver, colon, thyroid, prostate, bladder, kidney, brain, and skin) were analyzed with different algorithms, and Artificial Bee Colony (ABC) yielded better results than Ant Colony Optimization, Differential Evolution, and Particle Swarm Optimization. Among different cancer types, lower grade brain glioma had the highest accuracy (99.1%) based on the ABC algorithm (Bhowmick et al., 2019). However, the robustness of feature selection and classification modeling methods still needs to be comprehensively evaluated; different algorithms might result in different results depending on their model (Chopra et al., 2010; Bhowmick et al., 2019). Therefore, it is necessary to design a robust classification algorithm based on NGS data that can yield accurate cancer type classification and supplement clinical examination.

In the present study, genome-wide gene expression profiles were established based on comprehensive RNA-Seq data. The gene expression data of ∼8,000 tumor samples were used to identify gene signatures for 15 common human cancer types (lung, breast, liver, colorectal, gastroesophageal, ovarian, cervical, endometrial, pancreatic, bladder, head and neck, thyroid, prostate, kidney, and brain). To screen gene features and evaluate cancer classifiers, the Pearson correlation Neural Network (NN) algorithm was implemented in this study to identify tumor origins.

MATERIALS AND METHODS

RNA-Seq Datasets

NGS-based gene expression profiling data of 7,480 tumor samples were collected from The Cancer Genome Atlas (TCGA, release version v26),[1] and the tissue origins of those samples were confirmed through histopathological analysis. The downloaded data offered RNA-seq data of 21 cancer types that belongs to projects from United States, which is sequenced using the same protocols. Among them, melanoma had a distinct distribution from other cancer types (80 samples were sampled from primary tumor and 352 were sampled from metastatic tumor) and was excluded. Thus, the expression profiles of 15 common cancer types (lung, breast, liver, colorectal, gastroesophageal, ovarian, cervical, endometrial, pancreatic, bladder, head and neck, thyroid, prostate, kidney, and brain) were studied in this work. The normalized expression value of expression data was downloaded from TCGA and provided the expression levels of 20,501 unique genes for the 15 chosen cancer types.

To perform the bioinformatics analysis in this study, the transcript level of genes was normalized again to form a matrix with rows of sample numbers and columns of gene numbers.

[1] https://dcc.icgc.org/releases/release_26

The normalization was done by dividing the sum of the gene expression value of each sample. Normalized gene expression data were extracted and represented as a matrix with '*m*' rows and '*n*' columns, such that '*m*' represented 7,480 tumor samples and '*n*' represented the expression levels of 20,501 unique genes.

For log transformation, we used $\log_2$ to transform the original dataset after replacing zeros to global minimum × 0.1. No normalizations were done after feature selection.

Among all the samples, 7,460 samples were sampled from primary tumors, remaining 20 samples sampled from the metastatic tumors.

Gene Feature Identification

To identify an optimal gene signature, we introduced a strategy of feature selection and multi-class classification modeling in this study. According to the mechanism of feature selection, the sets of genes were screened by the Pearson Correlation algorithm (Hall, 1998; Saeys et al., 2007). This study consisted of the following steps: (i) create an array to binarize rows for each cancer type (*C* columns) for the *m* tumor samples, labeling the sample as "true" if the sample belongs to the cancer type, otherwise the sample was labeled as "False," where *C* is the total cancer types and *m* is the sample number; (ii) calculate the correlation of gene expression level with samples labeled "true" for each cancer type, then sort in decreasing order according to their correlation; (iii) take the most important signatures, appeared top *N* of the list, for each cancer type, where *N* is an integer; and (iv) combine *C* lists of the top *N* genes and remove the redundant genes, generating a gene set. Gene expression values from the gene set will be extracted for further usage.

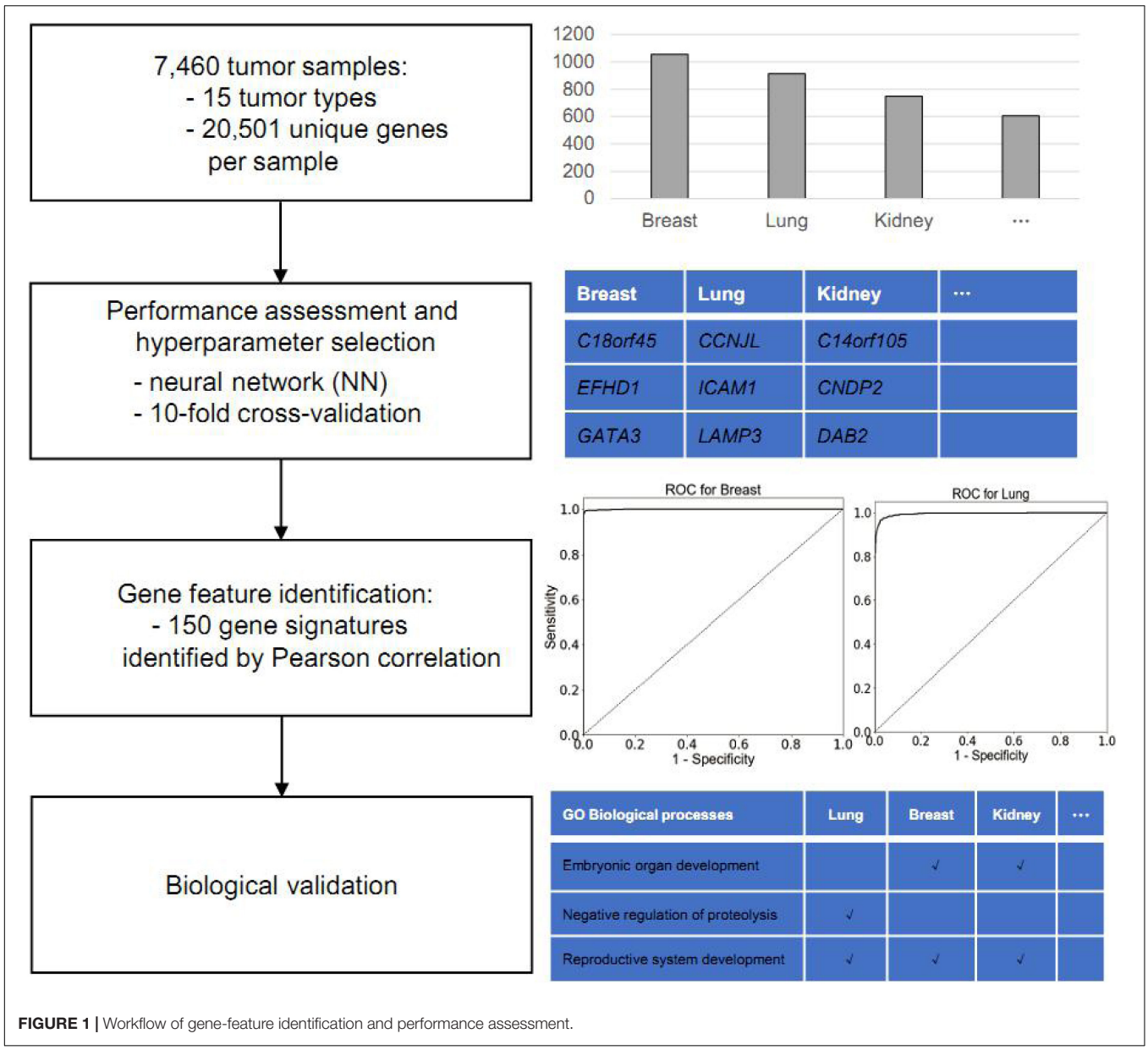


FIGURE 1 | Workflow of gene-feature identification and performance assessment.

Feature Performance Assessment

We used a NN (Hinton, 1989) to train the classification model. The gene expression values were used as input signatures for the NN. The NN was designed with three layers, in which the input layer has *N* units, the hidden layer has 50 units, and output layer has 15 units corresponding to each cancer where *N* is the gene number of the input matrix. The output layer of the NN was used as the input for the Softmax function to obtain the probabilities for each cancer type. To prevent overfitting, L2 penalty was set to 0.0001. For comparison, we used logistic regression as a baseline method. The parameter C was set to 10,000 for logistic regression. The algorithms were implemented using scikit-learn package (Pedregosa et al., 2011).

Gene Ontology Analysis

To perform the Gene ontology (GO) analysis of the identified gene features, GO consortium (Ashburner et al., 2000) was used. The enrichment result was generated by clusterProfiler, which performs a hyper geometric test between the tested genes and gene sets in GO terms (Yu et al., 2012). The biological significances of the selected genes were examined by GO enrichment analysis to identify the most enriched biological-process terms. Benjamini–Hochberg was used to adjust the *p* value.

RESULTS

Collection of Gene Expression Datasets of Common Human Cancer Types

The main objective in this study is to identify putative gene biomarkers to classify cancer type. The workflow of the present study is shown in **Figure 1**. For this analysis, the TCGA was used to obtain gene expression profiles of 15 common solid tumor cancer types via NGS-based RNA-Seq, including lung, gastroesophageal, colorectal, liver, breast, thyroid, cervical, brain, pancreatic, ovarian, endometrial, bladder, kidney, head and neck, and prostate. In total, the expression data of 7,480 tumor samples were collected. Among those, the gene expression profiles of lung adenocarcinoma and lung squamous cell carcinoma samples were merged into lung cancer; those of colon adenocarcinoma and rectum adenocarcinoma were merged into colorectal cancer; those of kidney renal clear cell carcinoma and kidney renal papillary cell carcinoma were merged into kidney cancer; and those of glioblastoma multiforme and lower grade glioma were merged into brain cancer.

Around 20 of the 7,480 samples were sampled from metastatic tumors, whereas 7,460 were sampled from primary tumors. Thus, we split the dataset into the 7,460-sample training dataset and the 20-sample test dataset according to the sampling tumor type. All cancer types in the training dataset had more than 100 samples; the largest sample size was that of breast cancer (1,056 samples), whereas, the smallest sample size was that of pancreatic cancer (142 samples). **Table 1** summarizes the datasets and provides information on the tumor samples.

TABLE 1 | Summary of samples used in the experiments.

Sampling site	Cancer type	Code	Sample size	Percentage (%)
Primary	Lung	LUAD + LUSC	914	12.25
	Gastroesophageal	STAD	415	5.56
	Colorectal	COAD + READ	604	8.10
	Liver	LIHC	294	3.94
	Breast	BRCA	1056	14.16
	Thyroid	THCA	500	6.70
	Cervical	CESC	258	3.46
	Brain	GBM + LGG	529	7.94
	Pancreatic	PAAD	142	1.90
	Ovary	OV	261	3.50
	Endometrial	UCEC	516	6.92
	Bladder	BLCA	301	4.03
	Kidney	KIRC + KIRP	748	10.03
	Head and Neck	HNSC	480	6.43
	Prostate	PRAD	379	5.08
	Total for primary tumors		7,460	100
Metastatic	Breast	BRCA	7	35.00
	Cervical	CESC	2	10.00
	Colorectal	COAD + READ	1	5.00
	Head and Neck	HNSC	2	10.00
	Thyroid	THCA	8	40.00
	Total for metastatic tumors		20	100

Hundred and Fifty as a Feature Number Works Well With the Neural Network

A classification modeling database of 15 common cancer types was established based on the expression data of 20,501 unique genes obtained from TCGA. However, having a large number of samples per cancer type might result in variations due to intra-tumor heterogeneity; hence, it is critical to identify the gene expression features from high-dimension datasets. Pearson correlation-based feature selection represents a multivariable filter method for high-dimension data analysis (Hall, 1998; Saeys et al., 2007), which is fast in operation and simple in complex computation; they are used to assess the correlation between cancer type and corresponding gene-expression features. Here, we used Pearson correlation to select the gene-expression signature from NGS-based mRNA expression data for each cancer type. In this study, we used integers from 1 to 20 as candidates for gene number for each cancer type, which might give rise to 20 possible gene sets of 15, 30, ..., 300 with a step of 15.

The regression model is an important mathematical model for classification. NNs, as types of deep learning algorithms, are advanced techniques that can analyze complex and high-dimensional data. NNs have been applied in protein classification (Asgari and Mofrad, 2015) and anomaly classification (Suk and Shen, 2013; Plis et al., 2014; Hua et al., 2015). Here, we used NNs as the classification model to assess the performance of different numbers of features. The gene expressions levels were the input layer for the NN; 15 cancer types were the output layer obtained from NNs.

Usually, 10-fold cross-validation is used for minimizing the over-fitting issues and obtaining good performance. Hence, to avoid overfitting of the NN algorithm, we ran a 10-fold cross-validation 10 times using the 7,460-sample training dataset to obtain relatively stable and reliable results, possibly minimizing the percentage of false positives and false negatives. The 10-fold cross validation was performed as follows. (a) Split the whole training dataset into 10 disjoint parts randomly. (b) Use 9 parts as the training set (9/10 training set). (c) Choose N genes using Pearson correlation from the 9/10 training set, where N is the gene number which might be 15, 30, . . ., 300 with a step of 15. (d) Train a model using the selected genes using the 9/10 training set. (e) Use the remaining one part as test set as the validation set of the previously trained model. (f) Repeat b–e 10 times with each part being the test set, until all the samples are predicted once. Finally, (g) merge the results from the test parts and evaluate the metrics.

The cross validation was done using different gene number and the accuracies from each 10-fold cross validation are plotted. For comparison, we also used logistic regression as a baseline model (**Figure 2**). We achieved a good accuracy when the selected gene number is 150. Though a better accuracy could be achieved using the 200 or more as the feature number, the growth curve of number-accuracy is slowing down. The 150 could be seen as a turning point for this curve. Thus, we finally chose the number 150 as the feature number. The results was calculated by averaging the results of 10 times of 10-fold cross validations and showed that the overall accuracy of each cancer type was

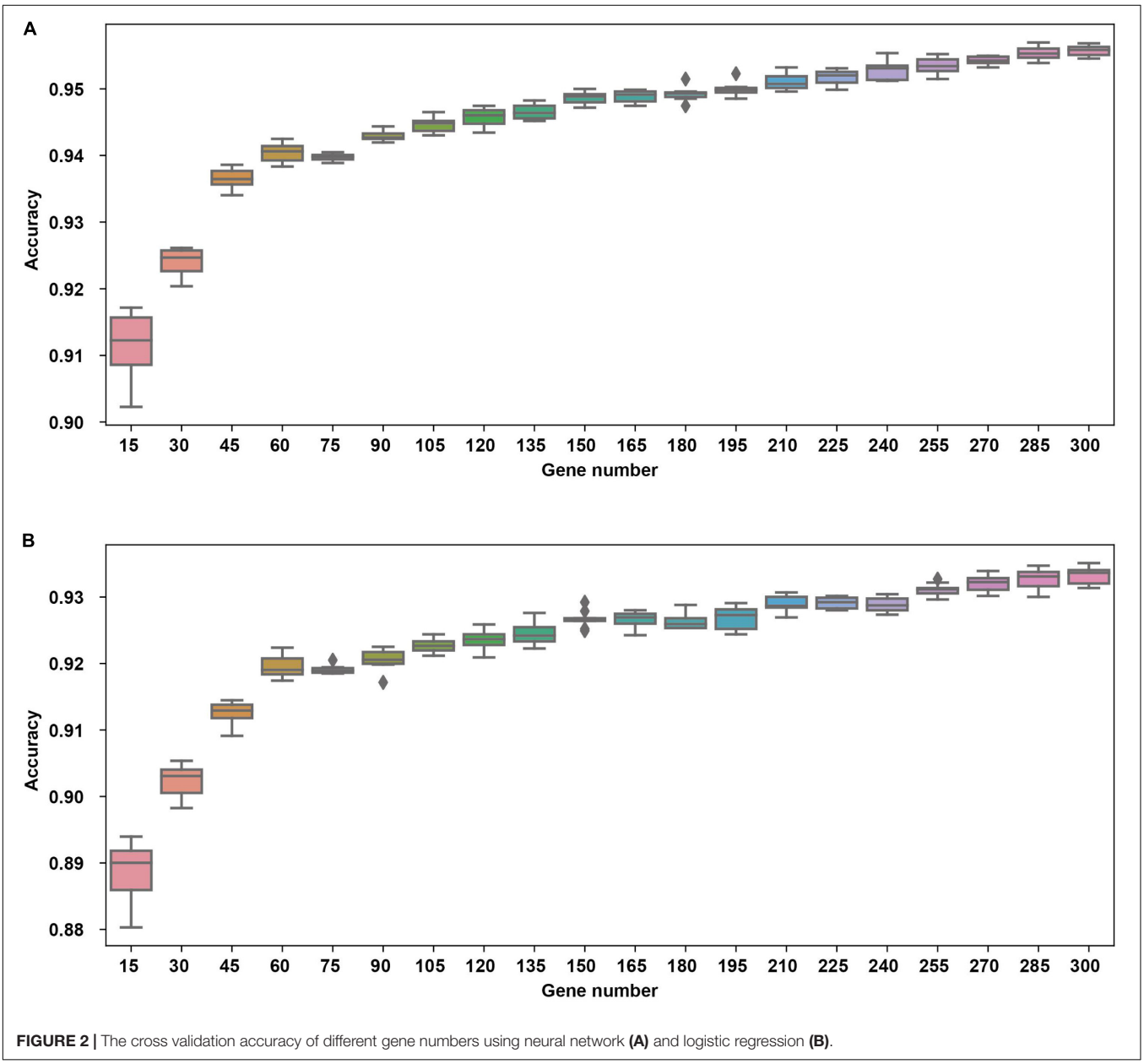


FIGURE 2 | The cross validation accuracy of different gene numbers using neural network **(A)** and logistic regression **(B)**.

TABLE 2 | Sensitivity and precision assessment for each cancer type.

	Sensitivity (%)	Precision (%)
Lung	91.87	92.76
Gastroesophageal	94.89	96.33
Colorectal	98.06	96.88
Liver	97.99	98.80
Breast	98.43	97.98
Thyroid	99.38	99.58
Cervical	71.63	76.38
Brain	99.32	99.41
Pancreatic	91.76	94.63
Ovarian	97.55	97.15
Endometrial	95.54	94.85
Bladder	74.75	88.36
Kidney	98.42	98.54
Head and Neck	90.83	79.39
Prostate	100.00	100.00
Average	93.36	94.07

94.87% using 150 as the feature number; the sensitivity was on average 93.36%, while the precision was on average 94.07%, corresponding to the actual numbers of cancer samples (**Table 2**). Among the 15 cancer types, the classifier sensitivity of 13 cancer types (lung, breast, liver, colorectal, gastroesophageal, ovarian, endometrial, pancreatic, head and neck, thyroid, prostate, kidney, and brain) was more than 90%, with that of prostate cancer having the highest sensitivity (100%). On the contrary, the remaining two cancer types had a sensitivity of <90% (74.75% for bladder cancer and 71.63% for cervical cancer) (**Figure 3** and **Table 2**).

We also attempted to use the log-transformed data for in the cross validation since log-transformation was a common transformation for gene expression profile. For a reasonable comparison, we selected 10 genes for each cancer in each fold of cross validation. However, the overall accuracy by 10 times of 10-fold cross validations only reached 80.90% (**Supplementary Table S1**), which is not satisfactory. In contrast, the data by the previously described transformation method output the result of 94.87%, showing more optimization shall be done for a better result using the log-transformed data.

The Identified Genes Were Enriched in Several Organ-Specific Pathways

A 150-gene set was identified using the whole training dataset for subsequent processing (**Table 3**). To understand how frequently those genes will show up in the cross validation phase, we counted the genes in all the 100 gene sets used in the cross validation and found that 117 genes out of the 150 gene showed up in all gene sets validation, showing the robustness of the feature selection method based on Pearson correlation (**Supplementary Table S2**). To investigate the biological processes of the involved signature genes, GO enrichment analysis was performed. We saw that the most functionally enriched processes related to our 150-gene panel by GO analysis were biological processes (**Figure 4** and **Table 4**). Among those, GO:0048568 Embryonic organ development, GO:0061458 Reproductive system development, GO:0007389 Pattern specification process, GO:0043062 Extracellular structure organization, GO:0002009 Morphogenesis of an epithelium, and GO:0048732 Gland development were related to tissue or organ morphogenesis. Our signature genes were involved in these biological processes and might be useful for classifying distinct cancer types. Hence, the enrichment analysis in the present study might provide a basis to improve our understanding of lung, gastroesophageal, colorectal, liver, breast, thyroid, cervical, brain, pancreatic, ovarian, endometrial, bladder, kidney, head and neck, and prostate cancers.

Reference diagnoses	Predicted cancer type														
	Lung	Gastroesophagus	Colorectal	Liver	Breast	Thyroid	Cervix	Brain	Pancreas	Ovary	Endometrium	Bladder	Kidney	Head and neck	Prostate
Lung	837	0	1	1	7	0	10	0	3	0	5	5	1	44	0
Gastroesophagus	0	395	9	1	0	0	1	0	2	1	0	3	0	3	0
Colorectal	1	6	593	0	0	0	0	0	2	0	0	1	0	1	0
Liver	0	1	0	287	1	0	0	0	0	0	1	1	3	0	0
Breast	2	0	0	0	1040	0	3	1	0	0	1	4	1	4	0
Thyroid	2	0	0	0	0	497	0	0	1	0	0	0	0	0	0
Cervix	14	1	1	0	3	0	191	0	1	0	11	4	0	32	0
Brain	1	0	0	0	1	0	0	588	0	0	1	0	1	0	0
Pancreas	2	5	3	0	1	0	0	0	128	0	1	2	0	0	0
Ovary	1	2	0	0	0	0	0	1	0	255	2	0	0	0	0
Endometrium	4	0	1	0	2	0	4	1	0	5	493	3	1	2	0
Bladder	10	0	3	0	2	0	22	0	0	0	3	228	4	29	0
Kidney	1	0	0	1	2	0	0	1	0	0	2	4	737	0	0
Head and neck	18	0	0	0	2	0	17	0	0	0	0	4	0	439	0
Prostate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	379
Sensitivity	91.58%	95.18%	98.18%	97.62%	98.48%	99.40%	74.03%	99.32%	90.14%	97.70%	95.54%	75.75%	98.53%	91.46%	100.00%
Specificity	99.14%	99.79%	99.74%	99.96%	99.67%	100.00%	99.21%	99.94%	99.88%	99.92%	99.61%	99.57%	99.84%	98.35%	100.00%

FIGURE 3 | Prediction of cancer type by confusion matrix analysis. The confusion matrix is from one 10-fold cross validation and displayed the relationship between reference diagnosis and the predicted cancer type. The first column represents reference diagnoses; the predicted cancer types by transcript levels of the 150 genes are shown across the top row.

TABLE 3 | Gene signatures, as identified by Pearson correlation.

Rank	Lung	Gastro-esophageal	Colo-rectal	Liver	Breast	Ovary	Cervix	Endometrial	Pancreatic	Bladder	Head and Neck	Thyroid	Prostate	Kidney	Brain
1	*CCNJL*	*BRI3BP*	*C2orf89*	*AMBP*	*C18orf45*	*BCAM*	*C1orf14*	*ASRGL1*	*CASR*	*C10orf116*	*BNC1*	*APLP2*	*C17orf93*	*C14orf105*	*BAALC*
2	*ICAM1*	*CCDC109A*	*CDH17*	*APOB*	*EFHD1*	*C10orf41*	*C9orf53*	*DLX5*	*CTRB1*	*FER1L4*	*CSDAP1*	*CTSB*	*HOXB13*	*CNDP2*	*CACNG7*
3	*LAMP3*	*CDC42EP1*	*CDX1*	*APOC2*	*GATA3*	*GPR27*	*CENPW*	*DLX6*	*CTRB2*	*GRHL3*	*GJB5*	*DAPK2*	*KLK2*	*DAB2*	*CTNND2*
4	*LPCAT1*	*GATA6*	*CDX2*	*ASGR1*	*IRX5*	*HOXD4*	*LOC642587*	*FLJ39739*	*CUZD1*	*KRT7*	*KRT14*	*HHEX*	*KLK3*	*FBXO17*	*FEZ1*
5	*NAPSA*	*GNL3L*	*EPS8L3*	*ASGR2*	*LMX1B*	*KCNK15*	*PSMC3IP*	*LOC442459*	*FFAR1*	*LOC100188947*	*KRT5*	*LCN12*	*KLK4*	*GALNT14*	*GPM6B*
6	*ROS1*	*HIATL2*	*GPA33*	*F2*	*PRLR*	*KLHL14*	*RASIP1*	*LOC643387*	*FOXL1*	*PLA2G2F*	*KRT6A*	*MUC15*	*NKX3-1*	*PKD2*	*LRRC4B*
7	*SFTA2*	*PIAS1*	*GPR35*	*ITIH1*	*TBC1D9*	*WIT1*	*SERPINB3*	*LYPLA2P1*	*INS*	*PPARG*	*KRT6B*	*NKX2-1*	*SLC45A3*	*SLC22A2*	*MGC42105*
8	*SFTPA1*	*POU2F1*	*HEPH*	*PROC*	*TFAP2A*	*WT1*	*SERPINB4*	*MSX1*	*PNLIPRP1*	*SNCG*	*PKP1*	*NPC2*	*STEAP2*	*SLC28A1*	*PPP2R2B*
9	*SFTPA2*	*ZBTB7A*	*NOX1*	*SERPINA10*	*TRPS1*	*ZFP92*	*SMC1B*	*SOX17*	*PRSS3*	*UPK1A*	*PTTG3P*	*TSHR*	*TMPRSS2*	*SLC3A1*	*REEP2*
10	*SFTPB*	*ZFPM1*	*VIL1*	*VTN*	*XBP1*	*ZNF503*	*TCAM1P*	*STX18*	*TRY6*	*UPK2*	*SFN*	*ZBED2*	*TRPV6*	*TMEM140*	*SYT11*

The Trained Neural Network Showed High Accuracy on Independent Metastatic Tumor Dataset

We further sought to validate our model on the 20-sample metastatic dataset as a test set. We trained the NN model and the logistic regression model on the whole training dataset using the 150-gene set, which was then used for predicting the test set. The prediction accuracy of NNs was 80%, while the prediction accuracy of the logistic regression model was 70%. The detailed predictions are shown in **Table 5**.

DISCUSSION

Inferring cancer TOO is important for CUP patients and might serve well for minimizing misdiagnosis, even if the cancer origin is diagnosed by pathological observation. Hence, it is critical to develop a method to classify TOO of common cancer types. This study was possible because of the great advancements in NGS technologies and the general application of NGS in clinical experiments, along with the efforts made by researchers who have contributed to the TCGA, from where huge gene expression datasets can be obtained. In the present study, we utilized the NN method to comprehensively analyze high-dimensional RNA-Seq datasets of 15 common cancer types. The 150-gene panel of cancer classifiers demonstrated an average accuracy of 94.87%, corresponding to the actual numbers of cancer samples.

Several hallmarked studies indicated that the cellular origin signatures that are expressed in primary tissue are sufficiently retained even after primary cancer cells undergo dedifferentiation and colonization in different tissue types (Ma et al., 2005; Tothill et al., 2005). A recent study compared four different algorithms and indicated that the modeling performance differed between these algorithms when analyzing RNA-Seq data from 4,127 primary tumor tissue samples related to nine cancer types (Bhowmick et al., 2019). Among those, ABC yielded the best results; it had an average precision of 91.16% and an average sensitivity of 96.5% for nine cancer types (Bhowmick et al., 2019). However, our study demonstrated an average precision of 94.07% and an average sensitivity of 93.36%, corresponding to 7,460 cancer samples related to 15 common cancer types. Although the average sensitivity from our study was a bit lower than that of ABC algorithm, we managed to dramatically minimize the false-positive rate to 0.34% (**Table 2**). Moreover, the overall accuracy with an average of 94.87% is higher than that of other gene expression-based signatures, which ranged from 79–91% (Ma et al., 2005; Monzon et al., 2009; Kerr et al., 2012). Furthermore, the performance of the 150-gene panel was higher than that of the immunohistochemistry technique (75%), which represents the current clinical practice standard, as tested by a 10-antibody panel (Park et al., 2007).

In the present study, GO analysis revealed several over-represented biological processes related to tissue morphogenesis, such as embryonic organ development, reproductive system development, pattern specification process/regionalization, extracellular structure organization, epithelial morphogenesis,

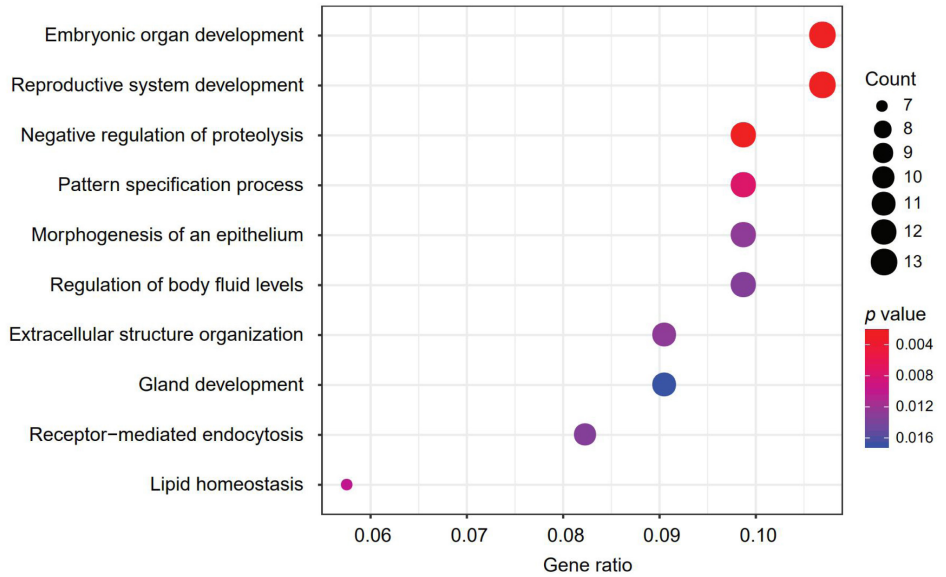


FIGURE 4 | The most represented biological processes associated with our signature genes. Dot plot displaying the number of signature genes involved in each biological process, determined by enrichment analysis. Dot size represents the number of genes, and dot color represents *p*-value; a lower *p*-value represents a higher probability of a biological process being enriched with the signature genes.

and glandular development (**Figure 4** and **Table 4**). Additionally, the expression patterns of several signature genes of the 150-gene panel were previously reported to be related to tissues of specific tumor types. For example, *GRHL3* (*Grainyhead-Like Transcription Factor 3*) encodes a cancer suppressor that is a member of the grainyhead-like transcription factor family (Darido et al., 2011). The downregulated *GRHL3* gene was associated with head and neck squamous cell carcinomas (Frisch et al., 2018); overexpression of the oncogenic mir21 was as result of decreased GRHL3 (Bhandari et al., 2013). In addition, *KLKs* (*Kallikrein-Related Peptidases*) are genes that encode serine proteases that exhibit a deregulated expression in prostate cancer. In our study, *KLK2, KLK3,* and *KLK4* were identified as gene signatures for prostate cancer; KLK3 is a prostate-specific antigen that is a gold-standard clinical biomarker widely employed in the diagnosis and monitoring of prostate cancer (Fuhrman-Luck et al., 2014); KLK2 showed promise as prostate cancer biomarker, as well. Additionally, the deregulated expression of KLKs has been utilized in designing novel therapeutic targets for prostate cancer (Fuhrman-Luck et al., 2014).

GATA DNA-binding proteins, commonly abbreviated as GATAs, are zinc-finger binding transcription factors that regulate tissue differentiation and specification (Chou et al., 2010; Zheng and Blobel, 2010). In our study, GATA3 and GATA6 transcripts were identified as gene signatures for breast cancer and gastroesophageal cancer, respectively. Previous studies have indicated that GATA3 was weakly expressed in a wide variety of normal tissues, while its expression was remarkably elevated in breast cancer (Yang and Nonaka, 2010; Liu et al., 2012); moreover, GATA3 has been identified as a novel clinical marker for detecting primary and metastatic breast cancer (Cimino-Mathews et al., 2013; Krings et al., 2014; Shield et al., 2014; Braxton et al., 2015; Sangoi et al., 2016; Yang et al., 2017). GATA6 was initially cloned from rat gastric tissue, designated as GATA-GT1 (Tamura et al., 1994); however, recent studies have indicated that GATA6 was frequently overexpressed and/or amplified in human gastroesophageal cancer (Sulahian et al., 2014; Chia et al., 2015; Song et al., 2018). There's some limitations about our studies. First, we assessed the model based on NGS RNA-Seq data from the formalin-fixed and paraffin-embedded materials, but not fresh materials. We did not evaluate it in fresh materials mainly due to the formalin-fixed and paraffin-embedded materials are most diagnostic materials in routine practice. Second, some solid tumor cancer types such as sarcoma was not included due to the unavailability of RNAseq data; besides, the non-solid tumors were currently excluded; melanoma was also excluded due to the data scarcity and the distinct distribution of its primary tumor sample number and metastatic tumor sample number. Thus, further efforts should be made for a broader application scope. Third, the training dataset could be further expanded. Since the final gene set contains some organ development-related genes, we can infer that the gene set does not only classify cancer types, but also organs. Staub et al. has already made efforts by expand the training dataset and achieved a better result (Staub et al., 2009). Thus, expression profiles from normal tissues could be further added to our training dataset for a better performance. Another limitation is that our method is based on the expression value without any manipulations. Recently, an algorithm called TSP was applied to this problem, which will generate gene pairs instead of single gene features, giving rise to a leap to the prediction accuracy (Shen et al., 2020). We believe that combining the

TABLE 4 | The overrepresented biological processes associated with identified gene-signatures, as obtained through GO enrichment analysis.

GO biological processes	Lung	Gastro-esophageal	Colo-rectal	Liver	Breast	Thyroid	Cervical	Brain	Pancreatic	Ovarian	Endo-metrial	Bladder	Kidney	Head and neck	Prostate
GO:0048568 Embryonic organ development		✓	✓		✓					✓	✓	✓	✓	✓	
GO:0045861 Negative regulation of proteolysis	✓		✓	✓		✓	✓		✓					✓	
GO:0061458 Reproductive system development	✓	✓	✓		✓	✓				✓		✓	✓	✓	✓
GO:0007389 Pattern specification process			✓		✓	✓				✓	✓	✓	✓		✓
GO:0055088 Lipid homeostasis				✓	✓	✓			✓			✓			
GO:0043062 Extracellular structure organization	✓		✓	✓				✓	✓	✓					✓
GO:0002009 Morphogenesis of an epithelium					✓	✓	✓		✓	✓	✓	✓	✓		✓
GO:0006898 Receptor-mediated endocytosis				✓				✓					✓		✓
GO:0050878 Regulation of body fluid levels		✓		✓	✓							✓	✓	✓	
GO:0048732 Gland development		✓		✓	✓	✓				✓	✓		✓		✓
GO:2001242 Regulation of intrinsic apoptotic signaling pathway			✓		✓	✓			✓		✓				✓
GO:0009755 Hormone-mediated signaling pathway		✓			✓	✓						✓	✓		✓

TABLE 5 | The performance on metastatic samples of the neural network trained on the primary samples.

Sample Id	predicted_ by_NN	predicted_by _logistic	true_label
TCGA-AC-A6IX-06A-11R-A32P-07	BRCA	BRCA	BRCA
TCGA-BH-A18V-06A-11R-A213-07	BRCA	BLCA	BRCA
TCGA-BH-A1ES-06A-12R-A24H-07	BRCA	LIHC	BRCA
TCGA-BH-A1FE-06A-11R-A213-07	KIDNEY	KIDNEY	BRCA
TCGA-E2-A15A-06A-11R-A12D-07	BRCA	BRCA	BRCA
TCGA-E2-A15E-06A-11R-A12D-07	BRCA	BRCA	BRCA
TCGA-E2-A15K-06A-11R-A12P-07	BRCA	BRCA	BRCA
TCGA-HM-A6W2-06A-22R-A33Z-07	UCEC	UCEC	CESC
TCGA-UC-A7PG-06A-11R-A42S-07	CESC	CESC	CESC
TCGA-NH-A8F7-06A-31R-A41B-07	COAD + READ	COADREAD	COAD + READ
TCGA-KU-A6H7-06A-21R-A31N-07	CESC	CESC	HNSC
TCGA-UF-A71A-06A-11R-A39I-07	LUNG	LUNG	HNSC
TCGA-DE-A4MD-06A-11R-A250-07	THCA	THCA	THCA
TCGA-EM-A2CS-06A-11R-A180-07	THCA	THCA	THCA
TCGA-EM-A2P1-06A-11R-A206-07	THCA	THCA	THCA
TCGA-EM-A3FQ-06A-11R-A21D-07	THCA	THCA	THCA
TCGA-EM-A3SU-06A-11R-A22U-07	THCA	THCA	THCA
TCGA-J8-A3O2-06A-11R-A23N-07	THCA	THCA	THCA
TCGA-J8-A3YH-06A-11R-A23N-07	THCA	THCA	THCA
TCGA-J8-A4HW-06A-11R-A250-07	THCA	THCA	THCA

neural network and the feature generation could further improve the performance for CUP problems.

CONCLUSION

In the present study, our 150-gene panel exhibited promising results as a tumor classifier for inferring the origin of tumor tissue. First, we obtained NGS-based RNA-Seq data for 7,460 tumor samples from TCGA. Second, we built a fine pipeline to identify gene signatures based on their transcript-levels for 15 common cancer types. Third, we utilized the Neural Network to evaluate the performance of the genes; on average, the precision was 94.07%, while the sensitivity was 93.36%. In addition, GO enrichment analysis revealed several biological processes, including tissue morphogenesis; notably, most of the gene signatures were involved in key oncogenic pathways, supporting our 150-gene panel. Therefore, the 150-gene biomarker signature in our study might prove to be clinically useful for identifying cancers of unknown origin and confirming initial clinical diagnoses. In future studies, we will focus on the application of this model in metastatic cancer patients, in addition to patients with cancer of unknown origin, to evaluate their therapy outcome.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://dcc.icgc.org/releases/release_26.

AUTHOR CONTRIBUTIONS

GT, JY, and HL conceived the concept of the work. BH, BW, YL, and JL performed the experiments. YZ wrote the manuscript. ZZ, HL, PB, LY, and DS reviewed the manuscript. All authors approved the final version of this manuscript.

FUNDING

This study was partially funded by Hunan Provincial Innovation Platform and Talents Program (No. 2018RS3105), the Natural Science Foundation of China (Nos. 61803151, 81560405, and 81960449), the Natural Science Foundation of Hunan province (Nos. 2018JJ2461, 2018JJ2463, and 2018JJ3570), the Project of Scientific Research Fund of Hunan Provincial Education Department (Nos. 19A060 and 19C0185), and the Talents Science and Technology Program of Changsha (No. kq1907035).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00737/full#supplementary-material

TABLE S1 | The result of 10 times of 10-fold cross validations using 10 genes for each cancer.

TABLE S2 | The gene sets from the cross validation phase and the occurrence in the final gene set.

REFERENCES

Asgari, E., and Mofrad, M. R. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One* 10:e0141287. doi: 10.1371/journal.pone.0141287

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29.

Bhandari, A., Gordon, W., Dizon, D., Hopkin, A. S., Gordon, E., Yu, Z., et al. (2013). The grainyhead transcription factor Grhl3/Get1 suppresses miR-21 expression and tumorigenesis in skin: modulation of the miR-21 target MSH2 by RNA-binding protein DND1. *Oncogene* 32, 1497–1507. doi: 10.1038/onc.2012.168

Bhowmick, S. S., Bhattacharjee, D., and Rato, L. (2019). Identification of tissue-specific tumor biomarker using different optimization algorithms. *Genes Genomics* 41, 431–443. doi: 10.1007/s13258-018-0773-2

Braxton, D. R., Cohen, C., and Siddiqui, M. T. (2015). Utility of GATA3 immunohistochemistry for diagnosis of metastatic breast carcinoma in cytology specimens. *Diagn. Cytopathol.* 43, 271–277. doi: 10.1002/dc.23206

Chia, N. Y., Deng, N., Das, K., Huang, D., Hu, L., Zhu, Y., et al. (2015). Regulatory crosstalk between lineage-survival oncogenes KLF5. GATA4 and GATA6 cooperatively promotes gastric cancer development. *Gut* 64, 707–719. doi: 10.1136/gutjnl-2013-306596

Chopra, P., Lee, J., Kang, J., and Lee, S. (2010). Improving cancer classification accuracy using gene pairs. *PLoS One* 5:e14305. doi: 10.1371/journal.pone.0014305

Chou, J., Provot, S., and Werb, Z. (2010). GATA3 in development and cancer differentiation: cells GATA have it! *J. Cell. Physiol.* 222, 42–49. doi: 10.1002/jcp.21943

Cimino-Mathews, A., Subhawong, A. P., Illei, P. B., Sharma, R., Halushka, M. K., Vang, R., et al. (2013). GATA3 expression in breast carcinoma: utility in triple-negative, sarcomatoid, and metastatic carcinomas. *Hum. Pathol.* 44, 1341–1349. doi: 10.1016/j.humpath.2012.11.003

Darido, C., Georgy, S. R., Wilanowski, T., Dworkin, S., Auden, A., Zhao, Q., et al. (2011). Targeting of the tumor suppressor GRHL3 by a miR-21-dependent proto-oncogenic network results in PTEN loss and tumorigenesis. *Cancer Cell* 20, 635–648. doi: 10.1016/j.ccr.2011.10.014

Frisch, A., Walter, T. C., Grieser, C., Geisel, D., Hamm, B., and Denecke, T. (2018). Performance survey on a new standardized formula for oral signal suppression in MRCP. *Eur. J. Radiol. Open* 5, 1–5. doi: 10.1016/j.ejro.2017.12.002

Fuhrman-Luck, R. A., Loessner, D., and Clements, J. A. (2014). Kallikrein-related peptidases in prostate cancer: from molecular function to clinical application. *Ejifcc* 25, 269–281.

Greco, F. A., Oien, K., Erlander, M., Osborne, R., Varadhachary, G., Bridgewater, J., et al. (2012). Cancer of unknown primary: progress in the search for improved and rapid diagnosis leading toward superior patient outcomes. *Ann. Oncol.* 23, 298–304. doi: 10.1093/annonc/mdr306

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. Waikato: The University of Waikato.

Handorf, C. R., Kulkarni, A., Grenert, J. P., Weiss, L. M., Rogers, W. M., Kim, O. S., et al. (2013). A multicenter study directly comparing the diagnostic accuracy of gene expression profiling and immunohistochemistry for primary site identification in metastatic tumors. *Am. J. Surg. Pathol.* 37, 1067–1075. doi: 10.1097/pas.0b013e31828309c4

Hinton, G. E. (1989). Connectionist learning procedures. *Artif. Intell.* 40, 185–234. doi: 10.1016/0004-3702(89)90049-0

Hua, K. L., Hsu, C. H., Hidayati, S. C., Cheng, W. H., and Chen, Y. J. (2015). Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets Ther.* 8, 2015–2022.

Hudis, C. A. (2007). Trastuzumab–mechanism of action and use in clinical practice. *N. Engl. J. Med.* 357, 39–51. doi: 10.1056/nejmra043186

Hyphantis, T., Papadimitriou, I., Petrakis, D., Fountzilas, G., Repana, D., Assimakopoulos, K., et al. (2013). Psychiatric manifestations, personality traits and health-related quality of life in cancer of unknown primary site. *Psycho Oncol.* 22, 2009–2015. doi: 10.1002/pon.3244

Kamposioras, K., Pentheroudakis, G., and Pavlidis, N. (2013). Exploring the biology of cancer of unknown primary: breakthroughs and drawbacks. *Eur. J. Clin. Invest.* 43, 491–500. doi: 10.1111/eci.12062

Kerr, S. E., Schnabel, C. A., Sullivan, P. S., Zhang, Y., Singh, V., Carey, B., et al. (2012). Multisite validation study to determine performance characteristics of a 92-gene molecular cancer classifier. *Clin. Cancer Res.* 18, 3952–3960. doi: 10.1158/1078-0432.ccr-12-0920

Krings, G., Nystrom, M., Mehdi, I., Vohra, P., and Chen, Y. Y. (2014). Diagnostic utility and sensitivities of GATA3 antibodies in triple-negative breast cancer. *Hum. Pathol.* 45, 2225–2232. doi: 10.1016/j.humpath.2014.06.022

Kurahashi, I., Fujita, Y., Arao, T., Kurata, T., Koh, Y., Sakai, K., et al. (2013). A microarray-based gene expression analysis to identify diagnostic biomarkers for unknown primary cancer. *PLoS One* 8:e63249. doi: 10.1371/journal.pone.0063249

Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. U.S.A.* 101, 811–816. doi: 10.1073/pnas.0304146101

Lazaridis, G., Pentheroudakis, G., Fountzilas, G., and Pavlidis, N. (2008). Liver metastases from cancer of unknown primary (CUPL): a retrospective analysis of presentation, management and prognosis in 49 patients and systematic review of the literature. *Cancer Treat. Rev.* 34, 693–700. doi: 10.1016/j.ctrv.2008.05.005

Liu, H., Shi, J., Wilkerson, M. L., and Lin, F. (2012). Immunohistochemical evaluation of GATA3 expression in tumors and normal tissues: a useful immunomarker for breast and urothelial carcinomas. *Am. J. Clin. Pathol.* 138, 57–64. doi: 10.1309/ajcp5uafmsa9zqbz

Liu, J., Ranka, S., and Kahveci, T. (2008). Classification and feature selection algorithms for multi-class CGH data. *Bioinformatics* 24, i86–i95. doi: 10.1093/bioinformatics/btn145

Ma, X. J., Patel, R., Wang, X., Salunga, R., Murage, J., Desai, R., et al. (2005). Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. *Arch. Pathol. Lab Med.* 130, 465–473.

MacReady, N. (2010). NICE issues guidance on cancer of unknown primary. *Lancet Oncol.* 11:824. doi: 10.1016/s1470-2045(10)70215-1

Meiri, E., Mueller, W. C., Rosenwald, S., Zepeniuk, M., Klinke, E., Edmonston, T. B., et al. (2012). A second-generation microRNA-based assay for diagnosing tumor tissue origin. *Oncol.* 17, 801–812. doi: 10.1634/theoncologist.2011-0466

Molina, R., Bosch, X., Auge, J. M., Filella, X., Escudero, J. M., Molina, V., et al. (2012). Utility of serum tumor markers as an aid in the differential diagnosis of patients with clinical suspicion of cancer and in patients with cancer of unknown primary site. *Tumour Biol.* 33, 463–474. doi: 10.1007/s13277-011-0275-1

Montezuma, D., Azevedo, R., Lopes, P., Vieira, R., Cunha, A. L., and Henrique, R. (2013). A panel of four immunohistochemical markers (CK7, CK20, TTF-1, and p63) allows accurate diagnosis of primary and metastatic lung carcinoma on biopsy specimens. *Virchows Archiv.* 463, 749–754. doi: 10.1007/s00428-013-1488-z

Monzon, F. A., Lyons-Weiler, M., Buturovic, L. J., Rigl, C. T., Henner, W. D., Sciulli, C., et al. (2009). Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin. *J. Clin. Oncol.* 27, 2503–2508. doi: 10.1200/jco.2008.17.9762

Mramor, M., Leban, G., Demsar, J., and Zupan, B. (2007). Visualization-based cancer microarray data classification analysis. *Bioinformatics* 23, 2147–2154. doi: 10.1093/bioinformatics/btm312

Oien, K. A., and Dennis, J. L. (2012). Diagnostic work-up of carcinoma of unknown primary: from immunohistochemistry to molecular profiling. *Ann. Oncol.* 23(Suppl. 10), x271–x277. doi: 10.1093/annonc/mds357

Park, S. Y., Kim, B. H., Kim, J. H., Lee, S., and Kang, G. H. (2007). Panels of immunohistochemical markers help determine primary sites of metastatic adenocarcinoma. *Arch. Pathol. Lab. Med.* 131, 1561–1567.

Pavlidis, N., and Fizazi, K. (2005). Cancer of unknown primary (CUP). *Crit. Rev. Oncol. Hematol.* 54, 243–250.

Pavlidis, N., and Pentheroudakis, G. (2012). Cancer of unknown primary site. *Lancet* 379, 1428–1435.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn Res.* 12, 2825–2830.

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229

Richardson, A., Wagland, R., Foster, R., Symons, J., Davis, C., Boyland, L., et al. (2015). Uncertainty and anxiety in the cancer of unknown primary patient journey: a multiperspective qualitative study. *BMJ Support. Palliat. Care* 5, 366–372. doi: 10.1136/bmjspcare-2013-000482

Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517. doi: 10.1093/bioinformatics/btm344

Sangoi, A. R., Shrestha, B., Yang, G., Mego, O., and Beck, A. H. (2016). The novel marker GATA3 is significantly more sensitive than traditional markers mammaglobin and GCDFP15 for identifying breast cancer in surgical and cytology specimens of metastatic and matched primary tumors. *Appl. Immunohistochem. Mol. Morphol.* 24, 229–237. doi: 10.1097/pai.0000000000000186

Shen, Y., Chu, Q., Yin, X., He, Y., Bai, P., Wang, Y., et al. (2020). TOD-CUP: a gene expression rank-based majority vote algorithm for tissue origin diagnosis of cancers of unknown primary. *Brief. Bioinform.* 8:bbaa031.

Shield, P. W., Papadimos, D. J., and Walsh, M. D. (2014). GATA3: a promising marker for metastatic breast carcinoma in serous effusion specimens. *Cancer Cytopathol.* 122, 307–312. doi: 10.1002/cncy.21393

Song, S. H., Jeon, M. S., Nam, J. W., Kang, J. K., Lee, Y. J., Kang, J. Y., et al. (2018). Aberrant GATA2 epigenetic dysregulation induces a GATA2/GATA6 switch in human gastric cancer. *Oncogene* 37, 993–1004. doi: 10.1038/onc.2017.397

Staub, E., Buhr, H. J., and Grone, J. (2009). WITHDRAWN: predicting the site of origin of tumors by a gene expression signature derived from normal tissues. *Oncogene* 29:3732. doi: 10.1038/onc.2010.184

Suk, H. I., and Shen, D. (2013). "Deep learning-based feature representation for AD/MCI classification," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. MICCAI 2013. Lecture Notes in Computer Science*, Vol. 8150, eds K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab (Berlin: Springer).

Sulahian, R., Casey, F., Shen, J., Qian, Z. R., Shin, H., Ogino, S., et al. (2014). An integrative analysis reveals functional targets of GATA6 transcriptional regulation in gastric cancer. *Oncogene* 33, 5637–5648. doi: 10.1038/onc.2013.517

Tamura, S., Wang, X. H., Maeda, M., and Futai, M. (1994). Gastric DNA-binding proteins recognize upstream sequence motifs of parietal cell-specific genes. *Proc. Natl. Acad. Sci. U.S.A.* 91:4609. doi: 10.1073/pnas.91.10.4609

Tomuleasa, C., Zaharie, F., Muresan, M. S., Pop, L., Fekete, Z., Dima, D., et al. (2017). How to diagnose and treat a cancer of unknown primary site. *J. Gastrointestin. Liver Dis.* 26, 69–79.

Tothill, R. W., Kowalczyk, A., Rischin, D., Bousioutas, A., Haviv, I., van Laar, R. K., et al. (2005). An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Res.* 65, 4031–4040. doi: 10.1158/0008-5472.can-04-3617

Tothill, R. W., Li, J., Mileshkin, L., Doig, K., Siganakis, T., Cowin, P., et al. (2013). Massively-parallel sequencing assists the diagnosis and guided treatment of cancers of unknown primary. *J. Pathol.* 231, 413–423. doi: 10.1002/path.4251

Varadhachary, G. R., Raber, M. N., Matamoros, A., and Abbruzzese, J. L. (2008). Carcinoma of unknown primary with a colon-cancer profile-changing paradigm and emerging definitions. *Lancet Oncol.* 9, 596–599. doi: 10.1016/s1470-2045(08)70151-7

Xu, Q., Chen, J., Ni, S., Tan, C., Xu, M., Dong, L., et al. (2016). Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. *Modern Pathol.* 29, 546–556. doi: 10.1038/modpathol.2016.60

Yang, M., and Nonaka, D. (2010). A study of immunohistochemical differential expression in pulmonary and mammary carcinomas. *Modern Pathol.* 23, 654–661. doi: 10.1038/modpathol.2010.38

Yang, Y., Lu, S., Zeng, W., Xie, S., and Xiao, S. (2017). GATA3 expression in clinically useful groups of breast carcinoma: a comparison with GCDFP15 and mammaglobin for identifying paired primary and metastatic tumors. *Ann. Diagn. Pathol.* 26, 1–5. doi: 10.1016/j.anndiagpath.2016.09.011

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterprofiler: an R package for comparing biological themes among gene clusters. *Omics* 16, 284–287. doi: 10.1089/omi.2011.0118

Zheng, R., and Blobel, G. A. (2010). GATA transcription factors and cancer. *Genes Cancer* 1, 1178–1188. doi: 10.1177/1947601911404223

Conflict of Interest: YZ, BW, YL, JL, HL, JY, and GT were employed by the company Geneis (Beijing) Co., Ltd.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 August 2020
doi: 10.3389/fgene.2020.00835

Genomics Score Based on Genome-Wide Network Analysis for Prediction of Survival in Gastric Cancer: A Novel Prognostic Signature

Zepang Sun†, Hao Chen†, Zhen Han†, Weicai Huang, Yanfeng Hu, Mingli Zhao, Tian Lin, Jiang Yu, Hao Liu, Yuming Jiang and Guoxin Li**

Department of General Surgery, Nanfang Hospital, Southern Medical University, Guangzhou, China

OPEN ACCESS

Edited by:
Xin Maizie Zhou,
Stanford University, United States

Reviewed by:
Hailin Tang,
Sun Yat-sen University Cancer Center (SYSUCC), China
Jie Tian,
Institute of Automation (CAS), China

****Correspondence:***
Yuming Jiang
jiangymbest@163.com
Guoxin Li
gzliguoxin@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *25 April 2020*
Accepted: *10 July 2020*
Published: *06 August 2020*

Citation:
Sun Z, Chen H, Han Z, Huang W, Hu Y, Zhao M, Lin T, Yu J, Liu H, Jiang Y and Li G (2020) Genomics Score Based on Genome-Wide Network Analysis for Prediction of Survival in Gastric Cancer: A Novel Prognostic Signature. Front. Genet. 11:835. doi: 10.3389/fgene.2020.00835

Purpose: Gastric cancer (GC) is a product of multiple genetic abnormalities, including genetic and epigenetic modifications. This study aimed to integrate various biomolecules, such as miRNAs, mRNA, and DNA methylation, into a genome-wide network and develop a nomogram for predicting the overall survival (OS) of GC.

Materials and Methods: A total of 329 GC cases, as a training cohort with a random of 150 examples included as a validation cohort, were screened from The Cancer Genome Atlas database. A genome-wide network was constructed based on a combination of univariate Cox regression and least absolute shrinkage and selection operator analyses, and a nomogram was established to predict 1-, 3-, and 5-year OS in the training cohort. The nomogram was then assessed in terms of calibration, discrimination, and clinical usefulness in the validation cohort. Afterward, in order to confirm the superiority of the whole gene network model and further reduce the biomarkers for the improvement of clinical usefulness, we also constructed eight other models according to the different combinations of miRNAs, mRNA, and DNA methylation sites and made corresponding comparisons. Finally, Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were also performed to describe the function of this genome-wide network.

Results: A multivariate analysis revealed a novel prognostic factor, a genomics score (GS) comprising seven miRNAs, eight mRNA, and 19 DNA methylation sites. In the validation cohort, comparing to patients with low GS, high-GS patients (HR, 12.886; $P < 0.001$) were significantly associated with increased all-cause mortality. Furthermore, after stratification of the TNM stage (I, II, III, and IV), there were significant differences revealed in the survival rates between the high-GS and low-GS groups as well ($P < 0.001$). The 1-, 3-, and 5-year C-index of whole genomics-based nomogram were 0.868, 0.895, and 0.928, respectively. The other models have comparable or relatively poor comprehensive performance, while they had fewer biomarkers. Besides that, DAVID 6.8 further revealed multiple molecules and pathways related to the genome-wide network, such as cytomembranes, cell cycle, and adipocytokine signaling.

Conclusion: We successfully developed a GS based on genome-wide network, which may represent a novel prognostic factor for GC. A combination of GS and TNM staging provides additional precision in stratifying patients with different OS prognoses, constituting a more comprehensive sub-typing system. This could potentially play an important role in future clinical practice.

Keywords: gastric cancer, genome-wide network, miRNA, mRNA, DNA methylation, nomogram

BACKGROUND

Gastric cancer (GC) is one of the most common malignant human tumors and the third leading cause of cancer-related mortalities worldwide. Reports estimate that nearly one million new cases and 800,000 deaths occur each year across the world (Torre et al., 2015). Despite the rapid research advancement, GC-related impacts on human life remain high around the globe. According to the global cancer burden data, hundreds of billions of dollars in economic losses are incurred each year due to GC. At the same time, stomach cancer has been reported to cause 19.1 million disability-adjusted life years, with 98% of these resulting from years of life lost and 2% from years lived with disability (Global Burden of Disease Cancer Collaboration et al., 2019).

Despite major breakthroughs in GC prevention, diagnosis, and treatment therapies reported over the past decade, prognosis remains a challenge at different TNM stages (Jiang et al., 2018a; Sun et al., 2019a,b). Notably, patients with similar clinical features and at the same tumor stage who receive uniform treatment have exhibited varying clinical outcomes (Bang et al., 2012; Jiang et al., 2018b). Such evidence indicates the existing challenges to traditional TNM staging (Serra et al., 2019), possibly due to a lack of molecular tools for effectively predicting the prognosis and the therapeutic effect of GC patients. Therefore, more rigorous and reliable systems that accurately reflect the heterogeneity of different patients and guide the development of treatment approaches are urgently needed (Duarte et al., 2018; Serra et al., 2019).

Tumors are a product of multiple genetic mutations, including genetic (gene expression) and epigenetic (DNA methylation and histone modification) modifications, as well as deregulations of tumor-suppressor genes and proto-oncogenes (Anna et al., 2018; Choi et al., 2019). In addition, changes in a set of genetic materials have been closely associated with cancer outcomes (Anna et al., 2018; Choi et al., 2019). Therefore, to effectively predict the prognosis of tumors, such as GC, a single biomarker is insufficient, necessitating the need for a gene network.

A variety of mRNAs have been associated with GC prognosis (Camargo et al., 2019), with microRNAs (miRNAs) also implicated in tumor prediction in the recent years (Li et al., 2010; Ueda et al., 2010; Camargo et al., 2019). These small, non-coding RNAs, comprising 22 nucleotides, primarily function to regulate protein translation by inhibiting the expression of target messenger RNAs (mRNAs). Apart from genetics, epigenetics is currently receiving considerable research attention. DNA methylation is the most common epigenetic event associated with cancer development and progression (Anna et al., 2018). Consequently, numerous studies have implicated DNA methylation in the diagnosis and the prognosis of various tumors, including GC (Camargo et al., 2019; Choi et al., 2019). Although these studies have revealed several biomarkers that have proved to be prognostic predictors in GC, only a handful have been adopted in clinical therapies or are used to build predictive models for the disease (Anna et al., 2018; Duarte et al., 2018; Camargo et al., 2019; Choi et al., 2019; Serra et al., 2019).

Previous studies have identified and recommended numerous biomarkers for GC. However, since malignant tumors often involve multiple layers and different levels of genetic changes, including the genome, transcriptome, and proteome, or even epigenetic content, selecting reasonable candidate factors from tens of thousands of biomarkers and comprehensively analyzing them as an independent feature is imperative to effectively develop a suitable prognostic target. Therefore, genetic networks containing a panel of abnormal factors from different regulatory levels represent the best chance for achieving prognostic value.

The whole genome-wide network analysis is reported in several other cancers, such as colorectal cancer, breast cancer, and lung cancer (Hou et al., 2018; Zhang et al., 2018), and it shows great value in differentiating the prognosis of these patients. Therefore, it is feasible and advantageous to apply genome-wide network analysis to GC.

In the current study, we performed a series of sophisticated statistical analyses and identified 33 genetic molecules that were highly correlated with the prognosis of GC. Specifically, we screened The Cancer Genome Atlas (TCGA), a genome project with 33 types of cancer, including gene expression, and DNA methylation as well as other biological information. Furthermore, we extended these independent prognostic factors to the "omics" concept. Then, a genome-wide network was constructed. Interestingly, the genomics score (GS) obtained herein could supplement TNM staging and enhance the prognostic value of different patients. Moreover, we developed multiple prognostic models, then validated, and compared them to ascertain their strengths and weaknesses. Finally, we performed pathway enrichment and gene oncology annotation analyses to elucidate the function of this gene network.

MATERIALS AND METHODS

Data Acquisition and Preprocessing

Level 3 data were downloaded from the TCGA database using TCGA-Assembler Module A, in January 2019, which was then pretreated with Module B. The dataset comprised of clinical variables from 443 patients, including age, sex, stage,

primary site, grade, treatment, and survival, as well as associated genome-wide data. In addition, the expression levels of 1,871 miRNAs, 20,531 mRNA, and 485,577 DNA methylation sites (Illumina methylation 450) were obtained from 384, 377, and 394 patients, respectively. Afterward, an intersection with a total of 332 samples was eventually retained. Furthermore, patients with missing active follow-up data were excluded from the analysis, leaving 329 patients in the final cohort (**Figure 1**). Moreover, genome-wide level 3 data whose expression levels for miRNAs, mRNA, and DNA methylation sites were missing in more than 50% of all samples were excluded from the final analysis. Finally, 329 GC patients with 566 miRNAs, 17,963 mRNA, and 396,081 DNA methylation sites were chosen for further analysis.

Genome-Wide Network Analysis

Gene expression and DNA methylation data were normalized using R package before subsequent processing. Univariate and least absolute shrinkage and selection operator (LASSO) Cox regression models were combined and used to identify the most useful prognostic factors in miRNAs, mRNA, and DNA methylation sites associated with survival. Firstly, univariate Cox regression was performed on each candidate miRNA, mRNA, and DNA methylation site to elucidate its role in patient survival, then signatures with *P* value less than 0.05 were retained for subsequent analysis. Thereafter, the LASSO Cox regression model was applied to select and shrink the data (**Supplementary Figure S3**; Tibshirani, 1997). Finally, a GS, based on a genome-wide network comprising seven miRNAs, eight mRNAs, and 19 DNA methylation sites, was constructed for predicting survival. A summary of the whole screening process is displayed in **Supplementary Figure S1**.

Development and Comparison of Individualized Prediction Models

The TCGA cohort with 329 cases was used as the training set, with a random 150 cases from the total cohort included as a validation group. The random number is 1,356. Firstly, we developed the original GS based on 34 biomarkers (seven miRNAs, eight mRNAs, and 19 DNA methylation sites). Then, considering the complexity of the original GS and difficult clinical application, in order to obtain a more concise and

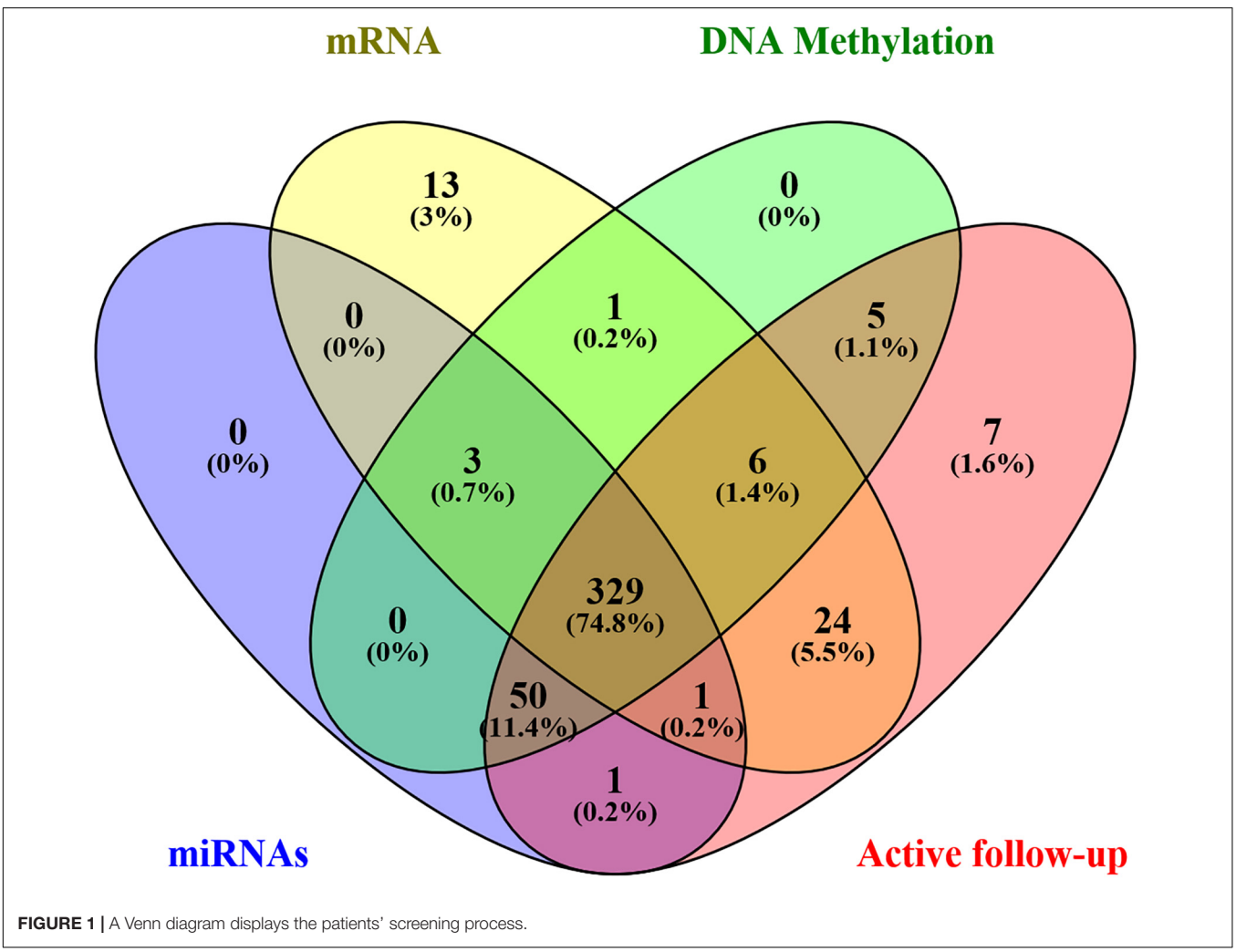


FIGURE 1 | A Venn diagram displays the patients' screening process.

effective GS, we also constructed eight other models according to the different combinations of miRNAs, mRNA, and DNA methylation and made corresponding comparisons. Finally, a total of nine GS models based on the genome-wide network from LASSO were adopted to screen for the most appropriate markers. These included the following models: genomics (seven miRNAs, eight mRNA, and 19 DNA methylation sites), miRNAs (seven miRNAs), mRNA (eight mRNA), methylation (19 DNA methylation sites), miRNAs + methylation (seven miRNAs and 19 DNA methylation sites), miRNAs + mRNA (seven miRNAs and eight mRNA), mRNA + methylation (eight mRNA and 19 DNA methylation sites), Cox-model 1 (two miRNAs, six mRNA, and nine DNA methylation sites), and Cox-model 2 (one miRNA, one mRNA, and seven DNA methylation sites). Among them, markers from Cox-model 1 were separately detected from miRNAs, mRNA, or DNA methylation sites using multivariate Cox regression analysis after LASSO (**Supplementary Tables S2–S4**). On the other hand, markers from Cox-model 2 depended on signatures from a multivariate Cox regression analysis combining the genome-wide network and the clinical characteristics (**Supplementary Table S5**). Thereafter, we constructed several nomograms by incorporating significant ($P < 0.05$) GS variables and other clinical features following multivariate Cox regression (Iasonos et al., 2008), and a clinical nomogram was built as a blank control. The equations used for calculating the GS of these models are listed in **Supplementary Table S6**.

To calculate the discrimination and the stability of different Cox regression models, we applied C-statistics and calibration. Additionally, we performed an analysis of time-dependent receiver operator characteristics (ROC), based on the 1-, 3-, and 5-year survival endpoints, to assess the prognostic accuracy of the different nomograms. Furthermore, we evaluated the potential net benefit of different predictive models using decision curve analysis (DCA). DCA compares the clinical usefulness of different indicators by calculating the potential net benefit of each decision strategy at each threshold probability. Thus, DCA was a significant novel approach for comparing the old and the new models (Vickers and Elkin, 2006).

Screening for Potential miRNA Target Genes

We predicted the potential target genes of the seven miRNAs, from LASSO, by screening the miRTarBase, miRDB, and TargetScan databases. Common genes from each miRNA across the three databases were then used for subsequent studies. More than 90% of the miRNAs showed negative regulation to target genes. Consequently, the expression data from TCGA were used to perform a batch of correlation analysis of each miRNA, with corresponding target genes, and the three genes with the largest absolute negative correlation were retained as the most likely targets. Additionally, at least three potential target genes from miRTarBase, which is co-expressed with miRNAs, were considered as equally important and were subjected to Cytoscape (version 3.7.2) for identification of miRNA–target genes co-expression network analysis (**Supplementary Figure S2**).

Functional Enrichment Analysis

The potential target genes that were negatively correlated with miRNAs in TCGA, as well as the coding sequences for mRNA and DNA methylation sites, were used for functional enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology (GO) using DAVID 6.8 (**Supplementary Figure S2**). Functional enrichment analysis indicates why the gene network produces images on the survival of GC from a molecular mechanism. Visualization was then done using the "ggplot2" package implemented in R.

Statistical Analysis

The patients were divided into low-risk and high-risk groups by the median GS as the cutoff point. Survival estimates were obtained according to the Kaplan–Meier method and compared using the log-rank test. Variables that reached significance, with $P < 0.05$, were entered into the multivariable analyses using the Cox proportional hazards model, with an entry stepwise approach to identify covariates associated with increased all-cause mortality, and then hazard ratio with 95% confidence intervals (CIs) of each variable was achieved. All the statistical significance values were set as two-sided ($P < 0.05$). LASSO Cox regression was performed through the "glmnet" package. Time-dependent ROC analysis at different follow-up times was implemented using the "timeROC" package of R project in order to further expound the performance of different GS models, and DCA was used to compare their clinical use by "rmda" package. Finally, nomogram based on the Cox regression model was constructed using the "rms" package. C-index and calibration to calculate the discrimination and the stability of these models were performed using c-statistics and Bootstrap sample. Harrell's concordance index (C-index) indicated a better prognostic model if its value was closer to 1, and the calibration diagram showed that the better the prediction if the closer the correction line was to the diagonal. All statistical methods are applied to both the training group and the validation group. Statistical analyses were performed using SPSS statistical software (version 18.0) and R software (version 3.5.3).

RESULTS

Patient Characteristics

Among the 329 GC patients analyzed in this study, 212 (64.4%) were male, whereas 117 (35.6%) were female. The average age of the entire study population was 65.0 ± 10.6 years. In terms of pathological stage, 38 (11.6%) cases were identified as stage I, 117 (35.6%) were stage II, 155 (47.1%) were stage III, and 19 (5.8%) were at stage IV. With regards to treatment, 303 (92.1%) patients received surgery (280 cases of R0 surgery, 14 R1, and nine R2), whereas 146 (44.4%) were subjected to fluorouracil-based chemotherapy. The genomic nomogram classified 165 samples into low GS (GS ≤ -0.137) and 164 into high GS (GS > -0.137) groups based on the median cutoff (**Figure 2**). A detailed description of tumor location, pathology grade, and Lauren classification is outlined in **Supplementary Table S1**, while a heat map of the genomic scores layered by clinicopathological

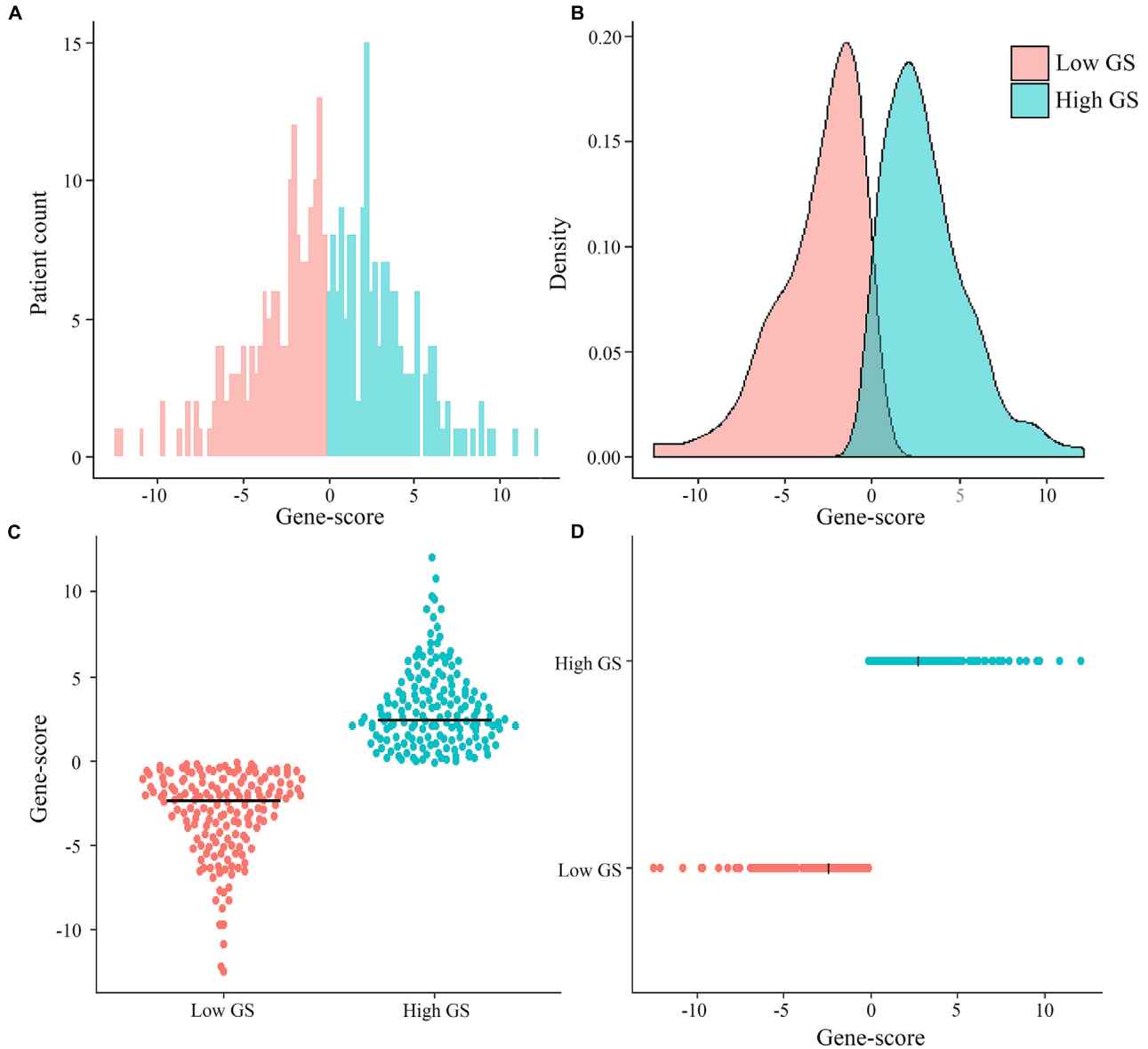


FIGURE 2 | Distribution of patient cases and density based on genomics score (GS) in the total The Cancer Genome Atlas cohort **(A,B)**. Scatter plots of genomics scores regarding the classification of low and high GS **(C,D)**, and the bold line represents the median.

factors is illustrated in **Supplementary Figure S4**. The median (mean; 95% CI) survival time for OS was 1,043 (670.2–1,415.8) days in the total cohort, 466 (370.6–561.4) days in the high-GS group, and 2,613 (mean; 2209.4–2017.5) days in the low-GS group (**Supplementary Figure S5**). Toward the last follow-up, a total of 129 deaths and 200 censoring were recorded. The estimated cumulative 1-, 3-, and 5-year OS in the total cohort were 78.9, 48.4, and 36.7%, respectively, although these rates increase to 95.1, 74.2, and 68.5%, respectively, in the low-GS group. Conversely, the 1-, 3-, and 5-year OS decreased to 63.3, 23.6, and 15.4%, respectively, in the high-GS group. The baseline information of the validation cohort is also listed in **Supplementary Table S1** and **Supplementary Figure S6**.

Survival Analysis

We identified a basic genome-wide network comprising seven miRNAs, eight mRNAs, and 19 DNA methylation sites as the prognostic factor for OS, from hundreds of thousands of univariate Cox regression and LASSO analyses. This network was then classified as other models in the training and the validation groups. Among the 34 features identified, poor prognosis was significantly associated with a high expression of seven miRNAs (hsa-mir-100, hsa-mir-1304, hsa-mir-136, hsa-mir-193b, hsa-mir-22, hsa-mir-653, and hsa-mir-6808), six mRNAs (NRP1|8829, RNF144A|9781, ZNF22|7570, DUSP1|1843, CPNE8|144402, MAGED1|9500, and LOC91450|91450), and seven DNA methylation sites

(cg07020967, cg08859156, cg12485556, cg15861578, cg15861578, cg25161386, and cg22740006). Conversely, poor prognosis was strongly associated with a low expression of SOX14|8403 and 12 DNA methylation sites, including cg02223323, cg00481239, cg14791193, cg15486740, cg20100408, cg20350671, cg22395807, cg24361571, cg25361506, cg22813794, cg26014401, and cg26856948 (**Table 2**). Univariate analysis performed on clinical characteristics revealed a significant association between age, pathological stage, TNM, and surgery with OS (**Table 1**). On the other hand, results from multivariable Cox regression showed that age, pathological stage, and GS were significantly associated with all-cause mortality in GC (**Table 1** and **Figure 3**). Furthermore, stratification of the pathological stage (I, II, III, and IV) revealed significant differences in survival rates between the high-GS and the low-GS groups (**Figure 3**). A similar result was found when the data were stratified by demographic variables (sex and age), clinical characteristics (primary site, grade, and Lauren classification) as well as treatments (surgery and chemotherapy; **Supplementary Figures S7, S8**). On the other hand, categorizing GS into high or low groups, using the median value across different models, indicated that the genomics nomogram had the highest HR value. Interestingly, HR was almost equal to miRNAs + methylation, mRNA + methylation, Cox-model 1, and Cox-model 2 nomograms, which contained fewer gene features. Moreover, the HR value showed a marked decrease in miRNAs, mRNA, methylation, and miRNAs + mRNA nomograms, which included the least characteristics (**Table 3**).

Nomograms Based on Genome-Wide Network

A genomics nomogram was first constructed based on the genome-wide network, comprising 34 gene features (**Figure 4**). To obtain a more concise and effective nomogram, we also built a Cox-model 1 (17 gene features) and Cox-model 2 (nine gene features) nomograms (**Supplementary Figures S9, S10**). Next, a clinical nomogram, based on stage and age, was built as a control (**Supplementary Figure S11**). Thereafter, we performed internal and external validation to evaluate the feasibility of all nomograms using a three-grouped random bootstrap sampling (**Figure 5** and **Supplementary Figures S9–S11**). We observed good predictive performance in the first three nomograms, but not in the simple clinical model.

Validation of the Nomograms Using ROC and DCA

To ensure a good comparison across different GS nomograms, we performed a time-dependent ROC (at 1, 3, and 5 years of follow-up) as well as DCA. In the validation group, genomics nomogram revealed the best comprehensive performance, with 1-, 3-, and 5-year area under the curve (AUC) values of 0.868, 0.895, and 0.928, respectively (**Table 4**), and Cox-model 1, miRNAs + methylation, and mRNA + methylation nomograms had a comparable performance, with 1-, 3-, and 5-year AUC values of 0.856–0.873, 0.884–0.905, and 0.907–0.919, respectively, but it had fewer biomarkers (**Table 4**). Although the Cox-model 2 nomograms had the least biomarkers, including miRNA, mRNA, and DNA methylation sites, it had a relatively poor performance with 1-, 3-, and 5-year AUC values of 0.835, 0.859, and 0.785, respectively. Besides that, the miRNA, mRNA, methylation, miRNAs + methylation, and miRNAs + mRNA nomograms recorded 1-, 3-, and 5-year AUC values of 0.729–0.877, 0.656–0.805, and 0.721–0.894, respectively. Finally, we found that, compared to miRNA (0.641, 0.729, and 0.736) and mRNA nomogram (0.806, 0.785, and 0.843), methylation nomogram had higher 1-, 3-, and 5-year AUC values of 0.866, 0.877, and 0.894. Nevertheless, all of them showed better performance than the clinical nomogram, which recorded 1-, 3-, and 5-year AUC values of 0.638, 0.598, and 0.721, respectively (**Figures 6A,B** and **Supplementary Figure S12**). The C-index based on different nomograms exhibited a similar effect (**Supplementary Table S8**). Additionally, DCA showed that the genomics, Cox-model 1, mRNA + methylation, and methylation nomograms had a significant net benefit compared to other GS models and the clinical nomogram (**Figures 6C,D**).

Potential miRNA Target Genes

A total of 72 hsa-mir-22, 39 hsa-mir-100, 56 hsa-mir-136, 58 hsa-mir-193b, 23 hsa-mir-653, 96 hsa-mir-1304, and 285 hsa-mir-6808 potential target genes were identified from the miRTarBase, miRDB, and TargetScan databases (**Supplementary Figure S14**). We then performed a correlation analysis between each target gene and miRNAs and finally generated a miRNA–potential target gene plot (**Supplementary Figure S15A**) as well as a miRNA–target gene co-expression network (**Supplementary Figure S15B**) using Cytoscape.

Functional Analysis of Genome-Wide Network

We imported the 301 potential target genes, mRNA, and DNA methylation site-coding sequences, identified above, into DAVID for KEGG and GO analyses and identified biological processes, molecular functions as well as cellular components (**Figures 7A–C**). Their corresponding KEGG pathways were also plotted (**Figure 7D**).

DISCUSSION

GC can be divided into two types or four main categories, according to the Lauren and World Health Organization (WHO) classifications (Lauren, 1965; Nagtegaal et al., 2019), although neither of these classifications is based on molecular markers. In the last decade, however, three novel molecular-based classification systems have been suggested for GC. The Singapore-Duke Group was the first to describe a classification with two intrinsic genomic subtypes, G-INT, and G-DIF, which had different gene expression (Tan et al., 2011; Serra et al., 2019). The subtypes have different levels of resistance to various chemotherapy drugs and show limited prognostic value. Later, TCGA used molecular evaluation to propose a new classification with four subtypes: EBV, MSI, GS, and CIN. The identification of these subtypes has provided a roadmap for patient stratification

TABLE 1 | Univariable and multivariable analyses of the genomics score and the clinicopathological characteristics for overall survival in the training group and the validation group.

Variables	Training group, *n* = 329						Validation group, *n* = 150					
	Univariable analysis			Multivariable analysis			Univariable analysis			Multivariable analysis		
	HR	95% CI	*P* value	HR	95% CI	*P* value	HR	95% CI	*P* value	HR	95% CI	*P* value
Age at diagnosis, tears												
<65	1	1	NA	1	1	NA	1	1	NA	1	1	NA
≥65	1.674	1.167–2.402	0.005	2.043	1.407–2.966	<0.001	1.464	0.856–2.505	0.164	2.029	1.086–3.789	0.026
Pathological stage												
I	1	1	NA	1	1	NA	1	1	NA	1	1	NA
II	1.627	0.784–3.377	0.192	1.644	0.777–3.481	0.194	1.412	0.510–3.905	0.507	1.377	0.480–3.951	0.552
III	2.240	1.116–4.496	0.023	1.880	0.924–3.825	0.082	1.637	0.635–4.219	0.308	1.926	1.344–2.487	0.878
IV	7.801	3.247–18.745	<0.001	5.119	1.832–14.303	0.002	8.106	2.527–26.005	<0.001	9.364	2.267–38.672	0.002
Surgery												
R0	1	1	NA	1	1	NA	1	1	NA	1	1	NA
R1	1.556	0.755–3.209	0.231	1.214	0.578–2.547	0.608	1.240	0.286–5.372	0.774	0.937	0.209–4.209	0.932
R2	6.944	3.163–15.246	<0.001	1.686	0.615–4.621	0.310	12.906	3.796–43.886	<0.001	1.316	0.285–6.075	0.725
Unknown	2.373	1.347–4.182	0.003	2.006	1.115–3.607	0.020	2.309	1.030–5.175	0.042	2.000	0.852–4.693	0.111
Genomics score[a]												
Low	1	1	NA	1	1	NA	1	1	NA	1	1	NA
High	6.304	4.079–9.744	<0.001	6.093	3.910–9.493	<0.001	10.906	5.452–21.817	<0.001	12.886	6.158–26.963	0.000
T staging												
T1	1	1	NA				1	1	NA			
T2	7.604	1.022–56.585	0.048				5.008	0.638–39.304	0.125			
T3	7.278	1.003–52.802	0.050				3.895	0.524–28.976	0.184			
T4	9.473	1.312–68.368	0.026				3.951	0.536–29.143	0.178			
N staging												
N0	1	1	NA				1	1	NA			
N1	1.424	0.869–2.335	0.161				1.563	0.722–3.393	0.257			
N2	1.642	0.930–2.898	0.087				1.612	0.658–3.953	0.296			
N3	2.200	1.369–3.535	0.001				2.509	1.252–5.029	0.009			
M staging												
M0	1	1	NA				1	1	NA			
M1	4.224	2.309–7.726	<0.001				5.499	2.446–12.364	<0.001			
Sex												
Female	1	1	NA				1	1	NA			
Male	1.449	0.989–2.123	0.057				1.126	0.648–1.956	0.674			
Primary site												
Cardia	1	1	NA				1	1	NA			
Fundus/body	0.844	0.543–1.312	0.451				0.605	0.320–1.144	0.122			
Antrum	0.822	0.530–1.274	0.380				0.763	0.394–1.476	0.422			
Unknown	0.183	0.025–1.343	0.095				0.152	0.003–0.254	0.976			
Pathology grade												
I–II	1	1	NA				1	1	NA			
III–IV	1.361	0.939–1.971	0.103				1.590	0.902–2.805	0.109			
Unknown	1.881	0.673–5.257	0.228				2.534	0.854–7.524	0.094			
Lauren classification												
Intestinal type	1	1	NA				1	1	NA			
Diffused type	1.245	0.805–1.925	0.326				1.416	0.728–2.756	0.305			
Unknown	1.156	0.770–1.736	0.484				1.923	1.061–3.485	0.031			
Chemotherapy												
Yes	1	1	NA				1	1	NA			
No	1.305	0.919–1.852	0.136				1.361	0.806–2.299	0.249			

[a]*Based on 34 biomarkers: seven miRNAs, eight mRNAs, and 19 DNA methylation sites.*

TABLE 2 | miRNAs, mRNA, and DNA methylation whose expression levels showed a significant association with overall survival in least absolute shrinkage and selection operator.

Molecular (probe)	ID (reference gene)	Coefficient	HR	95% CI	SE	*z* value	*p* value
miRNAs	hsa-mir-100	0.234	1.263	1.102–1.449	0.070	3.345	<0.001
	hsa-mir-1304	0.113	1.120	1.006–1.247	0.055	2.060	0.039
	hsa-mir-136	0.235	1.265	1.097–1.458	0.072	3.243	0.001
	hsa-mir-193b	0.241	1.272	1.116–1.450	0.067	3.612	<0.001
	hsa-mir-22	0.248	1.281	1.101–1.490	0.077	3.210	0.001
	hsa-mir-653	0.148	1.160	1.046–1.287	0.053	2.804	0.005
	hsa-mir-6808	0.180	1.197	1.027–1.396	0.078	2.297	0.022
mRNA	NRP1|8829	0.291	1.338	1.178–1.519	0.065	4.492	<0.001
	RNF144A|9781	0.313	1.367	1.186–1.576	0.073	4.313	<0.001
	ZNF22|7570	0.302	1.353	1.160–1.579	0.079	3.843	<0.001
	SOX14|8403	−0.464	0.629	0.470–0.841	0.148	−3.126	0.002
	DUSP1|1843	0.360	1.434	1.240–1.657	0.074	4.874	<0.001
	CPNE8|144402	0.342	1.407	1.203–1.646	0.080	4.269	<0.001
	MAGED1|9500	0.291	1.338	1.165–1.537	0.071	4.112	<0.001
	LOC91450|91450	0.278	1.320	1.156–1.509	0.068	4.083	<0.001
cg02223323	MAP7D2	−0.360	0.697	0.590–0.824	0.085	−4.225	<0.001
cg00481239	SHC4;EID1	−0.668	0.513	0.338–0.777	0.212	−3.153	0.002
cg07020967	TMEM117	0.380	1.462	1.215–1.759	0.094	4.026	<0.001
cg08859156	RPS4X	0.409	1.505	1.305–1.736	0.073	5.609	<0.001
cg12485556	PREP	0.435	1.545	1.196–1.994	0.130	3.333	<0.001
cg14791193	C1orf144	−0.400	0.671	0.581–0.773	0.073	−5.496	<0.001
cg15861578	ZC3H10	0.329	1.390	1.169–1.652	0.088	3.731	<0.001
cg15486740	ACOT13;TTRAP	−0.363	0.695	0.573–0.843	0.098	−3.697	<0.001
cg20100408	HLA-DPB1	−0.357	0.700	0.605–0.810	0.074	−4.799	<0.001
cg20350671	IL1RAPL1	−0.390	0.677	0.561–0.817	0.096	−4.080	<0.001
cg22395807	ATXN10	−0.443	0.643	0.476–0.867	0.153	−2.898	0.004
cg24361571	MIR365-2	−0.340	0.712	0.611–0.829	0.078	−4.374	<0.001
cg25361506	Unconfirmed	−0.362	0.696	0.587–0.825	0.087	−4.167	<0.001
cg25622155	Unconfirmed	0.331	1.392	1.176–1.647	0.086	3.851	<0.001
cg25161386	NUFIP2	0.305	1.357	1.159–1.590	0.081	3.787	<0.001
cg22740006	PC;LRFN4	0.342	1.407	1.149–1.723	0.103	3.303	<0.001
cg22813794	STYXL1;MDH2	−0.351	0.704	0.508–0.976	0.166	−2.106	0.035
cg26014401	Unconfirmed	−0.430	0.651	0.539–0.786	0.097	−4.453	<0.001
cg26856948	GOLGA3	−0.334	0.716	0.609–0.842	0.083	−4.042	<0.001

as well as targeted therapeutic trials (Cancer Genome Atlas Research, 2014). However, initial data on disease outcomes from this cohort did not show differences in survival among the four groups. A series of positive studies on prognosis based on TCGA classification was also reported (Sohn et al., 2017). In addition, the Asian Cancer Research group divided GC into four subtypes, MSI, EMT, MSS/TP53+, and MSS/ TP53-, based on gene expression data and found significantly different survival outcomes across them (Cristescu et al., 2015; Serra et al., 2019). Despite the significant milestones of these studies, they are all mainly based on the analysis of gene expression (mRNA). Besides that, a 2019 study proposed a five-miRNA model, while it had a C-index of 0.72 only (Zhang et al., 2019). In the current study, we included methylation data and performed functional enrichment analysis, making our work stronger. The aforementioned classifications are also complicated and need further optimization to increase clinical applicability. Furthermore, they focused on typing and finding new targets, whereas our study reports on prognostic analysis.

Some of the biomarkers we identified herein, including hsa-mir-22, hsa-mir-100, hsa-mir-136, hsa-mir-193b, hsa-mir-1304, NRP1, DUSP1, and MAP7D2 (cg02223323), have previously been reported in GC (Grandclement and Borg, 2011; Chen et al., 2014; Mu et al., 2014; Zuo et al., 2015; Zheng et al., 2017; Chen et al., 2018; Kurata and Lin, 2018; Liu K.T. et al., 2018; Song et al., 2018; Teng et al., 2018; Liu et al., 2019; Pan et al., 2019; Wang et al., 2019). Others, such as CPNE8, MAGED1, RNF144A, SOX14, ACOT13 (cg15486740), EID1 (cg00481239), RPS4X (cg08859156), and TTRAP (cg15486740), have been identified in various tumors other than GC (Kamio et al., 2010; Zeng et al., 2012; Zhou et al., 2013; Kuang et al., 2017; Stanisavljevic et al., 2017; Liu X. et al., 2018; Tosic et al., 2018; Nagasawa et al., 2019; Yang et al., 2019). The remaining biomarkers, including hsa-mir-653, hsa-mir-6808, LOC91450,

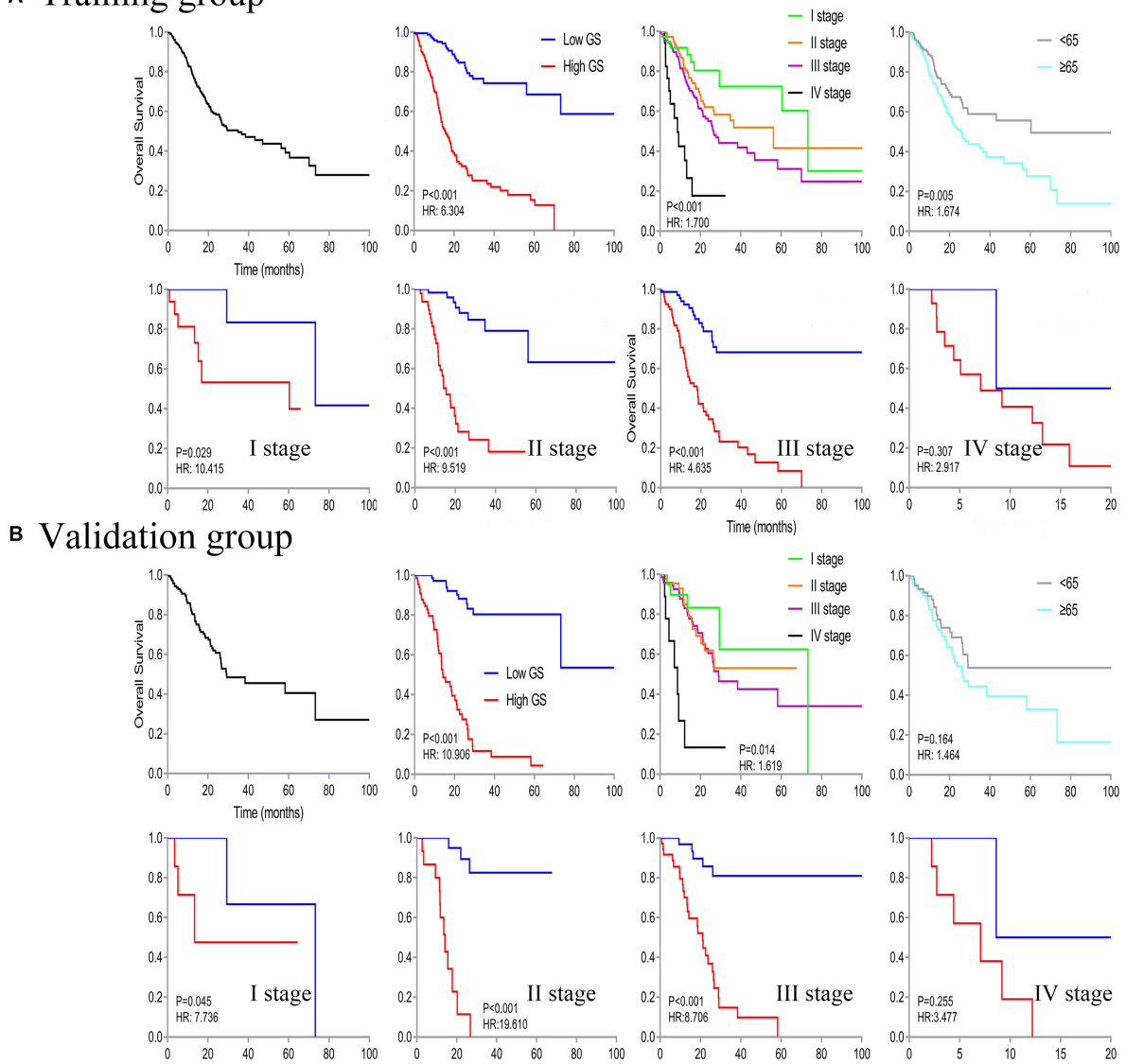


FIGURE 3 | Kaplan–Meier curve of overall survival in all patients, then stratified by genomics score (GS), pathological stage, and age. Survival analysis in the low- and high-GS groups was further divided based on stages (stages I–IV).

ZNF22, C1orf144 (cg14791193), GOLGA3 (cg26856948), HLA-DPB1 (cg20100408), LRFN4 (cg22740006), MDH2 (cg22813794), MIR365-2 (cg24361571), NUFIP2 (cg25161386), PREP (cg12485556), STYXL1 (cg22813794), TMEM117 (cg07020967), ZC3H10 (ZC3H10), IL1RAPL1 (cg20350671), PC (cg22740006), SHC4 (cg00481239), and ATXN10 (cg22395807), have not been previously reported.

Currently, focus has been directed on identifying prognostic miRNAs for GC. Particularly, one miRNA can regulate multiple targets, while multiple miRNAs can regulate a single mRNA. Therefore, a single miRNA may play an opposite role in cancer progression by regulating different target genes. For example, Mir-22 and Mir-100 were found to be tumor suppressors in various cancers, including GC (Chen et al., 2014; Zuo et al., 2015). Similarly, a high expression of Mir-136 was found to promote proliferation and invasion in GC cell lines by inhibiting PTEN expression (Chen et al., 2018), while a contrasting result was reported when HOXC10 was targeted (Zheng et al., 2017). Similarly, Mir-193b reportedly induced GC proliferation or apoptosis by mediating different mRNA expressions (Mu et al., 2014; Song et al., 2018), whereas a high Mir-1304 expression in GC was reported as a negative predictor for prognosis of lung and thyroid cancers (Kurata and Lin, 2018; Liu et al., 2019; Pan et al., 2019). However, the function of Mir-653 and Mir-6808 has not been previously reported. In the current study, we found an association between a high expression of all miRNAs and poor survival. Different outcomes may be observed in our study, relative to previous reports, owing to the huge

TABLE 3 | Comparison of different genomics score models (based on the median value) for overall survival in the training group and the validation group.

Variables	Training group, *n* = 329			Validation group, *n* = 150		
	Hazard ratio	95% CI	*P* value	Hazard ratio	95% CI	*P* value
Genomics nomogram						
Age	1.897	1.322–2.724	0.001	1.547	0.903–2.652	0.112
Pathological stage	1.489	1.155–1.920	0.002	1.267	0.846–1.897	0.252
Genomics score[a]	6.153	3.971–9.535	<0.001	10.141	5.011–20.520	<0.001
Clinical nomogram						
Age	1.760	1.225–2.528	<0.002	1.778	1.025–3.085	0.041
Pathological stage	1.754	1.349–2.282	<0.001	1.788	1.194–2.677	0.005
miRNAs nomogram						
Age	1.664	1.157–2.392	0.006	1.597	0.917–2.780	0.098
Pathological stage	1.748	1.345–2.273	<0.001	1.771	1.181–2.656	0.006
Genomics score[b]	2.011	1.402–2.883	<0.001	2.474	1.416–4.324	0.001
mRNA nomogram						
Age	2.140	1.484–3.086	<0.001	1.862	1.076–3.222	0.026
Pathological stage	1.716	1.314–2.241	<0.001	1.610	1.058–2.449	0.026
Genomics score[c]	3.222	2.209–4.700	<0.001	4.834	2.671–8.781	<0.001
Methylation nomogram						
Age	1.798	1.253–2.580	0.001	1.834	1.070–3.144	0.027
Pathological stage	1.599	1.231–2.078	<0.001	1.362	0.903–2.055	0.140
Genomics score[d]	4.627	3.058–7.002	<0.001	7.271	3.694–14.313	<0.001
miRNAs + methylation nomogram						
Age	1.750	1.220–2.511	0.002	1.692	0.987–2.899	0.056
Pathological stage	1.539	1.193–1.986	0.001	1.414	0.954–2.096	0.084
Genomics score[e]	5.009	3.291–7.624	<0.001	9.080	4.399–18.739	<0.001
miRNAs + mRNA nomogram						
Age	1.932	1.343–2.778	<0.001	1.824	1.057–3.148	0.031
Pathological stage	1.676	1.291–2.177	<0.001	1.546	1.027–2.326	0.037
Genomics score[f]	2.894	1.993–4.203	<0.001	3.431	1.969–5.979	<0.001
mRNA + methylation nomogram						
Age	1.939	1.351–2.784	<0.001	1.768	1.032–0.3.031	0.038
Pathological stage	1.523	1.181–1.965	0.001	1.322	0.882–1.979	0.176
Genomics score[g]	5.050	3.330–7.658	<0.001	7.553	3.911–14.586	<0.001
Cox-model 1 nomogram						
Age	1.908	1.329–2.740	<0.001	1.878	1.091–3.233	0.023
Pathological stage	1.642	1.276–2.112	<0.001	1.688	1.126–2.530	0.011
Genomics score[h]	5.034	3.334–7.601	<0.001	9.334	4.671–18.652	<0.001
Cox-model 2 nomogram						
Age	2.033	1.415–2.921	<0.001	1.777	1.030–3.067	0.039
Pathological stage	1.647	1.261–2.151	<0.001	1.722	1.142–2.595	0.009
Genomics score[i]	5.481	3.602–8.341	<0.001	8.679	4.347–17.325	<0.001

[a]*Based on 34 biomarkers (seven miRNAs, eight mRNA, and 19 DNA methylation sites).* [b]*Based on seven miRNAs.* [c]*Based on eight mRNA.* [d]*Based on 19 DNA methylation sites.* [e]*Based on seven miRNAs + 19 DNA methylation sites.* [f]*Based on seven miRNAs + eight mRNA.* [g]*Based on eight mRNA + 19 DNA methylation sites.* [h]*Based on two miRNAs + six mRNA + nine DNA methylation sites.* [i]*Based on one miRNAs + one mRNA + seven DNA methylation sites.*

number of corresponding miRNA target genes herein and lack of evidence on their role in GC development.

Messenger RNAs have been reported to play an essential role in GC cancer. For example, high NRP1 expression and hypermethylation were associated with poor GC prognosis (Wang et al., 2019), whereas another study indicated that it could be an anti-tumor target (Grandclement and Borg, 2011). In addition, high DUSP1 expression levels were found to promote progression, drug resistance, and poor prognosis of GC (Teng et al., 2018). On the other hand, SOX14 showed opposite prognostic values in cervical cancer and leukemia, with anti-tumor and carcinogenic effects, respectively (Stanisavljevic et al., 2017; Tosic et al., 2018). Studies have also implicated CPNE8, MAGED1, and RNF144A in ovarian and breast cancers (Zeng et al., 2012; Nagasawa et al., 2019; Yang et al., 2019). However, LOC91450 and ZNF22 have not been reported in cancer.

Accumulating evidence indicates that DNA methylation plays a significant role in cancer progression. However, only a handful of studies have described the relationship between levels of

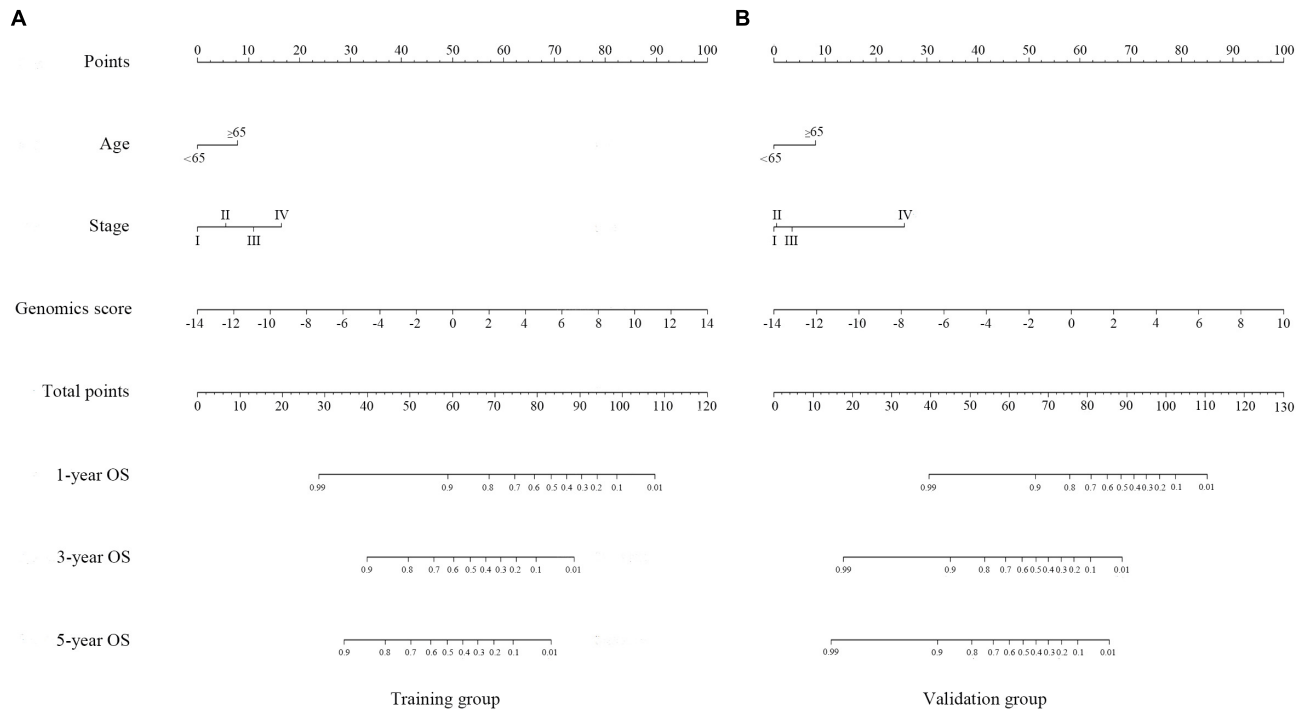


FIGURE 4 | Genomics nomogram to predict the probability of 1-, 3-, and 5-year overall survival (OS) in the training cohort **(A)** and the validation cohort **(B)**: to determine how many points for each variable to the probability of OS, locate the variable on its axis, draw a line straight upward to the point axis, repeat this process for each variable, sum up the points achieved for each of the risk factors, locate the final sum on the total point axis, and draw a line straight down to find the patient's probability of OS.

single-site methylation and GC prognosis. Particularly, high expressions of MAP7D2, ACOT13, EID1, RPS4X, and TTRAP have been associated with poor prognosis in gastric, lung, and pancreatic cancers as well as hepatic carcinoma, respectively, while a high TTRAP expression reportedly inhibits the growth of osteosarcoma (Kamio et al., 2010; Zhou et al., 2013; Kuang et al., 2017; Liu K.T. et al., 2018; Liu X. et al., 2018). Notably, the relationship between methylation levels and corresponding gene expression profiles is unknown, necessitating further research. Furthermore, the remaining DNA methylation sites and their corresponding genes have not been reported. Lastly, no study has described the prognostic significance using a genome-wide network.

Last but not least, in general, no study concerning their prognostic significance as a genome-wide network has been reported yet.

Tumorigenesis involves multiple interacting biological processes. In addition, an integrated genetic network is better at reflecting intra-tumor heterogeneity compared to a single biomarker. In the current study, we identified a novel, prognostic, signature genome-wide network, consisting of seven miRNAs, eight mRNA, and 19 DNA methylation sites after screening the entire TCGA cohort using training and random cohorts. This network was further divided into several other models.

Our results revealed that the integrative signature was an independent prognostic factor for survival in GC patients and performed better than any single biomarker or clinical characteristic. Moreover, stratification by other clinicopathological features, such as stage, age, sex, primary site, pathology grade, Lauren classification, and treatments, revealed significantly different prognosis values based on different GSs. In addition, staging was still an effective prognostic factor after dividing into low- and high-genomics-score groups, suggesting that GS and traditional staging can complement each other, and the genetic network could add prognostic value to traditional staging. Exclusion of patients with I staging showed that chemotherapy is a significant prognostic factor because I staging does not always need additional chemotherapy for effective prognosis.

We also developed and validated nomograms based on the GS. Particularly, results from ROC and DCA indicated that all of them had significantly better predictive performances than the traditional clinical nomogram. Comprehensive property (similar C-index) was not significantly different in genomics nomogram and Cox-model 1 nomogram, and compared to the genomics nomogram, Cox-model 1 nomogram had fewer biomarkers. In addition, Cox-model 1 nomogram performed well, with a higher positive net reclassification improvement (NRI). Therefore, Cox-model 1 nomogram might be more suitable for clinical application, which deserved further study. Besides that, the Cox-model 2 nomogram had the least feature (nine biomarkers) including miRNAs, mRNA, and DNA methylation sites for constructing a genome-wide network, while it had a lower C-index and a negative NRI. The other six

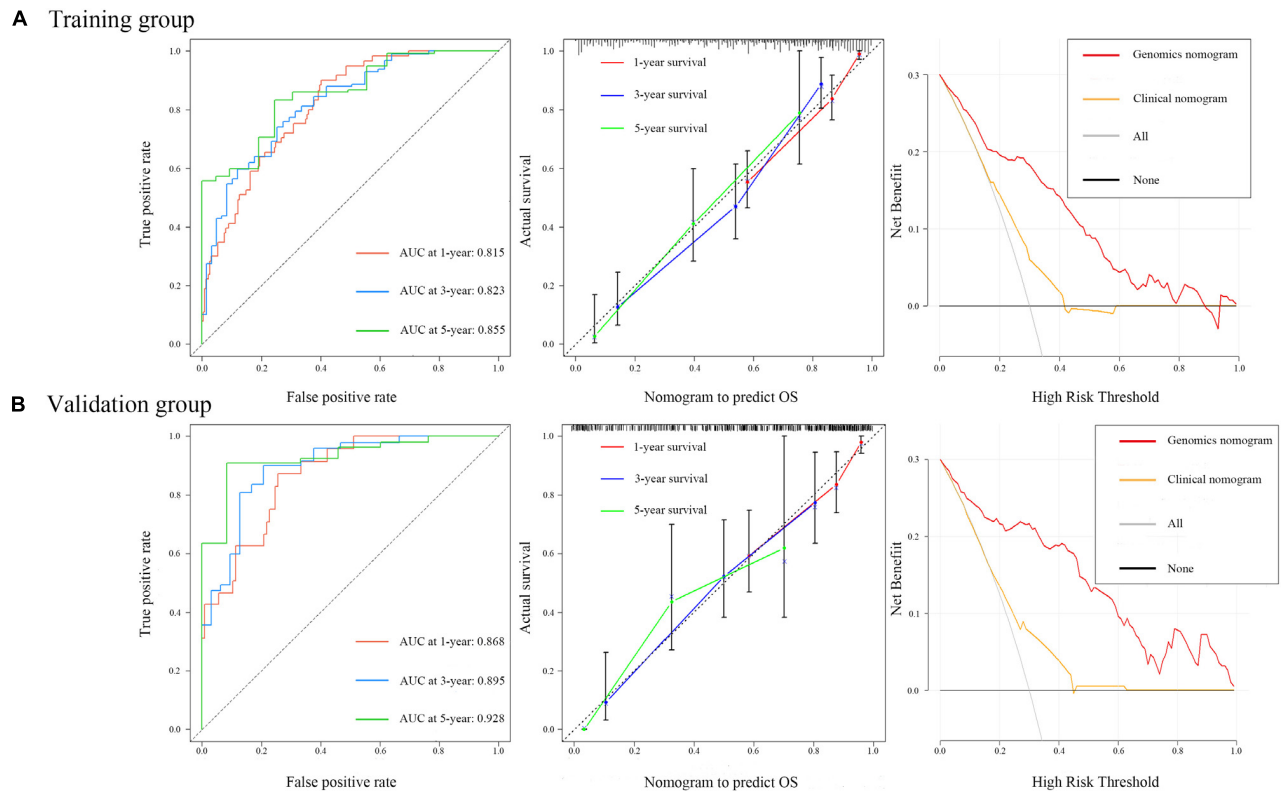


FIGURE 5 | Time-dependent receiver operating characteristic curves on 1, 3, and 5 years of genomics nomogram. Calibration plot showing 1-, 3-, and 5-year overall survival (OS) probability: the nomogram-estimated OS is plotted on the x-axis, and the actual OS is plotted on the *y*-axis. The diagonal dotted line is a perfect estimation by an ideal model, in which the estimated outcome perfectly corresponds to the actual outcome. The solid line is the performance of the nomogram: A closer alignment with the diagonal dotted line represents a better estimation. Decision curve analysis for genomics nomogram and clinical nomogram: the y-axis measures the net benefit. The gray line or the horizontal black line represents a follow-up of all patients or no patients. The model makes more benefit with a higher position in curve.

TABLE 4 | The area under the curve (AUC) values of different genomics score models in the training group and the validation group.

Models	Training group, *n* = 329			Validation group, *n* = 150		
	1-year OS	3-year OS	5-year OS	1-year OS	3-year OS	5-year OS
	AUC (95% CI)	AUC (95% CI)	AUC (95% CI)	AUC (95% CI)	AUC (95% CI)	AUC (95% CI)
Genomics nomogram	0.815 (0.787–0.843)	0.823 (0.785–0.861)	0.855 (0.799–0.911)	0.868 (0.832–0.900)	0.895 (0.851–0.939)	0.928 (0.886–0.970)
Clinical nomogram	0.609 (0.571–0.647)	0.615 (0.573–0.657)	0.642 (0.582–0.702)	0.638 (0.577–0.699)	0.598 (0.528–0.659)	0.721 (0.626–0.816)
miRNAs nomogram	0.621 (0.582–0.660)	0.656 (0.610–0.703)	0.717 (0.650–0.779)	0.641 (0.581–0.701)	0.729 (0.670–0.788	0.736 (0.656–0.817)
mRNA nomogram	0.747 (0.713–0.781)	0.711 (0.666–0.756)	0.728 (0.65.6–0.79.8)	0.806 (0.761–0.851)	0.785 (0.724–0.846)	0.843 (0.766–0.918)
Methylation nomogram	0.799 (0.768–0.830)	0.813 (0.774–0.852)	0.845 (0.781–0.909)	0.866 (0.827–0.905)	0.877 (0.830–0.923)	0.894 (0.821–0.966)
miRNAs + methylation nomogram	0.796 (0.765–0.827)	0.819 (0.781–0.857)	0.850 (0.787–0.911)	0.856 (0.817–0.895)	0.895 (0.854–0.939)	0.908 (0.856–0.961)
miRNAs + mRNA nomogram	0.743 (0.710–0.776)	0.731 (0.687–0.775)	0.771 (0.707–0.835)	0.803 (0.758–0.848)	0.825 (0.772–0.878)	0.883 (0.826–0.939)
mRNA + methylation nomogram	0.819 (0.791–0.847)	0.818 (0.780–0.856)	0.849 (0.794–0.904)	0.873 (0.837–0.909)	0.884 (0.836–0.932)	0.919 (0.867–0.971)
Cox-model 1 nomogram	0.833 (0.804–0.862)	0.851 (0.821–0.881)	0.833 (0.778–0.888)	0.869 (0.832–0.906)	0.905 (0.866–0.944)	0.907 (0.858–0.956)
Cox-model 2 nomogram	0.795 (0.764–0.826)	0.805 (0.767–0.843)	0.736 (0.662–0.810)	0.835 (0.793–0.877)	0.859 (0.797–0.921)	0.785 (0.667–0.903)

models showed a relatively poor performance in ROC or DCA, with limited application value. What is more, it is possible that DNA methylation was the highest contributor to the survival prediction of this gene network. We suspect that this may be related to the larger number of DNA methylation sites compared to miRNA and mRNA.

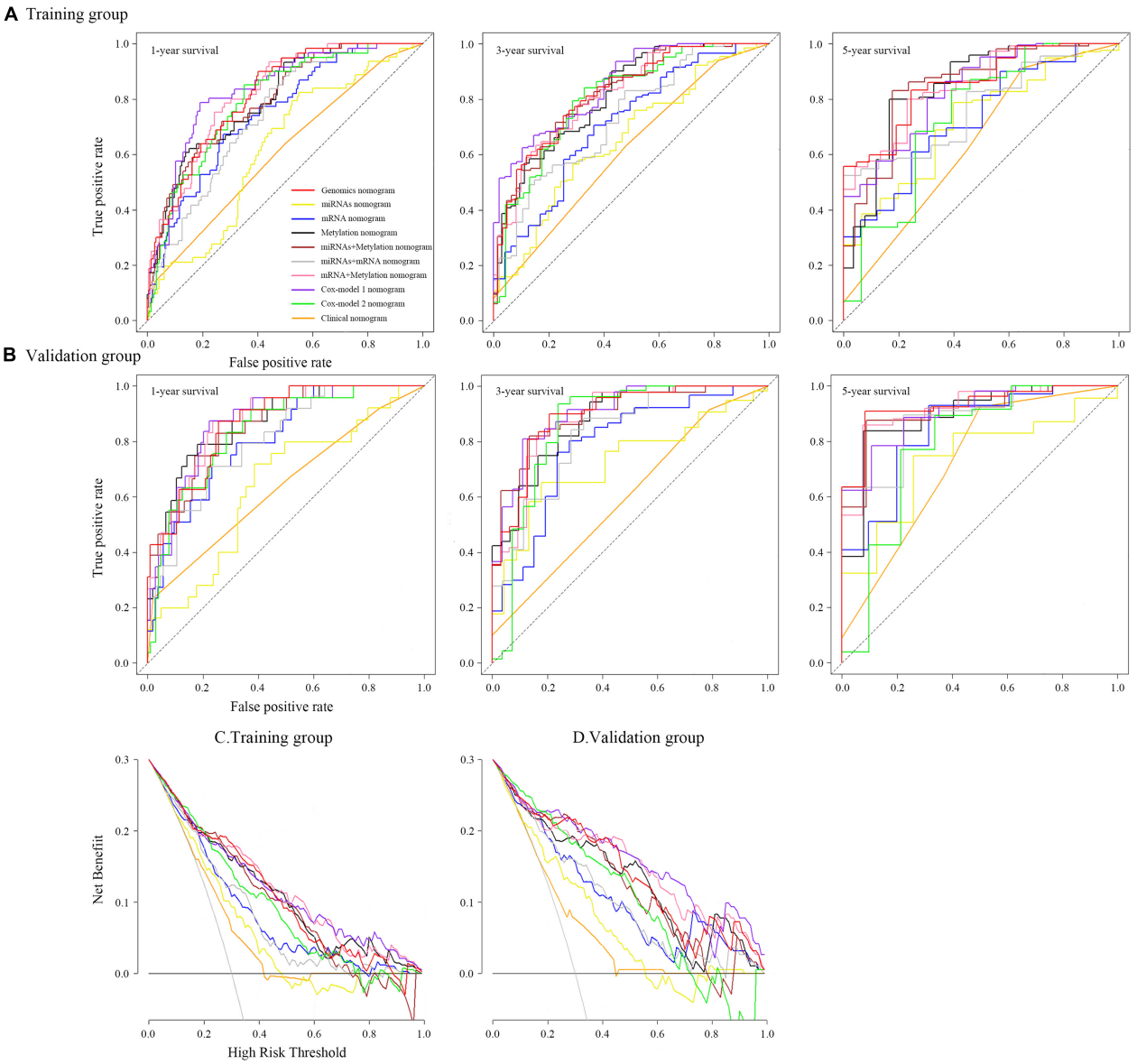


FIGURE 6 | Time-dependent receiver operating characteristic curves on 1, 3, and 5 years of each nomogram and decision curve analysis for each nomogram.

We adopted GO and KEGG analyses to assess the influence of genome-wide network in the prognosis of GC. Generally, biological processes mainly involve various biological functions, such as methylation, phosphorylation, and endocrine regulation. Methylation pathway was related to the occurrence and the development of GC, which was consistent with our results. Besides that, functional enrichment analysis revealed that phosphorylation pathway was significantly enriched as well, which got more and more attention these years. On the other hand, the main components of participation included organelles, cytomembranes, extrinsic to membranes, nuclear and synapses, whereas molecular functions comprise nucleoside, ATP, RNA, and transcription factor binding as well as activity of various enzymes. Abnormal cell composition is closely related to the development of tumor. The abnormal protein may act on the nucleus, membrane, or cell matrix, thereby leading to the progression of cancer, such as NRP1 protein (Wang et al., 2019). In the current study, KEGG analysis indicated that the gene network function was a relevant pathway in cancer, cell cycle, and adipocytokine signaling, while the other pathways had been reported in small cell lung and bladder cancers. Further experiments to reveal the biological function of this gene network are needed.

We also employed a series of complex statistical analyses to construct and validate a genome-wide network based on different biomarkers and then divided it into different models. We recommend the resulting GS despite it not being an absolute representative of tumor heterogeneity. This network

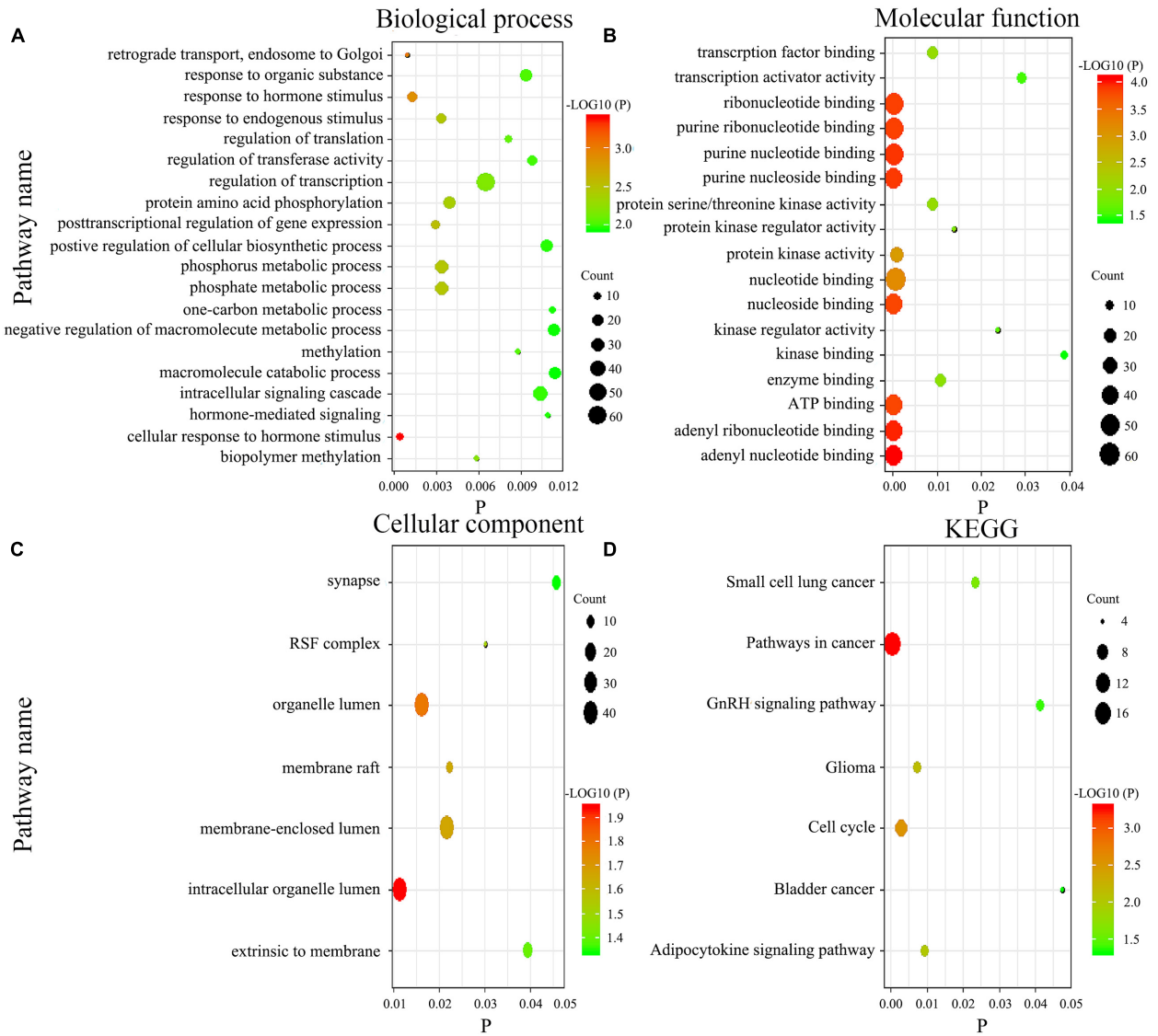


FIGURE 7 | Gene enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for the genome-wide network. The biological process **(A)**, molecular function **(B)**, cellular component **(C)**, and KEGG pathways **(D)**.

could complement the deficiency of traditional staging and generate a more accurate prediction of survival rates in GC patients. Additionally, it effectively distinguishes patients who could benefit from chemotherapy, thereby reducing unnecessary treatments. It is also possible that the network could be used to identify novel therapeutic targets for GC, although this requires further investigation.

Limitation

This study had several limitations. Firstly, information relating to patient co-morbidities and performance status was not available in the TCGA database. Secondly, the systemic chemotherapy regimens were not uniform, and most of them were based on fluoropyrimidines. Thirdly, the gene network contains too many biomarkers, increasing the difficulty of clinical use. Lastly, this was a retrospective study, without any independent external patient datasets as test. Despite some limitations, it was the first, to the best of our knowledge, to integrate miRNAs, mRNA, and DNA methylation sites as a genome-wide network to predict the OS of patients with GC, and we would try to design a validation in our hospital.

CONCLUSION

In summary, we used a TCGA cohort to develop and validate a novel genome-wide network comprising seven miRNAs, eight mRNAs, and 19 DNA methylation sites for the prognosis of GC. A combination of GS and TNM staging enhances its prognostic value, proposing a more comprehensive sub-typing system. The

developed network is expected to aid in predicting GC patients who may benefit from chemotherapy to some degree.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed during the current study are available in the TCGA database [https://portal.gdc.cancer.gov/].

ETHICS STATEMENT

Since TCGA was a public-use database, no additional permission was required from the Ethics Committee. In addition, this study was deemed exempt from institutional review board approval by Nanfang Hospital of Southern Medical University (Guangzhou, China).

AUTHOR CONTRIBUTIONS

All the authors listed had made a substantial contribution to this work. YJ and GL put forward the conception and designed the study. WH, TL, and MZ collected and collated the data. ZS, HC, and ZH analyzed the data and wrote the manuscript together. HL, JY, and YH made contribution to proofread the article. Finally, all the authors took responsibility of the final manuscript and approved it for publication.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (81872013 and 81672446) and The Outstanding Youths Development Scheme of Nanfang Hospital, Southern Medical University (2018J007).

ACKNOWLEDGMENTS

We would like to thank the staff members of TCGA program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00835/full#supplementary-material

FIGURE S1 | Screening process for the genome-wide network.

FIGURE S2 | Screening process for the Gene Ontology analysis and Kyoto Encyclopedia of Genes and Genomes pathways.

FIGURE S3 | Least absolute shrinkage and selection operator Cox regression performed for miRNAs, mRNA, and DNA methylation sites.

FIGURE S4 | A heat map based on genomic scores layered by clinicopathological factors.

FIGURE S5 | Median survival for patients in the low-genomics score (GS) and the high-GS groups.

FIGURE S6 | Distribution of patient cases and density based on genomics score in the validation group **(A,B)**. Scatter plots of genomics score (GS) regarding the classification of low- and high-GS **(C,D)**, and the bold line represents the median.

FIGURE S7 | Kaplan–Meier curve of overall survival for the low- and the high-genomics-score groups divided by age (young and old), sex (male and female), primary site (cardia, fundus/body, and antrum), pathology grade (I–II and III–IV), Lauren classification (intestinal type and diffused type), and treatment (R0 surgery, R1 surgery, chemotherapy, and no chemotherapy) in the training group.

FIGURE S8 | Kaplan–Meier curve of overall survival for the low- and the high-genomics-score groups divided by age (young and old), sex (male and female), primary site (cardia, fundus/body, and antrum), pathology grade (I-II and III-IV), Lauren classification (intestinal type and diffused type), and treatment (R0 surgery, R1 surgery, chemotherapy, and no chemotherapy) in the validation group.

FIGURE S9 | Cox-model 1 nomogram and its calibration plot in the training group and the validation group.

FIGURE S10 | Cox-model 2 nomogram and its calibration plot in the training group and the validation group.

FIGURE S11 | Clinical nomogram and its calibration plot in the training group and the validation group.

FIGURE S12 | Time-dependent receiver operating characteristic curves on 1, 3, and 5 years for each nomogram in the training group.

FIGURE S13 | Time-dependent receiver operating characteristic curves on 1, 3, and 5 years for each nomogram in the validation group.

FIGURE S14 | Venn diagrams for the target genes of each miRNA.

FIGURE S15 | Three most potential target genes for each miRNA **(A)** and the miRNA–target genes co-expression network **(B)**.

FIGURE S16 | Time-dependent receiver operating characteristic curves on 1, 3, and 5 years for each signature in the genome-wide network. Each color represents a different signature: red 2 (genomics nomogram), orange (clinical nomogram), yellow (hsa-mir-100), blue (hsa-mir-1304), black (hsa-mir-193b), brown (hsa-mir-22), gray (hsa-mir-136), pale violet red 1 (hsa-mir-653), purple 2 (hsa-mir-6808), green 2 (NRP1|8829), dark red (SOX14|8403), dark orchid 3 (CPNE8|144402), dark orange 2 (MAGED1|9500), cyan (RNF144A|9781), corn silk 4 (ZNF22|7570), hot pink 1 (DUSP1|1843), chartreuse (LOC91450|91450), aquamarine 2 (cg02223323), dark green (cg00481239), maroon 1 (cg07020967), pale green (cg08859156) gold (cg12485556), purple 1 (cg14791193), antique white (cg15861578), khaki 2 (cg15486740) sea green 2 (cg20100408), light sky blue (cg20350671), maroon (cg22395807), Indian red 2 (cg24361571), tan 1 (cg25361506), peach puff (cg25622155), turquoise 1 (cg25161386), orchid 2 (cg22740006), dark golden rod 2 (cg22813794), dark golden rod 4 (cg26014401), and violet red 1 (cg26856948).

FIGURE S17 | Kaplan–Meier curve of overall survival in the low-genomics-score group stratified by clinical features (e.g., stage).

FIGURE S18 | Kaplan-Meier curve of overall survival in the high-genomics-score group stratified by clinical features (e.g., stage).

FIGURE S19 | Kaplan–Meier curve of overall survival in stages II and III patients.

TABLE S1 | Descriptive statistics of gastric cancer patients.

TABLE S2 | Multivariable analysis for miRNAs, mRNA, and DNA methylation sites and clinical characteristics.

TABLE S3 | Multivariable analysis for miRNAs.

TABLE S4 | Multivariable analysis for mRNA.

TABLE S5 | Multivariable analysis for DNA methylation sites.

TABLE S6 | The equation of different models.

TABLE S7 | The C-index value of different models.

TABLE S8 | The area under the curve values for each individual of the genome-wide network.

REFERENCES

Anna, C. D. S., Junior, A. G. F., Soares, P., Tuji, F., Paschoal, E., Chaves, L. C., et al. (2018). Molecular biology as a tool for the treatment of cancer. *Clin. Exper. Med.* 18, 457–464.

Bang, Y. J., Kim, Y. W., Yang, H. K., Chung, H. C., Park, Y. K., Lee, K. H., et al. (2012). Adjuvant capecitabine and oxaliplatin for gastric cancer after D2 gastrectomy (CLASSIC): a phase 3 open-label, randomised controlled trial. *Lancet* 379, 315–321. doi: 10.1016/s0140-6736(11)61873-4

Camargo, M. C., Figueiredo, C., and Machado, J. C. (2019). Review: gastric malignancies: basic aspects. *Helicobacter* 24(Suppl. 1):e12642.

Cancer Genome Atlas Research (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202–209. doi: 10.1038/nature13480

Chen, J., Zheng, B., Wang, C., Chen, Y., Du, C., Zhao, G., et al. (2014). Prognostic role of microRNA-100 in various carcinomas: evidence from six studies. *Tumor Biol.* 35, 3067–3071. doi: 10.1007/s13277-013-1398-3

Chen, X., Huang, Z., and Chen, R. (2018). Microrna-136 promotes proliferation and invasion ingastric cancer cells through Pten/Akt/P-Akt signaling pathway. *Oncol. Lett.* 15, 4683–4689.

Choi, R. S., DiNardo, J. A., and Brown, M. L. (2019). Current and future molecular diagnostics of gastric cancer. *Expert. Rev. Mol. Diagn.* 19, 863–874.

Cristescu, R., Lee, J., Nebozhyn, M., Kim, K. M., Ting, J. C., Wong, S. S., et al. (2015). Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat. Med.* 21, 449–456.

Duarte, H. O., Gomes, J., Machado, J. C., and Reis, C. A. (2018). Gastric cancer: basic aspects. *Helicobacter* 23(Suppl. 1):e12523. doi: 10.1111/hel.12523

Global Burden of Disease Cancer Collaboration, Fitzmaurice, C., Akinyemiju, T. F., Al Lami, F. H., Alam, T., Alizadeh-Navaei, R., et al. (2019). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: a systematic analysis for the global burden of disease study. *JAMA Oncol.* 5, 1749–1768.

Grandclement, C., and Borg, C. (2011). Neuropilins: a new target for cancer therapy. *Cancers* 3, 1899–1928. doi: 10.3390/cancers3021899

Hou, X., He, X., Wang, K., Hou, N., Fu, J., Jia, G., et al. (2018). Genome-wide network-based analysis of colorectal cancer identifies novel prognostic factors and an integrative prognostic index. *Cell Physiol. Biochem.* 49, 1703–1716. doi: 10.1159/000493614

Iasonos, A., Schrag, D., Raj, G. V., and Panageas, K. S. (2008). How to build and interpret a nomogram for cancer prognosis. *J. Clin. Oncol.* 26, 1364–1370. doi: 10.1200/jco.2007.12.9791

Jiang, Y., Chen, C., Xie, J., Wang, W., Zha, X., Lv, W., et al. (2018a). Radiomics signature of computed tomography imaging for prediction of survival and chemotherapeutic benefits in gastric cancer. *EBio Med.* 36, 171–182. doi: 10.1016/j.ebiom.2018.09.007

Jiang, Y., Zhang, Q., Hu, Y., Li, T., Yu, J., Zhao, L., et al. (2018b). ImmunoScore signature: a prognostic and predictive tool in gastric cancer. *Ann. Surg.* 267, 504–513. doi: 10.1097/sla.0000000000002116

Kamio, Y., Maeda, K., Moriya, T., Takasu, N., Takeshita, A., Hirai, I., et al. (2010). Clinicopathological significance of cell cycle regulatory factors and differentiation-related factors in pancreatic neoplasms. *Pancreas* 39, 345–352. doi: 10.1097/mpa.0b013e3181bb9204

Kuang, J., Li, Q. Y., Fan, F., Shen, N. J., Zhan, Y. J., Tang, Z. H., et al. (2017). Overexpression of the X-linked ribosomal protein S4 predicts poor prognosis in patients with intrahepatic cholangiocarcinoma. *Oncol. Lett.* 14, 41–46. doi: 10.3892/ol.2017.6137

Kurata, J. S., and Lin, R. J. (2018). MicroRNA-focused CRISPR-Cas9 library screen reveals fitness-associated miRNAs. *RNA* 24, 966–981. doi: 10.1261/rna.066282.118

Lauren, P. (1965). The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification. *Acta Pathol. Microbiol. Scand.* 64, 31–49. doi: 10.1111/apm.1965.64.1.31

Li, X., Zhang, Y., Zhang, Y., Ding, J., Wu, K., Fan, D., et al. (2010). Survival prediction of gastric cancer by a seven-microRNA signature. *Gut* 59, 579–585. doi: 10.1136/gut.2008.175497

Liu, G. H., Shi, H., Deng, L., Zheng, H., Kong, W., Wen, X., et al. (2019). Circular RNA circ-FOXM1 facilitates cell progression as ceRNA to target PPDPF and MACC1 by sponging miR-1304-5p in non-small cell lung cancer. *Biochem. Biophys. Res. Commun.* 513, 207–212. doi: 10.1016/j.bbrc.2019.03.213

Liu, K. T., Yeh, I. J., Chou, S. K., Yen, M. C., and Kuo, P. L. (2018). Regulatory mechanism of fatty acid-CoA metabolic enzymes under endoplasmic reticulum stress in lung cancer. *Oncol. Rep.* 40, 2674–2682.

Liu, X., Wu, J., Zhang, D., Bing, Z., Tian, J., Ni, M., et al. (2018). Identification of potential key GENES associated with the pathogenesis and prognosis of gastric cancer based on integrated bioinformatics analysis. *Front. Genet.* 9:265. doi: 10.3389/fonc.2019.00265

Mu, Y. P., Tang, S., Sun, W. J., Gao, W. M., Wang, M., Su, X. L., et al. (2014). Association of miR-193b down-regulation and miR-196a up-regulation with clinicopathological features and prognosis in gastric cancer. *Asian Pac. J. Cancer Prev.* 15, 8893–8900. doi: 10.7314/apjcp.2014.15.20.8893

Nagasawa, S., Ikeda, K., Horie-Inoue, K., Sato, S., Itakura, A., Takeda, S., et al. (2019). Systematic identification of characteristic genes of ovarian clear cell carcinoma compared with high-grade serous carcinoma based on RNA-sequencing. *Int. J. Mol. Sci.* 20:4330. doi: 10.3390/ijms20184330

Nagtegaal, I. D., Odze, R. D., Klimstra, D., Paradis, V., Rugge, M., and Schirmacher, P. (2019). The 2019 WHO classification of tumours of the digestive system. *Histopathology* 76, 182–188. doi: 10.1111/his.13975

Pan, Y., Xu, T., Liu, Y., Li, W., and Zhang, W. (2019). Upregulated circular RNA circ_0025033 promotes papillary thyroid cancer cell proliferation and invasion via sponging miR-1231 and miR-1304. *Biochem. Biophys. Res. Commun.* 510, 334–338. doi: 10.1016/j.bbrc.2019.01.108

Serra, O., Galán, M., Ginesta, M. M., Calvo, M., Sala, N., Salazar, R., et al. (2019). Comparison and applicability of molecular classifications for gastric cancer. *Cancer Treat. Rev.* 77, 29–34. doi: 10.1016/j.ctrv.2019.05.005

Sohn, B. H., Hwang, J. E., Jang, H. J., Lee, H. S., Oh, S. C., Shim, J. J., et al. (2017). Clinical significance of four molecular subtypes of gastric cancer identified by the cancer genome atlas project. *Clin. Cancer Res.* 23, 4441–4449. doi: 10.1158/1078-0432.ccr-16-2211

Song, B., Du, J., Song, D. F., Ren, J. C., and Feng, Y. (2018). Dysregulation of NCAPG, KNL1, miR-148a-3p, miR-193b-3p, and miR-1179 may contribute to the progression of gastric cancer. *Biol. Res.* 51:44.

Stanisavljevic, D., Petrovic, I., Vukovic, V., Schwirtlich, M., Gredic, M., Stevanovic, M., et al. (2017). SOX14 activates the p53 signaling pathway and induces apoptosis in a cervical carcinoma cell line. *PLoS One* 12:e0184686. doi: 10.1371/journal.pone.0184686

Sun, Z., Liu, H., Yu, J., Huang, W., Han, Z., Lin, T., et al. (2019a). Frequency and prognosis of pulmonary metastases in newly diagnosed gastric cancer. *Front. Oncol.* 9:671. doi: 10.3389/fonc.2019.00671

Sun, Z., Zheng, H., Yu, J., Huang, W., Li, T., Chen, H., et al. (2019b). Liver metastases in newly diagnosed gastric cancer: a population-based study from SEER. *J. Cancer* 10, 2991–3005. doi: 10.7150/jca.30821

Tan, I. B., Ivanova, T., Lim, K. H., Ong, C. W., Deng, N., Lee, J., et al. (2011). Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy. *Gastroenterology* 141, 476–485.

Teng, F., Xu, Z., Chen, J., Zheng, G., Zheng, G., Lv, H., et al. (2018). DUSP1 induces apatinib resistance by activating the MAPK pathway in gastric cancer. *Oncol. Rep.* 40, 1203–1222.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16, 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., Jemal, A., et al. (2015). Global cancer statistics, 2012. *CA Cancer J. Clin.* 65, 87–108. doi: 10.3322/caac.21262

Tosic, N., Petrovic, I., Grujicic, N. K., Davidovic, S., Virijevic, M., and Vukovic, N. S. (2018). Prognostic significance of *SOX*2, SOX3, SOX11, SOX14 and SOX18 gene expression in adult de novo acute myeloid leukemia. *Leuk Res.* 67, 32–38. doi: 10.1016/j.leukres.2018.02.001

Ueda, T., Volinia, S., Okumura, H., Shimizu, M., Taccioli, C., Rossi, S., et al. (2010). Relation between microRNA expression and progression and prognosis of gastric cancer: a microRNA expression analysis. *Lancet Oncol.* 11, 136–146. doi: 10.1016/s1470-2045(09)70343-2

Vickers, A. J., and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* 26, 565–574. doi: 10.1177/0272989x06295361

Wang, G. H., Shi, B., Fu, Y., Zhao, S., Qu, K., Guo, Q., et al. (2019). Hypomethylated gene NRP1 is co-expressed with PDGFRB and associated with poor overall

survival in gastric cancer patients. *Biomed. Pharmacother.* 111, 1334–1341. doi: 10.1016/j.biopha.2019.01.023

Yang, Y. L., Zhang, Y., Li, D. D., Zhang, F. L., Liu, H. Y., Liao, X. H., et al. (2019). RNF144A functions as a tumor suppressor in breast cancer through ubiquitin ligase activity-dependent regulation of stability and oncogenic functions of HSPA2. *Cell Death Differ.* 27, 1105–1118. doi: 10.1038/s41418-019-0400-z

Zeng, Z. L., Wu, W. J., Yang, J., Tang, Z. J., Chen, D. L., Qiu, M. Z., et al. (2012). Prognostic relevance of melanoma antigen D1 expression in colorectal carcinoma. *J. Transl. Med.* 10:181. doi: 10.1186/1479-5876-10-181

Zhang, Y., Yang, W., Li, D., Yang, J. Y., Guan, R., Yang, M. Q., et al. (2018). Toward the precision breast cancer survival prediction utilizing combined whole genome-wide expression and somatic mutation analysis. *BMC Med. Genom.* 11(Suppl. 5):104. doi: 10.1186/s12920-018-0419-x

Zhang, Z., Dong, Y., Hua, J., Xue, H., Hu, J., Jiang, T., et al. (2019). A five-miRNA signature predicts survival in gastric cancer using bioinformatics analysis. *Gene* 699, 125–134. doi: 10.1016/j.gene.2019.02.058

Zheng, J., Ge, P., Liu, X., Wei, J., Wu, G., Li, X., et al. (2017). MiR-136 inhibits gastric cancer-specific peritoneal metastasis by targeting HOXC10. *Tumour Biol.* 39:1010428317706207.

Zhou, C., Shen, Q., Xue, J., Ji, C., and Chen, J. (2013). Overexpression of TTRAP inhibits cell growth and induces apoptosis in osteosarcoma cells. *BMB Rep.* 46, 113–118. doi: 10.5483/bmbrep.2013.46.2.150

Zuo, Q. F., Cao, L. Y., Yu, T., Gong, L., Wang, L. N., Zhao, Y. L., et al. (2015). MicroRNA-22 inhibits tumor growth and metastasis in gastric cancer by directly targeting MMP14 and Snail. *Cell Death Dis.* 6:e2000. doi: 10.1038/cddis.2015.297

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 August 2020
doi: 10.3389/fgene.2020.00989

Identification of the Prognostic Value of Immune-Related Genes in Esophageal Cancer

Xiong Guo[1†], Yujun Wang[2†], Han Zhang[3], Chuan Qin[4], Anqi Cheng[1], Jianjun Liu[1], Xinglong Dai[1] and Ziwei Wang[1]*

[1] Department of Gastrointestinal Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [2] Department of Pathology, Daping Hospital, Army Military Medical University, Chongqing, China, [3] Department of Digestive Oncology, Three Gorges Hospital, Chongqing University, Chongqing, China, [4] Department of Gastrointestinal Surgery, Three Gorges Hospital, Chongqing University, Chongqing, China

Esophageal cancer (EC) is a serious malignant tumor, both in terms of mortality and prognosis, and immune-related genes (IRGs) are key contributors to its development. In recent years, immunotherapy for tumors has been widely studied, but a practical prognostic model based on immune-related genes (IRGs) in EC has not been established and reported. This study aimed to develop an immunogenomic risk score for predicting survival outcomes among EC patients. In this study, we downloaded the transcriptome profiling data and matched clinical data of EC patients from The Cancer Genome Atlas (TCGA) database and found 4,094 differentially expressed genes (DEGs) between EC and normal esophageal tissue ($p < 0.05$ and fold change >2). Then, the intersection of DEGs and the immune genes in the "ImmPort" database resulted in 303 differentially expressed immune-related genes (DEIRGs). Next, through univariate Cox regression analysis of DEIRGs, we obtained 17 immune genes related to prognosis. We detected nine optimal survival-associated IRGs (*HSPA6, CACYBP, DKK1, EGF, FGF19, GAST, OSM, ANGPTL3, NR2F2*) by using Lasso regression and multivariate Cox regression analyses. Finally, we used those survival-associated IRGs to construct a risk model to predict the prognosis of EC patients. This model could accurately predict overall survival in EC and could be used as a classifier for the evaluation of low-risk and high-risk groups. In conclusion, we identified a practical and robust nine-gene prognostic model based on immune gene dataset. These genes may provide valuable biomarkers and prognostic predictors for EC patients and could be further studied to help understand the mechanism of EC occurrence and development.

Keywords: esophageal cancer, immune-related gene, TCGA, prognostic model, bioinformatics analysis

OPEN ACCESS

Edited by:
Yuanwei Zhang, University of Science and Technology of China, China

Reviewed by:
Yan Wang, Jilin University, China
Daniel Guariz Pinheiro, São Paulo State University, Brazil

****Correspondence:***
Ziwei Wang
ziweiwang1@sina.com;
wangziwei571@sina.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *28 March 2020*
Accepted: *05 August 2020*
Published: *21 August 2020*

Citation:
Guo X, Wang Y, Zhang H, Qin C, Cheng A, Liu J, Dai X and Wang Z (2020) Identification of the Prognostic Value of Immune-Related Genes in Esophageal Cancer. Front. Genet. 11:989. doi: 10.3389/fgene.2020.00989

INTRODUCTION

Esophageal cancer (EC) is ranked 7th and 6th in incidence and mortality, respectively (Bray et al., 2018). It is one of the most aggressive types of cancer. Although the addition of neoadjuvant or perioperative therapy provides a modest improvement in overall survival in resectable cases, the prognosis of patients with advanced EC is still very poor (Cunningham et al., 2006; Allum et al., 2009; van Hagen et al., 2012; Noble et al., 2017). Due to recurrence, extensive invasion

and metastasis, the overall 5-year survival rate of EC is lower than 13% after initial diagnosis (Khalil et al., 2016; Vo et al., 2019). Hence, identifying biomarkers for the treatment and prognostic prediction of EC could lead to better interventions for patients with an otherwise poor prognosis.

Immune disorders in tumor is regarded as a promoting factor during tumorigenesis and development. In recent years, immunotherapy has become a promising potential therapy for various cancers in addition to surgery and radiotherapy (Khalil et al., 2016; Zhao et al., 2019). EC cells harbor abundant tumor antigens, including tumor-associated antigens and neoantigens, which have the ability to initiate dendritic cell-mediated tumor-killing cytotoxic T lymphocytes in the early stage of cancer development. As EC cells battle the immune system, they obtain an ability to suppress antitumor immunity through immune checkpoints, secreted factors, and negative regulatory immune cells (Huang and Fu, 2019). Immune checkpoint inhibitors (ICIs) have been investigated in various types of cancers and provide a new treatment landscape (Tanaka et al., 2017). ICIs have been reported to attenuate tumor growth mainly by reducing the immune escape of cancer cells, and programed death 1 (*PDL1*) is one of the immune checkpoints that is the most commonly used target for immunotherapy in EC (Shaib et al., 2016). However, at present, EC immunotherapies always lead to mixed results, which are partially caused by the absence of reliable markers that are predictive of treatment response (Ohashi et al., 2015). Molecular profiles of tumor cells and cancer-related cells within their microenvironments represent promising candidates for predictive and prognostic biomarkers. Despite vigorous efforts have been made with major breakthroughs in high-throughput genomic technologies (Li et al., 2017). Increasing evidence suggests that the expression of IRGs may be related to the prognosis of tumors. Qiu et al. (2020) identified and verified of an individualized prognostic signature of bladder cancer based on seven immune related genes. Zhang et al. (2020) discovered a novel immune-related gene signature for risk stratification and prognosis of survival in lower-grade glioma. And Zhao et al. (2020) used immune score to predict survival in early-stage lung adenocarcinoma patients.

Similarly, the prognostic characteristics based on these IRGs may help in the diagnosis and individualized treatments for EC (Gentles et al., 2015). However, several studies have reported the relationship of IRGs with the prognosis of patients with EC (Turato et al., 2019; Yan et al., 2019). In addition, there is currently no systematic description or study of IRGs and the tumor immune microenvironment in large samples of patients with EC. Therefore, a systematic description and analysis of the tumor immune microenvironment and IRGs impact on prognosis is necessary for EC immunotherapy and patient prognosis. In this study, we analyzed 182 samples of EC in the TCGA database, and 303 differentially expressed IRGs were found. Through multivariate Cox regression analysis, we found 9 immune-related prognosis genes. An accurate model for evaluating the prognosis of patients was established, and we investigated the clinical utility of this model in patients with EC. In addition, we calculated the correlation between immune cell infiltration and risk score in the tumor microenvironment. Our study identified new biomarkers and prognostic factors for EC, thus provides some new therapeutic targets in EC.

MATERIALS AND METHODS

Data Acquisition and Processing

The RNA-Seq gene expression profiles of patients with EC, including the Fragments Per Kilobase of transcript per Million Mapped reads (FPKM) based on the Illumina (San Diego, CA, United States) HiSeq 2000 RNA sequencing platform, were downloaded from the TCGA database using the GDC-client download tool[1] (Cao et al., 2019). The workflow type is HTSeq-FPKM. Then, the "limma" package of R software was utilized for the normalization of RNA expression profiles and averaged the duplicate data to remove the batch effects. Clinical data for the corresponding EC patients were also retrieved from the TCGA database, which included gender, age, tumor stage, and survival information. The patient's TCGA ID was used to distinguish between a tumor sample and a normal sample. The detailed characteristics and histopathological features of the EC patients and their TCGA IDs are summarized in **Supplementary Table S1**.

Immunologically relevant list of genes curated with functions and Gene Ontology terms (immune-related gene list) were download from the resources section of the "ImmPort" database[2] (Bhattacharya et al., 2018). It contains a total of 2,496 genes defined as immune-related. Data regarding 318 cancer-associated transcription factors (TFs) were obtained from the "Cistrome" project[3] (Mei et al., 2017).

Criteria of Enrolled Patients for the Construction of Risk Signature

The inclusive criteria of patients with EC for model construction were as follows: (1) patients primarily diagnosed with EC, (2) with only adenocarcinoma or squamous cell carcinoma as pathological type, (3) only samples with RNA-sequencing data, (4) patients with complete clinicopathological parameters, (5) overall survival time is more than 30 days.

Identification of Differentially Expressed Genes, Differentially Expressed IRGs

Differentially expressed genes (DEGs) between EC and normal tissues were identified using Wilcoxon test after within-array replicate probes were replaced with their average via "limma" package in the R software (version 3.6.2). $|\text{Log}_2$ fold change (FC)$| > 2.0$ and false discovery rate (FDR) adjusted to less than 0.05 were set as the cutoff criteria. Then, the DEGs were intersected with the immune-related gene list to obtain the DEIRGs. Those significant DEGs are visualized using heatmaps and volcano plots via "pheatmap" package in the R software. In addition, an online database, GEPIA 2.0 (Tang et al., 2019), was used to analyze

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

[2]https://www.immport.org/home

[3]http://www.cistrome.org/

differential expression of prognostic genes between 286 GTEx normal samples and 182 TCGA tumor samples.

Functional Annotations and Signaling Pathway Enrichment Analysis

"Clusterprofiler" R package (Yu et al., 2012) was used for Gene Ontology (GO) annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of DEGs and IRGs. The results of GO annotation and KEGG pathway analyses were visualized using the "GOplot" package in R platform. Gene Set Enrichment Analysis (GSEA) software (version 4.0.1) was used to analyze pathway activation and inhibition in high-risk and low-risk patients.

Risk Score Calculation and Survival Analysis

To explore candidate prognostic biomarkers of EC, a joint cox regression analysis was performed. Firstly, we merged the expression levels of IRGs with the corresponding survival time and survival status data of EC patients. Then, a univariate Cox proportional hazard regression analysis was used to identify the candidate survival-associated IRGs when p-value < 0.05. Next, the least absolute shrinkage and selection operator (LASSO) Cox regression analysis was used to identify the genetic model with the best prognostic value by using "glmnet" package in R software. Finally, multivariate Cox regression analysis was employed to construct the prognosis signature for predicting the prognosis in EC patients. We calculated the risk score of each patient using the expression of DEIRGs and the regression coefficients obtained in the regression model. The coefficient of the gene is multiplied by the expression of the gene and then summed to obtain each patient's risk score. The calculation formula is below (Wan et al., 2019):

$$Risk\ score(patients) = \sum_{i=1}^{n} \text{coefficien } t\ (\text{gene}_i)\ \text{expression value of } (\text{gene}_i) \quad (1)$$

Here, "gene_i" is the ith selected gene, and "coefficient (gene_i)" is the estimated regression coefficient of gene_i from the Cox proportional hazards regression analysis. Time-dependent receiver operating characteristic (ROC) curves were used to assess the accuracy of prognostic prediction models. The area under the ROC curve (AUC) >0.60 was considered an acceptable prediction, and an AUC >0.75 was recognized as an excellent predictive value. For survival analysis, patients were divided into low- and high-risk groups according to the median risk score calculated by this prognostic model, and then log-rank tests were used to analyze the survival data.

Construction of Cancer-Associated TFs and IRG Regulatory Networks

Differentially expressed transcription factors (DETFs) were derived from the intersection of tumor-associated TFs and DEGs. DETFs and survival-associated IRGs samples with the same TCGA patient ID were then used for correlation testing. $p < 0.05$ and cor ≥ 0.3 were considered significant correlations. Then, cytoscape software (Shannon et al., 2003) was used to draw the regulatory network.

Construction of a Predictive Nomogram Based on the IRGs

A nomogram encompassing the risk score based on expression of prognostic IRGs and clinicopathological factors was constructed using the "rms" R package. Based on the different clinicopathological characteristics and the risk score of each patient, we calculated the total score to predict 1, 2, and 3-year prognosis of EC patients.

Clinical Correlation Analysis

Univariate regression analysis and multivariate regression analysis were used to identify factors (including gender, age, TNM stage and risk score) affecting survival and independent prognostic factors in patients with EC. The correlation between survival-associated IRGs and clinicopathological characteristics was analyzed in R platform. $p < 0.05$ was considered to have a significant correlation.

Relationship Between Risk Score and Immune Cell Infiltration

The immune cell infiltrate data were collected from Tumor Immune Estimation Resource (TIMER)[4] (Liu et al., 2011) database. The database includes 10,897 samples across 32 cancer types from TCGA to estimate the abundance of six subtypes of tumor-infiltrating immune cells, including B cells, CD8 T cells, CD4 T cells, dendritic cells (DCs), macrophages, and neutrophils. Based on the same patient's ID as TCGA, the correlation between patient immune infiltrated cells and risk score was calculated in R software.

Statistical Analyses

All data were processed with R (version 3.6.2) and Perl (5.30.1) software. DEGs were identified using the Wilcox test. Survival analyses were performed using the Kaplan-Meier method and the log-rank test.

RESULTS

Differentially Expressed IRGs in EC

The analysis process for this study is shown in **Figure 1**. A total of 182 patients were involved in the development and validation of the prognostic signature, including 95 squamous cell neoplasms, 87 adenomas and adenocarcinomas. Of these, 111 were white people, 46 were Asian, five were African American, and 20 were unreported. The TCGA IDs for the 182 patients were presented in **Supplementary Table S1**. Initially, we downloaded and normalized the mRNA expression data of 182 patients with EC from the TCGA database and eliminated partial incomplete

[4]http://timer.cistrome.org/

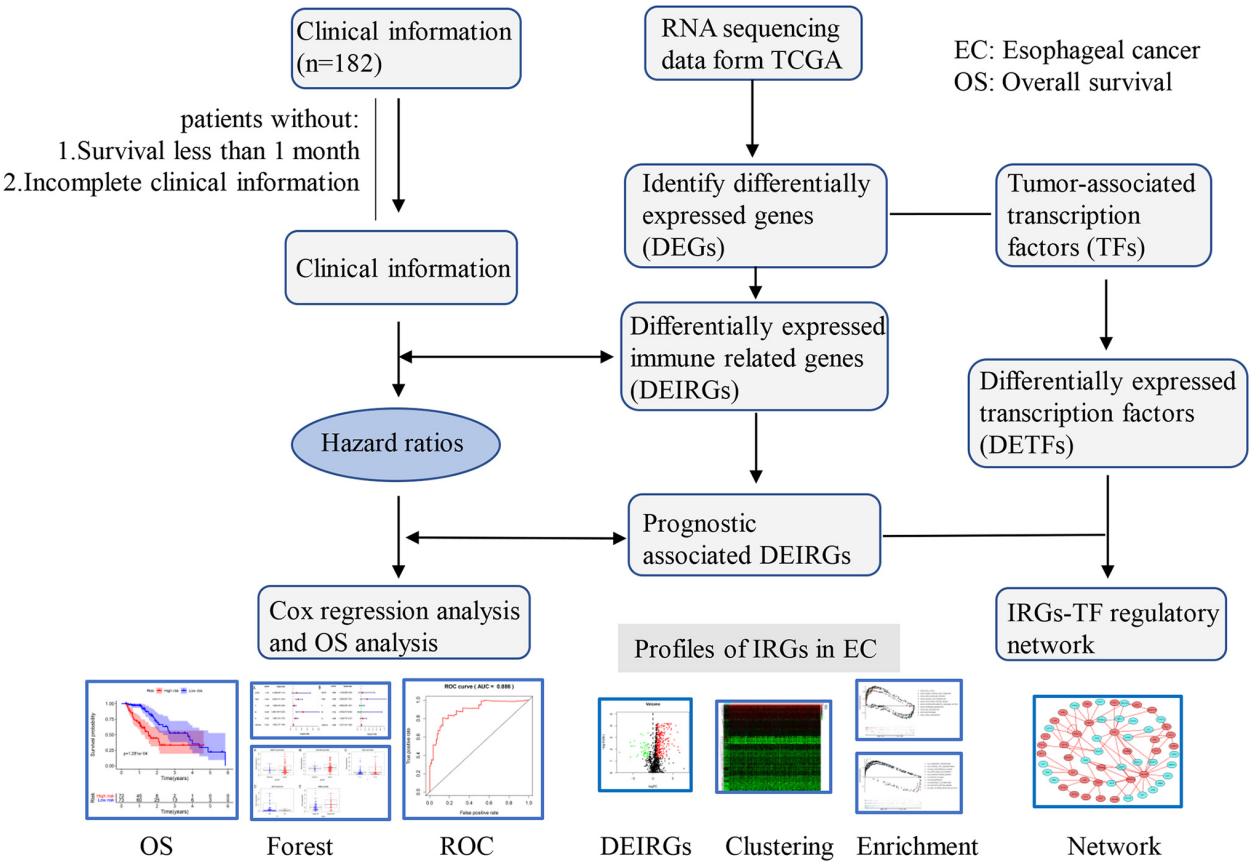


FIGURE 1 | Flowchart of the study. RNA-Seq data and corresponding clinical information of EC cohort were downloaded from the TCGA data portal. After excluding patients with incomplete clinical data and duplications, the complete data was used for subsequent analysis.

data. Then, we performed a differential expression analysis using Wilcoxon test with a $\log_2(FC) > 1$ and $p < 0.05$. We found 4,094 DEGs between 10 normal samples and 162 tumor samples (**Figures 2A,B**). The DEGs list, including $\log_2$FC and the FDR adjusted *p*-values of each gene was provided in **Supplementary Table S2**. Then, we performed GO and KEGG pathway analysis for the DEGs and the top 10 GO and KEGG pathway enrichment terms shown in **Figures 2C,D**. The KEGG analysis indicated that the genes were mainly involved in cytokine-cytokine receptor interaction and cell cycle signaling pathway, which are pivotal in the regulation of immune responses (Murphy and Murphy, 2010; Zhang J. et al., 2018). Next, we downloaded the list of IRGs from the "ImmPort" database. These IRGs intersect with the DEGs, and 303 differentially expressed IRGs were obtained (**Figure 3A**), including 56 down-regulated and 247 up-regulated genes (**Figures 3B,C**).

Prognostic Immune Signatures in EC

Clinical EC data corresponding to RNA sequencing data were downloaded from the TCGA database, and data with a survival time of less than 1 month were excluded. Then, we merged the survival time and survival status of each patient with gene expression data. Then, we set filter criteria of $p < 0.05$ and used univariate Cox regression analysis. Seventeen (*HSPA1A, HSPA1B, HSPA6, IL1B, FABP3, CST4, CACYBP, CCL3, CCL3L1, DKK1, EGF, FGF19, GAST, OSM, ANGPTL3, NR2F2*, and *OXTR*) prognostic immune signatures were obtained (**Figure 4**).

Establishment and Verification of Prognostic Model

Through further analysis via Lasso and multivariate Cox proportional hazards regression analysis, we ultimately obtained 9 optimal prognostic immune genes and incorporated them into the prognostic risk model: *HSPA6, CACYBP, DKK1, EGF, FGF19, GAST, OSM, ANGPTL3*, and *NR2F2*. All the 9 genes are high-risk genes, as shown in **Table 1**. We used gene mRNA levels and risk estimate regression coefficients to calculate risk score for each patient to explore the significance of prognostic genes. The calculation formula is described in the methods. Risk score = (-0.008235 × expression of *HSPA6*) + (0.492 × expression of *CACYBP*)+ (0.014939 × expression of *DKK1*) + (0.29151 × expression of *EGF*) +(0.004 × expression of *FGF19*) + (0.03515 × expression of *GAST*) + (0.327446 × expression of *OSM*) + (0.732285 × expression of *ANGPTL3*) + (0.018484 × expression of *NR2F2*). Then, those prognostic genes were verified between 182 tumor samples

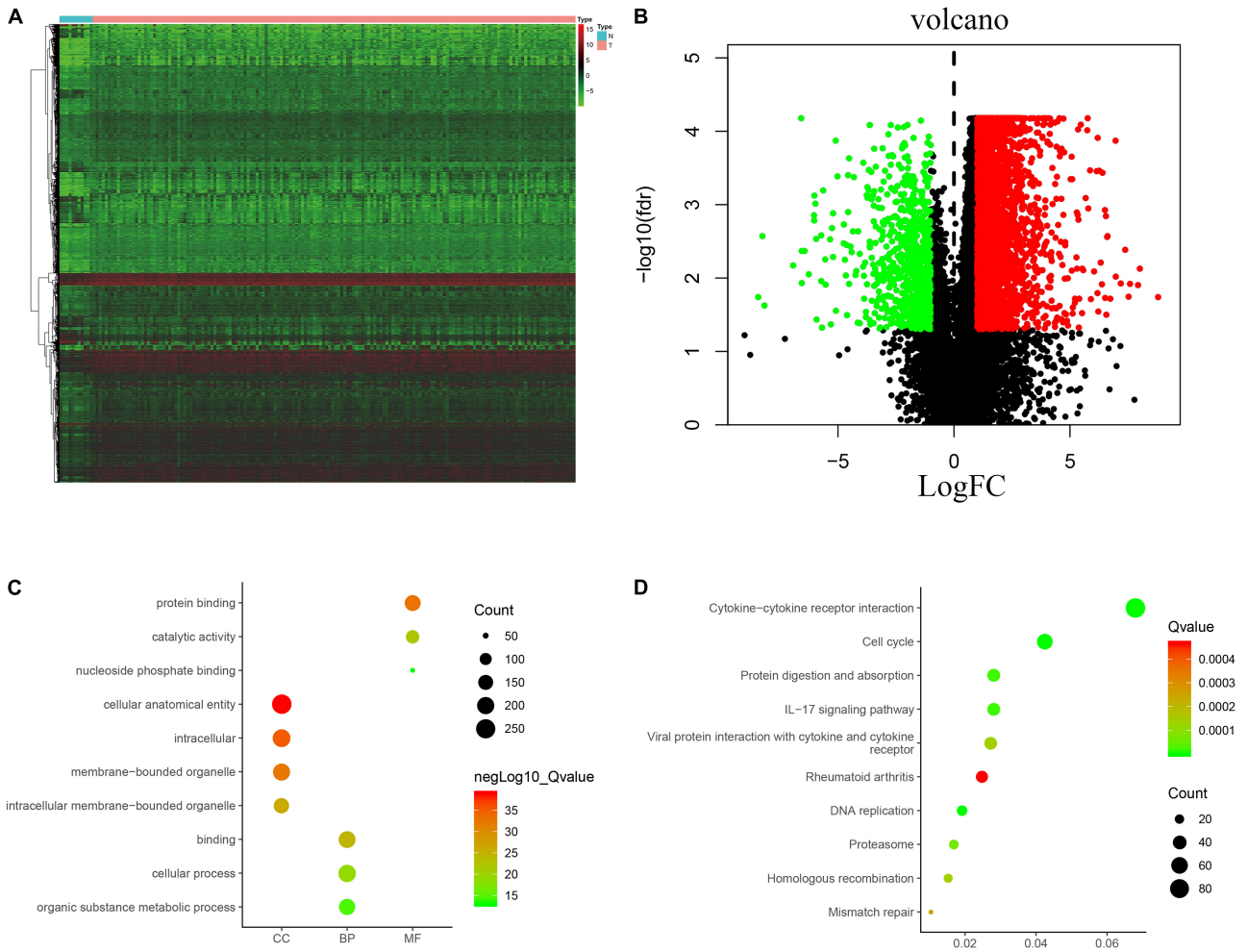


FIGURE 2 | Expression of genes and function enrichment. Heatmap **(A)** and volcano plot **(B)** showing the DEGs between EC and normal esophageal specimen. Red dots represent up-regulated and green dots represent down-regulated DEGs, black dots represent no difference, respectively (fold change >2, $p < 0.05$). GO **(C)** and KEGG **(D)** showing the differentially expressed immune-related genes. **(C)** GO analysis results showing that DEGs were particularly enriched in BP, CC, and MF. **(D)** The significantly enriched pathways of the DEGs determined by KEGG analysis. GO, gene ontology; BP, biological process; CC, cell component; MF, molecular function; KEGG, Kyoto Encyclopedia of Genes and Genomes.

of TCGA database and matched 286 normal samples from GETx database (**Figure 5**). Thus, we found *HSPA6, CACYBP, DKK1, GAST, OSM* were up-regulated in EC tissues ($p < 0.05$ and logFC > 1).

Then, patients were divided into a low-risk group and a high-risk group according to the median risk score. We used the log-rank test to plot survival curves to evaluate the difference in OS between the two groups. As shown in **Figure 6A**, the prognosis of the low-risk group was significantly better than that of the high-risk group ($p = 1.281e-04$). The 1-year survival rates for the high-risk and low-risk groups were 67% (95% CI: 56.8–79.5%) and 95% (95% CI: 90.14–100%), respectively. The 2-year survival rates for the high-risk and low-risk groups were 38% (95% CI: 25.1–59.9%) and 69% (95% CI: 56.79–84.7%), respectively. Here, because of the poor prognosis in the high-risk group, we could not obtain a complete 5-year survival rate. In order to test the predictive accuracy of the model, we constructed a ROC curve. The AUC value for the prognostic model was 0.886, which illustrates the accuracy of the model (**Figure 6B**). Then, we ranked patients according to their risk score and analyzed their distribution using the median risk score as the cut-off (**Figure 6C**). It can be seen that after patients were sorted according to risk score, as the risk score increases, more and more patients die, i.e., the higher the risk score, the greater was the number of deaths. Similarly, the higher the risk score, the shorter the survival time of the patient. The distribution of survival status, survival time and risk score were shown in **Figure 6D**. As the risk score increases, the expression of high-risk genes also increases, and vice versa. Expression patterns of risk genes in the low-risk group and high-risk group are shown in a heat map (**Figure 6E**). The risk score in the high-risk group was significantly higher than that in the low risk group (**Figure 6F**), and the survival time of patients in the high-risk group was significantly lower than that in the low risk group

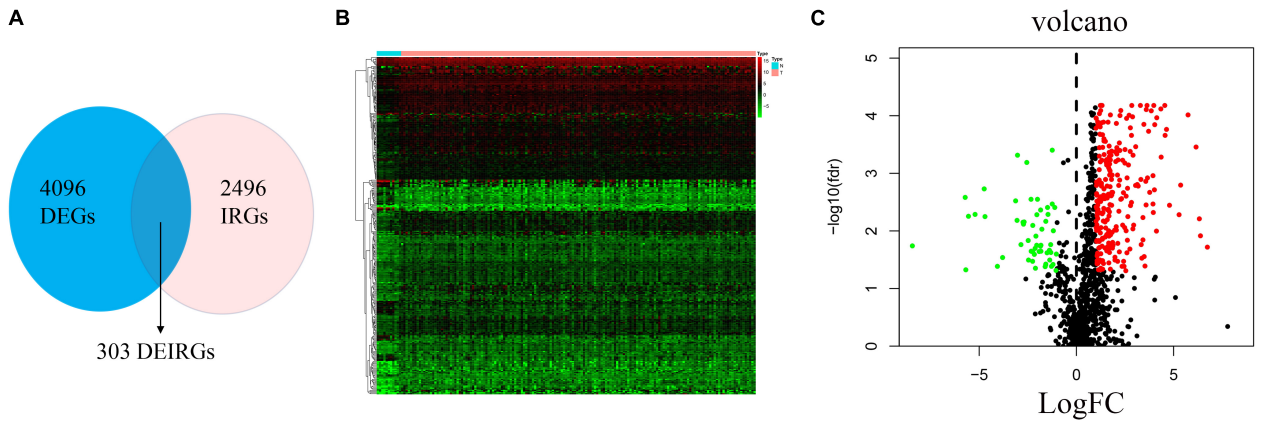


FIGURE 3 | Differential expression of immune-related genes. **(A)** The intersection of DEGs and IRGs. Heatmap **(B)** and volcano plot **(C)** showing the DEGs between EC and normal esophageal specimen. Red dots represent up-regulated and green dots represent down-regulated DEGs, black dots represent no difference, respectively (fold change >2, $p < 0.05$).

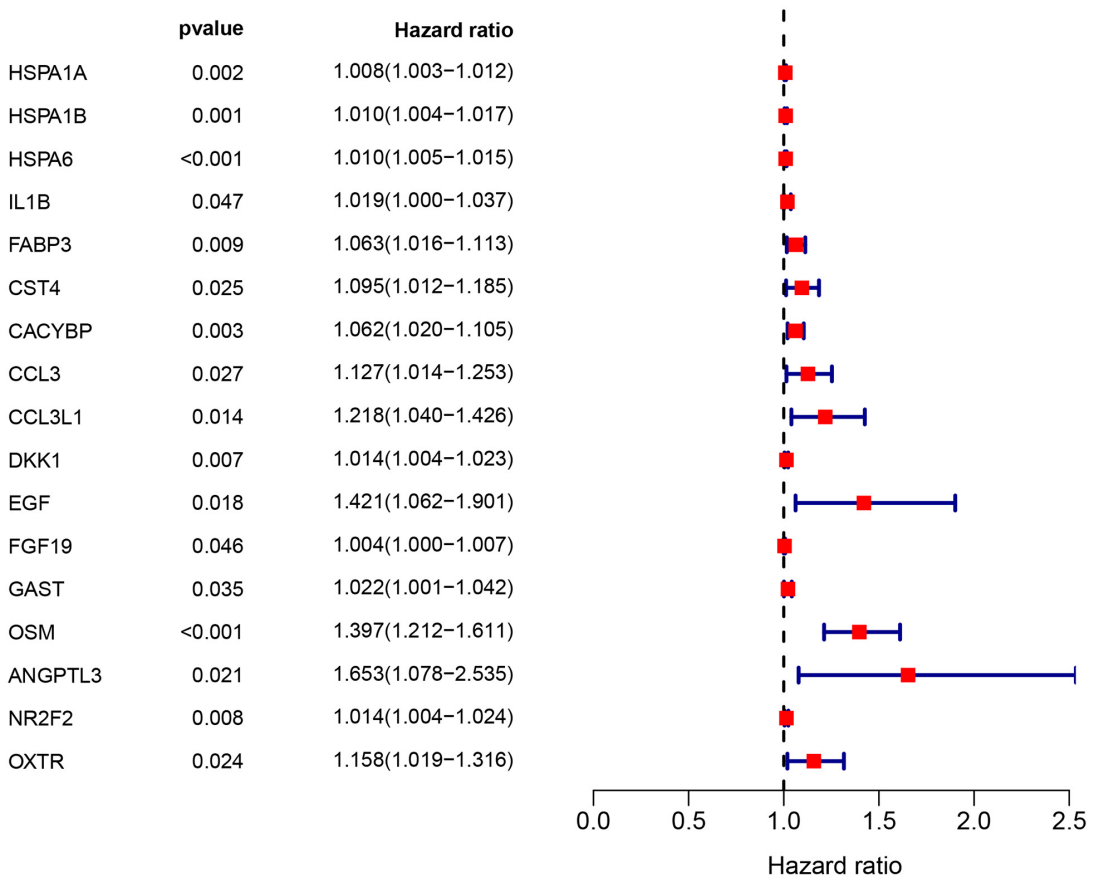


FIGURE 4 | The prognostic value of prognostic associated IRGs in EC. Univariate regression analysis of IRGs related to survival. $p < 0.05$ indicates a significant correlation between genes and prognosis, hazard ratio (HR) value >1 means that the gene is a high-risk gene, and HR <1 means a low-risk gene.

(**Figure 6G**), and the risk score was negatively correlated with the survival time of patients (**Figure 6H**). Those results show that the risk score in the model has an accurate predictive effect on the prognosis of patients.

Independent Prognostic Value of the Risk Model

First, we used univariate regression analysis to determine the correlation between clinical characteristics (age, gender, stage,

TABLE 1 | Coefficients and multivariable Cox model results for immune related genes in esophageal cancer.

Gene symbol	Coef	HR	(95%CI)	*p*-value
HSPA6	0.008235	1.008269	(1.001731–014852)	0.013119
CACYBP	0.043103	1.044046	(0.99238–1.098401)	0.095996
DKK1	0.014939	1.015051	(1.004806–1.025401)	0.003942
EGF	0.291513	1.338447	(0.993541–1.803087)	0.055194
FGF19	0.004144	1.004148	(1.000211–1.008102)	0.038915
GAST	0.034152	1.03474	(1.013293–1.05664)	0.001395
OSM	0.327446	1.387419	(1.178695–1.633105)	8.27E-05
ANGPTL3	0.732285	2.079828	(1.319571–3.278099)	0.001607
NR2F2	0.018484	1.018656	(1.00547–1.032014)	0.005427

coef, coefficient; HR, hazard ratio; CI, confidence interval.

and TNM staging) and prognosis. We found that age ($p = 0.007$), stage ($p < 0.001$), M staging ($p < 0.001$), N staging ($p = 0.005$) and risk score ($p < 0.001$) were significantly correlated with prognosis (**Figure 7A**). Then, we used multivariate analysis to determine the independent prognostic value of the risk model, and the results showed that age ($p = 0.001$), stage ($p = 0.021$), and risk score ($p = 0.005$) were independently associated with prognosis (**Figure 7B**). These results indicate that the prognostic risk model can be used to predict the prognosis of patients with EC accurately and independently. Subsequently, we used ROC curves to verify the accuracy of risk score in evaluating prognosis. The fact that the AUC is 0.850 also indicates the exactitude of our model (**Figure 7C**). Meanwhile, for better prediction of the prognosis of patients with EC at 1, 2, and 3 years after diagnosis, we constructed a new nomogram based on OS-related variables (age, sex, stage, and risk score). The higher the patient's total score, the worse is their prognosis (**Figure 7D**).

Correlation Between the Prognostic Factors and Clinicopathologic Parameters

To confirm our model's ability to predict EC progression, we also analyzed the potential relationship between the risk genes (*HSPA6, CACYBP, DKK1, EGF, FGF19, GAST, OSM, ANGPTL3*, and *NR2F2*), risk score and clinicopathologic parameters, including patient sex, tumor grade, and TNM staging. As shown in **Figures 8A,B**, *ANGPTL1* and *CACYBP* were significantly overexpressed in female patients. As the expression of DKK1 increases, the risk of T staging increases in patients with EC (**Figure 8C**). However, as *FGF19* expression decreased, the risk of distant metastasis decreased (**Figure 8D**). High expression of *OSM* was significantly correlated with high stage (**Figure 8E**). These results suggest

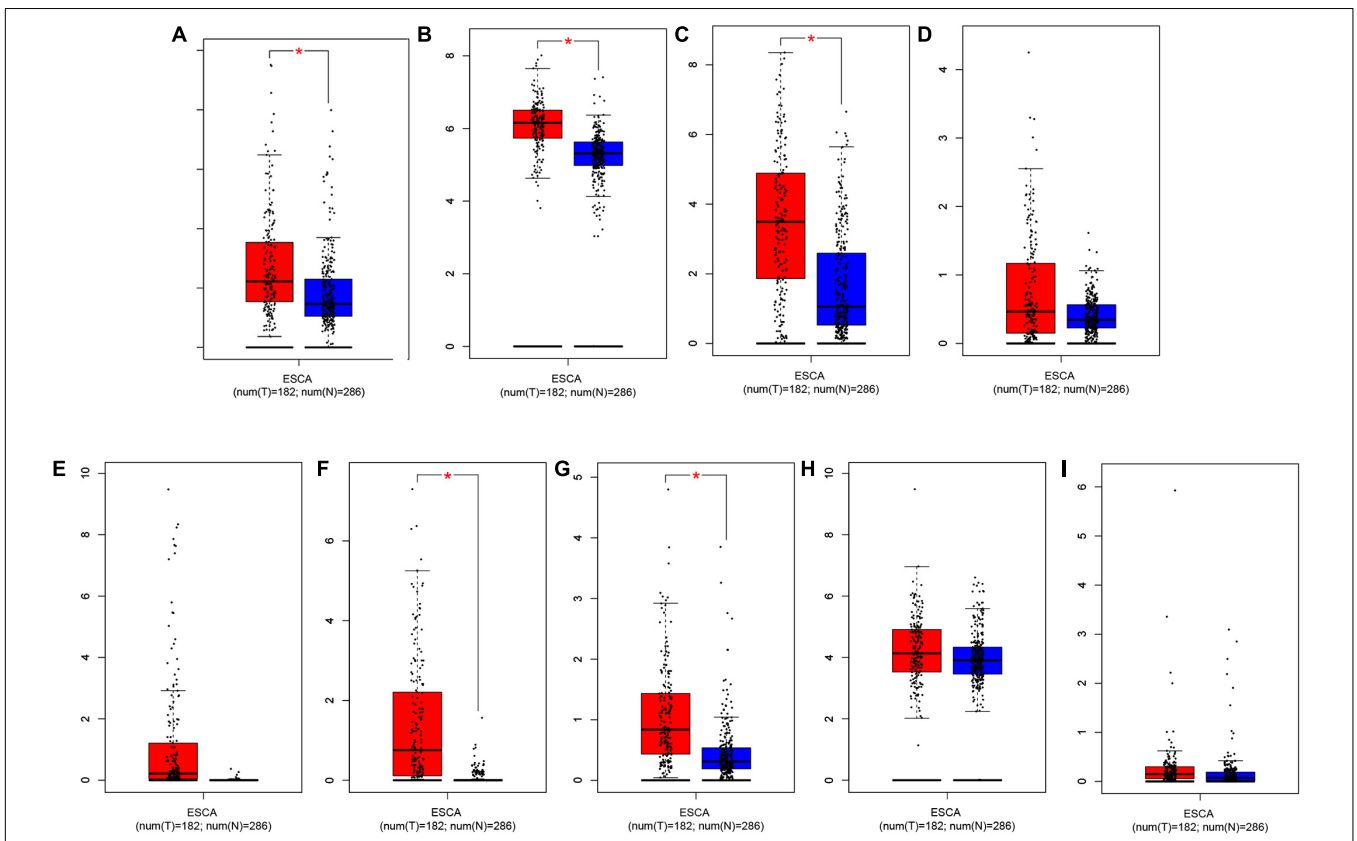


FIGURE 5 | Relative expression of prognostic-related IRGs between EC sample in TCGA database ($n = 182$) and normal esophageal sample form GTEx database ($n = 286$). **(A)** *HSPA6,* **(B)** *CACYBP*, **(C)** *DKK1*, **(D)** *EGF*, **(E)** *FGF19*, **(F)** *GAST*, **(G)** *OSM*, **(H)** *ANGPTL3*, and **(I)** *NR2F2* *$p < 0.05$.

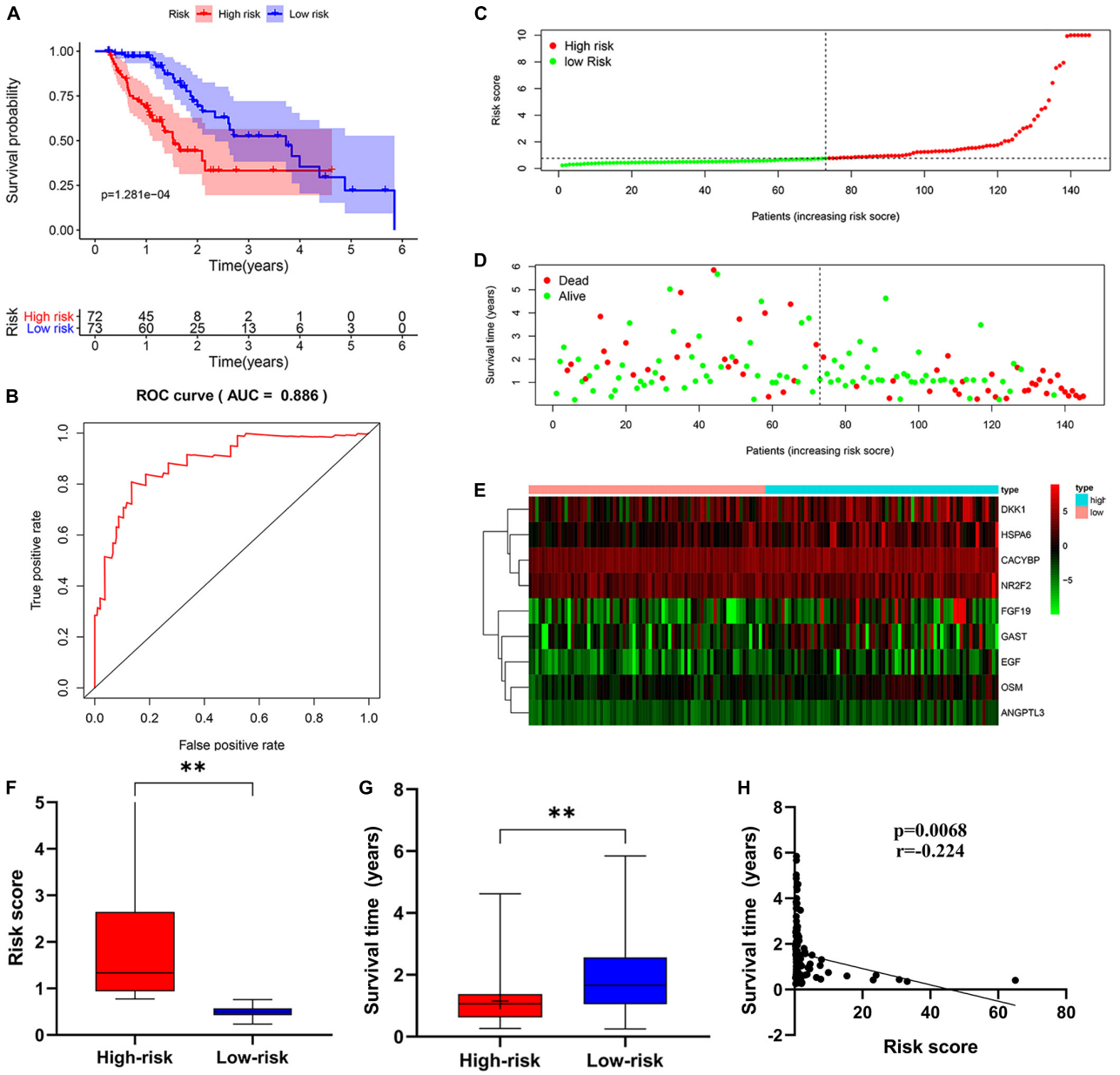


FIGURE 6 | The prognostic value of the immune-related risk score. **(A)** Patients in high-risk group suffered shorter OS. The blue represents the overall survival of patients in the low-risk group; the red represents the overall survival of patients in the low-risk group. **(B)** Survival-dependent receiver operating characteristic (ROC) curve validation of prognostic value of the prognostic index. **(C)** The risk score distribution. Green dots represent risk score for low-risk patients; red dots represent risk score for high-risk patients. **(D)** The relationship between survival status and risk score. The abscissa represents the number of patients, and the ordinate is the risk score. Red dots represent dead patients, green dots are living patients. **(E)** Risk gene expression and risk score (abscissa) in EC patients. **(F)** Risk score in high and low-risk group. **(G)** Patient survival time in high and low-risk group. **(H)** The correlation between survival time and risk score. *$p < 0.05$, **$p < 0.001$.

that the development of EC may be related to dysregulated expression of IRGs.

Immune Cell Infiltration Analysis

To determine whether there is a correlation between risk score and tumor infiltration with immune cells ($CD8^+$ T cells, $CD4^+$ T cells, B cells, macrophages, neutrophils and dendritic cells), we conducted a correlation test between immune cell infiltration and risk score, as shown in **Figure 9**. The risk score had no significant correlation with B cells (p = 0.434), $CD4^+$ T cell (p = 0.666) or $CD8^+$ T cells (p = 0.385) (**Figures 9A–C**). However, the risk score positively correlated with the levels of dendritic cell infiltration (cor = 0.180, p-value = 0.030) (**Figure 9D**), macrophage cells (cor = 0.191, p-value = 0.021) (**Figure 9E**) and neutrophil cells (cor = 0.394, p-value = 9.348e-07) (**Figure 9F**).

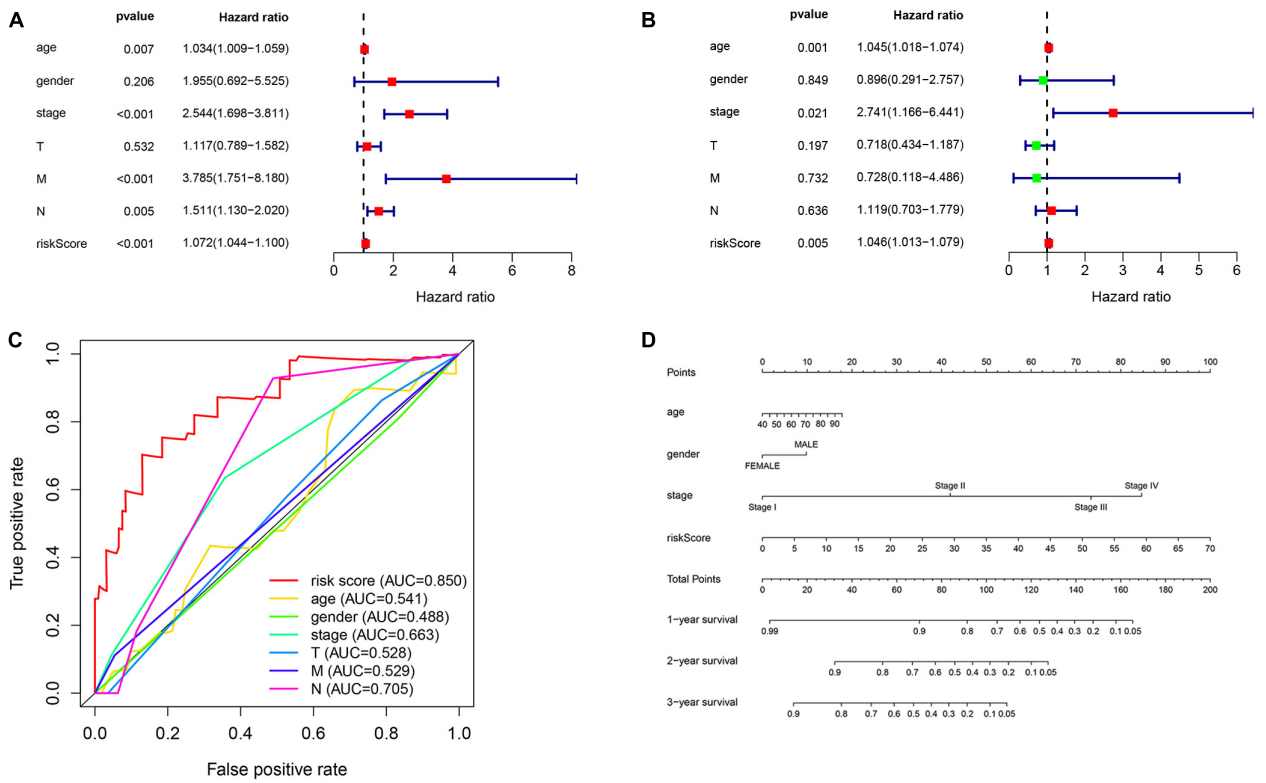


FIGURE 7 | Independent prognostic value of the risk model. **(A)** Univariate and **(B)** multivariate regression analysis of clinical characteristics and risk score as independent prognostic factors. **(C)** The ROC curve evaluated the accuracy of independent prognostic factors for EC. **(D)** A nomogram predict the outcome of EC patients based on their clinical characteristics.

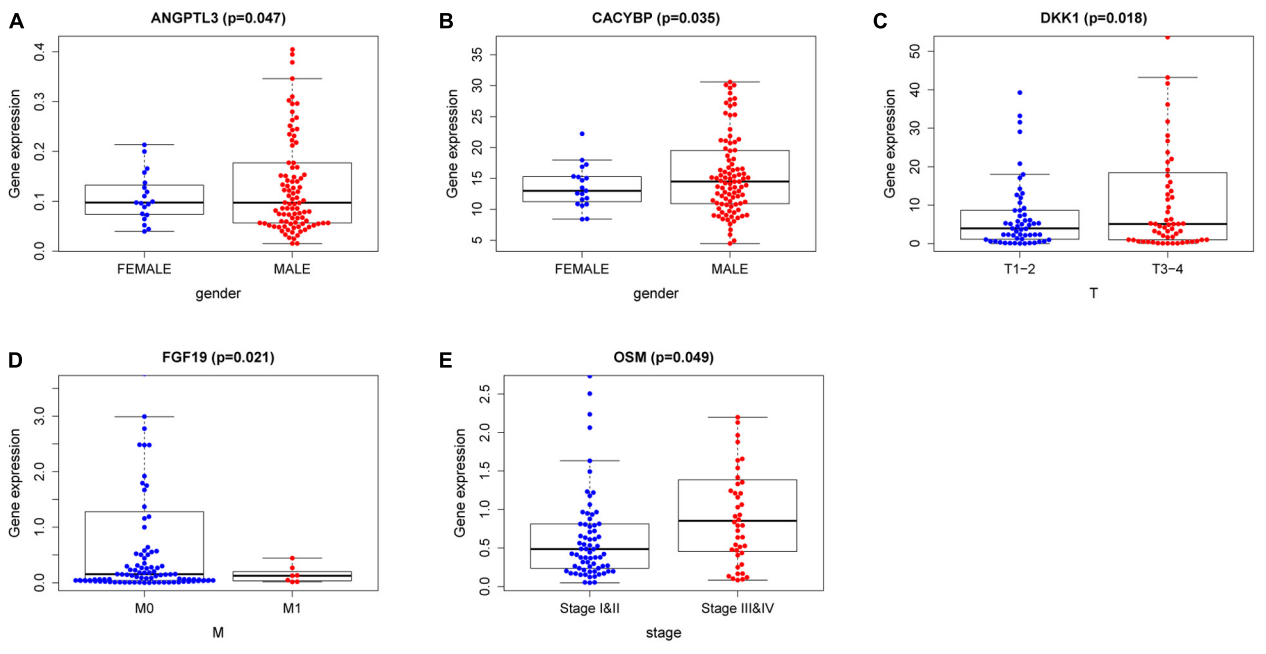


FIGURE 8 | Relationships of the variables in the model with the clinical characteristics of patients. **(A)** *ANGPTL3* expression and gender. **(B)** *CACYBP* expression and gender. **(C)** *DKK1* expression and T staging. **(D)** *FGF19* expression and M stage. **(E)** *OSM* expression and pathological stage. The three horizontal lines in each picture means mean ± SD.

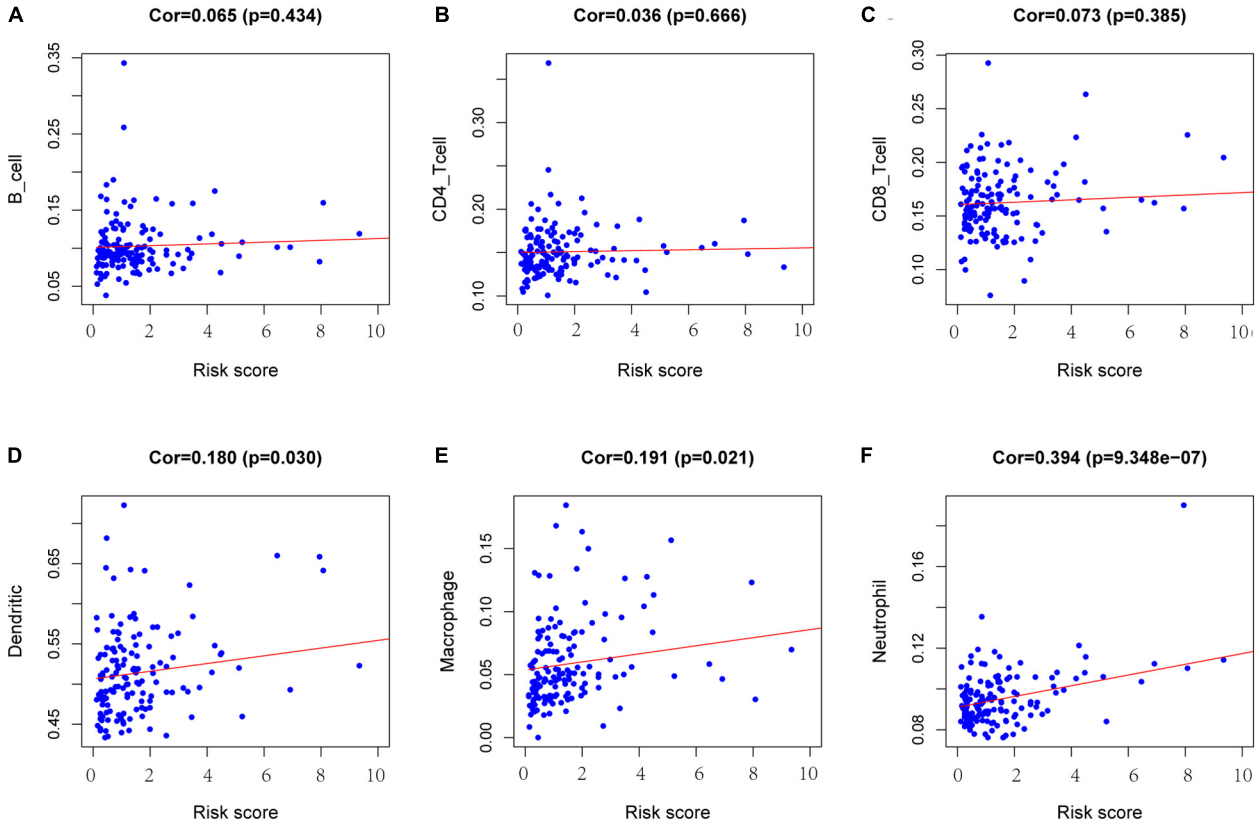


FIGURE 9 | Analysis of the correlation between the risk score and immune cell infiltration. **(A)** B cells. **(B)** CD4+ T cells. **(C)** CD8+ T cells. **(D)** Dendritic cells. **(E)** Macrophages. **(F)** Neutrophils. Cor >0.4 and $p < 0.05$ was used for correlation test.

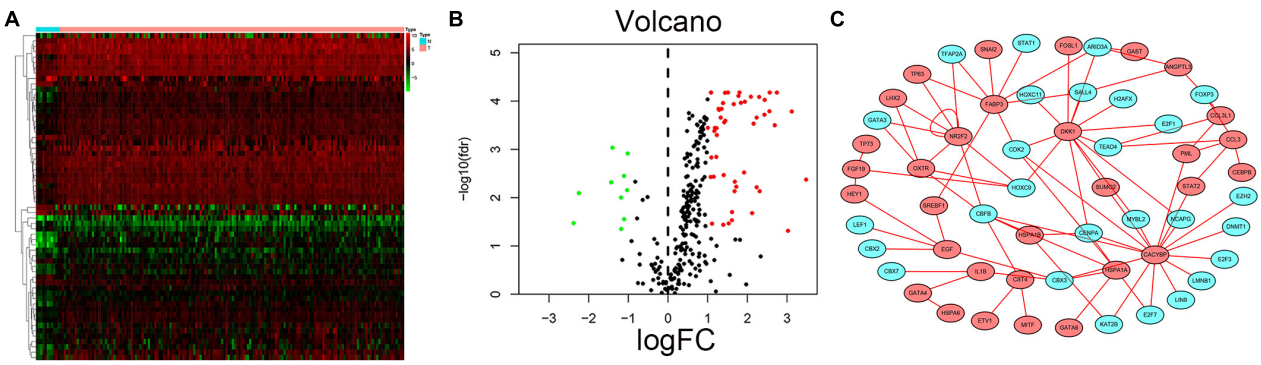


FIGURE 10 | Prognostic associated IRGs and TFs regulatory network. Heatmap **(A)** and volcano plot **(B)** show the differentially expressed transcription factors between EC and esophageal normal specimen. Red dots represent up-regulated and green dots represent down-regulated DEGs, black dots represent no difference, respectively. **(C)** Regulatory network of TFs and prognostic related IRGs; the green nodes represent TFs and the red nodes represent prognostic related IRGs. Correlation coefficient >0.3 and $p < 0.05$.

Construction of a Survival-Associated IRG and TF Regulatory Network

Transcription factors play an important role in the regulation of genes. To explore possible mechanisms of survival-associated IRG dysregulation in EC, we analyzed the correlation between tumor-related transcription factors (TFs) and survival-associated IRG expression. We screened 60 (FDR < 0.05, $\log_2$FC > 2) TFs that were differentially expressed between EC and normal tissues from 318 transcription factors in the "Cistrome" database (**Figures 10A,B**). Next, we used a p-value < 0.05 and correlation coefficient >0.3 as the cut-off values to analyze the correlations between the 60 TFs and survival-associated IRGs. Among the 60 TFs, 27 were significantly associated with survival-associated

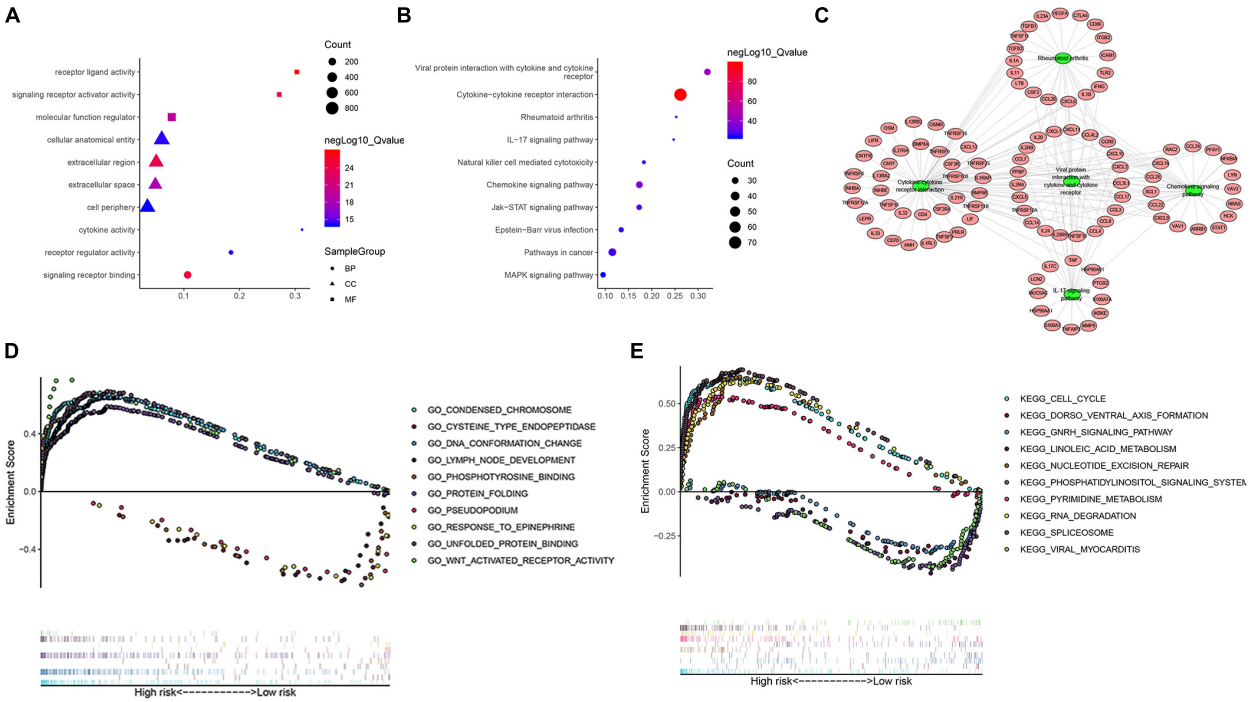


FIGURE 11 | The function enrichment analysis of the IRGs. Differentially expressed IRGs **(A)** Gene Ontology (GO) analysis, **(B)** Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. **(C)** The networks between IRGs and top 5 enrichment pathway. **(D)** GO analysis of Gene Set Enrichment Analysis (GSEA) in high-risk and low-risk groups, **(E)** KEGG analysis of GSEA in high-risk and low-risk groups.

IRGs. To better explain the regulatory relationship, Cytoscape software was used to draw the regulatory network, as shown in **Figure 10C**.

Enrichment Analysis of IRGs

To further study the potential function and mechanism of IRGs, we performed Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) analysis by using "clusterprofiler" R packages. The top 10 GO enrichment terms included biological process (BP), molecular function (MF) and cell component (CC), as shown in **Figure 11A**. The KEGG enrichment analysis results show that it is mainly enriched in some key immune-related pathways, such as chemokine signaling pathway, cytokine-cytokine receptor interaction and JAK-STAT signaling pathways (**Figure 11B**). Based on the relationship between IRGs and KEGG pathways, we constructed a network using Cytoscape to show the genes enriched in the top 5 pathways (**Figure 11C**). In addition, we also observed which pathways were enriched in patients in the high-risk and low-risk groups by using Gene Set Enrichment Analysis (GSEA) software. The top five GO terms enriched in the high-risk and low-risk groups are shown in **Figure 11D**, and the top 5 pathways enriched in the high-risk and low-risk groups are shown in **Figure 11E**. The results showed that key important pathways, such as the cell cycle, pyrimidine metabolism and RNA degradation, were significantly activated in the high-risk group. The GNRH signaling pathway, viral myocarditis, spliceosome pathway and other pathways were active in the low-risk group.

DISCUSSION

Esophageal cancer (EC) is a clinically challenging disease that requires a multidisciplinary approach (Lagergren et al., 2017). The high fatality rate of EC is a cause of concern around the world. Despite incremental advances in diagnostics and therapeutics, EC still carries a poor prognosis, and thus, there remains a need to elucidate the molecular mechanisms underlying this disease. Increasing evidence shows that a comprehensive understanding of EC requires attention not only to tumor cells but also to the tumor microenvironment (Lin et al., 2016). Further study on the relationship between immune signals and EC occurrence and development will help to develop new and specific targeted therapy strategies, especially in combination therapy, with great potential (Li et al., 2017).

In this study, we performed a comprehensive analysis of IRGs and immune infiltrating cells in EC and linked the data to clinical outcomes and prognosis of patients with EC. First, we systematically studied the IRGs in EC. We identified 303 differentially expressed IRGs. They are mainly enriched in the chemokine signaling pathway, cytokine-cytokine receptor interaction, *NF-κB* signaling pathway and JAK-STAT signaling pathway. Recent research reported that tumor cell-secreted IL-6 and IL-8 impair the activity and function of NK cells via *STAT3* signaling, and contribute to esophageal squamous cell carcinoma malignancy (Wu et al., 2019). *NF-κB* is overexpressed in many solid and liquids tumors, including both ESCC and EAC (Karin et al., 2002). Our results are the same as before, and some

of these pathways play an important role in EC (Izzo et al., 2006). Zhang B. et al. (2018) reported on IRGs, specifically that *TSPAN15* interacts with BTRC to promote esophageal squamous cell carcinoma metastasis by activating *NF-κB* signaling and indicated that *TSPAN15* may serve as a new biomarker and/or provide a novel therapeutic target for patients with OSCC. This suggests that IRGs can be used as prognostic biomarkers. To study the underlying mechanisms of EC development, we constructed an IRG-TF regulatory network and found 27 TFs related to prognostic genes; among them, *NR2F2* is both an IRG and TF and is involved in transcriptional regulation.

It makes sense to stratify patients and find predictive prognostic markers. Yuting He et al. found that a new model based on IRGs was effective in predicting prognosis, evaluating disease state, and identifying treatment options for patients with hepatocellular carcinoma (He et al., 2020). Therefore, we used univariate regression analysis to identify IRGs associated with prognosis and tested the value of these survival-associated IRGs for the prognostic stratification of patients. We finally identified the nine best candidate genes (*HSPA6, CACYBP, DKK1, EGF, FGF19, GAST, OSM, ANGPTL3*, and *NR2F2*) through a combination of Cox regression analyses and Lasso regression. These genes were used to construct a Cox regression risk model. This model can predict the outcome of high- and low-risk groups. The accuracy of the model was tested by ROC curve analysis. Then, we found that the risk score could be used as an independent prognostic factor by using univariate and multivariate regression analysis to determine the correlation between clinical characteristics, risk score and prognosis. A nomogram analysis suggested that by combining the clinical characteristics with the risk score, the 1, 2, and 3-year survival rates for EC can be predicted based on the patient's score.

An increasing number of studies about the tumor microenvironment (TME) have been published in the field of cancer immunotherapy (Fidler, 2003). For example, it has been reported in lung cancer (Shi et al., 2020), endometrial cancer (Chen et al., 2020), cervical squamous cell carcinoma (Pan et al., 2019) and so on. Tumor escape from antitumor immunity is essential for tumor survival and progression. Tumor cells can suppress the antitumor immune response via recruitment of various immune cell populations or expression of inhibitory molecular factors. Therefore, we explored the correlation between risk score and immune infiltrating cells and found that risk score in the model were not correlated with $CD8^+$ T cells, B cells, or $CD4^+$ T cells but were significantly correlated with dendritic cells, macrophage cells and neutrophil cells. The positive correlation between high risk score and immune cells also confirmed the accuracy of the model.

In conclusion, we constructed a prognostic model of EC based on IRGs that can accurately predict the prognosis of patients with EC. Furthermore, this model may help to identify new therapeutic targets for advanced EC and provide individualized immunotherapy for patients with EC. Further study of these survival-associated IRGs may shed light on the pathogenesis of EC.

DATA AVAILABILITY STATEMENT

Transcriptomic data and matching clinical data were downloaded from the TCGA GDC portal (https://portal.gdc.cancer.gov/). The 2498 immune genes were obtained from the ImmPort databae (https://www.immport.org/home) (Bhattacharya et al., 2018). Transcription factors (TFs) associated with cancer data and immune cell infiltrate data (including the abundances of CD8+T cells, B cells, macrophages, CD4+T cells, dendritic cells and neutrophils) were both obtained from the Cistrome project (http://www.cistrome.org/) (Liu et al., 2011).

AUTHOR CONTRIBUTIONS

XG and YW conceived, designed this research, and assisted in writing the manuscript. HZ, CQ, and XD conducted the data and statistics analysis. JL, AC, and ZW edited and revised the manuscript. ZW was responsible for supervising the study. All authors read and gave final approval of the manuscript.

FUNDING

This study was supported by the Chongqing Municipal Key Discipline Funding (201128GRJFJY).

ACKNOWLEDGMENTS

We thank the TCGA network for its generous sharing of large amounts of data. We thank the reviewers for their constructive comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00989/full#supplementary-material

REFERENCES

Allum, W. H., Stenning, S. P., Bancewicz, J., Clark, P. I., and Langley, R. E. (2009). Long-term results of a randomized trial of surgery with or without preoperative chemotherapy in Esophageal Cancer. *J. Clin. Oncol.* 27, 5062–5067. doi: 10.1200/JCO.2009.22.2083

Bhattacharya, S., Dunn, P., Thomas, C. G., Smith, B., Schaefer, H., Chen, J., et al. (2018). Immport, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* 5:180015. doi: 10.1038/sdata.2018.15

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global Cancer statistics 2018: globocan estimates of incidence and

mortality worldwide for 36 Cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Cao, J., Yang, X., Li, J., Wu, H., Li, P., Yao, Z., et al. (2019). Screening and identifying immune-related cells and genes in the tumor microenvironment of bladder urothelial carcinoma: based on TCGA database and bioinformatics. *Front. Oncol.* 9:1533. doi: 10.3389/fonc.2019.01533

Chen, P., Yang, Y., Zhang, Y., Jiang, S., Li, X., and Wan, J. (2020). Identification of prognostic immune-related genes in the tumor microenvironment of endometrial cancer. *Aging* 12, 3371–3387. doi: 10.18632/aging.102817

Cunningham, D., Allum, W. H., Stenning, S. P., Thompson, J. N., Van de Velde, C. J., Nicolson, M., et al. (2006). Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. *N. Engl. J. Med.* 355, 11–20. doi: 10.1056/NEJMoa055531

Fidler, I. J. (2003). The pathogenesis of cancer metastasis: the 'Seed and Soil'. Hypothesis Revisited. *Nat. Rev. Cancer* 3, 453–458. doi: 10.1038/nrc1098

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., et al. (2015). The prognostic landscape of genes and infiltrating immune cells across human Cancers. *Nat. Med.* 21, 938–945. doi: 10.1038/nm.3909

He, Y., Dang, Q., Li, J., Zhang, Q., Yu, X., Xue, M., et al. (2020). Prediction of hepatocellular carcinoma prognosis based on expression of an immune-related gene set. *Aging* 12, 965–977. doi: 10.18632/aging.102669

Huang, T. X., and Fu, L. (2019). The immune landscape of esophageal Cancer. *Cancer Commun.* 39:79. doi: 10.1186/S40880-019-0427-Z

Izzo, J. G., Correa, A. M., Wu, T. T., Malhotra, U., Chao, C. K., Luthra, R., et al. (2006). Pretherapy nuclear factor-kappab status, chemoradiation resistance, and metastatic progression in esophageal carcinoma. *Mol. Cancer Ther.* 5, 2844–2850. doi: 10.1158/1535-7163.MCT-06-0351

Karin, M., Cao, Y., Greten, F. R., and Li, Z. W. (2002). Nf-Kappab in Cancer: from innocent bystander to major culprit. *Nat. Rev. Cancer* 2, 301–310. doi: 10.1038/nrc780

Khalil, D. N., Smith, E. L., Brentjens, R. J., and Wolchok, J. D. (2016). The future of Cancer treatment: immunomodulation, cars and combination immunotherapy. *Nat. Rev. Clin. Oncol.* 13, 273–290. doi: 10.1038/nrclinonc.2016.65

Lagergren, J., Smyth, E., Cunningham, D., and Lagergren, P. (2017). Oesophageal Cancer. *Lancet* 390, 2383–2396. doi: 10.1089/omi.2011.0118016/S014

Li, Y., Lu, Z., Che, Y., Wang, J., Sun, S., Huang, J., et al. (2017). Immune signature profiling identified predictive and prognostic factors for esophageal squamous cell carcinoma. *Oncoimmunology* 6:e1356147. doi: 10.1080/2162402X.2017.1356147

Lin, E. W., Karakasheva, T. A., Hicks, P. D., Bass, A. J., and Rustgi, A. K. (2016). The tumor microenvironment in esophageal Cancer. *Oncogene* 35, 5337–5349. doi: 10.1016/S0140-6736(17)31462-9

Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., et al. (2011). Cistrome: an integrative platform for transcriptional regulation Studies. *Genome Biol.* 12:R83. doi: 10.1186/gb-2011-12-8-r83

Mei, S., Meyer, C., Zheng, R., Qin, Q., Wu, Q., Jiang, P., et al. (2017). Cistrome Cancer: a web resource for integrative gene regulation modeling in Cancer. *Cancer Res.* 77, e19–e22. doi: 10.1158/0008-5472.CAN-17-0327

Murphy, T. L., and Murphy, K. M. (2010). Slow down and survive: enigmatic immunoregulation by BTLA and HVEM. *Annu. Rev. Immunol.* 28, 389–411. doi: 10.1146/annurev-immunol-030409-101202

Noble, F., Lloyd, M. A., Turkington, R., Griffiths, E., O'Donovan, M., O'Neill, J. R., et al. (2017). Multicentre cohort study to define and validate pathological assessment of response to neoadjuvant therapy in oesophagogastric adenocarcinoma. *Br. J. Surg.* 104, 1816–1828. doi: 10.1002/bjs.10627

Ohashi, S., Miyamoto, S., Kikuchi, O., Goto, T., Amanuma, Y., and Muto, M. (2015). Recent advances from basic and clinical studies of esophageal squamous cell carcinoma. *Gastroenterology* 149, 1700–1715. doi: 10.1053/j.gastro.2015.08.05

Pan, X. B., Lu, Y., Huang, J. L., Long, Y., and Yao, D. S. (2019). prognostic genes in the tumor microenvironment in cervical squamous cell carcinoma. *Aging* 11, 10154–10166. doi: 10.18632/aging.102429

Qiu, H., Hu, X., He, C., Yu, B., Li, Y., and Li, J. (2020). Identification and validation of an individualized prognostic signature of bladder cancer based on seven immune related genes. *Front. Genet.* 11:12. doi: 10.3389/fgene.2020.00012

Shaib, W. L., Nammour, J. P., Gill, H., Mody, M., and Saba, N. F. (2016). The future prospects of immune therapy in gastric and esophageal adenocarcinoma. *J. Clin. Med.* 5:100. doi: 10.3390/jcm5110100

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shi, X., Li, R., Dong, X., Chen, A. M., Liu, X., Lu, D., et al. (2020). Irgs: an immune-related gene classifier for lung adenocarcinoma prognosis. *J. Transl. Med.* 18:55. doi: 10.1186/s12967-020-02233-y

Tanaka, T., Nakamura, J., and Noshiro, H. (2017). Promising Immunotherapies for Esophageal Cancer. *Expert Opin. Biol. Ther.* 17, 723–733. doi: 10.1080/14712598.2017.1315404

Tang, Z., Kang, B., Li, C., Chen, T., and Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* 47, W556–W560. doi: 10.1093/nar/gkz430

Turato, C., Scarpa, M., Kotsafti, A., Cappon, A., Quarta, S., Biasiolo, A., et al. (2019). Squamous cell carcinoma antigen 1 is associated to poor prognosis in esophageal cancer through immune surveillance impairment and reduced chemosensitivity. *Cancer Sci.* 110, 1552–1563. doi: 10.1111/cas.13986

van Hagen, P., Hulshof, M. C., van Lanschot, J. J., Steyerberg, E. W., van Berge Henegouwen, M. I., Wijnhoven, B. P., et al. (2012). Preoperative chemoradiotherapy for esophageal or junctional Cancer. *N. Engl. J. Med.* 366, 2074–2084. doi: 10.1056/NEJMoa1112088

Vo, J. N., Cieslik, M., Zhang, Y., Shukla, S., Xiao, L., Zhang, Y., et al. (2019). The landscape of circular rna in Cancer. *Cell* 176, 869.e13–881.e13. doi: 10.1016/j.cell.2018.12.021

Wan, B., Liu, B., Huang, Y., Yu, G., and Lv, C. (2019). Prognostic value of immune-related genes in clear cell renal cell carcinoma. *Aging* 11, 11474–11489. doi: 10.18632/aging.102548

Wu, J., Gao, F. X., Wang, C., Qin, M., Han, F., Xu, T., et al. (2019). Il-6 and Il-8 secreted by tumour cells impair the function of Nk Cells Via the Stat3 pathway in oesophageal squamous cell carcinoma. *J. Exp. Clin. Cancer Res.* 38:321. doi: 10.1186/s13046-019-1310-0

Yan, T., Cui, H., Zhou, Y., Yang, B., Kong, P., Zhang, Y., et al. (2019). Multi-region sequencing unveils novel actionable targets and spatial heterogeneity in esophageal squamous cell carcinoma. *Nat. Communi.* 10:1670. doi: 10.1038/s41467-019-09255-1

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). Clusterprofiler: an R Package for comparing biological themes among gene clusters. *OmicsJ. Integr. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, B., Zhang, Z., Li, L., Qin, Y. R., Liu, H., Jiang, C., et al. (2018). Tspan15 interacts with btrc to promote oesophageal squamous cell carcinoma metastasis via activating Nf-Kb signaling. *Nat. Commun.* 9:1423. doi: 10.1038/s41467-018-03716-9

Zhang, J., Bu, X., Wang, H., Zhu, Y., Geng, Y., Nihira, N. T., et al. (2018). Cyclin D-CDK4 kinase destabilizes PD-L1 via cullin 3-SPOP to control cancer immune surveillance. *Nature* 553, 91–95. doi: 10.1038/nature25015

Zhang, M., Wang, X., Chen, X., Zhang, Q., and Hong, J. (2020). Novel immune-related gene signature for risk stratification and prognosis of survival in lower-grade glioma. *Front. Genet.* 11:363. doi: 10.3389/fgene.2020.00363

Zhao, Q., Yu, J., and Meng, X. (2019). A good start of immunotherapy in esophageal Cancer. *Cancer Med.* 8, 4519–4526. doi: 10.1002/cam4.2336

Zhao, Z., Zhao, D., Xia, J., Wang, Y., and Wang, B. (2020). Immunoscore predicts survival in early-stage lung adenocarcinoma patients. *Front. Oncol.* 10:691. doi: 10.3389/fonc.2020.00691

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 August 2020
doi: 10.3389/fgene.2020.00940

A Multi-Gene Model Effectively Predicts the Overall Prognosis of Stomach Adenocarcinomas With Large Genetic Heterogeneity Using Somatic Mutation Features

Xianming Liu[1†], Xinjie Hui[2†], Huayu Kang[2†], Qiongfang Fang[2], Aiyue Chen[2], Yueming Hu[2], Desheng Lu[2], Xianxiong Chen[2] and Yejun Wang[2*]*

[1] Department of Gastrointestinal Surgery, Shenzhen People's Hospital, The Second Clinical Medical College of Jinan University, Shenzhen, China, [2] School of Basic Medicine, Shenzhen University Health Science Center, Shenzhen, China

OPEN ACCESS

Edited by:
Bailiang Li,
Stanford University, United States

Reviewed by:
Edwin Wang,
University of Calgary, Canada
Deli Liu,
Weill Cornell Medicine, United States

***Correspondence:**
Xianxiong Chen
gzcxx@szu.edu.cn
Yejun Wang
wangyj@szu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 31 March 2020
Accepted: 28 July 2020
Published: 26 August 2020

Citation:
Liu X, Hui X, Kang H, Fang Q, Chen A, Hu Y, Lu D, Chen X and Wang Y (2020) A Multi-Gene Model Effectively Predicts the Overall Prognosis of Stomach Adenocarcinomas With Large Genetic Heterogeneity Using Somatic Mutation Features. Front. Genet. 11:940. doi: 10.3389/fgene.2020.00940

Background: Stomach adenocarcinoma (STAD) is one of the most common malignancies worldwide with poor prognosis. It remains unclear whether the prognosis is associated with somatic gene mutations.

Methods: In this research, we collected two independent STAD cohorts with both genetic profiling and clinical follow-up data, systematically investigated the association between the prognosis and somatic mutations, and analyzed the influence of heterogeneity on the prognosis-genetics association.

Results: Typical association was identified between somatic mutations and overall prognosis for individual cohorts. In The Cancer Genome Atlas (TCGA) cohort, a list of 24 genes was also identified that tended to mutate within cases of the poorest prognosis. The association showed apparent heterogeneity between different cohorts, although common signatures could be identified. A machine-learning model was trained with 20 common genes that showed a similar mutation rate difference between prognostic groups in the two cohorts, and it classified the cases in each cohort into two groups with significantly different prognosis. The model outperformed both single-gene models and TNM-based staging system significantly.

Conclusion: The study made a systematic analysis on the association between STAD prognosis and somatic mutations, identified signature genes that showed mutation preference in different prognostic groups, and developed an effective multi-gene model that can effectively predict the overall prognosis of STAD in different cohorts.

Keywords: stomach adenocarcinoma, prognosis, prediction, multi-gene model, heterogeneity

INTRODUCTION

Stomach adenocarcinoma (STAD) represents the global fifth most common malignancy and the third leading cause of cancer mortality, with estimated 1,033,701 newly diagnosed cases and 782,685 deaths in 2018 (Bray et al., 2018). Screening of STADs at early stages with endoscopy and biopsy sampling remains the most effective approach to improve prognosis and reduce the mortality

(Banks et al., 2019). However, the majority of STADs worldwide except Japan and Korea were diagnosed at a late stage, due to the lack of symptoms at early stages, invasiveness of endoscopy, and unsound early-screening programs (Banks et al., 2019). Surgical resection and chemotherapy remain the main treatment regimens (Charalampakis et al., 2018). Although new therapies, such as targeted and immune therapies, have been applied to STADs, the overall outcome was only improved moderately (Tran et al., 2017; Charalampakis et al., 2018).

Multi-omics studies disclosed a high heterogeneity of STADs in genetics (Cancer Genome Atlas Research Network, 2014; Cristescu et al., 2015; Oh et al., 2018), gene expression (Boussioutas et al., 2003; Tan et al., 2011; Lei et al., 2013), and other molecular levels (Ooi et al., 2016; Ni et al., 2019; Zhang et al., 2019). The molecular heterogeneity could be associated with the complexity of anatomic regions of stomach, cell origins, and etiologies (Cancer Genome Atlas Research Network, 2014; Choi et al., 2014; Cristescu et al., 2015; Waldum and Fossmark, 2018; Ni et al., 2019; Zhang et al., 2019). STADs could originate from different anatomic sites such as cardia or gastroesophageal junction, fundus, lesser curvature, greater curvature, angular incisures, antrum, and pylorus, each with different cell compositions (Soybel, 2005; Choi et al., 2014). STADs are divided by the Lauren classification system into intestinal and diffuse types, the latter of which show poor clinical outcomes generally (Laurén, 1965; Shen et al., 2013). The World Health Organization proposed an alternative system, dividing STADs into papillary, tubular, mucinous (colloid), and poorly cohesive carcinomas (Bosman et al., 2010). Recently, genome-based molecular signatures were comprehensively identified and employed by The Cancer Genome Atlas (TCGA) to classify STADs into four subtypes, namely, Epstein–Barr virus positive (EBV), microsatellite instable (MSI), genome stable (GS), and chromosomal instability (CIN) (Cancer Genome Atlas Research Network, 2014). A gene expression-based study from the Asian Cancer Research Group (ACRG) also classified STADs into two major subtypes, MSI and microsatellite stable (MSS), while MSS STADs were further subdivided into three subtypes, epithelial-to-mesenchymal transition (EMT), TP53 active (TP53+), and TP53 inactive (TP53-) (Cristescu et al., 2015). The new molecular classification schemes could have more prospective clinical utilities in guiding STAD therapies and prognosis.

For a variety of tumors, prognosis has been reported to be associated with somatic gene mutations (Loi et al., 2013; Lee et al., 2017; Zhang et al., 2017; Cho et al., 2018; Yu et al., 2019). Despite the large heterogeneity of STADs, common genetic factors (e.g., BRCA2 and MUC16) were still identified and reported to be associated with the prognosis (Chen et al., 2015; Li et al., 2018). Currently, there is still a lack of systemic exploration of the association between STAD prognosis and somatic mutations. To achieve this goal, here, we collected the publically available data from two STAD cohorts that contained both genetic mutation profiles and clinical follow-up information (Cancer Genome Atlas Research Network, 2014; Chen et al., 2015), analyzed the STAD prognosis–genetics association globally and the influence of heterogeneity on the prognosis–genetics association, and identified a list of common genetic signatures that can be used widely for the guidance of STAD prognosis.

MATERIALS AND METHODS

Datasets, Stratification, and Mutation Frequency Comparison

Two STAD cohorts were used in this study, the TCGA cohort and a Chinese cohort (Cancer Genome Atlas Research Network, 2014; Chen et al., 2015). The TCGA cases were multiethnic but mostly white people, while the Chinese cohort was comprised by Chinese patients exclusively. Both the clinical data and the somatic mutation data were downloaded. Mutations causing codon changes, frame-shifts, and premature translational terminations were retrieved for further analysis. For prognosis–genetics association analysis, first, the cases were removed that received targeting therapies. Furthermore, only the ones with both somatic mutation data and corresponding prognostic follow-up information were recruited. The included cases were classified into two categories according to prognosis ("good" or "poor"). The "good" prognosis group included the patients surviving through the preset follow-up period while the "poor" one indicated the patients died within the observed period. The TNM (tumor-nodal-metastasis) staging system was used for stratification, and for the sake of convenience in binary classification, two categories, "early" (Stages I and II) and "later" (Stages III and IV) were predefined. In addition, considering the possible effects of different anatomic sites of tumor on prognosis, subdivisions were used for stratification as well. To compare the somatic gene mutation frequency between prognostic groups, a matrix was prepared to record the mutations of all the genes for each case, followed by counting the number of cases with mutations for each gene in each group. A genome-wide rate comparison test (EBT) proposed recently that could balance statistic power and precision was adopted to compare the gene mutation rates (Hui et al., 2017). To test the robustness of gene mutation signatures identified by EBT tests, a repeated resampling strategy was adopted, by which a subset (70% of the total sample size) of the training cases was randomly selected for 100 rounds, gene mutation rates were compared for each round, and the signature genes were observed for the recurrence among the top 50 genes with smallest p-values for each round (Hui et al., 2017).

Feature Extraction, Representation, and Model Training

Two strategies were adopted for the feature extraction in this research, p-value based and empirical. For the p-value-based strategy, the top n genes with the most significant mutation frequency difference were used as the genetic features. For the empirical strategy, the difference of mutation rates was calculated per gene between the two prognostic groups and ordered, and the genes with a minimal 10% (or any indicated percentage) mutation rate difference and with recurrent mutations in either group were retrieved as candidate features.

For each case, P_j (j = 1, 2,., m_i) belonging to a certain category C_i, where i equaled to 1 or 0, and m_i represented the total number of cases of the category C_i, the genetic features were represented as a binary vector F_j (g_1, g_2,., g_n) in which g_k (k = 1, 2,., n) represented the kth genetic feature, taking the value of 1 if the corresponding gene was mutated and 0 otherwise. There was an m_i*n matrix for category C_i. When stage was used as an additional feature, the size of the matrix was enlarged to m_i*(n + 1), and the stage feature was also represented in a binary form in the additional column, for which 1 and 0 represented "early" and "later," respectively. The anatomic sites were represented as two-bit features, i.e., "cardia/gastro-esophagus junction," "fundus/corpus," and "antrum/pylorus" being represented as "00," "01," and "10," respectively.

An R package, "e1071," was used for training SVM models using each training dataset[1]. During the training stage, all four kernels, "Radial Base Function (RBF)," "linear," "polynomial," and "sigmoid," were tested and the parameters were optimized based on a 10-fold cross-validation grid search. The best kernel with optimized parameters was selected for further model training.

Model Performance Assessment

A 5-fold cross-validation strategy was used in this study. The original feature-represented matrix for each category was randomly split into five parts with identical size. Every four parts of each category were combined and served as a training dataset while the rest one of each category was used for testing and performance evaluation.

The Receiver Operating Characteristic (ROC) curve, the area under the ROC curve (AUC), Accuracy, Sensitivity, Specificity, and Mathews Correlation Coefficient (MCC) were utilized to assess the predictive performance. In the following formula, Accuracy denotes the percentage of both positive instances ("good prognosis") and negative instances ("poor prognosis") correctly predicted. Specificity and Sensitivity represent the true negative rate and true positive rate, respectively, while the default threshold value from "e1070" (0.0) was used to define the Sensitivity and Specificity in the research. An ROC curve is a plot of Sensitivity versus (1 – Specificity) and is generated by shifting the decision threshold. AUC gives a measure of classifier performance. MCC takes into account true and false positives and false negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}),$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}),\ \text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}),$$

$$\text{MCC} = ((\text{TP}^*\text{TN}) - (\text{FN}^*\text{FP}))/\text{Sqrt}((\text{TP} + \text{FN})^*$$

$$(\text{TN} + \text{FP})^*(\text{TP} + \text{FP})^*(\text{TN} + \text{FN})).$$

Survival Analysis

The follow-up survival information of STAD cases was annotated. To evaluate the survival of prediction results of each model, all the 5-fold cross-validation testing results were collected and grouped, followed by the survival analysis for each predicted group. Kaplan–Meier overall survival analysis was performed with R survival package[1]. The Gehan–Breslow–Wilcoxon test was used to compare the difference of overall survival curves, and the significance level was set as 0.05.

[1]https://cran.r-project.org/

TML Analysis

Both Tumor Mutation Load (TML) and Missense TML were analyzed for STAD cases of different prognostic groups. TML is defined as logarithm transformation of mutation rate per megabase, while Missense TML only counts the mutations causing amino acid changes. The Wilcoxon rank-sum test was performed to compare the distributions of TML or Missense TML, with the preset significance level as 0.05.

RESULTS

Somatic Mutation Profile Difference Between Prognostic Groups of TCGA STADs

In total, 142 TCGA STAD cases remained after filtering the duplicates, the ones missing somatic mutation or clinical information and those treated with targeting therapies. The general clinical properties are shown in **Supplementary Table S1**. A somatic mutation profile analysis for these cases disclosed a list of genes with high mutation rates (>30%), including *TTN*, *PCDHAC2*, *PCDHGC5*, *TP53*, *MUC16*, *SYNE1*, and *CSMD* (**Supplementary Table S2**). The cases were also stratified according to sex and anatomic site, and the somatic mutation profiles were compared among the corresponding strata. Six genes were found with significant somatic mutation rates between male and female (EBT, $p < 0.05$), while 5 genes showed marginally significant somatic mutation rates among different anatomic sites (EBT, $p < 0.10$) (**Supplementary Table S2**).

The overall survival of the included TCGA STAD cases appeared poor, with a median of 805 days (**Figure 1A**). The cases were classified into good and poor prognostic groups with identical sample size (each with 40 cases) based on a cutoff survival period (573 days), and somatic gene mutation rates were compared between the groups (**Supplementary Figure S1A**). In total, 52 genes were identified with most striking difference (EBT, $p < 0.20$) (**Supplementary Table S3**). A random resampling procedure further indicated that these genes were stably associated with STAD prognosis (50/52 with the largest recurrence among the top 50 genes of smallest p-values for each resampling test; **Supplementary Table S3**). Genes involved in collagen chain trimerization were significantly enriched (**Supplementary Figure S1B**; Fisher's Exact, FDR = 0.013). Most of the genes (82.7%) were reported to be associated with cancers and 16 (30.8%) with gastric cancer, including MUC16, for which higher mutation rates were recently found to be associated with prognosis and the immune therapy outcome of gastric cancer (Li et al., 2018; **Supplementary Table S3**). With subsets or all of the 52 genes as features, SVM models were trained to predict the

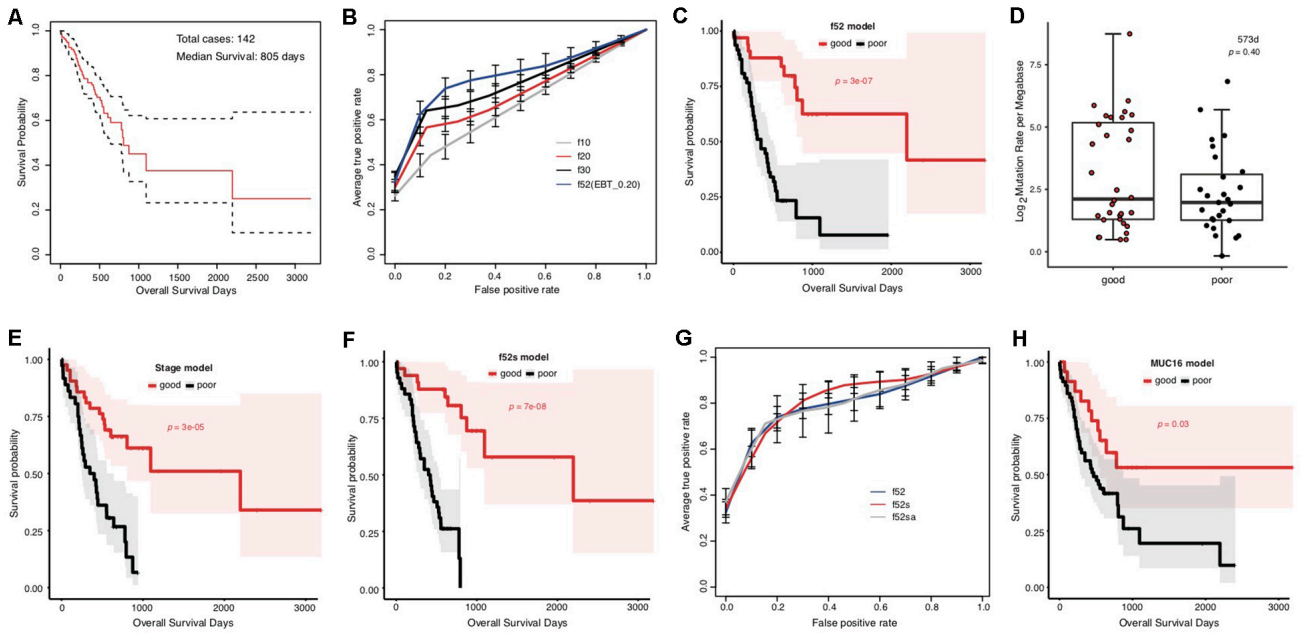


FIGURE 1 | Association between the overall prognosis of TCGA STADs and somatic gene mutations. **(A)** The Kaplan–Meier overall survival curve of the TCGA STAD cases. The median was indicated with an arrow. The analysis throughout the figure used the median survival as stratification cutoff of STAD prognosis. **(B)** The 5-fold cross-validated ROC curves of genetic models predicting the prognosis of STADs. **(C)** The Kaplan–Meier overall survival curves of the TCGA STAD cases classified by f52 with a 5-fold cross-validation strategy. **(D)** The distribution of TMLs for TCGA cases of good and poor prognosis groups. The *p*-value of the Wilcoxon rank-sum test was indicated. **(E,F)** The Kaplan–Meier overall survival curves of the TCGA STAD cases stratified by TNM stage information **(E)** or the combined f52s model with a 5-fold cross-validation strategy **(F)**. **(G)** The 5-fold cross-validated ROC curves of the genetic model f52 and the combined models f52s and f52sa predicting the prognosis of STADs. **(H)** The Kaplan–Meier overall survival curves of the TCGA STAD cases stratified by presence or absence of MUC16 mutations. Gehan–Breslow–Wilcoxon tests were performed to compare the overall survivals, and the *p*-values were shown in context.

tumor prognosis. Generally, the model performance improved as the number of features increased (**Figure 1B**). The 52-gene model (f52) could classify the cases into good and poor prognostic groups most accurately, with average 5-fold cross-validated accuracy (*ACC*), area under the receiver operating characteristic (ROC) curves (*AUC*), and Mathews Correlation Coefficient (*MCC*) of 0.81, 0.82, and 0.64, respectively (**Supplementary Table S4**). Cases classified by the model f52 showed significantly different overall survival (**Figure 1C**; Gehan–Breslow–Wilcoxon test, p = 3e-07).

To test whether the observed mutation–prognosis association was biased by tumor mutation load (TML), we compared the TML distribution between the cases with good and poor prognosis. However, neither total TML nor missense TML showed significant difference between the two groups of cases either classified by the median survival time or predicted by the f52 model (**Figure 1D** and **Supplementary Figure S2**; Wilcoxon rank-sum test, $p > 0.05$). Distribution analysis on clinical factors of the training cases demonstrated that clinical TNM stage could be a significant co-founding factor (**Supplementary Figure S3**). We developed a model featured by stage information, and found that its performance was far inferior to that of f52, despite its ability in classifying the cases into two groups with significantly different overall survival (**Supplementary Figure S4**, **Supplementary Table S4**, and **Figure 1E**). A model combined the 52 genetic features and stage information, f52s, but achieved better performance (**Supplementary Table S4**), which could classify the cases into two groups with more significant survival difference (**Figure 1F**; Gehan–Breslow–Wilcoxon test, p = 7e-08). The models further integrated with other clinical information-based features (e.g., anatomic site, f52sa), however, performed not better than f52s (**Figure 1G** and **Supplementary Table S4**).

MUC16 was recently reported to be associated with the prognosis of gastric cancer (Li et al., 2018). The gene was also included in our multi-gene feature list. We also found that the MUC16 prognosis-prediction model can classify the cases into two prognostic groups, but the significance was much lower than our multi-gene models (**Figure 1H**; Gehan–Breslow–Wilcoxon test, p = 0.03). Other performance measures further demonstrated the superiority of multi-gene models over the individual MUC16 model (**Supplementary Figure S5** and **Supplementary Table S4**).

Taking together, we identified a list of genes, whose somatic mutation profile could be used for effective prediction of prognosis for TCGA STAD cases.

Somatic Mutation Indicators for Poor Prognosis of TCGA STADs

We noticed that the f52 model showed lower classifying power for short-term prognosis of TCGA STAD cases (**Figure 1C**). All of the 52 genes were also found with higher mutation rates in

cases with good prognosis (**Supplementary Table S3**). Finally, the stratification for STAD prognosis was based on the median overall survival, and it would be also interesting to observe the dynamic changes of gene mutation rates between groups stratified with different cutoff survival times. To this end, we grouped the cases using overall survival of 1, 2, and 3 years as prognosis cutoff respectively besides the median and compared the gene mutation rates. Similar to the comparison results based on 573 days, an absolute majority of top significant genes showed higher mutation rates in the group of good prognosis than that of poor prognosis stratified by 2-year overall survival (**Figure 2A**; 3-year not shown due to the very limitation of case number for good prognosis). For 1-year stratification, however, the results demonstrated a contrary trend, i.e., most of the top significant genes showing higher mutation rates in the poor-prognosis group (**Figure 2A**). The consistent intersect between the top significant genes (50, 100, or 200) of 573-day and 2-year stratification was much larger than that between 573-day or 2-year and 1-year (**Figure 2B**).

To further explore the possible factors causing the observed contrary trends, we identified genes with the most strikingly different mutation rates (with a minimal difference of 10%) between poor and good prognostic groups stratified by 1 year, and observed the mutation rate changes along with overall survival time (**Figures 2C,D** and **Supplementary Table S5**). The results suggested that all of these (24) genes inclined to mutate in cases with the poorest prognosis (<1-year overall survival) (**Figures 2C,D**). As control, the genes showed no or much fewer

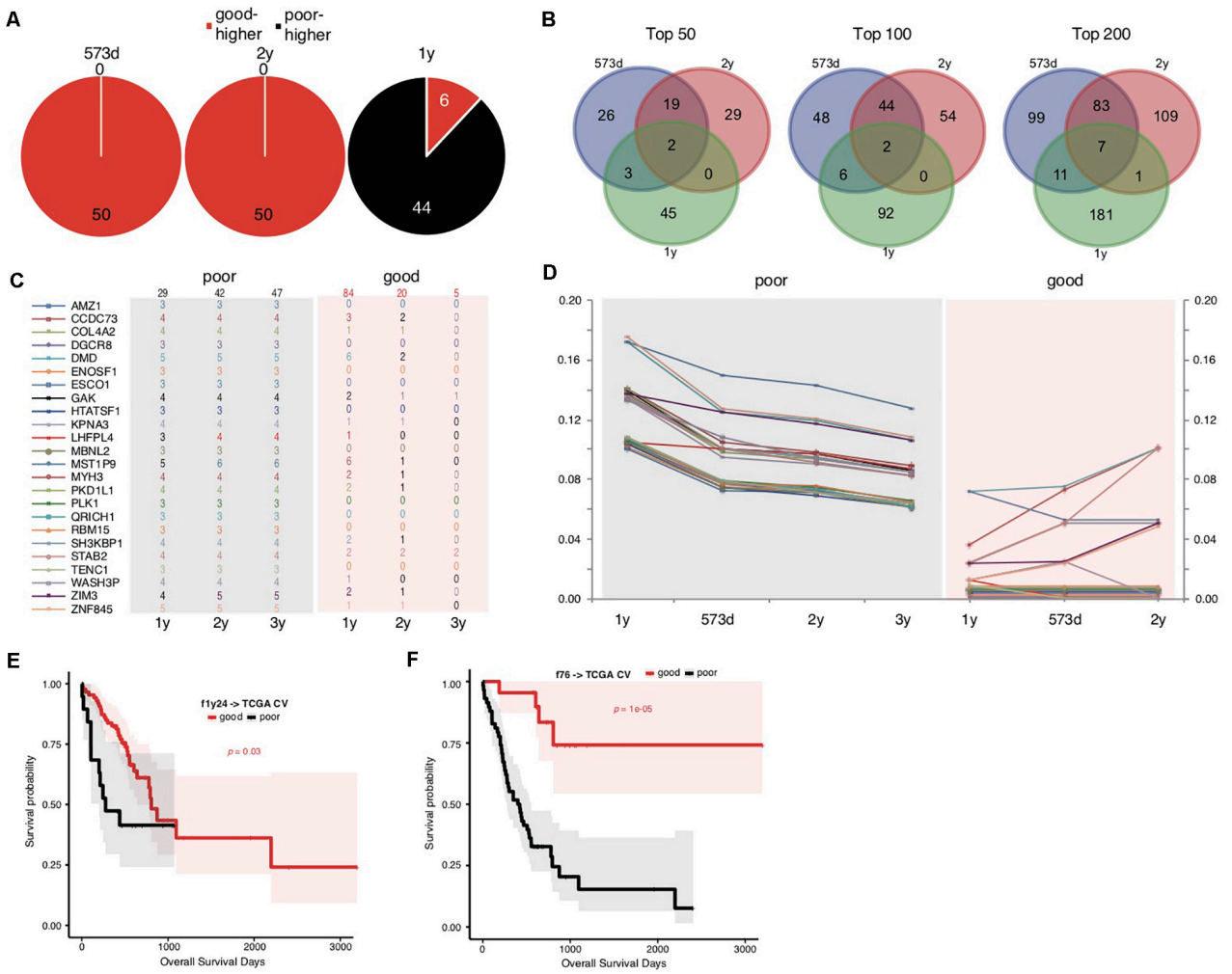


FIGURE 2 | Somatic mutation indicators for poor prognosis of TCGA STADs. **(A)** The distribution of top 50 significant genes of good and poor prognosis groups stratified with different cutoff survival time. Genes with higher mutation rates were counted for either the good or poor prognosis group and represented as "good-higher" or "poor-higher," respectively. The total number was also indicated. **(B)** The consistent intersect among the top significant genes (50, 100, or 200) identified by 573-day, 1-year, and 2-year stratifications. **(C,D)** The number of cases with mutation **(C)** and the mutation rate changes **(D)** of the top 24 significant genes selected in the good and poor prognostic groups stratified by 1 year along with overall survival time. The total case number for either prognostic group stratified by each survival cutoff was shown on the top of **(C)**. **(E,F)** The Kaplan–Meier overall survival curves of the TCGA STAD cases stratified by f1y24 **(E)** or the combined f76 model **(F)** with a 5-fold cross-validation strategy. Gehan–Breslow–Wilcoxon tests were performed to compare the overall survivals, and the *p*-values were shown in context.

mutations in cases with good prognosis, and the case number with mutations or mutation rates decreased generally for the patients with longer prognosis (**Figures 2C,D**).

The above results suggested that these gene mutations could be indicators for poorest prognosis. As validation, we used these genes as features and trained models based on 1-year-stratified TCGA training data (**Supplementary Table S4**). The 5-fold cross-validated results suggested that the optimized model (f1y24) could distinguish the cases with different prognoses in spite of a weaker distinguishing power (**Figure 2E**; Gehan–Breslow–Wilcoxon test, p = 0.03). Compared to f52, f1y24 did show better performance for the short-term prognosis classification (**Figure 2E**). Combination of the 24 short-term gene markers and 52 medium- and long-term markers generated a new model, f76, which showed a balanced classification power for both short-term and long-term prognosis classification, although the general significance was not comparable to f52 (**Supplementary Table S4** and **Figure 2F**; Gehan–Breslow–Wilcoxon test, p = 1e-05).

Heterogeneity of Prognosis-Associated Genetic Signatures Between TCGA and Chinese STAD Cohorts

The overall survival of the Chinese cohort with 78 STAD cases appeared better than the TCGA cohort, with a median of 1353 days (**Figure 3A**). We also stratified the Chinese cases into good and poor prognostic groups according to 1-, 2-, and 3-years, and median overall survival, respectively. Different from TCGA results, the top significant genes showed large consistence between 1-year and other survival time stratifications (**Figures 3B,C**).

To our surprise, the Chinese and TCGA cohorts showed an unexpected heterogeneity on the prognosis-associated gene mutation signatures. Very few common genes were identified in both cohorts with either higher or lower mutation rates in good prognostic groups (**Figure 3D**, upper; 4 with higher and 5 with lower mutation rates in good prognostic groups). More genes even showed the contrary trends in the TCGA and Chinese cohorts, e.g., higher mutation rates in the good prognostic group of TCGA cohort and the poor prognostic group of Chinese cohort (**Figure 3D**, lower; 48 genes). Further analysis for ethnicity stratification of the TCGA cases was precluded since the number of included Asian cases was too limited, and the secondary prognosis stratification and mutation rate comparison were infeasible.

This dramatic genetic heterogeneity could likely make the TCGA-based prognosis prediction models ineffective in application for the Chinese cohort. The application of the f52, f1y24, and f76 models confirmed the following assumption: none of them could well classify the Chinese cases into groups with

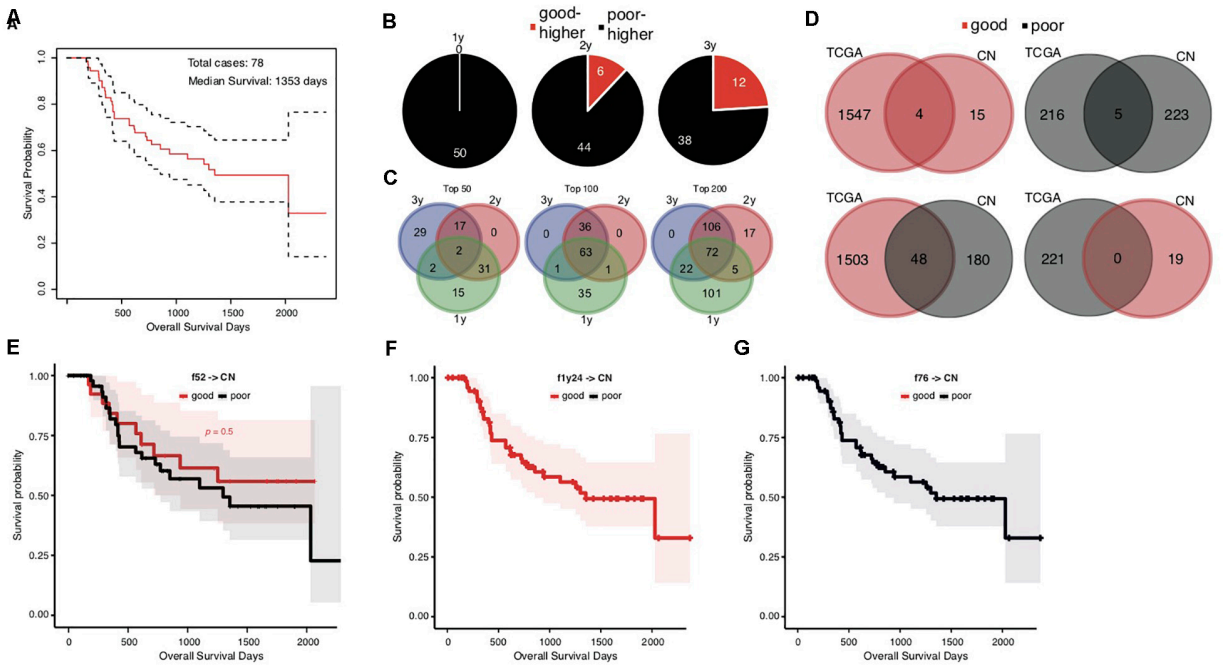


FIGURE 3 | Heterogeneity of prognosis-associated genetic signatures between TCGA and Chinese STAD cohorts. **(A)** The Kaplan–Meier overall survival curve of the Chinese STAD cases, with a median of 1353 days. **(B)** Distribution of the top 50 significant genes with different mutation rates between the good and poor prognostic groups stratified by different cutoff survival times. Genes with higher mutation rates were counted for either the good or poor prognosis group and represented as "good-higher" or "poor-higher," respectively. The total number was also indicated. **(C)** The consistent intersect among the top significant genes (50, 100, or 200) identified by 1-, 2-, and 3-years stratifications. **(D)** The consistent intersect between the prognosis-associated gene mutation signatures of TCGA and Chinese STAD cohorts. Genes were merged for the 1-year, 2-year, and 573-day stratifications for the TCGA cohort and merged for the 1-year, 2-year, and 3-year stratifications for the Chinese cohort in the first place. **(E–G)** The Kaplan–Meier overall survival curves of the Chinese STAD cases classified with different models built on the TCGA training data.

different prognosis (**Figures 3E–G**; Gehan–Breslow–Wilcoxon test for f52, p = 0.5; f1y24 and f76 classifying all Chinese cases as good and poor prognosis, respectively).

Common Signatures Effectively Predict STAD Prognosis of Both TCGA and Chinese Cohorts

To overcome the generalization drawbacks of the prognosis prediction models based on individual cohorts due to the genetic heterogeneity, we came up with a new strategy to identify and test a list of new signatures by screening the genes with the same change trend of somatic mutation rates between prognostic groups in the TCGA and Chinese cohorts. Genes were extracted with different levels of mutation rate difference (≥15%, ≥10%, and ≥5%) between prognostic groups for both cohorts stratified, respectively, and the common ones were further identified correspondingly to serve as signatures. To reduce the biases caused by imbalanced sample size between groups, the prognostic groups were stratified by an overall survival period of 576 days for the TCGA cohort and 3 years for the Chinese cohort, respectively, with which the two groups in either cohort showed the identical sample size. There were 0, 4 (*MUC16*, *ATP10A*, *MPDZ*, and *VPS13A*) and 20 genes showing ≥15%, ≥10%, and ≥5% mutation rate differences between prognostic groups for both cohorts with the same direction (**Supplementary Table S6**). Furthermore, the 20 genes (with ≥ 5% mutation rate difference) were tested for the prognosis prediction performance as signatures. With these common feature genes, we trained a prognosis prediction model (cf20) based on the TCGA training data stratified with the median survival time. The 5-fold cross-validation performance on TCGA data was not comparable to f52, however, it remained to be effective in classifying the data into two prognostic groups (**Figure 4A**; Gehan–Breslow–Wilcoxon test, p = 0.008). The model appeared much more effective in prognosis prediction of the Chinese cohort (**Figure 4B**; Gehan–Breslow–Wilcoxon test, p = 4e-06). It consistently showed good performance to predict the different stratifications of prognosis for the Chinese cohort, especially for 3- and 2-year prognosis (**Figure 4C**).

We also integrated the TNM staging information in cf20 to generate a new model, cf20s. For TCGA data and based on the 5-fold cross-validation evaluation, cf20s apparently outperformed cf20 (**Figure 4D**; Gehan–Breslow–Wilcoxon test for cf20s, p = 4e-06). When testing in the Chinese cohort, however, the performance deteriorated, in spite that it remained effective to predict the prognosis (**Figure 4E**; Gehan–Breslow–Wilcoxon test, p = 4e-06). Both cf20 and cf20s, however, outperformed the single-gene models in predicting the prognosis of the STAD cases (**Figure 4F**).

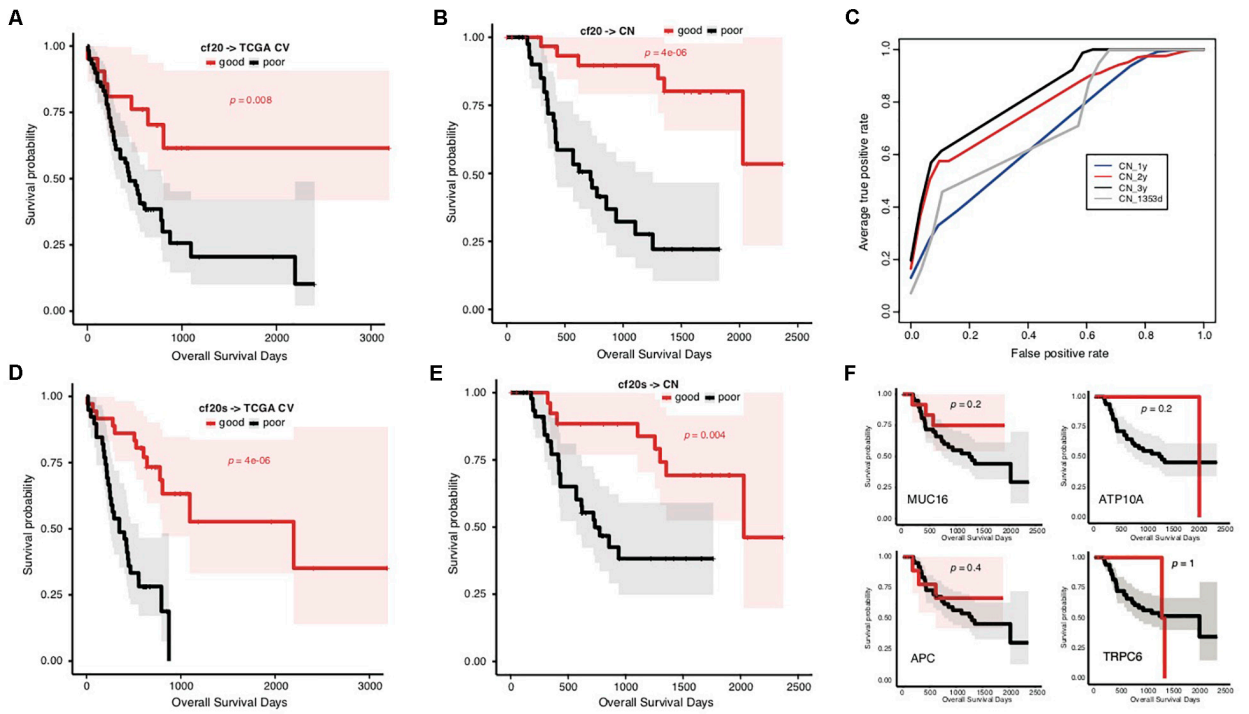


FIGURE 4 | Prediction of STAD prognosis with the genetic models based on 20 common somatic mutation features. **(A)** The Kaplan–Meier overall survival curves of the TCGA STAD cases classified by cf20 with a 5-fold cross-validation strategy. **(B)** The Kaplan–Meier overall survival curves of the Chinese STAD cases stratified by the cf20 model. **(C)** The 5-fold cross-validated ROC curves of cf20 model predicting the prognosis of Chinese STAD cohorts stratified by different cutoff survival time. **(D)** The Kaplan–Meier overall survival curves of the TCGA STAD cases classified by the cf20s model which was combined 20 common-gene features and TNM stage information. **(E)** The Kaplan–Meier overall survival curves of the Chinese STAD cases stratified by the combined cf20s model. **(F)** The Kaplan–Meier overall survival curves of the Chinese STAD cases stratified by models based on example individual genes, including MUC16, ATP10A, APC, and TRPC6.

DISCUSSION

In this research, we found the association between overall prognosis of STADs and somatic gene mutations from the TCGA cohort. Despite that the rate comparison-based feature extraction strategy involved division of the cases into different prognostic groups according to a survival cutoff preset subjectively, the model (f52, median survival as cutoff) could well classify the cases into two groups with significantly differentiated survival (**Figures 1B,C**). It is noteworthy that 5-fold cross validation was used for assessment of model performance, independent between each training subset and testing subset, and the survival comparison was performed between the predicted groups for all testing cases. Therefore, the observed association was not biased by the model-training scheme. Except tumor stage, other possible co-founding clinical factors, including sex, anatomic site, and histopathology, did not show a biased distribution between the prognostic groups. The TNM staging system could predict STAD prognosis independently but was not comparable to the f52 genetic model (**Figure 1E** and **Supplementary Figure S4**). Combination of the genetic features and stage information improved the prognosis-classifying performance significantly (**Figure 1F**). Therefore, as for the TCGA cohort, the prognosis is associated with genetic factors.

It was noted that all the 52 genes with most significant difference showed higher mutation rates in the good prognosis group of TCGA cases stratified by the median survival time (**Figure 2A**). It was consistent with previous findings in lung adenocarcinomas (Yu et al., 2019). Recently, a study identified the association between higher MUC16 gene mutation rate and better prognosis of STADs. Meanwhile, the more frequent MUC16 mutation was associated with a higher TML (Li et al., 2018). Maruvka et al. argued that a larger MUC16 mutation frequency could only be an accompanying result of high TML (Maruvka et al., 2019). MUC16 was also present in our 52-gene list. We suspected that the list of signature genes with different mutation rates in prognostic groups could be merely caused by different TMLs. However, no significant difference was detected for either TMLs or missense TMLs between the prognosis groups of the TCGA data (**Figure 1D** and **Supplementary Figure S2**). Interestingly, we noticed that, for TML or missense TML, although there was no difference in the medians or lower quartile, the good prognosis group always showed a larger upper quartile (**Figure 1D**). Therefore, a higher TML could be an important but not unique factor predicting better prognosis. The identified signatures could partially represent TML difference and also represent other unknown mechanisms influencing the prognosis of STAD.

With the analytic strategy in this study, we also got an interesting finding that the composition of genes and the direction of mutation rate difference between groups stratified by 1-year survival were totally different from those identified for median (573-day) or 2-year stratification (**Figures 2A,B**). The latter two stratifications showed larger consistency between each other (**Figures 2A,B**). A list of genes was identified with different mutation rates between prognostic groups stratified by 1 year, which showed more frequent mutations in the group of poor prognosis (**Figure 2C**). These genes tend to mutate in cases of poorest prognosis (**Figures 2C,D**), unlike those identified in median (or 2-year) stratification for which the mutation rate showed a linear correlation with overall survival period generally. The model trained with the 1-year gene features could only distinguish the cases with poorest prognosis (**Figure 2E**), further demonstrating that the mutations of these signatures could be the indicators of very poor prognosis of STAD.

Heterogeneity of STADs and their genetics was not surprisingly identified between cohorts, and yet the dramatic difference of prognosis-associated genetic signatures between the TCGA and Chinese cohort was unexpected (**Figure 3D**). Direct application of the signatures and models trained in the TCGA cohort showed an awful performance in prognosis prediction of the Chinese cohort (**Figures 3E–G**). There was a large heterogeneity of genes with mutation rate difference identified from the two cohorts. Many genes even showed a contrary trend for the mutation rate in prognostic groups (**Figure 3D**). We attempted to isolate the Asian cases from the TCGA cohort but failed to evaluate the gene mutation rates within different prognosis groups due to the very limited number of the cases. It remains to be clarified whether the heterogeneity between cohorts is related with ethnicity of STAD cases. Two prognostic biomarker genes, BRCA2 and MUC16 (Chen et al., 2015; Li et al., 2018), were found with a mutation rate difference between the good and poor prognostic groups, and with the same trend in the two cohorts. We modified the signature-identification strategy, with an attempt to find out all the common genes with a consistent mutation difference between prognostic groups within each cohort. In total, 20 genes were identified, including MUC16 and BRCA2. A model (cf20) was trained with these genes as features and the TCGA cohort as training data. The model well predicted the prognosis of both TCGA cases based on a cross-validation evaluation, and the Chinese cases independently (**Figures 4A–C**). The multi-gene model also outperformed the ones based on individual genes strikingly (**Figure 4F**). However, the problems of over-fitting cannot be totally excluded despite of the use of only TCGA data for model training and cross-validation evaluation, and the Chinese cohort as an independent validation dataset, because the signatures were identified using both the cohorts. The effective sample sizes for the cohorts (especially the Chinese cohort) were too small so that they were hardly further divided, and therefore resampling or cross-validation-based feature identification strategies appeared difficult. It would be better to, but currently we cannot, find one or more independent STAD datasets (with both gene mutation profiling data and clinical follow-up information) to make further assessment. New larger datasets are also in need to further evaluate the potential heterogeneity caused by human ethnicity and develop more ethnicity-specific models like the f52 for the TCGA population.

Besides somatic mutation signatures, germline variants could also be associated with tumor prognosis. Recently, Milanese et al. reported different germline variants in recurred and non-recurred patients of breast cancers (Milanese et al., 2019). These signature germline variants could potentially facilitate the formation of the pro-tumorigenic environment by impairing adaptive and innate immune pathways and could be used

for prediction of breast cancer outcomes (Milanese et al., 2019). In another study, Xu et al. (2019) observed negative associations between the number of germline defective genes in natural killer cells and survival time in a variety of cancer types. It is interesting to understand whether there is also heterogeneity between different cohorts for the associations between germline mutations and STAD prognosis. Combination of both germline variants and somatic mutations as well as other signatures, e.g., hypermethylation signatures, and RNA markers, could also further improve the model prediction performance on STAD prognosis.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

YW, XC, and DL conceived the project. XL, XH, and YW coordinated the project. XH, HK, QF, and AC collected the data. XL, XH, HK, YH, and YW performed the analysis. XH, HK, and YW developed the models. XL, XH, HK, and YW wrote the first draft. XL, XH, HK, DL, XC, and YW revised the manuscript. All authors approved the final version of manuscript.

FUNDING

The study was supported by Natural Science Funding of Shenzhen (JCYJ20190808165205582 and JCYJ201607115221141) and a Shenzhen Peacock Innovation Team Project Grant KQTD20140630100658078.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00940/full#supplementary-material

FIGURE S1 | The general prognosis of TCGA STAD cases and the functional enrichment analysis of the prognosis-associated genes. **(A)** The good and poor prognostic groups of STAD cases within different survival periods, including 1, 2, and 3 years and the median of 573 days. **(B)** Gene Ontology (GO) enrichment analysis of the top 52 genes with significant mutation rate difference between the prognostic groups stratified by the median survival time.

FIGURE S2 | TML distribution for different prognostic groups of the TCGA cases. The distribution of TMLs **(A)** and missense TMLs **(B,C)** for TCGA cases of good and poor prognosis groups for the raw TCGA training dataset stratified by 573-day survival or classified by the f52 model. The *p*-value of Wilcoxon rank-sum test was indicated.

FIGURE S3 | Distribution analysis on clinical factors of the training cases. Sex **(A)**, anatomic regions of stomach **(B)**, clinical TNM stage **(C)**, and the two main histological types of gastric carcinoma **(D)** were involved. Chi-square tests were performed and the *p*-values were indicated.

FIGURE S4 | Performance comparison of the prognosis prediction models based on 52 somatic mutation features and clinical TNM stage information. Specificity (Sp), Sensitivity (Sn), Accuracy (ACC), and Mathews Correlation Coefficient (MCC) were utilized to assess the predictive performance. The model f52 was based on the 5-fold cross-validation results. Pairwise one-tail Student's *t*-tests were performed, and the *p*-values were indicated.

FIGURE S5 | Performance comparison of the prognosis prediction models based on 52 somatic mutation features and MUC16. Specificity (Sp), Sensitivity (Sn), Accuracy (ACC), and Mathews Correlation Coefficient (MCC) were utilized to assess the predictive performance. The model f52 was based on the 5-fold cross-validation results. Pairwise one-tail Student's *t*-tests were performed, and the *p*-values were indicated.

REFERENCES

Banks, M., Graham, D., Jansen, M., Gotoda, T., Coda, S., di Pietro, M., et al. (2019). British Society of Gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. *Gut* 68, 1545–1575. doi: 10.1136/gutjnl-2018-318126

Bosman, F. T., Carneiro, F., Hruban, R. H., and Theise, N. D. (2010). *WHO Classification of Tumours of the Digestive System*, 4th Edn. Geneva: World Health Organization.

Boussioutas, A., Li, H., Liu, J., Waring, P., Lade, S., Holloway, A. J., et al. (2003). Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer. *Cancer Res.* 63, 2569–2577.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 cof untries. *CA Cancer J Clin.* 68, 394–424. doi: 10.3322/caac.21492

Cancer Genome, Atlas Research, and Network. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202–209. doi: 10.1038/nature13480

Charalampakis, N., Economopoulou, P., Kotsantis, I., Tolia, M., Schizas, D., Liakakos, T., et al. (2018). Medical management of gastric cancer: a 2017 update. *Cancer Med.* 7, 123–133. doi: 10.1002/cam4.1274

Chen, K., Yang, D., Li, X., Sun, B., Song, F., Cao, W., et al. (2015). Mutational landscape of gastric adenocarcinoma in Chinese: implications for prognosis and therapy. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1107–1112. doi: 10.1073/pnas.1422640112

Cho, H. J., Lee, S., Ji, Y. G., and Lee, D. H. (2018). Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. *PLoS One* 13:e0207204. doi: 10.1371/journal.pone.0207204

Choi, E., Roland, J. T., Barlow, B. J., O'Neal, R., Rich, A. E., Nam, K. T., et al. (2014). Cell lineage distribution atlas of the human stomach reveals heterogeneous gland populations in the gastric antrum. *Gut* 63, 1711–1720. doi: 10.1136/gutjnl-2013-305964

Cristescu, R., Lee, J., Nebozhyn, M., Kim, K. M., Ting, J. C., Wong, S. S., et al. (2015). Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat. Med.* 21, 449–456.

Hui, X., Hu, Y., Sun, M. A., Shu, X., Han, R., Ge, Q., et al. (2017). EBT: a statistic test identifying moderate size of significant features with balanced power and precision for genome-wide rate comparisons. *Bioinformatics* 33, 2631–2641. doi: 10.1093/bioinformatics/btx294

Laurén, P. (1965). The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. *Acta Pathol. Microbiol. Scand.* 64, 31–49. doi: 10.1111/apm.1965.64.1.31

Lee, D. W., Han, S. W., Cha, Y., Bae, J. M., Kim, H. P., Lyu, J., et al. (2017). Association between mutations of critical pathway genes and survival outcomes according to the tumor location in colorectal cancer. *Cancer* 123, 3513–3523. doi: 10.1002/cncr.30760

Lei, Z., Tan, I. B., Das, K., Deng, N., Zouridis, H., Pattison, S., et al. (2013). Identification of molecular subtypes of gastric cancer with different responses

to PI3-kinase inhibitors and 5-fluorouracil. *Gastroenterology* 145, 554–565. doi: 10.1053/j.gastro.2013.05.010

Li, X., Pasche, B., Zhang, W., and Chen, K. (2018). Association of MUC16 mutation with tumor mutation load and outcomes in patients with gastric Cancer. *JAMA Oncol.* 4, 1691–1698.

Loi, S., Michiels, S., Lambrechts, D., Fumagalli, D., Claes, B., Kellokumpu-Lehtinen, P. L., et al. (2013). Somatic mutation profiling and associations with prognosis and trastuzumab benefit in early breast cancer. *J. Natl. Cancer Inst.* 105, 960–967. doi: 10.1093/jnci/djt121

Maruvka, Y. E., Haradhvala, N. J., and Getz, G. (2019). Analyzing frequently mutated genes and the association with tumor mutation load. *JAMA Oncol.* 5:577. doi: 10.1001/jamaoncol.2019.0127

Milanese, J. S., Tibiche, C., Zou, J., Meng, Z., Nantel, A., Drouin, S., et al. (2019). Germline variants associated with leukocyte genes predict tumor recurrence in breast cancer patients. *NPJ Precis Oncol.* 3:28.

Ni, X., Tan, Z., Ding, C., Zhang, C., Song, L., Yang, S., et al. (2019). A region-resolved mucosa proteome of the human stomach. *Nat. Commun.* 10:39.

Oh, S. C., Sohn, B. H., Cheong, J. H., Kim, S. B., Lee, J. E., Park, K. C., et al. (2018). Clinical and genomic landscape of gastric Cancer with a mesenchymal phenotype. *Nat. Commun.* 9:1777.

Ooi, W. F., Xing, M., Xu, C., Yao, X., Ramlee, M. K., Lim, M. C., et al. (2016). Epigenomic profiling of primary gastric adenocarcinoma reveals super-enhancer heterogeneity. *Nat. Commun.* 7:12983.

Shen, L., Shan, Y. S., Hu, H. M., Price, T. J., Sirohi, B., Yeh, K. H., et al. (2013). Management of gastric cancer in asia: resource-stratified guidelines. *Lancet Oncol.* 14:e535-47.

Soybel, D. I. (2005). Anatomy and physiology of the stomach. *Surg. Clin. North Am.* 85, 875–894. doi: 10.1016/j.suc.2005.05.009

Tan, I. B., Ivanova, T., Lim, K. H., Ong, C. W., Deng, N., Lee, J., et al. (2011). Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy. *Gastroenterology* 141, 476–485.

Tran, P. N., Sarkissian, S., Chao, J., and Klempner, S. J. (2017). PD-1 and PD-L1 as emerging therapeutic targets in gastric cancer: current evidence. *Gastrointest. Cancer* 7, 1–11. doi: 10.2147/gictt.s113525

Waldum, H. L., and Fossmark, R. (2018). Types of gastric carcinomas. *Int. J. Mol. Sci.* 19:4109. doi: 10.3390/ijms19124109

Xu, X., Li, J., Zou, J., Feng, X., Zhang, C., Zheng, R., et al. (2019). Association of germline variants in natural killer cells with tumor immune microenvironment subtypes, tumor-infiltrating lymphocytes, immunotherapy response, clinical outcomes, and Cancer risk. *JAMA Netw. Open* 2:e199292. doi: 10.1001/jamanetworkopen.2019.9292

Yu, J., Hu, Y., Xu, Y., Wang, J., Kuang, J., Zhang, W., et al. (2019). LUADpp: an effective prediction model on prognosis of lung adenocarcinomas based on somatic mutational features. *BMC Cancer* 19:263. doi: 10.1186/s12885-019-5433-7

Zhang, P., Yang, M., Zhang, Y., Xiao, S., Lai, X., Tan, A., et al. (2019). Dissecting the Single-Cell transcriptome network underlying gastric premalignant lesions and early gastric Cancer. *Cell Rep.* 27, 1934–1947.

Zhang, S., Xu, Y., Hui, X., Yang, F., Hu, Y., Shao, J., et al. (2017). Improvement in prediction of prostate cancer prognosis with somatic mutational signatures. *J. Cancer* 8, 3261–3267. doi: 10.7150/jca.21261

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 August 2020
doi: 10.3389/fgene.2020.00934

Integrated Analysis of a Risk Score System Predicting Prognosis and a ceRNA Network for Differentially Expressed lncRNAs in Multiple Myeloma

Sijie Zhou[1†], Jiuyuan Fang[2†], Yan Sun[1] and Huixiang Li[1,2]*

[1] The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [2] School of Basic Medical Sciences, Zhengzhou University, Zhengzhou, China

OPEN ACCESS

Edited by:
Xin Maizie Zhou,
Stanford University, United States

Reviewed by:
Cheng Liang,
Shandong Normal University, China
Jie Sun,
Wenzhou Medical University, China

***Correspondence:**
Huixiang Li
lihuixiang19@sina.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *19 April 2020*
Accepted: *27 July 2020*
Published: *27 August 2020*

Citation:
Zhou S, Fang J, Sun Y and Li H (2020) Integrated Analysis of a Risk Score System Predicting Prognosis and a ceRNA Network for Differentially Expressed lncRNAs in Multiple Myeloma. Front. Genet. 11:934. doi: 10.3389/fgene.2020.00934

Long non-coding RNAs (lncRNAs) are non-protein-coding RNAs longer than 200 nucleotides. Accumulating evidence demonstrates that lncRNA is a potential biomarker for cancer diagnosis and prognosis. However, there are no prognostic biomarkers and lncRNA models for multiple myeloma (MM). Hence, it is necessary to screen novel lncRNA that can potentially participate in the initiation and progression of MM and consequently construct a risk score system for the disease. Raw microarray datasets were obtained from the Gene Expression Omnibus website. Weighted gene co-expression network analysis and principal component analysis identified 12 lncRNAs of interest. Then, univariate, least absolute shrinkage and selection operator Cox regression and multivariate Cox hazard regression analysis identified two lncRNAs (LINC00996 and LINC00525) that were formulated to construct a risk score system to predict survival. Receiver operating characteristic analysis certificated the superior performance in predicting 3-year overall survival (area under the curve = 0.829). The similar prognostic values of the two-lncRNA signature were also observed in the tested The Cancer Genome Atlas dataset. Furthermore, two other lncRNAs (LINC00324 and LINC01128) were differentially expressed between CD138+ plasma cells from normal donors and MM patients and were verified to be associated with cancer stage in the Gene Expression Omnibus dataset. A lncRNA-mediated competing endogenous RNA network, including 2 lncRNAs, 12 mitochondrial RNAs, and 103 target messenger RNAs, was constructed. In conclusion, we developed a two-lncRNA expression signature to predict the prognosis of MM and constructed a key lncRNA-based competing endogenous RNA network in MM. These lncRNAs were associated with survival and are probably involved in the occurrence and progression of MM.

Keywords: long non-coding RNA, biomarkers, multiple myeloma, weighted gene co-expression network analysis, principal component analysis, competing endogenous RNA network, prognostic long non-coding RNA expression signature

INTRODUCTION

Multiple myeloma (MM) is the second most common hematological malignancy. It is caused by the clonal proliferation of malignant plasma cells in the bone marrow (BM) (Laubach et al., 2011). MM is characterized by renal impairment, lytic bony lesions, anemia, and bone pain. The survival of MM patients ranges from a few weeks to more than 10 years (Decaux et al., 2008; Chen W. C. et al., 2017; Cowan et al., 2018).

As a newly discovered type of non-coding RNA, long non-coding RNAs (lncRNAs) function as imperative regulators involved in tumorigenesis, tumor suppression (Poliseno et al., 2010; Hung and Chang, 2010), and many biological processes (Geisler and Coller, 2013; Fatica and Bozzoni, 2014). Many lncRNAs involved in the initiation and progression of MM have been identified. Furthermore, lncRNAs can also regulate gene expression by interacting with mitochondrial RNA (miRNA) at miRNA-binding sites (MREs). For example, MALAT1 is an lncRNA that inhibits the proliferation and adhesion of myeloma cells by upregulating the expression of miR-181a-5p (Sun et al., 2019a). The aberrant expression of urothelial cancer associated 1 lncRNA affords it the ability to promote proliferation and inhibits apoptosis by regulating miR-1271-5p and hepatocyte growth factor in MM cells (Yang and Chen, 2019). Abnormally expressed lncRNA NR_046683 in patients of different MM subtypes and stages indicated that it could be used as a new indicator for potential drug target and prognosis (Dong et al., 2019). Although several lncRNA prognostic models have been identified in uterine corpus endometrial carcinoma (Ouyang et al., 2019), hepatocellular carcinoma (Sun et al., 2019b), cervical cancer (Wu et al., 2019), and lung adenocarcinoma (Zhou et al., 2019), the clinical implication of most lncRNAs in MM remains unclear.

Weighted gene co-expression network analysis (WGCNA) is an algorithm that is frequently used to cluster highly synergistically altered gene sets into separate modules. This can establish connections with clinical traits and thus screen out candidate indicator genes or therapeutic targets (Langfelder and Horvath, 2008; Shi et al., 2010). Principal component analysis (PCA) is another mathematical algorithm. It is a powerful technique that is widely applied in bioinformatics and other fields. It can reduce the dimensionality of the data while retaining most of the variations that are uncorrelated in the data set. These unrelated variables are called principal components (PCs) (Ringner, 2008). After identifying new variables, the PCs, with a sample-like pattern and a weight for each gene, further exploration can be done by building a link with clinical data, and candidate genes can be obtained by comparing component loadings. In the present study, the Gene Expression Omnibus (GEO) public integrated database provided an application platform of genomic sequencing data along with the clinical information of each MM patient. WGCNA and PCA were performed to explore public sequencing data and clinical information of MM patients.

A few key gene modules associated with tumor stage and PCs correlated with risk score and proliferation index were identified, and 12 lncRNAs in the intersection were identified. We found a two-lncRNA signature that might act as an independent prognostic factor to identify MM patients that are at higher risk of poor clinical outcome. Furthermore, using other datasets, we recognized database of essential genes (DEG) and constructed a competing endogenous RNA (ceRNA) network in MM based on two of the 12 abnormally expressed lncRNAs. These two lncRNAs may participate in tumorigenesis or serve as clinical indicators of the progression of MM.

RESULTS

Weighted Gene Co-expression Network Analysis Identification of Clinically Significant Modules

A total of 32 MM samples with a known stage of cancer were utilized to conduct the hierarchical clustering analysis using the WGCNA package. The sample dendrogram and clinical trait heatmap of GSE16791 is displayed in **Figure 1A**. No obvious outlier was evident in the sample clustering. The information of two clinical traits of 32 MM samples, including age and cancer stage, is presented in **Figure 1A**. Selecting the best soft-thresholding powers is imperative to obtain relatively balanced scale independence and mean connectivity. As presented in **Supplementary Figure S2A**, we selected $\beta = 8$ (scale-free $R^2 = 0.81$) as a soft-threshold to construct a scale-free network, and a total of 21 modules were detected (**Figure 1B**). As the overall gene expression level of the corresponding module, the module eigengenes were calculated to assess the relationship between modules and clinical information by Pearson's correlation analysis. The results indicated that the stage was negatively associated with blue and green modules (**Figure 1C**). Scatterplots of gene significance of stage vs. module membership in the blue and green modules revealed that they were highly correlated (**Supplementary Figure S2B**). Also, we calculated eigengenes of all modules and clustered them on the base of their correlations. A module eigengenes dendrogram indicated that the blue and green modules were clustered together, and the eigengene network heatmap revealed similar results (cor = 0.65, P = 5e-05; **Supplementary Figure S2C**). Therefore, we chose blue and green modules for further analysis.

Principal Component Analysis Determination of Interesting Principal Components Associated With Clinical Traits

Principal component analysis was performed on the 52 samples in GSE17306. In this dataset, the gene expression profiling (GEP)-risk score and proliferation index of each sample were calculated according to the GEP (Zhou et al., 2010). Initially, PCA created 52 composite variables (PCs) by reducing the dimensionality of numerous genes. The first 33 components, which explained 80% of the variability among the 52 samples, were retained to correlate clinical traits (**Figure 2A**). These 33 composite variables are enough to explain the sample differences to the greatest extent. Next, to ascertain the capability of PCs to differentiate risk score level and proliferation index level, the pairs plot was conducted

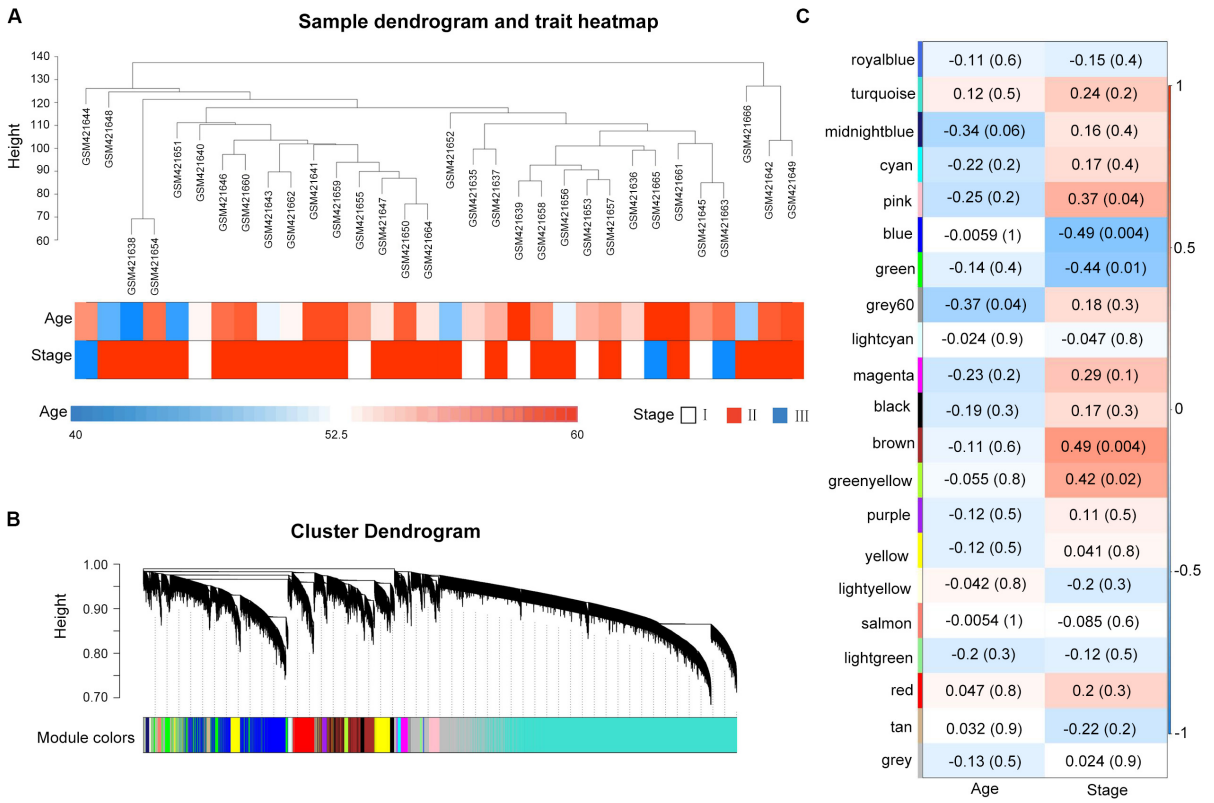


FIGURE 1 | Weighted gene co-expression network of multiple myeloma and module–trait relationships. **(A)** Cluster analysis of samples and clinical traits. All the samples were in the clusters. **(B)** Gene dendrogram obtained by clustering all genes from GSE16791. Each branch in the figure corresponds to one gene, and each color to one co-expression gene module. **(C)** Module clinical associations. Each row represents a module eigengene, and each column represents a clinical trait. Each cell contains the corresponding correlation coefficient and the *P*-value. The blue and green modules were significantly correlated with a stage.

to compare PC1 with PC8 on a pairwise basis (**Figure 2B**). Additionally, a bi-plot of PC1 versus PC6 indicated that PC6 could roughly distinguish the high-risk group from the low-risk group (**Figure 2D**). Next, we correlated the PCs back to the clinical data, including the GEP-risk score and proliferation index, to identify interesting PCs. PC6 and PC8 were negatively associated with risk score and proliferation index in all the 33 PCs retained (**Figure 2C**). PC11 and PC12 were positively correlated with the proliferation index. For each PC of interest, "plotloadings" determined the genes ranked in the top 20 of the loadings range and then created a final consensus list of these (**Figure 2E**).

Construction a Risk Assessment Model

To construct a lncRNA scoring system that is predictive of survival in the MM patients, we extracted lncRNAs from the blue and green modules and PC6 and PC8 based on the Genecode annotation[1]. Finally, a total of 12 lncRNAs were obtained from the intersection of the interesting modules and PCs (**Figure 3A**). The expression levels of 12 lncRNAs were extracted from GSE57317 to conduct the univariate Cox regression analysis. The results of the univariate Cox analysis of 12 prognostic lncRNAs from the discovery cohort are shown in **Table 1**. After this, six significant lncRNAs ($P < 0.05$) were identified and were included in the least absolute shrinkage and selection operator (LASSO) model; cross-validation was adopted to select the penalty parameters (**Figures 3B,C**). Two lncRNAs were identified based on lambda.1se values (**Supplementary Table S1**). The quantitative real-time polymerase chain reaction (qRT-PCR) results showed that the expression of LINC00525 was significantly downregulated in Roswell Park Memorial Institute (RPMI)-8226 and KM3 cell lines, whereas LINC00996 was significantly upregulated in KM3 cell line compared with normal plasma cells (**Supplementary Figures S3A,B**). We further included expression levels of the two lncRNAs in a multivariate Cox model. The risk score = (−0.3647) × (expression value of LINC00996) + (−0.4266) × (expression value of LINC00525). The details of the two lncRNAs are depicted in **Figure 4B**. We used the median of the risk score as the cutoff to define the groups of MM patients with high and low scores (**Figure 4A**). The survival time and overall survival (OS) status in the training dataset are presented in the middle panel of **Figure 4A**. Compared with those in the low-risk score group, patients in the high-risk score group displayed an obviously worse OS (**Figure 4C**). The 3-year survival receiver operating characteristic (ROC) curve was also plotted. The area under the curve of the

[1]https://www.gencodegenes.org/

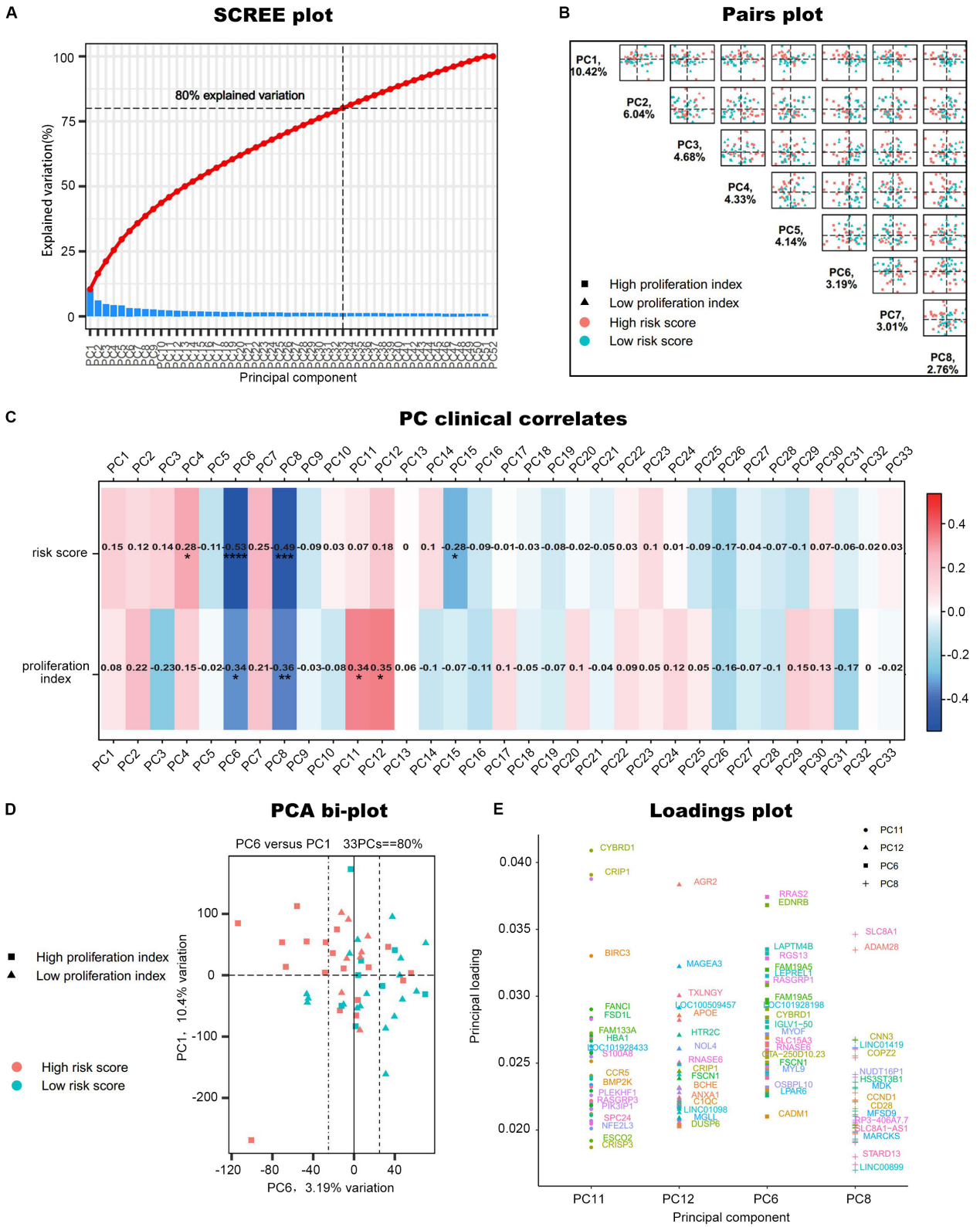


FIGURE 2 | PCA of GSE17306. **(A)** PCs accounted for 80% of the explained variation in the dataset, and the first 33 PCs were responsible for the same. **(B)** A plot comparing PC1–PC8 on a pairwise basis. PC1 is usually the most important part of PCA. **(C)** Correlation of the principal components (PCs) to the clinical data. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$. **(D)** A bi-plot of PC1 versus PC6. **(E)** Determine the variables that drive variation among each PC. Components have a sample-like pattern with a weight called component loading for each gene. Genes ranked the top 20 of the loadings range were presented.

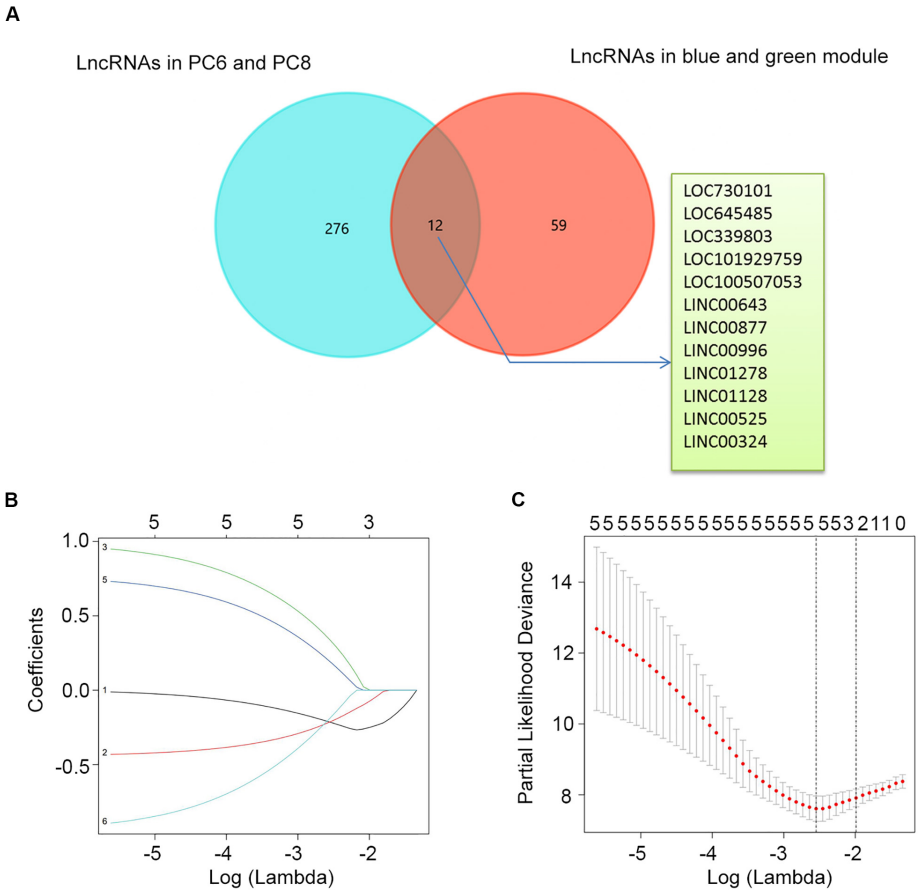


FIGURE 3 | Determination of candidate lncRNAs and LASSO regression analysis. **(A)** Venn diagram of candidate lncRNAs in blue and green modules and PC6 and PC8. **(B)** LASSO coefficient profiles of the six candidate lncRNAs. **(C)** Ten-fold cross-validation used to tune parameter selection in the LASSO model. A vertical line is drawn at the value chosen by 10-fold cross-validation.

risk score reached 0.829 (**Figure 4D**), revealing that the risk score based on the two lncRNAs is a good indicator of prognosis. The results of univariate and multivariate Cox regression analyses indicated that the risk score ($P < 0.001$ and $P = 0.006$) was an independent prognostic indicator (**Supplementary Table S2**). To further examine the accuracy of the lncRNA risk score model developed in the training dataset, the performance of the risk score was also evaluated in The Cancer Genome Atlas (TCGA) dataset. The result of multivariate Cox regression analysis for the expression level of two lncRNAs in the TCGA dataset is presented in **Supplementary Figure S4B**. The risk survival status, score distribution, and expression pattern of the two lncRNAs in the 787 MM patients in the TCGA dataset are displayed in **Supplementary Figure S4A**. Also, corresponding to our previous conclusion, the OS was significantly shorter in the high-risk group compared with that in the low-risk group (**Supplementary Figure S4C**), and the AUC of the risk score reached 0.584 (**Supplementary Figure S4D**). Univariate Cox regression analyses were conducted to detect various factors correlated with prognosis. The results revealed that age ($P = 0.009$), tumor stage ($P < 0.001$), and risk score ($P = 0.002$) were significantly associated with the OS of the MM patients. A subsequent multivariate Cox regression analysis indicated that the tumor stage ($P < 0.001$) and risk score ($P = 0.001$) were independent prognostic indicators (**Supplementary Table S3**).

TABLE 1 | Univariate Cox analysis of 12 prognostic lncRNAs from the discovery cohort.

lncRNA name	Type	HR	*P*
LINC01128	Bad	5.324513	0.015396
LOC339803	Bad	2.530873	0.031322
LOC100507053	Bad	1.474999	0.259866
LINC01278	Bad	1.437495	0.173067
LINC00643	Bad	1.097326	0.578361
LINC00877	Good	0.876417	0.512653
LOC645485	Good	0.85804	0.474318
LINC00996	Good	0.602728	7.11E-05
LINC00525	Good	0.575173	0.001695
LOC101929759	Good	0.464721	0.040177
LINC00324	Good	0.382086	0.004307

HR, hazard ratio. Type represents bad survival lncRNAs and good survival lncRNAs. All statistical tests were two-sided.

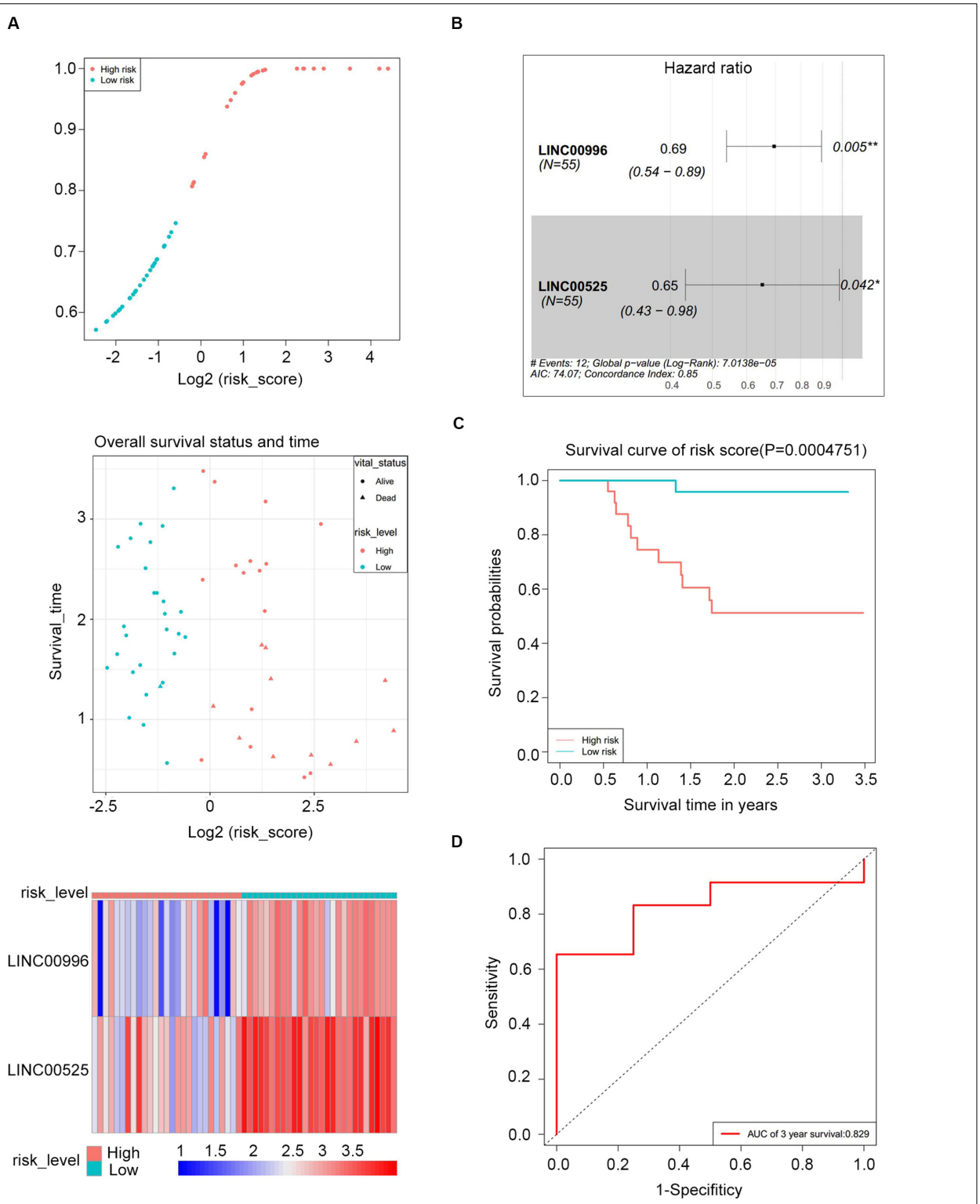


FIGURE 4 | Risk score performance in the GSE57317 (training) datasets. **(A)** Risk score of the two lncRNAs in 55 MM patients (top); overall survival status and duration (middle); heatmap of the two lncRNA expression in MM patients (bottom). **(B)** Forest plot showing the hazard ratios with 95% confidence interval of the multivariate Cox regression results. **(C)** Overall survival of the high- and low-risk score groups. **(D)** Three-year survival receiving operating characteristic curve (ROC) according to the two-lncRNA signature risk score (red).

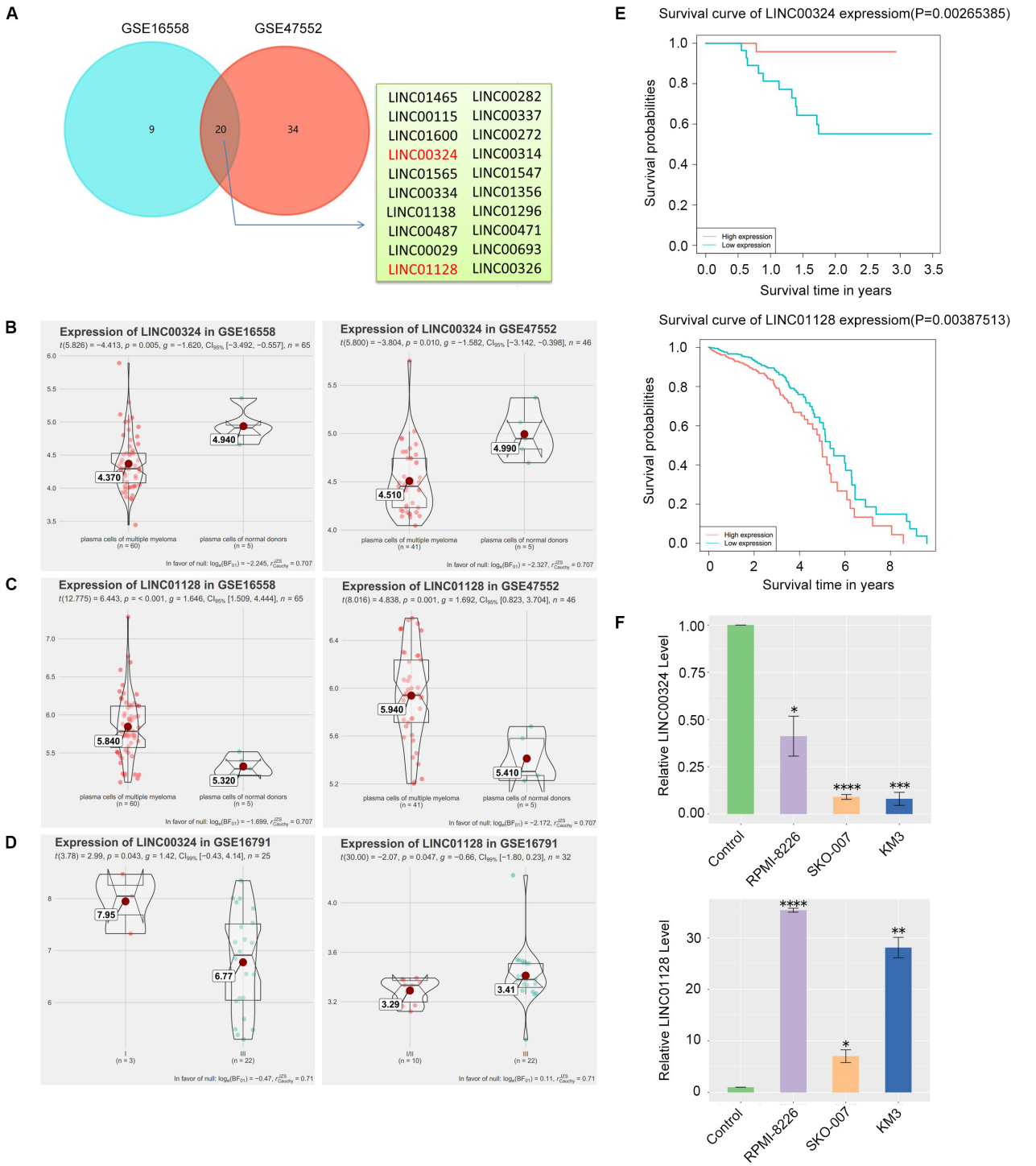


FIGURE 5 | Detection and validation of differentially expressed lncRNAs (DELs). **(A)** Venn plot of DELs between GSE16558 and GSE47552. **(B)** GSE16558 and GSE47552 indicated the lower expression of LINC00324 in CD138+ plasma cells of MM patients compared with normal donors. **(C)** GSE16558 and GSE47552 indicated the higher expression of LINC01128 in CD138+ plasma cells of MM patients compared with normal donors. **(D)** Relationship between the two lncRNAs and cancer stages in GSE16791. **(E)** GSE57317 dataset (left) and TCCA dataset (right) revealed that MM patients with low expression of LINC00324 and high expression of LINC01128 had an obviously poorer overall survival. **(F)** Expression of LINC00324 (top) and LINC01128 (bottom) in human multiple myeloma cell lines (RPMI-8226, SKO-007, KM3) as well as normal plasma cells. Data are presented as the mean ± standard deviation. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

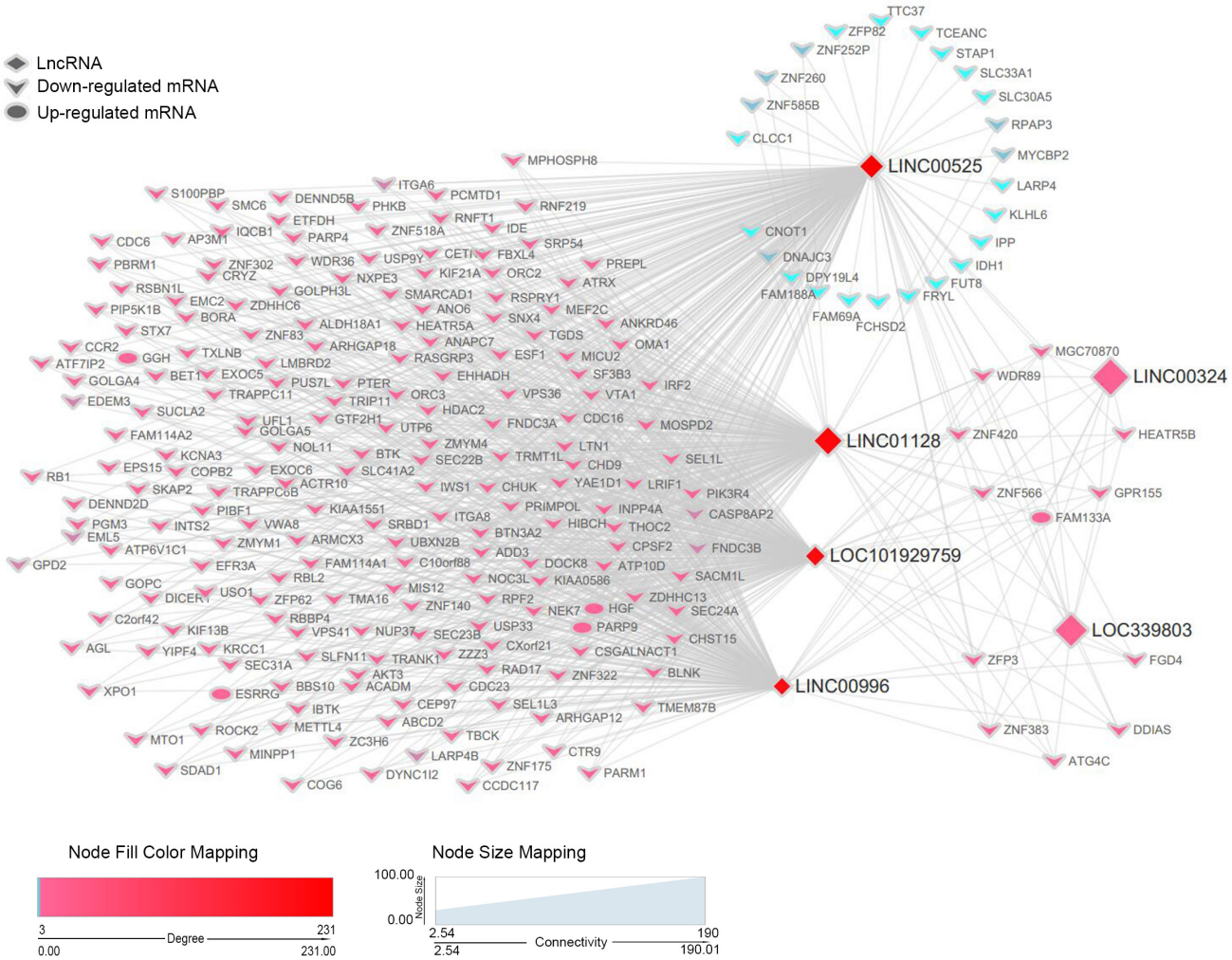
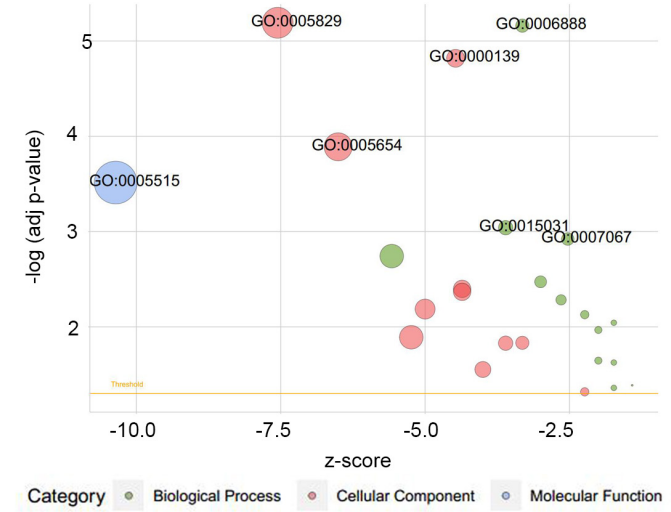


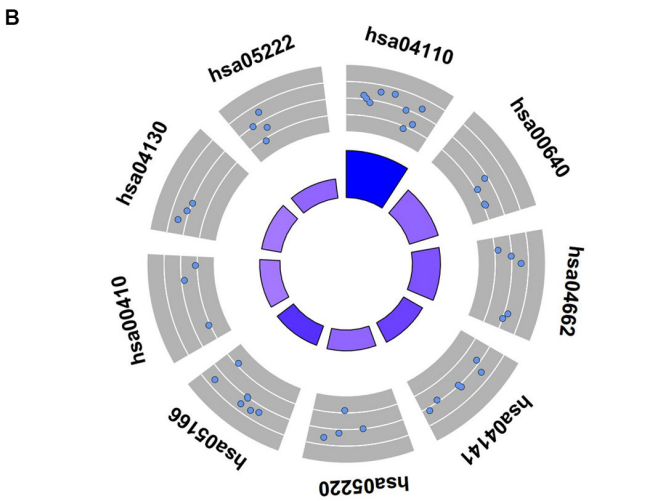
FIGURE 6 | A co-expression network of six lncRNAs and DEmRNAs in the blue module. For simplicity, only DEmRNAs with a connection with the interesting lncRNAs in the blue module were retained to construct the co-expression subnetwork. lncRNAs are depicted by diamonds, whereas DEmRNAs are indicated by the rounded rectangles (upregulation) or vs. (downregulation). Size of genes is related to the intra-modular connectivity, and the color is related to the weighted score of the interactions.

Detection and Validation of Differentially Expressed Long Non-coding RNAs

CD138+ plasma cells obtained from healthy donors in GSE16558 and GSE47552 were analyzed. Based on the cutoff criteria of $P < 0.05$, 20 DELs were detected (**Figure 5A**). Surprisingly, among the 12 prognostic lncRNAs we identified earlier, LINC00324 and LINC01128 are abnormally expressed (**Figures 5B,C**). The relationship between the two lncRNAs and cancer stages in GSE16791 is displayed in **Figure 5D**. Expression levels of the two lncRNAs among patients with different stages were compared, and statistical differences were calculated using Student's *t*-test. Corresponding to our previous WGCNA and PCA results, patients with poorly differentiated stage III cancer displayed significantly lower LINC00324 expression levels compared with patients with moderately differentiated cancer of less advanced stage I. Furthermore, increased expression of LINC01128 was correlated with advanced MM stage. Also, to determine the prognostic value of these two lncRNAs in MM, the survival data of MM patients were obtained from the TCGA database and GSE57317. As presented in **Figure 5E**, patients with high LINC01128 expression exhibited a significantly poorer OS rate compared with patients with high LINC01128 expression. On the contrary, we observed that patients with higher LINC00324 expression had better OS than those with lower LINC00324 expression. These results indicate that LINC00324 may be a tumor suppressor gene, whereas LINC01128 may be a cancer gene. The qRT-PCR results also showed that the expression pattern of the two lncRNAs in MM cells and normal plasma cells was similar to the microarray results (**Figure 5F**). LINC01128 was upregulated, whereas LINC00324 was significantly downregulated in the three MM cell lines.



ID	Description
GO:0006888	ER to Golgi vesicle–mediated transport
GO:0015031	protein transport
GO:0007067	mitotic nuclear division
GO:0005829	cytosol
GO:0000139	Golgi membrane
GO:0005654	nucleoplasm
GO:0005515	protein binding



ID	Description
hsa04110	Cell cycle
hsa00640	Propanoate metabolism
hsa04662	B cell receptor signaling pathway
hsa04141	Protein processing in endoplasmic reticulum
hsa05220	Chronic myeloid leukemia
hsa05166	HTLV−I infection
hsa00410	beta−Alanine metabolism
hsa04130	SNARE interactions in vesicular transport
hsa05222	Small cell lung cancer

FIGURE 7 | Functional annotation analysis of DEmRNAs co-expressed with LINC01128. **(A)** Gene ontology (GO) terms enrichment analysis was performed using the Database for Annotation, Visualization, and Integrated Discovery and visualized by GOplot. Significance of a term is indicated by the adjusted *P*-value (adj_p-val). logFC of selected genes is taken from GSE16558. *z*-score provided a hint if the biological process (/molecular function/cellular components) was more likely to be decreased (negative value) or increased (positive value). *z*-score is assigned to the *x*-axis and the negative logarithm of the adjusted *P*-value to the *y*-axis. Area of the displayed circles is proportional to the number of genes assigned to the term, and the color corresponds to the category. A threshold for the labeling is set as log(adj_p-value) > 2.8. **(B)** Plot of the enriched KEGG pathway. Outer circle shows a scatter plot for each term of the logFC of the assigned genes. Red circles display upregulation and blue ones downregulation by default. There were no upregulated DEmRNAs co-expressed with LINC01128 enriched.

Co-expression Network of Key Long Non-coding RNAs and Differentially Expressed Messenger RNAs in the Blue Module

Based on the previous results, we recognized six lncRNAs (LINC00525, LINC00996, LINC01128, LINC00324, LINC101929759, and LINC339803) as potential biomarkers or prognostic indicators. These lncRNAs were all in the blue module. To further dissect the role of six lncRNAs in MM, we created a gene co-expression subnetwork for the genes in the blue module according to their topology overlap matrix similarity; messenger RNAs (mRNAs) connected to six lncRNAs are too much to display perfectly; thus, we selected only differentially expressed mRNAs (DEmRNAs) to construct a network. Our lncRNAs may potentially regulate these co-expressed DEmRNAs through the ceRNA mechanism. DEmRNAs were obtained from GSE16558 and GSE47552 based on the cutoff criteria of a *P*-value < 0.05; $|\log(FC)| > 1.680$ DEmRNAs that overlapped in GSE16558 and GSE47552 were identified (**Supplementary Table S4**). Finally, the connections between the six lncRNAs and DEmRNAs are displayed in

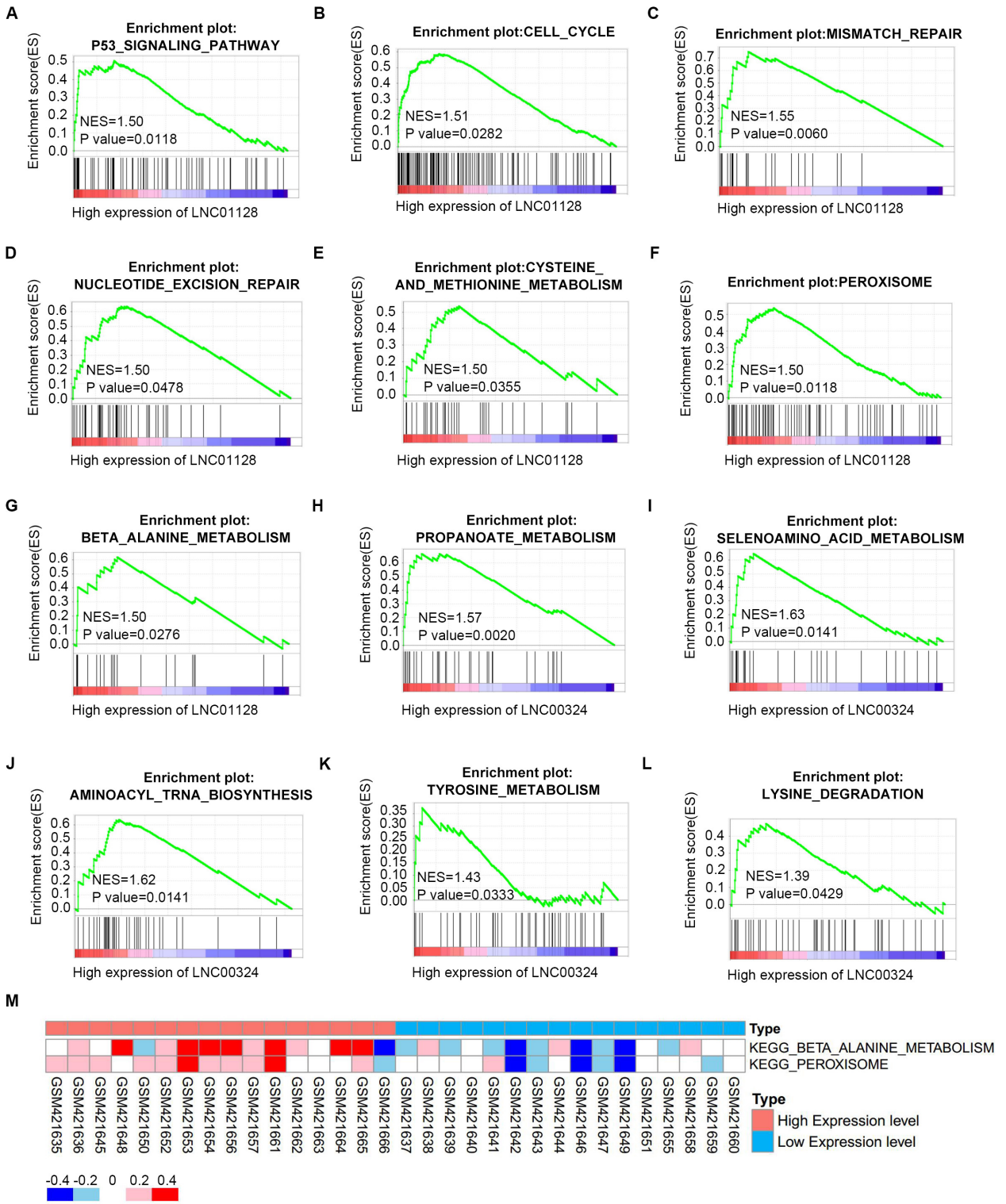


FIGURE 8 | Gene set enrichment analysis (GSEA) and gene set variation analysis (GSVA) of two lncRNAs in GSE16791. **(A–G)** GSEA results of c2 reference gene sets for high LINC01128 expression groups in GSE16791. **(H–L)** GSEA results of c2 reference gene sets for high LINC00324 expression groups in GSE16791. **(M)** GSVA-derived clustering heatmaps of differentially expressed pathways for LINC01128 in GSE16791.

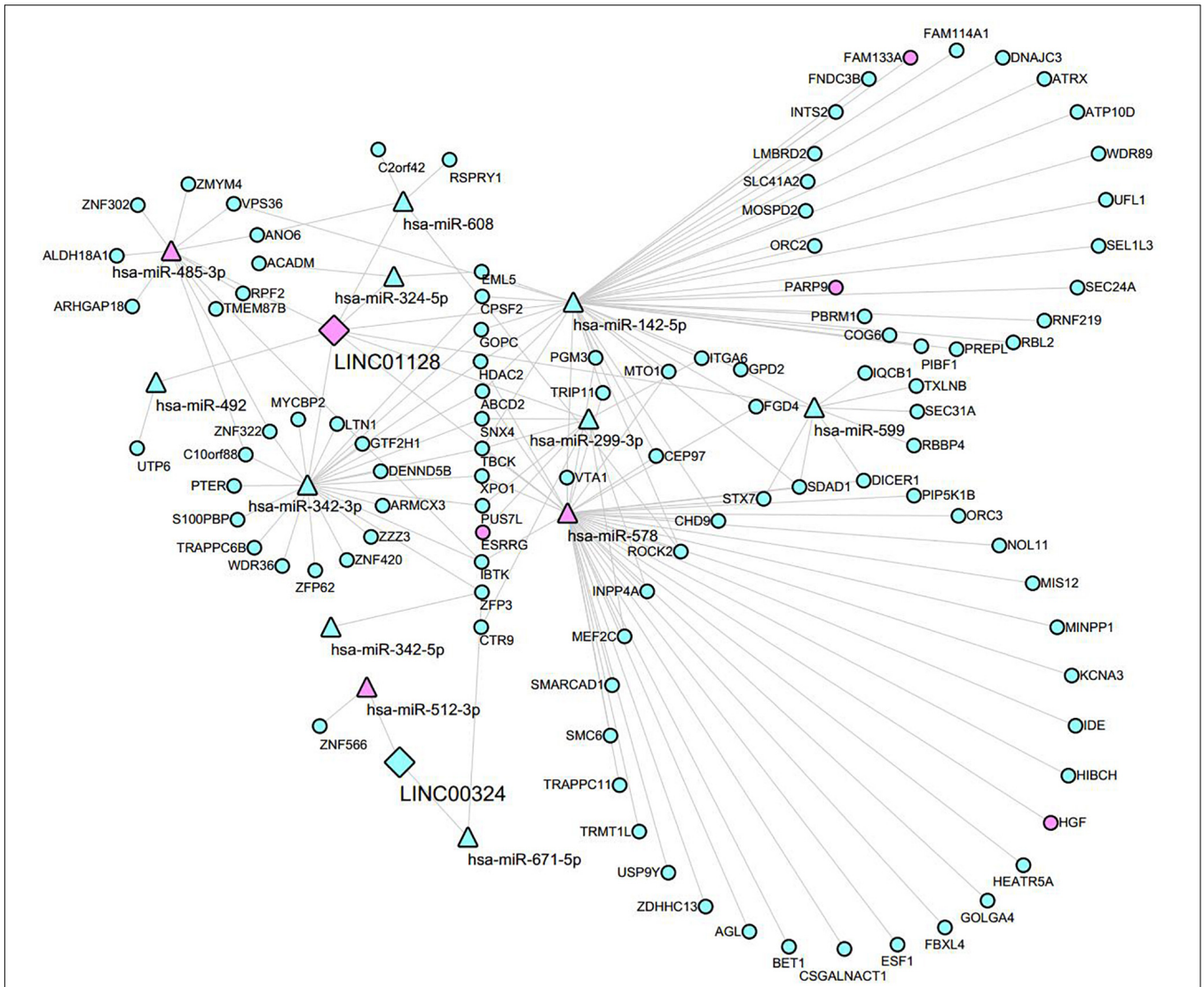


FIGURE 9 | Global view of the ceRNA network in MM. Network consists of two lncRNA nodes, 12 miRNA nodes, and 103 mRNA nodes. Diamonds indicate lncRNAs, triangles indicate miRNA, and ellipses indicate mRNA. Pink and blue represent up- and downregulation, respectively.

Figure 6. LncRNAs are shown by diamonds, whereas DEmRNAs are represented by round rectangles (upregulation) or vs. (downregulation). The size of the nodes reflects the strength of connectivity, and the color is related to the weighted score of the interactions.

Functional Annotation

The preceding findings indicated that the LINC00324 and LINC01128 were potentially involved in the occurrence and progression of MM. To more precisely understand the biological relevance and function of these two lncRNAs, we uploaded DEmRNAs, which were co-expressed with key lncRNAs in the blue module into the Database for Annotation, Visualization, and Integrated Discovery to conduct Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses. The results were visualized using the GOplot R package. The results of the differential analysis were used to calculate a *z*-score for presenting enriched KEGG pathways (**Supplementary Table S4**). Regarding enriched GO terms, DEmRNAs co-expressed with LINC01128 were mainly enriched in the endoplasmic reticulum to Golgi vesicle-mediated transport, protein transport, mitotic nuclear division, cytosol, Golgi membrane, nucleoplasm, and protein binding (**Figure 7A**). Regarding the enriched KEGG pathways, there were no upregulated DEmRNAs co-expressed with LINC00324 enriched, and other downregulated DEmRNAs were significantly enriched in the cell cycle, propanoate metabolism, B-cell receptor signaling pathway, protein processing in the endoplasmic reticulum, chronic myeloid leukemia, human T-cell lymphotropic virus type 1 infection, and beta-alanine metabolism (**Figure 7B**). There were no significant results for LINC00324 because too few mRNAs are connected with it.

Gene Set Enrichment Analysis and Gene Set Variation Analysis Reveal a Close Relationship Between Key Long Non-coding RNAs, Multiple Cancer-Related Pathways, and Metabolic Pathways

To further investigate the potential functions of LINC01128 and LINC00324, we performed gene set enrichment analysis (GSEA) and gene set variation analysis (GSVA) on the GSE16791 dataset. We divided these samples into two groups based on the expression levels of these two lncRNAs. As shown in **Figures 8A–G**, samples in GSE16791 with high expression of LINC01128 were enriched in multiple cancer-related pathways, including the P53 signaling pathway, cell cycle, mismatch repair, nucleotide excision repair, and several metabolic pathways, including cysteine and methionine metabolism, peroxisome, and beta-alanine metabolism. Also, our previous finding that the DEmRNAs co-expressed with LINC01128 were enriched in beta-alanine metabolism was, surprisingly, verified by GSEA and GSVA results (**Figures 8G,M**). The expression level of LINC00324 was also extracted for enrichment analysis. Genes in the high expression groups of LINC00324 were mainly involved in multiple metabolic pathways, including propanoate metabolism, selenoamino acid metabolism, aminoacyl-tRNA biosynthesis, tyrosine metabolism, and lysine degradation (**Figures 8H–L**).

Long Non-coding RNA-Mediated Competing Endogenous RNA Network Revealed Potential Mechanisms of LINC01128 and LINC00324

To investigate the interaction between the lncRNA and mRNAs, the lncRNA–miRNA–mRNA network was constructed according to the ceRNA hypothesis by integrating expression profile data and their regulatory relationships. We obtained DEmRNAs, DEmiRNAs based on the criteria mentioned in section "Materials and Methods." The interaction between the two lncRNAs and miRNAs were first predicted through Starbase3.0 and the RNA22 tool. We then predicted that the potential DEmiRNAs can target LINC01128 and LINC00324 co-expressed DEmRNAs in the blue module using DIANA TOOLS (**Supplementary Table S5**). Finally, a total of 12 miRNAs overlapped in our prediction results; 2 lncRNAs and 103 mRNAs were included in the ceRNA network (**Supplementary Table S6**), and their regulatory relationships were visualized by Cytoscape (**Figure 9**). In this network, different shapes represent different RNA types, with pink and blue denoting up- and downregulation, respectively.

DISCUSSION

Multiple myeloma is the most common primary bone cancer among 70-year-old and older American adults (Reisenbuckler, 2014). Although genetic and epigenetic events contributing to the occurrence and progression of MM have been increasingly identified, the diagnosis, treatment, and clinical outcome of MM remain mostly unclear (Prideaux et al., 2014). More recently, aberrant lncRNA expression in MM was observed and further validated to be involved in epigenetic, transcriptional, and posttranscriptional regulation (Meng et al., 2018). Several lncRNA prognostic models have been identified in multiple cancers, including hepatocellular carcinoma (Zhang et al., 2020), bladder cancer (Zhou et al., 2020), non-small cell lung cancer (Zhou et al., 2015a; Sun et al., 2020), breast cancer (Shen et al., 2020), glioma (Wang et al., 2018), glioblastoma (Zhou et al., 2018), and diffuse large B-cell lymphoma (Zhou et al., 2017). These studies had highlighted the diagnostic and prognostic roles of lncRNAs, and the lncRNA signatures they constructed had an imperative value for survival predicting for different cancer patients. Therefore, identifying new and effective prognostic biomarkers and establishing a reliable prognostic model based on lncRNA expression signature are critical for patients with MM.

WGCNA is a powerful algorithm that has not yet been utilized to analyze the expression profile of MM samples. Presently, a total of 32,216 genes, which were not all DEGs, were selected to conduct WGCNA analysis in case of missing significant information. Furthermore, we applied PCA for the first time to correlate PCs with clinical traits to find key lncRNAs. Then, 12 key lncRNAs that were associated with cancer stage, risk score, and proliferation index were identified in the intersection of key modules and PCs. Univariate Cox regression analysis retained six significant lncRNAs ($P < 0.05$) for further analysis. A co-expression network of six lncRNAs and co-expressed DEmRNAs in the blue module was constructed to present the co-expression pattern and the relationship between key lncRNAs and DEmRNAs. This network can provide insights for identifying possible targets of key lncRNAs. After the LASSO and Cox proportional hazard regression analysis, we detected a prognostic formula for predicting survival based on the two lncRNAs, including LINC00525 and LINC00996, and verified it in the testing set. The patients were ultimately divided into high- or low-risk patients according to the median risk value. Kaplan–Meier analysis showed that the patients in the high-risk score group displayed obviously worse OS compared with those in the low-risk score group. Furthermore, ROC curve analysis revealed the stability and accuracy of the two-lncRNA signature in predicting patient prognosis. Further analysis showed that the two-lncRNA risk score signature is an independent predictor of MM patient prognosis. Indeed, prior studies had established several lncRNA prognostic signatures that can provide a comprehensive clinical assessment of MM prognosis (Zhou et al., 2015b; Samur et al., 2018). Significantly, instead of simply utilizing survival associated lncRNAs to construct lncRNA prognostic signatures, it Is our first time to combine WGCNA and PCA to select prognostic lncRNAs that could be further used to establish a survival model. Subsequently, we performed a series of rigorous analyses, including univariate, LASSO Cox regression, and multivariate Cox hazard regression analysis to realize exact survival prediction. Additionally, in contrast to the earlier lncRNA model in MM (Zhu et al., 2020), we utilized an external dataset to examine the accuracy of our lncRNA signature.

Many recent studies have indicated that lncRNAs can regulate gene expression by interacting with the miRNA via MREs in MM (Sun et al., 2019a; Yang and Chen, 2019). Thus, it is imperative to recognize MM-specific lncRNAs as biomarkers and determine their potential mechanisms. These lncRNAs may be essential in the initiation and development of MM. Firstly, we identified 12 interesting lncRNAs, which may participate in the development of MM. To further select MM-specific lncRNAs, we screened DElncRNAs that overlapped in GSE16558 and GSE47552. Surprisingly, our PCR and microarray results indicated that 2 of the 12 lncRNAs (LINC01128 and LINC00324) were differentially expressed. LINC00324 can promote proliferation and metastasis but can inhibit cell apoptosis of lung adenocarcinoma cells by sponging miR-615-5p to promote AKT1 expression (Pan et al., 2018). Similar results were also found where LINC00324 can promote gastric cancer cell proliferation by binding with HuR and stabilizing FAM83B expression (Zou et al., 2018). It can also be used to predict the prognosis in patients with thymoma (Gong et al., 2018). There are no references for LINC01128. Its potential function remains to be determined. Next, GO analysis revealed that those LINC01128 co-expressed DEmRNAs were associated with protein transport and protein binding processes. KEGG pathway analysis demonstrated that they were enriched in cancer-related pathways, including cell cycle, chronic myeloid leukemia, small cell lung cancer, and metabolism-related pathways, including propanoate metabolism and beta-alanine metabolism. To further explore the underlying mechanism of LINC00324 and LINC01128, we formulated a ceRNA network based on predicted interactions between DEmiRNAs and DEmRNAs. Based on our network and the ceRNA mechanism, we speculated that LINC01128 might act as a tumor suppressor in MM through multiple mechanisms, including miR-142-5p/PARP9 or FAM133A axis, and the miR-299-3p/estrogen-related receptor gamma axis. The cancer–testis antigen FAM133A is a downstream target of miR-155 and is a negative regulator of glioma invasion and migration (Huang et al., 2018). Estrogen-related receptor gamma is a tumor suppressor as well as an activator of multiple cancers, including gastric cancer (Kang et al., 2018), breast cancer (Kumari et al., 2018), laryngeal squamous cell carcinoma (Shen et al., 2019), and liver cancer (Kim et al., 2016). LINC00324 may exert tumor-promoting functions in MM through targeting the miR-512-3p/ZNF566 axis. However, this remains to be verified. Finally, GSEA revealed that samples with high expression of LINC01128 were in multiple cancer-related pathways, including the P53 signaling pathway, cell cycle, mismatch repair, nucleotide excision repair, and several metabolic pathways, including cysteine and methionine metabolism, peroxisome, and beta-alanine metabolism. Several studies have reported that the cell cycle, P53 signaling, and DNA repair-related pathways are important tumor biological mechanisms (Balint and Vousden, 2001; Jackson and Bartek, 2009). Also, high beta-alanine concentrations are linked with cancer (Pine et al., 1982; Nishimura et al., 2012). Our findings suggested that the high expression of LINC01128 may be crucial in tumorigenesis and progression of MM, probably by regulating the cell cycle, DNA damage, or amino acid metabolism. Corresponding with our predicted mechanism of LINC01128, the mutation of the NAD+ binding site in PARP9 has been reported to increase the DNA repair activity of the heterodimer (Yang et al., 2017). On the other hand, genes in high expression groups of LINC00324 were mainly involved in multiple metabolic pathways, including propanoate metabolism, selenoamino acid metabolism, aminoacyl-tRNA biosynthesis, tyrosine metabolism, and lysine degradation. These observations can be explained by the hypothesis that LINC00324 suppresses tumorigenesis of MM by interfering with carbohydrate metabolism, amino acid metabolism, and protein translation.

In conclusion, WGCNA and PCA were performed to correlate the gene expression profile of patients with MM to the corresponding clinical traits. We identified lncRNAs that may potentially be involved in the initiation and development of MM. Finally, a two-lncRNA risk score model was formulated, and its precise prediction value was demonstrated. We also identified two lncRNAs as biomarkers and predicted their possible function as ceRNAs. These findings provide fundamental insights for further basic studies.

MATERIALS AND METHODS

Gene Expression Profile Data and Clinical Characteristics

The overall design and workflow of this study are presented in **Supplementary Figure S1**. The RNA expression profiles of MM patients and normal donors were identified from the GEO database[2] (**Table 2**). GSE16791 was utilized to conduct a WGCNA analysis for this study. This series of microarray experiments include 16,325 mRNA and 1,137 lncRNA expression profiles of purified plasma cells (PCs) obtained from 32 newly diagnosed MM. GSE17306 is a microarray analysis that contains 16,401 mRNA, 556 miRNA, and 1,146 lncRNA expression profiles of MM patients with corresponding clinical information, including mRNA-based GEP-risk score and proliferation index (Shaughnessy et al., 2007). It was used here to implement the PCA algorithm to correlate clinical traits with gene expression patterns. Corresponding clinical information, including survival time and vital status, was obtained from the GSE57317, including 16,325 mRNA and 1,137 lncRNA expression profiles of 55 MM patients, and TCGA RNA-Seq dataset contains 56,753 mRNA, 1,881 miRNA, and 14,142 lncRNA expression profiles of 765 MM patients to construct lncRNA risk score system. GSE16558, including 18,966 mRNA, 382 miRNA, and 431 lncRNA expression profiles of 60 MM patients and 5 healthy donors, GSE47552, including 18,966 mRNA and 431 lncRNA expression profiles of 41 MM patients and 5 healthy donors, and GSE17498, including 722 miRNA expression profiles of 40 MM patients and 3 healthy donors, were used to screen DEGs including DElncRNAs, DEmiRNAs, and DEmRNAs. Microarray annotation information was utilized to match probes with corresponding genes, and

[2]https://www.ncbi.nlm.nih.gov/geo/

TABLE 2 | Summary of included datasets.

Dataset ID	Sample size		Age (year)	Gender		Tumor stage			Vital status	
	Multiple myeloma	Normal		Male	Female	I	II	III	Alive	Dead
GSE16791	32	0	40–65	–	–	3	7	22	–	–
GSE17306	52	2	–	–	–	–	–	–	–	–
GSE57317	55	0	–	–	–	–	–	–	43	12
GSE16558	60	5	–	–	–	–	–	–	–	–
GSE17498	40	3	39–85	23	17					
GSE47552	41	5	–	–	–	–	–	–	–	–
TCGA	765	0	27–88	449	316	266	276	223	609	156

lncRNA expression was obtained based on the annotation of Genecode (see footnote 1).

Weighted Co-expression Network Analysis

A total of 32,216 genes identified in each sample of GSE16791 were utilized to construct a gene co-expression network using the WGCNA R package (Langfelder and Horvath, 2008; Chen L. et al., 2017). Sample clustering of all genes was applied to check if they were good genes and good samples. A scale-free co-expression network was achieved when the soft-threshold power was set as 8 (scale-free $R^2 = 0.81$), cut height as 0.25, and minimal module size as 30. Then, to evaluate co-expression levels between genes, Pearson correlations were performed and then weighted by raising their absolute value to a power. Hierarchical clustering dendrograms visualized gene modules in different colors. Modules with the highest correlation with cancer stage were selected for further analysis.

Principal Component Analysis

Principal component analysis compresses all the original variables into a smaller subset of composite variables (PCs) instead of ignoring or discarding variables. PCA tools, a useful R package that provides functions for data exploration, were applied to analyze GSE17306 dataset[3]. At first, PCA helped us to determine PCs, accounting for 80% of the explained variation. Secondly, we correlated the PCs back to the clinical data, including mRNA-based GEP-risk score and proliferation index, to gain interesting PCs. Finally, the plotLoadings function could contribute to determining the variables ranked top 5% of the loadings range.

Identification and Evaluation of a Risk Assessment Model

The prognostic value of 12 lncRNAs in the intersection of blue and green modules, and PC6 and PC8, were evaluated by a univariate Cox model with a statistical level of significance set at $P < 00.05$. Critical prognostic lncRNAs were further identified by the LASSO regression method (Gao et al., 2010). LASSO regression is a penalized regression method that is often used in machine learning to select the subset of variables. The R glmnet software package was adopted to carry out the LASSO Cox analysis (Tibshirani, 1997). Also, lncRNAs obtained in these steps were then enrolled into a multivariate Cox regression model using a survival R package, and prognosis-associated lncRNAs were selected. The risk score of each patient was calculated based on the summation of each lncRNA and its coefficient, and we distinguished high- from low-risk patients according to the median risk score. The Kaplan–Meier method was applied to analyze the difference of OS between two groups, and a ROC analysis was adopted to estimate the predictive power of this lncRNA risk score system. The TCGA dataset served as a testing set for further validation.

Construction of Co-expression Network of Key Long Non-coding RNAs and Differentially Expressed Messenger RNAs in the Blue Module

The multivariate Cox regression analysis identified six lncRNAs with $P < 0.05$, which were considered as key lncRNAs. We created a gene co-expression subnetwork for the genes in the blue module according to their topology overlap matrix similarity; DEmRNAs connected to key lncRNAs were selected to construct a co-expression network using Cytoscape. DEmRNAs that overlapped in GSE16558 and GSE47552 ($n = 680$) were identified based on the cutoff criteria of $P < 0.05$ and $|\log(FC)| > 1$. The size of the nodes reflected the strength of connectivity, and the color was related to the weighted score of the interactions.

Screening of Database of Essential Genes and Survival Analysis

The Limma package in R (Ritchie et al., 2015) was used to identify the DEGs from GSE16558 and GSE47552. We identified DElncRNAs and DEmiRNAs according to the criterion that adjusted $P < 0.05$. Abnormally expressed miRNAs in GSE17306, GSE16558, and GSE17498 were all selected for constructing the ceRNA network. The two DElncRNAs were utilized to perform Kaplan–Meier analysis and log-rank test to identify whether they were correlated with OS using the GSE57317 and TCGA datasets. Log-rank test with $P < 0.05$ was set as statically significant.

[3]https://github.com/kevinblighe/PCAtools

Cell Lines and Clinical Specimens

The RPMI-8226, SKO-007, and KM3 MM cell lines were a generous gift of Prof. Yumin Huang, Department of Hematology, First Affiliated Hospital of Zhengzhou University. Cells were maintained in RPMI-1640 medium (Sigma-Aldrich, St. Louis, MO, United States) with 10% fetal bovine serum at 37°C in an atmosphere of 5% CO_2. BM was obtained from three healthy controls from a pool of volunteers without any diseases. All volunteers provided written informed consent, and the research ethics committee of the First Affiliated Hospital of Zhengzhou University approved the study (2019-KY-357). Flow cytometry was performed using the CD138 antibody (PE, BD Bioscience, United States) to isolate CD138-positive PC from BM samples according to the manufacturer's protocol.

Quantitative Real-Time Polymerase Chain Reaction

Total RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, United States). A NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, United States) was utilized to detect RNA purity and concentration. RT-PCR was performed using a FastStart Universal SYBR Green Master (Servicebio, Wuhan, China) Kit. Actin was used as an internal control. Primers were synthesized by Servicebio (Wuhan, China). Primer sequences were: LINC01128: Forward 5′-AGGACATAGGCCAGCCAGTAC-3′, Reverse 5′-GTCTTTGGTCCCAGATCACTCC-3′; LINC00324: Forward 5′-ACCTACGGTTTCTGGTCAGCG-3′, Reverse 5′-GACGACGGCAGCCATTACTTT-3′; ACTIN: Forward 5′-CACCCAGCACAATGAAGATCAAGAT-3′, Reverse 5′-CCAGTTTTTAAATCCTGAGTCAAGC-3′. LINC00525: Forward 5′-GCTTTGGAAACTTACTCAGGGTG-3′, Reverse 5′-CTTGAGGCACCAGTGCAAATAC-3′; LINC00996: Forward 5′-GAGGGCACTTTGTCTTACTTGGC-3′, Reverse 5′-ATTCTTCATGCCAATCCTCTCAC-3′. Relative expression was calculated using the 2-△△Ct method. Student's t-test was conducted by SPSS 25.0 software (SPSS Inc., Chicago, IL, United States) to determine the significance of the differences in mean values.

Construction of Interesting Differentially Expressed Long Non-coding RNA-Based Competing Endogenous RNA Network

The ceRNA hypothesis posits that lncRNAs can regulate gene expression by interacting with miRNA at miRNA-binding sites (MREs). It is vital to match the DEmRNAs, miRNAs, and lncRNAs to figure out a novel molecular mechanism involved in the development of MM. The MIRanda database[4], Starbase3.0[5], and RNA22 tool[6] were used to predict the interactions between DElncRNAs and miRNAs. The miRNAs that potentially target DEmRNAs were predicted by DIANA Tools[7]. DElncRNAs, DEmRNAs, and DEmiRNAs that overlapped with the predicted miRNAs were selected to construct a ceRNA network and were visualized with Cytoscape version 3.6.1.

Functional Annotation of Long Non-coding RNA Target Genes

The GO and KEGG enrichment analyses for DEmRNAs, which were co-expressed with LINC00324 and LINC01128, were analyzed using the Database for Annotation, Visualization, and Integrated Discovery database (Huang et al., 2007) and visualized by the GOplot R package (Walter et al., 2015). The *z*-score is a value that can be easily calculated and reveals whether the biological process (molecular function/cellular components) is more likely to be decreased (negative value) or increased (positive value). It is calculated as *z*-score = (up-down)/$\sqrt{}$count. Up or down represents the number of upregulated or downregulated genes, respectively. The count represents the number of genes that belong to each term. A threshold for the labeling is set as log (adjust *p*-value) > 2.8.

Gene Set Enrichment Analysis and Gene Set Variation Analysis

The GSE16791 dataset was used to conduct GSEA according to expression levels of two lncRNAs (high expression vs. low expression) (Subramanian et al., 2005). Annotated gene sets c2.cp.kegg. v 7.0.symbols.gmt was chosen as the reference gene sets[8]. The nominal *P*-value estimates the statistical significance of the enrichment score, and a nominal *P*-value $\leq$ 0.05 was set as the cutoff criterion.

DATA AVAILABILITY STATEMENT

The datasets GSE16791, GSE17306, GSE57317, GSE16558, and GSE47552 for this study can be found in the GEO database (https://www.ncbi.nlm.nih.gov/geo/). The TCGA datasets were downloaded from TCGA (https://cancergenome.nih.gov/) database.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of the First Affiliated Hospital of Zhengzhou University (2019-KY-357). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SZ and JF contributed equally to this work, they were responsible for study design, data collection, and data analyzing. YS was involved in manuscript preparation and literature searching. HL took part in study design and manuscript revision.

[4]http://www.microrna.org/

[5]http://starbase.sysu.edu.cn/index.php

[6]https://cm.jefferson.edu/

[7]http://diana.imis.athena-innovation.gr/DianaTools/index.php

[8]http://software.broadinstitute.org/gsea/msigdb/index.jsp

All the authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors would like to thank the TCGA and GEO databases developed by the National Institutes of Health.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00934/full#supplementary-material

FIGURE S1 | Overall design and workflow of this study.

FIGURE S2 | Soft threshold determination and the relationship between these two modules and clinical traits. **(A)** Determination of soft-thresholding power in Wgcna. **(B)** Scatter plot of module eigengenes in blue and green modules. **(C)** Module eigengene dendrogram and interactions among different gene coexpression modules.

FIGURE S3 | Relative quantification of Linc00525 and Linc00996 expression by qRt-Pcr. The expression of Linc00525 **(A)** and Linc00996 **(B)** in human multiple myeloma cell lines (Rpmi-8226, Sko-007, Km3) as well as normal plasma cells. Data are presented as the mean ± standard deviation. The ns represents not significant, * represents $P < 0.05$, ** represents $P < 0.01$, *** represents $P < 0.001$ and **** represents $P < 0.0001$.

FIGURE S4 | The risk score performance in the Tcga (testing) datasets. **(A)** Risk score of the 2 lncRnas in 787 Mm patients (top); overall survival status and duration (middle); heatmap of the 2 lncRnas expression in Mm patients (Bottom). **(B)** The forest plot showed the hazard ratios (Hr) with 95% confidence interval (95%Ci) according to the multivariate Cox regression results. **(C)** The overall survival of high-risk score group and low-risk score group. **(D)** The 3-year survival receiving operating characteristic curve (Roc) of according to 2 lncRna signature risk score (red).

REFERENCES

Balint, E. E., and Vousden, K. H. (2001). Activation and activities of the p53 tumour suppressor protein. *Br.* J. Cancer 85, 1813–1823. doi: 10.1054/bjoc.2001.2128

Chen, L., Yuan, L., Wang, Y., Wang, G., Zhu, Y., Cao, R., et al. (2017). Co-expression network analysis identified FCER1G in association with progression and prognosis in human clear cell renal cell carcinoma. *Int. J. Biol. Sci.* 13, 1361–1372.

Chen, W. C., Kanate, A. S., Craig, M., Petros, W. P., and Hazlehurst, L. A. (2017). Emerging combination therapies for the management of multiple myeloma: the role of elotuzumab. *Cancer Manag. Res.* 9, 307–314. doi: 10.2147/CMAR.S117477

Cowan, A. J., Allen, C., Barac, A., Basaleem, H., Bensenor, I., Curado, M. P., et al. (2018). Global burden of multiple myeloma: a systematic analysis for the global burden of disease study 2016. *JAMA Oncol.* 4, 1221–1227.

Decaux, O., Lode, L., Magrangeas, F., Charbonnel, C., Gouraud, W., Jezequel, P., et al. (2008). Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosomal instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients: a study of the Intergroupe Francophone du Myelome. *J. Clin. Oncol.* 26, 4798–4805. doi: 10.1200/JCO.2007.13.8545

Dong, H., Jiang, S., Fu, Y., Luo, Y., Gui, R., and Liu, J. (2019). Upregulation of lncRNA NR_046683 Serves as a prognostic biomarker and potential drug target for multiple myeloma. *Front. Pharmacol.* 10:45. doi: 10.3389/fphar.2019.00045

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi: 10.1038/nrg3606

Gao, J., Kwan, P. W., and Shi, D. (2010). Sparse kernel learning with LASSO and Bayesian inference algorithm. *Neural Netw.* 23, 257–264. doi: 10.1016/j.neunet.2009.07.001

Geisler, S., and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* 14, 699–712. doi: 10.1038/nrm3679

Gong, J., Jin, S., Pan, X., Wang, G., Ye, L., Tao, H., et al. (2018). Identification of long non-coding RNAs for predicting prognosis among patients with Thymoma. *Clin. Lab.* 64, 1193–1198.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., et al. (2007). DAVID Bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* 35, W169–W175. doi: 10.1093/nar/gkm415

Huang, G. H., Du, L., Li, N., Zhang, Y., Xiang, Y., Tang, J. H., et al. (2018). Methylation-mediated miR-155-FAM133A axis contributes to the attenuated invasion and migration of IDH mutant gliomas. *Cancer Lett.* 432, 93–102.

Hung, T., and Chang, H. Y. (2010). Long noncoding RNA in genome regulation: prospects and mechanisms. *RNA Biol.* 7, 582–585. doi: 10.4161/rna.7.5.13216

Jackson, S. P., and Bartek, J. (2009). The DNA-damage response in human biology and disease. *Nature* 461, 1071–1078. doi: 10.1038/nature08467

Kang, M. H., Choi, H., Oshima, M., Cheong, J. H., Kim, S., Lee, J. H., et al. (2018). Estrogen-related receptor gamma functions as a tumor suppressor in gastric cancer. *Nat. Commun.* 9:1920.

Kim, J. H., Choi, Y. K., Byun, J. K., Kim, M. K., Kang, Y. N., Kim, S. H., et al. (2016). Estrogen-related receptor gamma is upregulated in liver cancer and its inhibition suppresses liver cancer cell proliferation via induction of p21 and p27. *Exp. Mol. Med.* 48:e213. doi: 10.1038/emm.2015.115

Kumari, K., Adhya, A. K., Rath, A. K., Reddy, P. B., and Mishra, S. K. (2018). Estrogen-related receptors alpha, beta and gamma expression and function is associated with transcriptional repressor EZH2 in breast carcinoma. *BMC Cancer* 18:690. doi: 10.1186/s12885-018-4586-0

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Laubach, J., Richardson, P., and Anderson, K. (2011). Multiple myeloma. *Annu. Rev. Med.* 62, 249–264. doi: 10.1146/annurev-med-070209-175325

Meng, H., Han, L., Hong, C., Ding, J., and Huang, Q. (2018). Aberrant lncRNA expression in multiple Myeloma. *Oncol. Res.* 26, 809–816.

Nishimura, S., Uno, M., Kaneta, Y., Fukuchi, K., Nishigohri, H., Hasegawa, J., et al. (2012). MRGD, a MAS-related G-protein coupled receptor, promotes tumorigenisis and is highly expressed in lung cancer. *PLoS One* 7:e38618. doi: 10.1371/journal.pone.0038618

Ouyang, D., Li, R., Li, Y., and Zhu, X. (2019). A 7-lncRNA signature predict prognosis of Uterine corpus endometrial carcinoma. *J. Cell Biochem.* 120, 18465–18477. doi: 10.1002/jcb.29164

Pan, Z. H., Guo, X. Q., Shan, J., and Luo, S. X. (2018). LINC00324 exerts tumor-promoting functions in lung adenocarcinoma via targeting miR-615-5p/AKT1 axis. *Eur. Rev. Med. Pharmacol. Sci.* 22, 8333–8342.

Pine, M. J., Kim, U., and Ip, C. (1982). Free amino acid pools of rodent mammary tumors. *J. Natl. Cancer Inst.* 69, 729–735.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465, 1033–1038. doi: 10.1038/nature09144

Prideaux, S. M., Conway O'Brien, E., and Chevassut, T. J. (2014). The genetic architecture of multiple myeloma. *Adv. Hematol.* 2014:864058.

Reisenbuckler, C. (2014). Multiple myeloma and diagnostic imaging. *Radiol. Technol.* 85, 391–410.

Ringner, M. (2008). What is principal component analysis? *Nat. Biotechnol.* 26, 303–304. doi: 10.1038/nbt0308-303

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47.

Samur, M. K., Minvielle, S., Gulla, A., Fulciniti, M., Cleynen, A., Aktas Samur, A., et al. (2018). Long intergenic non-coding RNAs have an independent impact on survival in multiple myeloma. *Leukemia* 32, 2626–2635.

Shaughnessy, J. D. Jr., Zhan, F., Burington, B. E., Huang, Y., Colla, S., Hanamura, I., et al. (2007). A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* 109, 2276–2284.

Shen, Y., Peng, X., and Shen, C. (2020). Identification and validation of immune-related lncRNA prognostic signature for breast cancer. *Genomics* 112, 2640–2646. doi: 10.1016/j.ygeno.2020.02.015

Shen, Z., Hu, Y., Zhou, C., Yuan, J., Xu, J., Hao, W., et al. (2019). ESRRG promoter hypermethylation as a diagnostic and prognostic biomarker in laryngeal squamous cell carcinoma. *J. Clin. Lab. Anal.* 33, e22899.

Shi, Z., Derow, C. K., and Zhang, B. (2010). Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. *BMC Syst. Biol.* 4:74. doi: 10.1002/jcla.22899

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550.

Sun, J., Zhang, Z., Bao, S., Yan, C., Hou, P., Wu, N., et al. (2020). Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. *J. Immunother. Cancer* 8:e000110.

Sun, Y., Jiang, T., Jia, Y., Zou, J., Wang, X., and Gu, W. (2019a). LncRNA MALAT1/miR-181a-5p affects the proliferation and adhesion of myeloma cells via regulation of Hippo-YAP signaling pathway. *Cell Cycle* 18, 2509–2523. doi: 10.1080/15384101.2019.1652034

Sun, Y., Zhang, F., Wang, L., Song, X., Jing, J., Zhang, F., et al. (2019b). A five lncRNA signature for prognosis prediction in hepatocellular carcinoma. *Mol. Med. Rep.* 19, 5237–5250.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16, 385–395. doi: 10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

Walter, W., Sanchez-Cabo, F., and Ricote, M. (2015). GOplot. *Bioinformatics* 31, 2912–2914. doi: 10.1093/bioinformatics/btv300

Wang, W., Zhao, Z., Yang, F., Wang, H., Wu, F., Liang, T., et al. (2018). An immune-related lncRNA signature for patients with anaplastic gliomas. *J. Neurooncol.* 136, 263–271.

Wu, W., Sui, J., Liu, T., Yang, S., Xu, S., Zhang, M., et al. (2019). Integrated analysis of two-lncRNA signature as a potential prognostic biomarker in cervical cancer: a study based on public database. *PeerJ* 7:e6761.

Yang, C. S., Jividen, K., Spencer, A., Dworak, N., Ni, L., Oostdyk, L. T., et al. (2017). Ubiquitin Modification by the E3 Ligase/ADP-Ribosyltransferase Dtx3L/Parp9. *Mol. Cell* 66, 503.e5–516.e5. doi: 10.1016/j.molcel.2017.04.028

Yang, Y., and Chen, L. (2019). Downregulation of lncRNA UCA1 facilitates apoptosis and reduces proliferation in multiple myeloma via regulation of the miR-1271-5p/HGF axis. *J. Chin. Med. Assoc.* 82, 699–709. doi: 10.1097/JCMA.0000000000000145

Zhang, Y., Zhang, L., Xu, Y., Wu, X., Zhou, Y., and Mo, J. (2020). Immune-related long noncoding RNA signature for predicting survival and immune checkpoint blockade in hepatocellular carcinoma. *J. Cell Physiol.* doi: 10.1002/jcp.29730 Online ahead of print

Zhou, H., Zhang, H., Chen, J., Cao, J., Liu, L., Guo, C., et al. (2019). A seven-long noncoding RNA signature predicts relapse in patients with early-stage lung adenocarcinoma. *J. Cell Biochem.* 120, 15730–15739.

Zhou, M., Guo, M., He, D., Wang, X., Cui, Y., Yang, H., et al. (2015a). A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J. Transl. Med.* 13:231.

Zhou, M., Zhang, Z., Bao, S., Hou, P., Yan, C., Su, J., et al. (2020). Computational recognition of lncRNA signature of tumor-infiltrating B lymphocytes with potential implications in prognosis and immunotherapy of bladder cancer. *Brief. Bioinform.* 8:bbaa047.

Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2018). An Immune-related six-lncRNA signature to improve prognosis prediction of Glioblastoma Multiforme. *Mol. Neurobiol.* 55, 3684–3697.

Zhou, M., Zhao, H., Wang, Z., Cheng, L., Yang, L., Shi, H., et al. (2015b). Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. *J. Exp. Clin. Cancer Res.* 34:102.

Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., and Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Mol. Cancer* 16:16.

Zhou, Y., Chen, L., Barlogie, B., Stephens, O., Wu, X., Williams, D. R., et al. (2010). High-risk myeloma is associated with global elevation of miRNAs and overexpression of EIF2C2/AGO2. *Proc Natl Acad Sci U S A* 107, 7904–7909.

Zhu, F. X., Wang, X. T., Ye, Z. Z., Gan, Z. P., and Lai, Y. R. (2020). Construction of a prognosisassociated long noncoding RNAmRNA network for multiple myeloma based on microarray and bioinformatics analysis. *Mol. Med. Rep.* 21, 999–1010.

Zou, Z., Ma, T., He, X., Zhou, J., Ma, H., Xie, M., et al. (2018). Long intergenic non-coding RNA 00324 promotes gastric cancer cell proliferation via binding with HuR and stabilizing FAM83B expression. *Cell Death Dis.* 9:717. doi: 10.1038/s41419-018-0758-8

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 September 2020
doi: 10.3389/fgene.2020.00895

Identification of Hub Genes Associated With Hepatocellular Carcinoma Using Robust Rank Aggregation Combined With Weighted Gene Co-expression Network Analysis

Hao Song[1,2], Na Ding[3], Shang Li[3], Jianlong Liao[3], Aimin Xie[3], Youtao Yu[2], Chunlong Zhang[3]* and Caifang Ni[1]**

[1] Department of Interventional Radiology, The First Affiliated Hospital of Soochow University, Suzhou, China, [2] Department of Intervention Therapy, The Fourth Medical Center of PLA General Hospital, Beijing, China, [3] Department of Computational Biology, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Bailiang Li,
Stanford University, United States

Reviewed by:
Ankush Sharma,
University of Oslo, Norway
Jiri Vohradsky,
Academy of Sciences of the Czech Republic, Czechia

***Correspondence:**
Youtao Yu
yuyoutao@126.com
Chunlong Zhang
zhangchunlong@hrbmu.edu.cn
Caifang Ni
caifangnisdfyy@163.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *22 April 2020*
Accepted: *20 July 2020*
Published: *30 September 2020*

Citation:
Song H, Ding N, Li S, Liao J, Xie A, Yu Y, Zhang C and Ni C (2020) Identification of Hub Genes Associated With Hepatocellular Carcinoma Using Robust Rank Aggregation Combined With Weighted Gene Co-expression Network Analysis. Front. Genet. 11:895. doi: 10.3389/fgene.2020.00895

Background: Bioinformatics provides a valuable tool to explore the molecular mechanisms underlying pathogenesis of hepatocellular carcinoma (HCC). To improve prognosis of patients, identification of robust biomarkers associated with the pathogenic pathways of HCC remains an urgent research priority.

Methods: We employed the Robust Rank Aggregation method to integrate nine qualified HCC datasets from the Gene Expression Omnibus. A robust set of differentially expressed genes (DEGs) between tumor and normal tissue samples were screened. Weighted gene co-expression network analysis was applied to cluster DEGs and the key modules related to clinical traits identified. Based on network topology analysis, novel risk genes derived from key modules were mined and biological verification performed. The potential functions of these risk genes were further explored with the aid of miRNA–mRNA regulatory networks. Finally, the prognostic ability of these genes was assessed by constructing a clinical prediction model.

Results: Two key modules showed significant association with clinical traits. In combination with protein–protein interaction analysis, 29 hub genes were identified. Among these genes, 19 from one module showed a pattern of upregulation in HCC and were associated with the tumor node metastasis stage, and 10 from the other module displayed the opposite trend. Survival analyses indicated that all these genes were significantly related to patient prognosis. Based on the miRNA-mRNA regulatory network, 29 genes strongly linked to tumor activity were identified. Notably, five of the novel risk genes, ABAT, DAO, PCK2, SLC27A2, and HAO1, have rarely been reported in previous studies. Gene set enrichment analysis for each gene revealed regulatory roles in proliferation and prognosis of HCC. Least absolute shrinkage and selection operator

regression analysis further validated DAO, PCK2, and HAO1 as prognostic factors in an external HCC dataset.

Conclusion: Analysis of multiple datasets combined with global network information presents a successful approach to uncover the complex biological mechanisms of HCC. More importantly, this novel integrated strategy facilitates identification of risk hub genes as candidate biomarkers for HCC, which could effectively guide clinical treatments.

Keywords: weighted gene co-expression network analysis (WGCNA), hub genes, hepatocellular carcinoma (HCC), biomarker, progression and prognosis

INTRODUCTION

Hepatocellular carcinoma (HCC) is the sixth most common malignant tumor type and the fourth leading cause of cancer-related deaths worldwide, with approximately 841,000 new cases and 782,000 deaths each year (Bray et al., 2018). Although multiple therapies have been recently developed for HCC, prognosis remains unsatisfactory due to disease progression, recurrence, and metastasis (Budhu et al., 2006). Abnormal expression of several genes is critical in tumorigenesis and development of HCC. Recent research has shown that tumor necrosis factor-α-induced protein 8 (TNFAIP8) increases HCC cell survival by blocking apoptosis, promoting greater resistance to the anticancer drugs sorafenib and regorafenib (Niture et al., 2020). High expression of ATP/GTP binding protein like 2 (AGBL2) is associated with significantly enhanced survival and proliferation of HCC cells *in vitro* and tumor growth *in vivo* (Wang L. L. et al., 2018). Although these single genes affect the phenotype of HCC, it is not known whether they constitute the hub genes. Integration of multiple datasets and network topology structures may therefore facilitate the identification of more robust biomarkers.

Owing to the substantial improvements in high-throughput gene microarray and next-generation sequencing technologies, bioinformatics analyses are increasingly applied to explore the biological characteristics of cancers. To avoid the potential large bias caused by analysis of a single dataset, many researchers have focused on analysis of multiple datasets for HCC. Recently, Li and colleagues examined the intersection of differentially expressed genes (DEGs) of three datasets (Li and Xu, 2020) and merged the multiple datasets for analysis (Li and Xu, 2020; Li et al., 2020). In the current study, we adopted the Robust Rank Aggregation (RRA) method for the analysis of multiple integrated datasets (Kolde et al., 2012).

We downloaded nine eligible microarray datasets from the Gene Expression Omnibus (GEO), which were subjected to meta-analysis to identify robust DEGs between HCC and matched normal tissues using the RRA method. Next, weighted gene co-expression network analysis (WGCNA) was performed with the DEGs to identify the most significant modules related to clinical traits of HCC. After screening the protein–protein interaction (PPI) network (Szklarczyk et al., 2015), the 29 hub genes uploaded to miRNet[1] were screened to construct miRNA–mRNA regulatory networks and explore their potential functions. In an external test dataset from The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) collection, 28 of these hub genes were associated with the prognosis and progression of HCC. Gene set enrichment analysis (GSEA) was further performed on the independent dataset (TCGA-LIHC) to determine the potential functions of the identified hub genes. Least absolute shrinkage and selection operator (LASSO) regression was applied to construct clinical predictive models with the aim of verifying the prognostic capability of these genes in patients with HCC. In summary, integrated analysis of multiple datasets was initially conducted, followed by comprehensive screening of hub genes strongly related to HCC using a variety of efficient bioinformatics methods and verification of the results in an external dataset. Our overall findings contribute to the elucidation of the molecular mechanisms underlying pathogenesis and identification of novel prognostic biomarkers for HCC.

MATERIALS AND METHODS

Data Sources

We downloaded nine microarray datasets from the GEO database for RRA[2]. Access numbers of the included datasets are as follows: GSE36376 (Lim et al., 2013), GSE39791 (Kim et al., 2014), GSE45114 (Wei et al., 2014), GSE57957 (Mah et al., 2014), GSE60502 (Wang et al., 2014), GSE76297 (Chaisaingmongkol et al., 2017), GSE76427 (Grinchuk et al., 2018), GSE84005, and GSE14520 (Roessler et al., 2010). Datasets were collected up to February 1, 2020, and were included based on the following criteria: (1) gene expression data from HCC and adjacent normal tissue samples were evaluated; (2) at least 15 pairs of tumor and paracancerous tissue samples were assessed; and (3) the number of genes in a single dataset was >10,000. GSE14520 contained adequate clinical information and the largest HCC sample number (471 samples) for WGCNA and LASSO regression. Detailed information on these datasets is provided in **Table 1**. Additionally, the TCGA-LIHC dataset containing 374 HCC and 50 normal samples was utilized as the external validation dataset and GSEA was performed.

[1]https://www.mirnet.ca/

[2]https://www.ncbi.nlm.nih.gov/gds/?term=

TABLE 1 | Details of the eight GEO datasets about HCC.

GEO	No. of samples		Platform	References
	T	N		
GSE36376	249	193	GPL10558	Lim et al., 2013
GSE39791	72	72	GPL10558	Kim et al., 2014
GSE45114	24	25	GPL5918	Wei et al., 2014
GSE57957	39	39	GPL10558	Mah et al., 2014
GSE60502	18	18	GPL96	Wang et al., 2014
GSE76297	61	58	GPL17586	Chaisaingmongkol et al., 2017
GSE76427	115	52	GPL10558	Grinchuk et al., 2018
GSE84005	38	38	GPL5175	NA
GSE14520	471	459	GPL571&GPL3921	Roessler et al., 2010

GEO, Gene Expression Omnibus; GPL, Gene Expression Omnibus Platform; GSE, Gene Expression Omnibus Series; T, tumor samples; N, paracancerous normal samples. There is no reference information in GSE84005. NA, not available.

Identification of Robust DEGs

The input data of WGCNA is usually less than 5000 genes. Therefore, preliminary screening of genes is required. In addition, DEGs (tumor vs normal tissue) can better reflect the differences in biological characteristics between tumors and normal liver tissues (Sarathi and Palaniappan, 2019).We employed "limma" (R package) to normalize and analyze the differences of each dataset downloaded from the GEO (HCC and normal samples) under a false discovery rate threshold (FDR) < 0.05 (Ritchie et al., 2015). Results from each dataset were ranked according to the fold change value of each gene. Next, "RobustRankAggreg" (R package) was implemented to analyze the results of the nine datasets for the identification of robust DEGs with adjusted *P*-values < 0.05 (Kolde et al., 2012).

Construction of the WGCNA Network and Enrichment Analysis of Key Modules

Weighted gene co-expression network analysis was used to identify modules highly correlated with clinical traits. We applied "WCGNA" (R package) to cluster all the robust DEGs identified from the GSE14520 HCC dataset with the largest sample size (471 HCC samples) and sufficient clinical information (Langfelder and Horvath, 2008). The resulting adjacency matrix was transformed into a topological overlap matrix (TOM). Differentially expressed genes were subsequently grouped into different modules based on the TOM-based dissimilarity measure. A soft-thresholding power of 7 (scale-free R = 0.90) and minimal module size of 30 were applied. The cut height was set as 0.4 to merge similar modules.

After clustering the genetic modules, key modules associated with clinicopathological variables were determined using Pearson's correlation coefficient, including age, hepatitis B virus (HBV) activity, alanine aminotransferase (ALT) level ($\leq$ and $>$50 U/L), primary tumor size ($\leq$ and $>$5 cm), multinodular characteristics, cirrhosis, tumor node metastasis (TNM) stage, Barcelona Clinic Liver Cancer (BCLC) stage, Cancer of the Liver Italian Program (CLIP) stage, AFP level ($\leq$ and $>$300 ng/mL), survival status, survival time (months), recurrence status, and recurrence time (months). We selected the modules that were highly correlated with clinical traits. To establish the biological functions of the key modules, R package "clusterprofiler" was applied to perform Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses on individual genes. *P*-values <0.05 were indicative of significant enrichment.

Identification of Hub Genes Based on WGNCA Combined With PPI and Construction of miRNA–mRNA Regulatory Networks

After the identification of the key modules, genes with gene significance (GS) > 0.3 and module membership (MM) > 0.8 were taken as core genes in WGCNA. Initially, the top 100 genes with high connectivity from each module were screened, of which the top 30 were marked as "hub genes in WGCNA". Next, we uploaded the top 100 connectivity genes to the STRING[3] database for PPI network analysis (Szklarczyk et al., 2017). The "TSV: tab separated values" file was downloaded in the "Exports" option and imported into the Cytoscape software (version 3.7.0), whereby the top 30 genes were screened as "hub genes in PPI" by "Degree" using the "cytoHubba" (Chin et al., 2014) app. GeneMANIA is a common tool for PPI network analysis and predicting the functions of preferred genes (Warde-Farley et al., 2010). The program displays genes or gene lists using bioinformatics methods, including gene co-expression, physical interactions, gene co-location, gene enrichment analysis, and website prediction. We observed the interaction types among the hub genes and visualized the gene networks with the aid of GeneMANIA. Finally, the intersecting results of both analytical methods were used to obtain hub genes, which were uploaded to miRNet[4] to generate a miRNA–mRNA regulatory network for establishing their potential functions.

Verification of Hub Genes

First, GEPIA2 was employed to visualize the differential expression of the hub genes between HCC and normal tissues (one-way ANOVA). Next, we used "ggpubr" (R package) to analyze the expression patterns at different TNM stages (Kruskal–Wallis test). Stage IV samples were excluded owing to a small size of less than five samples. In addition, we used the "survival" R package to perform Kaplan–Meier (K-M) survival or unique Cox regression analysis. Results were validated in the external verification dataset TCGA-LIHC.

GSEA and LASSO

To further explore the potential functions of the genes rarely reported in HCC, we utilized the "clusterprofiler" R package to perform GSEA for each gene. In the TCGA-LIHC dataset (normalized with the "edgeR" package), 374 HCC samples were used as the gene expression matrix. Gene lists were generated according to the order of correlation with the expression of each hub gene. The C2 reference gene sets were downloaded

[3]https://string-db.org/

[4]https://www.mirnet.ca/

from the Molecular Signatures Database (MSigDB)[5]. We set an adjusted *P*-value < 0.05 as the cut-off criterion. LASSO regression is widely used in the construction of clinical prediction models (Tibshirani, 1997). Next, "glmnet" (R package) was applied to verify the potential of these genes as biomarkers. GSE14520 was used as the training set and TCGA-LIHC as the test set for the LASSO regression analysis. Each cohort was divided into two groups according to the best cutoff risk score. Finally, results were visualized with K-M and ROC curves.

RESULTS

Overall Study Design

A flow chart of the study, divided into four steps, is presented in **Figure 1A**. Firstly, we used the RRA method to integrate and analyze the nine GEO datasets to obtain robust DEGs (Step 1). These DEGs were used to construct a WGCNA network using the GSE14520 dataset, and the key modules displaying a significant correlation with clinical traits were identified (Step 2). Hub genes were screened according to the WGCNA and PPI networks (Step 3). Finally, the hub genes were validated (Step 4).

[5]http://software.broadinstitute.org/gsea/msigdb/index.jsp

RRA-Based Identification of Robust DEGs Between HCC and Normal Tissues

A total of 4244 robust DEGs (2674 significantly upregulated and 1570 significantly downregulated) were identified from the nine datasets integrated using RRA (adjusted *P*-value < 0.05). As shown in **Figure 1B**, the 20 most significant DEGs were consistently identified among most of the datasets evaluated, signifying the robustness of the results. The majority of these genes are associated with HCC. For example, TOP2A displaying the most significant upregulation has been identified as a biomarker for HBV-related HCC (Liao et al., 2019) and APOF with the most significant downregulation is considered a tumor suppressor in HCC (Wang Y. B. et al., 2019). Significantly, AKR1B10 was not included in the GSE84005 dataset or NCAPG and LAPTM4B in the GSE45114 dataset. However, close association of these three genes with the progression of HCC has been recently reported (DiStefano and Davis, 2019; Gong et al., 2019; Wang F. et al., 2019). The RRA method effectively maximizes the retention of hub genes.

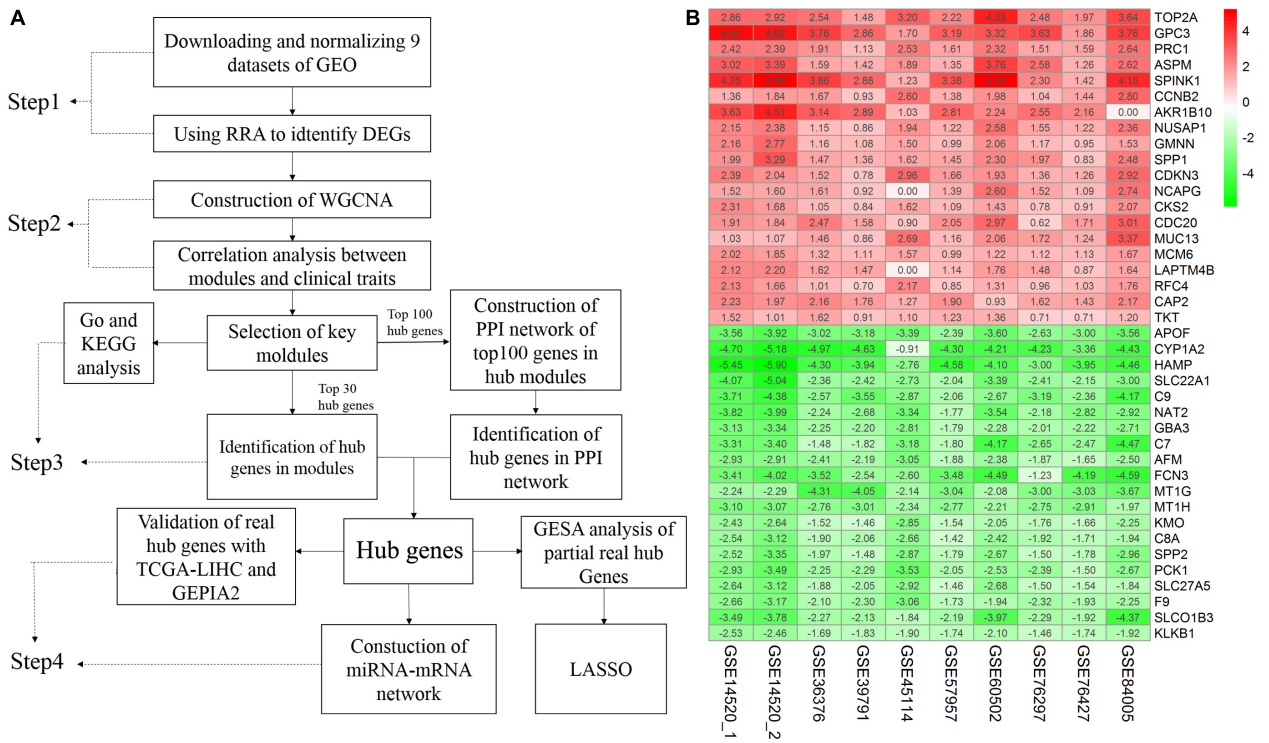


FIGURE 1 | Study workflow and heatmap of the top 20 robust DEGs. **(A)** Study workflow. GEO, Gene Expression Omnibus; RRA, Robust Rank Aggregation; DEGs, differentially expressed genes; WGCNA, weighted gene co-expression network analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology; PPI, protein– protein interaction. GSEA, Gene Set Enrichment Analysis. GEPIA2, gene expression profiling interactive analysis. LASSO, least absolute shrinkage and selection operator **(B)** Robust DEGs analyzed by RRA. The top 20 up and down regulated genes according to the fold change value of the RRA analysis are shown in this heatmap. The row names are DEGs. The column names are GEO datasets. The numbers in each cell represent logarithmic fold change in each dataset calculated by the "limma" R package. "0" indicates that the gene corresponding to the row is missing in the data set corresponding to the column. Red indicates that DEGs are upregulated in HCC samples, while green indicates the opposite. DEGs, differentially expressed genes; GEO: Gene Expression Omnibus; RRA: robust rank aggregation. GSE14520_1, dataset from GPL571, GSE14520_2, dataset from GPL3921.

Identification of Key Modules

To acquire the key modules, "WGCNA" (R package) was used to examine the co-expression network with the GSE14520 dataset. All DEGs derived from the RRA analysis were used as input. As shown in **Supplementary Figure 1A**, when the soft-thresholding power was 7 or 8, R^2 was >0.9 (red line). Here, a power of $\beta = 7$ (scale-free $R^2 = 0.9$) was selected as the soft-thresholding power to ensure a scale-free network. After applying threshold values, a total of eight modules were obtained for subsequent analysis (**Supplementary Figures 2C,D**). As determined from evaluation of module-trait relationships (**Figure 2A**), the brown and turquoise modules showed greater significance in relation

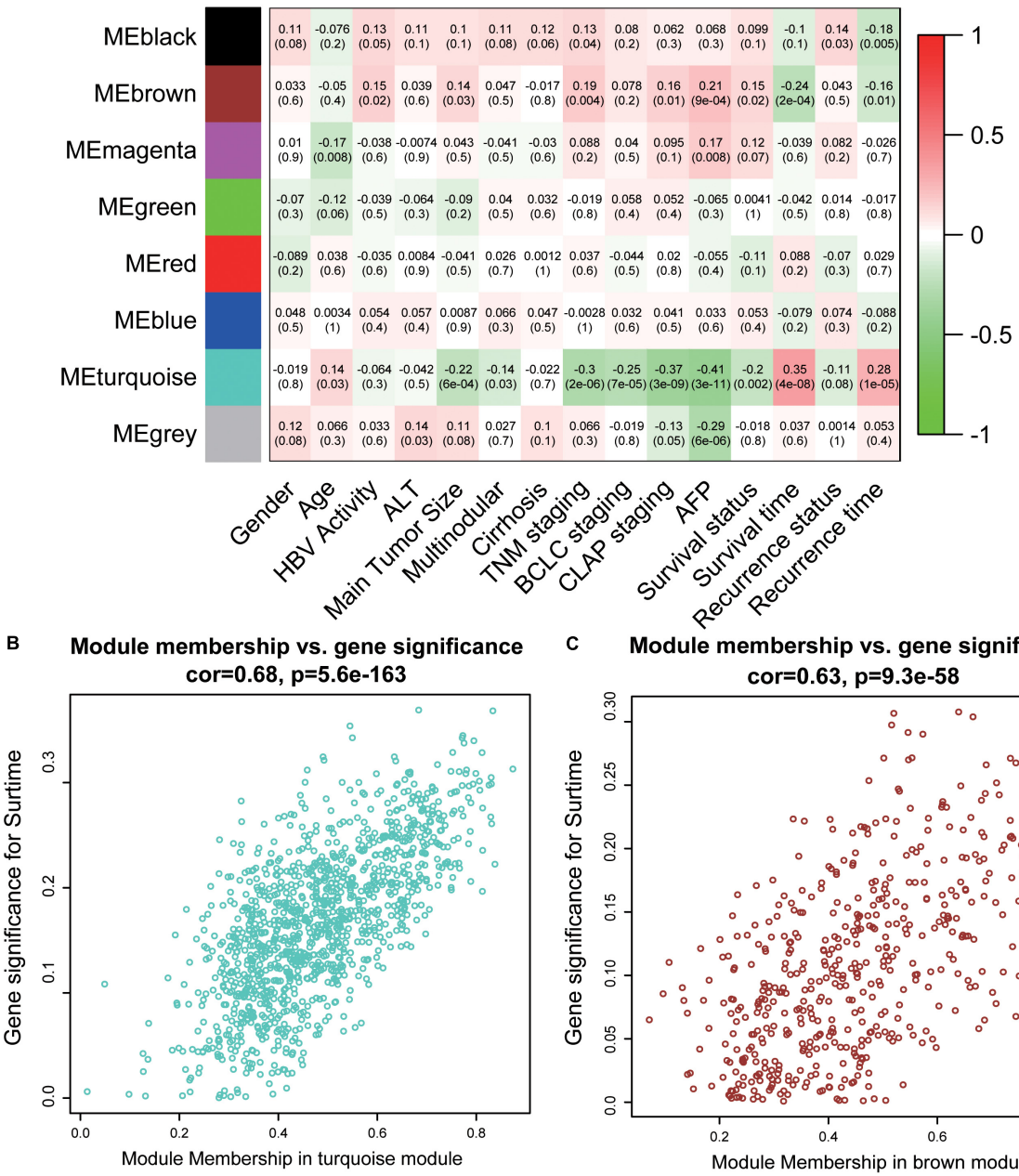


FIGURE 2 | Identification of key modules. **(A)** The heatmap shows the correlation between the genes module and clinical traits of HCC. Pearson's correlation coefficient between the gene modules and clinicopathological variables are shown, accompanied by the corresponding *P* value in brackets. Red represents positive correlation and green represents negative correlation. **(B)** The scatter plot of module eigengenes in the turquoise module. **(C)** The scatter plot of module eigengenes in the brown module. ALT, alanine aminotransferase; HBV, hepatitis B virus; TNM, tumor node metastasis; BCLC, Barcelona Clinic Liver Cancer; CLIP, Cancer of the Liver Italian Program.

to clinical information, compared with the other modules, in particular, main tumor size, TNM stage, and AFP level critical for prognosis of HCC patients (Han et al., 2014; Zhang et al., 2016) (**Figures 2B,C**).

Functional Enrichment Analysis of Genes Within the Key Modules and Identification of Hub Genes

To clarify the functions of genes from the two modules, we performed separate GO and KEGG analyses. In the brown module, "DNA replication," "cell cycle," "p53 signaling pathway," and "cellular senescence" were enriched in the KEGG pathway analysis (**Supplementary Figure 2A**) while in the turquoise module, "drug metabolism – cytochrome P450" and "chemical carcinogenesis" were enriched (**Supplementary Figure 2B**). These findings were consistent with previous studies reporting the involvement of the above functions in tumorigenesis of HCC. For example, Xie et al. (2018) showed that DNA replication is associated with tumor cell proliferation and prognosis of patients with HCC. Moreover, genetic variations in cell cycle pathway genes affect the disease-free survival of patients with HCC (Liu et al., 2017). The TP53 mutation is considered one of the molecular mechanisms of HCC pathogenesis (Hussain et al., 2007). Abnormal cellular senescence is a characteristic phenotype of various cancers (Chen S. L. et al., 2019). Cytochrome P450 is severely damaged and dysregulated in HCC (Yan et al., 2015). The collective findings validate the functional association of the key modules in this study with HCC. The significant biological process (BP), cellular component (CC), and molecular function (MF) GO terms of the two modules are presented in **Supplementary Tables 1–6**.

To further screen for the most significant hub genes, we used a combination of two methods (WGCNA and PPI networks, see section "Materials and Methods"). The PPI network of the top 100 connectivity genes from the brown module is shown in **Figure 3A**. According to degree (high to low), the positions of genes are arranged from the inside to outside, and the top 30 considered "hub genes in PPI". Interaction analysis of hub genes in PPI was further performed using GeneMANIA to clarify the correlations among colocalization, shared protein domains, co-expression, prediction, and pathways. As revealed by the protein–protein interaction network generated with GeneMANIA (**Figure 3C**), co-expression interactions accounted for the largest proportion (83.83%), consistent with the results of WGCNA. "Hub genes in WGCNA" and their correlated expression levels are shown in **Figure 3B**. The hub genes were obtained by selecting the intersecting results with the two methods (**Figure 3D**). The hub genes of the turquoise module were obtained with the same method (**Supplementary Figures 3A–D**). Overall, we identified a total of 29 core genes from the two key modules.

Construction of the miRNA–mRNA Regulatory Network

Interactions between miRNA and mRNA are an increasing focus of research attention. To further explore the functions of hub genes from a global perspective, a miRNA–mRNA regulatory network was constructed via miRNet (**Figure 4**). Previous studies suggest that a number of these miRNAs are related to HCC. For example, exosome hsa-miR-335 was identified as a therapeutic target for HCC (Wang F. et al., 2018). Furthermore, according to web-based KEGG analysis, this network is enriched in multiple tumor-related pathways (**Supplementary Table 7**), such as cell cycle and p53 signaling (Hussain et al., 2007; Sanchez-Vega et al., 2018; Ikeno et al., 2019). Thus our group of hub genes may play important roles in HCC through the miRNA–mRNA regulatory network.

Verification of Hub Genes Based on the TCGA-LIHC Dataset

In total, 29 hub genes were obtained. Interestingly, TOP2A was consistently ranked first. Ten of the genes were filtered from the turquoise module (**Figure 3C**), which were further verified in TCGA-LIHC and GEPIA2 based on three parameters: (1) differential expression (HCC sample vs paracancerous sample), (2) TMN staging, and (3) survival analysis. In terms of expression, hub genes from the turquoise module were downregulated in HCC relative to normal samples. Notably, F13B was excluded due to lack of statistical significance. These results were validated using an external dataset (**Figure 5A** and **Supplementary Figure 4A**). Additionally, genes were differentially expressed in HCC samples with different TNM stages to a significant extent. A higher expression of these genes was correlated with an earlier TNM stage (**Figure 5B** and **Supplementary Figure 4B**). Survival analysis revealed an association of low expression of these genes with poor prognosis (**Figure 5C** and **Supplementary Table 8**). Using the same method, hub genes of the brown module (**Supplementary Figure 3D**) were validated, which showed an opposite trend to genes of the turquoise module (**Supplementary Figures 5, 6** and **Supplementary Table 8**). Our collective data support critical roles of 28 of the 29 hub genes in HCC.

GSEA of Tumor Suppressor Roles of Hub Genes

The majority of the hub genes for HCC have already been reported (**Supplementary Table 9**). However, DAO, SLC27A2, GYS2, HAO1, and PCK2 have not been previously studied in association with HCC. To analyze their potential functions in HCC, we performed GSEA on TCGA-LIHC RNA sequencing data. As shown in **Supplementary Figure 7**, three gene sets associated with tumors were defined. In samples showing a significant negative correlation of HAO1 and SLC27A2 expression with HCC, "epithelial-mesenchymal transition" (EMT) and "PI3K/Akt/mTOR" were enriched. "Wnt/beta-catenin signaling" and "MYC target v1" were significantly enriched in samples showing a negative correlation of PCK2 and DAO expression with HCC. The gene set "DNA repair" was enriched in samples showing negative correlation of ABAT and SLC27A2 expression with HCC. These mechanisms are typical tumor-associated pathways. For instance, EMT is reported to coordinate the occurrence of liver fibrosis, carcinogenesis, and proliferation and invasion of HCC cells

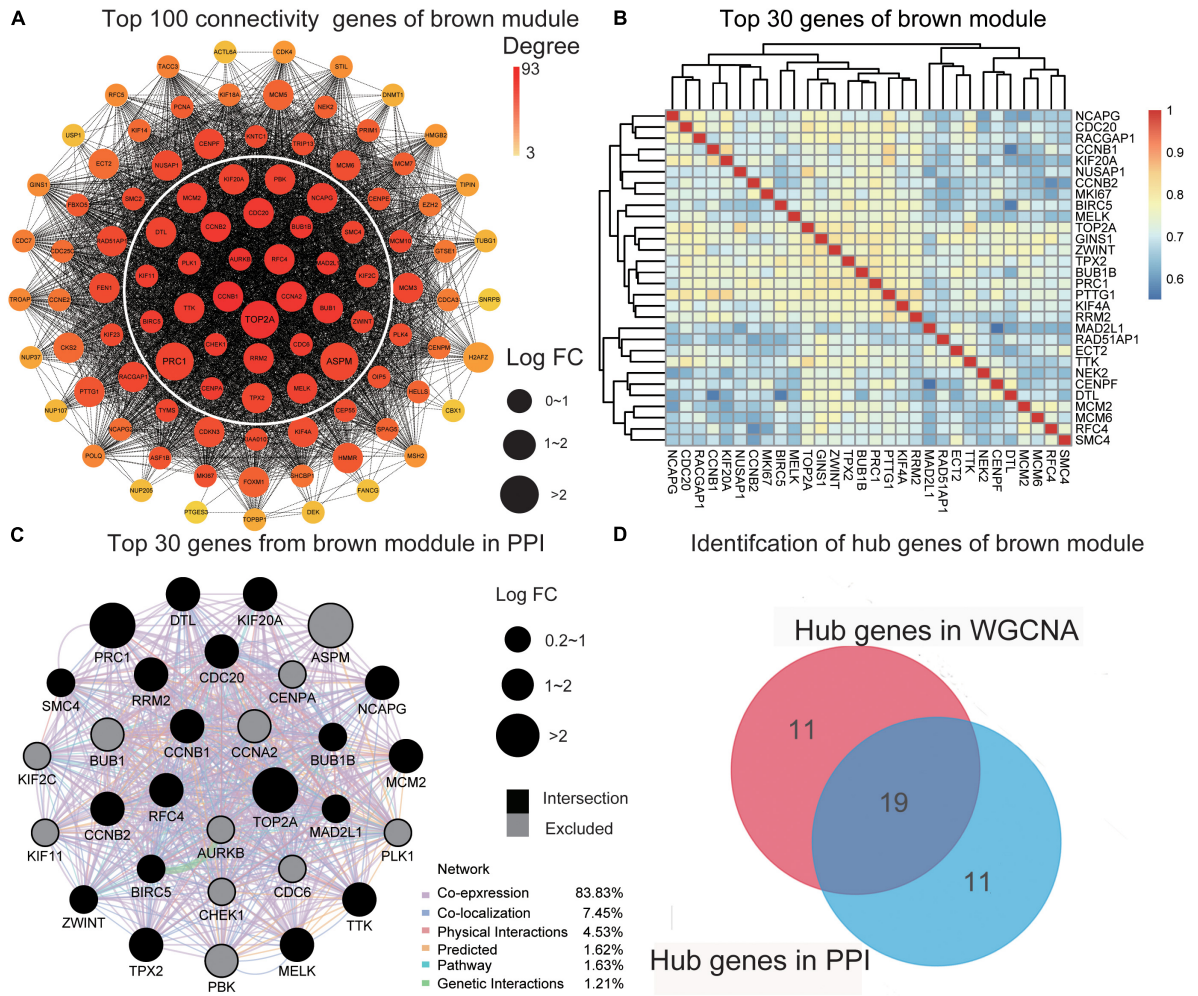


FIGURE 3 | Identification of hub genes. **(A)** The PPI network of the top 100 connectivity genes from the brown module. According to degree from high to low, the gene is arranged from the center to the edge. "Hub genes in PPI" is inside the white circle. **(B)** The top 30 genes gained in WGCNA from the brown module by setting MM) > 0.8 and GS > 0.3. The correlation between these genes is shown. **(C)** The PPI network (GeneMANIA) of the top 30 genes in the brown module. Different colors of the network edge indicate the bioinformatics methods applied: physical interactions, co-expression, predicted, co-localization, pathway, genetic interactions, and shared protein domains. The size of each node indicates the value of LogFC according to the result of RRA. The black nodes present the intersection of hub genes in PPI and WCGNA. **(D)** The hub genes in the brown module are selected by PPI network and co-expression network. PPI, protein–protein interaction; MM, module membership; GS, gene significance. LogFC, log fold change; RRA, Robust Rank Aggregation. WGCNA, weighted gene co-expression network analysis.

(Giannelli et al., 2016). The activation of PI3K/AKT signaling has been shown to promote EMT (Liu et al., 2018). "Wnt/beta-catenin signaling", "MYC target v1", and "DNA repair" are closely related to tumorigenesis and the development of HCC (Dolezal et al., 2017; Dimri and Satyanarayana, 2020; Pardini et al., 2020). Taken together, the findings clearly suggest that these genes are closely associated with the mechanisms underlying HCC cell proliferation.

Construction of the Novel Hub Gene Signature for Survival Prediction

Finally, we included the above five hub genes in the LASSO regression analysis to construct a survival prediction model for HCC patients. GSE14520 was used as the training set to generate a prediction model comprising three of the genes, specifically, OSPCK2, DAO, and HAO1. The formula for calculating the prognostic risk score was as follows: (−0.0179 × expression HAO1) + (−0.0221 × expression PCK2) + (−0.1209 × expression DAO). The results of this scoring system were depicted using a K–M curve (**Figures 6A,B**). The high-risk group had shorter OS, both in the training (P = 0.002) and test (P < 0.001) datasets. In addition, we generated time-dependent ROC curves to evaluate the predictive effects of the three-gene signature based on the area under the curve (AUC) value. In the training cohort, one-year and three-year AUC values were 0.673 and 0.605, respectively. In the verification cohort, AUC for one year was 0.605 and that for

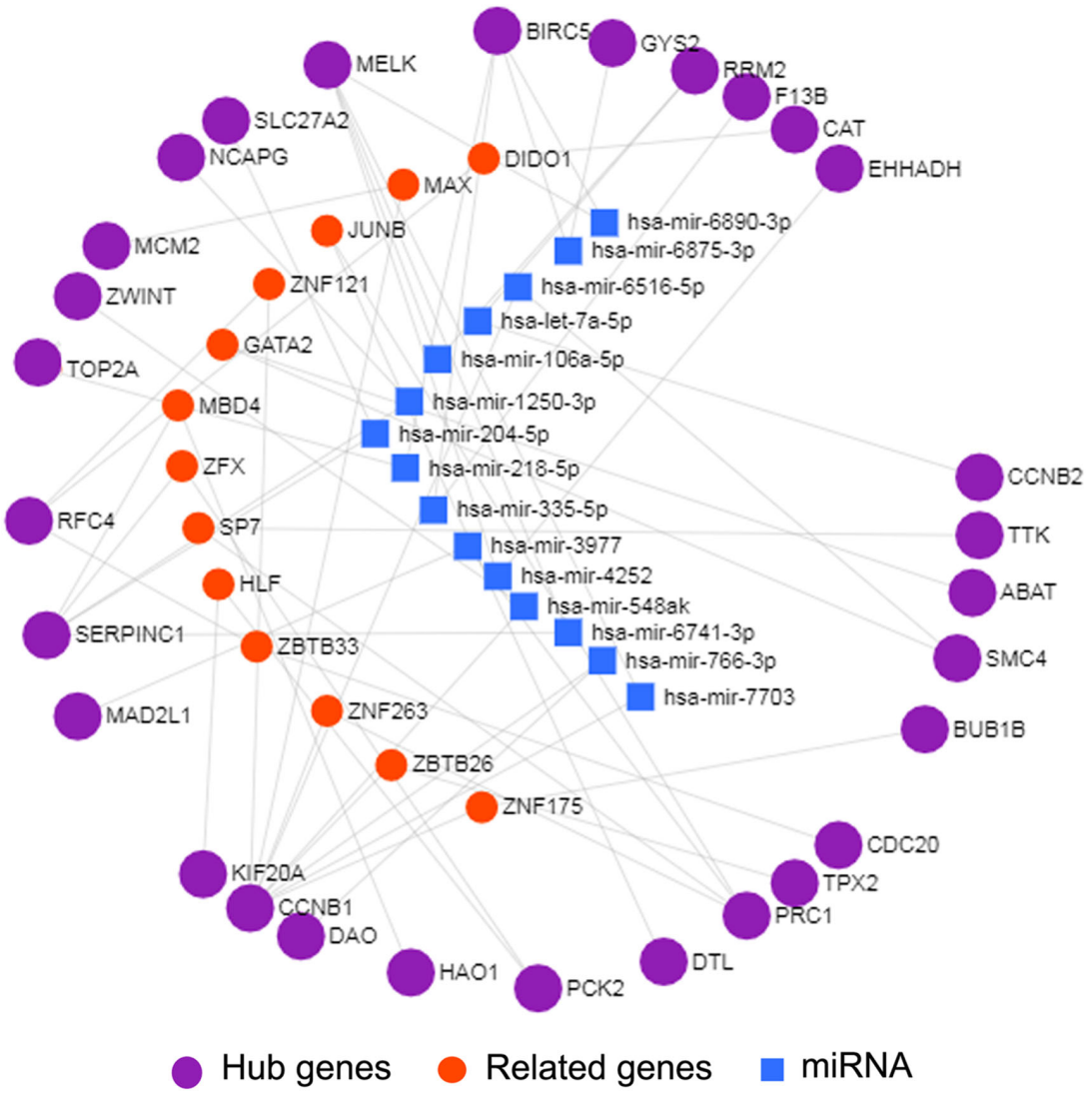


FIGURE 4 | miRNA-mRNA network of 29 hub genes by miRNet. The purple spheres represent 29 hub genes; the blue squares represent the miRNA associated with the hub genes, and the red spheres represent the genes associated with the hub genes.

three years was 0.672 (**Figures 6C,D**). Based on the results, we propose that this novel three-gene signature can serve as a reliable predictor of OS in HCC patients.

DISCUSSION

In this study, we used multiple bioinformatics methods to establish the biological mechanisms of HCC. To avoid the potential bias caused by DEGs in a single database, numerous studies have focused on evaluating multiple datasets (Xu et al., 2016). In the process of merging data, gene symbols that are not detected in only one dataset may be lost. For example, as shown in **Figure 1B**, AKR1B10, NCAPG, and LAPTM4B exist in multiple datasets and would therefore be lost if the datasets were simply merged. However, these genes are closely related to the progression of HCC (DiStefano and Davis, 2019; Gong et al., 2019; Wang F. et al., 2019). Furthermore, in dataset GSE39791, logFC values of some of the top 20 DEGs were less than 1. However, in combination with other datasets, the RRA method suggests that these genes are robust DEGs. Potential bias of results due to inclusion of only one dataset should be avoided. In the current investigation, RRA was applied to analyze nine groups of datasets to minimize bias, avoid missing hub genes, and obtain the most robust DEGs.

Weighted gene co-expression network analysis is based on the correlation between modules and clinical features, and the screening results are highly reliable and biologically meaningful (Yin et al., 2018). To our knowledge, the current study is the first to combine the RRA method with WGCNA for efficient identification of hub genes associated with HCC. Among the

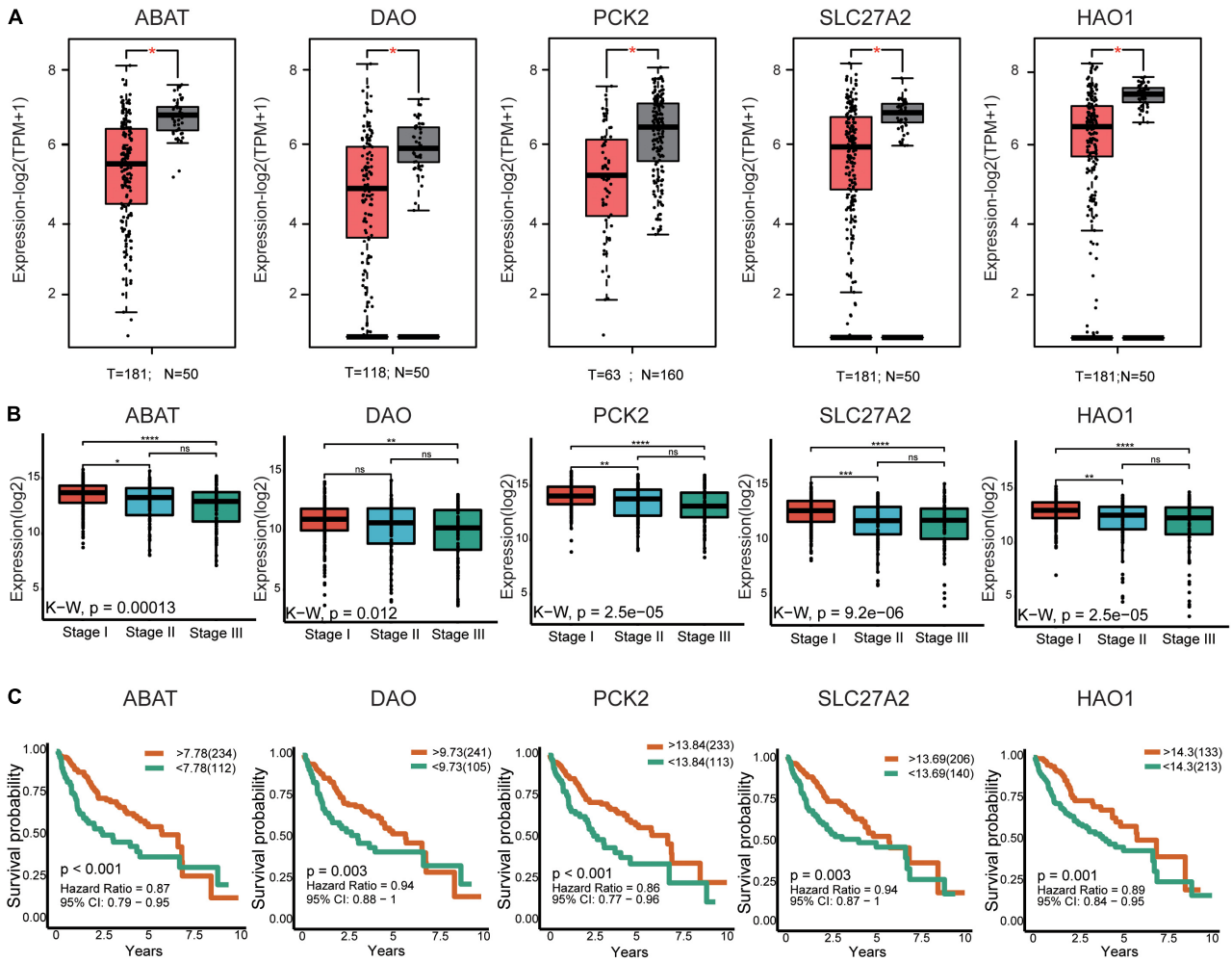


FIGURE 5 | External validation of the partial hub genes in the turquoise module. **(A)** Partial hub genes, rarely reported on HCC, expression differences between HCC and adjacent normal tissues in GPEIA 2. ABAT, DAO, PCK2, SLC27A2, and HAO are downregulated in HCC tissues. "*" represents *P* value < 0.05. **(B)** Expression of ABAT, DAO, PCK2, SLC27A2, and HAO1 in HCC samples with different TNN stages. The lower the expression level of these genes indicates the more advanced stage of HCC. "*" represents *P* value < 0.05; "**" represents *P* value < 0.01; "****" represents *P* value < 0.001; "*****" represents *P* value < 0.0001. **(C)** The association between ABAT, DAO, PCK2, SLC27A2, and HAO1 expression and overall survival time in the TCGA-LIHC dataset. The yellow line indicates high expression groups and the green line represents the low expression group. *T*, number of HCC samples, *N*, number of normal samples.

eight gene modules, brown and turquoise modules were closely related to clinical characteristics, such as primary tumor size, AFP level, TNM stage, and overall survival time. In addition, GO and KEGG analyses showed enrichment of both modules in multiple tumor-related pathways. For instance, DNA replication is associated with tumor cell proliferation and the prognosis of HCC (Xie et al., 2018), variations in cell cycle pathway genes affect disease-free survival of patients with HCC (Liu et al., 2017), TP53 mutation is considered one of the critical molecular mechanisms of HCC pathogenesis (Hussain et al., 2007), an abnormal cellular senescence phenotype is observed in various cancer types (Chen S. L. et al., 2019), and cytochrome P450 is severely damaged and dysregulated in HCC (Yan et al., 2015).

Next, we combined co-expression and PPI networks to screen for hub genes. After a series of strict screening steps, 29 hub genes (10 from the turquoise module and 19 from the brown module) were isolated. To explore the functions of this group of genes from the global network, miRNA–mRNA regulatory networks were generated using miRNet (**Figure 4**). As shown in **Supplementary Table 7**, specific pathways, such as cell cycle, and p53 signaling, were highlighted, both of which are closely related to tumorigenesis and the development of HCC (Hussain et al., 2007; Sanchez-Vega et al., 2018; Ikeno et al., 2019). Importantly, we used TCGA-LIHC, a dataset containing 374 HCC samples, to validate the predictive power of these hub genes in the progression and prognosis of HCC. Among the genes examined, only one (F13B) failed verification.

The involvement of the majority of these genes in HCC has been confirmed in earlier experiments (**Supplementary Table 9**), supporting the efficacy of our screening strategy. Among the hub genes, TOP2A, RFC, and the CCMB family have received considerable research attention. DNA topoisomerase

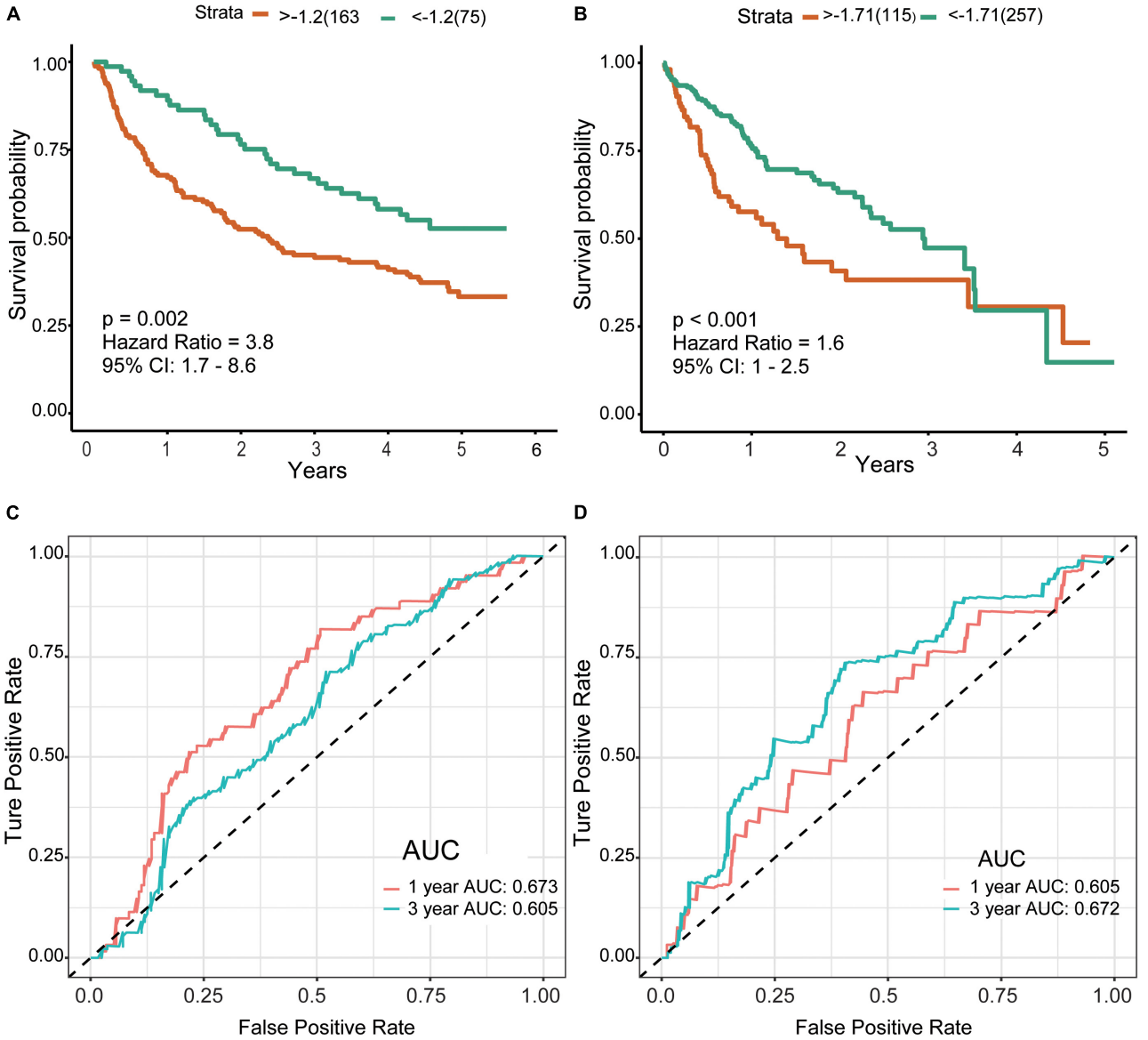


FIGURE 6 | Risk score and survival analysis in training and validation datasets.**(A,B)** K–M curves for patients in GSE14520 **(A)** and TCGA-LIHC **(B)** datasets divided into high- and low-risk groups by the best cutoff values. Patients of a high risk group exhibited poorer prognosis in both cohorts. **(C,D)** ROC curves showed the predictive risk signature for patients in GSE14520 **(C)** and TCGA-LIHC **(D)** datasets on the survival rate. K–M, Kaplan–Meier; ROC, receiver operating characteristic, AUC, area under curve.

II alpha (TOP2A) is abundantly expressed in testis, lymph node tissues, and a variety of tumor tissues, including liver cancer. Several bioinformatics analyses have validated TOP2A as a biomarker for HCC, in particular, HBV-related HCC (Liao et al., 2019). Panvichian et al. (2015) reported overexpression of TOP2A in 72.5% of tumor tissues and its significant association with the hepatitis B surface antigen (HBsAg) in serum. In addition, results of a phase III prospective randomized study showed that TOP2A is associated with the histological grade of liver cancer, microvascular invasion, early onset of malignant tumors (≤40 years), and chemotherapy resistance (Wong et al., 2009). Replication factor C subunit 4 (RFC4) has recently been identified as a hub gene affecting prognosis of patients with HCC (Kong et al., 2019). The knockdown of endogenous RFC4 suppresses HCC cell growth and enhances the chemosensitivity of HepG2 cells (Arai et al., 2009). Cyclin B1 (CCNB1) and TOP2A are considered key genes for early diagnosis of HCC (Wu et al., 2019).

Interestingly, DAO, ABAT, SL27AL, PCK2, and HAO1, all from the turquoise module, have not been shown to be associated with HCC to date, either *in vivo* or *in vitro*. However, several studies support inhibitory roles of these genes in other tumors. The peroxisomal enzyme D-amino acid oxidase (DAO) is highly expressed in the kidney, liver, and brain in mammals

(Fang et al., 2008) and plays a critical role in the pathophysiology of schizophrenia (Liu et al., 2016). Earlier reports suggest that DAO inhibits glioma cell growth by inhibiting angiogenesis (El Sayed et al., 2012) and inducing apoptosis (Li et al., 2008). 4-Aminobutyrate aminotransferase (ABAT) is mainly responsible for decomposing γ-aminobutyric acid (GABA), an inhibitory neurotransmitter, into succinic semialdehyde. In basal-like breast cancer (BLBC) cells, GABA increases the intracellular Ca^{2+} concentration and effectively activates nuclear factor 1-4 (NFAT1). Consequently, ABAT expression inhibits the tumorigenicity and metastasis of BLBC cells *in vitro* and *in vivo*. Conversely, the downregulation of ABAT promotes the progression of BLBC (Chen X. et al., 2019) and resistance to endocrine therapy of inflammatory breast cancer (Jansen et al., 2015). Moreover, ABAT has been identified as a prognostic factor for renal cell carcinoma and hepatic adenocarcinoma (Reis et al., 2015; Lu et al., 2020). Chen et al. (2017) screened six genes related to HCC metastasis and prognosis through a co-expression network analysis, which led to the identification of DAO and ABAT. However, their mechanisms of action in HCC have not been clarified. Solute carrier family 27 member 2 (SLC27A2), also designated FATP2, improves the efficiency of cancer therapy by inhibiting the activity of polymorphonuclear myeloid-derived suppressor cells (PMN-MDSCs) (Veglia et al., 2019). Phosphoenolpyruvate carboxykinase 2 (PCK2) encodes a key mitochondrial enzyme for gluconeogenesis in the liver. The overexpression of PCK2 inhibits melanoma cell growth *in vitro* and prevents tumorigenesis *in vivo* (Luo et al., 2017). More recent experiments have demonstrated an association of decreased PCK2 expression with metastasis and the recurrence of osteosarcoma (Zhang et al., 2019). Upon suppression of autophagy, levels of glucose-6-phosphatase (G6Pase) and phosphoenolpyruvate carboxykinase (PEPCK, a protein encoded by PCK2) are reduced in the human HCC cell line HepG2 (Jeon et al., 2015). Hypoxia-inducible factor 1α (HIF-1α) can promote the growth of human breast tumor-repopulating cells by downregulating PCK2 (Tang et al., 2019). However, a number of studies have reported that PEPCK coordinates the regulation of central carbon metabolism to promote tumor cell growth (Montal et al., 2015). Therefore, the biological characteristics of PCK2 in HCC requires further investigation. Hydroxyacid oxidase 1 (HAO1) is expressed mainly in the liver and pancreas. An earlier genome-wide association study in Korea showed that a single nucleotide polymorphism in HAO1 is one of the risk factors for childhood acute lymphoblastic leukemia (Han et al., 2010).

In our study, GSEA consistently supported the tumor suppressor roles of these genes in multiple carcinogenic pathways in HCC datasets. Further research is warranted to establish the mechanisms of action of these genes in HCC. The collective evidence to date suggests that these genes play a suppressive roles in the biological processes of tumors. In addition, the clinical prediction model generated using a three-gene signature showed efficacy in predicting the survival of patients with HCC and the potential as a robust biomarker. Our study has some limitations, such as the fact that the nine datasets of the training set are all microarrays and lack RNA-seq datasets. The data diversity is insufficient.

CONCLUSION

Systematic analysis of the genes involved in pathogenesis of HCC using a novel integrated strategy led to the identification of two risk modules and several representative hub genes. Among these, HAO1, SCL27A2, DAO, ABAT, and PCK2, rarely reported in HCC to date, were validated as novel hub genes that may serve as effective clinical diagnostic and prognostic markers as well as therapeutic targets for HCC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The datasets analyzed in the current study are available in the TCGA repository (http://cancergenome.nih.gov/) and GEO (https://www. ncbi.nlm.nih.gov/geo/).

AUTHOR CONTRIBUTIONS

YY contributed to the project design. CN and CZ revised and verified the article. ND, SL, JL, and AX provided the bioinformatics method support and reviewed the manuscript. HS analyzed the data and wrote the article. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by findings from the National Natural Scientific Foundation of China (No. 31701145).

ACKNOWLEDGMENTS

In this study, the authors express their gratitude to the scholars and institutions that submitted the raw datasets.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00895/full#supplementary-material

Supplementary Figure 1 | Determination of soft-thresholding power and cut height in the WGCNA. **(A)** Analysis of scale-free index (left) and mean connectivity (right) for different soft-thresholding power (β) red line indicates signed R^2 = 0.9. **(B)** When β = 7 histogram of connectivity distribution (left) and scale-free topology R^2 = 0.9 (right). **(C)** Clustering of module eigengenes. Set the cut height as 0.4 (red line) to merge similar modules. **(D)** Dendrogram of all DEGs clustered based on a dissimilarity measure (1-TOM). Each color represents a set of gene modules.

Supplementary Figure 2 | The screening of hub genes in the turquoise module. **(A)** The PPI network of top 100 connectivity genes from the brown module. "Hub

genes in PPI" is inside the black circle. **(B)** The top 30 hub genes gained in WGCNA from the brown module by setting MM) > 0.8 and GS > 0.3. Correlation between these genes is shown. **(C)** PPI network (GeneMANIA) of the top 30 genes in the brown module. **(D)** Selection of hub genes that occur in both the PPI network and WGCNA. PPI, protein–protein interaction; MM, module membership; GS, gene significance. WGCNA, weighted gene co-expression network analysis.

Supplementary Figure 3 | KEGG analysis for the key modules. **(A)** Brown module. **(B)** Turquoise module. KEGG, Kyoto Encyclopedia of Genes and Genomes.

Supplementary Figure 4 | External validation of the rest of the hub genes in the turquoise module. **(A)** The rest of the hub genes in the turquoise module expression differences between HCC and adjacent normal tissues in GPEIA2. "*" represents *P* value I < 0.05. **(B)** Expression of CAT, F13B, EHHADH, GYC2, and SERPINC1 in HCC samples with different TNN stages. "*" represents *P* value < 0.05; "**" represents *P* value I < 0.01; "***" represents *P* value < 0.001.

Supplementary Figure 5 | External validation of hub genes in the brown module. The hub genes from the brown module expression differences between HCC and adjacent normal tissues in GPEIA2. "*" represents *P* value I < 0.05.

Supplementary Figure 6 | External validation of hub genes in the brown module. Expression of these genes in HCC samples with different TNN stages. "*" represents *P* value < 0.05; "**" represents *P* value I < 0.01; "***" represents *P* value < 0.001.

Supplementary Figure 7 | Gene sets related to cancer. Results of GSEA related to cancer in samples negatively correlated with PCK2 **(A)**, ABAT **(B)**, HAO1 **(C)**, SLC27A2 **(D)**, and DAO **(E)** expression. Highlight 3 gene sets for each gene.

Supplementary Figure 8 | Risk score distribution, survival status, and heatmaps for patients in the GSE14520 **(A)** and TCGA-LIHC **(B)** datasets divided into high- and low-risk groups.

Supplementary Figure 9 | Node degree distribution plot of differentially expressed genes.

Supplementary Table 1 | BP of GO analysis for brown module.

Supplementary Table 2 | CC of GO analysis for brown module.

Supplementary Table 3 | MF of GO analysis for brown module.

Supplementary Table 4 | BP of GO analysis for turquoise module.

Supplementary Table 5 | CC of GO analysis for turquoise module.

Supplementary Table 6 | MF of GO analysis for turquoise module.

Supplementary Table 7 | Enriched function of the miRNA-mRNA network by miRNet.

Supplementary Table 8 | Survival analysis for the rest of the hub genes.

Supplementary Table 9 | Biological functions of the hub genes in HCC.

REFERENCES

Arai, M., Kondoh, N., Imazeki, N., Hada, A., Hatsuse, K., Matsubara, O., et al. (2009). The knockdown of endogenous replication factor C4 decreases the growth and enhances the chemosensitivity of hepatocellular carcinoma cells. *Liver Int.* 29, 55–62. doi: 10.1111/j.1478-3231.2008.01792.x

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Budhu, A., Forgues, M., Ye, Q. H., Jia, H. L., He, P., Zanetti, K. A., et al. (2006). Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. *Cancer Cell* 10, 99–111. doi: 10.1016/j.ccr.2006.06.016

Chaisaingmongkol, J., Budhu, A., Dang, H., Rabibhadana, S., Pupacdi, B., Kwon, S. M., et al. (2017). Common molecular subtypes among asian hepatocellular carcinoma and cholangiocarcinoma. *Cancer Cell* 32, 57.e3–70.e3. doi: 10.1016/j.ccell.2017.05.009

Chen, P., Wang, F., Feng, J., Zhou, R., Chang, Y., Liu, J., et al. (2017). Co-expression network analysis identified six hub genes in association with metastasis risk and prognosis in hepatocellular carcinoma. *Oncotarget* 8, 48948–48958. doi: 10.18632/oncotarget.16896

Chen, S. L., Zhang, C. Z., Liu, L. L., Lu, S. X., Pan, Y. H., Wang, C. H., et al. (2019). A GYS2/p53 negative feedback loop restricts tumor growth in HBV-related hepatocellular carcinoma. *Cancer Res.* 79, 534–545. doi: 10.1158/0008-5472.CAN-18-2357

Chen, X., Cao, Q., Liao, R., Wu, X., Xun, S., Huang, J., et al. (2019). Loss of ABAT-mediated GABAergic system promotes basal-like breast cancer progression by activating Ca2+-NFAT1 axis. *Theranostics* 9, 34–47. doi: 10.7150/thno.29407

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8(Suppl. 4):S11. doi: 10.1186/1752-0509-8-S4-S11

Dimri, M., and Satyanarayana, A. (2020). Molecular signaling pathways and therapeutic targets in hepatocellular carcinoma. *Cancers* 12:491. doi: 10.3390/cancers12020491

DiStefano, J. K., and Davis, B. (2019). Diagnostic and prognostic potential of akr1b10 in human hepatocellular carcinoma. *Cancers* 11:486. doi: 10.3390/cancers11040486

Dolezal, J. M., Wang, H., Kulkarni, S., Jackson, L., Lu, J., Ranganathan, S., et al. (2017). Sequential adaptive changes in a c-Myc-driven model of hepatocellular carcinoma. *J. Biol. Chem.* 292, 10068–10086. doi: 10.1074/jbc.M117.782052

El Sayed, S. M., El-Magd, R. M., Shishido, Y., Yorita, K., Chung, S. P., Tran, D. H., et al. (2012). D-Amino acid oxidase-induced oxidative stress, 3-bromopyruvate and citrate inhibit angiogenesis, exhibiting potent anticancer effects. *J. Bioenerg. Biomembr.* 44, 513–523. doi: 10.1007/s10863-012-9455-y

Fang, J., Deng, D., Nakamura, H., Akuta, T., Qin, H., Iyer, A. K., et al. (2008). Oxystress inducing antitumor therapeutics via tumor-targeted delivery of PEG-conjugated D-amino acid oxidase. *Int. J. Cancer* 122, 1135–1144. doi: 10.1002/ijc.22982

Giannelli, G., Koudelkova, P., Dituri, F., and Mikulits, W. (2016). Role of epithelial to mesenchymal transition in hepatocellular carcinoma. *J. Hepatol.* 65, 798–808. doi: 10.1016/j.jhep.2016.05.007

Gong, C., Ai, J., Fan, Y., Gao, J., Liu, W., Feng, Q., et al. (2019). NCAPG promotes the proliferation of hepatocellular carcinoma through PI3K/AKT signaling. *Onco. Targets Ther.* 12, 8537–8552. doi: 10.2147/OTT.S217916

Grinchuk, O. V., Yenamandra, S. P., Iyer, R., Singh, M., Lee, H. K., Lim, K. H., et al. (2018). Tumor-adjacent tissue co-expression profile analysis reveals pro-oncogenic ribosomal gene signature for prognosis of resectable hepatocellular carcinoma. *Mol. Oncol.* 12, 89–113. doi: 10.1002/1878-0261.12153

Han, J. H., Kim, D. G., Na, G. H., Kim, E. Y., Lee, S. H., Hong, T. H., et al. (2014). Evaluation of prognostic factors on recurrence after curative resections for hepatocellular carcinoma. *World J. Gastroenterol.* 20, 17132–17140. doi: 10.3748/wjg.v20.i45.17132

Han, S., Lee, K. M., Park, S. K., Lee, J. E., Ahn, H. S., Shin, H. Y., et al. (2010). Genome-wide association study of childhood acute lymphoblastic leukemia in Korea. *Leuk. Res.* 34, 1271–1274. doi: 10.1016/j.leukres.2010.02.001

Hussain, S. P., Schwank, J., Staib, F., Wang, X. W., and Harris, C. C. (2007). TP53 mutations and hepatocellular carcinoma: insights into the etiology and pathogenesis of liver cancer. *Oncogene* 26, 2166–2176. doi: 10.1038/sj.onc.1210279

Ikeno, S., Nakano, N., Sano, K., Minowa, T., Sato, W., Akatsu, R., et al. (2019). PDZK1-interacting protein 1 (PDZK1IP1) traps Smad4 protein and suppresses transforming growth factor-β (TGF-β) signaling. *J. Biol. Chem.* 294, 4966–4980. doi: 10.1074/jbc.RA118.004153

Jansen, M. P., Sas, L., Sieuwerts, A. M., Van Cauwenberghe, C., Ramirez-Ardila, D., Look, M., et al. (2015). Decreased expression of ABAT and STC2 hallmarks ER-positive inflammatory breast cancer and endocrine therapy resistance in advanced disease. *Mol. Oncol.* 9, 1218–1233. doi: 10.1016/j.molonc.2015.02.006

Jeon, J. Y., Lee, H., Park, J., Lee, M., Park, S. W., Kim, J. S., et al. (2015). The regulation of glucose-6-phosphatase and phosphoenolpyruvate carboxykinase by autophagy in low-glycolytic hepatocellular carcinoma cells. *Biochem. Biophys. Res. Commun.* 463, 440–446. doi: 10.1016/j.bbrc.2015.05.103

Kim, J. H., Sohn, B. H., Lee, H. S., Kim, S. B., Yoo, J. E., Park, Y. Y., et al. (2014). Genomic predictors for recurrence patterns of hepatocellular carcinoma: model derivation and validation. *PLoS Med.* 11:e1001770. doi: 10.1371/journal.pmed.1001770

Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 28, 573–580. doi: 10.1093/bioinformatics/btr709

Kong, J., Wang, T., Zhang, Z., Yang, X., Shen, S., and Wang, W. (2019). Five core genes related to the progression and prognosis of hepatocellular carcinoma identified by analysis of a coexpression network. *DNA Cell Biol.* 38, 1564–1576. doi: 10.1089/dna.2019.4932

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Li, C., and Xu, J. (2020). Identification of potentially therapeutic target genes of hepatocellular carcinoma. *Int. J. Environ. Res. Public Health* 17:1053. doi: 10.3390/ijerph17031053

Li, H., Wei, N., Ma, Y., Wang, X., Zhang, Z., Zheng, S., et al. (2020). Integrative module analysis of HCC gene expression landscapes. *Exp. Ther. Med.* 19, 1779–1788. doi: 10.3892/etm.2020.8437

Li, J., Shen, Y., Liu, A., Wang, X., and Zhao, C. (2008). Transfection of the DAAO gene and subsequent induction of cytotoxic oxidative stress by D-alanine in 9L cells. *Oncol. Rep.* 20, 341–346.

Liao, X., Yu, T., Yang, C., Huang, K., Wang, X., Han, C., et al. (2019). Comprehensive investigation of key biomarkers and pathways in hepatitis B virus-related hepatocellular carcinoma. *J. Cancer* 10, 5689–5704. doi: 10.7150/jca.31287

Lim, H. Y., Sohn, I., Deng, S., Lee, J., Jung, S. H., Mao, M., et al. (2013). Prediction of disease-free survival in hepatocellular carcinoma by gene expression profiling. *Ann. Surg. Oncol.* 20, 3747–3753. doi: 10.1245/s10434-013-3070-y

Liu, S., Yang, T. B., Nan, Y. L., Li, A. H., Pan, D. X., Xu, Y., et al. (2017). Genetic variants of cell cycle pathway genes predict disease-free survival of hepatocellular carcinoma. *Cancer Med.* 6, 1512–1522. doi: 10.1002/cam4.1067

Liu, X., Liao, W., Yuan, Q., Ou, Y., and Huang, J. (2015). TTK activates Akt and promotes proliferation and migration of hepatocellular carcinoma cells. *Oncotarget* 6, 34309–34320. doi: 10.18632/oncotarget.5295

Liu, Y. L., Wang, S. C., Hwu, H. G., Fann, C. S., Yang, U. C., Yang, W. C., et al. (2016). Haplotypes of the D-amino acid oxidase gene are significantly associated with schizophrenia and its neurocognitive deficits. *PLoS One* 11:e0150435. doi: 10.1371/journal.pone.0150435

Liu, Z., Chen, M., Xie, L. K., Liu, T., Zou, Z. W., Li, Y., et al. (2018). CLCA4 inhibits cell proliferation and invasion of hepatocellular carcinoma by suppressing epithelial-mesenchymal transition via PI3K/AKT signaling. *Aging* 10, 2570–2584. doi: 10.18632/aging.101571

Lu, J., Chen, Z., Zhao, H., Dong, H., Zhu, L., Zhang, Y., et al. (2020). ABAT and ALDH6A1, regulated by transcription factor HNF4A, suppress tumorigenic capability in clear cell renal cell carcinoma. *J. Transl. Med.* 18:101. doi: 10.1186/s12967-020-02268-1

Luo, S., Li, Y., Ma, R., Liu, J., Xu, P., Zhang, H., et al. (2017). Downregulation of PCK2 remodels tricarboxylic acid cycle in tumor-repopulating cells of melanoma. *Oncogene* 36, 3609–3617. doi: 10.1038/onc.2016.520

Mah, W. C., Thurnherr, T., Chow, P. K., Chung, A. Y., Ooi, L. L., Toh, H. C., et al. (2014). Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. *PLoS One* 9:e104158. doi: 10.1371/journal.pone.0104158

Montal, E. D., Dewi, R., Bhalla, K., Ou, L., Hwang, B. J., Ropell, A. E., et al. (2015). PEPCK coordinates the regulation of central carbon metabolism to promote cancer cell growth. *Mol. Cell* 60, 571–583. doi: 10.1016/j.molcel.2015.09.025

Niture, S., Gyamfi, M. A., Lin, M., Chimeh, U., Dong, X., Zheng, W., et al. (2020). TNFAIP8 regulates autophagy, cell steatosis, and promotes hepatocellular carcinoma cell proliferation. *Cell Death Dis.* 11:178. doi: 10.1038/s41419-020-2369-4

Panvichian, R., Tantiwetrueangdet, A., Angkathunyakul, N., and Leelaudomlipi, S. (2015). TOP2A amplification and overexpression in hepatocellular carcinoma tissues. *Biomed. Res. Int.* 2015:381602. doi: 10.1155/2015/381602

Pardini, B., Corrado, A., Paolicchi, E., Cugliari, G., Berndt, S. I., Bezieau, S., et al. (2020). DNA repair and cancer in colon and rectum: Novel players in genetic susceptibility. *Int. J. Cancer* 146, 363–372. doi: 10.1002/ijc.32516

Reis, H., Padden, J., Ahrens, M., Pütter, C., Bertram, S., Pott, L. L., et al. (2015). Differential proteomic and tissue expression analyses identify valuable diagnostic biomarkers of hepatocellular differentiation and hepatoid adenocarcinomas. *Pathology* 47, 543–550. doi: 10.1097/PAT.0000000000000298

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Roessler, S., Jia, H. L., Budhu, A., Forgues, M., Ye, Q. H., Lee, J. S., et al. (2010). A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res.* 70, 10202–10212. doi: 10.1158/0008-5472.CAN-10-2607

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell* 173, 321–337.e10. doi: 10.1016/j.cell.2018.03.035

Sarathi, A., and Palaniappan, A. (2019). Novel significant stage-specific differentially expressed genes in hepatocellular carcinoma. *BMC Cancer* 19:663. doi: 10.1186/s12885-019-5838-3

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D362. doi: 10.1093/nar/gkw937

Tang, K., Yu, Y., Zhu, L., Xu, P., Chen, J., Ma, J., et al. (2019). Hypoxia-reprogrammed tricarboxylic acid cycle promotes the growth of human breast tumorigenic cells. *Oncogene* 38, 6970–6984. doi: 10.1038/s41388-019-0932-1

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16, 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

Veglia, F., Tyurin, V. A., Blasi, M., De Leo, A., Kossenkov, A. V., Donthireddy, L., et al. (2019). Fatty acid transport protein 2 reprograms neutrophils in cancer. *Nature* 569, 73–78. doi: 10.1038/s41586-019-1118-2

Wang, F., Li, L., Piontek, K., Sakaguchi, M., and Selaru, F. M. (2018). Exosome miR-335 as a novel therapeutic strategy in hepatocellular carcinoma. *Hepatology* 67, 940–954. doi: 10.1002/hep.29586

Wang, F., Wu, H., Zhang, S., Lu, J., Lu, Y., Zhan, P., et al. (2019). LAPTM4B facilitates tumor growth and induces autophagy in hepatocellular carcinoma. *Cancer Manag. Res.* 11, 2485–2497. doi: 10.2147/CMAR.S201092

Wang, L. L., Jin, X. H., Cai, M. Y., Li, H. G., Chen, J. W., Wang, F. W., et al. (2018). AGBL2 promotes cancer cell growth through IRGM-regulated autophagy and enhanced Aurora A activity in hepatocellular carcinoma. *Cancer Lett.* 414, 71–80. doi: 10.1016/j.canlet.2017.11.003

Wang, Y. B., Zhou, B. X., Ling, Y. B., Xiong, Z. Y., Li, R. X., Zhong, Y. S., et al. (2019). Decreased expression of ApoF associates with poor prognosis in human hepatocellular carcinoma. *Gastroenterol. Rep.* 7, 354–360. doi: 10.1093/gastro/goz011

Wang, Y. H., Cheng, T. Y., Chen, T. Y., Chang, K. M., Chuang, V. P., and Kao, K. J. (2014). Plasmalemmal Vesicle Associated Protein (PLVAP) as a therapeutic target for treatment of hepatocellular carcinoma. *BMC Cancer* 14:815. doi: 10.1186/1471-2407-14-815

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220. doi: 10.1093/nar/gkq537

Wei, L., Lian, B., Zhang, Y., Li, W., Gu, J., He, X., et al. (2014). Application of microRNA and mRNA expression profiling on prognostic biomarker discovery for hepatocellular carcinoma. *BMC Genomics* 15(Suppl. 1):S13. doi: 10.1186/1471-2164-15-S1-S13

Wong, N., Yeo, W., Wong, W. L., Wong, N. L., Chan, K. Y., Mo, F. K., et al. (2009). TOP2A overexpression in hepatocellular carcinoma correlates with early age onset, shorter patients survival and chemoresistance. *Int. J. Cancer* 124, 644–652. doi: 10.1002/ijc.23968

Wu, M., Liu, Z., Li, X., Zhang, A., Lin, D., and Li, N. (2019). Analysis of potential key genes in very early hepatocellular carcinoma. *World J. Surg. Oncol.* 17:77. doi: 10.1186/s12957-019-1616-6

Xie, X. W., Wang, X. Y., Liao, W. J., Fei, R., Cong, X., Chen, Q., et al. (2018). Effect of upregulated DNA replication and sister chromatid cohesion 1 expression on proliferation and prognosis in hepatocellular carcinoma. *Chin. Med. J.* 131, 2827–2835. doi: 10.4103/0366-6999.246076

Xu, X., Zhou, Y., Miao, R., Chen, W., Qu, K., Pang, Q., et al. (2016). Transcriptional modules related to hepatocellular carcinoma survival: coexpression network analysis. *Front. Med.* 10:183–190. doi: 10.1007/s11684-016-0440-4

Yan, T., Lu, L., Xie, C., Chen, J., Peng, X., Zhu, L., et al. (2015). Severely Impaired and dysregulated cytochrome P450 expression and activities in hepatocellular carcinoma: implications for personalized treatment in patients. *Mol. Cancer Ther.* 14, 2874–2886. doi: 10.1158/1535-7163.MCT-15-0274

Yang, J., Xie, Q., Zhou, H., Chang, L., Wei, W., Wang, Y., et al. (2018). Proteomic analysis and NIR-II imaging of MCM2 protein in hepatocellular carcinoma. *J. Proteome Res.* 17, 2428–2439. doi: 10.1021/acs.jproteome.8b00181

Yin, L., Cai, Z., Zhu, B., and Xu, C. (2018). Identification of Key Pathways and Genes in the Dynamic Progression of HCC Based on WGCNA. *Genes* 9:92. doi: 10.3390/genes9020092

Zhang, T. T., Zhao, X. Q., Liu, Z., Mao, Z. Y., and Bai, L. (2016). Factors affecting the recurrence and survival of hepatocellular carcinoma after hepatectomy: a retrospective study of 601 Chinese patients. *Clin. Transl. Oncol.* 18, 831–840. doi: 10.1007/s12094-015-1446-0

Zhang, Y., Zhao, H., Xu, W., Jiang, D., Huang, L., and Li, L. (2019). High expression of PQBP1 and low expression of PCK2 are associated with metastasis and recurrence of osteosarcoma and unfavorable survival outcomes of the patients. *J. Cancer* 10, 2091–2101. doi: 10.7150/jca.28480

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 February 2021
doi: 10.3389/fgene.2021.590377

Single-Cell Transcriptional Profiling Reveals Sex and Age Diversity of Gene Expression in Mouse Endothelial Cells

Xianxi Huang[1,2,3], *Wenjun Shen*[4,5], *Stefan Veizades*[2,3,6], *Grace Liang*[2,3,7], *Nazish Sayed*[3] *and Patricia K. Nguyen*[2,3,7]*

[1] Department of Critical Care Medicine, The First Affiliated Hospital of Shantou University Medical College, Shantou, China, [2] Division of Cardiovascular Medicine, Stanford University, Stanford, CA, United States, [3] Stanford Cardiovascular Institute, Stanford, CA, United States, [4] Department of Bioinformatics, Shantou University Medical College, Shantou, China, [5] Center for Biomedical Informatics Research, Stanford University, Stanford, CA, United States, [6] Edinburgh Medical School, College of Medicine and Veterinary Medicine, University of Edinburgh, Edinburgh, United Kingdom, [7] Cardiology Section, Department of Veteran Affairs, Palo Alto, CA, United States

OPEN ACCESS

Edited by:
Lu Zhang,
Hong Kong Baptist University, Hong Kong

Reviewed by:
Yunpeng Xu,
Rutgers, The State University of New Jersey, United States
Ralf Adams,
Max Planck Institute for Molecular Biomedicine, Germany

***Correspondence:**
Patricia K. Nguyen
pknguyen@stanford.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *01 August 2020*
Accepted: *05 January 2021*
Published: *17 February 2021*

Citation:
Huang X, Shen W, Veizades S, Liang G, Sayed N and Nguyen PK (2021) Single-Cell Transcriptional Profiling Reveals Sex and Age Diversity of Gene Expression in Mouse Endothelial Cells. Front. Genet. 12:590377. doi: 10.3389/fgene.2021.590377

Although it is well-known that sex and age are important factors regulating endothelial cell (EC) function, the impact of sex and age on the gene expression of ECs has not been systematically analyzed at the single cell level. In this study, we performed an integrated characterization of the EC transcriptome of five major organs (e.g., fat, heart-aorta, lung, limb muscle, and kidney) isolated from male and female C57BL/6 mice at 3 and 18 months of age. A total of 590 and 252 differentially expressed genes (DEGS) were identified between females and males in the 3- and 18-month subgroups, respectively. Within the younger and older group, there were 177 vs. 178 DEGS in fat, 305 vs. 469 DEGS in heart/aorta, 22 vs. 37 DEGS in kidney, 26 vs. 439 DEGS in limb muscle, and 880 vs. 274 DEGS in lung. Interestingly, LARS2, a mitochondrial leucyl tRNA synthase, involved in the translation of mitochondrially encoded genes was differentially expressed in all organs in males compared to females in the 3-month group while S100a8 and S100a9, which are calcium binding proteins that are increased in inflammatory and autoimmune states, were upregulated in all organs in males at 18 months. Importantly, findings from RNAseq were confirmed by qPCR and Western blot. Gene enrichment analysis found genes enriched in protein targeting, catabolism, mitochondrial electron transport, IL 1- and IL 2- signaling, and Wnt signaling in males vs. angiogenesis and chemotaxis in females at 3 months. In contrast, ECs from males and females at 18-months had up-regulation in similar pathways involved in inflammation and apoptosis. Taken together, our findings suggest that gene expression is largely similar between males and females in both age groups. Compared to younger mice, however, older mice have increased expression of genes involved in inflammation in endothelial cells, which may contribute to the development of chronic, non-communicable diseases like atherosclerosis, hypertension, and Alzheimer's disease with age.

Keywords: single-cell sequencing, endothelial cells, sex, age, cardiovascular disease

INTRODUCTION

The endothelium comprises a single monolayer of cells that lines the cardiovascular and lymphatic system, serving as the interface between tissue walls and the blood and lymph, respectively. Although once considered a passive conduit for nutrient and waste exchange, endothelial cells (ECs) are now recognized as active regulators of coagulation, inflammation, vascular tone, metabolism, and tissue repair. Endothelial cells, however, are not identical in their structure and function across organ systems. An organ's phenotype as well as its microenvironment play an important role in shaping vascular development during embryogenesis as well as vascular repair after injury, which in turn alters the morphology and behavior of ECs within individual organs and across various organs. Interestingly, EC heterogeneity may be maintained even after removal from their microenvironment as shown in previous studies showing that ECs from different organs respond uniquely to pro-inflammatory cytokines, such as tumor necrosis factor alpha, interleukin-1 beta and the bacterial product lipopolysaccharide, when administered *in vitro* (Booth et al., 2004; Gutierrez et al., 2013). Whether age and sex add additional layers of heterogeneity has not been systematically evaluated.

As the gatekeepers of vascular health, it is not surprising that injury to endothelial cells in the arteries, capillaries and veins has been associated with a myriad of diseases affecting the brain [e.g., multiple sclerosis (Barak et al., 2017), stroke (Budhiraja et al., 2004)], cardiovascular system [e.g., coronary artery disease (Johnson and Nangaku, 2016), vasculitis (Perticone et al., 2010)], lung [e.g., asthma, COPD (Dabiré et al., 2012), pulmonary hypertension (Timmerman and Volpi, 2013)], kidney [e.g., diabetic kidney disease (Molema, 2010), hypertensive kidney disease (Muller et al., 2002)], and muscle [e.g., Duchenne's muscular dystrophy (Derada Troletti et al., 2019), age-associated sarcopenia (Cereda et al., 2013)]. Importantly, many of these diseases have age (e.g., old vs. young) and sex dimorphisms in their prevalence, manifestation, and outcome. The biological reasons underlying these clinical observations remain poorly understood.

We hypothesize that the phenotypic similarities and differences in EC structure and function across various organs are reflected in their global gene expression and show a pattern of age and sexual dimorphism. While we are not the first group to evaluate the EC global gene expression across organs (Feng et al., 2019), we provide an unbiased, systematic, and comprehensive comparison of EC transcriptomics based on sex and age within the tissue microenvironment of 5 major organs (e.g., fat, heart and aorta, lung, limb muscle, and kidney) harvested from the same mice, using state-of-the-art single cell technology. We identify shared and organ-specific gene signatures for ECs in males and females across different age groups. Findings from this study will not only provide a reference guide for the gene expression of ECs across multiple organs in males and females, but may also provide valuable insight into the potential mechanisms that underlie why the patterns of certain diseases may vary by sex and age and may facilitate the development of personalized approach to diagnosis and treatment.

MATERIALS AND METHODS

Data Source and Identification of Differentially Expressed Genes

Single cell transcript data was obtained from the database generated by the Tabula Muris Consortium et al. (2018) and Tabula Muris Consortium (2020) (https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733).

The data was processed using Seurat V2. The expression of Cdh5 and Pecam1 were used to identify the endothelial cells in each tissue in males and females in the young and old cohort. Of the 20 organs, only the following five organs contained sufficient cell numbers for analysis: (1) fat, (2) heart and aorta, (3) lung, (4) limb muscle, and (5) kidney. Only cells that expressed both transcripts in these five organs were merged into a Seurat object and analyzed for differential expression at a cut off of log2 fold change >1 and adjusted $p < 0.05$ (Benjamin Hochberg).

GO and Pathway Enrichment Analysis

Gene ontology and pathway enrichment analysis was performed using EnrichR (https://amp.pharm.mssm.edu/Enrichr3; Chen et al., 2013; Kuleshov et al., 2016), *ClusterProfiler*, and the Gene Ontology Consortium website (http://geneontology.org/). Specifically, all differentially expressed genes for males and females from all organs were analyzed to generate **Figures 4–9**. The *p*-value is computed using the Fisher exact test and represents the probability of having at least × genes out of y total genes in the list annotated by the GO term, given the proportion of genes in the whole genome annotated by the GO term. For **Figure 4** for the analysis of all ECs, we used Enrichr, which uses a combined score that is computed by multiplying the unadjusted *P*-value with the *Z*-score that is calculated by assessing the deviation from the expected rank. For **Figures 5–9** for the analysis of ECs in each organ, we used Cluster Profiler, which computes a gene ratio that represents the number of genes in our input list associated with a given GO term divided by the total number of input genes. All graphs were plotted in Prism 7, and a heatmap was plotted in R 3.6.1 (https://www.r-project.org).

Aortic Endothelial Cell Isolation, qPCR, and Western Blot

To verify findings from single cell transcriptomics, we performed qPCR on aortic ECs isolated from young and old C57Bl/6J mice. To obtain ECs for *in vitro* culture, we extracted aortas from mice ($n = 6$ per group) and digested them using Liberase (e.g., 5 mg dissolved in 10 ml DMEM/F12 medium to achieve a concentration of 1 mg/ml). We collected the cell pellet and resuspended it in EGM medium with 5%FBS and Pen strep. The resuspended cells were then placed in 24 well-plates coated with gelatin. Media was changed daily. After 1 week, wells containing confluent cells were trypsinized and re-plated into six well-plates. After another week of expansion, cells were trypsinized and collected for RNA extraction, cDNA synthesis, and qPCR using standard protocols. We used the following primers: (1) Lars2, (2)

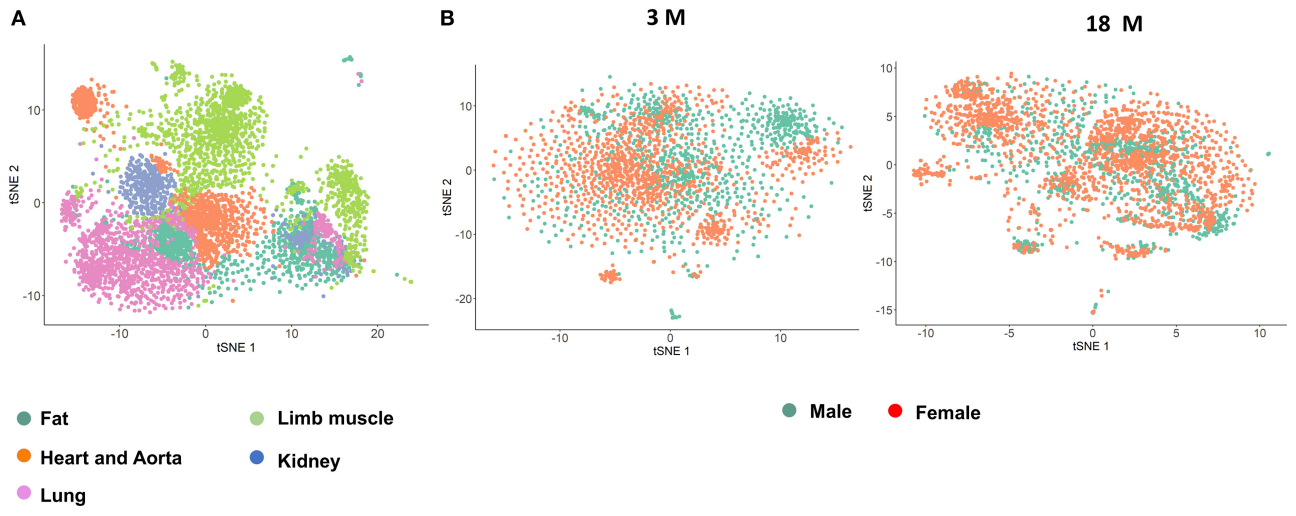


FIGURE 1 | *T-SNE* visualization of endothelial cells (ECs) from single cell sequencing. **(A)** *t*-SNE plot of all ECs collected by single cell transcriptomics and colored by organ; **(B)** *t*-SNE plot of all ECs colored by sex in the 3 month (*left*) and 18 month (*right*) age groups.

S100a8, (3) S100a9, and (4) genes belonging to the WNT pathway (e.g., FZD4, PFN1, PSMA2, PSMA7, PSMB8, and PSMB).

In addition to PCR, using standard protocols, we performed Western blot on aortic ECs to determine the protein expression of selected genes. Briefly, endothelial cells were harvested and lysed for Western blot analysis. Protein was loaded onto 4–15% Tris gels (Bio-red) at 100 V for 60 min. The separated proteins were transferred onto a polyvinylidene fluoride (PVDF) membrane (Bio-red, 0.2 μm). The PVDF membrane was then blocked with 5% skim milk powder at room temperature for 2 h, washed with PBS for 3 min, and incubated overnight at room temperature with the following rabbit anti-mouse antibodies: (1) anti-Lars2 (Proteintech, 1:500, 17170-1-AP), (2) anti-Profilin-1 (Invitrogen, 1: 1,000, 11680-1-AP), (3) anti-Frizzled4 (Invitrogen, 1: 300, PA5-41972), (4) anti-S100a8 (Invitrogen, 1:100PA5-79948), (5) anti-S100a9 (Invitrogen, 1:200, 14226-1-AP), and (6) anti-GAPDH (Invitrogen, 1:2,000, # 39-8600). The membrane was then incubated with anti-rabbit secondary antibody (Jackson immunoresearch, 1:5,000, 111-035-144) for 2 h in room temperature and washed by TBST three times. Protein expression was detected using enhanced chemiluminescence. The relative expression of the target protein was defined as the ratio of average OD value of target protein bands to that of the internal reference GAPDH.

Statistical Analysis

All statistical analysis was performed by GraphPad Prism software (version 7) and R software (version 3.6.1). All *p*-values were adjusted for multiple comparisons. Adjusted $p < 0.05$ was considered statistically significant.

RESULTS

Data Source and Analysis

Original data from the Tabula Muris Consortium (e.g., *Tabula Muris* and *Aging Transcriptomic Atlas*) was obtained (Tabula Muris Consortium et al., 2018; Tabula Muris Consortium, 2020). Information on the following 5 organs were analyzed from 4 male and 3 female mice in the 3 month group, and 2 males and 4 females in the 18 month group: fat, heart and aorta, lung, limb muscle, and kidney (**Supplementary Figure 1A**). To identify distinct cell populations based on shared and unique patterns of gene expression, we performed dimensionality reduction and unsupervised cell clustering methods. EC lineage genes, Pecam1 and Cdh5, were used as markers to identify the ECs (**Figure 1**; Tabula Muris Consortium, 2020). Cell counts for each organ stratified by age and sex are shown in **Supplementary Figure 1B**. Profiles of 4,883 cells analyzed in the Seurat V2 by unsupervised analysis revealed that most cells are grouped by their parent organs. There was a sub-cluster of ECs from the heart and aorta as well as another sub-cluster composed of all organs except the heart and aorta that diverged from the primary cluster. Analysis by sex in each age group, did not reveal distinct clusters, suggesting that the majority of the transcriptome in male and female ECs is similar across age groups.

Differentially Expressed Genes in All ECs Based on Age and Sex

Subsequent analyses focused on comparing the patterns of differential gene expression between male and female in young and old age group. Consistent with tSNE visualization analysis of all ECs by age and sex, density plots showed that the majority of genes have <50% difference in expression between males and females ($\log_2$ fold <1) (**Figure 2A**). A total of 590 and 252 differentially expressed genes (DEGS) were identified between females and males in the 3- and 18-month subgroups, respectively, with 59 shared genes between young and older mice that were sexually dimorphic (**Figure 2B** and **Supplementary Table 1**). These genes are involved in angiogenesis (e.g., Acvrl1, Lrg1, Ptprb, and Tmem100), immunity and inflammation (e.g., Adamts1, Cd74, Cebpb, Ctla2a, DCN,

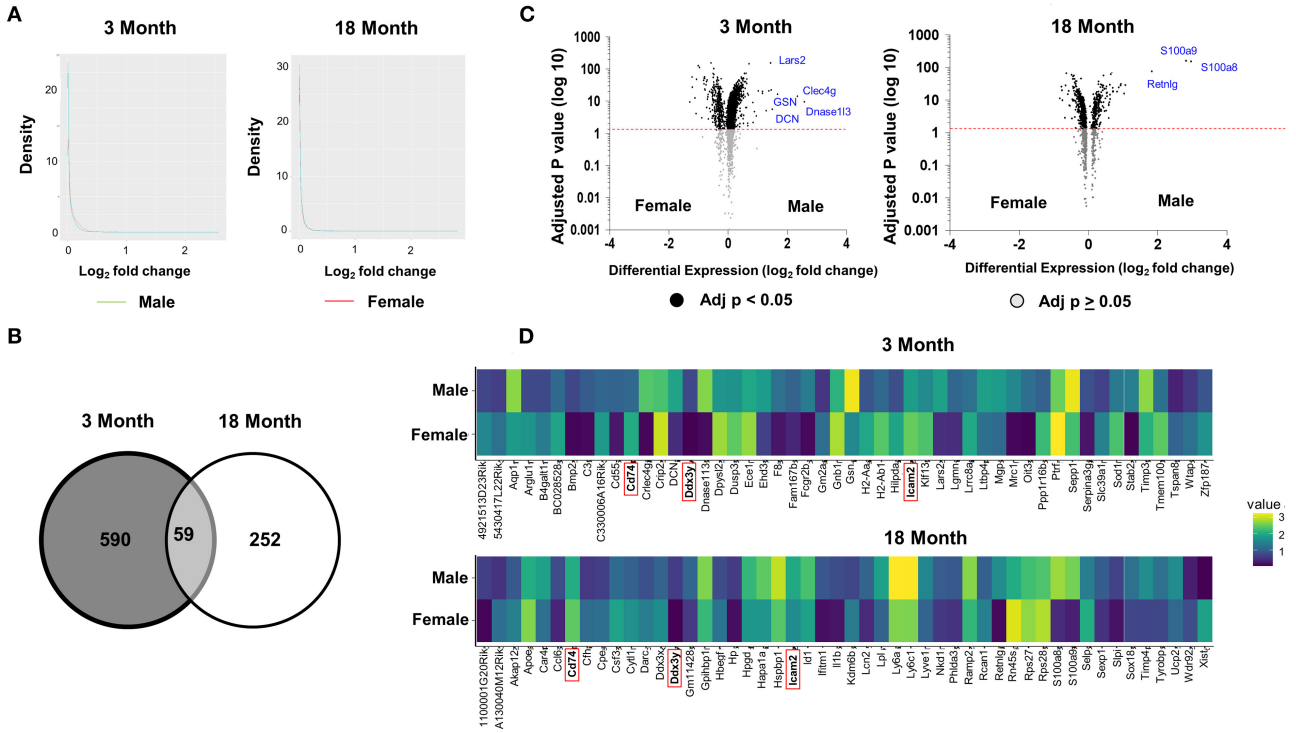


FIGURE 2 | Differential analysis reveals age and sex dimorphism in ECs. **(A)** A comparison of density plots of differential gene expression reveals the expression of the majority of genes was not significantly different in ECs isolated from males and females for both age groups (log fold change <1). Genes up-regulated in male mice plotted in aqua. Genes up-regulated in female mice plotted in red. **(B)** Venn diagram showing 590 and 252 genes that are differentially expressed (adjusted $p < 0.05$) between females and males in the 3 and 18 month group, respectively. Of those genes that are differentially expressed, there are 59 shared genes that are differentially expressed between the sexes in both age groups. **(C)** The most statistically upregulated genes with the largest fold-change in males at 3 months were: Lars2, GSN, Clec4g, DCN, and Dnase1l3 at 3 months. The most statistically upregulated genes with the largest fold-change in males at 18 months was S100a9, S100a8, and Retnlg. The most statistically significant genes with the larges fold change include Xist, Ddx3x, and Cfh, which are upregulated in females, and S1009, S1008, and Retinlg, which are upregulated in males at 3 months *(left)*. The most statistically significant genes with the larges fold change include H2Ab1, Ppp1r16b, and Tmem100B in females and Dnase1l3, Clec4g, Lars2, GSN, and DCN in males at 18 months *(right)*. **(D)** Heat map showing the expression levels of 50 genes with the largest fold difference between males and females at 3 months (*top)* and 18 months (*bottom*). Of the 50 top genes, the following three genes have the largest fold difference and shared between the young and old group (*highlighted in red*): Cd74, DDx3y, and Icam2.

Fcgrt, H2-Ab1, Icam2, Kdm6b, Lcn2, Nfkbia, Nfkbiz, and Sgk1), cellular chemotaxis (e.g., Ecscr, Gpr56, Pcdh1, and Tmsb10), endothelial specific function (e.g., Apold1), smooth muscle cell differentiation (e.g., Crip2), and cellular growth and development (e.g., Bmpr2, Ccdc85b, Egr1, Fosb. Id3, Oaz1, Pfkbfb3, and Tspan8), apoptosis (e.g., Gas5 and Phlda3), and lipid metabolism (e.g., Thrsp). Of note, genes with the largest fold change (>1.4-fold) included Dnase1l3, Clec4g, Lars2, GSN, and DCN, which were significantly upregulated in males at 3 months. Of these, Clec4g and GSN are important in the immune response, DCN is related to angiogenesis, Lars2 is involved in mitochondrial function, and DNASE1l3 regulates apoptosis (**Figure 2C** and **Supplementary Table 2**). In contrast, three genes including RETNLG, S100A8, and S100A9, which are involved in the regulation of immune function and inflammation, were significantly upregulated in males >18 months. When comparing the gene expression profiles across the age groups, we find the following two genes with shared sexual dimorphism across age: (1) Cd74, a cell surface receptor for cytokine macrophage migration inhibitory factor that is involved in apoptosis, immune response and cell migration (Fan et al., 2011; Le Hiress et al., 2015; Gil-Yarom et al., 2017); and (2) ICAM2, intracellular adhesion molecule 2, that mediates adhesive interactions important for immune response and surveillance and angiogenesis (**Figure 2D**; Huang et al., 2005; Halai et al., 2014).

Differentially Expressed Genes of ECs in the Five Major Organs Based on Age and Sex

In order to explore if the tissue microenvironment affects the differential expression of ECs, we performed an organ-specific analysis based on age and sex. We found gene expression signatures that distinguish each organ and appear to alter more with age than sex (**Figure 3** and **Supplementary Figures 2–6**, **Supplementary Tables 3**, **4**). Interestingly, we found that the genes Lars2 is differentially expressed in males compared to females in all five organs in the 3-month group. In contrast, S100A8 and S100A9 are upregulated in all organs from males compared to females at 18 months. In addition to various other functions, these genes are involved in regulating immunity and inflammation.

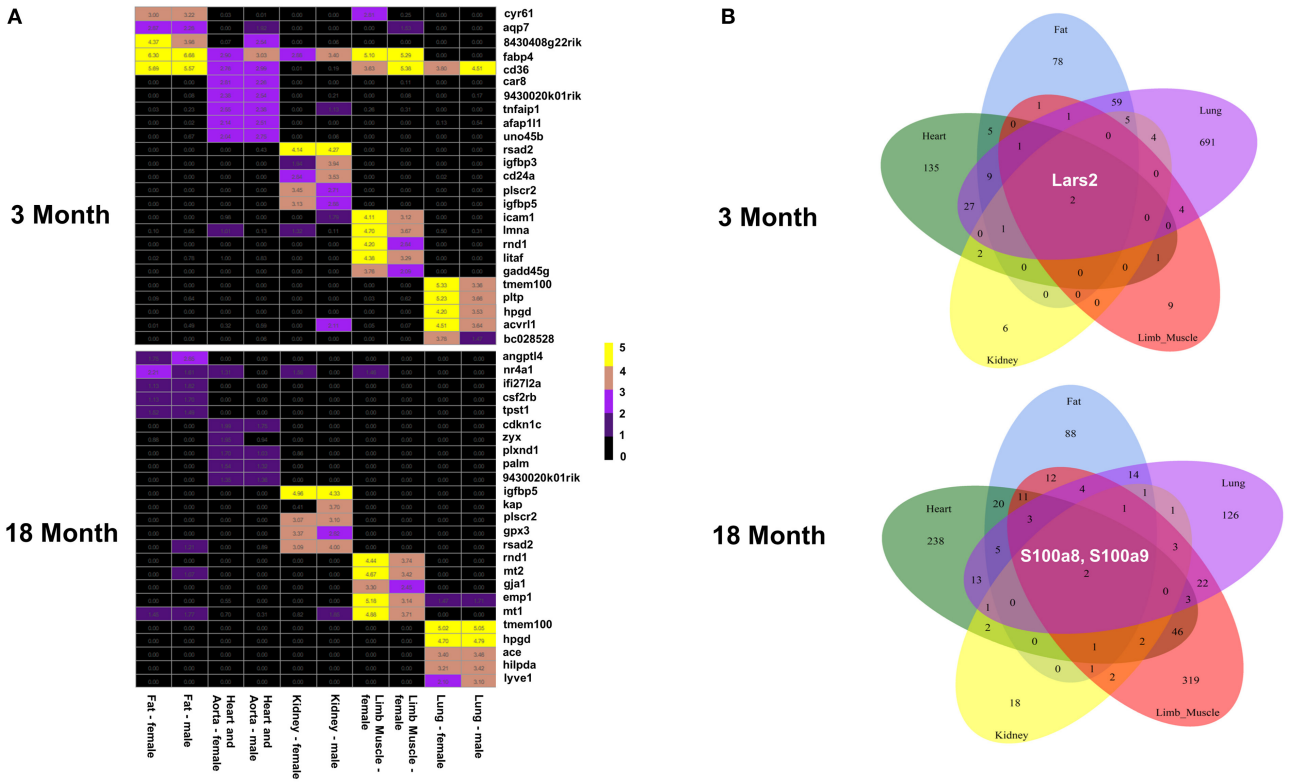


FIGURE 3 | EC gene expression signatures in the five organs. **(A)** Heat map showing the genes with the highest expression values in males and females in fat, heart aorta, kidney, limb and lung at 3 months (top) and 18 months (bottom). **(B)** Venn diagram showing that Lars2 is differentially expressed at 3 months (top) and S100A8 and S100A9 are differentially expressed at 18 months (bottom) in all five organs.

GO Analysis of DEGS in ECs Based on Age and Sex

To further explore differences in functional characteristics of ECs based on age and sex, DEGS were submitted to gene ontology pathway analysis. The overall analysis revealed enrichment in pathways involving protein targeting, catabolism, mitochondrial electron transport, IL 1- and IL 2- signaling, and WNT signaling at 3 months (**Figure 4**). In contrast, genes involved in angiogenesis and chemotaxis were enriched in females at 3 months. ECs from males and females at 18 months, however, had up-regulation in similar pathways involved in inflammation and apoptosis. When analyzing DEGS stratified by organ (**Figures 5–9**), we find that genes enriched in pathways regulating inflammation and immunity pathways were upregulated in fat and lung from females. In contrast, these pathways were upregulated in both male and female ECs from fat, in male ECs from heart and aorta, in both male and female ECs from the lung, and in male ECs from the kidney. ECs from limb muscles for both sexes as well as ECs from the heart and aorta were enriched in genes involved in apoptosis.

Confirmation of Single Cell RNAseq Findings by qPCR and Western Blot

Importantly, findings from RNAseq were verified by qPCR including up-regulation of Lars2 in 3-month old males as well as S100A8 and S100A9 in 18-month old males, respectively (**Figure 10**). We also confirmed differential expression of genes involved in selected pathways (**Figure 10**). Further analysis with Western blot showed increased protein expression in Lars2 and selected proteins involved in the Wnt pathway (e.g., FZD4 and PFN1) in 3-month males as well as S100A8 and S100A9 proteins in the 18-month males.

DISCUSSION

Age and sex are major risk factors for many diseases associated with endothelial dysfunction including obesity (Palmer and Clegg, 2015; Jura and Kozak, 2016), metabolic syndrome (Chella Krishnan et al., 2018), coronary artery disease (Nguyen et al., 2011; Madhavan et al., 2018), diabetes (DECODE Study Group, 2003), hypertension (Gillis and Sullivan, 2016), emphysema (Barnes, 2016), pulmonary artery hypertension (Lakshmanan et al., 2020), sarcopenia (Tay et al., 2015), and chronic kidney disease (Yu et al., 2012). It has been suggested that the female sex chromosome increases survival and lifespan although the exact mechanisms remain unclear (Davis et al., 2019). Biological changes associated with normal chronological aging—including alterations in the immune system, changes in hormone secretion, and defects in the cell repair systems as a result of telomere shortening or cellular mutations—can result in the deterioration of micro- and macrovascular function that can lead to disease

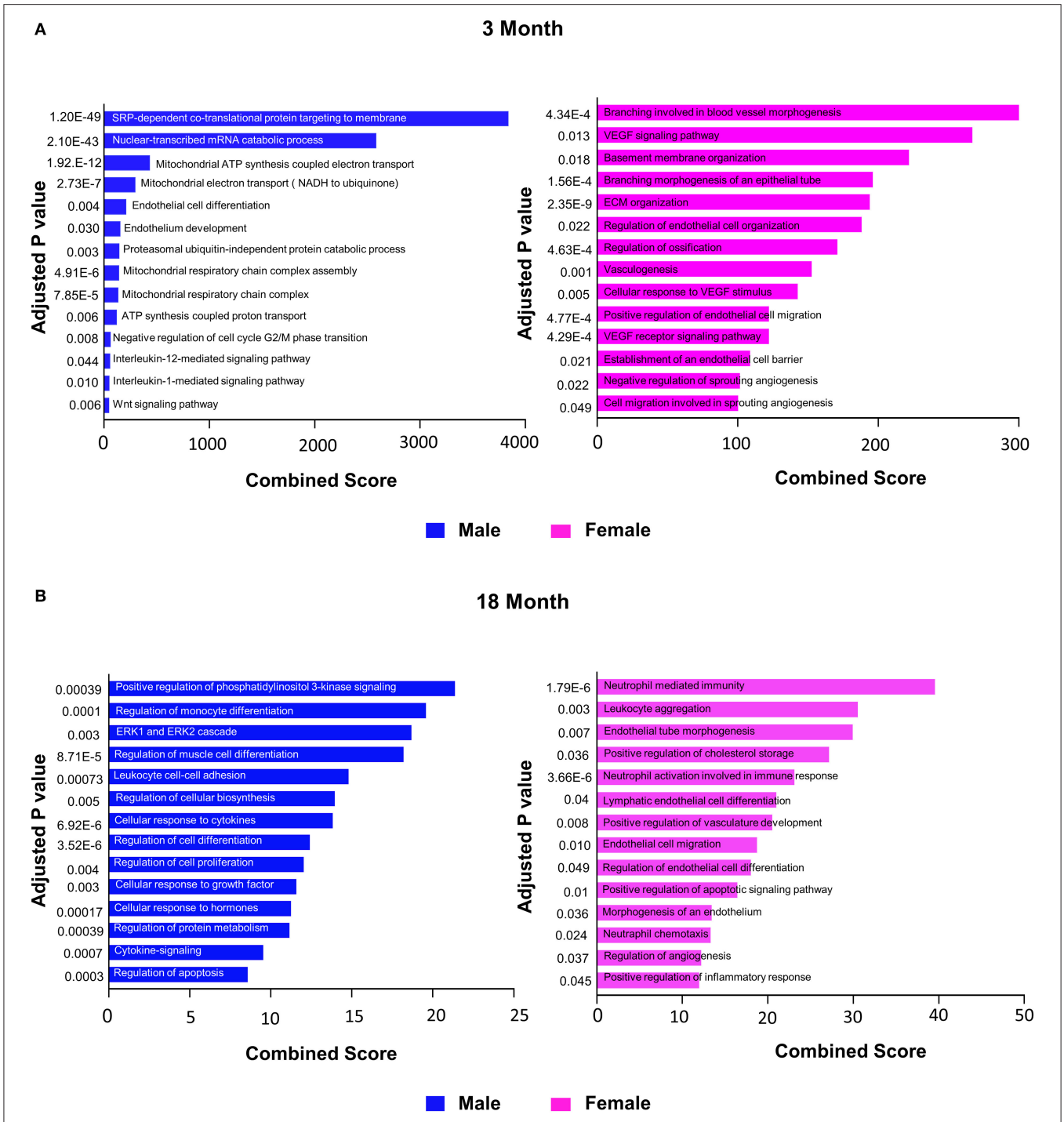


FIGURE 4 | Comparison of biological pathways that are upregulated in ECs from all organs in males and females in the young and old cohort. **(A)** Bar graph showing the top significant biological pathways that are upregulated in males and females at 3 months. **(B)** Bar graph showing significant biological pathways that are upregulated in males and females at 18 months (adjusted $P < 0.05$).

development (López-Otín et al., 2013). To evaluate whether age- and sex related differences in EC function are reflected in the EC transcriptome, we performed an unbiased analysis of ECs isolated from fat, heart and aorta, lung, limb muscle, and kidney obtained from the same male and female mice at 3 months (equivalent to the human age of 20–30 years old) and 18 months of age (equivalent to the human age of 50–60 years old).

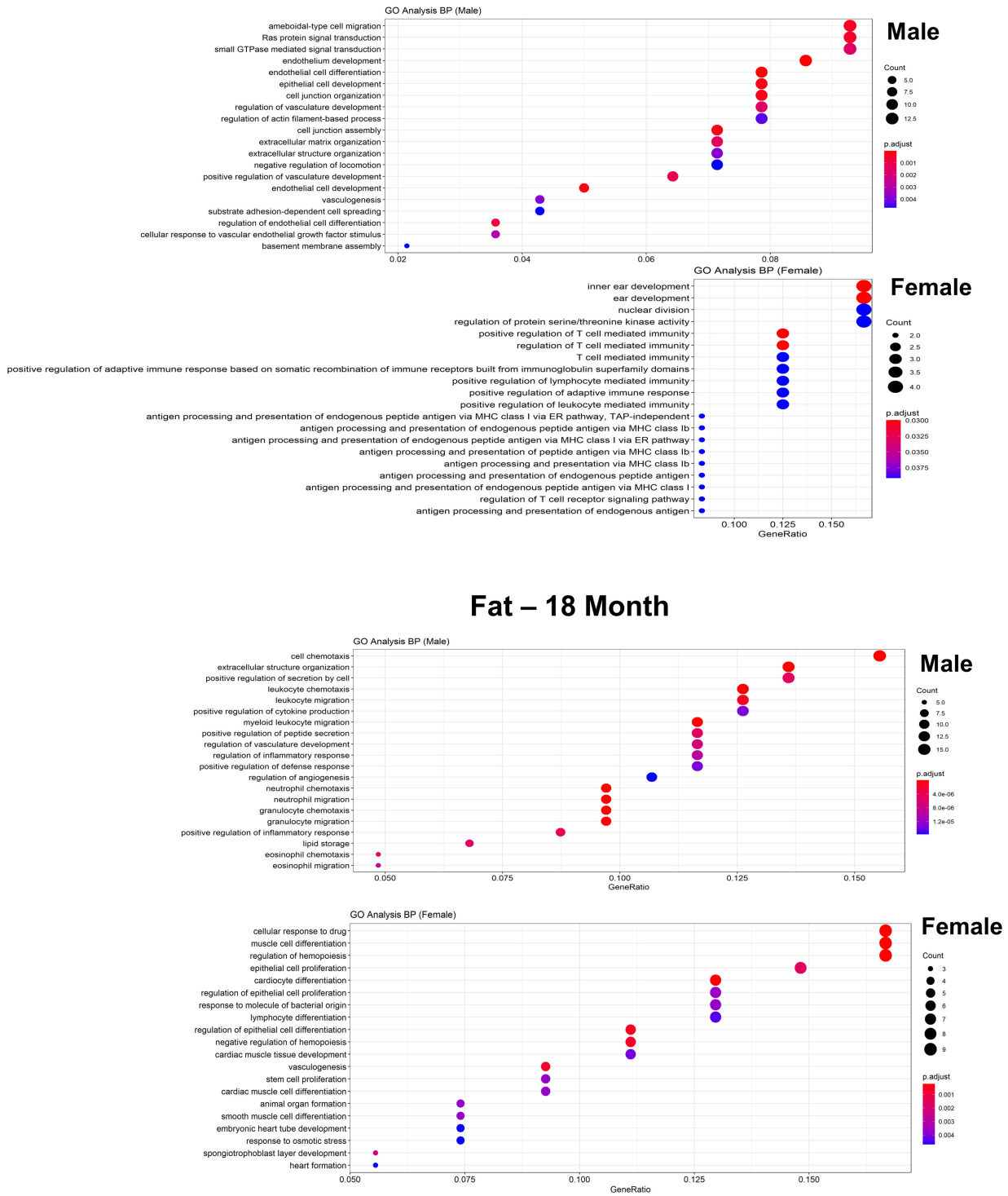


FIGURE 5 | Comparison of the biological pathways that are upregulated in ECs isolated from fat in males and females in the young and old cohort. **(A)** Dot plot showing significant biological pathways that are upregulated in males and females in ECs isolated from fat at 3 months. **(B)** Dot plot showing significant biological pathways that are upregulated in males and females in ECs isolated from fat at 18 months (adjusted $P < 0.05$).

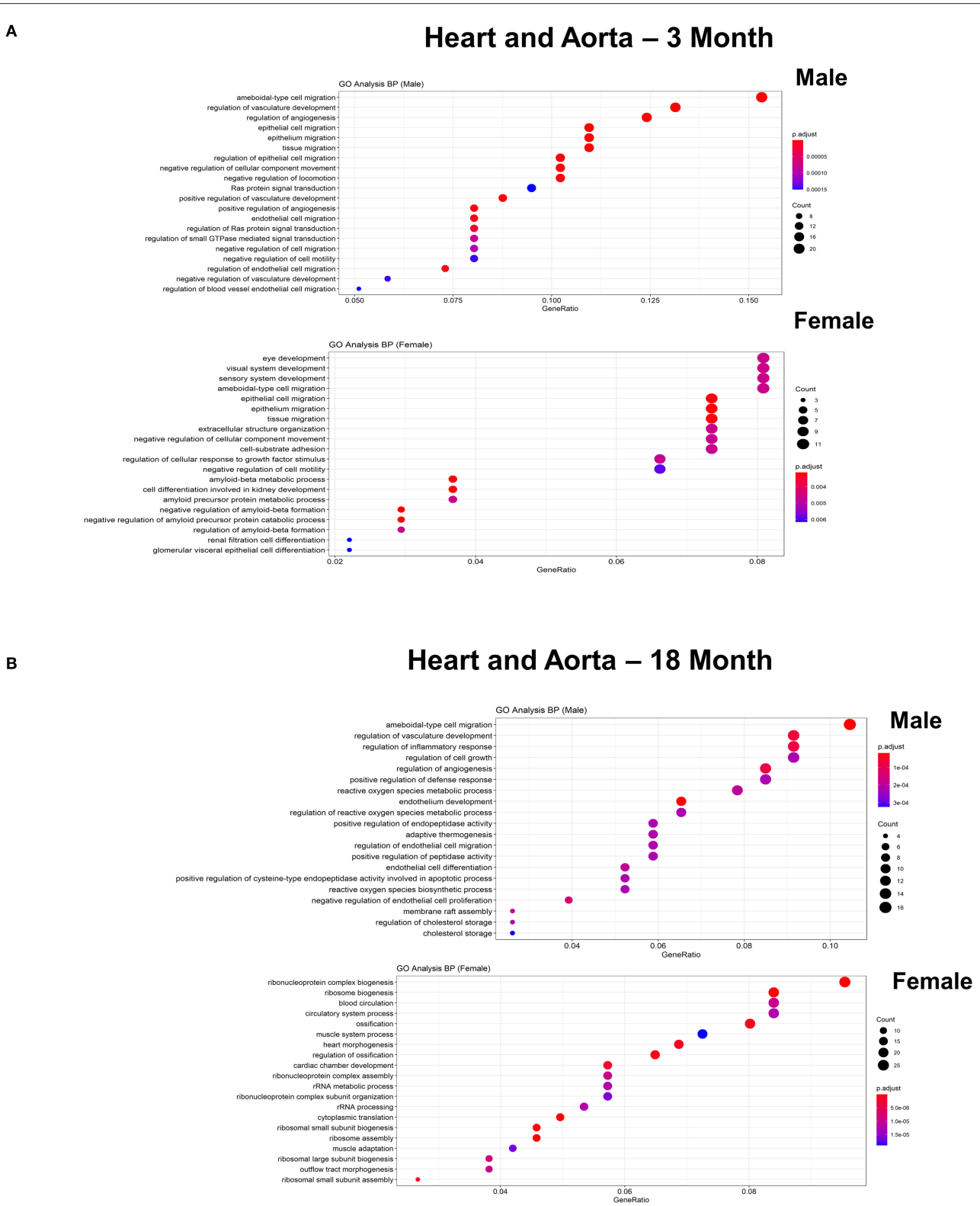


FIGURE 6 | Comparison of the biological pathways that are upregulated in ECs isolated from heart and aorta in males and females in the young and old cohort. **(A)** Dot plot showing significant biological pathways that are upregulated in males and females in ECs isolated from heart and aorta at 3 months. **(B)** Dot plot showing significant biological pathways that are upregulated in males and females in ECs isolated from heart and aorta at 18 months.

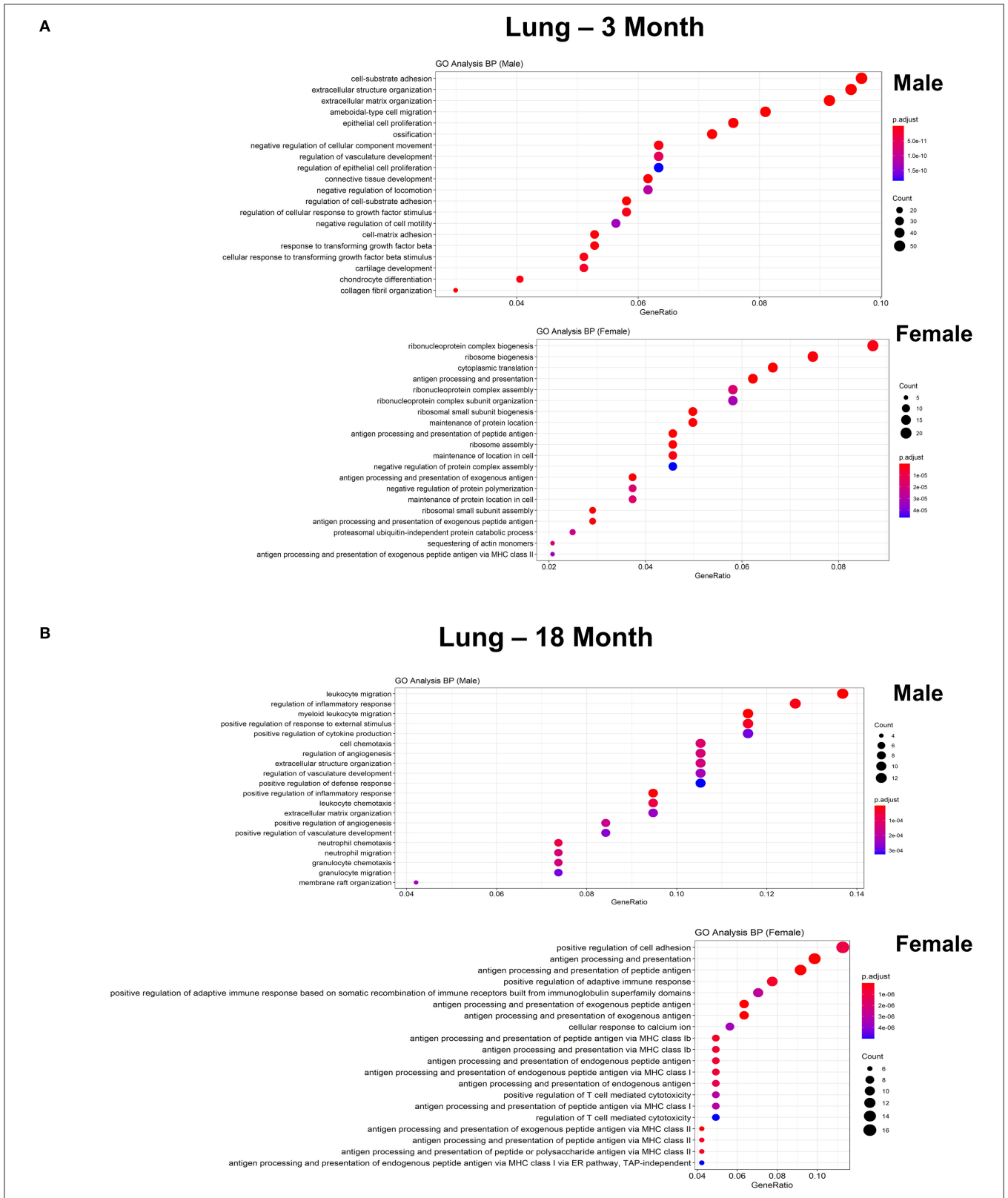


FIGURE 7 | Comparison of biological pathways that are upregulated in ECs isolated from lung in males and females in the young and old cohort. **(A)** Dot plot showing significant biological pathways that are upregulated in males and females in ECs isolated from lung at 3 months. **(B)** Dot plot showing significant pathways that are upregulated in males and females in ECs isolated from lung at 18 months.

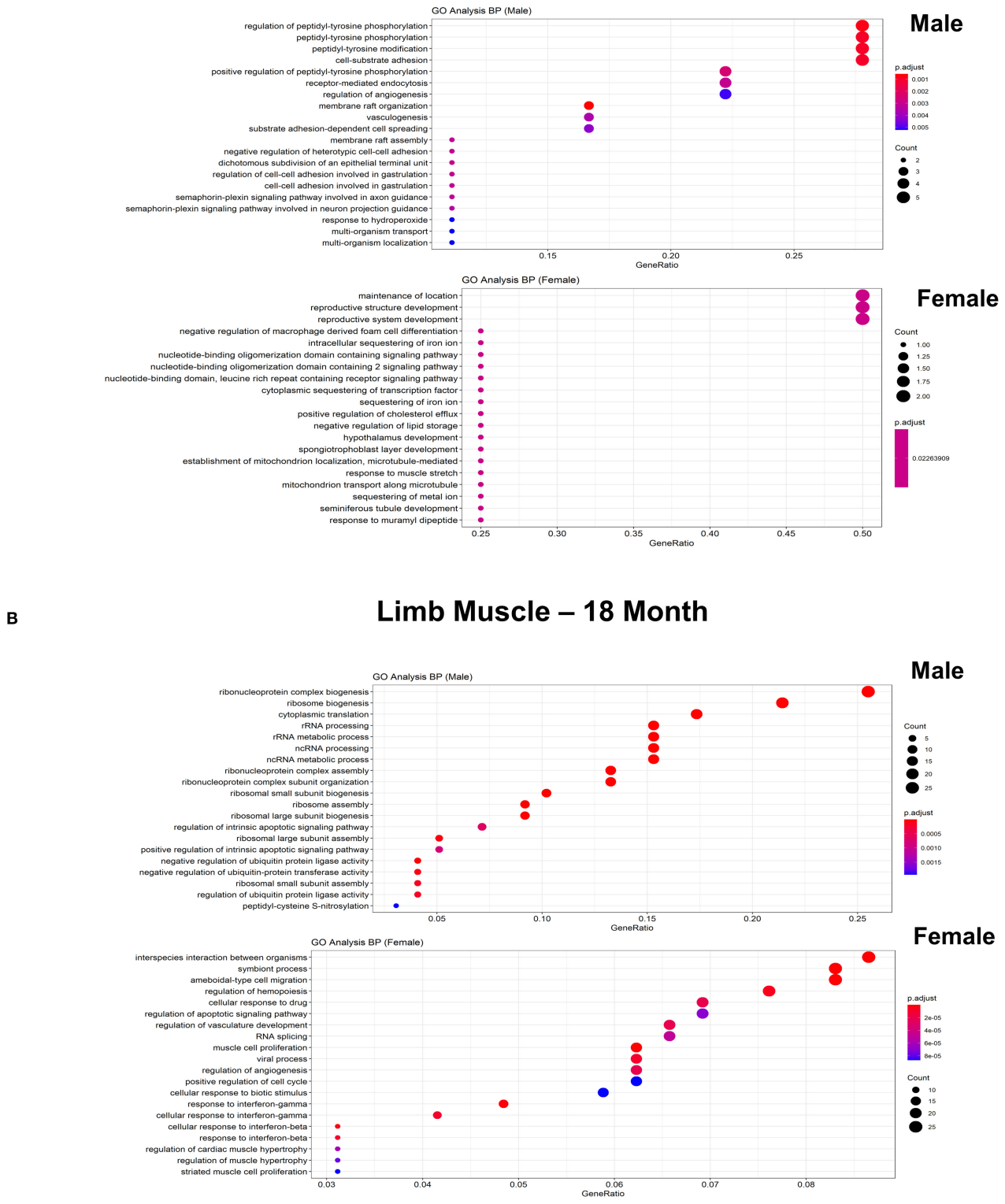


FIGURE 8 | Comparison of biological pathways that are upregulated in ECs isolated from limb muscle in males and females in the young and old cohort. **(A)** Dot plot showing the significant biological pathways that are upregulated in males and females in ECs isolated from limb muscle at 3 months. **(B)** Dot plot showing significant biological pathways that are upregulated in males and females in ECs isolated from limb muscle at 18 months.

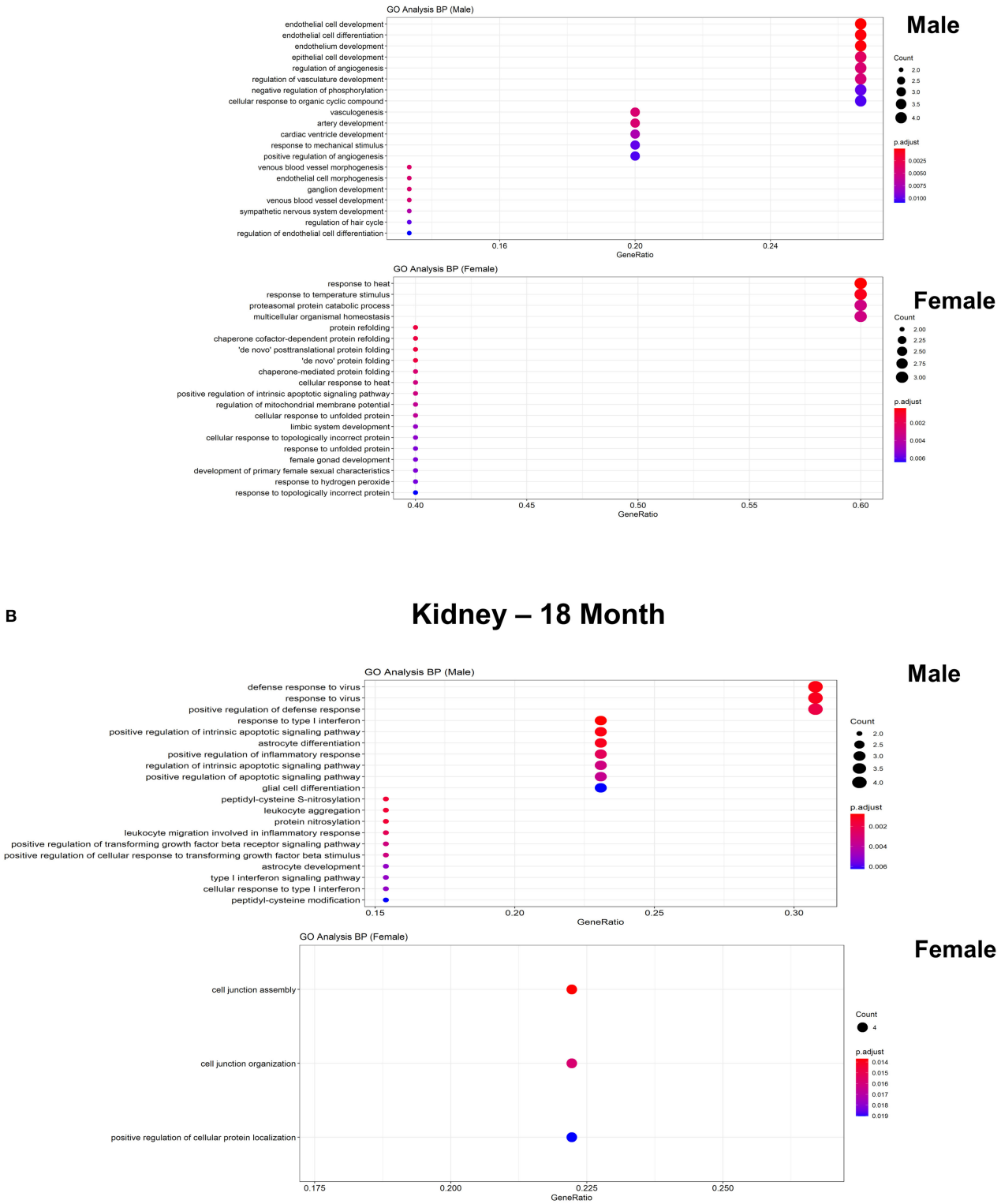


FIGURE 9 | Comparison of biological pathways that are upregulated in ECs isolated from kidney in males and females in the young and old cohort. **(A)** Dot plot showing significant biological pathways that are upregulated in males and females in ECs isolated from kidney at 3 months. **(B)** Dot plot showing significant biological pathways that are upregulated in males and females in ECs isolated from kidney at 18 months.

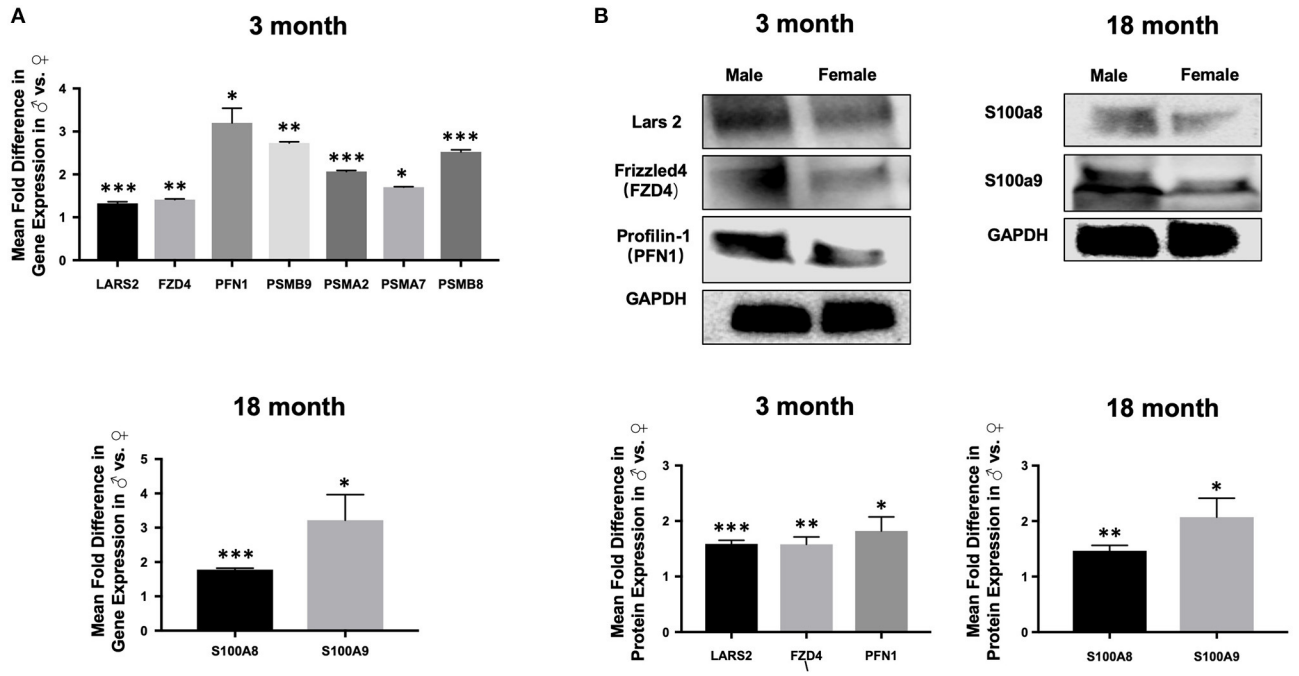


FIGURE 10 | Expression of selected genes and proteins in aortic ECs. **(A)** Bar graphs displaying upregulation of the expression of Lars2 as well as genes involved in the Wnt pathway at 3 months (top) and upregulation of S100A8 and S100A9 at 18-months in males relative to females by quantitative PCR (bottom). **(B)** Western blot analysis confirming the expression of Lars2 as well as genes involved in the Wnt pathway are expressed at the protein level (top) and displayed as relative fold-difference between males and females in a bar graph (bottom). ***$P < 0.001$; **$P < 0.01$; and *$P < 0.01$.

Our analysis revealed that changes in the EC transcriptome are largely similar between the sexes in each age group. Consistent with previous studies using mouse and human tissues (InanlooRahatloo et al., 2017; Kassam et al., 2019), the relative fold-difference in gene expression of ECs between males and females in the majority of genes is <50% (average fold change <1), a finding that is similar across all organs evaluated. Interestingly, in the entire transcriptome of these organs, Lars2 is the one somatic gene appears to be consistently up-regulated in males compared to females in the younger age group. Lars2 encodes for a mitochondrial leucyl-tRNA synthetase that affect aminoacyl-tRNA ligase activity in mitochondria ('t Hart et al., 2005; Carminho-Rodrigues et al., 2020). As a part of a unique group of enzymes that catalyzes the ligation of amino acids to their cognate tRNAs, Lars2 as well as other aminoacyl-tRNA synthetases determine the genetic code that is essential for protein synthesis and cell viability. Abnormalities in aminacyl-tRNA synthetases have been implicated in the development of neurological disease, cancer, and auto immune disease (Park et al., 2008). Interestingly, single nucleotide mutations in Lars2, perhaps induced by the accumulation of oxidative stress stimulated by episodes of hyperglycemia and hyperinsulinemia, has been implicated as a novel type 2 diabetes susceptibility gene (Kassam et al., 2019). Mutations in Lars2 have also been associated with sensorineural hearing loss, hydrops, lactic acidosis, sideroblastic anemia, and multisystem failure (Riley et al., 2016; Xia et al., 2018). Although the exact function of Lars2 in ECs is unclear, given its basic function in protein synthesis in mitochondria, further study is warranted to investigate whether the expression levels of Lars2 mediate phenotypic differences between younger males and females.

In contrast to their younger counterparts, across all the organs, the older male mice had upregulation in S100A8 and S100A9 (Vogl et al., 2007), which are Ca^{2+} binding proteins in the S100 family that regulate apoptosis, proliferation, differentiation, migration, energy metabolism, calcium balance, protein phosphorylation, and inflammation. During cellular stress, S100A8 and S100A9 is released as a heterodimer (e.g., calprotectin) into the extracellular space where it binds to TLR4 and initiate a signaling cascade that regulates inflammation, cell proliferation, differentiation, and tumor development in an NF-κB-dependent manner (Turovskaya et al., 2008). Alternatively, calprotectin can interact with receptor for advanced glycation end products (RAGE), which activates NF-κB to induce production of pro-inflammatory cytokines that result in the migration of neutrophils, monocytes, and macrophages (Yen et al., 1997; Sorci et al., 2013). Although predominantly expressed in immune cells, expression of S100A8 and S100A9 is increased in activated endothelial cells under conditions of oxidative stress, hyperglycemia, and pro-inflammatory stimuli (McCormick et al., 2005; Sroussi et al., 2009; Yao and Brownlee, 2010; Furman et al., 2019). Taken together, these findings suggest that age-related changes in the EC tissue microenvironment in males can promote inflammation, which could account for the

increased incidence of endothelial dysfunction and its associated diseases among older middle aged males compared to their female counterparts.

Consistent with these findings, we found that the differentially expressed genes in older mice were enriched in pathways related to inflammation. Aging has long been associated with the development of inflammation (Donato et al., 2015). Previous studies have demonstrated that aging endothelial cells acquire a senescent phenotype characterized by increased secretion of pro-inflammatory cytokines and chemokines into the micro-environment (Hoffmann et al., 2001; Uraoka et al., 2011). Previous studies have shown that senescent endothelial cells do not migrate, proliferate, or sprout; they have limited capacity to form new vessels and have reduced numbers of endothelial progenitor cells; and they do not respond appropriately to hypoxia (e.g., reduced expression of HIF-1 alpha and angiogenic factors) (Lin et al., 2015; Rudnicki et al., 2018). These senescent cells contribute to many non-communicable age-related chronic diseases including insulin resistance, CVD, pulmonary arterial hypertension, chronic obstructive pulmonary disorder, emphysema, Alzheimer's and Parkinson's diseases, macular degeneration, osteoarthritis, and cancer (Lin et al., 2015). Although the exact reasons why these cells develop this senescent phenotype is unclear, studies suggest that both endogenous factors related to biological aging (e.g., oxidative stress, telomere shortening, and DNA damage) and environmental factors (e.g., diet, stress, and chronic infection) may contribute (Uraoka et al., 2011).

Unlike their older counterparts, younger female mice had activation of pathways associated with angiogenesis including activation of genes involved in blood vessel morphogenesis, VEGF signaling, and endothelial cell migration and organization. This finding is consistent with previous studies that have shown that young female mice produce higher levels of proangiogenic factors and vascularity in response to stress than male mice (Xu et al., 2019). Angiogenesis is an important adaptive response to physiological stress and an endogenous repair mechanism after injury that can be impaired with aging. In contrast to young female mice, young male mice showed increased expression of genes involved in the Wnt signaling pathway, which has been shown to be an important regulator of lifespan, especially in the earlier stages of life (MacDonald et al., 2009; Franco et al., 2016). In endothelial cells, Wnt ligands have been shown to regulate vascular remodeling through their regulation of endothelial cell survival and proliferation (MacDonald et al., 2009). Although further study is needed, these findings suggest that vascular morphogenesis in males and females are regulated by diverse pathways.

In summary, our unbiased, integrated analysis of the gene transcriptome has revealed that the EC transcriptome is largely similar in male and female mice. Older mice, especially males, have increased expression of genes involved in immunity and inflammation, which could contribute to the increased prevalence of age-related non-communicable diseases associated with endothelial dysfunction in older men. Future studies are needed to further elucidate the role of DEGS identified in this study in the development of disease.

Limitations

The major limitation of this study is that not all of the organs were collected from both males and females in both age groups. The five organs that we analyzed, however, represent major tissues with important physiological function for health. Another limitation is that single cell sequencing was performed using different techniques for the young (e.g., plate-seq) and old group (e.g., dropseq). In the *Tabula Muris and Tabula Muris Senis* project, gene expression data from 20 organs were performed using these two sequencing methods and compared. The study showed close agreement between the genes, defining each organ-specific cells cluster for both methods. Moreover, gene expression analysis showed several hundred genes were differentially expressed to a similar degrees across organs using both methods. To address the differences in sequencing methods, in our study, we perform DEG and pathway analysis separately for each age group. Within each age group, we calculated the relatively gene expression only for males and females. Any comparisons between age groups was performed only on the output of the differential analysis. Importantly, we performed qPCR on selected genes to confirm results from the RNAseq analysis.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733.

ETHICS STATEMENT

This animal study was reviewed and approved by Stanford APLAC.

AUTHOR CONTRIBUTIONS

PN: conceptualization and supervision. WS, SV, and XH: methodology. WS: data curation. NS: qPCR and Western blot. GL: writing and reviewing. XH and PN: writing review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Institutes of Health (Grant No. 3R01 HL 134830-01) and American Heart Association No. 20POST35210415.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.590377/full#supplementary-material

Supplementary Figure 1 | Schematic of study. **(A)** Endothelial cells (ECs) defined by expression of PECAM and Cdh5 were identified from the Tabula Muris dataset and analyzed by organ specific expression, by age group, and by sex. **(B)** Five organs from 6 males (4 at 3 months and 2 at 18 months) and 7 females (3 at 3 months and 4 at 18 months) were analyzed. Data for 3 and 18 month were generated from single cell plateseq and dropseq data from the Tabula Muris Project, respectively.

Supplementary Figure 2 | Age and sex differences in gene expression in ECs from fat. **(A)** *T-SNE* visualization of endothelial cells from fat. **(B)** Violin plots showing the top genes, defined by their fold change, which were differentially expressed in fat in males vs. females and in young vs. old.

Supplementary Figure 3 | Age and sex differences in gene expression in ECS from the heart and aorta. **(A)** *T-SNE* visualization of endothelial cells from heart and aorta. **(B)** Violin plots showing the top genes that were differentially expressed in heart and aorta in males vs. females and in young vs. old.

Supplementary Figure 4 | Age and sex differences in gene expression in ECS from the lung. **(A)** *T-SNE* visualization of endothelial cells from lung. **(B)** Violin plots showing the top genes that were differentially expressed in lung in males vs. females and in young vs. old.

Supplementary Figure 5 | Age and sex differences in gene expression in ECS from the limb muscle. **(A)** *T-SNE* visualization of endothelial cells from limb muscle. **(B)** Violin plots showing the top genes that were differentially expressed in limb muscle in males vs. females and in young vs. old.

Supplementary Figure 6 | Age and sex differences in gene expression in ECS from kidney. **(A)** *T-SNE* visualization of endothelial cells from kidney. **(B)** Violin plots showing the top genes that were differentially expressed in kidney in males vs. females and in young vs. old.

Supplementary Table 1 | Genes differentially expressed in both 3 month and 18 month groups.

Supplementary Table 2 | Organ-specific gene signatures.

Supplementary Table 3 | Significant difference in gene expression in each organ between male and female in 3 Month.

Supplementary Table 4 | Significant difference in gene expression in each organ between male and female in 18 Month.

REFERENCES

't Hart, L. M., Hansen, T., Rietveld, I., Dekker, J. M., Nijpels, G., Janssen, G. M., et al. (2005). Evidence that the mitochondrial leucyl tRNA synthetase (LARS2) gene represents a novel type 2 diabetes susceptibility gene. *Diabetes* 54, 1892–1895. doi: 10.2337/diabetes.54.6.1892

Barak, O. F., Mladinov, S., Hoiland, R. L., Tremblay, J. C., Thom, S. R., Yang, M., et al. (2017). Disturbed blood flow worsens endothelial dysfunction in moderate-severe chronic obstructive pulmonary disease. *Sci. Rep*. 7:16929. doi: 10.1038/s41598-017-17249-6

Barnes, P. J. (2016). Sex differences in chronic obstructive pulmonary disease mechanisms. *Am. J. Respir Crit. Care Med*. 193, 813–814. doi: 10.1164/rccm.201512-2379ED

Booth, A. D., Jayne, D. R., Kharbanda, R. K., McEniery, C. M., Mackenzie, I. S., Brown, J., et al. (2004). Infliximab improves endothelial dysfunction in systemic vasculitis: a model of vascular inflammation. *Circulation* 109, 1718–1723. doi: 10.1161/01.CIR.0000124720.18538.DD

Budhiraja, R., Tuder, R. M., and Hassoun, P. M. (2004). Endothelial dysfunction in pulmonary hypertension. *Circulation* 109, 159–165. doi: 10.1161/01.CIR.0000102381.57477.50

Carminho-Rodrigues, M. T., Klee, P., Laurent, S., Guipponi, M., Abramowicz, M., Cao-van, H., et al. (2020). LARS2-Perrault syndrome: a new case report and literature review. *BMC Med. Genet*. 21:109. doi: 10.1186/s12881-020-01028-8

Cereda, C. W., Tamisier, R., Manconi, M., Andreotti, J., Frangi, J., Pifferini, V., et al. (2013). Endothelial dysfunction and arterial stiffness in ischemic stroke: the role of sleep-disordered breathing. *Stroke* 44, 1175–1178. doi: 10.1161/STROKEAHA.111.000112

Chella Krishnan, K., Mehrabian, M., and Lusis, A. J. (2018). Sex differences in metabolism and cardiometabolic disorders. *Curr. Opin. Lipidol*. 29, 404–410. doi: 10.1097/MOL.0000000000000536

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14:128. doi: 10.1186/1471-2105-14-128

Dabiré, H., Barthélémy, I., Blanchard-Gutton, N., Sambin, L., Sampedrano, C. C., Gouni, V., et al. (2012). Vascular endothelial dysfunction in Duchenne muscular dystrophy is restored by bradykinin through upregulation of eNOS and nNOS. *Basic Res. Cardiol*. 107:240. doi: 10.1007/s00395-011-0240-6

Davis, E. J., Lobach, I., and Dubal, D. B. (2019). Female XX sex chromosomes increse surviva and extend lifespan in aging mice. *Aging Cell* 18:e12871. doi: 10.1111/acel.12871

DECODE Study Group (2003). Age- and sex-specific prevalences of diabetes and impaired glucose regulation in 13 European Cohorts. *Diabetes Care* 26, 61–69. doi: 10.2337/diacare.26.1.61

Derada Troletti, C., Fontijn, R. D., Gowing, E., Charabati, M., van Het Hof, B., and Didouh, I. (2019). Inflammation-induced endothelial to mesenchymal transition promotes brain endothelial cell dysfunction and occurs during multiple sclerosis pathophysiology. *Cell Death Dis*. 10:45. doi: 10.1038/s41419-018-1294-2

Donato, A. J., Morgan, R. G., Walker, A. E., and Lesniewski, L. A. (2015). Cellular and molecular biology of aging endothelial cells. *J. Mol. Cell. Cardiol*. 89(Pt B):122–135. doi: 10.1016/j.yjmcc.2015.01.021

Fan, H., Hall, P., Santos, L. L., Gregory, J. L., Fingerle-Rowson, G., Bucala, R., et al. (2011). Macrophage migration inhibitory factor and CD74 regulate macrophage chemotactic responses via MAPK and Rho GTPase. *J. Immunol*. 186, 4915–4924. doi: 10.4049/jimmunol.1003713

Feng, W., Chen, L., Nguyen, P. K., Wu, S. M., and Li, G. (2019). Single cell analysis of endothelial cells identified organ-specific molecular signatures and heart-specific cell populations and molecular features. *Front. Cardiovasc. Med*. 6:165. doi: 10.3389/fcvm.2019.00165

Franco, C., Jones, M., Bernnabeu, M., and Vion, A. (2016). Non-canonical Wnt signaling modulates the endothelial shear stress flow in vascular remodeling. *Elife* 5:e07727. doi: 10.7554/eLife.07727

Furman, D., Campisi, J., Verdin, E., Carrera-Bastos, P., Targ, S., Franceschi, C., et al. (2019). Chronic inflammation in the etiology of disease across the life span. *Nat. Med*. 25, 1822–1832. doi: 10.1038/s41591-019-0675-0

Gillis, E. E., and Sullivan, J. C. (2016). Sex differences in hypertension: recent advances. *Hypertension* 68, 1322–1327. doi: 10.1161/HYPERTENSIONAHA.116.06602

Gil-Yarom, N., Radomir, L., Sever, L., Kramer, M. P., Lewinsky, H., Bornstein, C., et al. (2017). Cd74 is a novel transcription regulator. *Proc. Natl. Acad. Sci. U.S.A*. 114, 562–567. doi: 10.1073/pnas.1612195114

Gutierrez, E., Flammer, A. J., Lerman, L. O., Elizaga, J., Lerman, A., and Fernandez-Aviles, F. (2013). Endothelial dysfunction over the course of coronary artery disease. *Eur. Heart J*. 34, 3175–3181. doi: 10.1093/eurheartj/eht351

Halai, K., Whiteford, J., Ma, B., Nourshargh, S., and Woodfin, A. (2014). ICAM-2 facilitates luminal interactions between neutrophils and endothelial cells *in vivo*. *J. Cell Sci*. 127(Pt 3):620–629. doi: 10.1242/jcs.137463

Hoffmann, J., Haendeler, J., Aicher, A., Rössig, L., Vasa, M., Zeiher, A. M., et al. (2001). Aging enhances the sensitivity of endothelial cells toward apoptotic stimuli: important role of nitric oxide. *Circ. Res*. 89, 709–715. doi: 10.1161/hh2001.097796

Huang, M. T., Mason, J. C., Birdsey, G. M., Amsellem, V., Gerwin, N., and Haskard, D. (2005). Endothelial intercellular adhesion molecule (ICAM)-2 regulates angiogenesis. *Blood* 106, 1636–1643. doi: 10.1182/blood-2004-12-4716

InanlooRahatloo, K., Liang, G., Vo, D., Ebert, A., Nguyen, I., and Nguyen, P. K. (2017). Sex-based differences in myocardial gene expression in recently deceased organ donors with no prior cardiovascular disease. *PLoS ONE* 12:e0183874. doi: 10.1371/journal.pone.0183874

Johnson, R. J., and Nangaku, M. (2016). Endothelial dysfunction: the secret agent driving kidney disease. *J. Am. Soc. Nephrol*. 27, 3–5. doi: 10.1681/ASN.2015050502

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

Kassam, I., Wu, Y., Yang, J., Visscher, P. M., and McRae, A. F. (2019). Tissue-specific sex differences in human gene expression. *Hum. Mol. Genet*. 28, 2976–2986. doi: 10.1093/hmg/ddz090

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Lakshmanan, S., Jankowich, M., Wu, W. C., Blackshear, C., Abbasi, S., and Choudhary, G. (2020). Gender differences in risk factors associated with pulmonary artery systolic pressure, heart failure, and mortality in blacks: Jackson Heart Study. *J. Am. Heart Assoc.* 9:e013034. doi: 10.1161/JAHA.119.013034

Le Hiress, M., Tu, L., Ricard, N., Phan, C., Thuillet, R., Fadel, E., et al. (2015). Proinflammatory signature of the dysfunctional endothelium in pulmonary hypertension. Role of the macrophage migration inhibitory factor/CD74 complex. *Am. J. Respir Crit. Care Med.* 192, 983–997. doi: 10.1164/rccm.201402-0322OC

Lin, J. R., Shen, W. L., Yan, C., and Gao, P. J. (2015). Downregulation of dynamin-related protein 1 contributes to impaired autophagic flux and angiogenic function in senescent endothelial cells. *Arterioscler. Thromb. Vasc. Biol.* 35, 1413–1422. doi: 10.1161/ATVBAHA.115.305706

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell* 153, 1194–1217. doi: 10.1016/j.cell.2013.05.039

MacDonald, B. T., Tamai, K., and He, X. (2009). Wnt/beta-catenin signaling: components, mechanisms, and diseases. *Dev. Cell.* 17, 9–26. doi: 10.1016/j.devcel.2009.06.016

Madhavan, M. V., Gersh, B. J., Alexander, K. P., Granger, C. B., and Stone, G. W. (2018). Coronary artery disease in patients ≥80 years of age. *J. Am. Coll. Cardiol.* 71, 2015–2040. doi: 10.1016/j.jacc.2017.12.068

McCormick, M. M., Rahimi, F., Bobryshev, Y. V., Gaus, K., Zreiqat, H., Cai, H., et al. (2005). S100A8 and S100A9 in human arterial wall. Implications for atherogenesis. *J. Biol. Chem.* 280, 41521–4159. doi: 10.1074/jbc.M509442200

Molema, G. (2010). Heterogeneity in endothelial responsiveness to cytokines, molecular causes, and pharmacological consequences. *Semin. Thromb. Hemost.* 36, 246–264. doi: 10.1055/s-0030-1253448

Muller, A. M., Hermanns, M. I., Cronen, C., and Kirkpatrick, C. J. (2002). Comparative study of adhesion molecule expression in cultured human macro- and microvascular endothelial cells. *Exp. Mol. Pathol.* 73, 171–180. doi: 10.1006/exmp.2002.2446

Nguyen, P. K., Nag, D., and Wu, J. C. (2011). Sex differences in the diagnostic evaluation of coronary artery disease. *J. Nucl. Cardiol.* 18, 144–152. doi: 10.1007/s12350-010-9315-2

Palmer, B. F., and Clegg, D. J. (2015). The sexual dimorphism of obesity. *Mol. Cell. Endocrinol.* 402, 113–119. doi: 10.1016/j.mce.2014.11.029

Park, S. G., Schimmel, P., and Kim, S. (2008). Aminoacyl tRNA synthetases and their connections to disease. *Proc. Natl. Acad. Sci. U.S.A.* 105, 11043–11049. doi: 10.1073/pnas.0802862105

Perticone, F., Maio, R., Perticone, M., Sciacqua, A., Shehaj, E., Naccarato, P., et al. (2010). Endothelial dysfunction and subsequent decline in glomerular filtration rate in hypertensive patients. *Circulation* 122, 379–384. doi: 10.1161/CIRCULATIONAHA.110.940932

Riley, L. G., Rudinger-Thirion, J., Schmitz-Abe, K., Thorburn, D. R., Davis, R. L., Teo, J., et al. (2016). LARS2 variants associated with hydrops, lactic acidosis, sideroblastic anemia, and multisystem failure. *JIMD Rep.* 28, 49–57. doi: 10.1007/8904_2015_515

Rudnicki, M., Abdifarkosh, G., Rezvan, O., Nwadozi, E., Roudier, E., and Haas, T. L. (2018). Female mice have higher angiogenesis in perigonadal adipose tissue than males in response to high-fat diet. *Front. Physiol.* 9:1452. doi: 10.3389/fphys.2018.01452

Sorci, G., Riuzzi, F., Giambanco, I., and Donato, R. (2013). RAGE in tissue homeostasis, repair and regeneration. *Biochim. Biophys. Acta* 1833, 101–109. doi: 10.1016/j.bbamcr.2012.10.021

Sroussi, H. Y., Kohler, G. A., Agabian, N., Villines, D., and Palefsky, J. M. (2009). Substitution of methionine 63 or 83 in S100A9 and cysteine 42 in S100A8 abrogate the antifungal activities of S100A8/A9: potential role for oxidative regulation. *FEMS Immunol. Med. Microbiol.* 55, 55–61. doi: 10.1111/j.1574-695X.2008.00498.x

Tabula Muris Consortium (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* 583, 590–595. doi: 10.1038/s41586-020-2496-1

Tabula Muris Consortium, Overall Coordination, Logistical Coordination, Organ Collection and Processing, Library Preparation and Sequencing, Computational Data Analysis, et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562, 367–372. doi: 10.1038/s41586-018-0590-4

Tay, L., Ding, Y. Y., Leung, B. P., Ismail, N. H., Yeo, A., Yew, S., et al. (2015). Sex-specific differences in risk factors for sarcopenia amongst community-dwelling older adults. *Age* 37:121. doi: 10.1007/s11357-015-9860-3

Timmerman, K. L., and Volpi, E. (2013). Endothelial function and the regulation of muscle protein anabolism in older adults. *Nutr. Metab. Cardiovasc. Dis.* 23(Suppl. 1), S44–S50. doi: 10.1016/j.numecd.2012.03.013

Turovskaya, O., Foell, D., Sinha, P., Vogl, T., Newlin, R., Nayak, J., et al. (2008). RAGE, carboxylated glycans and S100A8/A9 play essential roles in colitis-associated carcinogenesis. *Carcinogenesis* 29, 2035–2043. doi: 10.1093/carcin/bgn188

Uraoka, M., Ikeda, K., Kurimoto-Nakano, R., Nakagawa, Y., Koide, M., Akakabe, Y., et al. (2011). Loss of bcl-2 during the senescence exacerbates the impaired angiogenic functions in endothelial cells by deteriorating the mitochondrial redox state. *Hypertension* 58, 254–263. doi: 10.1161/HYPERTENSIONAHA.111.176701

Vogl, T., Tenbrock, K., Ludwig, S., Leukert, N., Ehrhardt, C., Van Zoelen, M. A., et al. (2007). Mrp8 and Mrp14 are endogenous activators of toll-like receptor 4, promoting lethal, endotoxin-induced shock. *Nat. Med.* 13, 1042–1049. doi: 10.1038/nm1638

Xia, C., Braunstein, Z., Toomey, A. C., Zhong, J., and Rao, X. (2018). S100 proteins as an important regulator of macrophage inflammation. *Front. Immunol.* 8:1908. doi: 10.3389/fimmu.2017.01908

Xu, Y., He, Z., Song, M., Zhou, Y., and She, Y. (2019). A microRNA switch controls dietary restriction-induced longevity through Wnt signaling. *EMBO Rep.* 20:e46888. doi: 10.15252/embr.201846888

Yao, D., and Brownlee, M. (2010). Hyperglycemia-induced reactive oxygen species increase expression of the receptor for advanced glycation end products (RAGE) and RAGE ligands. *Diabetes* 59, 249–255. doi: 10.2337/db09-0801

Yen, T., Harrison, C. A., Devery, J. M., Leong, S., Iismaa, S. E., Yoshimura, T., et al. (1997). Induction of the S100 chemotactic protein, CP-10, in murine microvascular endothelial cells by proinflammatory stimuli. *Blood* 90, 4812–4821. doi: 10.1182/blood.V90.12.4812

Yu, M. K., Lyles, C. R., Bent-Shaw, L. A., Young, B. A., and the Pathways Authors (2012). Risk factor, age and sex differences in chronic kidney disease prevalence in a diabetic cohort: the pathways study. *Am. J. Nephrol.* 36, 245–251. doi: 10.1159/000342210

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 February 2021
doi: 10.3389/fgene.2020.580138

Construction of a MicroRNA-Based Nomogram for Prediction of Lung Metastasis in Breast Cancer Patients

Leyi Zhang[1,2,3†], Jun Pan[1,2,3†], Zhen Wang[1,2,3], Chenghui Yang[1,2,3] and Jian Huang[1,2,3]*

[1]Key Laboratory of Tumor Microenvironment and Immune Therapy of Zhejiang Province, Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [2]Cancer Institute (Key Laboratory of Cancer Prevention Intervention, National Ministry of Education), Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [3]Department of Breast Surgery, Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

OPEN ACCESS

Edited by:
Yuanwei Zhang, University of Science and Technology of China, China

Reviewed by:
Xinhua Xie, Sun Yat-sen University Cancer Center (SYSUCC), China
Jiangyang Zhao, ATGC Inc, United States

***Correspondence:**
Jian Huang
drhuangjian@zju.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *10 July 2020*
Accepted: *30 December 2020*
Published: *19 February 2021*

Citation:
Zhang L, Pan J, Wang Z, Yang C and Huang J (2021) Construction of a MicroRNA-Based Nomogram for Prediction of Lung Metastasis in Breast Cancer Patients. Front. Genet. 11:580138. doi: 10.3389/fgene.2020.580138

The lung is one of the most common sites of distant metastasis in breast cancer (BC). Identifying ideal biomarkers to construct a more accurate prediction model than conventional clinical parameters is crucial. MicroRNAs (miRNAs) data and clinicopathological data were acquired from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database. miR-663, miR-210, miR-17, miR-301a, miR-135b, miR-451, miR-30a, and miR-199a-5p were screened to be highly relevant to lung metastasis (LM) of BC patients. The miRNA-based risk score was developed based on the logistic coefficient of the individual miRNA. Univariate and multivariate logistic regression selected tumor node metastasis (TNM) stage, age at diagnosis, and miRNA-risk score as independent predictive parameters, which were used to construct a nomogram. The Cancer Genome Atlas (TCGA) database was used to validate the signature and nomogram. The predictive performance of the nomogram was compared to that of the TNM stage. The area under the receiver operating characteristics curve (AUC) of the nomogram was higher than that of the TNM stage in all three cohorts (training cohort: 0.774 vs. 0.727; internal validation cohort: 0.763 vs. 0.583; external validation cohort: 0.925 vs. 0.840). The calibration plot of the nomogram showed good agreement between predicted and observed outcomes. The net reclassification improvement (NRI), integrated discrimination improvement (IDI), and decision-curve analysis (DCA) of the nomogram showed that its performances were better than that of the TNM classification system. Functional enrichment analyses suggested several terms with a specific focus on LM. Subgroup analysis showed that miR-30a, miR-135b, and miR-17 have unique roles in lung metastasis of BC. Pan-cancer analysis indicated the significant importance of eight predictive miRNAs in lung metastasis. This study is the first to establish and validate a comprehensive lung metastasis predictive nomogram based on the METABRIC and TCGA databases, which provides a reliable assessment tool for clinicians and aids in appropriate treatment selection.

Keywords: breast cancer, lung metastasis, microRNA, nomogram, the cancer genome atlas, METABRIC dataset, risk score

INTRODUCTION

Breast cancer (BC) is the most common cancer diagnosed (excluding skin cancers) and is the second leading cause of cancer death among United States women (DeSantis et al., 2019) and worldwide. Most BC-related deaths are caused by distant metastases, which become lethal even after the primary lesion being removed (Knott et al., 2018). BC tends to metastasize distantly to the bone, brain, liver, lung, and distant lymph nodes. Lung metastases particularly tend to occur within the initial 5 years of BC diagnosis and significantly impact patients' prognosis (Medeiros and Allan, 2019). Therefore, it is of great clinical importance to select patients who are prone to have lung metastasis so that they can benefit from prevention treatment and early diagnosis.

Currently, the traditional tumor node metastasis (TNM) staging system is a standard tool for risk evaluation in clinical practice. However, BC patients with the same stage can have varying clinical outcomes (Wang et al., 2019). The TNM staging system is mainly based on anatomical information, which fails to incorporate important pathological parameters and biological changes that happened in BC. The mechanisms of the lymphatic dissemination and hematogenous dissemination are different, which may be one of the reasons for the poor metastasis prediction ability of the TNM staging system. Hence, new methods to identify patients who are likely to have lung metastasis are needed.

MicroRNAs (miRNAs) are small, non-coding single-stranded RNAs (18–25 nucleotides) and negatively regulate gene expression by binding to complementary sequences in the 3' untranslated region (3' UTR) of mRNAs (Lin and Gregory, 2015). Accumulating evidence suggests that miRNAs play critical roles in various physiological and pathological processes, including many proposed mechanisms of cancer metastasis (Pencheva and Tavazoie, 2013). Previous studies have presented the association of certain miRNAs with lung metastasis, including miR-629-3p, miR-106b-5p, and so on (Schrijver et al., 2017; Wang et al., 2017). However, due to the biological heterogeneity of BC, a comprehensive prediction model incorporating multiple biomarkers, rather than a single parameter, can improve predictive accuracy. Nomograms constructed on the basis of known predictive variables are being widely used to predict the specific outcome for an individual patient (Iasonos et al., 2008). There have been reports that clinical variables-based nomogram and miRNA signature could be used to predict distant metastasis in BC patients (Delpech et al., 2015; Rohan et al., 2019), yet there is no literature concerning comprehensive lung metastasis prediction model. We hypothesized that our new model based on the combination of predictive miRNAs and clinicopathological variables could improve the accuracy in predicting lung metastasis and prolong survival in BC patients.

Therefore, the purpose of this study was to establish and validate a comprehensive nomogram that incorporated both the miRNAs signature and clinical-related risk features for the individual prediction of lung metastasis status of BC patients. The new prediction model was compared with the traditional TNM staging system in order to determine its reliability. Aiding with this model, clinicians might be able to evaluate the lung metastasis risk of BC patients, thus choosing appropriate medical examinations and optimizing therapeutic regimen.

MATERIALS AND METHODS

Datasets Selection and Data Processing

To identify lung metastasis-related miRNA and mRNA in BC, public datasets with matched miRNA, mRNA, and clinical data were used in this study. A European Genome-phenome Archive (RRID: SCR_004944),[1] EGAS00000000122 (Molecular Taxonomy of Breast Cancer International Consortium, METABRIC miRNA landscape; Curtis et al., 2012; Dvinge et al., 2013), contains a total of 1,302 BC patients with matched mRNA (EGAD00010000434) and miRNA (EGAD00010000438) data. The inclusion criteria included: (1) samples had lung metastasis or no metastasis (NM); (2) samples had both mRNA and miRNA expression data; and (3) samples had intact clinical data. Around 439 patients were selected in subsequent analysis. Among them (n = 439), 327 samples were randomly assigned as a training cohort and the rest were assigned as an internal validation cohort based on a computer number generator (**Supplementary Table S1**). About 449 of 1,109 BC patients from The Cancer Genome Atlas (TCGA) dataset (RRID: SCR_003193) were selected according to the same inclusion criteria as an external validation cohort (Network, 2012; **Supplementary Table S1**).[2] The method of acquisition and application complied with the guidelines and policies. It is not necessary to obtain informed patient consent for data obtained from the METABRIC and TCGA databases since they do not include information that can be used to identify individual patients.

Development of a miRNA-Based Risk Score

Among the 439 BC patients in the METABRIC dataset, two subsets of patients were defined based on their metastasis status: a lung metastasis group (those who had lung metastasis) and an NM group (those who did not report metastasis until the last follow-up). We identified 853 miRNAs annotated in the METABRIC dataset, and differentially expressed miRNAs (DEmiRNAs) between the two groups were identified using the *LIMMA* package of R (Ritchie et al., 2015; LIMMA, RRID: SCR_010943). Of the top 20 DEmiRNAs with the most significant foldchanges, four miRNAs were dropped from highly correlated pairs (r > 0.8, Wei and Simko, 2017). The least absolute shrinkage and selection operator (LASSO) method (Friedman et al., 2010) was used to select the most useful predictive miRNAs from the 16 lung metastasis-related DEmiRNAs in the training cohort and constructed an eight-miRNA based risk score for predicting lung metastasis status of BC patients in the training set. The risk score was calculated

[1]https://www.ebi.ac.uk/ega/home
[2]https://portal.gdc.cancer.gov/

for each patient *via* a linear combination of selected miRNAs that were weighted by their respective coefficients.

$$Risk\ score = \sum_{i=1}^{8} \beta_i \times Exp_i$$

An optimal cut-off point was determined using receiver operating characteristic (ROC) curve, to classify samples into low (≤0.168) and high risk (>0.168) group. The Kaplan-Meier (KM) survival analysis with a log-rank test was implemented to compare the survival difference between the two groups (Kassambara et al., 2017). Then KM analysis with the log-rank test was also implemented to show the relationship between the expression of predictive miRNAs and prognosis in external validation cohort.

Construction and Validation of miRNA-Based LM Predictive Nomogram

Univariate logistic regression analysis was performed to compare the predictive power of the eight-miRNA risk score and clinical parameters including age at diagnosis, tumor size, TNM stage, grade, estrogen receptor (ER) status, progesterone receptor (PR) status, human epidermal growth factor receptor 2 (HER2) status, and hormone therapy. Furthermore, we used a multivariate logistic regression analysis to determine whether the eight-miRNA risk score could be an independent predictive factor for lung metastasis in BC patients. Other clinical parameters with values of p less than 0.1 in the univariate logistic regression analysis were also incorporated in the analysis. A composite nomogram was constructed based on all independent predictive parameters screened by multivariate logistic regression analysis above to predict the rate of lung metastasis (Harrell, 2013), and to be a graphic representation of the prediction model.

The ROC curves were plotted to assess the sensitivity and the specificity of independent predictive parameters including eight-miRNA signature, age at diagnosis, TNM stage, and miRNA-based nomogram in predicting lung metastasis (Sing et al., 2005). The area under the receiver operating characteristics curve (AUC) was also calculated to make a comparison for the discriminatory ability of the above predictive parameters. Calibration curves were implemented to assess the calibration ability of the miRNA-based nomogram, accompanied by the Hosmer-Lemeshow test (Kramer and Zimmerman, 2007). The predicted and observed outcomes of the nomogram could be compared in the calibration curve, while the 45-degree diagonal line represented the ideal prediction. The net reclassification improvement (NRI) and integrated discrimination improvement (IDI) were used to quantify the improvement in sensitivity and specificity offered by our miRNA-based nomogram compared to the TNM staging system (Kundu et al., 2011). NRI was based on reclassification tables composed of patients with and without events and could quantify the correct reclassification in categories (Pencina et al., 2011). IDI summarized the extent to which a new model increased risk in patients with events and decreased risk in patients without events (Pencina et al., 2008; Chipman and Braun, 2017). We used decision-curve analysis (DCA) to test the clinical applicability of our miRNA-based nomogram model by quantifying the net benefits at different threshold probabilities. DCA was conducted by adding the benefits (true positives) and subtracting the harms (false positives; Vickers and Elkin, 2006; Vickers et al., 2008).

Identification of Potential Targets for Predictive miRNAs and Construction a miRNA-mRNA Network Associated With Lung Metastasis

The target genes of eight predictive miRNAs were first predicted and analyzed using miRWalk3.0 (RRID: SCR_016509; Sticht et al., 2018),[3] miRDB (RRID: SCR_010848; Chen and Wang, 2020),[4] TargetScan (RRID: SCR_010845; Nam et al., 2014),[5] and miRTarBase (RRID: SCR_017355; Chou et al., 2018).[6] An mRNA would be considered as a target of a miRNA if the mRNA was predicted to be the target in all three *in silico* prediction algorithms (miRWalk, miRDB, and miRTarBase) or could be found in a experimentally validated database (miRTarBase). We also acquired matched mRNA transcriptome data (RRID: SCR_004944, EGAD00010000434) of the patients enrolled in the analysis of identifying DEmiRNAs.[7] 3,791 differentially expressed mRNAs (DEmRNAs) between the lung metastasis group and no metastasis group were identified using the *LIMMA* package of R. CytoHubba plugin (RRID: SCR_017677) in Cytoscape (RRID: SCR_003032) was used to predict the hub genes among the target genes of upregulated or downregulated miRNAs (Chin et al., 2014).[8] miRNA-mRNA networks were also visualized with the Cytoscape software.

Functional Enrichment Analysis of Target Genes of Predictive miRNAs

For the screened overlapped target genes of each miRNA separately or hub genes for upregulated or downregulated miRNAs, gene ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways analysis were performed (clusterProfiler, RRID: SCR_016884; Yu et al., 2012; Walter et al., 2015). Statistically significant GO and KEGG terms ($p < 0.05$) related to cancer and metastasis were identified.

Identification of miRNAs Unique to Lung Metastasis or BC

MicroRNA transcriptome data of BC patients from the TCGA dataset were selected to perform two differential miRNA expression analyses between different subgroups of BC patients. Around 48 DEmiRNAs between patients with lung metastasis only and patients with distant metastasis (except for the lung) were identified using the DESeq2 package of R (DESeq2, RRID: SCR_015687; Love et al., 2014). Around 90 DEmiRNAs between patients with distant metastasis (except for the lung) and patients without metastasis were identified.

[3]http://mirwalk.umm.uni-heidelberg.de/
[4]http://www.mirdb.org/
[5]http://www.targetscan.org/vert_72/
[6]http://mirtarbase.cuhk.edu.cn/php/index.php
[7]https://www.ebi.ac.uk/ega/home
[8]http://cytoscape.org

The miRNA expression data and corresponding clinical data of the patients of six cancer types [adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), sarcoma (SARC), skin cutaneous melanoma (SKCM), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), and stomach adenocarcinoma (STAD)] were downloaded from the TCGA database. DEmiRNAs between patients with lung metastasis and patients without metastasis were identified in each type of cancer using the DESeq2 package of R.

Statistical Analysis

All the statistical analyses were performed with the SPSS software (RRID: SCR_002865) and R software (version 4.0.0; RRID: SCR_001905).[9,10] A two-sided probability value of $p < 0.05$ was considered to be statistically significant.

RESULTS

Demographic and Clinicopathological Characteristics

A total of 479 BC patients from METABRIC and 449 BC patients from TCGA were included in this study. Baseline clinical and pathological characteristics of the study participants in the training and two validation cohorts were listed in **Table 1**. The median age of patients was 61.11, 60.57, and 60 years in the training and two validation cohorts, respectively. The rates of lung metastasis were 8.26, 7.24, and 3.56% in the training and two validation cohorts, respectively.

Identification Candidate Lung Metastasis-Related miRNAs in the Training Cohort

The METABRIC dataset includes 1,302 BC samples, of which 479 (36.79%) of them reached the inclusion criteria for the analysis of identifying DEmiRNAs. About 327 samples were randomly assigned as a training cohort and the rest were assigned as the internal validation cohort based on a computer number generator. The flow chart of the study design was showed in **Figure 1**. A total of 184 miRNAs ($p < 0.05$) were identified to be differentially expressed between patients with lung metastasis and patients without metastasis (**Figure 2A**; **Supplementary Table S2**). Around 20 most significantly upregulated and downregulated miRNAs were selected to conduct correlation analysis (upregulated in lung metastasis patients: miR-663, miR-210, miR-1,202, miR-1973, miR-17, miR-18a, miR-301a, miR-135b, miR-20a, miR-17*; down-regulated in lung metastasis patients: miR-451, miR-26b, miR-199b-5p, miR-30a*, miR-10a, miR-10b, miR-30a, miR-199a-3p, miR-199a-5p, and miR-99a; **Supplementary Figure S1**). Four miRNAs (miR-30a*, has-miR-199a-3p, miR-99a, and miR-18a) were dropped from highly correlated pairs ($r > 0.8$) to reduce multicollinearity and improve stability for subsequent model selection.

[9]http://www-01.ibm.com/software/uk/analytics/spss/
[10]http://www.Rproject.org

TABLE 1 | Demographics of the samples chosen for the study.

Variables	Training cohort (n = 327)	Internal validate cohort (n = 152)	External validate cohort (n = 449)
Median age at diagnosis in years (IQR)	61.11 (51.09–68.99)	60.57 (50.94–70.25)	60.00 (71.00–67.00)
Median follow up time from diagnosis in days (IQR)	3,318 (1916–4,719)	3,144 (1781–4,479)	343.5 (114–1,108)
Lung metastasis status			
No metastasis	300	141	433
Lung metastasis	27	11	16
Pam50 subtype			
Luminal A	151	78	205
Luminal B	83	37	66
HER2	26	7	21
Basal like	46	22	100
Normal breast-like	21	8	14
Unknown	0	0	43
TNM stage			
1	203	93	152
2	114	58	281
3	8	1	14
4	2	0	2
ER status			
Positive	259	118	304
Negative	68	34	124
Unknown	0	0	21
PR status			
Positive	187	87	270
Negative	142	65	156
Unknown	0	0	23
HER2 status			
Positive	41	9	53
Negative	286	143	248
Unknown			148
Menopausal state			
Pre	71	32	84
Post	256	120	312
Peri	0	0	19
Unknown	0	0	34
Vital status			
Alive	198	86	435
Dead	129	66	14

PAM50, prediction analysis of microarray 50; ER, estrogen receptor; PR, progesterone receptor; HER2, human epithelial growth factor receptor 2; and TNM, the tumor node metastasis.

Development of an Eight-miRNA Signature to Distinguish Lung Metastasis Status in BC Patients

In the training cohort, we used LASSO-based logistic regression and identified eight miRNAs from the 16 DEmiRNAs, which were as follows: miR-663, miR-210, miR-17, miR-301a, miR-135b, miR-451, miR-30a, and miR-199a-5p (**Figures 2B,C**). The eight-miRNA based risk score was calculated based on their logistic coefficients. An optimal cut-off point was determined according to ROC. We then divided samples into a low-risk (risk score $\leq$ 0.168) and a high-risk (risk score > 0.168) group. The distributions of the miRNA-based risk score, overall survival (OS), OS status, and the expression

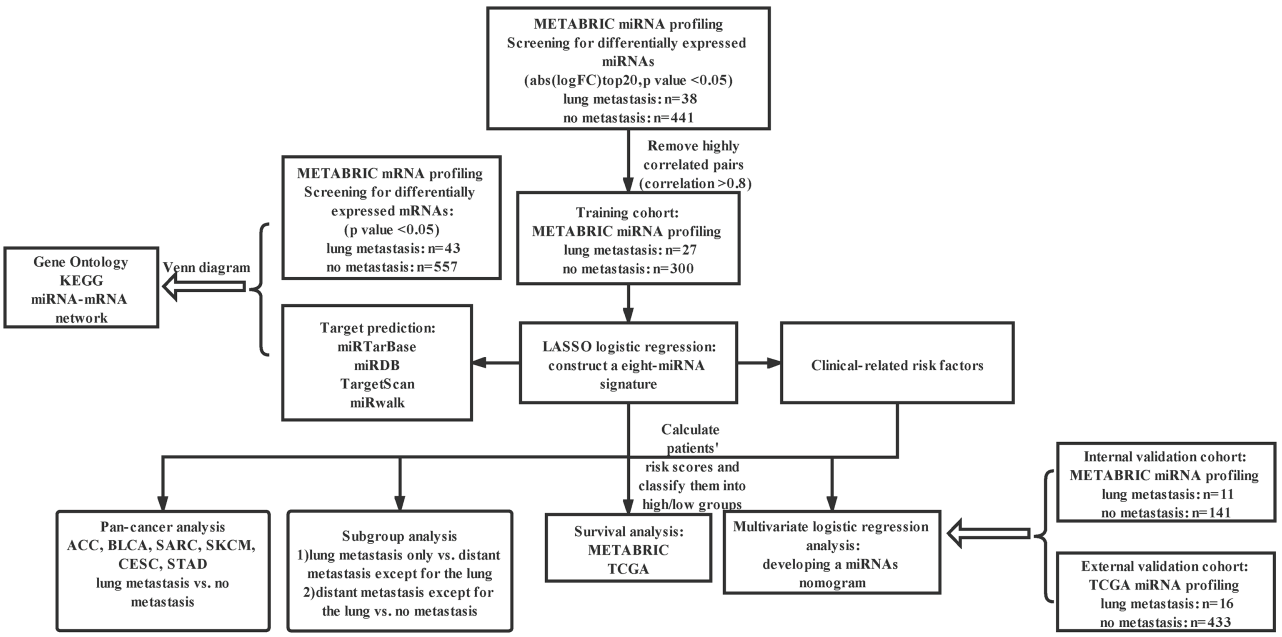


FIGURE 1 | Study design. METABRIC, molecular taxonomy of BC international consortium; LASSO, the least absolute shrinkage and selector operation; KEGG, Kyoto encyclopedia of genes and genomes; Abs, absolute value; FC, fold change; miRNA, microRNA; and TCGA, the cancer genome atlas.

profiles of eight miRNAs in the training cohort were shown in **Figure 2D**. The five risky upregulated miRNAs identified in lung metastasis cases exhibited high expression in the high-risk group and the three protective downregulated miRNAs had high expression in the low-risk group. And the patients with higher risk scores tended to have poorer prognoses, yet failed to reach a significant level (p = 0.078) (**Figure 2E**). Age stratified analysis indicated that miRNAs-based risk score predicted prognosis well in people aged 45–70 years (**Supplementary Figure S2**).

Establishment of a Nomogram for Predicting Lung Metastasis Status Incorporating miRNAs Signature and Clinical-Related Factors

In the training cohort, according to the results of univariate logistic regression analysis, the eight-miRNA signature, and five clinical risk factors (age at diagnosis, tumor size, grade, TNM stage, and HER2 status) with values of p less than 0.1 were included in multivariate regression analysis for assessing the independent risk factors for lung metastasis (**Table 2**). A multivariate logistic regression analysis was used to develop a nomogram model and found age at diagnosis, TNM stage, and the eight-miRNA signature significantly increased the likelihood of lung metastasis (**Figure 3**). The AUC of the miRNA-based nomogram model was 0.774 (95% CI, 0.669–0.879) in the training cohort (**Table 3**; **Figure 4A**). The calibration curve of the miRNA-based nomogram was very close to the standard 45-degree diagonal line, which showed good calibration in the training cohort (**Figure 4D**).

Assessment of the Eight-miRNA Signature and Nomogram Model in Validation Cohorts

We then examined the predictive ability of our eight-miRNA signature and nomogram model in two validation cohorts. The distributions of the miRNA-based risk score, OS, OS status, and the expression profiles of predictive miRNAs in the internal validation cohorts have been shown in **Supplementary Figure S3A**. The eight-miRNA signature and miRNA-based nomogram model displayed an AUC of 0.754 (95% CI, 0.561–0.946) and 0.763 (95% CI, 0.597–0.929) for lung metastasis risk prediction, respectively (**Table 3**; **Figure 4B**). The calibration curve of the miRNA-based nomogram also exhibited favorable accordance between the predicted risk and the actual risk in the internal validation cohort (**Figure 4E**).

An independent external validation cohort of 449 patients who fulfilled the same requirements as above was recruited from the TCGA dataset. A total of seven of the eight miRNAs identified in our study were found in the TCGA miRNA dataset (the exception being miR-663). The distributions of the miRNA-based risk score, OS, OS status, and the expression profiles of predictive miRNAs in the external validation cohorts has been shown in **Supplementary Figure S3B**. Among them, the elevated expression of four miRNAs was significantly associated with poorer OS and disease-free survival (DFS) (miR-210, miR-451a, miR-135b, and miR-17) (**Figures 5A–D,F–I**). In the meantime, the higher expression of miR-30a indicated better OS and DFS (**Figures 5E,J**). Due to the different sequence platforms used in the external validation cohort, the risk score of the external validation

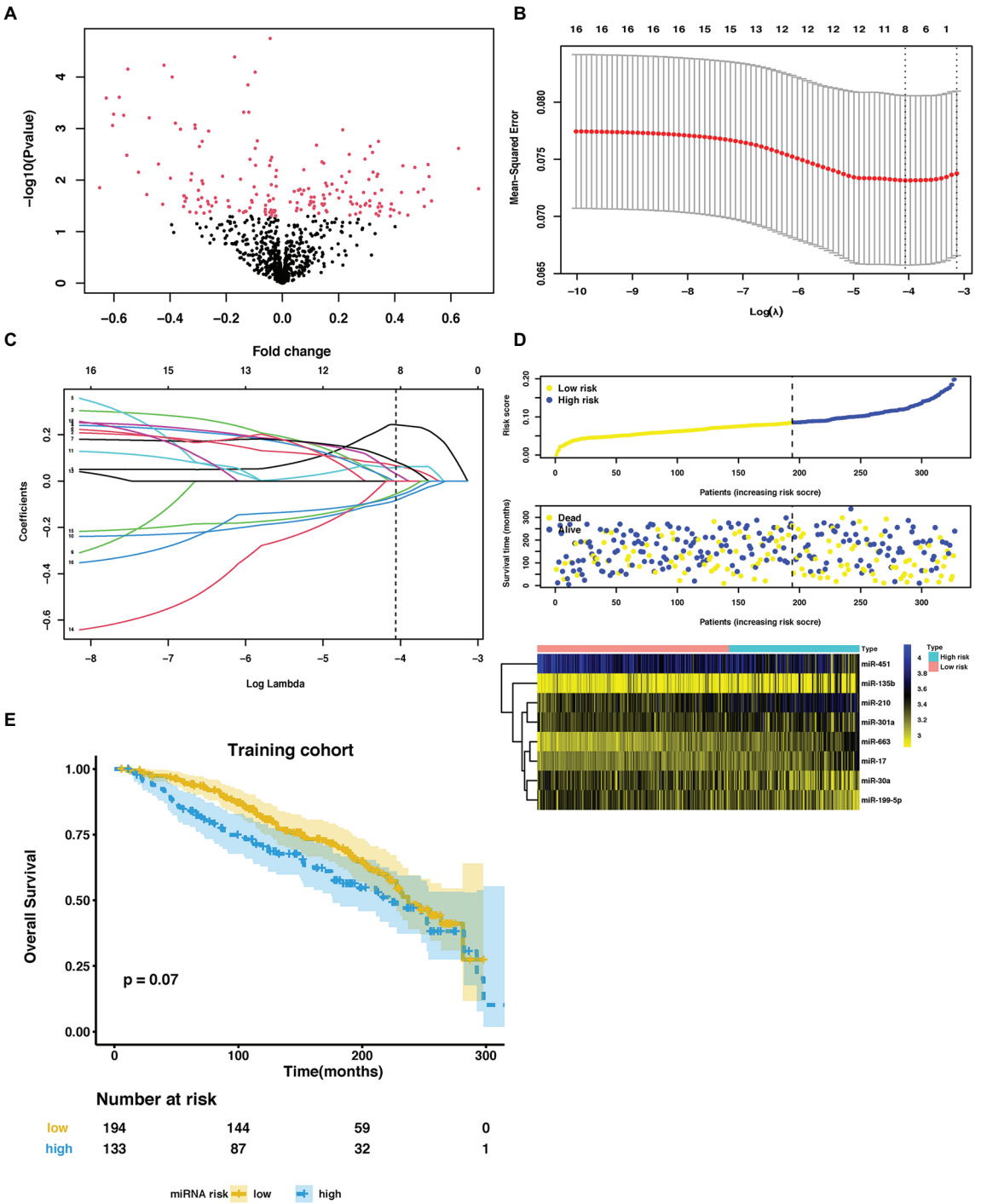


FIGURE 2 | Parameter selection to develop an eight-miRNA signature to distinguish lung metastasis status of breast cancer. **(A)** Volcano plot of miRNAs expression in the METABRIC dataset. **(B)** 3-fold cross-validation for parameter selection *via* minimum criteria in the LASSO model. Two dotted vertical lines were drawn at the optimal values by using the minimum criteria (the value of lambda that gives a minimum mean cross-validated error) and the one SE of the minimum criteria (the value of lambda that gives one SE away from the minimum error). **(C)** LASSO coefficient profiles of the 16 LM-related differentially expressed miRNAs (DEmiRNAs) in the training cohort. Each curve corresponds to a miRNA. The coefficient profile plot was against the log (lambda) sequence. The dotted vertical line was drawn at the value lambda = 0.01718646 selected by using 3-fold cross-validation *via* minimum criteria, where optimal lambda resulted in eight nonzero coefficients. **(D)** The distribution of risk score, overall survival (OS), vital status, and the expression profiles of eight-miRNA in the training cohort. **(E)** Kaplan-Meier (KM) curves of OS of breast cancer patients stratified by eight-miRNA risk score in the training cohort. METABRIC, molecular taxonomy of breast cancer international consortium; LASSO, the least absolute shrinkage and selector operation; miRNA, microRNA; LM, lung metastasis; and DEmiRNAs, differentially expressed miRNAs.

cohort was constructed using seven miRNAs. An optimal cut-off point was determined by ROC to dichotomize the samples into low and high-risk groups. Patients with higher miRNA risk scores tended to have a poorer prognosis than those with lower risk scores (**Figures 5K,L**). Other than predicting OS and DFS, the miRNA risk score was also significantly associated with the risk of lung metastasis in univariate and multivariate logistic regression analysis (**Table 4**). The miRNA signature and miRNA-based nomogram model displayed an AUC of 0.711 (95% CI, 0.608–0.815) and 0.925 (95% CI, 0.846–1.000) for the estimation of lung metastasis risk respectively (**Table 3**; **Figure 4C**). The calibration plot showed that the predicted risks of the nomogram were in good accordance with the actual risks (**Figure 4F**).

TABLE 2 | Risk factors for lung metastasis (LM) in training cohort.

	Univariate analysis		Multivariate analysis	
	OR (95% CI)	***p* value**	**OR (95% CI)**	***p* value**
miRNA score	1.898 (1.237–2.912)	0.0033	1.651 (1.046–2.606)	0.0311
Age at diagnosis	0.583 (0.330–1.020)	0.0587	0.486 (0.275–0.862)	0.0134
Tumor size	1.499 (1.148–1.958)	0.0030		
Grade	3.129 (0.824–11.884)	0.094		
TNM stage	3.494 (1.905–6.407)	<0.0001	4.025 (2.078–7.795)	<0.0001
ER status	0.738 (0.298–1.824)	0.511		
PR status	1.085 (0.487–2.416)	0.842		
HER2 status	2.759 (1.087–7.005)	0.0328		
Hormone therapy	0.563 (0.253–1.254)	0.1599		

LM, lung metastasis; miRNA, microRNA; ER, estrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2; and TNM, the tumor node metastasis.

Comparison With Other Prognostic Markers

Currently, the conventional TNM staging system is the standard tool for risk evaluation in clinical practice. When comparing the AUC, we found that the miRNA-based prediction nomogram achieved better predictive accuracy than the TNM stage in the training cohort and two validation cohorts (**Table 3**). NRI and IDI were employed to compare the discriminative ability between our model and the TNM stage. Compared the TNM stage alone, the NRI values for miRNA-based prediction nomogram were 0.216 (95% CI, 0.048–0.384, value of $p = 0.012$), 0.307 (95% CI, 0.020–0.594, value of $p = 0.036$) and 0.308 (95% CI, 0.081–0.535, value of $p = 0.008$) in the training cohort and two validation cohorts, respectively (**Table 5**). The IDI values for miRNA-based prediction nomogram were 0.065 (95% CI, 0.015–0.115, value of $p = 0.011$), 0.093 (95% CI,

TABLE 3 | Area under the receiver operating characteristics curve (AUC) of prognostic indicators for lung metastasis in breast cancer (BC).

Variables	**Training cohort**	**Internal validation cohort**	**External validation cohort**
miRNA score	0.681 (95% CI, 0.589–0.774)	0.754 (95% CI, 0.561–0.946)	0.711 (95% CI, 0.608–0.815)
Age at diagnosis	0.403 (95% CI, 0.290–0.516)	0.282 (95% CI, 0.117–0.448)	0.623 (95% CI, 0.479–0.768)
TNM stage	0.727 (95% CI, 0.628–0.825)	0.583 (95% CI, 0.407–0.759)	0.840 (95% CI, 0.716–0.963)
Nomogram model	0.774 (95% CI, 0.669–0.879)	0.763 (95% CI, 0.597–0.929)	0.925 (95% CI, 0.846–1.000)

TNM, the tumor node metastasis; AUC, area under the receiver operating characteristics curve.

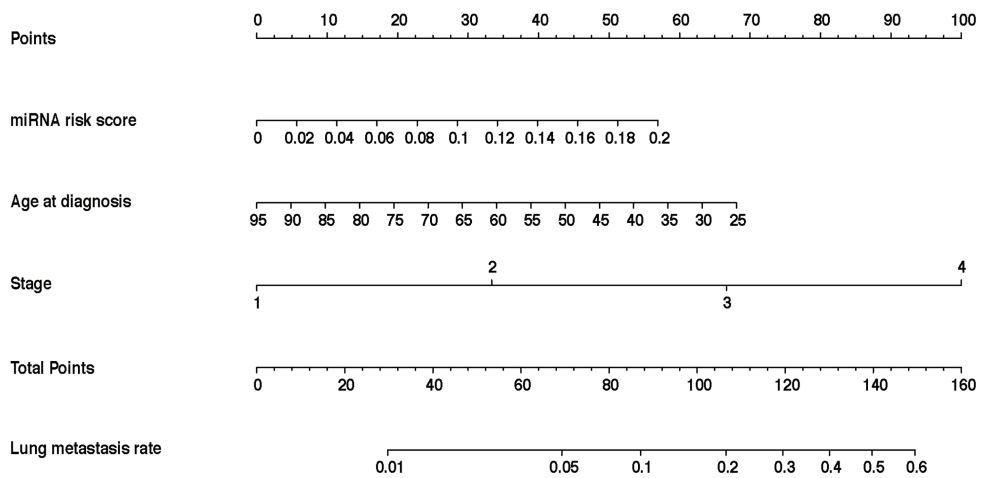


FIGURE 3 | Development and assessment of the miRNA-based nomogram. Constructed a miRNA-based nomogram to predict LM for BC patients in the training cohort, with age at diagnosis, stage, and eight-miRNA signature incorporated. LM, lung metastasis; BC, breast cancer; and miRNA, micro RNA.

0.021–0.165, value of p = 0.011), and 0.025 (95% CI, −0.048–0.098, value of p = 0.500) in the training cohort and two validation cohorts, respectively (**Table 5**). Both NRI and IDI indicated a superior predictive ability of our model compared to the TNM staging system.

Decision-curve analysis was conducted to compare the clinical use of our nomogram to that of the TNM staging system (Stewart et al., 2005; **Figures 4G–I**). The decision curves in both the training and external validation cohorts showed that if the threshold probability was between 0 and 0.60 (in the internal validation cohort, the threshold probability was between 0 and 0.40), using the miRNA-based nomogram to predict lung metastasis added more benefit than treating either all or no patients. DCA also indicated that the net benefit of the miRNA-based nomogram model was comparable, with several overlaps, or even superior to the TNM staging system. Overall, these results suggested the superiority of the miRNA-based nomogram for its lung metastasis predictive performance when compared to the TNM stage.

Identification of Potential Targets for Predictive miRNAs and Their Roles in Lung Metastasis

We identified the gene targets for predictive miRNAs using *in silico* predictions (TargetScan, miRWalk, and miRDB) and experimentally verified microRNA database (miRTarBase). We also acquired matched mRNA transcriptome data of the patients enrolled in the analysis of identifying DEmiRNAs. Around 3,791 genes were differentially expressed, of which 1,710 were upregulated and 2,081 were downregulated (**Figure 6A**; **Supplementary Table S3**). The benefit of using matched mRNA dataset was that it acted as an approach to be the functional validation of targets genes identified by the prediction algorithm (Krishnan et al., 2015). We further used Venn diagram to found the overlap between DEmRNAs and the gene targets for miRNAs and proceeded to the subsequent analysis (**Figure 6B**; **Supplementary Table S4**).

Gene ontology and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analyses were performed for the overlapped target genes of each predictive miRNAs. Among pro-metastatic miRNAs, miR-17 mainly interfered with cell cycle arrest (BP), mitotic G1/S transition checkpoint (BP), positive regulation of autophagy (BP), signal transduction by p53 class mediator (BP), focal adhesion (KEGG), signaling pathways regulating pluripotency of stem cells (KEGG), regulation of actin cytoskeleton (KEGG), and hippo signaling pathway (KEGG; **Supplementary Figures S4A,B**). miR-210 negatively influenced lactate metabolic process (BP), post-embryonic animal organ development (BP), and negative regulation of vascular permeability (BP; **Supplementary Figure S4C**). Another pro-metastatic miR-663 potentiated the invasion of tumor cells by targeting actin filament polymerization (BP), cell-substrate junction assembly (BP), cell-substrate junction assembly (BP), focal adhesion assembly (BP), and actin filament organization (BP; **Supplementary Figure S4D**). The protective miR-30a was found able to restrain PI3K-Akt signaling pathway (KEGG), Ras signaling pathway (KEGG), IL-17 signaling pathway (KEGG), estrogen signaling pathway (KEGG), MAPK signaling pathway (KEGG), Wnt signaling pathway (KEGG), and ERBB signaling pathway (KEGG; **Supplementary Figure S4E**). No terms were enriched in the enrichment analysis of other miRNAs alone.

These miRNAs functioned together in the organism, so then we tried to identify the role of five upregulated or three downregulated miRNAs as a whole. Hub genes of the target genes for five upregulated or three downregulated miRNAs were generated to identify central elements of pro-metastatic and anti-metastatic biological networks (**Supplementary Table S5**). miRNA-mRNA interaction networks of the hub genes of five upregulated or three downregulated miRNAs were plotted (**Figures 6C,D**). The metastatic cascade is composed of a series of sequential events that involve cell detachment from the primary tumor, invasion of these cells into surrounding tissue, intravasation migration, arrest, and extravasation into distant tissues, and formation of metastasis (Lambert et al., 2017). GO analysis was also performed for the hub genes of five upregulated or three downregulated miRNAs (**Figures 6E,F**; **Supplementary Table S6**). We found our predictive miRNAs participated in most of the above events and thereby promoting lung metastasis. They suppressed the adhesion between cancer cells and matrix facilitated the vasculature development and hematogenous metastasis, promoted proliferation, and then adapted to the lung so as to form the metastasis.

miR-30a and miR-135b Have Unique Roles in Lung Metastasis of BC

In order to determine whether these eight predictive miRNAs were unique to lung metastasis in BC patients, we first identified DEmiRNAs between patients with lung metastasis only and patients with distant metastasis except for the lung (**Supplementary Tables S7, S8**). Baseline clinical and pathological characteristics of the study participants in the comparison were listed in **Table 6**. Compared to patients with distant metastasis except for the lung, protective miR-30a was found to be downregulated in patients with lung metastasis only. On the contrary, miR-135b was upregulated in patients with lung metastasis only. In addition, we recognized DEmiRNAs between patients with distant metastasis except for the lung and patients without metastasis (**Table 6**; **Supplementary Tables S7, S8**). The expression levels of miR-135b and miR-17 were downregulated in patients with distant metastasis except for the lung. In order to further confirm whether these three miRNAs were lung-metastasis-specific in BC patients, we performed dot plots to see their expression levels in patients with distant metastasis except for the lung, patients with lung metastasis only, and patients without metastasis (**Figure 7**). The expression level of miR-30a was extremely low in BC patients with lung metastasis, while the expression level of miR-135b was extremely high in BC patients with lung metastasis. These analyses of identifying DEmiRNAs in different subgroups of BC patients showed the unique roles of miR-30a and miR-135b in lung metastasis of BC.

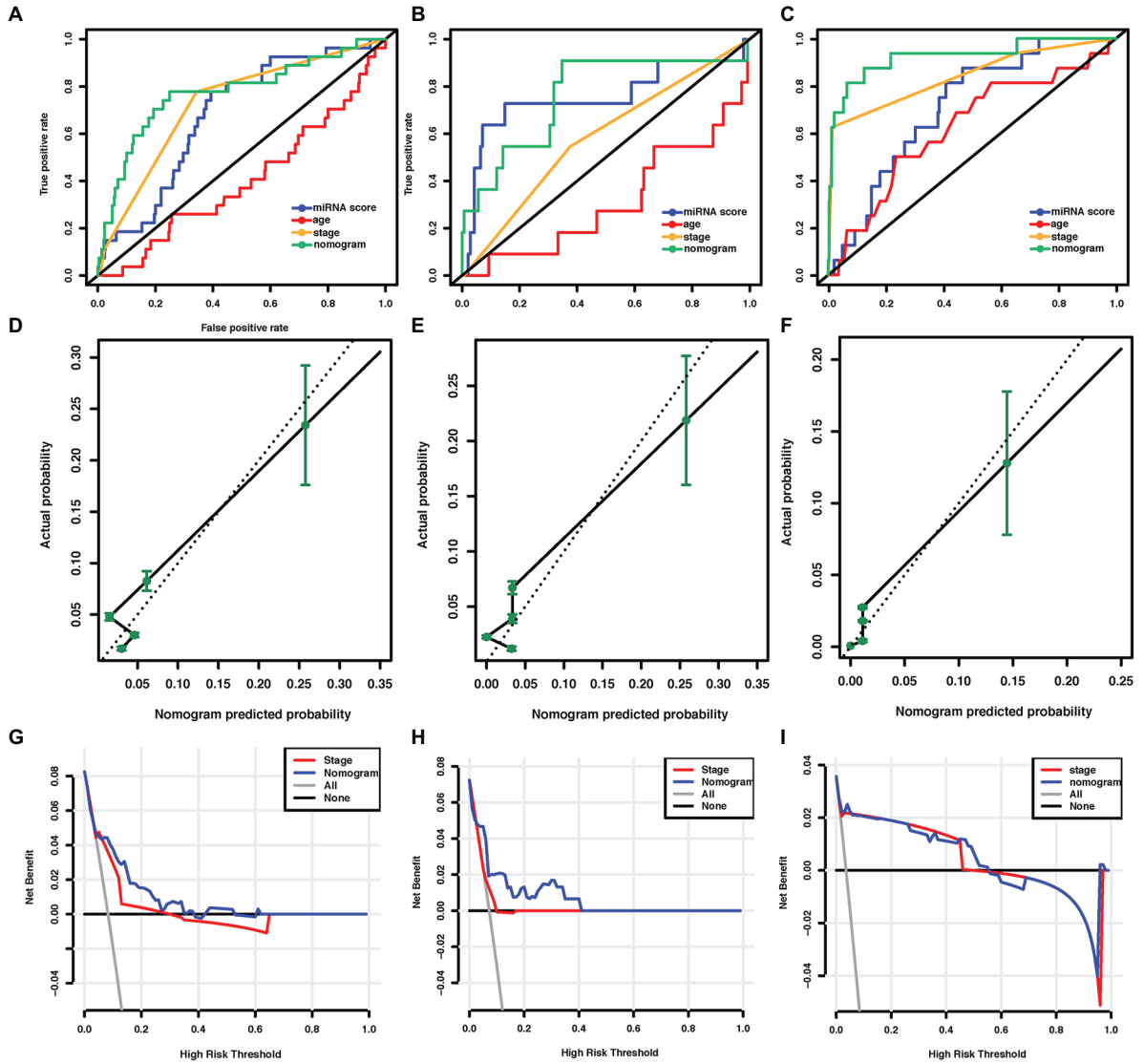


FIGURE 4 | Assessment of the miRNA-based nomogram. Receiver operating characteristic (ROC) curves of eight-miRNA signature, age at diagnosis, stage, and the miRNA-based nomogram model predicting LM in **(A)** training cohort, **(B)** internal validation cohort, and **(C)** external validation cohort. Calibration plots for miRNA-based nomogram model predicting LM in the **(D)** training cohort, **(E)** internal validation cohort, and **(F)** external validation cohort. Calibration curves depict the calibration of the model in terms of the agreement between the predicted risks of LM and the observed outcomes of LM. The y-axis represents the actual LM rate. The x-axis represents the predicted LM risk. The dashed line (the 45-degree diagonal line) represents a perfect prediction by an ideal model, and the black solid line represents the performance of the nomogram of which a closer fit to the diagonal dotted line represents a better prediction. Decision curve analysis of the miRNA-based nomogram model and tumor staging system in **(G)** training cohort, **(H)** internal validation cohort, and **(I)** external validation cohort. The y-axis displays the net benefit. Solid black line: net benefit when all breast cancer patients are considered as not having the LM; solid gray line: net benefit when all breast cancer patients are considered as having LM. Solid red line: net benefit when all breast cancer patients are considered according to the tumor staging system. Solid blue line: net benefit when all breast cancer patients are considered according to the miRNA-based nomogram model. The net benefit was calculated by subtracting the proportion of all patients who are false positive from the proportion who are truly positive, weighting by the relative harm of giving up treatment compared with the negative consequences of unnecessary treatment (Vickers et al., 2008). miRNA, microRNA; ROC, receiver operating characteristic; LM, lung metastasis; and BC, breast cancer.

Pan-Cancer Analysis of the Expression Levels of Eight Predictive miRNAs in Patients With Lung Metastasis and Without Metastasis

We performed differential miRNA expression analyses between patients with lung metastasis and patients without metastasis in six cancer types (ACC, BLCA, SARC, SKCM, CESC, and STAD; **Supplementary Tables S9, S10**). The expression level of miR-663 was not detected in these datasets. The distributions of the expression levels of these predictive miRNAs in six cancer types were also presented (**Supplementary Figures S5A–F**). Combined analyses indicated that compared to patients without

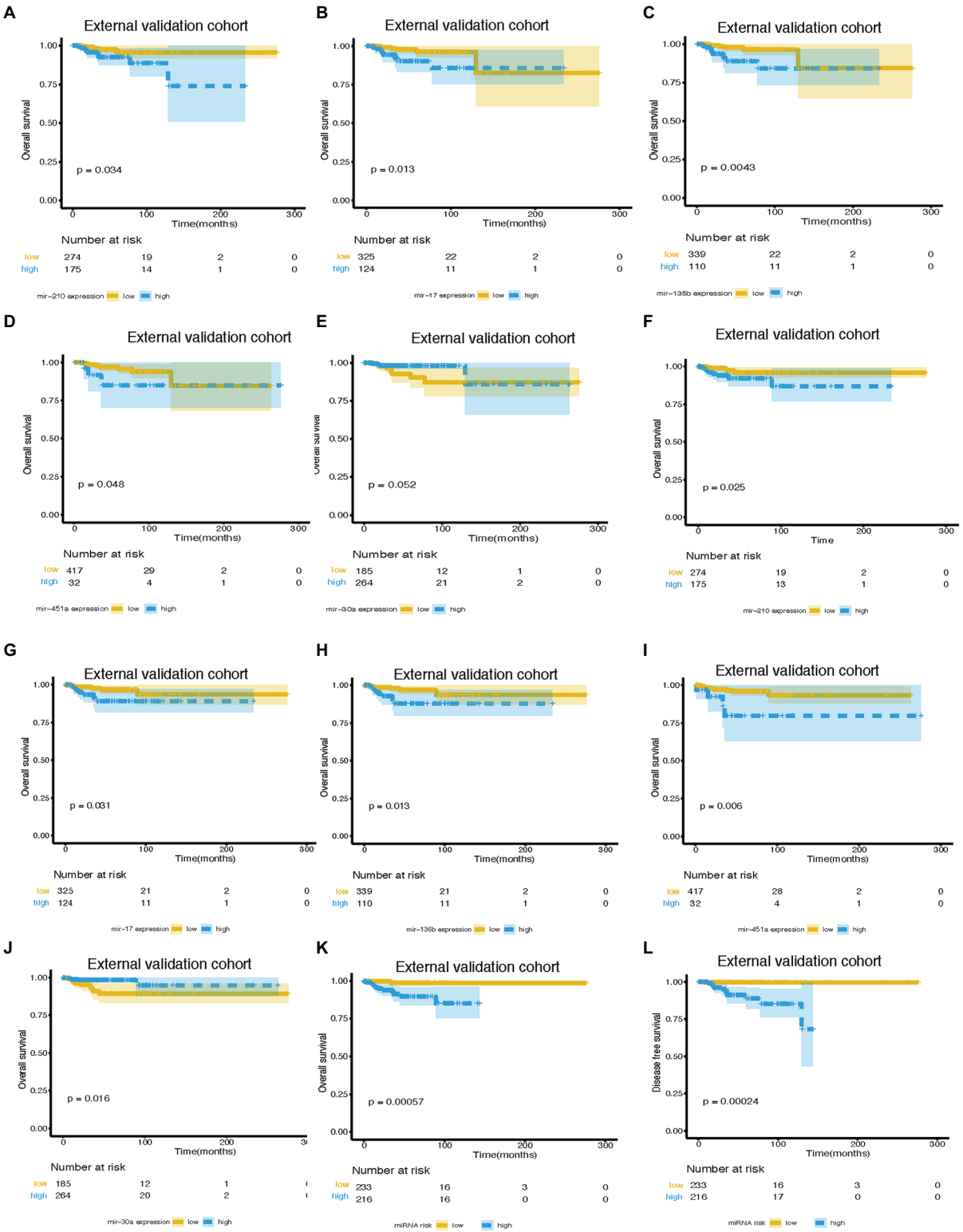


FIGURE 5 | Survival curves of BC patients stratified by different variables. KM curves of overall survival of breast cancer patients stratified by **(A)** mir-210 expression, **(B)** mir-17 expression, **(C)** mir-135b expression, **(D)** mir-451a expression, **(E)** mir-30a expression, and **(K)** miRNA risk score in the external validation cohort. Kaplan-Meier curves of disease-free survival of breast cancer patients stratified by **(F)** mir-210 expression, **(G)** mir-17 expression, **(H)** mir-135b expression, **(I)** mir-451a expression, **(J)** mir-30a expression, and **(L)** miRNA risk score in the external validation cohort. miRNA, microRNA; BC, breast cancer.

metastasis, miR-210 was upregulated in ACC and SARC patients with lung metastasis. The expression level of miR-199a-5p was higher in BLCA patients with lung metastasis, whereas the expression level of miR-199a-5p was lower in SARC patients with lung metastasis. miR-17 was upregulated in SARC patients with lung metastasis. Elevated expression levels of miR-135b were detected in ACC patients with lung metastasis. Compared to patients without metastasis, the expression level of miR-30a was suppressed in ACC patients with lung metastasis.

DISCUSSION

Based on Surveillance, Epidemiology, and End Results (SEER) database, the median survival time for BC patients with lung metastases was 21 months, and only 15.5% of the patients were alive for more than 3 years (Xiao et al., 2018). Once metastasis occurs, the disease is largely incurable. Identifying effective predictive biomarkers to construct an accurate nomogram model to predict the lung metastasis status of BC patients is an advisable choice applied in the clinical practice. At present, the TNM staging system is commonly used to assess the metastasis probability of BC patients. But as discussed above, a single clinical parameter has limited power of outcome prediction. We put forward the idea for the first time that BC patients with lung metastasis might have unique clinicopathologic characteristics and miRNA expression profiles, which could distinguish themselves from those who had no lung metastasis.

Subgroup analysis suggested that miR-30a and miR-135b have distinct roles in lung metastasis of BC patients. miR-30 has been reported to be able to stabilize pulmonary vessels and inhibit pulmonary vascular hyperpermeability in the premetastatic phase (Qi et al., 2015). The role of miR-135b in BC patients remains controversial. miR-135b reduces the proliferation of ERα-positive BC cells (Aakula et al., 2015), but promotes the proliferation and invasion of triple-negative breast cancer (TNBC) by downregulating *APC* expression (Lv et al., 2019). TNBC especially tends to metastasize to the lungs (Foulkes et al., 2010), which may partly explain the uniqueness of miR-135b to the lung metastasis. The precise roles of these miRNAs in the lung have been studied to some extent, yet further research is needed to fill the gap.

The significance of miRNAs is better appreciated from the aspect of their potential functional impact on biological pathways, as these influence the outcomes for the patient (Krishnan et al., 2015). Cancer metastasis is a complicated process, and the outcome of metastasis depends on the interactions between cancer cells and a given microenvironment. We could see that the targets for the identified miRNAs were enriched for cell proliferation, invasion, and migration, which participated in the whole regulatory process of metastasis. During lung metastasis, metastatic tumor cells will rewrite their biology and expression profiles to adapt to the distant microenvironment, which endows tumor cells with full competence for outgrowth in the lung. Therefore, we also identified some adaptations specific to the lung microenvironment. The target of miR-30a, *SEMA3A*, has been reported to modulate distal pulmonary epithelial cell development and alveolar septation, which has also been found upregulated in patients with lung metastasis (Becker et al., 2011). Transforming growth factor beta (TGFβ) promotes metastasis of BC to the lungs but it is dispensable to bone metastasis (Chen et al., 2018). We identified "positive regulation of TGFβ production" enriched in patients with lung metastasis. Terms concerning lung such as "lung development" and "epithelial tube branching involved in lung morphogenesis" have also been identified in GO analysis.

We also conducted a pan-cancer analysis to figure out whether the eight predictive miRNAs were specific to BC. Some of the miRNAs had consistent effects in different cancer types, such as miR-30a, miR-17, miR-451a, and miR-135b, while others showed controversial effects, such as miR-210, miR-301a, and miR-199a. Previous studies also identified the role of these predictive miRNAs in lung metastasis of other types of cancer (Qi et al., 2015; Kai et al., 2016; Jin et al., 2017; Xu et al., 2019; Wang et al., 2020). miR-17, miR-135b, and miR-210 facilitate cancer cells to metastasize to the lungs,

TABLE 4 | Risk factors for lung metastasis in external validation cohort.

	Univariate analysis		Multivariate analysis	
	OR (95% CI)	***p* value**	**OR (95% CI)**	***p* value**
miRNA score	2.748 (1.299–5.816)	0.0082	4.207 (1.440–12.290)	0.0086
Age at diagnosis	1.678 (0.861–3.268)	0.1277	1.748 (0.811–3.769)	0.1540
TNM stage	29.345 (9.153–94.086)	<0.0001	32.540 (8.986–117.830)	<0.0001

TNM, the tumor node metastasis.

TABLE 5 | The improvement of miRNA-based nomogram in predicting lung metastasis according to net reclassification improvement (NRI) and integrated discrimination improvement (IDI).

Training cohort				Internal validation cohort				External validation cohort			
NRI (95% CI)	***p***	**IDI (95% CI)**	***p***	**NRI (95% CI)**	***p***	**IDI (95% CI)**	***p***	**NRI (95% CI)**	***p***	**IDI (95% CI)**	***p***
0.216 (0.048–0.384)	0.012	0.065 (0.015–0.115)	0.011	0.307 (0.020–0.594)	0.036	0.093 (0.021–0.165)	0.011	0.308 (0.081–0.535)	0.008	0.025 (–0.048–0.098)	0.5

NRI, net reclassification improvement; IDI, the integrated discrimination improvement; and P, p value.

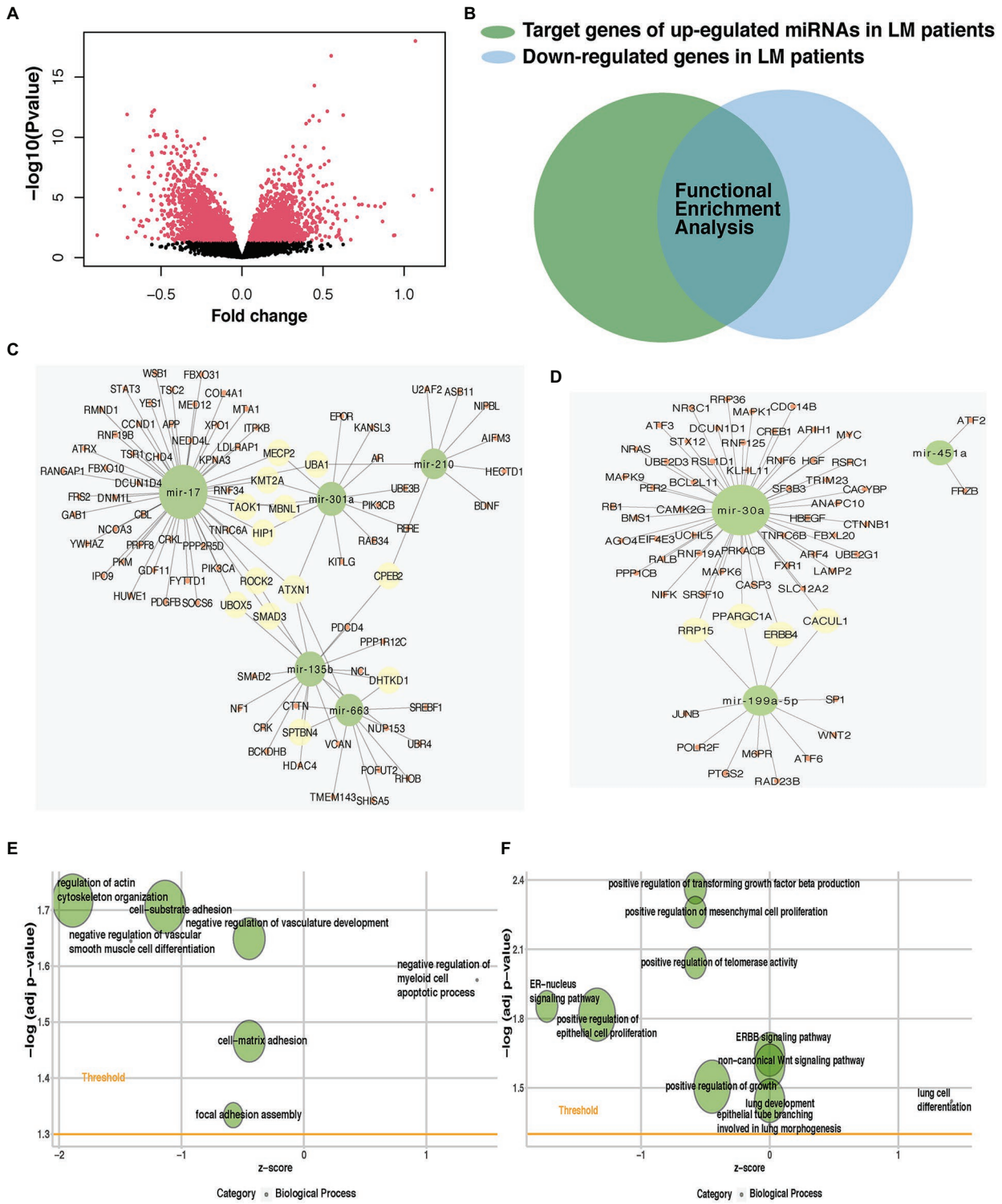


FIGURE 6 | Identification of potential targets for predictive miRNAs and their role in lung metastasis. **(A)** Volcano plot of mRNAs expression in the METABRIC dataset. **(B)** Venn diagram was plotted to show the overlap between differentially expressed mRNAs (DEmRNAs) and gene targets for predictive miRNAs. The overlap of each predictive miRNA was used in subsequent analysis. miRNA-mRNA interaction networks of the hub genes of **(C)** five upregulated or **(D)** three downregulated predictive miRNAs. Enriched metastasis-related gene ontology (GO) terms of the hub genes of **(E)** five upregulated or **(F)** three downregulated predictive miRNAs. miRNA, microRNA; METABRIC, molecular taxonomy of breast cancer international consortium; DEmRNAs, differentially expressed mRNAs; and GO, gene ontology.

TABLE 6 | Demographics of the samples recruited in subgroup analysis.

Variables	lung metastasis only (*n* = 6)	distant metastasis except for the lung (*n* = 54)	without metastasis (*n* = 433)
Median age at diagnosis in years (IQR)	65 (56–71.5)	57 (47–63.25)	60 (50.5–67.00)
Median follow up time from diagnosis in days (IQR)	1,233 (645.3–3,578)	1,096 (190.5–2,405)	343.5 (109.3–1,064)
Pam50 subtype			
Luminal A	0	22	201
Luminal B	1	11	64
HER2	0	5	21
Basal like	2	6	94
Normal breast-like	1	2	13
Unknown	2	8	40
TNM stage			
1	0	7	151
2	2	27	276
3	4	14	5
4	0	6	1
ER status			
Positive	1	37	115
Negative	5	12	297
Unknown	0	5	21
PR status			
Positive	1	31	266
Negative	5	19	144
Unknown	0	4	23
HER2 status			
Positive	1	2	52
Negative	1	17	241
Unknown	4	35	140
Menopausal state			
Pre	1	12	82
Post	5	32	299
Peri	0	2	19
Unknown	0	8	33
Patient metastatic sites			
Lung	*6*	*0*	*0*
Bone	*0*	*29*	*0*
Brain	*0*	*3*	*0*
Liver	*0*	*7*	*0*
Multi-organ Metastasis	*0*	*15*	*0*
No metastasis	*0*	*0*	*433*
Vital status			
Alive	1	16	433
Dead	5	38	0

PAM50, prediction analysis of microarray 50; ER, estrogen receptor; PR, progesterone receptor; HER2, human epithelial growth factor receptor 2; and TNM, the tumor node metastasis.

while miR-30a and miR-451a suppress lung metastasis, which exerts similar effects to our results. The lack of research and missing data of miR-663 suggests it can serve as an appealing target for future research. In addition, the notion that miRNAs exert both oncogenic and tumor-suppressive effects has been put forward (Rohan et al., 2019). An individual miRNA could regulate the expression of hundreds of genes. The effect of miRNA in each situation depends on the balance of the pro-tumor and anti-tumor pathways. Multiple biological factors can interfere with the balance, such as the interplay between cells and microenvironment, energy supply, and so on. Although two miRNAs have conflicting roles in pan-cancer analysis, the overall consistency indicated the significant importance of these eight miRNAs in lung metastasis.

The univariate and multivariate logistic regression analysis showed that the eight-miRNA signature could be an independent risk factor in training and validation cohorts. The AUC of eight-miRNA signature alone for lung metastasis prediction showed a little bit smaller than that of the TNM staging system in training and external validation cohort. Therefore, the comprehensive predictive nomogram was constructed integrating the risk score and conventional clinical parameters including stage and age, all of which were verified as an independent risk factor using univariate and multivariate logistic regression analysis for the lung metastasis status of BC patients. Apart from AUC, the calibration plot was also used to assess the discrimination performance of the nomogram model. Although the overall trend was in line with the 45-degree ideal diagonal line, yet the calibration plot showed some deviation, which may due to the limited events and thus affecting the power. NRI, IDI, and DCA were used to evaluate the prediction ability between miRNA-based nomogram and the TNM staging system. The results of NRI indicated the significant improvement of miRNA-based nomogram in all three cohorts, and the results of IDI suggested that the nomogram model improved the predictive power, yet failed to reach a significant level in the external validation cohort. DCA results also indicated that our miRNA-based nomogram improved current treatment standards, while the ideal model was the model with the positive net benefit at any given threshold.

However, several limitations of our study should be acknowledged. Firstly, due to the different sequence platforms, only seven of eight predictive miRNAs were identified in the external validation cohort, so we did not adopt the risk scores and cut-off points generated in the training set as previous research suggested (Volinia and Croce, 2013; Krishnan et al., 2015; Rohan et al., 2019). Secondly, the limited number of events in the cohorts may affect the statistical power. Among DEmiRNAs that were not selected by LASSO method, some have also been reported to be related to lung metastasis (Ma et al., 2010). HER2 overexpression has been proved to be a risk for the development of visceral-only metastasis including lung (Bartmann et al., 2017). However, HER2 status reached a significant level in univariate logistic regression but failed in multivariate logistic regression, so it was not included in the nomogram model. Last but not least, we have emphasized the complexity of miRNA regulation previously. Therefore, experiments for revealing and verification of their roles in lung metastasis are crucial in the future.

In this study, we constructed a nomogram model based on multiple lung metastasis-related miRNAs and clinical risk factors to predict the lung metastasis of BC patients. We screened the high-throughput sequence data from the METABRIC database to explore DEmiRNAs and used the LASSO method to identify an eight-miRNA signature. The risk score was calculated by

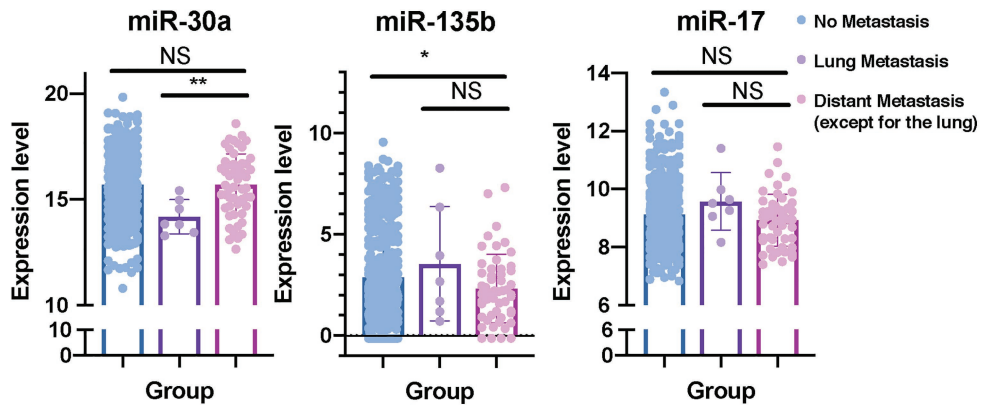


FIGURE 7 | miR-30a and miR-135b have unique roles in lung metastasis of BC. Dot plots were plotted to show the distributions of miR-30a, miR-135b, and miR-17 in BC patients with distant metastasis except for the lung, BC patients with lung metastasis only, and BC patients without metastasis. miRNA, microRNA; BC, breast cancer. *$p < 0.05$; **$p < 0.01$.

the multivariate logistic coefficient multiplied by the expression of the miRNA. Then the risk score and clinical risk factors were combined together to construct a miRNA-based nomogram, which was assessed by the calibration plot, ROC analysis, NRI, IDI, and DCA. Internal and external validation was also performed to evaluate the nomogram model. Functional enrichment analyses were performed to identify the potential biological roles of eight predictive miRNAs. Subgroup analysis of BC patients with different distant metastasis showed that miR-30a, miR-135b, and miR-17 have unique roles in lung metastasis of BC. Pan-cancer analysis of patients with lung metastasis or without metastasis in six types of cancer indicated the significant importance of eight predictive miRNAs in lung metastasis. A biomarker-based approach to accurately predict the metastasis status of BC patients is urgently needed in the era of precision medicine. Risk assessment is vital for making appropriate therapeutic decisions and follow-up strategies in BC patients. If a patient has a high probability to have lung metastasis in the future, we might recommend the patient to take a close inspection of the lung and adopt advanced treatment. This model might be able to perform well in all patients, for it was constructed based on large-scale datasets. In addition, this risk score was also a significant factor in affecting survival. Therefore, this nomogram could be used as a supportive graphic tool in clinical practice to facilitate treatment decisions of BC patients.

CONCLUSION

In our current study, we identified eight predictive miRNAs from publicly available data and constructed an eight-miRNA based nomogram that incorporated other clinical parameters including stage and age to predict the lung metastasis status of BC patients, whose prediction power was better than that of conventional TNM stage system. Subgroup analysis suggested that miR-30a, miR-135b, and miR-17 may have unique roles in lung metastasis of BC patients. On the basis of the GO, KEGG enrichment, and pan-cancer analyses, the eight miRNAs played crucial roles in lung metastasis cascade. Therefore, our eight-miRNA-based nomogram might be a vital tool for lung metastasis prediction in BC patients, aiding in developing personalized treatment strategies.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in The Cancer Genome Atlas (https://portal.gdc.cancer.gov/) and European Genome Archive (https://ega-archive.org/).

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

LZ initiated and organized the study. LZ and JP designed and carried out bioinformatics and statistical analyses, drew figures, and drafted the manuscript. ZW, CY, and JH participated in editing 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 (no. 81872317, 81520108024).

ACKNOWLEDGMENTS

The datasets of this study are obtained by the Cancer Genome Atlas database and European Genome-phenome Archive. We are grateful to them for the establishment and management of the databases.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.580138/full#supplementary-material

REFERENCES

Aakula, A., Leivonen, S., Hintsanen, P., Aittokallio, T., Ceder, Y., Børresen-Dale, A., et al. (2015). MicroRNA-135b regulates ERα, AR and HIF1AN and affects breast and prostate cancer cell growth. *Mol. Oncol.* 9, 1287–1300. doi: 10.1016/j.molonc.2015.03.001

Bartmann, C., Diessner, J., Blettner, M., Häusler, S., Janni, W., Kreienberg, R., et al. (2017). Factors influencing the development of visceral metastasis of breast cancer: a retrospective multi-center study. *Breast* 31, 66–75. doi: 10.1016/j.breast.2016.10.016

Becker, P., Tran, T., Delannoy, M., He, C., Shannon, J., and McGrath-Morrow, S. (2011). Semaphorin 3A contributes to distal pulmonary epithelial cell differentiation and lung morphogenesis. *PLoS One* 6:e27449. doi: 10.1371/journal.pone.0027449

Chen, W., Hoffmann, A., Liu, H., and Liu, X. (2018). Organotropism: new insights into molecular mechanisms of breast cancer metastasis. *NPJ Precis. Oncol.* 2:4. doi: 10.1038/s41698-018-0047-0

Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* 48, D127–D131. doi: 10.1093/nar/gkz757

Chin, C., Chen, S., Wu, H., Ho, C., Ko, M., and Lin, C. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 4:S11. doi: 10.1186/1752-0509-8-s4-s11

Chipman, J., and Braun, D. (2017). Simpson's paradox in the integrated discrimination improvement. *Stat. Med.* 36, 4468–4481. doi: 10.1002/sim.6862

Chou, C., Shrestha, S., Yang, C., Chang, N., Lin, Y., Liao, K., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46, D296–D302. doi: 10.1093/nar/gkx1067

Curtis, C., Shah, S., Chin, S., Turashvili, G., Rueda, O., Dunning, M., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Delpech, Y., Bashour, S., Lousquy, R., Rouzier, R., Hess, K., Coutant, C., et al. (2015). Clinical nomogram to predict bone-only metastasis in patients with early breast carcinoma. *Br. J. Cancer* 113, 1003–1009. doi: 10.1038/bjc.2015.308

DeSantis, C., Ma, J., Gaudet, M., Newman, L., Miller, K., Goding Sauer, A., et al. (2019). Breast cancer statistics, 2019. *CA Cancer J. Clin.* 69, 438–451. doi: 10.3322/caac.21583

Dvinge, H., Git, A., Gräf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., et al. (2013). The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature* 497, 378–382. doi: 10.1038/nature12108

Foulkes, W. D., Smith, I. E., and Reis-Filho, J. S. (2010). Triple-negative breast cancer. *N. Engl. J. Med.* 363, 1938–1948. doi: 10.1056/NEJMra1001389

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Harrell, F. J. (2013). rms: regression modeling strategies. R package version 4.0-0. Computer software. Available at: http://CRAN.R-project.org/package=rms (Accessed June 23, 2020).

Iasonos, A., Schrag, D., Raj, G., and Panageas, K. (2008). How to build and interpret a nomogram for cancer prognosis. *J. Clin. Oncol.* 26, 1364–1370. doi: 10.1200/jco.2007.12.9791

Jin, H., Luo, S., Wang, Y., Liu, C., Piao, Z., Xu, M., et al. (2017). miR-135b stimulates osteosarcoma recurrence and lung metastasis via notch and Wnt/β-catenin signaling. *Mol. Ther. Nucleic Acids* 8, 111–122. doi: 10.1016/j.omtn.2017.06.008

Kai, A., Chan, L., Lo, R., Lee, J., Wong, C., Wong, J., et al. (2016). Down-regulation of TIMP2 by HIF-1α/miR-210/HIF-3α regulatory feedback circuit enhances cancer metastasis in hepatocellular carcinoma. *Hepatology* 64, 473–487. doi: 10.1002/hep.28577

Kassambara, A. K., Kosinski, M., Biecek, P., and Fabian, S. (2017). Survminer: drawing survival curves using "ggplot2" R package version 0.4.4. Available at: https://CRAN.R-project.org/package=survminer (Accessed June 24, 2020).

Knott, S. R. V., Wagenblast, E., Khan, S., Kim, S. Y., Soto, M., Wagner, M., et al. (2018). Asparagine bioavailability governs metastasis in a model of breast cancer. *Nature* 554, 378–381. doi: 10.1038/nature25465

Kramer, A., and Zimmerman, J. (2007). Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. *Crit. Care Med.* 35, 2052–2056. doi: 10.1097/01.Ccm.0000275267.64078.B0

Krishnan, P., Ghosh, S., Wang, B., Li, D., Narasimhan, A., Berendt, R., et al. (2015). Next generation sequencing profiling identifies miR-574-3p and miR-660-5p as potential novel prognostic markers for breast cancer. *BMC Genomics* 16:735. doi: 10.1186/s12864-015-1899-0

Kundu, S., Aulchenko, Y. S., van Duijn, C. M., and Janssens, A. C. (2011). PredictABEL: an R package for the assessment of risk prediction models. *Eur. J. Epidemiol.* 26, 261–264. doi: 10.1007/s10654-011-9567-4.

Lambert, A., Pattabiraman, D., and Weinberg, R. (2017). Emerging biological principles of metastasis. *Cell* 168, 670–691. doi: 10.1016/j.cell.2016.11.037

Lin, S., and Gregory, R. (2015). MicroRNA biogenesis pathways in cancer. *Nat. Rev. Cancer* 15, 321–333. doi: 10.1038/nrc3932

Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Lv, Z., Xin, H., Yang, Z., Wang, W., Dong, J., Jin, L., et al. (2019). miR-135b promotes proliferation and metastasis by targeting APC in triple-negative breast cancer. *J. Cell. Physiol.* 234, 10819–10826. doi: 10.1002/jcp.27906

Ma, L., Reinhardt, F., Pan, E., Soutschek, J., Bhat, B., Marcusson, E. G., et al. (2010). Therapeutic silencing of miR-10b inhibits metastasis in a mouse mammary tumor model. *Nat. Biotechnol.* 28, 341–347. doi: 10.1038/nbt.1618

Medeiros, B., and Allan, A. (2019). Molecular mechanisms of breast cancer metastasis to the lung: clinical and experimental perspectives. *Int. J. Mol. Sci.* 20:2272. doi: 10.3390/ijms20092272

Nam, J., Rissland, O., Koppstein, D., Abreu-Goodger, C., Jan, C., Agarwal, V., et al. (2014). Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol. Cell* 53, 1031–1043. doi: 10.1016/j.molcel.2014.02.013

Network, C. G. A. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Pencheva, N., and Tavazoie, S. (2013). Control of metastatic progression by microRNA regulatory networks. *Nat. Cell Biol.* 15, 546–554. doi: 10.1038/ncb2769

Pencina, M., D'Agostino, R. Sr., D'Agostino, R. Jr., and Vasan, R. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* 27, 157–172. doi: 10.1002/sim.2929

Pencina, M., D'Agostino, R., and Steyerberg, E. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat. Med.* 30, 11–21. doi: 10.1002/sim.4085

Qi, F., He, T., Jia, L., Song, N., Guo, L., Ma, X., et al. (2015). The miR-30 family inhibits pulmonary vascular hyperpermeability in the premetastatic phase by direct targeting of Skp2. *Clin. Cancer Res.* 21, 3071–3080. doi: 10.1158/1078-0432.Ccr-14-2785

Ritchie, M., Phipson, B., Wu, D., Hu, Y., Law, C., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Rohan, T., Wang, T., Weinmann, S., Wang, Y., Lin, J., Ginsberg, M., et al. (2019). A miRNA expression signature in breast tumor tissue is associated with risk of distant metastasis. *Cancer Res.* 79, 1705–1713. doi: 10.1158/0008-5472.Can-18-2779

Schrijver, W., van Diest, P., and Moelans, C. (2017). Unravelling site-specific breast cancer metastasis: a microRNA expression profiling study. *Oncotarget* 8, 3111–3123. doi: 10.18632/oncotarget.13623

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* 21, 3940–3941. doi: 10.1093/bioinformatics/bti623

Stewart, G., Bariol, S., Grigor, K., Tolley, D., and McNeill, S. (2005). A comparison of the pathology of transitional cell carcinoma of the bladder and upper urinary tract. *BJU Int.* 95, 791–793. doi: 10.1111/j.1464-410X.2005.05402.x

Sticht, C., De La Torre, C., Parveen, A., and Gretz, N. (2018). miRWalk: an online resource for prediction of microRNA binding sites. *PLoS One* 13:e0206239. doi: 10.1371/journal.pone.0206239

Vickers, A., Cronin, A., Elkin, E., and Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Inform. Decis. Mak.* 8:53. doi: 10.1186/1472-6947-8-53

Vickers, A., and Elkin, E. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* 26, 565–574. doi: 10.1177/0272989x06295361

Volinia, S., and Croce, C. M. (2013). Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* 110, 7413–7417. doi: 10.1073/pnas.1304977110

Walter, W., Sánchez-Cabo, F., and Ricote, M. (2015). GOplot: an R package for visually combining expression data with functional analysis. *Bioinformatics* 31, 2912–2914. doi: 10.1093/bioinformatics/btv300

Wang, Q., Shang, J., Zhang, Y., Zhou, Y., and Tang, L. (2020). MiR-451a restrains the growth and metastatic phenotypes of papillary thyroid carcinoma cells via inhibiting ZEB1. *Biomed. Pharmacother.* 127:109901. doi: 10.1016/j.biopha.2020.109901

Wang, J., Song, C., Tang, H., Zhang, C., Tang, J., Li, X., et al. (2017). miR-629-3p may serve as a novel biomarker and potential therapeutic target for lung metastases of triple-negative breast cancer. *Breast Cancer Res.* 19:72. doi: 10.1186/s13058-017-0865-y

Wang, Y., Yang, Y., Chen, Z., Zhu, T., Wu, J., Su, F., et al. (2019). Development and validation of a novel nomogram for predicting distant metastasis-free survival among breast cancer patients. *Ann. Transl. Med.* 7:537. doi: 10.21037/atm.2019.10.10

Wei, T., and Simko, V. (2017). *R package "corrplot": visualization of a correlation matrix (Version 0.84)*. Available at: https://github.com/taiyun/corrplot

Xiao, W., Zheng, S., Liu, P., Zou, Y., Xie, X., Yu, P., et al. (2018). Risk factors and survival outcomes in patients with breast cancer and lung metastasis: a population-based study. *Cancer Med.* 7, 922–930. doi: 10.1002/cam4.1370

Xu, J., Meng, Q., Li, X., Yang, H., Xu, J., Gao, N., et al. (2019). Long noncoding RNA MIR17HG promotes colorectal cancer progression via miR-17-5p. *Cancer Res.* 79, 4882–4895. doi: 10.1158/0008-5472.Can-18-3880

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* 16, 284–287. doi: 10.1089/omi.2011.0118

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.

GLOSSARY

BC	Breast cancer
miRNAs	MicroRNAs
METABRIC	Molecular taxonomy of breast cancer international consortium
DEmiRNAs	Differentially expressed miRNAs
TCGA	The cancer genome atlas
AUC	Area under the receiver operating characteristics curve
TNM	The tumor node metastasis
NRI	Net reclassification improvement
IDI	The integrated discrimination improvement
DCA	Decision-curve analysis
LASSO	Least absolute shrinkage and selection operator
OS	Overall survival
3' UTR	3' untranslated region
KM	Kaplan-Meier
ROC	Receiver operating characteristic curve
ER	Estrogen receptor
PR	Progesterone receptor
HER2	Human epidermal growth factor receptor 2
DEmRNAs	Differentially expressed mRNAs
GO	Gene ontology
KEGG	Kyoto encyclopedia of genes and genomes
DFS	Disease-free survival
SEER	Surveillance, epidemiology, and end results
EMT	Epithelial-mesenchymal transition
TGFβ	Transforming growth factor beta
ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
STAD	Stomach adenocarcinoma
LM	Lung metastasis

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 March 2021
doi: 10.3389/fgene.2021.636867

A New Model for Caries Risk Prediction in Teenagers Using a Machine Learning Algorithm Based on Environmental and Genetic Factors

*Liangyue Pang[1], Ketian Wang[1], Ye Tao[1], Qinghui Zhi[1], Jianming Zhang[2] and Huancai Lin[1]**

[1] Guangdong Provincial Key Laboratory of Stomatology, Department of Preventive Dentistry, Guanghua School of Stomatology, Hospital of Stomatology, Sun Yat-sen University, Guangzhou, China, [2] Foshan Stomatology Hospital, School of Stomatology and Medicine, Foshan University, Foshan, China

OPEN ACCESS

Edited by:
Lu Zhang,
Hong Kong Baptist University, Hong Kong

Reviewed by:
Erika Kuchler, Universidade Positivo, Brazil
Alexandre Rezende Vieira, University of Pittsburgh, United States
Shuguo Zheng, Peking University School and Hospital of Stomatology, China
Xingyu Zhang, University of Michigan, United States

***Correspondence:**
Huancai Lin
linhc@mail.sysu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *02 December 2020*
Accepted: *19 February 2021*
Published: *11 March 2021*

Citation:
Pang L, Wang K, Tao Y, Zhi Q, Zhang J and Lin H (2021) A New Model for Caries Risk Prediction in Teenagers Using a Machine Learning Algorithm Based on Environmental and Genetic Factors. Front. Genet. 12:636867. doi: 10.3389/fgene.2021.636867

Dental caries is a multifactorial disease that can be caused by interactions between genetic and environmental risk factors. Despite the availability of caries risk assessment tools, caries risk prediction models incorporating new factors, such as human genetic markers, have not yet been reported. The aim of this study was to construct a new model for caries risk prediction in teenagers, based on environmental and genetic factors, using a machine learning algorithm. We performed a prospective longitudinal study of 1,055 teenagers (710 teenagers for cohort 1 and 345 teenagers for cohort 2) aged 13 years, of whom 953 (633 teenagers for cohort 1 and 320 teenagers for cohort 2) were followed for 21 months. All participants completed an oral health questionnaire, an oral examination, biological (salivary and cariostate) tests, and single nucleotide polymorphism sequencing analysis. We constructed a caries risk prediction model based on these data using a random forest with an AUC of 0.78 in cohort 1 (training cohort). We further verified the discrimination and calibration abilities of this caries risk prediction model using cohort 2. The AUC of the caries risk prediction model in cohort 2 (testing cohort) was 0.73, indicating high discrimination ability. Risk stratification revealed that our caries risk prediction model could accurately identify individuals at high and very high caries risk but underestimated risks for individuals at low and very low caries risk. Thus, our caries risk prediction model has the potential for use as a powerful community-level tool to identify individuals at high caries risk.

Keywords: caries, risk prediction model, preventive dentistry, biomarkers, biomedical informatics

INTRODUCTION

Permanent teeth caries was the most common chronic disease worldwide in 2016. A previous study reported that the global cost of dental diseases exceeded 540 billion dollars in 2015 and resulted in major health and financial burdens (Righolt et al., 2018). Therefore, there is an urgent need for effective caries control.

Accumulating evidence has shown a skewed distribution of caries; the majority of the disease was suffered by the minority teenagers in the population (Kaste et al., 1996). The conference of National Institutes of Health Consensus Development Conference Statement (2001) concluded that a focus on high-risk individuals was required for the prevention and control of dental caries (2001). Since caries is largely preventable, risk prediction models for early and accurate identification of teenagers at high risk of caries would be useful tools for designing more cost-effective caries control measures.

As a prerequisite for implementing minimally invasive treatment programs, caries risk prediction models (CRPMs) have huge potential in improving patient care because they allow individuals to choose appropriate non-invasive or invasive interventions (Domejean et al., 2017). There are four commonly used standardized caries risk assessment models at present: ADA (American Dental Association), CAT (Caries-Risk Assessment Tool), CAMBRA (Caries Management by Risk Assessment), and Cariogram. All these models included only environmental factors such as socio-demographic indicators, behavioral factors, plaque index, the number of *Streptococcus mutans*, and *Lactobacillus*, saliva flow, and salivary buffer capacity (Petersson and Twetman, 2015). Cariogram, one of the better CRPMs, has provided reliable results for few tests in children, but there is not enough evidence to prove its effectiveness in caries assessment and prediction. Cagetti et al. (2018) reported that the sensitivity of Cariogram in different samples ranged from 41.0 to 75.0%, while the specificity ranged from 65.8 to 88.0%.

Dental caries is a multifactorial disease caused by complex interactions between genetic and environmental risk factors. Environmental risk factors for caries included sugar-rich diet, poor oral hygiene, dental plaque, high numbers of cariogenic bacteria, inadequate salivary flow and so on (Selwitz et al., 2007). Genetic contribution to caries risk score variation has been reported to be 49.1–62.7% (Haworth et al., 2020). As a genetically complex phenotype, caries risk may be influenced by many loci with small contributions individually. These genetic factors that contribute to caries may include variants in loci for enamel formation, immune response, saliva, taste, and dietary habits (Vieira et al., 2014). Enamel formation was tested as being potentially involved in caries susceptibility. Patir et al. (2008) reported an association between enamelin (*ENAM*) and higher caries experience. Additionally, a relationship between the genetic variation of tuftelin (*TUFT1*) and caries could be detected only when the *Streptococcus mutans* levels were high (Slayton et al., 2005).

Therefore, CRPMs based on environmental factors alone may lead to the loss of useful information. Previous studies have suggested that constructing a disease risk prediction model with both environmental and genetic factors can stratify the disease risk more accurately than either of these factors alone (Li et al., 2019; Okubo et al., 2020). Accordingly, research is needed to construct CRPMs based on both genetic and environmental risk factors and evaluate their abilities to predict caries risk better. Thus, this prospective study aimed to construct a new CRPM including both genetic and environmental risk factors in teenagers of the Chinese population.

MATERIALS AND METHODS

Study Population

This study was approved by the Ethics Committee of the Guanghua School of Stomatology, Sun Yat-sen University (ERC-[2018]01). The analysis consisted of two cohorts that began from March to April 2018 and were followed up for 21 months until the end, from December 2019 to January 2020, in Foshan, southern China. The two cohorts included 710 and 345 teenagers aged 13–14 years. Cohort 1 was used to construct the model, which included teenagers from two urban and two rural schools. Cohort 2 was used to evaluate the caries risk prediction model and included teenagers from one urban and one rural school. All participants completed an oral health questionnaire, clinical examination, and donated saliva samples at baseline. Written informed consent was obtained from the guardians of every participant before the study.

Oral Health Questionnaire

Under the guidance of their guardians, the adolescents completed a well-designed oral health questionnaire consisting of three parts: Part 1 was mainly about demographic information, Part 2 was mainly about socioeconomic information, and Part 3 was mainly about oral health-related behaviors (Wang et al., 2020a). The specific variables are as follows:

The variables in part 1: sex, age, residence, whether the child is an only child in his/her family, and his/her primary caregiver.

The variables in part 2: family income, caregivers' education levels, and whether they have dental insurance.

The variables in part 3: frequency of tooth brushing, flossing or mouthwash habits, toothpaste containing fluoride or not, professional application of fluoride, frequency of snack consumption, sweet drink consumption, and attendance in a dental clinic in the past 6 months.

Clinical Examination

Plaque index (PlI) was evaluated using Silness and Löe's scale (Loe, 1967), and six dental indices were recorded. Plaque samples were collected with sterile swabs, according to the procedural instructions of the cariostat kit (GangDa Medical Technology Co. Ltd., Beijing, China). The swabs were then immersed in culture media in ampules and incubated at 37°C for 48 h. Finally, the color of the medium was compared with the reference colors in the color chart provided by the cariostat kit.

After air drying, each tooth was examined and recorded as decayed, missing, or filled (DMFT). The caries status was evaluated according to the International Caries Detection and Assessment System (ICDAS) criteria (Pitts and Ekstrand, 2013). Codes 3–6 in the ICDAS system were recorded as decayed teeth. We also recorded filled and missing teeth due to caries. Oral examinations were conducted at both the baseline and after 21 months in the classrooms.

The students rinsed their mouths before the collection of unstimulated saliva. Unstimulated saliva was collected for 15 min. Students were first asked to swallow all the saliva in the mouth, then spit all the saliva into the scaled tube every 3 min and

five times in total. The saliva flow rate (ml/min) was calculated, and saliva buffering capability was measured according to the Ericsson method. One milliliter of saliva was added to 3 ml of 3.3 mmol HCl within 5 min after collection and then allowed to stand for 20 mins. The final pH of the saliva was evaluated by an electrical pH meter (Wang et al., 2020b).

Selection of Candidate Genetic Markers and DNA Analysis

Single nucleotide polymorphisms (SNPs) were selected based on the results of previous studies on caries susceptibility (n = 4) and screening of tag SNPs (n = 19). We used a candidate gene approach or related-pathway strategies to screen tag SNPs. Caries-related pathway genes, such as those involved in enamel formation, immune responses, saliva secretion, and taste, were identified based on the pathogenesis of caries. The tag SNPs were screened as described in our previous study (Wang et al., 2020b). Thus, 23 target SNPs were detected in all study participants (**Table 1**).

From each participant, 2 ml of unstimulated saliva samples were collected and stored in Oragene DNA Self-Collection kits (Lang Fu, China) at room temperature until they were processed. Genomic DNA was extracted from saliva samples according to the manufacturer's instructions. DNA samples were first purified using MassARRAY Nanodispenser (Sequenom, United States) and then transferred to a SpectroCHIP (Sequenom, United States) chip. Finally, the SNP markers were sequenced by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) (Pang et al., 2017). First, 10 ng of genomic DNA were amplified by PCR in a final volume of 0.5 μL containing locus-specific primers at a final concentration of 10 μmol/L using 0.1-unit HotStarTaq DNA polymerase (Qiagen, Hilden, Germany). PCR conditions were 94°C for 3 min for hot start followed by 40 cycles of denaturation at 94°C for 30 s, annealing at 56°C for 25 s, and extension for 30 s at 72°C, and, finally, incubation at 72°C for 3 min. Then, PCR products were treated with shrimp alkaline phosphatase (Amersham, Freiburg, Germany) for 40 min at 37°C to remove excess deoxynucleotide triphosphates followed by 5 min at 85°C to inactivate shrimp alkaline phosphatase. Base extension reaction conditions were 94°C for 30 s followed by 40 cycles of 94°C for 5 s, 52°C for 5 s, and 80°C for 5 s, and, finally, incubation at 72°C for 3 min. The final base extension products were treated with SpectroCLEAN resin (Sequenom) to remove salts from the reaction buffer. A total of 10 nl of the reaction solution was dispensed onto a 384 format SpectroCHIP microarray (Sequenom, SanDiego, CA). The MassARRAY Analyzer Compac was used for data acquisitions from the MassARRAY SpectroCHIP. Genotyping calls were made in real-time with the Mass Array RT software (Sequenom) (Pang et al., 2020).

TABLE 1 | Candidate genetic markers evaluated in this study.

Gene	Chromosome	Marker public ID	Base pair exchange (MAF)	Most severe consequence
Enamel formation genes				
ENAM	4	rs12640848	A/G (0.33)	Intron variant
		rs3796703	C/T (0.01)	Missense(leu)
AMBN	4	rs13115627	A/G (0.30)	Intron variant
AMELX	X	rs946252	C/T (0.31)	Intron variant
TFIP11	22	rs134143	T/C (0.35)	Non-coding transcript exon variant
		rs2097470	C/T (0.29)	Intron variant
MMP20	11	rs1612069	G/T (0.48)	Intron variant
		rs1784418	C/T (0.42)	Intron variant
TUFT1	1	rs17640579	A/G (0.22)	Intron variant
		rs3790506	G/A (0.25)	Intron variant
Immune response genes				
DEFB1	8	rs11362	C/T (0.40)	5 prime UTR variant
		rs1800972	G/C (0.14)	5 prime UTR variant
LTF	3	rs4547741	C/T (0.07)	Intron variant
		rs1126478	C/T (0.37)	Missense variant
MBL2	10	rs1800450	C/T (0.12)	Missense variant
		rs11003125	G/C (0.31)	Intron variant, upstream variant 2 KB
MASP2	1	rs10779570	T/G (0.36)	Intron variant
Water channel protein gene				
AQP5	12	rs1996315	G/A (0.43)	Intron variant, upstream variant 2 KB
		rs923911	C/A (0.22)	Intron variant, upstream variant 2 KB
Saliva secretion gene				
CA6	1	rs2274327	C/T (0.27)	Intron variant, missense
Taste gene				
TAS1R2	1	rs35874116	T/C (0.27)	Missense variant
		rs9701796	C/G (0.20)	missense variant
TAS2R38	7	rs713598	G/C (0.50)	Missense variant

Statistical Analysis

Data of all teenagers in cohort 1 were used to construct a CRPM with random forest, and those of teenagers from cohort 2 were used to verify this newly constructed model. The logistic regression model was used as a reference for performance evaluation. When we analyzed the variables associated with the occurrence and development of caries, the independent variable included the environmental variables and SNPs. The dependent variable was DMFT increment (ΔDMFT) over 21 months of follow-up, which is the outcome of this study. A previous study conducted by Chaffee BW (Chaffee et al., 2015) found that the DMFT increment was about 1.01 in the low caries risk groups after 18 months of follow-up. Remember that individuals with DMFT increments of no more than one caries after 21 months of follow-up should be classified in the low caries risk group. Chi-square tests were used to identify SNPs associated with increased risk of caries, and univariate logistic analysis was used to select environmental factors associated with caries. Variables with $P < 0.1$ were considered statistically significant and used as predictors in the caries risk prediction

model. R 3.6.1 software was used to construct the model. Using the data of the training cohort (cohort1), the random forest package was used to train the random forest model, and the nTree and mtry parameters were debugged. The random forest prediction model was the most effective when nTree = 300 and mtry = 2. In the model constructed with cohort 1, we segmented the population into five different caries risk layers based on the 5-quantiles: very low, low, moderate, high, and very high caries risk. Then, we stratified the caries risk in the cohort 2 (testing cohort) population based on the cutoff value in cohort 1. The discrimination ability of the model was evaluated using receiver operator characteristic (ROC) curve analysis. The calibration ability of the model was measured via a risk stratification plot, which was used to demonstrate the similarity of the predicted absolute risk to the absolute observed risk at different risk levels.

RESULTS

Characteristics of Study Samples

In total, 1,055 teenagers (710 in cohort 1 and 345 in cohort 2) were recruited. The average age at baseline was 13.19 ± 0.40 years (Wang et al., 2020a). The questionnaire was completed by all teenagers. After 21 months, 953 teenagers (including 633 teenagers in cohort 1 and 320 teenagers in cohort 2) were followed up. During these 21 months, follow-up was lost for only 102 (9.66%) teenagers. The main reasons for loss of follow-up were absence in school or transfer from schools. The flow chart of the prospective longitudinal study is shown in **Figure 1**.

At baseline, 34.37% of the teenagers in cohort 1 and 39.88% of those in cohort 2 were affected by caries, and the mean (SD) DMFTs were 0.67 ± 1.25 and 0.84 ± 1.38, respectively. After 21 months, 57.66% of the teenagers in cohort 1 and 63.13% of those in cohort 2 developed more than one caries (ΔDMFT > 1). The mean (SD) increases in DMFTs after 21 months were 2.40 ± 2.97 in cohort 1 and 2.73 ± 3.21 in cohort 2.

Caries Risk Prediction Factors

Table 2 shows the results of a logistic analysis of the association between environmental variation and caries. Among the environmental variations, we found that "sex," "dental attendance in the past 6 months," "cariostat score," and "past caries experience" were significantly associated with the caries risk (all $P < 0.05$).

Table 3 shows the results of the chi-square tests on the association between SNPs and caries. Among all the SNPs, rs1996315 (*AQP5*), and rs3790506 (*TUFT1*) were significantly associated with caries risk (all $P < 0.05$).

CRPM Training and Validation

The CRPM has been developed using logistic regression and random forest. The performance of CRPM developed using logistic regression was 0.70 (0.66–0.74) for the training cohort (**Figure 2A**) and 0.74 (0.68–0.79) for the test cohort (**Figure 2B**). The performance of the random forest was 0.78 (0.75–0.82) for the training cohort (**Figure 3A**) and 0.73 (0.67–0.78) for

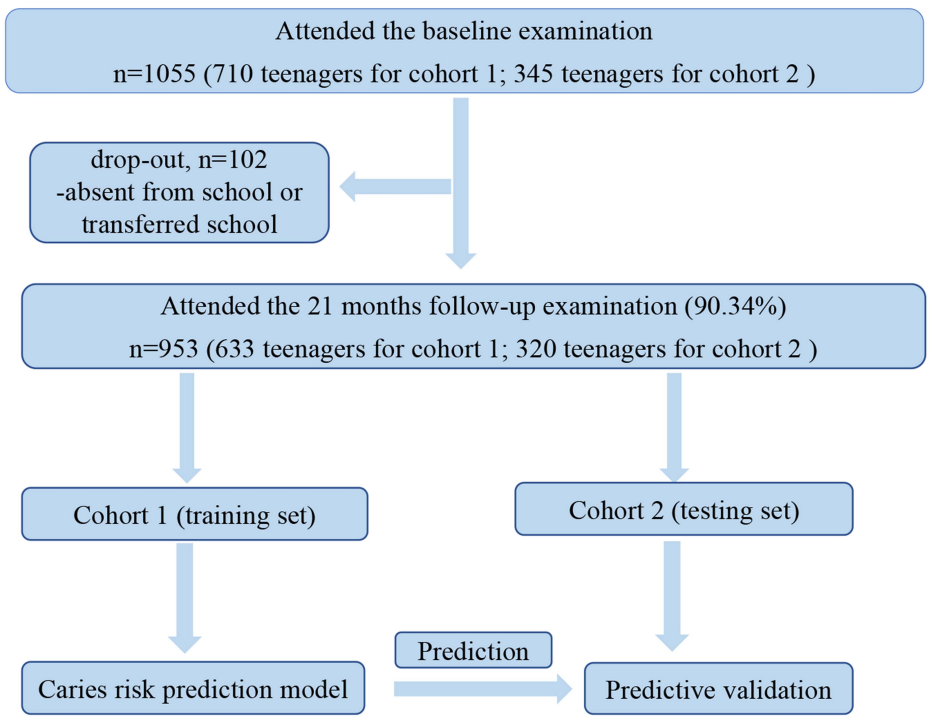


FIGURE 1 | Flow chart of the prospective longitudinal study.

TABLE 2 | Logistic analysis of the association between environmental factors and caries.

Characteristics	Level	ΔDMFT ≤ 1 (*n* = 328)	ΔDMFT > 1 (*n* = 305)	*P*-value
Pit and fissure sealant (%)	No	320 (97.6)	296 (97.0)	0.879
	Yes	8 (2.4)	9 (3.0)	
Sex (%)	Female	118 (36.0)	135 (44.3)	0.041*
	Male	210 (64.0)	170 (55.7)	
Frequency of tooth brushing (%)	<1 times/day	7 (2.1)	6 (2.0)	0.127
	1 times/day	146 (44.5)	112 (36.7)	
	2 times/day	175 (53.4)	187 (61.3)	
Toothpaste (%)	No	1 (0.3)	2 (0.7)	0.95
	Yes	327 (99.7)	303 (99.3)	
Mouthwash (%)	No	243 (74.1)	230 (75.4)	0.771
	Yes	85 (25.9)	75 (24.6)	
Dental flossing (%)	No	301 (91.8)	288 (94.4)	0.247
	Yes	27 (8.2)	17 (5.6)	
Professional application of fluoride (%)	No	313 (95.4)	294 (96.4)	0.68
	Yes	15 (4.6)	11 (3.6)	
Dental attendance in the past 6 months (%)	No	166 (50.6)	122 (40.0)	0.009*
	Yes	162 (49.4)	183 (60.0)	
One-child family (%)	No	250 (76.2)	252 (82.6)	0.059*
	Yes	78 (23.8)	53 (17.4)	
Activity (%)	No	108 (32.9)	107 (35.1)	0.625
	Yes	220 (67.1)	198 (64.9)	
Cariostat score (%)	Low	85 (25.9)	48 (15.7)	<0.001*
	Medium	198 (60.4)	183 (60.0)	
	High	45 (13.7)	74 (24.3)	
Plaque Index (%)	Low	31 (9.5)	23 (7.5)	0.057*
	Medium	119 (36.3)	139 (45.6)	
	High	178 (54.3)	143 (46.9)	
Residence (%)	Urban	171 (52.1)	151 (49.5)	0.561
	Rural	157 (47.9)	154 (50.5)	
Toothpaste (%)	Non-fluoride	79 (24.1)	91 (29.8)	0.123
	Fluoride	249 (75.9)	214 (70.2)	
Saliva buffering capability (pH) (%)	PH < 3.5	94 (28.7)	94 (30.8)	0.895
	PH 3.5–4.24	104 (31.7)	89 (29.2)	
	PH 4.25–4.75	50 (15.2)	48 (15.7)	
	PH > 4.75	80 (24.4)	74 (24.3)	
Dental insurance (%)	No	251 (76.5)	230 (75.4)	0.814
	Yes	77 (23.5)	75 (24.6)	
Caregiver (%)	Mother	194 (59.1)	192 (63.0)	0.151
	Father	48 (14.6)	28 (9.2)	
	Grandparents	17 (5.2)	11 (3.6)	
	Nursemaid	11 (3.4)	8 (2.6)	
	No regular caregiver	58 (17.7)	66 (21.6)	
Education of caregiver (%)	<9 years	293 (89.3)	272 (89.2)	1
	≥9 years	35 (10.7)	33 (10.8)	
Household monthly income (CNY) (%)	<3,000	54 (16.5)	48 (15.7)	0.97
	3,000–7,000	192 (58.5)	180 (59.0)	
	≥7,000	82 (25.0)	77 (25.2)	

(Continued)

TABLE 2 | Continued

Characteristics	Level	△DMFT ≤ 1	△DMFT>1	*P*-value
Frequency of snacks consuming (%)	<1 per day	215 (65.5)	211 (69.2)	0.374
	≥1 per day	113 (34.5)	94 (30.8)	
Saliva secretion(ml/min)	<0.1	31 (9.5)	33 (10.8)	0.801
	0.1–0.25	62 (18.9)	60 (19.7)	
	>0.25	235 (71.6)	212 (69.5)	
Frequency of sweet drinks consuming (%)	<1 per day	212 (64.6)	193 (63.3)	0.786
	≥1 per day	116 (35.4)	112 (36.7)	
Past caries experience (%)	No	273 (83.2)	170 (55.7)	<0.001*
	Yes	55 (16.8)	135 (44.3)	

*ΔDMFS, mean increment of decayed, missing, or filled surfaces over 21 months. Past caries experience means whether the individual had caries at the baseline examination or not. Univariate logistic regression was used to analyze the environmental factors related to the occurrence and development of caries. *P < 0.1.*

TABLE 3 | Chi-square test analysis of the association between SNPs and caries.

		ΔDMFT ≤ 1			ΔDMFT>1					
SNP	**Allele 1/2**	**11**	**12**	**22**	**11**	**12**	**22**	**OR**	**95% CI**	***P*-value**
rs10779570	G/T	21	111	196	17	105	183	0.97	0.75–1.26	0.824
rs11003125	C/G	55	173	100	48	170	87	1.02	0.81–1.29	0.860
rs1126478	C/T	163	133	32	144	121	40	1.15	0.91–1.45	0.231
rs11362	C/T	121	161	46	115	138	52	1.06	0.85–1.33	0.604
rs12640848	A/G	219	95	14	198	94	13	1.07	0.81–1.41	0.631
rs13115627	A/G	190	118	20	175	120	10	0.93	0.71–1.21	0.578
rs134143	T/C	152	132	44	129	140	36	1.05	0.83–1.31	0.699
rs1612069	G/T	84	177	67	77	176	52	0.93	0.74–1.18	0.567
rs17640579	A/G	176	133	19	156	121	28	1.17	0.91–1.5	0.214
rs1784418	C/T	95	168	65	77	169	59	1.07	0.85–1.35	0.548
rs1800450	C/T	239	82	7	235	63	7	0.85	0.61–1.16	0.305
rs1800972	G/C	260	60	8	240	63	2	0.93	0.66–1.31	0.671
rs1996315	G/A	110	160	58	116	154	35	0.79	0.62–0.99	0.042*
rs2097470	C/T	170	136	22	150	139	16	1.02	0.79–1.32	0.858
rs2274327	C/T	162	139	27	140	141	24	1.07	0.84–1.37	0.579
rs35874116	C/T	4	58	266	0	77	228	1.31	0.92–1.89	0.138
rs3790506	G/A	187	125	16	158	114	33	1.33	1.04–1.71	0.024*
rs3796703	C/T	309	15	4	287	13	5	1.06	0.64–1.76	0.830
rs457741	C/T	293	32	3	277	28	0	0.76	0.46–1.25	0.283
rs713598	C/G	30	137	161	31	125	149	1.03	0.81–1.31	0.811
rs923911	C/A	199	116	13	201	87	17	0.90	0.69–1.17	0.434
rs946252	C/T	136	54	138	127	66	112	0.93	0.79–1.11	0.440
rs9701796	C/G	204	112	12	194	99	12	0.97	0.74–1.29	0.857

*ΔDMFS, mean increment of decayed, missing, or filled surfaces over 21 months. Chi-square test was used to analyze the SNPs related to the occurrence and development of caries. *P < 0.05.*

the test cohort (**Figure 3B**). The results showed that the prediction performance of the CRPM constructed using Random Forest was stable.

The Gini coefficient of the random forest suggested that the selected variables in this prediction model could be arranged as follows according to their importance: "past caries experience," "cariostate score," "plaque index," "rs3790506," "rs1996315," "gender," and "whether they were only teenagers" (**Figure 4**).

The ability of the CRPM to identify caries risk in individuals was examined further. A risk stratification plot was created, in which the data from 320 patients in cohort 2 were sorted by increasing the predicted risk and separated into five risk layers: very low, low, medium, high, and very high. Then, the actual rate of caries incidence after 21 months was calculated for each risk layer. **Figure 5** shows the degree of discrepancy between the actual and predicted risks for each of the five risk layers.

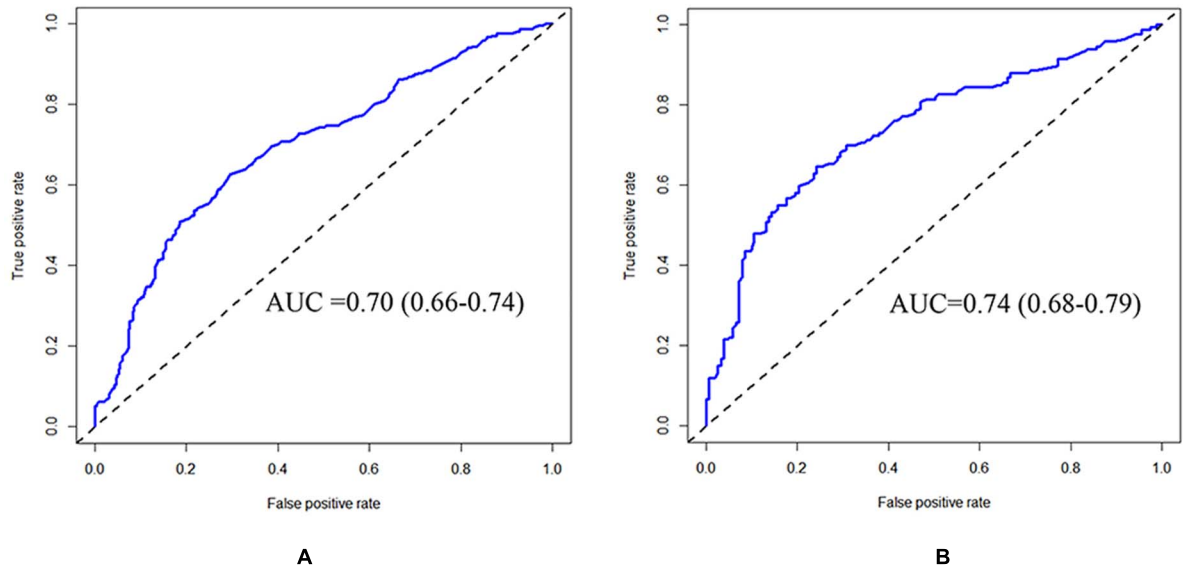


FIGURE 2 | ROC curve of training and testing cohort (Logistic Regression Model). Measurement of the discrimination ability of the caries risk prediction model (Logistic Regression) with ROC curve. The AUC (95%CI) of the training cohort was 0.70 (0.66–0.74) **(A)**, and the AUC (95% CI) of the testing cohort was 0.74 (0.68–0.79) **(B)**.

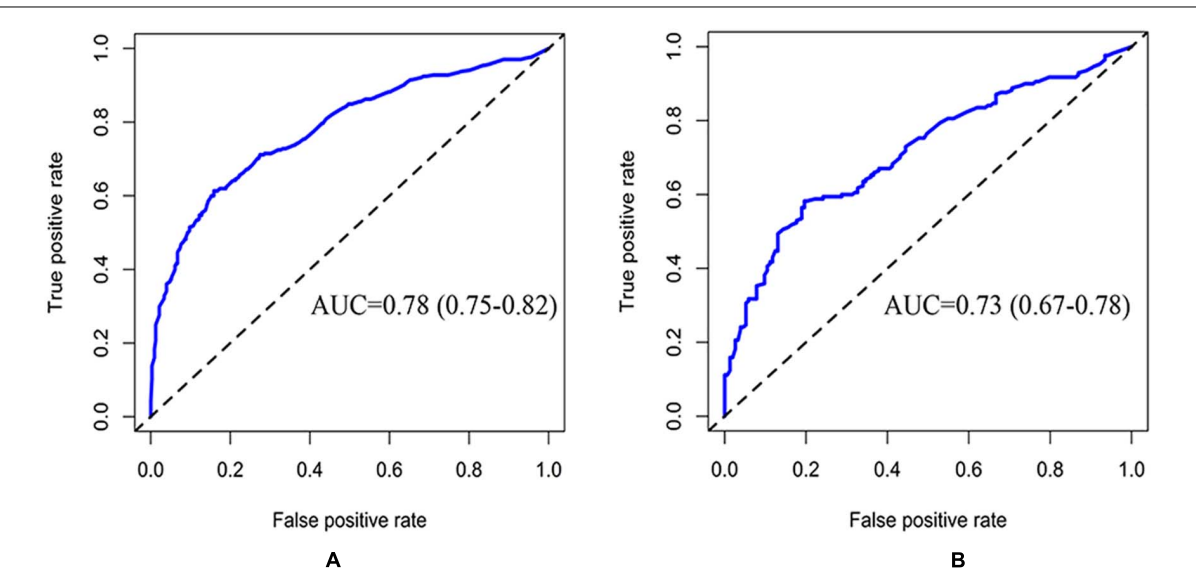


FIGURE 3 | ROC curve of training and testing cohort (Random Forest Model). Measurement of the discrimination ability of the caries risk prediction model (Random Forest) with ROC curves. The AUC of the training cohort was 0.78 (0.75–0.82) **(A)**, and the AUC of the testing cohort was 0.73 (0.67–0.78) **(B)**.

Using the CRPM constructed with the training cohort, we assigned the participants in cohort 1 into five risk groups based on the 5-quantiles of the predicted incidence probabilities as follows: very low, low, medium, high, and very high. The predicted incidence rates of caries after 21 months in cohort 1 for each risk layer were 5.60, 16.02, 33.29, 65.06, and 90.51%, respectively, and the actual incidence rates of caries after 21 months in cohort 1 for each risk layer were 18.25, 31. 71, 39. 34, 61. 94, and 87.50%, respectively (**Table 4**). The numbers of individuals in the caries layers of cohort 2, i.e., very low, low, medium, high, and very high, were 48,49,73,102, and 48, respectively, and the mean DMFT increment in each risk layer are shown in **Table 5**; the predicted incidence rates of caries after 21 months in each risk layer of cohort 2 were 5.41, 16.79, 33.56, 66.20, and 91.07%, respectively, and the actual incidence rates of caries after 21 months in each risk layer of cohort 2 were 27.08, 34.69, 47.95, 59.80, and 85.42%, respectively (**Table 5**). The risk of new caries was consistently reduced from the extremely

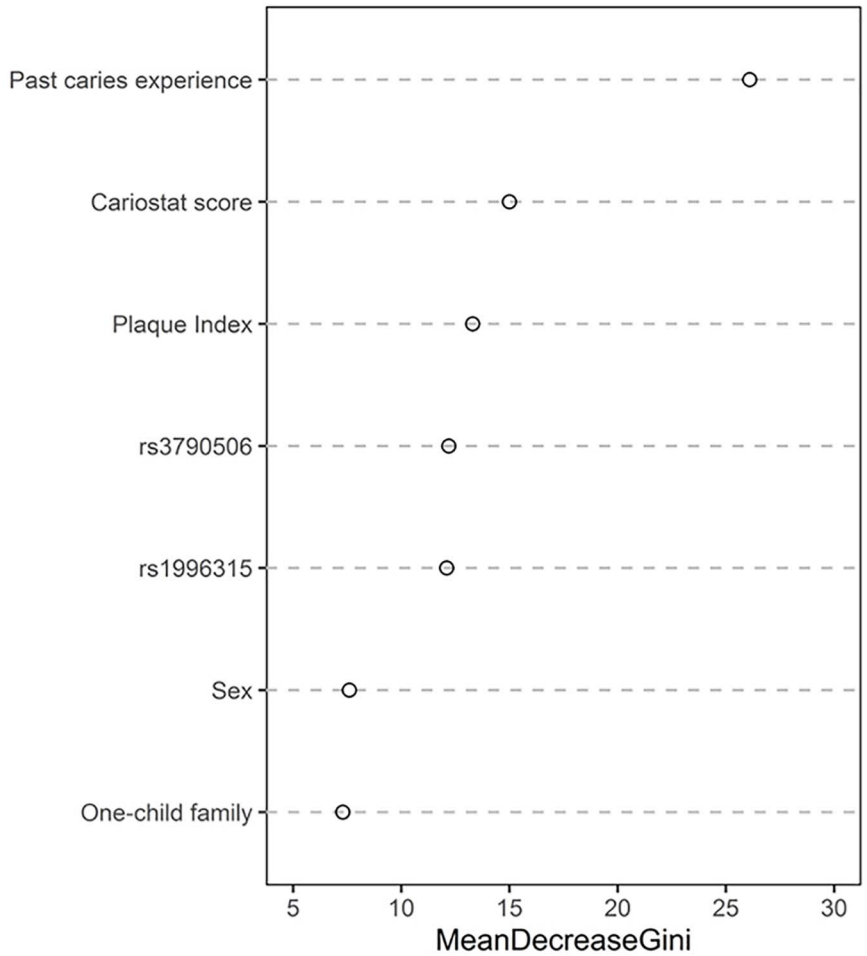


FIGURE 4 | The Gini coefficient of the random forest.

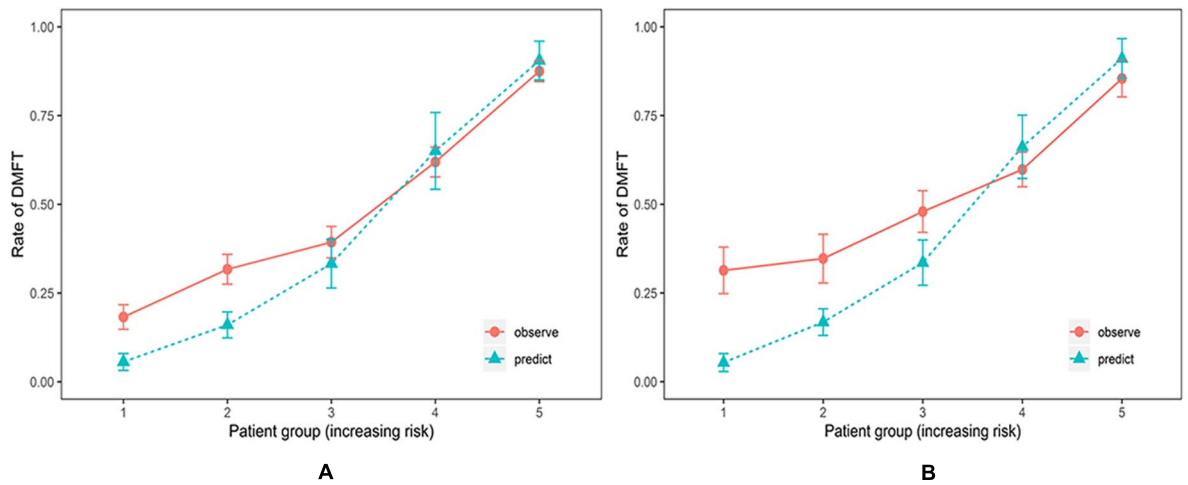


FIGURE 5 | Risk stratification plot for the training and testing cohort (Random Forest Model). Relationship between observed (orange, 95% confidence intervals) and predicted (green) scores of new carious lesions for 21 months for the training cohort **(A)** and the testing cohort **(B)**. The prediction model could accurately estimate risk for individuals at high and very high caries risk but underestimated risks for individuals at low and very low caries risk.

TABLE 4 | Actual number of new caries after 21 months: actual and predicted caries incidences in cohort 1.

Caries risk	Total number of participants in cohort 1 (n)	Actual number of new caries incidence in cohort 1 (n)	Actual caries incidence in cohort 1 (%)	Predicted caries incidence in cohort 1 (%)
Very low	126	23	18.25	5.60
Low	123	39	31. 71	16.02
Moderate	122	48	39. 34	33.29
High	134	83	61. 94	65.06
Very high	128	112	87. 50	90.51

TABLE 5 | Actual number of new caries after 21 months: actual and predicted caries incidences in cohort 2.

Caries risk	Total number of participants in cohort 2 (n)	Actual number of new caries incidence in cohort 2 (n)	Actual caries incidence in cohort 2 (%)	Predicted caries incidence in cohort 2 (%)	Caries increment mean (SD)
Very low	48	13	27.08	5.41	1.25 ± 2.12
Low	49	17	34.69	16.79	1.67 ± 2.63
Moderate	73	35	47.95	33.56	2.39 ± 2.93
High	102	61	59.80	66.20	3.43 ± 3.72
Very high	48	41	85.42	91.07	4.33 ± 2.90

high-risk category to the extremely low-risk category, reflecting the ability of our newly constructed CRPM to estimate future caries accurately.

The sensitivity, specificity, positive predictive value, and negative predictive value of cohorts 1 and 2 are displayed in **Table 6**. The positive predictive value was high (>73%) for those stratified into very high and high caries risk categories. When the "moderate caries risk" and "low caries risk "categories were used as a cutoff level, the negative predictive values were low.

DISCUSSION

In this study, a new caries risk prediction model was constructed, using both environmental risk factors, such as cariostate score, plaque index, and past caries experience, and genetic factors as predictors. To our knowledge, this is the first CRPM constructed with both environmental and genetic factors, using machine learning algorithms. We further verified the accuracy of this prediction model using another independent cohort, and the results demonstrated that this CRPM could effectively identify high caries-risk individuals.

It is well recognized that dental caries is a multifactorial disease. Environmental and genetic factors play important roles in the occurrence and development of caries (Yildiz et al., 2016). Combining genetic factors with environmental factors to explain the incidence of caries is both reasonable and necessary. Being a polygenetic disease, caries is difficult to predict based on a single SNP or SNPs of individual genes. Hence, it is necessary to select SNPs from different candidate genes. In this study, SNPs were selected based on the results of previous studies, combining tag SNP screening via related-pathway strategies and candidate gene approach (Opal et al., 2015). Finally, 23 SNPs from 16 candidate genes were included in this study. After analyzing the correlation of each SNP, two SNPs were found to be associated with caries in the Chinese population.

The SNPs included in the final CRPM described here were rs3790506 and rs1996315. Of these, rs3790506 is an SNP of *TUFT1*, which is involved in enamel development and mineralization. Previous studies have reported a relationship between *TUFT1* and caries incidence in both children and adults. Slayton et al. suggested that rs3790506 in *TUFT1* interacts with the *Streptococcus mutans* present in the oral cavity and further explained over a quarter of the factors affecting the variability of caries conditions in teenagers from Iowa, United States

TABLE 6 | Sensitivity, specificity, and predictive values for new caries lesions over 21 months.

Caries risk	Sensitivity (%)		Specificity (%)		PPV[a] (%)		NPV[b] (%)		Youden's index[c]	
	Cohort 1	Cohort 2	Cohort 1	Cohort 2	Cohort 1	Cohort 2	Cohort 1	Cohort 2	Cohort 1	Cohort 2
Very-high	67.8	65.8	75.0	57.2	95.0	90.0	25.0	22.2	0.43	0.23
High	54.2	59.0	68.7	68.3	73.8	73.5	47.9	52.8	0.23	0.27
Moderate	45.8	34.3	69.0	65.8	48.9	48.0	66.2	52.1	0.15	0.001
Low	41.0	29.4	73.9	62.5	42.1	29.4	72.9	62.5	0.15	0.08

[a] *PPV, positive predictive value.*
[b] *NPV, negative predictive value.*
[c] *Youden's index, sensitivity + specificity − 1.*

(Slayton et al., 2005). rs1996315 is a SNP of *AQP5*, which encodes a water channel protein expressed in lacrimal and salivary glands and epithelial cells. Aquaporins play a role in the generation of tears, saliva, and pulmonary secretions. *AQP5* protein also plays an important role in extracellular matrix hydration during tooth development (Felszeghy et al., 2004). It has been reported that variations in *AQP5* could contribute to the occurrence and development of caries (Wang et al., 2012; Anjomshoaa et al., 2015). Our previous study showed that gene-gene interaction between rs1996315 and rs923911 was significantly associated with molar-incisor hypomineralization (Pang et al., 2020). Both SNPs included in the CRPM constructed in this study were associated with enamel development. The etiological theory of dental caries states that enamel characteristics also affect the pathogenesis of dental caries, although it is not feasible to detect the physical and chemical characteristics of enamel *in vivo*. The identification of variations in enamel-related genes can indirectly reflect enamel characteristics associated with the occurrence of dental caries. Although genetic factors were included in this CRPM, it should be noted that environmental factors were more dominant than genetic factors. Silva et al. revealed that, compared to environmental factors, genetic factors have relatively little influence on the risk of dental caries, which is consistent with the results of our study (Silva et al., 2019).

In accordance with the results of traditional CRPMs, such as the Cariogram model, the CRPM constructed in this study using a machine learning algorithm identified "past caries experience" as the strongest predictor of individual risk. Besides the "past caries experience," "cariostate score," "plaque index," "gender," and "whether they were only teenagers in the family" were also included in this new CRPM. Unlike the Cariogram model, we used the "cariostate score" instead of "bacterial counts" to evaluate the cariogenic ability of the dental plaque. Cariostat uses a colorimetric test to evaluate the acid produced by bacteria in the plaque (Ramesh et al., 2013). The occurrence of carious lesions is a dynamic process in which acids produced by bacteria impact the demineralization of dental tissues (Richards et al., 2017). When the pH of the tooth surface decreases to a level < 5.5, the hydroxyapatite (HA) matrix of the tooth starts to demineralize; Cariostat can assess the activity of the caries microbiology. Unlike other cariogenic microbiology tests, such as Dentocult SM, Cariostat assesses bacteria in plaque instead of saliva, leading to higher accuracy because cariogenic bacteria act on tooth surfaces in the form of plaque.

An ideal but possibly unrealistic model will correctly distinguish individuals at risk of a caries event from those who are not at risk, without any instance of misdiagnosis (Alba et al., 2017). The extent to which a model can achieve this goal is represented by two related properties of discrimination and calibration (Alba et al., 2017). Discrimination refers to the extent to which the model distinguishes between high-risk and low-risk participants of an event, usually described by the receiver operating characteristic (ROC) curve. It is well recognized that an AUC < 0.6 represents poor discrimination, while an AUC ≥ 0.7 indicates high discrimination ability (Fontana et al., 2020). The training set resulted in an AUC of 0.78 in cohort 1 and 0.73 in cohort 2, indicating high discrimination ability.

Discrimination alone is not sufficient to evaluate the performance of a prediction model. The second essential characteristic of a prediction model is demonstrating the similarity of the predicted absolute risk to the absolute observed risk at different risk levels. Calibration is usually considered the most important characteristic of a prediction model because it reflects the extent to which a model correctly predicts the absolute risk (Alba et al., 2017). In terms of accurate estimation, the model is well-calibrated. The relationship between predicted and observed risk could be visually represented, allowing efficient evaluation of the calibration (Alba et al., 2017). We found that the CRPM constructed in this study can accurately estimate the risks of individuals at high and very high caries risks but underestimates those for individuals at low and very low caries risks. However, this poor calibration may not pose a problem for low-risk individuals because the purpose of this CRPM is to identify teenagers at high risk of developing caries for better prevention and intervention, and the underestimation of patients at lower risk would be rather irrelevant. Hence, our CRPM can be considered a useful tool for selecting high caries risk population in China.

Our study has several limitations. First, although the SNPs were selected based on the results of previous studies on caries susceptibility and through screening of tag SNPs from multiple genes, it cannot be ruled out that some key loci with powerful diagnostic performance were missed. As an infectious disease, caries risk will certainly be affected by microorganisms. Even if we use "cariostate score" to evaluate the cariogenic ability of the dental plaque, the prediction performance might be influenced by microbiome markers. Although the ICDAS system was used to record caries, earlier signs (ICDAS code 1 or 2) of caries were not detected in our study. In addition, despite external verification with an independent cohort, further multicenter research is also highly needed.

In conclusion, we constructed a CRPM based on both environmental and genetic factors using a machine learning algorithm. We also estimated the discrimination and calibration abilities of this CRPM using a separate independent cohort for validation, demonstrating that this CRPM can accurately identify a high caries risk population. Our CRPM included specific patient characteristics, such as SNPs, gender, and whether the participants were the only child of the respective families, to provide an estimate of the absolute risk of a specific caries outcome. Thus, our CRPM can be utilized as a powerful tool at the community level for identifying high caries risk groups, enabling policymakers to plan necessary preventive measures for the future.

DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in the European Variation Archive (EVA) repository, accession number PRJEB43233. The data will first be made available to download here: https://www.ebi.ac.uk/ena/data/view/PRJEB43233.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Guanghua School of Stomatology, Sun Yat-sen University (ERC- [2018]01). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

LP contributed to conception, design, and drafted the manuscript. KW contributed to data acquisition, analysis, and critically revised manuscript. YT contributed to design and critically revised manuscript. QZ contributed to conception and drafted manuscript. JZ contributed to design and critically revised manuscript. HL contributed to conception, design, and critically revised manuscript. All authors gave final approval and agreed to be accountable for all aspects of the work.

FUNDING

This research was supported by the National Natural Science Foundation of China (Grant No. 81903345).

REFERENCES

Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P. J., et al. (2017). Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* 318, 1377–1384. doi: 10.1001/jama.2017.12126

Anjomshoaa, I., Briseño-Ruiz, J., Deeley, K., Poletta, F. A., Mereb, J. C., Leite, A. L., et al. (2015). Aquaporin 5 interacts with fluoride and possibly protects against caries. *PLoS One* 10:e143068. doi: 10.1371/journal.pone.0143068

Cagetti, M. G., Bonta, G., Cocco, F., Lingstrom, P., Strohmenger, L., and Campus, G. (2018). Are standardized caries risk assessment models effective in assessing actual caries status and future caries increment? A systematic review. *BMC Oral Health* 18:123. doi: 10.1186/s12903-018-0585-4

Chaffee, B. W., Cheng, J., and Featherstone, J. (2015). Baseline caries risk assessment as a predictor of caries incidence. *J. Dent.* 43, 518–524. doi: 10.1016/j.jdent.2015.02.013

Domejean, S., Banerjee, A., and Featherstone, J. (2017). Caries risk/susceptibility assessment: its value in minimum intervention oral healthcare. *Br. Dent. J.* 223, 191–197. doi: 10.1038/sj.bdj.2017.665

Felszeghy, S., Módis, L., Németh, P., Nagy, G., Zelles, T., Agre, P., et al. (2004). Expression of aquaporin isoforms during human and mouse tooth development. *Arch. Oral Biol.* 49, 247–257. doi: 10.1016/j.archoralbio.2003.09.011

Fontana, M., Carrasco-Labra, A., Spallek, H., Eckert, G., and Katz, B. (2020). Improving caries risk prediction modeling: a call for action. *J. Dent. Res.* 99, 1215–1220. doi: 10.1177/0022034520934808

Haworth, S., Esberg, A., Lif Holgerson, P., Kuja-Halkola, R., Timpson, N. J., Magnusson, P. K. E., et al. (2020). Heritability of caries scores, trajectories, and disease subtypes. *J. Dent. Res.* 99, 264–270. doi: 10.1177/0022034519897910

Kaste, L. M., Selwitz, R. H., Oldakowski, R. J., Brunelle, J. A., Winn, D. M., and Brown, L. J. (1996). Coronal caries in the primary and permanent dentition of children and adolescents 1-17 years of age: united states, 1988-1991. *J. Dent. Res.* 75, 631–641. doi: 10.1177/00220345960750S03

Li, C., Sun, D., Liu, J., Li, M., Zhang, B., Liu, Y., et al. (2019). A prediction model of essential hypertension based on genetic and environmental risk factors in northern han chinese. *Int. J. Med. Sci.* 16, 793–799. doi: 10.7150/ijms.33967

Loe, H. (1967). The gingival index, the plaque index, and the retention index systems. *J. Periodontol.* 38, 610–616. doi: 10.1902/jop.1967.38.6.610

National Institutes of Health Consensus Development Conference Statement (2001). Diagnosis and management of dental caries throughout life, March 26-28, 2001. *J. Am. Dent. Assoc.*132, 1153–1161. doi: 10.14219/jada.archive.2001.0343

Okubo, Y., Nakano, Y., Ochi, H., Onohara, Y., Tokuyama, T., Motoda, C., et al. (2020). Predicting atrial fibrillation using a combination of genetic risk score and clinical risk factors. *Heart Rhythm* 17, 699–705. doi: 10.1016/j.hrthm.2020.01.006

Opal, S., Garg, S., Jain, J., and Walia, I. (2015). Genetic factors affecting dental caries risk. *Aust. Dent.* J. 60, 2–11. doi: 10.1111/adj.12262

Pang, L., Li, X., Wang, K., Tao, Y., Cui, T., Xu, Q., et al. (2020). Interactions with the aquaporin 5 gene increase the susceptibility to molar-incisor hypomineralization. *Arch. Oral Biol.* 111:104637. doi: 10.1016/j.archoralbio.2019.104637

Pang, L., Zhi, Q., Zhuang, P., Yu, L., Tao, Y., and Lin, H. (2017). Variation in enamel formation genes influences enamel demineralization in vitro in a *Streptococcus mutans* biofilm model. *Front. Physiol.* 8:851. doi: 10.3389/fphys.2017.00851

Patir, A., Seymen, F., Yildirim, M., Deeley, K., Cooper, M. E., and Marazita, M. L. (2008). Enamel formation genes are associated with high caries experience in Turkish children. *Caries Res.* 42, 394–400. doi: 10.1159/000154785

Petersson, G., and Twetman, S. (2015). Caries risk assessment in young adults: a 3 year validation of the Cariogram model. *BMC Oral Health* 27:17. doi: 10.1186/1472-6831-15-17

Pitts, N. B., and Ekstrand, K. R. (2013). International caries detection and assessment system (icdas) and its international caries classification and management system (iccms) – methods for staging of the caries process and enabling dentists to manage caries. *Community Dent. Oral Epidemiol.* 41, e41–e52. doi: 10.1111/cdoe.12025

Ramesh, K., Kunjappan, S., Ramesh, M., Shankar, S., and Reddy, S. (2013). Comparative evaluation of predictive value of three caries activity tests-snyder, lactobacillus count and cariostat in mixed dentition children with and without caries. *J. Pharm. Bioallied Sci.* 5, S63–S68. doi: 10.4103/0975-7406.113299

Richards, V. P., Alvarez, A. J., Luce, A. R., Bedenbaugh, M., Mitchell, M. L., Burne, R. A., et al. (2017). Microbiomes of site-specific dental plaques from children with different caries status. *Infect. Immun.* 85, e00106-17. doi: 10.1128/IAI.00106-17

Righolt, A. J., Jevdjevic, M., Marcenes, W., and Listl, S. (2018). Global-, regional-, and country-level economic impacts of dental diseases in 2015. *J. Dent. Res.* 97, 501–507. doi: 10.1177/0022034517750572

Selwitz, R. H., Ismail, A. I., and Pitts, N. B. (2007). Dental caries. *Lancet* 369, 51–59. doi: 10.1016/S0140-6736(07)60031-2

Silva, M. J., Kilpatrick, N. M., Craig, J. M., Manton, D. J., Leong, P., Burgner, D. P., et al. (2019). Genetic and early-life environmental influences on dental caries risk: a twin study. *Pediatrics* 143:e20183499. doi: 10.1542/peds.2018-3499

Slayton, R. L., Cooper, M. E., and Marazita, M. L. (2005). Tuftelin, mutans streptococci, and dental caries susceptibility. *J. Dent. Res.* 84, 711–714. doi: 10.1177/154405910508400805

Vieira, A. R., Modesto, A., and Marazita, M. L. (2014). Caries: review of human genetics research. *Caries Res.* 48, 491–506. doi: 10.1159/000358333

Wang, K., Pang, L., Fan, C., Cui, T., Yu, L., and Lin, H. (2020b). Enamel and dentin caries risk factors of adolescents in the context of the International Caries Detection and Assessment System (ICDAS): a longitudinal study. *Front. Pediatr.* 8:419. doi: 10.3389/fped.2020.00419

Wang, K., Pang, L., Tao, Y., Li, X., Zhang, J., Cui, T., et al. (2020a). Association of genetic and environmental factors with dental caries among adolescents in south china: a cross-sectional study. *Eur. J. Paediatr. Dent.* 21, 129–136. doi: 10.23804/ejpd.2020.21.02.07

Wang, X., Willing, M. C., Marazita, M. L., Wendell, S., Warren, J. J., Broffitt, B., et al. (2012). Genetic and environmental factors associated with dental caries

in children: the Iowa fluoride study. *Caries Res.* 46, 177–184. doi: 10.1159/000337282

Yildiz, G., Ermis, R. B., Calapoglu, N. S., Celik, E. U., and Turel, G. Y. (2016). Gene-environment interactions in the etiology of dental caries. *J. Dent. Res.* 95, 74–79. doi: 10.1177/0022034515605281

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 SZ declared a past co-authorship with one of the authors HL to the handling editor.

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 March 2021
doi: 10.3389/fgene.2021.589408

Integrated Transcriptomic Analysis of the miRNA–mRNA Interaction Network in Thin Endometrium

*Lu Zong†, Shengxia Zheng†, Ye Meng, Wenjuan Tang, Daojing Li, Zhenyun Wang, Xianhong Tong and Bo Xu**

Reproductive and Genetic Hospital, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China

Although the thin endometrium (TE) has been widely recognized as a critical factor in implantation failure, the contribution of miRNA–mRNA regulatory network to the development of disease etiology remains to be further elucidated. This study performed an integrative analysis of the miRNA–mRNA expression profiles in the thin and adjacent normal endometrium of eight patients with intrauterine adhesion to construct the transcriptomic regulatory networks. A total of 1,093 differentially expressed genes (DEGs) and 72 differentially expressed miRNAs (DEMs) were identified in the thin adhesive endometrium of the TE group compared with the control adjacent normal endometrial cells. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses showed that the DEGs and the target genes of DEM were significantly enriched in angiogenesis, cell growth regulation, and Wnt signaling pathway. Multiple hub genes (CAV1, MET, MAL2, has-mir-138, ARHGAP6, CLIC4, RRAS, AGFG1, has-mir-200, and has-mir-429) were identified by constructing the miRNA–mRNA regulatory networks. Furthermore, a miRNA–mRNA pathway function analysis was conducted, and the hub genes were enriched in the FoxO signaling pathway, cell growth regulation, inflammatory response regulation, and regulation of autophagy pathways. Our study is the first to perform integrated mRNA-seq and miRNA-seq analyses in the thin adhesive endometrium and the control adjacent normal endometrial cells. This study provides new insights into the molecular mechanisms underlying the formation of thin endometrium.

Keywords: thin endometrium, transcriptome analysis, miRNA, mRNA, regulatory

OPEN ACCESS

Edited by:
Maizie Xin Zhou,
Vanderbilt University, United States

Reviewed by:
Zhenjian Zhuo,
Guangzhou Medical University, China
Cuncong Zhong,
University of Kansas, United States

***Correspondence:**
Bo Xu
bioxubo@mail.ustc.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *30 July 2020*
Accepted: *28 January 2021*
Published: *16 March 2021*

Citation:
Zong L, Zheng S, Meng Y, Tang W, Li D, Wang Z, Tong X and Xu B (2021) Integrated Transcriptomic Analysis of the miRNA–mRNA Interaction Network in Thin Endometrium. Front. Genet. 12:589408. doi: 10.3389/fgene.2021.589408

INTRODUCTION

The endometrium is an indispensable factor for implantation and pregnancy, and an increase in endometrial thickness promotes an increased pregnancy rate. An endometrial thickness of <7 mm is usually regarded as sub-optimal for embryo transfer and results in a decreased probability of pregnancy (Shufaro et al., 2008). For patients with Asherman's syndrome (AS), repetitive curettage or invasive endometritis disrupts endometrial regeneration, thus resulting in a fibrotic and thin endometrium (TE) (Azizi et al., 2018). Patients suffering from thin or fibrotic endometrium are more susceptible to abnormal menstruation and, particularly, fertility impairments, such as a decreased pregnancy rate, unfavorable pregnancy outcomes, or recurrent pregnancy loss (Du et al., 2020). The current AS treatments aim to increase endometrial regeneration with low-dose aspirin, exogenous estrogen, vitamin E, vaginal sildenafil citrate, cytokines, and colony-stimulating

factors (CSFs). Nonetheless, these treatments are unable to attain a satisfactory clinical response in many patients with TE (Azizi et al., 2018). The definite etiology and physiopathology of thin endometrium remain largely unclear at present. Therefore, studies aiming to explore the related molecular mechanism of TE are urgently needed to guide disease therapy in the future.

A transcriptomic analysis is essential for understanding the occurrence and pathogenic mechanism of thin endometrium. Only one existing study has reported the global transcriptomic abnormalities in thin endometrium at the mid-luteal phase (Maekawa et al., 2017). The study compared the transcriptomic profiles between three patients and three normal subjects using the Gene Chip Human Genome U133 Plus 2.0 Array platform. Finally, 318 genes were upregulated in the thin endometrium, while 322 genes were downregulated. According to that study, implantation failure induced by thin endometrium might be related to the abnormal activation of the inflammatory environment, together with an abnormally reduced oxidative stress (OS) response. Nonetheless, researchers have not clearly elucidated the underlying mechanism of endometrial regeneration dysfunction in patients with thin endometrium. Additionally, more studies are needed to comprehensively characterize how thin endometrium affects the transcriptomic profiles.

MicroRNAs (miRNAs) are non-protein-coding RNA molecules with short (20–25) nucleotides. miRNAs bind to target mRNAs for transcription and translation regulation, including mRNA degradation, cleavage, or translational repression (Shukla et al., 2011; Li et al., 2019). miRNAs have been deemed to participate in the regulation of various cellular processes, including cellular proliferation, differentiation, apoptosis, and angiogenesis (Laurent, 2008; Nicoli et al., 2010; Hong et al., 2019). Recently, more and more miRNAs were found to be associated with endometrial receptivity (Altmae et al., 2013), endometrial stromal cell differentiation (Qian et al., 2009), embryo development (Laurent, 2008), and implantation (Paul et al., 2019). The expression of miR-27a-3p and miR-124-3p was downregulated in the endometrium of chronic endometritis (Di Pietro et al., 2018). The expression of hsa-miR-449a, hsa-miR-3135b, and hsa-miR-345-3p could promote endometrium receptivity in preparation for *in vitro* fertilization and embryo transplantation (Mu et al., 2020). miR-30 and miR-200 family members have been repeatedly recognized as important miRNAs in the regulation of endometrial receptivity (Rekker et al., 2018). Aberrant miR-200 expression may negatively regulate endometrial development and decidualization (Jimenez et al., 2016) and plays an important role in regulating normal endometrial development and disorders such as endometriosis and endometrial cancer (Panda et al., 2012). However, few studies have investigated the effect of miRNA on thin endometrium. The dysfunction of endometrium cells in TE and how miRNAs regulate the pathogenesis of TE remain to be elucidated.

Our article aimed to identify the miRNA–mRNA networks and molecular pathways in women experiencing intrauterine adhesion (IUA) and to provide additional insights into the underlying transcriptomic mechanisms by performing RNA-Seq. The differentially expressed miRNA–mRNA regulatory axis along with the gene pathway–function network interactions in thin endometrium was constructed. Our findings supply a basis to better investigate the biological mechanisms of thin endometrium and facilitate the formulation of molecular targeted treatments for thin endometrium.

MATERIALS AND METHODS

Tissue Sample Collection

Eight females aged 20–40 years old, with a history of severe IUAs (Grade III–V) as diagnosed by hysteroscopy at the Reproductive Medicine Center of The First Affiliated Hospital of the University of Science and Technology of China, were enrolled in the study. The severity of IUAs was determined according to the American Fertility Society classification system (1988 version) (1988). Scores of 9–12 represented severe adhesions. The thickness of the endometrium was determined through vaginal ultrasound (at mid-luteal phase) as the maximum distance between endometrial interfaces, and the endometrial thickness in all patients was <7 mm. The sample information is described in **Table 1**. The endometrial tissue from the IUA (TE group) and adjacent normal endometrium tissues (AJ-CN group) from eight patients with severe IUAs (Grade III–V) were analyzed in the present study. This study was approved and monitored by the Human Research Ethics Committee of the First Affiliated Hospital of the University of Science and Technology of China. Each patient was required to provide a written informed consent prior to participation in this study. Endometrial tissues were sampled at mid-luteal phase during the menstrual cycle. Afterwards, the collected endometrial tissue samples were rinsed with saline to remove blood and then stored in liquid nitrogen at -80°C until subsequent RNA isolation.

RNA Isolation and Library Construction

Total tissue RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, USA), pooled equally, and reverse-transcribed into cDNAs using the QuantiTect Reverse Transcription Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's

TABLE 1 | Information about the sample of patients with thin endometrium analyzed in our study.

Group	Age (years)	Endometrial thickness (mm)	Sample date (from LMP)	History of gestation	Score of IUA
TE 1	31	6.5	20	G3P1	9
TE 2	28	5.6	21	G3P1	10
TE 3	35	4.8	22	G2P1	10
TE 4	32	5.2	21	G2P1	11
TE 5	27	6.4	21	G2P1	10
TE 6	32	5.8	20	G4P1	12
TE 7	37	6.5	22	G3P1	10
TE 8	30	5	21	G2P1	10

LMP, last menstrual period; TE, thin endometrium; history of gestation, G for gestation, P for parturition.

specific instructions. The quantity and quality of the extracted RNA were measured using Nanodrop (Thermo Scientific). The cDNA library was constructed using KAPA Stranded RNA-Seq Library Preparation Kit (Illumina) following the manufacturer's protocol. The synthetic cDNAs were end-repaired by polymerase and ligated with "A-tailing" base adaptors. Suitable fragments were selected for polymerase chain reaction (PCR) amplification to construct the final cDNA library. The final double-stranded cDNA samples were verified with Agilent 2100 Bioanalyzer (Agilent Technologies). Sequencing was performed on an Illumina HiSeq 4000 sequencing platform with 150-bp paired-end sequencing.

Then, the combined RNA samples were separated using 15% (w/v) denaturing polyacrylamide gel electrophoresis.

TABLE 2 | Gene Ontology analysis of the 1,093 differentially expressed genes between thin endometrium and adjacent normal endometrium.

Term		Count	*P*-value	Genes
GO:0007155	Cell adhesion	62	1.92E-10	NRP2, MPZL3, CXCL12, PRKX, HMCN2, AZGP1, WISP2, WISP1, CTGF, COL12A1, AFDN, CEACAM1, EGFL6, ADGRE5, NECTIN4, GRHL2, CTNNA2, JUP, NCAM1, PGM5, CD36, LAMC3, VCAN, LAMC2, TGFB1I1, MFAP4, ADAM12, AOC3, OLFM4, ITGA11, PCDHGC3, ALCAM, LAMB3, SORBS1, ITGB8, COMP, MSLN, THBS1, ENTPD1, DPT, SPP1, HAPLN2, SELP, LPP, MCAM, EMILIN2, TINAGL1, COL4A6, LAMA2, ITGA9, CDH13, NME1-NME2, CDH16, DSG2, FREM2, CDON, ITGA7, NLGN4X, DSC2, PDZD2, OMG, MUC16
GO:0030198	Extracellular matrix organization	31	4.19E-07	MPZL3, ELF3, PDGFA, NPNT, ITGA11, CDH1, SOX9, SMOC2, LAMB3, HPSE, ITGB8, COMP, THBS1, SPP1, RXFP1, EGFL6, CCDC80, SPINT1, OLFML2A, COL4A6, COL4A5, LAMA2, ITGA9, BGN, LAMC3, KAZALD1, FBLN5, ITGA7, LAMC2, VCAN, MFAP5
GO:0090190	Positive regulation of branching involved in ureteric bud morphogenesis	9	3.85E-06	NOG, AGTR2, HOXB7, PAX8, SIX4, PAX2, GREM1, SOX9, WNT2B
GO:0045926	Negative regulation of growth	8	4.46E-05	HIF1A, MT1M, MT2A, MT1H, MT1X, MT1G, MT1F, IGFBP5
GO:0001525	Angiogenesis	29	4.63E-05	NRP2, SAT1, CAV1, HTATIP2, PDGFA, CSPG4, FGF10, PRKX, NOV, TYMP, OVOL2, UNC5B, CTGF, XBP1, HS6ST1, SOX17, RAMP1, CEACAM1, SCG2, KLF5, MCAM, ECM1, HOXB3, HIF1A, CLIC4, ID1, PROK1, HIF3A, RBPJ
GO:0086073	Bundle of His cell-Purkinje myocyte adhesion involved in cell communication	5	1.30E-04	JUP, DSG2, PKP2, DSC2, DSP
GO:0034329	Cell junction assembly	6	2.99E-04	LIMS2, FERMT2, ILK, FLNC, GRHL2, FLNA
GO:0051145	Smooth muscle cell differentiation	6	4.63E-04	MEF2C, WNT4, MYOCD, GATA6, HEY2, FGF10
GO:0030336	Negative regulation of cell migration	15	7.39E-04	PTPRJ, NOG, EPPK1, PLXNB3, DPYSL3, SLC9A3R1, TPM1, SLIT2, WNT4, PKP2, CLIC4, SFRP2, RRAS, STC1, IGFBP5
GO:0001558	Regulation of cell growth	13	0.001477124	NOV, PRKCQ, SGK1, WISP2, WISP1, CTGF, KAZALD1, FBLN5, FOXM1, RASGRP2, IGFBP6, CEACAM1, IGFBP5
GO:0055015	Ventricular cardiac muscle cell development	5	0.001518037	CCNB1, CDK1, HEY2, LMNA, FHL2
GO:0090027	Negative regulation of monocyte chemotaxis	4	0.001571138	NOV, MINOS1-NBL1, GREM1, SLIT2
GO:0070830	Bicellular tight junction assembly	8	0.001582947	OCLN, ACTN4, MARVELD2, CLDN3, MARVELD3, CRB3, ECT2, GRHL2
GO:0050679	Positive regulation of epithelial cell proliferation	11	0.001599078	NOG, OSR1, NME1-NME2, ID1, DLX6, DLX5, FGF10, ESRP2, PAX2, SOX9, IHH
GO:0002576	Platelet degranulation	15	0.001650782	SELP, ACTN4, PDGFA, ACTN1, ECM1, TIMP3, FLNA, CTSW, ORM1, CD36, LEFTY2, SERPINA3, SERPINA1, THBS1, ORM2
GO:0016055	Wnt signaling pathway	22	0.001694244	NKD1, SPIN1, FERMT2, TLE2, FRZB, SLC9A3R1, APCDD1, WNT2B, CCNE1, RNF43, DKK3, WNT4, WISP1, RSPO1, CPE, DACT3, SFRP2, RSPO3, KREMEN1, RNF138, TGFB1I1, LRP4
GO:0045216	Cell–cell junction organization	6	0.002445448	OCLN, LIMS2, MARVELD2, MARVELD3, NLGN4X, CXADR
GO:0030308	Cell growth negative regulation	16	0.002889171	PTPRJ, CRYAB, FHL1, FBP1, OSGIN2, FRZB, GREM1, SLIT2, RERG, NOV, AGTR2, MSX1, SFRP2, DACT3, CDKN2AIP, SOX17
GO:0030514	Negative regulation of BMP signal transduction pathway	9	0.003034824	RBPMS2, CAV1, NOG, CHRDL1, DKK1, MINOS1-NBL1, SFRP2, GREM1, TOB1

Subsequently, miRNA fragments with a size of ∼18–28 nt were separated by gel extraction, followed by RNA purification. The total RNA of each sample was used to prepare the miRNA sequencing library, which included the following steps: (1) 3′-adaptor ligation, (2) 5′-adaptor ligation, (3) cDNA synthesis, (4) PCR amplification, and (5) size selection of ∼135–155-bp PCR-amplified fragments (corresponding to ∼15–35 nt small RNAs). Libraries were quantified and validated with Agilent 2100 Bioanalyzer (Agilent Technologies). Thereafter, the small RNA library was sequenced using Illumina Hiseq 4000 (Illumina, San Diego, CA, USA), with a configuration of 50 cycles single reads according to the manufacturer's recommendations. All sequencing procedures were performed by Kang Chen Bio-tech (Shanghai, China).

mRNA Sequencing and Data Analysis

Raw data were pre-processed using Solexa CHASTITY and Cutadapt to remove adaptor sequences, ribosomal RNA, low-quality reads, and other contaminants that may interfere with assembly. The criteria for this filtering procedure were set as follows: (1) RNA 5′ and 3′ adapters were removed, respectively, (2) bases with a phred quality score below 20 were clipped from both ends of reads, (3) after low-quality bases were trimmed, reads containing over two "N" were discarded, (4) reads with a length shorter than 75 nt were discarded; and (5) the parameters for BWA v0.5.724 were set as recommended according to Fastq_clean instructions. Then, the sequence quality was examined using FastQC v0.11.7. Afterwards, Hisat2 was utilized to align those trimmed reads to the reference genome. StringTie (version 1.2.3) was used to reconstruct the transcriptome. Fragments per kilobase per million (FPKM) values of genes were normalized with Ballgown using the default parameters. FPKM ≥0.5 (Cuffquant) was considered as statistically significant for the next DEG analysis. RNA sequencing data were deposited into the Gene Expression Omnibus (GEO accession number GSE160635).

miRNA Sequencing and Data Analysis

The miRNA sequencing data from TE group and AJ-CN group endometrium cells were analyzed by our previously published tool, DeAnnIso (Zhang et al., 2016). Briefly, after sequencing, Bowtie was used for mapping reads into the reference genome. The aligned reads had no more than "N" mismatches (0–3, default is 2) in the first "L" bases (≥5, default is 10) of the left end. Thereafter, those precursor sequence-matched reads were aligned to the pooled pre-miRNA databases (known pre-miRNAs in miRBase v21) using the BLAST. The default E-value was set to 0.01 for BLAST. All the detected isomiRs were aligned with their canonical miRNAs, the numbers of mapped reads that were defined as the raw expression levels of that miRNA. To correct for the difference in read counts between samples, the read counts were scaled to reads per million (RPM). Small RNA sequencing data were deposited into the Gene Expression Omnibus (GEO accession number GSE108966).

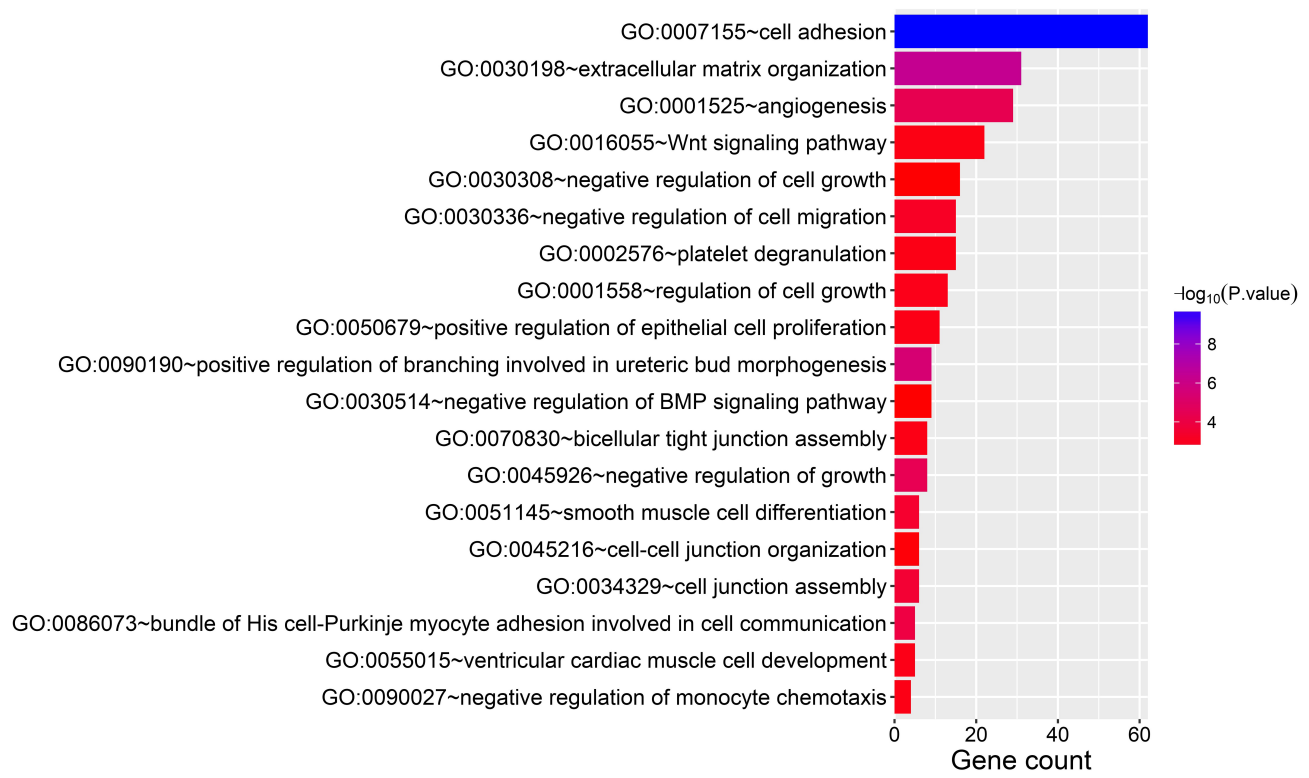


FIGURE 1 | Gene Ontology analysis of 1,093 differentially expressed genes between thin endometrium and adjacent normal endometrium.

Differential Expression Analysis

After excluding the transcripts with a low count, genes with an FPKM or RPM ≥5 in at least one sample were used for the analysis. Fold change (FC) and *P*-value for Fisher's exact test was calculated and used when comparing the differentially expressed mRNAs (DEGs) and miRNAs (DEMs) between the two groups. The $\log_2$FC derived from the comparisons of the FPKM or RPM values of the TE group with the AJ-CN group is depicted ($|\text{Log}_2\text{FC}| \geq 2$) and $P < 0.05$ were selected as the cutoff criteria to identify significant DEMs and DEGs. Additionally, TargetScan (Garcia et al., 2011) and miRDB (Wang and El Naqa, 2008) were used to predict mRNAs targeted by DEMs.

Functional Annotations

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed using the online analysis tool of Annotation Visualization and Integrated Discovery (https://david.ncifcrf.gov/). The *P*-value for Fisher's exact test was calculated as a result of enrichment degree. GO term enrichment of biological processes or KEGG pathway annotations with a *P*-value cutoff of 0.05 were identified as an important term in this study.

Construction of the Protein–Protein Interaction Network

The Search Tool for the Retrieval of Interacting Genes (STRING) database (http://www.string-db.org/) was used to construct the PPI network. The obtained interactions included both the known and the estimated interactions. A requisite confidence value (pooled score >0.4) was used as the threshold. In addition, Cytoscape v3.7.1 was utilized to visualize the PPI network, and CytoHubba functions were employed to identify the hub genes. Genes with Gene significance >0.2, module membership >0.8, and $P \leq 0.05$ were defined as hub genes.

TABLE 3 | Kyoto Encyclopedia of Genes and Genomes pathway analysis of 1,093 differentially expressed genes between thin endometrium and adjacent normal endometrium.

Term		Count	*P*-value	Genes
hsa04270	Vascular smooth muscle contraction	22	2.32E-06	KCNMA1, ACTA2, CALD1, MRVI1, PRKG1, ITPR3, KCNMB1, ITPR1, PRKCB, MYL9, ITPR2, PRKCQ, ACTG2, PLA2G4A, PLA2G2A, AVPR1A, PLA2G4F, CACNA1C, RAMP1, MYLK, ADRA1D, PPP1R14A
hsa04512	Extracellular matrix–receptor interaction	16	1.06E-04	ITGA11, COL4A6, COL4A5, HMMR, LAMA2, ITGA9, LAMB3, SDC1, CD36, ITGB8, LAMC3, COMP, ITGA7, LAMC2, THBS1, SPP1
hsa05410	Hypertrophic cardiomyopathy	15	1.14E-04	ITGA9, ACE, ACTC1, DES, TNNC1, ITGB8, DMD, ITGA7, ITGA11, LMNA, CACNB2, TPM2, CACNA1C, TPM1, SGCA
hsa04510	Focal adhesion	26	2.73E-04	CAV2, CAV1, ACTN4, PDGFA, MET, ITGA11, ACTN1, BIRC3, FLNC, COL4A6, FLNA, COL4A5, PRKCB, MYL9, LAMA2, ITGA9, LAMB3, LAMC3, ITGB8, COMP, ILK, ITGA7, LAMC2, THBS1, MYLK, SPP1
hsa04530	Tight junction	14	0.001241795	CLDN7, OCLN, CLDN4, ACTN4, CLDN3, CRB3, ACTN1, LLGL2, MYL9, CGN, MYH11, AFDN, MYH14, TJP3
hsa04514	Cell adhesion molecules	18	0.002948225	SELP, CLDN7, OCLN, CLDN4, CADM1, CLDN3, VTCN1, CD276, CDH1, HLA-DMB, ALCAM, NCAM1, ITGA9, SDC1, ITGB8, NLGN4X, VCAN, HLA-DOB
hsa04750	Inflammatory mediator regulation of transient receptor potential channels	14	0.003685506	IL1R1, CYP2J2, CAMK2G, F2RL1, ITPR3, ITPR1, PRKCB, ITPR2, PRKCQ, PLA2G4A, MAPK13, PLA2G4F, CAMK2A, MAP2K6
hsa04610	Complement and coagulation cascades	11	0.005535681	C7, CD55, MASP1, C4A, C4B, C3, F3, SERPINA5, SERPINA1, CFI, C4BPA
hsa04670	Leukocyte transendothelial migration	14	0.013940458	CLDN7, OCLN, CLDN4, ACTN4, CLDN3, ACTN1, CXCL12, PRKCB, MYL9, CTNNA2, EZR, CXCR4, MAPK13, AFDN
hsa04730	Long-term depression	9	0.020179859	GNAZ, PLA2G4A, GRIA2, PLA2G4F, PRKG1, ITPR3, ITPR1, PRKCB, ITPR2
hsa04911	Insulin secretion	11	0.022698704	KCNMA1, CAMK2G, ADCYAP1R1, SLC2A1, KCNN2, ITPR3, CACNA1C, CAMK2A, SNAP25, KCNMB1, PRKCB
hsa04020	Calcium signal transduction pathway	18	0.02706823	PTGER3, TNNC1, ERBB3, CAMK2G, ITPKB, PTGFR, ITPR3, ITPR1, PRKCB, ITPR2, GNAL, PLN, AVPR1A, CACNA1H, CACNA1C, CAMK2A, MYLK, ADRA1D
hsa00512	Mucin type O-glycan biosynthesis	6	0.029788794	GALNT3, GCNT3, GALNT4, GALNT18, GALNT12, ST6GALNAC1
hsa05150	*Staphylococcus aureus* infection	8	0.033164245	SELP, MASP1, C4A, C4B, C3, CFI, HLA-DMB, HLA-DOB
hsa04720	Long-term potentiation	9	0.033723842	GRIA2, RPS6KA1, CAMK2G, ITPR3, CACNA1C, CAMK2A, ITPR1, PRKCB, ITPR2
hsa04912	GnRH signal transduction pathway	11	0.034509876	PLA2G4A, MAPK13, CAMK2G, PLA2G4F, ITPR3, CACNA1C, CAMK2A, MAP2K6, ITPR1, PRKCB, ITPR2
hsa04115	p53 signal transduction pathway	9	0.036464983	CCNB1, CCNE1, CDK1, CCNB2, MDM4, SFN, THBS1, PERP, GTSE1
hsa04520	Adherens junction	9	0.04891771	PTPRJ, ACTN4, SORBS1, MET, ACTN1, CDH1, AFDN, NECTIN4, CTNNA2

Construction of the DEM–DEG Regulatory Network

TargetScan (http://www.targetscan.org/) and miRDB (http://www.mirdb.org/) were utilized to preliminarily predict DEM target genes. The co-predicted targets were used for further GO and KEGG pathway enrichment analyses. The genes shared between DEM targets and DEGs were used to analyze the miRNA–mRNA pairs, which were maintained to construct the DEM–DEG regulatory network with Cytoscape. Differentially expressed target genes were chosen for GO and KEGG pathway analyses to investigate the miRNA–mRNA regulatory networks in TE.

RESULTS

Genome-Wide Patterns of the mRNA Transcriptomic Landscape

Using Illumina Hiseq 4000, 18,354,811 original RNA reads were obtained from the thin endometrial cells of patients with IUA, and 21,755,164 reads were obtained from adjacent normal endometrial cells. After removing adaptor sequences and low-quality reads, 18,288,140 (thin endometrial cells from patients with IUA) and 21,745,564 (adjacent normal endometrial cells) clean reads remained. Then, the genes were normalized to FPKM, and 15,561 genes were expressed in endometrial tissues from those eight women.

In the thin adhesive endometrial tissue of the TE group, 374 genes were upregulated, while 719 genes were downregulated compared to the control adjacent normal endometrial cells (**Supplementary Tables 1, 2**). The GO analysis of 1,093 DEGs identified many genes that were significantly enriched in the cell adhesion process (GO: 0007155, $P = 1.92E\text{-}10$), negative regulation of growth (GO: 0045926, $P = 4.46E\text{-}05$), angiogenesis (GO: 0001525, $P = 4.63E\text{-}05$), cell junction assembly (GO: 0034329, $P = 2.99E\text{-}04$), negative regulation of cell migration (GO: 0030336, $P = 7.39E\text{-}04$), the Wnt signaling pathway (GO: 0016055, $P = 0.0017$), and negative regulation of the BMP signaling pathway (GO: 0030514, $P = 0.003$) (**Table 2** and **Figure 1**). A blockade angiogenesis was considered as the

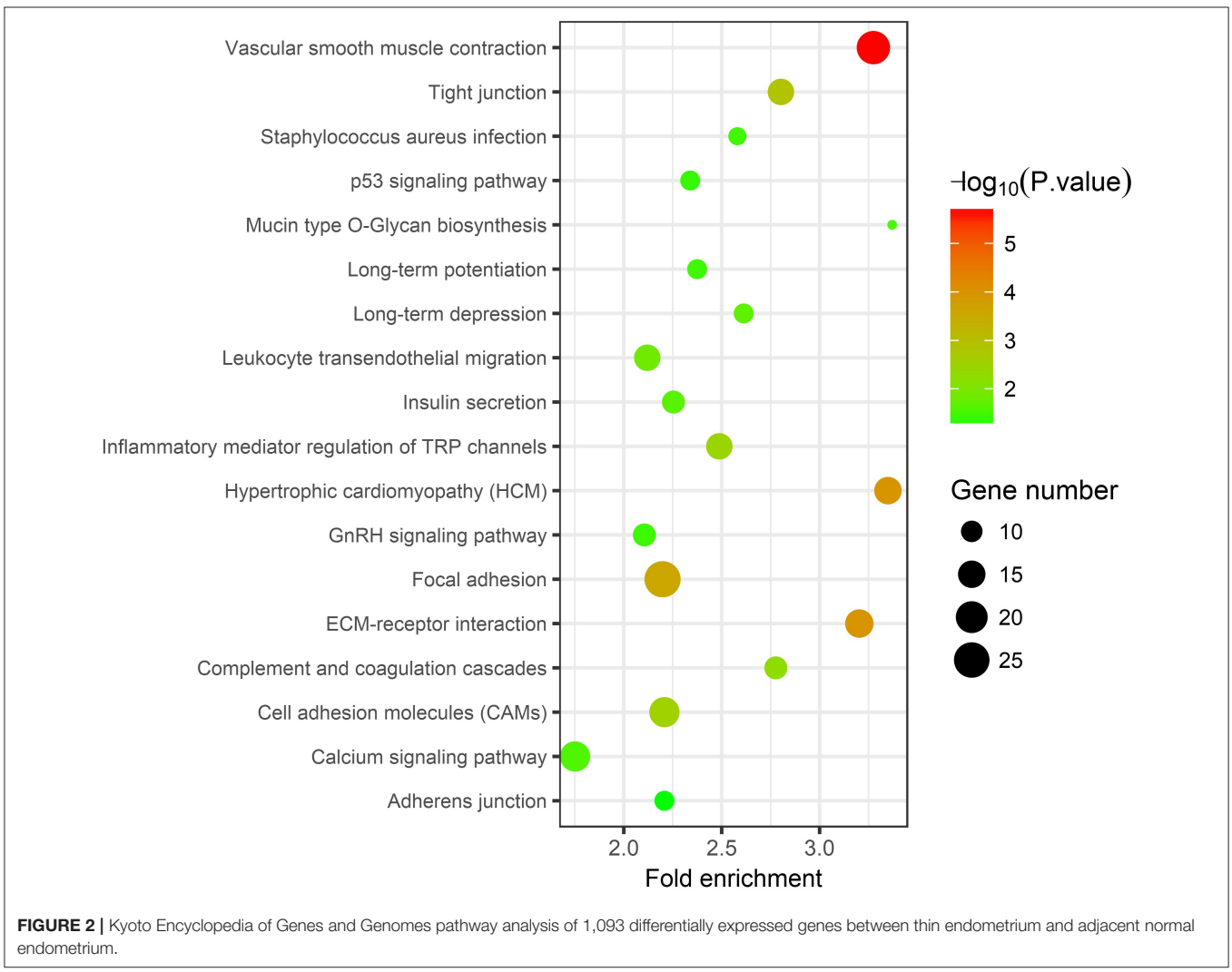


FIGURE 2 | Kyoto Encyclopedia of Genes and Genomes pathway analysis of 1,093 differentially expressed genes between thin endometrium and adjacent normal endometrium.

main pathological change in the scarred thin endometrium (Jiang et al., 2019). Moreover, this study identified several DEGs-related signaling pathways by performing KEGG pathway enrichment analysis, including the vascular smooth muscle contraction pathway, extracellular matrix–receptor interaction, focal adhesion, tight junction, cell adhesion molecules, calcium signal transduction pathway, p53 signal transduction pathway, and adherens junction pathway (**Table 3** and **Figure 2**). The 1,093 DEGs were also compared with the primary associated changes identified in the transcriptome of the thin endometrium (Maekawa et al., 2017), and nine commonly upregulated genes (PDLIM3, FABP3, HIF3A, FILIP1, DPP6, MYOCD, PRKCB, ALDH1B1, and TRNP1) and 65 commonly downregulated genes were identified (**Supplementary Table 3**). The expression of MYOCD (myocardin), a cardiac-specific co-activator of serum response factor, was upregulated in thin endometrium, while ADAM12 (a disintegrin and metalloproteinase 12) expression was decreased in thin endometrium, and these genes are associated with the fibrosis process (Li et al., 2018; Mittal et al., 2019; Nakamura et al., 2019).

Genome-Wide Patterns of the miRNA Transcriptomic Landscape

First, clean reads were mapped to the human genome, and then those mapped reads were further matched to miRbase (V22). Notably, 7,004,583 reads (TE sample) and 5,717,874 reads (AJ-CN sample) were aligned to human pre-miRNAs. A total of 1,244 known miRNAs were altogether identified in our endometrial samples. According to the results of the miRNA-seq analysis, 72 known miRNAs were deemed to be DEMs between the thin adhesive endometrium of the IUA group and the control adjacent normal endometrial cells. Among these DEMs, five miRNAs were upregulated and 67 were downregulated compared with the control adjacent normal endometrial cells (**Supplementary Table 4**). The five upregulated and top 10 downregulated DEMs are shown in **Table 4**.

TargetScan and miRDB were used to characterize the putative target mRNAs of the 72 candidate DEMs in thin endometrium and to better illustrate the functions of DEMs. TargetScan and miRDB were employed to identify 812 common candidate target genes for the 15 DEMs (**Supplementary Table 5**). Then, GO and KEGG analyses were performed for the 812 target genes. GO enrichment analyses suggested that the target genes of multiple DEMs were associated with the regulation of angiogenesis, MAPK activation, negative regulation of cell migration, negative regulation of stress fiber assembly, positive regulation of epithelial cell proliferation, regulation of the canonical Wnt signaling pathway, and positive regulation of cell proliferation (**Table 5** and **Supplementary Figure 1**). The KEGG pathways in which the DEM targeted genes are involved were discovered, which included the Ras signal transduction pathway, Hippo signal transduction pathway, MAPK signal

TABLE 4 | The five upregulated and top 10 downregulated differentially expressed miRNAs in thin endometrium.

MATURE-ID	PRE-ID	MATURE-SEQ	Log2 (fold change)
hsa-miR-1-3p	hsa-mir-1-2	UGGAAUGUAAAGAAGUAUGUAU	4.369546
hsa-miR-133a-3p	hsa-mir-133a-1	UUUGGUCCCCUUCAACCAGCUG	3.602664
hsa-miR-143-3p	hsa-mir-143	UGAGAUGAAGCACUGUAGCUC	2.285041
hsa-miR-133b	hsa-mir-133b	UUUGGUCCCCUUCAACCAGCUA	2.142958
hsa-miR-145-5p	hsa-mir-145	GUCCAGUUUUCCCAGGAAUCCCU	1.896256
hsa-miR-34c-5p	hsa-mir-34c	AGGCAGUGUAGUUAGCUGAUUGC	−6.13482
hsa-miR-200a-3p	hsa-mir-200a	UAACACUGUCUGGUAACGAUGU	−5.88598
hsa-miR-200c-3p	hsa-mir-200c	UAAUACUGCCGGGUAAUGAUGGA	−5.57759
hsa-miR-200b-3p	hsa-mir-200b	UAAUACUGCCUGGUAAUGAUGA	−5.57684
hsa-miR-375	hsa-mir-375	UUUGUUCGUUCGGCUCGCGUGA	−5.49063
hsa-miR-449c-5p	hsa-mir-449c	UAGGCAGUGUAUUGCUAGCGGCUGU	−5.24102
hsa-miR-429	hsa-mir-429	UAAUACUGUCUGGUAAAACCGU	−5.17173
hsa-miR-141-3p	hsa-mir-141	UAACACUGUCUGGUAAAGAUGG	−5
hsa-miR-449a	hsa-mir-449a	UGGCAGUGUAUUGUUAGCUGGU	−4.92875
hsa-miR-182-5p	hsa-mir-182	UUUGGCAAUGGUAGAACUCACACU	−4.89552
hsa-miR-34c-5p	hsa-mir-34c	AGGCAGUGUAGUUAGCUGAUUGC	−6.13482
hsa-miR-200a-3p	hsa-mir-200a	UAACACUGUCUGGUAACGAUGU	−5.88598
hsa-miR-200c-3p	hsa-mir-200c	UAAUACUGCCGGGUAAUGAUGGA	−5.57759
hsa-miR-200b-3p	hsa-mir-200b	UAAUACUGCCUGGUAAUGAUGA	−5.57684
hsa-miR-375	hsa-mir-375	UUUGUUCGUUCGGCUCGCGUGA	−5.49063
hsa-miR-449c-5p	hsa-mir-449c	UAGGCAGUGUAUUGCUAGCGGCUGU	−5.24102
hsa-miR-429	hsa-mir-429	UAAUACUGUCUGGUAAAACCGU	−5.17173
hsa-miR-141-3p	hsa-mir-141	UAACACUGUCUGGUAAAGAUGG	−5
hsa-miR-449a	hsa-mir-449a	UGGCAGUGUAUUGUUAGCUGGU	−4.92875

TABLE 5 | Gene Oncology analysis of the identified targets of differentially expressed miRNAs between thin endometrium and adjacent normal endometrium.

Term		Count	Fold enrichment	*P*-value	Targeted genes
GO:0007264	Small GTPase-mediated signal transduction	22	2.37	3.79E-04	RALGPS2, RAB3C, RAP2C, RAP1GDS1, RASGEF1B, RHOQ, ARF6, PLCE1, RAB43, ARF3, ARF4, ARHGAP1, YWHAQ, RAB5A, RAB14, RRAS, RHEB, RAB6B, RAP1B, RAB38, RIT2, RAB21
GO:0045944	Positive regulation of transcription from RNA polymerase II promoter	43	1.70	8.26E-04	FOSL2, HELZ2, LMO4, EDN1, RHOQ, INO80, EGLN1, PAX3, ZEB1, ASH2L, PAX7, RARB, PPP3CA, MYC, GABPB2, SATB2, RARG, KLF12, EPAS1, MET, EOMES, IGF1, DLL1, DDX5, NCL, TET1, RBMX, BCL2L12, RNF222, FOXP1, PPARGC1B, MYCN, ASCL1, RPS6KA4, EBF3, ETS1, SP3, JUN, ARF4, ZFPM2, TFAP2D, NR5A2, BMPR1B
GO:0045765	Regulation of angiogenesis	6	7.24	1.02E-03	ETS1, EFNA1, EGLN1, EMP2, VASH2, VASH1
GO:0000045	Autophagosome assembly	8	4.69	1.28E-03	GABARAPL2, GABARAPL1, ATG4B, MAP1LC3B, TRAPPC8, RB1CC1, WIPI2, TP53INP2
GO:0000187	Activation of MAPK activity	8	4.00	3.32E-03	MAP3K7, PLCE1, NTF3, EFNA1, IGF1, LPAR1, THBS1, FRS2
GO:0030336	Negative regulation of cell migration	10	3.20	3.64E-03	DLC1, RECK, TMEFF2, PTPRK, RAP2C, CLIC4, SULF1, RRAS, SRGAP1, SRGAP2
GO:0051592	Response to calcium ion	6	5.13	5.29E-03	CAV1, SLC25A13, ALG2, AHCYL1, PPP3CA, THBS1
GO:0000422	Mitophagy	6	4.39	1.04E-02	GABARAPL2, GABARAPL1, ATG4B, MAP1LC3B, RB1CC1, WIPI2
GO:0051497	Negative regulation of stress fiber assembly	4	6.31	2.28E-02	DLC1, TMEFF2, ARHGAP6, PPFIA1
GO:0016601	Rac protein signal transduction	4	6.31	2.28E-02	EPS8, WASF1, ELMO1, NCKAP1
GO:0050679	Positive regulation of epithelial cell proliferation	6	3.51	2.61E-02	WDR48, NOTCH1, FGF9, IGF1, MYC, FOXP1
GO:0008277	Modulation of G-protein coupled receptor protein signal transduction pathway	4	5.86	2.80E-02	PLCE1, GPR158, KCTD16, USP33
GO:0060828	Modulation of canonical Wnt signal transduction pathway	4	5.47	3.37E-02	AMER1, CCNY, CTNND2, CDK14
GO:0035556	Intracellular signal transduction	22	1.60	3.52E-02	ARHGEF3, SGK1, NUAK1, PREX1, DSTYK, SPSB4, ITSN1, PLCL2, RPS6KA4, SNRK, DGKE, PPP1R1C, STAC, GUCY1A3, DGKZ, RGS7, STK39, DCX, STK38L, PAG1, NET1, SHC4
GO:0007205	Protein kinase C-activating G-protein coupled receptor signaling pathway	4	5.13	4.01E-02	GRM5, DGKE, EDN1, DGKZ
GO:0008284	Positive regulation of cell proliferation	18	1.68	4.05E-02	RARG, PDCD10, NTF3, PTH1R, IGF1, DLL1, PTGFR, TET1, TGFB2, CRKL, KRAS, ASH2L, HBEGF, RARB, MAB21L1, EMP2, CSF1R, SHC4

transduction pathway, PI3K–Akt signal transduction pathway, gap junction, p53 signaling pathway, Wnt signal transduction pathway, and ErbB signal transduction pathway (**Figure 3** and **Table 6**). The PI3K/Akt signal transduction pathway is suggested to participate in endometrial regeneration induced by granulocyte macrophage–CSF therapy (Liu et al., 2020).

DEG–DEM Regulatory Network and Functional Assessment

For the establishment of the DEG-DEM regulatory network, 53 (21 upregulated genes and 32 downregulated genes) overlapping genes were discovered by comparing the target genes of DEMs (five were upregulated and 10 were downregulated) with DEGs, and they were deemed as consistently expressed genes (CEGs) (**Figure 4**). The STRING database was used to construct the PPI network using the CEG list. As shown in **Figure 5**, CAV1, MET, MAL2, has-mir-138, ARHGAP6, CLIC4, RRAS, AGFG1, has-mir-200, and has-mir-429 were the top 10 hub genes that interacted with the maximum number of nodes. Additionally, the gene pathway–function interactions were analyzed, and the identified hub genes showed significant enrichment in negative regulation of cell growth and inflammatory response regulation. For a better assessment of how this miRNA–mRNA regulatory network affected thin endometrium, a KEGG pathway analysis of CEGs was performed. The miRNA-mediated gene regulatory network in thin endometrium plays important roles in the regulation of the FoxO signaling pathway and the regulation of autophagy (**Table 7**).

DISCUSSION

IUA, which is characterized by endometrial fibrosis and thin endometrium, was always regarded as a major cause of female infertility and a major challenge to clinical therapy. Even

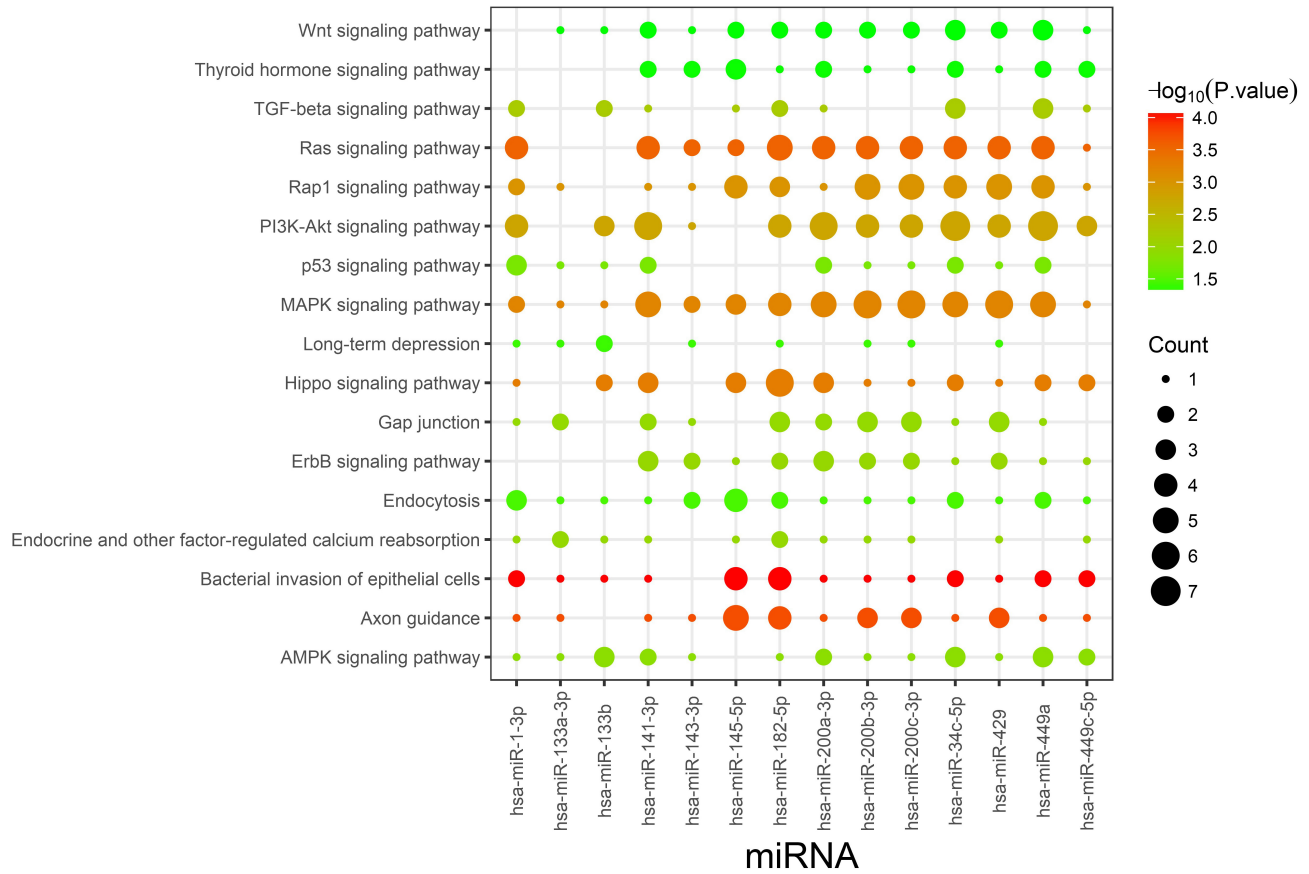


FIGURE 3 | Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis of the identified targets of differentially expressed miRNAs. Count indicates the number of predicted target genes.

through a surgical operation combined with hormone treatment, TE with severe endometrial injuries is difficult to restore. The previous transcriptomic microarray analysis discovered 318 upregulated genes and 322 downregulated genes in thin endometrium and revealed the abnormal activation of the inflammatory environment and an abnormal decrease in the OS response in thin endometrium (Maekawa et al., 2017). Current knowledge about the pathogenesis and involvement of miRNA–mRNA networks in thin endometrium is limited. In this study, gene expression patterns of thin endometrium along with the matched control endometrial tissues from women were explored, and we revealed the abnormal activation of the inflammatory environment and an abnormal decrease in the OS response in thin endometrium. To our knowledge, this study is the first to employ self-controlled transcriptomic analysis to investigate the regulatory functions of miRNA–mRNA networks of cells from the mid-secretory thin endometrium and adjacent normal endometrial cells.

As revealed in our results, some genes were abnormally expressed at the time of disease onset, revealing that thin endometrium may have occurred as a type of endometrial disorder due to the abnormal expression of genes within endometrial tissues prior to lesion occurrence. Indeed 1,093 genes were significantly differentially expressed in thin endometrium. A total of 74 DEGs associated with TE in our study were consistent with a previous study performed in thin and control endometrial samples using a microarray (Maekawa et al., 2017), including those that were up-regulated. Furthermore, our DEG functional enrichment analysis also revealed the involvement of angiogenesis and negative regulation of growth and cell migration in thin endometrium. Typically, during each menstrual cycle, angiogenesis promotes new blood vessel formation and is crucial for endometrial regeneration by supplying a vascularized and receptive endometrium for embryo implantation. Previous studies also show that the vascular endothelial growth factor (VEGF) could be a regulator of endometrial angiogenesis. Thus, the differential expression of VEGF and the blockade of angiogenesis in our study could be considered as pathological changes of the scarred thin endometrium (Jiang et al., 2019).

Interestingly, consistent with previous miRNA expression profiles reported for the recurrent implantation failure endometrium (Vilella et al., 2015; Rekker et al., 2018), some upregulated DEMs in TE in our study also belonged

TABLE 6 | Kyoto Encyclopedia of Genes and Genomes pathway analysis of the identified targets of differentially expressed miRNAs between thin endometrium and adjacent normal endometrium.

Term		Count	Fold enrichment	*P*-value	Targeted genes
cfa05100	Bacterial invasion of epithelial cells	13	3.92	9.32E-05	ACTB, CAV1, CLTA, WASF1, MET, CLTC, CD2AP, ELMO1, ACTG1, CTTN, CRKL, GAB1, SHC4
cfa04360	Axon guidance	16	3.08	1.86E-04	GNAI3, EFNA1, MET, NTNG1, L1CAM, EPHA2, SLIT2, SEMA6A, KRAS, CFL2, CFL1, SEMA3A, PPP3CA, RASA1, SRGAP1, SRGAP2
cfa04014	Ras signaling pathway	22	2.41	2.81E-04	FGF9, GRB2, EFNA1, MET, IGF1, ARF6, EPHA2, KDR, PLCE1, KRAS, ETS1, GAB1, PDGFRA, RAB5A, RAPGEF5, RRAS, RAP1B, PRKACB, ABL2, RASA1, CSF1R, SHC4
cfa04390	Hippo signaling pathway	17	2.69	5.28E-04	ACTB, MOB1B, YWHAZ, MPP5, LEF1, SMAD1, TGFB2, AJUBA, ACTG1, YWHAG, CCND2, PPP2CA, PPP2CB, YWHAQ, BMPR1B, MYC, FBXW11
cfa04010	MAPK signaling pathway	23	2.21	6.38E-04	LAMTOR3, NTF3, FGF9, GRB2, MAP2K4, CACNB3, CACNB4, TGFB2, MAP3K7, BDNF, CRKL, KRAS, RPS6KA4, DUSP1, JUN, PDGFRA, RRAS, RAP1B, PRKACB, PPP3CA, MYC, RASA1, DUSP6
cfa04015	Rap1 signaling pathway	20	2.30	9.89E-04	ACTB, GNAI3, FGF9, EFNA1, MET, IGF1, LPAR1, EPHA2, KDR, ACTG1, PLCE1, CRKL, KRAS, GNAQ, PDGFRA, RAPGEF5, RRAS, RAP1B, THBS1, CSF1R
cfa04151	PI3K–Akt signaling pathway	27	1.91	1.71E-03	YWHAZ, PPP2R3A, EFNA1, GRB2, FGF9, LPAR1, FOXO3, CCNE2, KRAS, PPP2CA, PPP2CB, PIK3AP1, THBS1, MYC, CSF1R, SGK1, MET, IGF1, IL6R, EPHA2, KDR, YWHAG, EIF4E, CCND2, YWHAQ, PDGFRA, RHEB
cfa04350	TGF-beta signaling pathway	10	2.94	6.36E-03	E2F5, PPP2CA, PPP2CB, TGIF2, SMAD1, SKP1, THBS1, BMPR1B, MYC, TGFB2
cfa04012	ErbB signaling pathway	10	2.74	1.01E-02	CRKL, KRAS, GRB2, JUN, GAB1, MAP2K4, HBEGF, MYC, ABL2, SHC4
cfa04961	Endocrine and other factor-regulated calcium reabsorption	7	3.71	1.04E-02	CLTA, AP2B1, ATP1B3, GNAQ, PTH1R, PRKACB, CLTC
cfa04540	Gap junction	10	2.71	1.08E-02	GRM5, GNAI3, KRAS, GNAQ, GRB2, PDGFRA, GJA1, GUCY1A3, PRKACB, LPAR1
cfa04152	AMPK signaling pathway	12	2.35	1.27E-02	MAP3K7, PPP2R3A, HNF4A, PFKFB3, PPP2CA, PPP2CB, RAB14, ADIPOR2, IGF1, RHEB, FOXO3, SCD5
cfa04115	p53 signaling pathway	8	2.94	1.81E-02	CCNE2, CCND2, ZMAT3, SHISA5, IGF1, MDM4, THBS1, SESN1
cfa04144	Endocytosis	17	1.74	3.41E-02	CAV1, CLTA, PSD3, VPS37B, ARF6, SNX4, ASAP3, CLTC, DAB2, AP2B1, CHMP1A, ARF3, PDGFRA, RAB5A, GIT2, STAM, EPN1
cfa04730	Long-term depression	7	2.78	3.82E-02	GNAI3, KRAS, GNAQ, PPP2CA, PPP2CB, GUCY1A3, IGF1
cfa04919	Thyroid hormone signal transduction pathway	10	2.11	4.61E-02	ACTB, ACTG1, SLC16A2, PLCE1, NOTCH1, KRAS, ATP1B3, RHEB, PRKACB, MYC
cfa04310	Wnt signal transduction pathway	11	2.00	4.69E-02	MAP3K7, CTBP2, CCND2, JUN, LEF1, PRKACB, PPP3CA, SKP1, DAAM1, MYC, FBXW11

to the miR-200 family, including miR-200a-3p, miR-200c-3p, miR-200c-5p, miR-141-3p, and miR-429. The miR-200 family has been suggested to target multiple genes that are involved in cell proliferation, invasion, and inflammation. Thus, the aberrant expression of miR-200 may negatively regulate the endometrial development which would result in endometriosis or endometrial cancer (Panda et al., 2012).

Through analyzing the interactions between DEMs and their targets, some vital pathways, including MAPK, p53, PI3K–Akt, and Wnt signal transduction, were found to participate in TE. As endometrial thickness has been recognized as an important indicator of endometrial receptivity (Ledee-Bataille et al., 2002), we thus assume that the abnormalities of these pathways may compromise the development of the endometrium. For example, rapid activation of PI3K/Akt signaling cascades by growth factors and estrogen is involved in the migration of normal endometrial stromal cells (Gentilini et al., 2007). However, the expression of DEGs in the PI3K/AKT pathway, including EFNA1, FGF9, LPAR1, CCNE2, SGK1, MET, IL6R, and PDGFRA, was decreased in thin endometrium, which suggests that the repair ability of the thin endometrium was impaired during the proliferative phase (Le et al., 2016). Similarly, the abnormal Wnt/beta-catenin signal pathway would also impair the proliferation of estrogen-dependent endometrial cells (Tepekoy et al., 2015).

In the present study, our miRNA–mRNA regulatory networks provided a complete profile for the underlying mechanism of thin endometrium formation, and the hub genes identified in the networks may play certain roles in the development of thin endometrium. CAV1 expression is associated with cell survival and proliferation (Zhao et al., 2013). MET, the receptor

for insulin-like growth factor, potentially affects the functions of the endometrium (Satterfield et al., 2008). Therefore, the present study may provide useful information for understanding of the miRNA-mediated changes in mRNA expression in thin endometrium, and a further understanding of the functions of **miRNA–mRNA** networks can provide a new perspective for future studies examining potentially novel biomarkers and therapeutic targets.

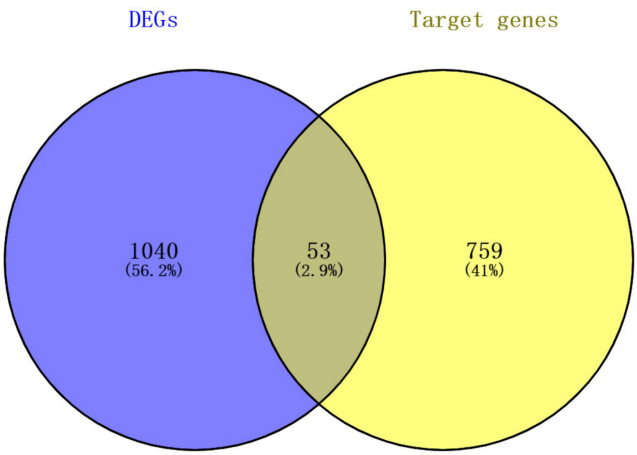


FIGURE 4 | The intersecting mRNAs between the common predicted target mRNAs and differentially expressed genes.

TABLE 7 | Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analyses of consistently expressed genes.

Term	Count	Fold enrichment	*P*-value	Genes
GO:0030308—cell growth negative regulation	2	76.69298	0.025132	OSGIN2, SLIT2
GO:0006954—inflammatory response	3	10.30204	0.032433	SGK1, THBS1, PTGFR
GO:0001558—cell growth regulation	2	20.00686	0.093047	SGK1, IGFBP5
GO:0007605—sensory perception of sound	2	20.00686	0.093047	CLIC4, MARVELD2
xtr04068: FoxO signaling pathway	3	6.969466	0.059369	SGK1, GABARAPL1
xtr04140: regulation of autophagy	2	24.34667	0.074102	GABARAPL1

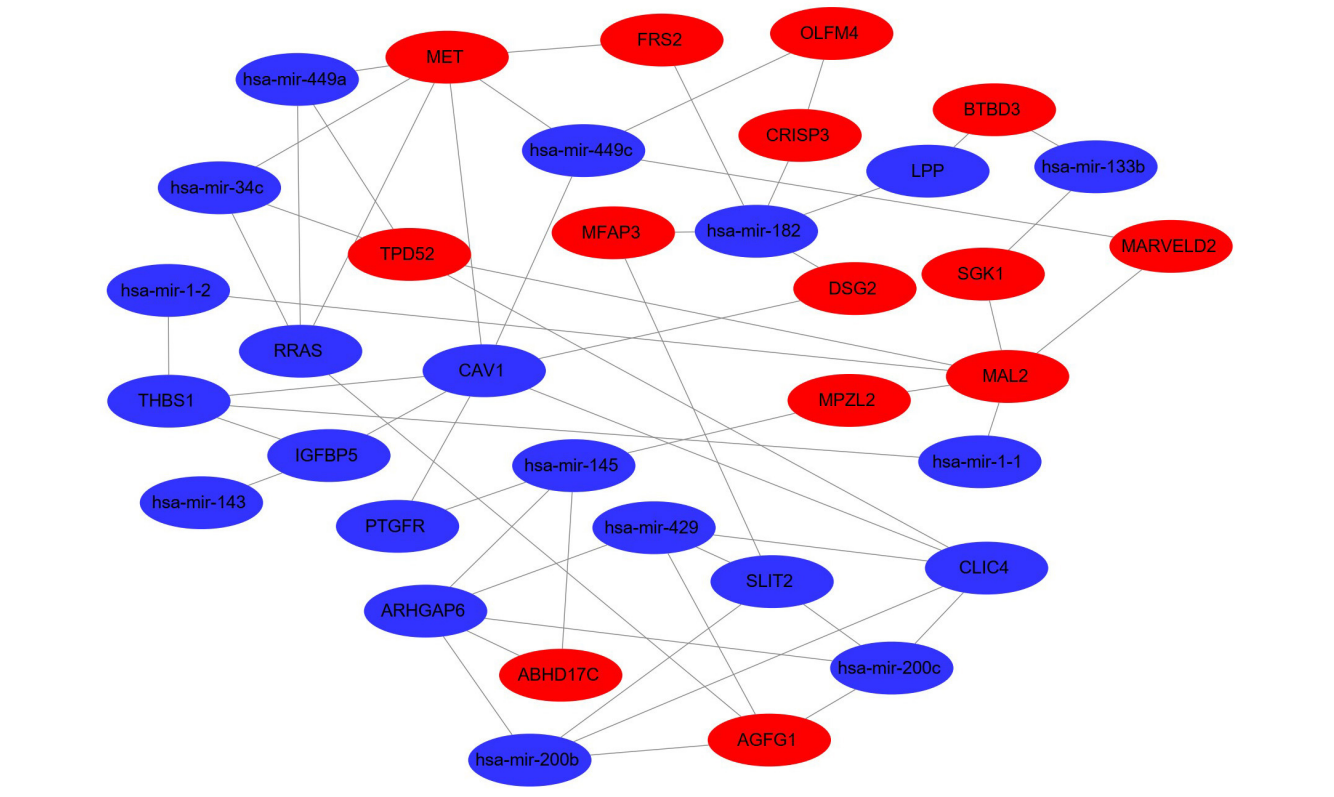


FIGURE 5 | Differentially expressed miRNA–differentially expressed gene regulatory network. The red and green colors denote upregulation and downregulation, respectively.

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 Materials**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committees on Human Research of the First Hospital Affiliated with University of Science and Technology of China. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DL, ZW, and WT contributed to sample collection. XT and YM designed the experiment. SZ, LZ, and BX performed the experiment, data analysis, and manuscript preparation. XT and BX revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by grants from the National Key Research and Development Project (grant number: 2019YFA0802600), the National Natural Science Foundation of China (grant numbers: 81971333 and 81971339), the Anhui Natural Science Foundation (grant numbers: S2012010009664 and 1808085MH242), and the Fundamental Research Funds for the Central Universities (grant number: WK9110000140).

ACKNOWLEDGMENTS

The authors also acknowledge the assistance provided by Dr. Xiaohua Jiang for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.589408/full#supplementary-material

REFERENCES

(1988). The American Fertility Society classifications of adnexal adhesions, distal tubal occlusion, tubal occlusion secondary to tubal ligation, tubal pregnancies, mullerian anomalies and intrauterine adhesions. *Fertil. Steril.* 49, 944–955. doi: 10.1016/S0015-0282(16)59942-7

Altmae, S., Martinez-Conejero, J.A., Esteban, F.J., Ruiz-Alonso, M., Stavreus-Evers, A., Horcajadas, J.A., et al. (2013). MicroRNAs miR-30b, miR-30d, and miR-494 regulate human endometrial receptivity. *Reprod. Sci.* 20, 308–317. doi: 10.1177/1933719112453507

Azizi, R., Aghebati-Maleki, L., Nouri, M., Marofi, F., Negargar, S., and Yousefi, M. (2018). Stem cell therapy in Asherman syndrome and thin endometrium: stem cell- based therapy. *Biomed. Pharmacother.* 102, 333–343. doi: 10.1016/j.biopha.2018.03.091

Di Pietro, C., Caruso, S., Battaglia, R., Iraci Sareri, M., La Ferlita, A., Strino, F., et al. (2018). MiR-27a-3p and miR-124-3p, upregulated in endometrium and serum from women affected by Chronic Endometritis, are new potential molecular markers of endometrial receptivity. *Am. J. Reprod. Immunol.* 80:e12858. doi: 10.1111/aji.12858

Du, J., Lu, H., Yu, X., Dong, L., Mi, L., Wang, J., et al. (2020). The effect of icariin for infertile women with thin endometrium: a protocol for systematic review. *Medicine* 99:e19111. doi: 10.1097/MD.0000000000019111

Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., and Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.* 18, 1139–1146. doi: 10.1038/nsmb.2115

Gentilini, D., Busacca, M., Di Francesco, S., Vignali, M., Vigano, P., and Di Blasio, A.M. (2007). PI3K/Akt and ERK1/2 signalling pathways are involved in endometrial cell migration induced by 17beta-estradiol and growth factors. *Mol. Hum. Reprod.* 13, 317–322. doi: 10.1093/molehr/gam001

Hong, L., Liu, R., Qiao, X., Wang, X., Wang, S., Li, J., et al. (2019). Differential microRNA expression in porcine endometrium involved in remodeling and angiogenesis that contributes to embryonic implantation. *Front. Genet.* 10:661. doi: 10.3389/fgene.2019.00661

Jiang, P., Tang, X., Wang, H., Dai, C., Su, J., Zhu, H., et al. (2019). Collagen-binding basic fibroblast growth factor improves functional remodeling of scarred endometrium in uterine infertile women: a pilot study. *Sci. China Life Sci.* 62, 1617–1629. doi: 10.1007/s11427-018-9520-2

Jimenez, P.T., Mainigi, M.A., Word, R.A., Kraus, W.L., and Mendelson, C.R. (2016). miR-200 regulates endometrial development during early pregnancy. *Mol. Endocrinol.* 30, 977–987. doi: 10.1210/me.2016-1050

Laurent, L.C. (2008). MicroRNAs in embryonic stem cells and early embryonic development. *J. Cell. Mol. Med.* 12, 2181–2188. doi: 10.1111/j.1582-4934.2008.00513.x

Le, A.W., Shan, L.L., Dai, X.Y., Xiao, T.H., Li, X.R., Wang, Z.H., et al. (2016). PI3K, AKT, and P-AKT levels in thin endometrium. *Genet. Mol. Res.* 15. doi: 10.4238/gmr.15017184

Ledee-Bataille, N., Olivennes, F., Lefaix, J.L., Chaouat, G., Frydman, R., and Delanian, S. (2002). Combined treatment by pentoxifylline and tocopherol for recipient women with a thin endometrium enrolled in an oocyte donation programme. *Hum. Reprod.* 17, 1249–1253. doi: 10.1093/humrep/17.5.1249

Li, M., Jia, F., Zhou, H., Di, J., and Yang, M. (2018). Elevated aerobic glycolysis in renal tubular epithelial cells influences the proliferation and differentiation of podocytes and promotes renal interstitial fibrosis. *Eur. Rev. Med. Pharmacol. Sci.* 22, 5082–5090. doi: 10.26355/eurrev_201808_15701

Li, Y., Zhuo, Z.J., Zhou, H., Liu, J., Xiao, Z., Xiao, Y., et al. (2019). miR-34b/c rs4938723 T>C decreases neuroblastoma risk: a replication study in the hunan children. *Dis. Markers* 2019:6514608. doi: 10.1155/2019/6514608

Liu, J., Ying, Y., Wang, S., Li, J., Xu, J., Lv, P., et al. (2020). The effects and mechanisms of GM-CSF on endometrial regeneration. *Cytokine* 125:154850. doi: 10.1016/j.cyto.2019.154850

Maekawa, R., Taketani, T., Mihara, Y., Sato, S., Okada, M., Tamura, I., et al. (2017). Thin endometrium transcriptome analysis reveals a potential mechanism of implantation failure. *Reprod. Med. Biol.* 16, 206–227. doi: 10.1002/rmb2.12030

Mittal, A., Rana, S., Sharma, R., Kumar, A., Prasad, R., Raut, S.K., et al. (2019). Myocardin ablation in a cardiac-renal rat model. *Sci. Rep.* 9:5872. doi: 10.1038/s41598-019-42009-z

Mu, Y., Li, Q., Cheng, J., Shen, J., Jin, X., Xie, Z., et al. (2020). Integrated miRNA-seq analysis reveals the molecular mechanism underlying the effect of acupuncture on endometrial receptivity in patients undergoing fertilization: embryo transplantation. *3 Biotech.* 10:16. doi: 10.1007/s13205-019-1990-3

Nakamura, Y., Kita, S., Tanaka, Y., Fukuda, S., Obata, Y., Okita, T., et al. (2019). A disintegrin and metalloproteinase 12 prevents heart failure by regulating cardiac hypertrophy and fibrosis. *Am. J. Physiol. Heart Circ. Physiol.* 318, H238–H251. doi: 10.1152/ajpheart.00496.2019

Nicoli, S., Standley, C., Walker, P., Hurlstone, A., Fogarty, K.E., and Lawson, N.D. (2010). MicroRNA-mediated integration of haemodynamics and Vegf signalling during angiogenesis. *Nature* 464, 1196–1200. doi: 10.1038/nature08889

Panda, H., Pelakh, L., Chuang, T.D., Luo, X., Bukulmez, O., and Chegini, N. (2012). Endometrial miR-200c is altered during transformation into cancerous states and targets the expression of ZEBs, VEGFA, FLT1, IKKbeta, KLF9, and FBLN5. *Reprod. Sci.* 19, 786–796. doi: 10.1177/1933719112438448

Paul, A.B.M., Sadek, S.T., and Mahesan, A.M. (2019). The role of microRNAs in human embryo implantation: a review. *J. Assist. Reprod. Genet.* 36, 179–187. doi: 10.1007/s10815-018-1326-y

Qian, K., Hu, L., Chen, H., Li, H., Liu, N., Li, Y., et al. (2009). Hsa-miR-222 is involved in differentiation of endometrial stromal cells *in vitro*. *Endocrinology* 150, 4734–4743. doi: 10.1210/en.2008-1629

Rekker, K., Altmae, S., Suhorutshenko, M., Peters, M., Martinez-Blanch, J.F., Codoner, F.M., et al. (2018). A two-cohort RNA-seq study reveals changes in endometrial and blood miRNome in fertile and infertile women. *Genes* 9:574. doi: 10.3390/genes9120574

Satterfield, M.C., Hayashi, K., Song, G., Black, S.G., Bazer, F.W., and Spencer, T.E. (2008). Progesterone regulates FGF10, MET, IGFBP1, and IGFBP3 in the endometrium of the ovine uterus. *Biol. Reprod.* 79, 1226–1236. doi: 10.1095/biolreprod.108.071787

Shufaro, Y., Simon, A., Laufer, N., and Fatum, M. (2008). Thin unresponsive endometrium–a possible complication of surgical curettage compromising ART outcome. *J. Assist. Reprod. Genet.* 25, 421–425. doi: 10.1007/s10815-008-9245-y

Shukla, G.C., Singh, J., and Barik, S. (2011). MicroRNAs: processing, maturation, target recognition and regulatory functions. *Mol. Cell. Pharmacol.* 3, 83–92.

Tepekoy, F., Akkoyunlu, G., and Demir, R. (2015). The role of Wnt signaling members in the uterus and embryo during pre-implantation and implantation. *J. Assist. Reprod. Genet.* 32, 337–346. doi: 10.1007/s10815-014-0409-7

Vilella, F., Moreno-Moya, J.M., Balaguer, N., Grasso, A., Herrero, M., Martinez, S., et al. (2015). Hsa-miR-30d, secreted by the human endometrium, is taken up by the pre-implantation embryo and might modify its transcriptome. *Development* 142, 3210–3221. doi: 10.1242/dev.124289

Wang, X., and El Naqa, I.M. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* 24, 325–332. doi: 10.1093/bioinformatics/btm595

Zhang, Y., Zang, Q., Zhang, H., Ban, R., Yang, Y., Iqbal, F., et al. (2016). DeAnnIso: a tool for online detection and annotation of isomiRs from small RNA sequencing data. *Nucleic Acids Res.* 44, W166–W175. doi: 10.1093/nar/gkw427

Zhao, L., Zhou, S., Zou, L., and Zhao, X. (2013). The expression and functionality of stromal caveolin 1 in human adenomyosis. *Hum. Reprod.* 28, 1324–1338. doi: 10.1093/humrep/det042

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 June 2021
doi: 10.3389/fgene.2021.663449

Associations Between Sleep Quality and Health Span: A Prospective Cohort Study Based on 328,850 UK Biobank Participants

*Muhammed Lamin Sambou[1†], Xiaoyu Zhao[1†], Tongtong Hong[1†], Jingyi Fan[1], Til Bahadur Basnet[1], Meng Zhu[1], Cheng Wang[1], Dong Hang[1], Yue Jiang[1,2] and Juncheng Dai[1,2]**

[1] Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China, [2] Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China

OPEN ACCESS

Edited by:
Yuanwei Zhang,
University of Science and Technology of China, China

Reviewed by:
Shaohua Xie,
Karolinska Institutet (KI), Sweden
Liangdan Sun,
Anhui Medical University, China

***Correspondence:**
Juncheng Dai
djc@njmu.edu.cn

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

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 02 February 2021
Accepted: 10 May 2021
Published: 15 June 2021

Citation:
Sambou ML, Zhao X, Hong T, Fan J, Basnet TB, Zhu M, Wang C, Hang D, Jiang Y and Dai J (2021) Associations Between Sleep Quality and Health Span: A Prospective Cohort Study Based on 328,850 UK Biobank Participants. Front. Genet. 12:663449. doi: 10.3389/fgene.2021.663449

Objective: To examine the associations between sleep quality and health span using a prospective cohort design based on the UK Biobank (UKB).

Materials and Methods: This longitudinal cohort study enrolled 328,850 participants aged between 37 and 73 years from UKB to examine the associations between sleep quality and risk of terminated health span. End of health span was defined by eight events strongly associated with longevity (cancer, death, congestive heart failure, myocardial infarction, chronic obstructive pulmonary disease, stroke, dementia, and diabetes), and a sleep score was generated according to five sleep behavioral factors (sleep duration, chronotype, sleeplessness, daytime sleepiness, and snoring) to characterize sleep quality. The hazard ratio (HR) and 95% confidence intervals (CIs) were calculated by multivariate-adjusted Cox proportional hazards model. Moreover, we calculated population attributable risk percentage (PAR%) to reflect the public health significance of healthy sleep quality.

Results: Compared with poor sleep quality, participants with healthy sleep quality had a 15% (HR: 0.85, 95% CI: 0.81–0.88) reduced risk of terminated health span, and those of less-healthy sleep quality had a 12% (HR: 0.88, 95% CI: 0.85–0.92) reduced risk. Linear trend results indicated that the risk of terminated health span decreased by 4% for every additional sleep score. Nearly 15% health span termination events in this cohort would have been prevented if a healthy sleep behavior pattern was adhered to (PAR%: 15.30, 95% CI: 12.58–17.93).

Conclusion: Healthy sleep quality was associated with a reduced risk of premature end of health span, suggesting healthy sleep behavior may extend health span. However, further studies are suggested for confirmation of causality and potential mechanism.

Keywords: sleep quality, sleep score, health span, aging, population attributable risk percent, UK Biobank

INTRODUCTION

Health span is a significant phenotype that enables individuals to age in good health without chronic diseases or disability (Zenin et al., 2019). Although global life expectancy has increased (Gbd 2016 Mortality Collaborators, 2017), aging populations often suffer functional health loss, and absolute expansion of morbidity (Jagger et al., 2008; Gbd 2015 DALYs and Hale Collaborators, 2016). Due to the significance of sleep and the fact that humans spend one-third of their lives asleep, there is growing interest in sleep behavior as a determinant of health span (Kay and Dzierzewski, 2015). Moreover, an alarming number of individuals suffer from sleeping problems and sleep deprivation worldwide. It was estimated that over 36% of the global population are suffering from sleep loss (Kryger et al., 2017), and nearly 50–70 million Americans chronically suffer from sleeplessness and sleep-related disorders, which hinder daily functioning and adversely affect health and longevity (Institute of Medicine Committee on Sleep Medicine and Research, 2006).

Poor sleep moderates biological responses such as increased oxidative stress, altered inflammatory and coagulatory responses, neural autonomic control changes, and accelerated atherosclerosis (Kryger et al., 2017; Tobaldini et al., 2017), which also show the profound impact of sleep on maintaining individual health status. Recent studies revealed that sleep quality is associated with cardiometabolic health and mortality (Karthikeyan et al., 2019; Fan et al., 2020), as well as epigenetic and skin aging, frailty, and mental health (Lo et al., 2014; Oyetakin-White et al., 2015; Carskadon et al., 2019; Sun et al., 2020). Abnormal sleep duration (both short and long sleep duration) was associated with a higher risk of total cardiovascular diseases (CVDs), chronic heart disease (CHD), stroke, and myocardial infarction (MI) (Mesas et al., 2010; Cappuccio et al., 2011; Daghlas et al., 2019). Therefore, it is essential to pay particular attention to sleep problems.

Although the associations of sleep behavioral factors with morbidity and mortality risk are documented (Merikanto et al., 2013), the evidence related to health span is still insufficient and uncertain, especially from the perspective of integrating multiple sleep behaviors. Most of the studies were limited by their modest sample sizes, the inclusion of patients with certain diseases at baseline, short follow-up, or insufficient confounder control, leading to inconsistency in the findings (Cappuccio et al., 2011; Yin et al., 2017). To fill this void, we integrated eight predominant health span-terminating events (Zenin et al., 2019) and adopted a sleep score consisting of five sleep behavioral factors (chronotype, sleep duration, sleeplessness/insomnia, snoring, and daytime sleepiness) as a measurement for sleep quality. Therefore, our study aimed to assess the associations between sleep quality and health span based on a large-scale prospective cohort [UK Biobank (UKB)].

MATERIALS AND METHODS

Study Population

The study population was composed of 328,850 participants of the UKB, a large-scale prospective cohort study with over 500,000 participants recruited between 2006 and 2010 across the United Kingdom. A detailed description of the UKB project is reported elsewhere (Sudlow et al., 2015). Briefly, the participants (aged from 37 to 73 years) attended one of 22 assessment centers in England, Wales, and Scotland, where they completed baseline questionnaires, underwent various physical assessments, and reported medical conditions. The North West Multicenter Research Ethical Committee approved the UKB project, and participants' consent was obtained.

Before performing the analysis in this study, we pruned the data for suitability. First, we excluded 72,477 and 29,027 participants whose health span had terminated prior to the baseline according to in-patient hospital admissions data (UKB data category 2000) and self-reported diagnoses obtained *via* verbal interview (UKB data category 100074), respectively. Additionally, 72,153 participants with missing sleep-related data were excluded. Finally, 328,850 participants of the UKB were included in this study (**Supplementary Figure 1**).

Ascertainment of Sleep Behaviors

The self-reported sleep behaviors (chronotype, sleep duration, sleeplessness/insomnia, snoring, and daytime sleepiness) were measured in the UKB using a standardized questionnaire. Except for sleep duration and snoring, the responses were measured on Likert scales from "never/rarely" to "usually" experiences (Fan et al., 2020). Chronotype means the tendency for earlier or later timing of sleep. An individual who prefers going to bed and waking earlier is considered a "morning person," while a person who prefers going to bed and waking late is considered an "evening person" (Jones et al., 2019). Chronotype preference was assessed with the question "Do you consider yourself to be (i) "definitely a morning person," (ii) "more a morning than evening person," (iii) "more an evening than morning person," or (iv) "definitely an evening person." For sleep duration, participants responded to the question "About how many hours sleep do you get in every 24 h? (including naps)" with responses in hourly increments. Experience of sleeplessness/insomnia symptom was assessed with the question "Do you have trouble falling asleep at night or do you wake up in the middle of the night?" and the responses were given in 3-point Likert scale (never/rarely; sometimes; usually). Habitual snoring was assessed with the question "Does your partner or a close relative or friend complain about your snoring?" with responses of (i) yes or (ii) no. The question for subjective daytime sleepiness was "How likely are you to doze off or fall asleep during the daytime when you don't mean to (for example, when working, reading, or driving)?," with responses of (i) never/rarely, (ii) sometimes, (iii) often, or (iv) all of the time.

Definition of Sleep Score and Sleep Quality

According to an epidemiologic study associated with sleep patterns and incident cardiovascular disease (Fan et al., 2020), for each sleep behavior, participants with the low-risk sleep behavior were assigned a score of 1, while those classified as high risk earn the score of zero (0). Then, all component scores were

summed to acquire a total sleep score ranging from 0 to 5, with higher scores indicating healthier sleep patterns. Furthermore, we defined "sleep quality" as three levels: "healthy" (sleep scores 4–5), "less-healthy" (sleep scores 2–3), and "poor" (sleep scores 0–1) (Fan et al., 2020).

Here, the low-risk sleep behaviors include early chronotype ("morning" or "morning than evening") (Merikanto et al., 2013), sleep duration 7–8 h per day (Gallicchio and Kalesan, 2009; Itani et al., 2017), never or rarely experience sleeplessness/insomnia symptoms (Hsu et al., 2015; Javaheri and Redline, 2017), no self-reported snoring (Li et al., 2014), and no frequent daytime sleepiness ("never/rarely" or "sometimes") (Gangwisch et al., 2014).

Ascertainment of Outcome

Health span is defined generally as aging without functional health loss (GBD, 2015; Li et al., 2020). In this study, health span was defined based on eight predominant health-terminating events strongly associated with longevity, such as congestive heart failure (CHF), myocardial infarction (MI), chronic obstructive pulmonary disease (COPD), stroke, dementia, diabetes, cancer, and death (Zenin et al., 2019). Health span was considered "terminated" for only participants first time diagnosed with any of these conditions during the UKB follow-up.

For each selected condition, except for cancer and death, we compiled a list of hospital data codes [International Classification of Diseases, 10th Revision (ICD-10)] and self-reported data codes (UKB data coding 6) to define these conditions in our study (**Supplementary Table 1**). We used the "National cancer registries linkage to UKB" (UKB data category 100092) to define cancer, and the "National death registries linkage to UKB" (UKB data category 100093) to define death event. The National cancer registries linkage to UKB was updated until December 14, 2016, earlier than the other two databases (inpatient hospital admissions data: March 31, 2017; National death registries linkage to UKB: February 14, 2018). To ensure consistency for the three databases, we set December 14, 2016, as the end date of follow-up in this study. Therefore, we calculated the personal follow-up time from the date of attending assessment center until the date of health span termination or December 14, 2016, whichever occurred first.

Statistical Analysis

We applied descriptive statistics (mean, SD, and proportion) to explore the baseline characteristics of the participants and estimated multivariate-adjusted hazard ratio (HR) for terminated health span using Cox proportional hazards regression models (Chandola et al., 2010; Boden-Albala et al., 2012). The proportional hazards assumptions for the Cox model were tested using Schoenfeld residuals method (Weisberg, 2010). In the basic model, we adjusted for age, sex, and ethnicity and further adjusted, in the fully adjusted model, for Townsend Deprivation Index, education, body mass index (BMI), smoking status, alcohol consumption, physical activity, healthy diet, family history of diseases [cancer and cardiac–cerebrovascular disease (CCVD)], and medication (sleep-related drugs and aspirin/ibuprofen). More details of the covariates can be found in the section "**Supplementary Method**"). Furthermore, we calculated the population attributable risk percentage (PAR%) for high-risk sleep behaviors using the "epi2by2" function in "epiR" package of R (Stevenson et al., 2020). Stratified analyses were conducted according to age (<50, 50–60, and >60 years), gender (male and female), BMI (<30 and ≥30 kg/m^2), smoking status (never and ever), alcohol intake frequency (≥once a week and <once a week), physical activity (low and moderate&high), healthy diet intake (yes and no), college degree (yes and no), and Townsend Deprivation Index (≥median and <median) to examine heterogeneity across these subgroups.

Additionally, in sensitivity analysis, we constructed a weighted sleep score of five sleep behaviors using the following equation: weighted sleep score = (β1 × factor 1 + β2 × factor 2 +. + β5 × factor 5) × (5/sum of the β coefficients) to evaluate the reliability of the results (Fan et al., 2020). To validate the robustness of our findings, we further performed sensitivity analyses: (1) excluding participants with terminated health span within the first 2 years of follow-up, (2) excluding those with poor self-reported health status at baseline, (3) further adjustment for principal components (PC1–3) and genotype chip. All analyses were performed using R (version 4.1.0), and statistical significance was defined as two-sided p-value ≤ 0.05.

RESULTS

In total, 49,772 participants of the 328,850 participants had terminated health span during the follow-up period, and approximately half of the events were caused by cancer (46.38%), followed by MI (17.73%) and death (10.99%) (**Supplementary Table 2**). The median follow-up time was 7.66 years (interquartile range: 6.80–8.42 years).

The baseline characteristics of 328,850 participants are summarized in **Table 1**. Overall, 4.08% of the participants had poor sleep quality (sleep scores 0–1), 57.59% had less-healthy sleep quality (sleep scores 2–3), and 38.33% had healthy sleep quality (sleep scores 4–5), with corresponding 18.87%, 15.75%, and 13.81% terminated health span, respectively. The female population was slightly higher among the healthy sleep quality group (58.45%). More participants with healthy (37.70%) sleep quality attained higher education than those with less-healthy (32.27%) or poor (27.89%) sleep quality. Besides, participants in the healthy sleep quality group had a relatively lower mean BMI (26.38 kg/m^2), and approximately 37.64% of them engaged in high physical exercise. Participants who reported "currently smoking" seldom had healthy sleep quality (7.46%) compared to "never smoked" (61.45%). Participants with healthy sleep quality were more likely to have a healthy diet intake (79.15%) and less likely to have a family history of cardiovascular diseases (61.23%) and cancer (33.58%). Similarly, compared to poor sleep quality, participants with less-healthy and healthy sleep quality were less likely to take sleep-related drugs and aspirin/ibuprofen.

In **Table 2**, associations for sleep quality with risk of terminated health span were exhibited. Compared to poor sleep quality, participants with healthy sleep quality and less-healthy sleep quality had 15% (HR: 0.85, 95% CI: 0.81–0.88) and 12%

TABLE 1 | Baseline characteristics of 328,850 participants according to sleep quality.

Characteristics (%)	Sleep quality		
	Poor (*n* = 13,432)	Less-healthy (*n* = 189,371)	Healthy (*n* = 126,047)
Terminated health span	2,535 (18.87)	29,834 (15.75)	17,403 (13.81)
Age, years, mean (SD)	55.32 (7.89)	55.91 (7.99)	55.35 (8.27)
Townsend Index, mean (SD)	−0.91 (3.26)	−1.41 (3.03)	−1.64 (2.89)
BMI, kg/m^2, mean (SD)	29.22 (5.25)	27.50 (4.65)	26.38 (4.23)
Sex, female	6,359 (47.34)	1103,036 (54.41)	73,680 (58.45)
Ethnicity, white race	112,381 (92.18)	1179,379 (94.72)	1119,806 (95.05)
College or university degree	3,746 (27.89)	61,111 (32.27)	47,514 (37.70)
Smoking status			
Current smokers	2,320 (17.27)	21,568 (11.39)	9,397 (7.46)
Never smokers	6,077 (45.24)	1101,789 (53.75)	77,450 (61.45)
Alcohol intake frequency			
>3 times/week	5,915 (44.04)	87,335 (46.12)	55,260 (43.84)
Special occasions only/Never	2,665 (19.84)	32,431 (17.13)	22,404 (17.77)
Physical activity			
Low	2,841 (21.15)	30,115 (15.90)	15,907 (12.62)
High	3,754 (27.95)	61,972 (32.73)	47,442 (37.64)
Healthy diet	8,825 (65.70)	1140,048 (73.95)	99,760 (79.15)
Family history of CCVD	8,615 (64.14)	1119,497 (63.10)	77,184 (61.23)
Family history of cancer	4,909 (36.55)	66,667 (35.20)	42,323 (33.58)
Sleep-related drugs use	227 (1.69)	1,565 (0.83)	404 (0.32)
Aspirin/ibuprofen use	3,826 (28.48)	46,494 (24.55)	26,286 (20.85)
Having low-risk sleep factors (%)			
Early chronotype	879 (6.54)	93,740 (49.50)	1111,806 (88.70)
Sleep 7–8 h/day	548 (4.08)	1107,747 (56.90)	1119,576 (94.87)
Never/rarely insomnia	143 (1.06)	22,324 (11.79)	60,545 (48.03)
No self-reported snoring	612 (4.56)	96,133 (50.76)	1111,032 (88.09)
No frequent daytime sleepiness	10,639 (79.21)	1184,742 (97.56)	1125,777 (99.79)

The chi-square test for categorical variables and Kruskal–Wallis test for continuous variables were used to calculate the p values across the sleep quality, and all the variables had p value < 0.001.
SD, standard deviation; BMI, body mass index; CCVD, cardiac–cerebrovascular disease.

(HR: 0.88, 95% CI: 0.85–0.92) reduced risk of terminated health span, respectively. The corresponding PAR% for less-healthy and healthy sleep quality was 1.29% (PAR%: 1.29, 95% CI: 1.01–1.58) and 3.41% (PAR%: 3.41; 95% CI: 2.95–3.88), respectively.

From the perspective of sleep score, we found the participants with the score 5 had the lowest risk of premature end of health span (HR: 0.84, 95% CI: 0.80–0.88), and the trend analysis also revealed that the risk of terminated health span decreased by 4% (HR: 0.96, 95% CI: 0.96–0.97) for every additional sleep score (**Figure 1A**). Moreover, the corresponding PAR% for score 5 was nearly 15% (PAR%:14.31; 95% CI: 12.45–16.13) compared to those with the lowest sleep scores (**Figure 1B**).

Additionally, we also demonstrated the cumulative hazard curves between sleep situation and terminated health span. **Figure 2A** showed that with increasing follow-up time, the cumulative hazard of terminated health span increased more in individuals with poor sleep quality than those with less-healthy or healthy sleep quality. Similar results were observed for sleep score, showing distinct risk between sleep scores 0–1 and high scores (**Figure 2B**).

Then, we further explored the effects of related sleep traits on health span (**Table 3**). Participants with the low-risk sleep behaviors such as "sleep duration 7–8 h/day" (HR: 0.94, 95% CI: 0.92-0.95), "Never/rarely insomnia" (HR: 0.94, 95% CI: 0.92–0.96), and "rarely daytime sleepiness" (HR: 0.83, 95% CI: 0.79–0.87) had decreased risk of terminated health span. Furthermore, PAR% for terminated health span suggests that nearly 15% (PAR%: 15.30, 95% CI: 12.58–17.93) of terminated health span in this cohort would not have occurred if all participants had been in the low-risk group for all five sleep factors.

In stratified analyses, we observed that the associations of sleep quality with the risk of terminated health span were largely consistent across subgroups, except the smoking status. The ever smokers had a stronger association than never smokers (**Supplementary Figure 2**). Additionally, we constructed a weighted sleep score to reevaluate its association with the risk of terminated health span. We found that high-grade weighted sleep score (weighted sleep score 4∼5 vs. 0∼<1, HR: 0.76, 95% CI: 0.71–0.82) also reduced the risk of terminated health span (**Table 4**).

Moreover, further sensitivity analyses were performed by respectively excluding participants with terminated health span within the first 2 years of follow-up and those with poor self-reported health status at baseline. The associations were largely similar to our previous findings (**Supplementary Tables 4, 5**). Given that complicated structure of population in UKB, we additionally adjusted the top 3 principal components (PC1–3) and genotype chip to offset the potential effect. Similarly, the results were consistent and supported the robustness of the observed findings in our study (**Supplementary Table 6**).

DISCUSSION

In this large-scale prospective cohort study, we examined the associations of sleep quality/sleep score with risk of terminated health span based on 328,850 participants of the UKB. Participants with a healthy sleep quality had a 15% lower risk of terminated health span. The PAR% further suggested that nearly 15% of terminated health span in this cohort would not have occurred if all participants had low-risk sleep behavior for all five sleep behavioral factors. Besides, four sensitivity analyses implemented in this study indicated that the associations we found are robust and reliable to some degree.

TABLE 2 | Associations for sleep quality with risk of terminated health span among 328,850 participants.

Sleep quality	Total N (%)	Cases N (%)	Basic model*	Fully adjusted model†	PAR% (95% CI)
			HR (95% CI)[a]	HR (95% CI)[a]	
Poor	13,432 (4.08)	2,535 (18.87)	ref	ref	ref
Less-healthy	189,371 (57.59)	29,834 (15.75)	0.80 (0.77–0.83)	0.88 (0.85–0.92)	1.29 (1.01–1.58)
Healthy	126,047 (38.33)	17,403 (13.81)	0.72 (0.69–0.75)	0.85 (0.81–0.88)	3.41 (2.95–3.88)

N, number; HR, hazard ratio; 95% CI, 95% confidence interval; ref, reference; PAR%, population attributable risk percentage; BMI, body mass index; CCVD, cardiac–cerebrovascular disease.

**Basic model: adjusted for age, sex, and ethnicity.*

†Fully adjusted model: additionally adjusted for Townsend Deprivation Index, education, BMI, smoking status, alcohol consumption, physical activity, healthy diet, family history of diseases (cancer and CCVD), and medication (sleep-related drugs and aspirin/ibuprofen).

[a]Each group was compared to participants with poor sleep quality.

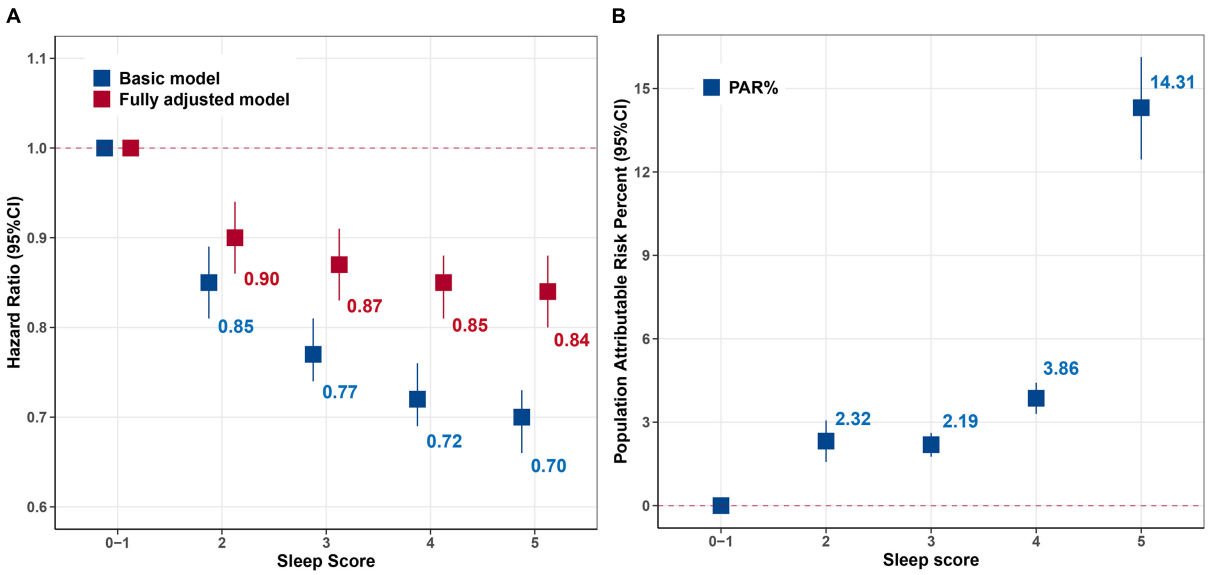


FIGURE 1 | Hazard ratio (HR) and population attributable risk percentage (PAR%) for risk of terminated health span according to sleep score. **(A)** HR [95% confidence interval (95% CI)]. **(B)** Population attributable risk percentage (PAR%). Basic model: adjusted for age, sex, and ethnicity. Fully adjusted model further adjusted for Townsend Deprivation Index, education, body mass index (BMI), smoking status, alcohol consumption, physical activity, healthy diet, family history of diseases [cancer and cardiac–cerebrovascular disease (CCVD)], and medication (sleep-related drugs and aspirin/ibuprofen). Sleep scores 0–1 were the reference. The values were the point estimation of HRs and PAR%.

Our results are in line with other comparable findings that sleep behavior affects health and wellbeing (Cappuccio et al., 2011; Gangwisch et al., 2014). Although sleep behavioral factors separately have a bearing on health, it was significant to evaluate a combination of sleep behavioral factors due to their synchrony that could jointly increase the risk of health span termination (Javaheri and Redline, 2017). For instance, a previous study showed that insomnia/sleeplessness was related to sleep duration and excessive daytime sleepiness (EDS), and late chronotype reduced sleep duration (Depner et al., 2014). Thus, we generated a sleep score integrating five sleep behaviors to comprehensively assess sleep quality and its association with health span, which was characterized by a host of eight health events that are commonly involved in health span termination (Zenin et al., 2019). Our study showed that healthy sleep quality reduced the risk of terminated health span, suggesting that healthy sleep behavior can improve health span. In agreement with our finding, previous studies showed that insomnia accompanied short sleep duration (Hsu et al., 2015; He et al., 2017; Javaheri and Redline, 2017), and habitual snoring with EDS increased the risk of hypertension, lung cancer (Liu et al., 2019; Campos et al., 2020), vascular death (Blachier et al., 2012; Boden-Albala et al., 2012), atherosclerosis (Sands et al., 2013; Javaheri and Redline, 2017), and diabetes (Li et al., 2015).

We also observed a decreased risk of terminated health span for single low-risk sleep behaviors, such as "sleep duration 7–8 h/day," "no daytime sleepiness," "never/rarely insomnia/sleeplessness," and "early chronotype." Similarly, high-risk sleep behavioral factors including "late chronotype" (Merikanto et al., 2013; Erren et al., 2016), "abnormal sleep duration" (Mesas et al., 2010; Cappuccio et al., 2011; Rudnicka et al., 2017; Daghlas et al., 2019), "frequently experience insomnia/sleeplessness" (Hsu et al., 2015; Javaheri and Redline, 2017), "habitual snoring" (Seidel et al., 2012; Sands et al., 2013;

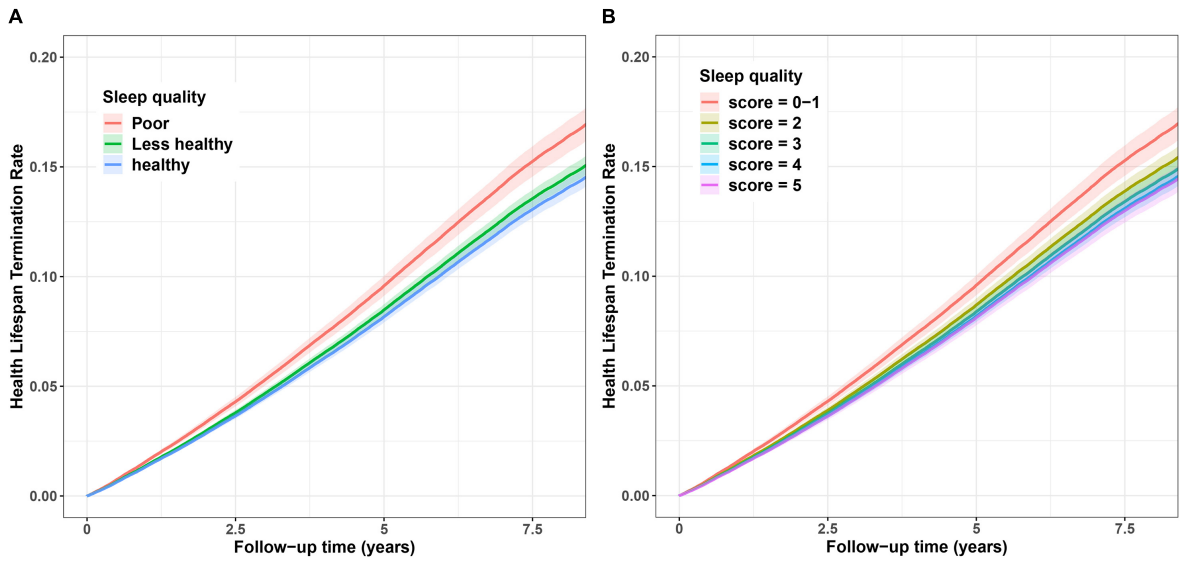


FIGURE 2 | Cumulative hazard curve for the associations between sleep quality and risk of terminated health span. **(A)** Sleep quality (poor, less–healthy, and healthy). **(B)** Sleep score (scores 0–1, 2, 3, 4, and 5). The *Y*-axis represents cumulative hazard of terminated health span, the while *X*-axis represents the follow-up time (years). Shaded regions represent the 95% confidence intervals (95% CIs). Cumulative hazard curves were based on the fully adjusted model.

TABLE 3 | Associations for low-risk sleep behaviors with risk of terminated health span among 328,850 participants.

Sleep behaviors	Total (N)	Cases (N)	Basic model*	Fully adjusted model†	PAR% (95% CI)
			HR (95% CI)[a]	HR (95% CI)[a]	
Early chronotype	206,425 (62.77)	31,686 (15.35)	0.95 (0.94–0.97)	1.00 (0.98–1.01)	−1.42 (−2.04 to −0.80)
Sleep 7–8 h/day	227,871 (69.29)	33,049 (14.50)	0.88 (0.87–0.90)	0.94 (0.92–0.95)	4.17 (3.62–4.72)
Never/rarely insomnia	83,012 (25.24)	11,510 (13.87)	0.92 (0.90–0.94)	0.94 (0.92–0.96)	8.39 (7.02–9.74)
No self-reported snoring	207,777 (63.18)	29,664 (14.28)	0.93 (0.91–0.95)	0.99 (0.97–1.01)	5.67 (5.04–6.30)
Rarely daytime sleepiness	321,158 (97.66)	48,208 (15.01)	0.77 (0.73–0.81)	0.83 (0.79–0.87)	0.82 (0.68–0.96)
All five factors (overall)[b]	24,548 (7.46)	3,147 (12.82)	0.89 (0.86–0.92)	0.96 (0.93–1.00)	15.30 (12.58–17.93)

**Basic model: adjusted for age, sex, and ethnicity.*
[a]Compared with all other participants not in this low-risk group.
[b]All low-risk factors were included simultaneously in the same model, and participants without all five low-risk behaviors were set as the reference
†Fully adjusted model: additionally adjusted for Townsend Deprivation Index, education, BMI, smoking status, alcohol consumption, physical activity, healthy diet, family history of diseases (cancer and CCVD), and medication (sleep-related drugs and aspirin/ibuprofen).

Li et al., 2015), and "excessive daytime sleepiness" (Blachier et al., 2012; Boden-Albala et al., 2012; Barfield et al., 2019) were associated with increased risk of chronic disease morbidity and mortality. If all these five high-risk sleep behaviors were rectified appropriately, nearly 15% of terminated health span would have been prevented, highlighting the importance of adhering to healthy sleep behaviors. However, it is worth noting that due to the multiple-center and large-scale design of UKB, these 328,850 participants aged from 37 to 73 years were nationwide. Besides, sleep traits were collected by trained volunteers according to a standard questionnaire. Therefore, the exposure distribution could represent the general population of United Kingdom, indicating the reliability of the PAR% we calculated at some degree, although a further validation in other cohorts was still necessary.

Biologically, sleep regulates many important pathways in the human physiology, such as autonomic, sympathetic, cardiometabolic, and immunologic responses (Cappuccio et al., 2011; Tobaldini et al., 2017), which support that adopting healthy sleep behavior to the circadian rhythm would enhance health and quality of life (Backhaus et al., 2015; Jones et al., 2019). On the other hand, poor sleep quality affects functional health in both young and adult, including disruption of cognitive performance and diurnal alertness (Skeldon et al., 2016). Moreover, it is essential to be cautious about shared and non-shared environmental determinants of ill-sleep habit (Gregory et al., 2016), including lifestyle, such as alcohol dependence, smoking, obesity, physical inactivity, and stress, that could upset healthy sleep behaviors (Chakravorty et al., 2016; Christie et al., 2016; Liao et al., 2019; Garcia-Marin et al., 2020).

Moreover, sleep disorders and poor sleep habits are alarming health threats warranting more public attention and are necessary to take appropriate action to promote sleep quality and health status, particularly for those with irregular sleep patterns, such

TABLE 4 | Associations for weighted sleep score with risk of terminated health span among 328,850 participants.

Weighted score	Total N (%)	Cases N (%)	Basic model*	Fully adjusted model†
			HR (95% CI)[a]	HR (95% CI)[a]
0∼ < 1	3,409 (1.04)	746 (21.88)	ref	ref
1∼ < 2	3,864 (1.18)	749 (19.38)	0.83 (0.75–0.92)	0.89 (0.81–0.99)
2∼ < 3	28,229 (8.58)	5,055 (17.91)	0.80 (0.74–0.86)	0.84 (0.77–0.90)
3∼ < 4	120,549 (36.66)	19,418 (16.11)	0.72 (0.67–0.78)	0.79 (0.74–0.85)
4∼5	172,799 (52.55)	23,804 (13.78)	0.65 (0.60–0.70)	0.76 (0.71–0.82)
Overall (continuous)	328,850 (100.00)	49,772 (15.14)	0.91 (0.90–0.92)	0.94 (0.93–0.95)

N, number; HR, hazard ratio; 95% CI, 95% confidence interval; ref, reference; BMI, body mass index; CCVD, cardiac–cerebrovascular disease.
**Basic model: adjusted for age, sex, and ethnicity.*
†Fully adjusted model: additionally adjusted for Townsend Deprivation Index, education, BMI, smoking status, alcohol consumption, physical activity, healthy diet, family history of diseases (cancer and CCVD), and medication (sleep-related drugs and aspirin/ibuprofen).
[a]Each group was compared to participants with 0–1 sleep scores.

as shift workers (Palermo et al., 2015; Redeker et al., 2019). A previous study among shift workers showed that resting and napping lowered the levels of sleepiness at the end of the shift (Barthe et al., 2015), which means more efficiency at work and less chances of accidents due to sleepiness (Ruggiero and Redeker, 2014; Geiger-Brown et al., 2016) and ultimately beneficial for the extension of health span. In addition to potential contributions to the individual's life quality, a healthy sleep quality may also mitigate extravagant medical costs associated with chronic disease morbidity as well as lighten the heavy burden of social demands on health services. Therefore, we not only aim to investigate the potential effects of sleep quality on health span but also hope to call for more attention to individual sleep problems and correct improper sleep patterns as far as possible.

Here, the definition of health span we adopted is a promising longevity phenotype, reflecting individual aging and health status. Based on the richness and accuracy of clinical information in UKB, the construction of the health span phenotype is reliable and robust. Thus, we have a chance to assess the associations between sleep quality and risk of premature health span termination for the first time. The sleep score constructed by five sleep behaviors is an effective way to measure the sleep quality quantitatively. Meanwhile, the reliable data, large-scale sample, and long-term follow-up time of UKB provide sufficient power for our study. However, all the sleep behaviors are self-reported, which may lead to misclassification of exposures inevitably. To our knowledge, misclassification will underestimate the associations we observed. Although we have adjusted the sociodemographic characteristics, lifestyles, and other confounding factors in the full model, residual confounding from unknown or unmeasured factors still remains possible. Thus, the effects of associations and the PAR% we calculated are essential to be further validated in other perspective cohorts. Thirdly, a single measurement of sleep behaviors at baseline is not satisfactory to reflect the dynamic change of sleep factors during the following time, which means that the evaluation of effects of changing sleep patterns on health span requires repeated measurements of sleep traits. Moreover, in our observational study, the potential causality is hard to determine, and further work is necessary. Finally, most of the study participants are white, and generalizing the findings to other populations should warrant caution.

In summary, we tentatively explored the effect of sleep quality on health life span in this study. A healthy sleep quality plays an important role in individual health status, aging, and diseases. Sleep problem is not only related to individual physical and mental health but also a public health and social problem, which deserves more attention and early intervention.

CONCLUSION

In this large-scale prospective study that enrolled 328,850 participants, we found that healthy sleep quality was associated with a reduced risk of premature end of health life span, suggesting that healthy sleep behaviors may be beneficial to extend health life span. Therefore, sleep problems deserve more attention and early intervention. However, further studies are suggested for confirmation of causality and potential mechanism.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the North West Multi-Centre Research Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JD contributed to the conception and design of the study. MS, XZ, and TH conducted the statistical

analysis and wrote the first draft. TH, JF, and TB helped apply for the permission to use data and offered technical support during the study. MZ, CW, DH, and YJ critically revised the manuscript for important intellectual content. All authors reviewed and approved the final manuscript.

FUNDING

This work was funded by the National Natural Science of China (81941020).

ACKNOWLEDGMENTS

This study was conducted using the UK Biobank resource (Application Number 64689). We thank the study participants and research staff for their contributions and commitment to this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.663449/full#supplementary-material

REFERENCES

Backhaus, W., Kempe, S., and Hummel, F. C. (2015). The effect of sleep on motor learning in the aging and stroke population - a systematic review. *Restor. Neurol. Neurosci.* 34, 153–164. doi: 10.3233/rnn-150521

Barfield, R., Wang, H., Liu, Y., Brody, J. A., Swenson, B., Li, R., et al. (2019). Epigenome-wide association analysis of daytime sleepiness in the multi-ethnic study of atherosclerosis reveals African-American-specific associations. *Sleep* 42:zsz101.

Barthe, B., Tirilly, G., Gentil, C., and Toupin, C. (2015). Job demands and resting and napping opportunities for nurses during night shifts: impact on sleepiness and self-evaluated quality of healthcare. *Ind. Health* 54, 157–162. doi: 10.2486/indhealth.2015-0002

Blachier, M., Dauvilliers, Y., Jaussent, I., Helmer, C., Ritchie, K., Jouven, X., et al. (2012). Excessive daytime sleepiness and vascular events: the three city study. *Ann. Neurol.* 71, 661–667. doi: 10.1002/ana.22656

Boden-Albala, B., Roberts, E. T., Bazil, C., Moon, Y., Elkind, M. S., Rundek, T., et al. (2012). Daytime sleepiness and risk of stroke and vascular disease: findings from the Northern manhattan study (NOMAS). *Circ. Cardiovasc. Qual. Outcomes* 5, 500–507. doi: 10.1161/circoutcomes.111.963801

Campos, A. I., García-Marín, L. M., Byrne, E. M., Martin, N. G., Cuéllar-Partida, G., and Rentería, M. E. (2020). Insights into the aetiology of snoring from observational and genetic investigations in the UK Biobank. *Nat. Commun.* 11:817.

Cappuccio, F. P., Cooper, D., D'Elia, L., Strazzullo, P., and Miller, M. A. (2011). Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. *Eur. Heart J.* 32, 1484–1492. doi: 10.1093/eurheartj/ehr007

Carskadon, M. A., Chappell, K. R., Barker, D. H., Hart, A. C., Dwyer, K., Gredvig-Ardito, C., et al. (2019). A pilot prospective study of sleep patterns and DNA methylation-characterized epigenetic aging in young adults. *BMC Res. Notes* 12:583.

Chakravorty, S., Chaudhary, N. S., and Brower, K. J. (2016). Alcohol dependence and its relationship with insomnia and other sleep disorders. *Alcohol. Clin. Exp. Res.* 40, 2271–2282. doi: 10.1111/acer.13217

Chandola, T., Ferrie, J. E., Perski, A., Akbaraly, T., and Marmot, M. G. (2010). The effect of short sleep duration on coronary heart disease risk is greatest among those with sleep disturbance: a prospective study from the whitehall II cohort. *SLEEP* 33, 739–744. doi: 10.1093/sleep/33.6.739

Christie, A. D., Seery, E., and Kent, J. A. (2016). Physical activity, sleep quality, and self-reported fatigue across the adult lifespan. *Exp. Gerontol.* 77, 7–11. doi: 10.1016/j.exger.2016.02.001

Daghlas, I., Dashti, H. S., Lane, J., Aragam, K. G., Rutter, M. K., Saxena, R., et al. (2019). Sleep duration and myocardial infarction. *J. Am. Coll. Cardiol.* 74, 1304–1314. doi: 10.1016/j.jacc.2019.07.022

Depner, C. M., Stothard, E. R., and Wright, K. P. (2014). Metabolic consequences of sleep and circadian disorders. *Curr. Diab. Rep.* 14:507.

Erren, T. C., Morfeld, P., Foster, R. G., Reiter, R. J., Groß, J. V., and Westermann, I. K. (2016). Sleep and cancer: synthesis of experimental data and meta-analyses of cancer incidence among some 1,500,000 study individuals in 13 countries. *Chronobiol. Int.* 33, 325–350. doi: 10.3109/07420528.2016.1149486

Fan, M., Sun, D., Zhou, T., Heianza, Y., Lv, J., Li, L., et al. (2020). Sleep patterns, genetic susceptibility, and incident cardiovascular disease: a prospective study of 385 292 UK biobank participants. *Eur. Heart J.* 41, 1182–1189. doi: 10.1093/eurheartj/ehz849

Gallicchio, L., and Kalesan, B. (2009). Sleep duration and mortality: a systematic review and meta-analysis. *J. Sleep Res.* 18, 148–158. doi: 10.1111/j.1365-2869.2008.00732.x

Gangwisch, J. E., Rexrode, K., Forman, J. P., Mukamal, K., Malaspina, D., and Feskanich, D. (2014). Daytime sleepiness and risk of coronary heart disease and stroke: results from the Nurses' Health study II. *Sleep Med.* 15, 782–788. doi: 10.1016/j.sleep.2014.04.001

Garcia-Marin, L. M., Campos, A. I., Martin, N. G., Cuellar-Partida, G., and Renteria, M. E. (2020). Inference of causal relationships between sleep-related traits and 1,527 phenotypes using genetic data. *Sleep* 44:zsaa154.

Gbd 2015 DALYs and Hale Collaborators (2016). Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* 388, 1603–1658.

Gbd 2016 Mortality Collaborators (2017). Global, regional, and national under-5 mortality, adult mortality, age-specific mortality, and life expectancy, 1970-2016: a systematic analysis for the Global burden of disease study 2016. *Lancet* 390, 1084–1150.

GBD (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 386, 743–800.

Geiger-Brown, J. P., Sagherian, K., Zhu, S., Blair, L., Warren, J., et al. (2016). Napping on the night shift: a two-hospital implementation project. *Am. J. Nursing* 116, 26–33. doi: 10.1097/01.naj.0000482953.88608.80

Gregory, A. M., Rijsdijk, F. V., Eley, T. C., Buysse, D. J., Schneider, M. N., Parsons, M., et al. (2016). A longitudinal twin and sibling study of associations between insomnia and depression symptoms in young adults. *Sleep* 39, 1985–1992. doi: 10.5665/sleep.6228

He, Q., Sun, H., Wu, X., Zhang, P., Dai, H., Ai, C., et al. (2017). Sleep duration and risk of stroke: a dose-response meta-analysis of prospective cohort studies. *Sleep Med.* 32, 66–74. doi: 10.1016/j.sleep.2016.12.012

Hsu, C. Y., Chen, Y. T., Chen, M. H., Huang, C. C., Chiang, C. H., Huang, P. H., et al. (2015). The association between insomnia and increased future cardiovascular events: a nationwide population-based study. *Psychosom. Med.* 77, 743–751. doi: 10.1097/psy.0000000000000199

Institute of Medicine Committee on Sleep Medicine and Research (2006). "The national academies collection: reports funded by national institutes of health," in *Sleep Disorders and Sleep Deprivation: an Unmet Public Health Problem*, eds H. R. Colten and B. M. Altevogt (Washington, DC: National Academies Press).

Itani, O., Jike, M., Watanabe, N., and Kaneita, Y. (2017). Short sleep duration and health outcomes: a systematic review, meta-analysis, and meta-regression. *Sleep Med.* 32, 246–256. doi: 10.1016/j.sleep.2016.08.006

Jagger, C., Gillies, C., Moscone, F., Cambois, E., Van Oyen, H., Nusselder, W., et al. (2008). Inequalities in healthy life years in the 25 countries of the European Union in 2005: a cross-national meta-regression analysis. *Lancet* 372, 2124–2131. doi: 10.1016/s0140-6736(08)61594-9

Javaheri, S., and Redline, S. (2017). Insomnia and risk of cardiovascular disease. *Chest* 152, 435–444. doi: 10.1016/j.chest.2017.01.026

Jones, S. E., Lane, J. M., Wood, A. R., van Hees, V. T., Tyrrell, J., Beaumont, R. N., et al. (2019). Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. *Nat. Commun.* 10:343.

Karthikeyan, R., Spence, D. W., and Pandi-Perumal, S. R. (2019). The contribution of modern 24-hour society to the development of type 2 diabetes mellitus: the role of insufficient sleep. *Sleep Sci.* 12, 227–231.

Kay, D. B., and Dzierzewski, J. M. (2015). Sleep in the context of healthy aging and psychiatric syndromes. *Sleep Med. Clin.* 10, 11–15. doi: 10.1016/j.jsmc.2014.11.012

Kryger, M. H., ROth, T., and Dement, W. C. (2017). *Principles and Practice of Sleep Medicine*. Amsterdam: Elsevier.

Li, D., Liu, D., Wang, X., and He, D. (2014). Self-reported habitual snoring and risk of cardiovascular disease and all-cause mortality. *Atherosclerosis* 235, 189–195. doi: 10.1016/j.atherosclerosis.2014.04.031

Li, M., Li, K., Zhang, X. W., Hou, W. S., and Tang, Z. Y. (2015). Habitual snoring and risk of stroke: a meta-analysis of prospective studies. *Int. J. Cardiol.* 185, 46–49. doi: 10.1016/j.ijcard.2015.03.112

Li, Y., Schoufour, J., Wang, D. D., Dhana, K., Pan, A., Liu, X., et al. (2020). Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: prospective cohort study. *Bmj* 368:l6669. doi: 10.1136/bmj.l6669

Liao, Y., Xie, L., Chen, X., Kelly, B. C., Qi, C., Pan, C., et al. (2019). Sleep quality in cigarette smokers and nonsmokers: findings from the general population in central China. *BMC Public Health* 19:808.

Liu, W., Luo, M., Fang, Y. Y., Wei, S., Zhou, L., and Liu, K. (2019). Relationship between occurrence and progression of lung cancer and nocturnal intermittent hypoxia. apnea and daytime sleepiness. *Curr. Med. Sci.* 39, 568–575. doi: 10.1007/s11596-019-2075-6

Lo, J. C., Sim, S. K., and Chee, M. W. (2014). Sleep reduces false memory in healthy older adults. *Sleep* 37, 665-671, 671A.

Merikanto, I., Lahti, T., Puolijoki, H., Vanhala, M., Peltonen, M., Laatikainen, T., et al. (2013). Associations of chronotype and sleep with cardiovascular diseases and type 2 diabetes. *Chronobiol. Int.* 30, 470–477. doi: 10.3109/07420528.2012.741171

Mesas, A. E., López-García, E., León-Muñoz, L. M., Guallar-Castillón, P., and Rodríguez-Artalejo, F. (2010). Sleep duration and mortality according to health status in older adults. *J. Am. Geriatr. Soc.* 58, 1870–1877. doi: 10.1111/j.1532-5415.2010.03071.x

Oyetakin-White, P., Suggs, A., Koo, B., Matsui, M. S., Yarosh, D., Cooper, K. D., et al. (2015). Does poor sleep quality affect skin ageing? *Clin. Exp. Dermatol.* 40, 17–22. doi: 10.1111/ced.12455

Palermo, T. A., Rotenberg, L., Zeitoune, R. C., Silva-Costa, A., Souto, E. P., and Griep, R. H. (2015). Napping during the night shift and recovery after work among hospital nurses. *Rev. Lat. Am. Enfermagem* 23, 114–121. doi: 10.1590/0104-1169.0147.2532

Redeker, N. S., Caruso, C. C., Hashmi, S. D., Mullington, J. M., Grandner, M., and Morgenthaler, T. I. (2019). Workplace interventions to promote sleep health and an alert, healthy workforce. *J. Clin. Sleep Med.* 15, 649–657. doi: 10.5664/jcsm.7734

Rudnicka, A. R., Nightingale, C. M., Donin, A. S., Sattar, N., Cook, D. G., Whincup, P. H., et al. (2017). Sleep duration and Risk of Type 2 diabetes. *Pediatrics* 140, 529–537.

Ruggiero, J. S., and Redeker, N. S. (2014). Effects of napping on sleepiness and sleep-related performance deficits in night-shift workers: a systematic review. *Biol. Res. Nursing* 16, 134–142. doi: 10.1177/1099800413476571

Sands, M., Loucks, E. B., Lu, B., Carskadon, M. A., Sharkey, K., Stefanick, M., et al. (2013). Self-reported snoring and risk of cardiovascular disease among postmenopausal women (from the women's health initiative). *Am. J. Cardiol.* 111, 540–546. doi: 10.1016/j.amjcard.2012.10.039

Seidel, S., Frantal, S., Oberhofer, P., Bauer, T., Scheibel, N., Albert, F., et al. (2012). Morning headaches in snorers and their bed partners: a prospective diary study. *Cephalalgia* 32, 888–895. doi: 10.1177/0333102412453950

Skeldon, A. C., Derks, G., and Dijk, D. J. (2016). Modelling changes in sleep timing and duration across the lifespan: changes in circadian rhythmicity or sleep homeostasis? *Sleep Med. Rev.* 28, 96–107. doi: 10.1016/j.smrv.2015.05.011

Stevenson, M., Nunes, T., Heuer, C., Marshall, J., Sanchez, J., and Thornton, R. (2020). *epiR: Tools for the Analysis of Epidemiological Data*. Avaialble online at: https://CRAN.R-project.org/package=epiR

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12:e1001779. doi: 10.1371/journal.pmed.1001779

Sun, X. H., Ma, T., Yao, S., Chen, Z. K., Xu, W. D., Jiang, X. Y., et al. (2020). Associations of sleep quality and sleep duration with frailty and pre-frailty in an elderly population Rugao longevity and ageing study. *BMC Geriatr.* 20:9.

Tobaldini, E., Costantino, G., Solbiati, M., Cogliati, C., Kara, T., Nobili, L., et al. (2017). Sleep, sleep deprivation, autonomic nervous system and cardiovascular diseases. *Neurosci. Biobehav. Rev.* 74, 321–329. doi: 10.1016/j.neubiorev.2016.07.004

Weisberg, S., and Fox, J. (2018). *Cox Proportional-Hazards Regression for Survival Data in R. Appendix to an R Companion to Applied Regression,* 3rd Edn. Available online at: https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices/Appendix-Cox-Regression.pdf

Yin, J., Jin, X., Shan, Z., Li, S., Huang, H., Li, P., et al. (2017). Relationship of sleep duration with all-cause mortality and cardiovascular events: a systematic review and dose-response meta-analysis of prospective cohort studies. *J. Am. Heart. Assoc.* 6:e005947.

Zenin, A., Tsepilov, Y., Sharapov, S., Getmantsev, E., Menshikov, L. I., and Fedichev, P. O. (2019). Identification of 12 genetic loci associated with human healthspan. *Commun. Biol.* 2:41.

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership