BIOMEDICAL IMAGE OR GENOMIC DATA CHARACTERIZATION AND RADIOGENOMIC/IMAGE-OMICS

EDITED BY: Ming Fan, Jiangning Song and Zhaowen Qiu
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-83250-093-4
DOI 10.3389/978-2-83250-093-4

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

BIOMEDICAL IMAGE OR GENOMIC DATA CHARACTERIZATION AND RADIOGENOMIC/IMAGE-OMICS

Topic Editors:
Ming Fan, Hangzhou Dianzi University, China
Jiangning Song, Monash University, Australia
Zhaowen Qiu, Northeast Forestry University, China

Citation: Fan, M., Song, J., Qiu, Z., eds. (2022). Biomedical Image or Genomic Data Characterization and Radiogenomic/Image-omics.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-093-4

Table of Contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 17 August 2022
DOI 10.3389/fgene.2022.994880

OPEN ACCESS

EDITED AND REVIEWED BY
Jared C. Roach,
Institute for Systems Biology (ISB),
United States

*CORRESPONDENCE
Ming Fan,
ming.fan@hdu.edu.cn

SPECIALTY SECTION
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

RECEIVED 15 July 2022
ACCEPTED 22 July 2022
PUBLISHED 17 August 2022

CITATION
Fan M, Song J and Qiu Z (2022), Editorial: Biomedical image or genomic data characterization and radiogenomic/image-omics. *Front. Genet.* 13:994880. doi: 10.3389/fgene.2022.994880

Editorial: Biomedical image or genomic data characterization and radiogenomic/image-omics

Ming Fan[1]*, Jiangning Song[2] and Zhaowen Qiu[3]

[1]Institute of Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou, China, [2]Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Melbourne, VIC, Australia, [3]Institute of information Computer Engineering, Northeast Forestry University, Harbin, China

KEYWORDS

radiomics, radiogenomics, image-omics, medical imaging, feature analysis, precision medicine, biomarker identification

Editorial on the Research Topic
Biomedical image or genomic data characterization and radiogenomic/image-omics

Precision medicine has emerged as a practical solution for disease care thanks to advances in high-throughput data generation and analysis. Much of the emphasis in discussions about precision medicine or personalized medicine has been focused on the molecular characterization of tissues. However, as genetics differ between and within tumors and are quite heterogeneous, molecular characterizations are limited. Furthermore, there is no easy methodology yet to unravel why tumors with similar characteristics respond differently to a targeted therapy.

Imaging is relatively noninvasive and is often used in routine clinical practice for disease diagnosis, treatment, and prognosis. Medical imaging can provide a comprehensive view of entire tumor lesions; it is commonly used in clinical practice to monitor the progress of the cancer during treatment. The imaging includes but is not limited to ultrasound, X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET).

Radiomics refers to the conversion of images to high-dimensional data and subsequent mining for the characterization of biology and ultimately to improve disease management for patients. Radiogenomics investigates relationships between imaging features and genomics, which represents the correlation between the anatomical-histological level and the genomic level.

With advanced artificial intelligence methods, especially deep learning, data processing, feature extraction and data integration, medical image- or genomic data-based precision medicine has been greatly improved. There are 15 papers in this Research Topic: "*Biomedical image or genomic data characterization and radiogenomic/image-omics*." The articles focus on machine learning methods-based biomarker identification

from genomics or biomedical imaging to predict disease diagnosis, treatment, and prognosis.

For genomic-based biomarkers in precision medicine, we include nine papers focused on identifying molecular signatures by proposing machine learning algorithms in precise disease diagnosis and treatment management. Machine learning methods, such as gene–gene interactions and classification/regression models, have been developed to identify diagnostic/prognosis biomarkers.

We present one paper on building gene signatures in cancer prognostic analysis. Liang et al. established a ferroptosis-related gene-based prognostic model to investigate the prognosis of lung adenocarcinoma. Seven ferroptosis-related genes (FRGs) with prognostic significance were identified for dividing patients with lung adenocarcinoma into high-risk and low-risk groups. The results demonstrate the prognostic significance of FRGs in patients with lung adenocarcinoma, which may regulate tumor progression through a variety of pathways.

We also present two papers that employ a gene–gene interaction-based machine learning algorithm in predicting disease statuses. Cope et al. proposed a machine learning algorithm that use large amounts of data to find gene–gene interactions that they showed outperformed a polygenic risk score for predicting Parkinson's disease status. This work advances the state of art in prediction of susceptibility to complex traits or diseases.

Liu et al. established oncogene Aurora kinase A (AURKA)-related gene signatures for predicting the prognosis of patients with hepatocellular carcinoma (HCC) by a protein–protein interaction network analysis. Eight AURKA-related genes were thus identified that can effectively stratify the risk of HCC patients with differing survival rates. Additionally, patients in the high-risk group showed a higher percentage of immune cell infiltration and higher immune checkpoints. The identified gene signatures can be used as a candidate prognostic marker and therapeutic target in patients with HCC.

We also present one paper that analyzes gene pathways in predicting treatment response. Dai et al. investigated the roles of the Keap1-Nrf2 and PI3K-Akt pathways in esophageal squamous cell carcinoma (ESCC) treated with chemoradiotherapy. The results demonstrate that patients with dysregulated PI3K-Akt pathway exhibit a better survival outcome than patients with an intact PI3K-Akt pathway. This study highlighted the prognostic implications of aberrant cancer pathways in ESCC patients, which may be valuable in guidance of chemoradiotherapy management and treatment-induced toxicity.

We include one paper identifying lncRNA biomarkers for cancer survival analysis. Li et al. explored the biological functions and prognostic significance of epithelial-mesenchymal transition (EMT)-related lncRNAs in patients with colorectal cancer (CRC). A clinical factors and risk signature-based predictive nomogram was established for survival analysis. This signature was verified by predicting the immune-related phenotype and was found to be associated with immune cell infiltration and tumor mutation burden. This study indicated the clinical significance of the identified 11-EMT-lncRNA signature in predicting survival and immunotherapeutic response in CRC.

We also include papers analyzing immune subtypes in disease management. Yu et al. identified three immune-related subgroups for predicting immune checkpoint blockade (ICB) therapy response in lung adenocarcinoma (LUAD). The immune subgroup with higher infiltration scores exhibited a good response to IBC therapy and a better survival, whereas the subgroup with lower scores for immune checkpoint-related genes but higher infiltration scores for suppressive immune cells is more likey to be resistance to ICB therapy and have a poor prognosis. The identified immune subgroup can be promising in preoperatively discriminating LUAD patients with differing ICB therapy responses for a better guidance in treatment management.

Zhang et al. investigated the clinical implications of biglycan in gastric cancer prognosis. They identified biglycan-related differentially expressed genes (DEGs) by comparing the expression of biglycan in gastric cancer and normal tissues. The differential expression was verified through real-time PCR and immunohistochemistry. The constructed nomogram can accurately predict the survival outcomes of patients with gastric cancer. This study demonstrates that biglycan may be important in the occurrence and progression of gastric cancer.

Chen et al. identified human myometrial transcriptome and established the Competing endogenous RNA (ceRNA) regulatory network depending on labor duration. This study highlights the roles of dynamic changes that occur at ceRNAs during parturition in functional changes in human myometrium at labor.

We also included one study that used multiomics biomarkers in disease prognosis analysis. Xiao et al. aims to reveal the prognostic nontumor cell landscape in glioblastoma niches by a multiomics analysis. The biomarkers of nonmalignant cells in the microenvironment of glioblastoma multiforme (GBM) were identified, which separate patients into negative or positive immune response clusters with significantly different survival rates. Negative immune response markers were particularly enriched.

We included six radiological image-based studies in predicting tumor status, survival outcomes, metastases and treatment response. Quantitative mining of data from radiological images, including MRI, ultrasound and CT, was performed, with applications in precise disease diagnosis and prognosis analyses.

We include one paper using radiomics extracted from ultrasound for the diagnosis of ovarian neoplasms. Qi et al. established a nomogram integrating ultrasound-based radiomics signatures and clinical factors, named combined clinical-radiomics (CCR), to discriminate between benign, borderline, and malignant serous ovarian tumors. This CCR-based model

shows better prediction performance than a clinical factor-based model.

We present two papers on disease treatment response prediction based on radiomic signatures. Xia et al. developed and validated a nomogram integrating radiomics, MRI findings, and clinicopathological factors to predict survival in triple-negative breast cancer patients treated with neoadjuvant chemotherapy. The proposed signatures significantly stratified patients into high- and low-risk groups with different survival rates. These signatures were further validated in an external validation group. Three indicators, including the multifocal/centric disease status, pathological complete response status, and Rad-score, were independently associated with survival. The results demonstrated that the integrated signature-based nomogram improved the accuracy of survival prediction.

Wang et al. identified MRI features for predicting antipsychotic medication treatment outcomes in schizophrenia. To this end, nine categories of MRI measures and the polygenic risk score (PRS) were combined to separate the responders and nonresponders. The results showed that the PRS was better in prediction performance than measures of cortical thickness, cortical volume, and surface sulcal depth but lower than GMV, ALFF, and surface curvature.

We include one paper on brain metastasis prediction using imaging features derived from CT. Wang et al. identified pretreatment thoracic CT biomarkers for predicting brain metastases in patients with ALK-rearranged non-small cell lung cancer (NSCLC). A machine learning method was proposed to identify the radiomic features extracted from pretreatment thoracic CT images, which achieved good performance in predicting brain metastases within 1 year after detection of the primary tumor.

Wang and Li performed a unified framework for estimating inattention, which is one of the most useful clinical symptoms in attention deficit hyperactivity disorder (ADHD). To improve the classical brain-behavior models, the phase synchrony features were identified from resting state functional MRI (fMRI) using a machine learning method. Among the brain networks, the bilateral subcortical-cerebellum networks exhibit the most predictive phase synchrony patterns for inattention estimation.

We thank the authors and the reviewers for their contribution to this Research Topic. This Research Topic of articles may serve as an inspiring compendium for future research in biomedical imaging or genomic data characterization and radiogenomic/image-omics.

Author contributions

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

Conflict of interest

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

Publisher's note

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 September 2021
doi: 10.3389/fgene.2021.741325

Multiomics Analysis Reveals the Prognostic Non-tumor Cell Landscape in Glioblastoma Niches

Zixuan Xiao[1†], Wei Zhang[2,3†], Guanzhang Li[2,3†], Wendong Li[1], Lin Li[1], Ting Sun[1], Yufei He[1], Guang Liu[1], Lu Wang[1], Xiaohan Han[1], Hao Wen[1], Yong Liu[1], Yifan Chen[1], Haoyu Wang[1], Jing Li[1], Yubo Fan[1] and Jing Zhang[1*]*

[1]Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Centre for Biomedical Engineering, School of Engineering Medicine, School of Biological Science and Medical Engineering, Beihang University, Beijing, China, [2]Department of Molecular Neuropathology, Beijing Neurosurgical Institute, Capital Medical University, Beijing, China, [3]Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing, China

A comprehensive characterization of non-tumor cells in the niches of primary glioblastoma is not fully established yet. This study aims to present an overview of non-malignant cells in the complex microenvironment of glioblastoma with detailed characterizations of their prognostic effects. We curate 540 gene signatures covering a total of 64 non-tumor cell types. Cell type-specific expression patterns are interrogated by normalized enrichment score across four large gene expression profiling cohorts of glioblastoma with a total number of 967 cases. The glioblastoma multiforms (GBMs) in each cohort are hierarchically clustered into negative or positive immune response classes with significantly different overall survival. Our results show that astrocytes, macrophages, monocytes, NKTs, and MSC are risk factors, while CD8 T cells, CD8 naive T cells, and plasma cells are protective factors. Moreover, we find that the immune system and organogenesis are uniformly enriched in negative immune response clusters, in contrast to the enrichment of nervous system in positive immune response clusters. Mesenchymal differentiation is also observed in the negative immune response clusters. High enrichment status of macrophages in negative immune response clusters is independently validated by analyzing scRNA-seq data from eight high-grade gliomas, revealing that negative immune response samples comprised 46.63 to 55.12% of macrophages, whereas positive immune response samples comprised only 1.70 to 8.12%, with IHC staining of samples from six short-term and six long-term survivors of GBMs confirming the results.

Keywords: glioblastoma, tumor microenvironment, immunology, prognosis, tumor-infiltrating cells

HIGHLIGHTS

1) A comprehensive characterization of non-tumor cells in the niches of primary glioblastoma.
2) Astrocytes, macrophages, monocytes, NKTs, and MSC are risk factors, while CD8 T cells, CD8 naive T cells, and plasma cells are protective factors.
3) Mesenchymal differentiation is observed in the negative immune response clusters.
4) High enrichment status of macrophages is in negative immune response clusters of glioblastomas.

OPEN ACCESS

Edited by:
Ming Fan,
Hangzhou Dianzi University, China

Reviewed by:
Beibei Xin,
China Agricultural University, China
Y. Penghui,
People's Liberation Army General Hospital, China
Xiangyu Liu,
Shenzhen University, China

****Correspondence:***
Jing Zhang
jz2716@126.com;
jz2716@buaa.edu.cn
Yubo Fan
yubofan@buaa.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: *14 July 2021*
Accepted: *11 August 2021*
Published: *16 September 2021*

Citation:
Xiao Z, Zhang W, Li G, Li W, Li L, Sun T, He Y, Liu G, Wang L, Han X, Wen H, Liu Y, Chen Y, Wang H, Li J, Fan Y and Zhang J (2021) Multiomics Analysis Reveals the Prognostic Non-tumor Cell Landscape in Glioblastoma Niches. Front. Genet. 12:741325. doi: 10.3389/fgene.2021.741325

INTRODUCTION

Gliomas account for 70% of all brain tumors (Ohgaki and Kleihues, 2005) and are categorized into four types: Grade I pilocytic astrocytoma and grade II astrocytoma are low-grade gliomas, whereas grade III anaplastic astrocytoma and grade IV glioblastoma multiform (GBM) are malignant tumors (Kleihues et al., 1993). The GBMs have poor prognosis with a median survival rate of 1 year after diagnosis and a 2-year survival rate of only 12.7 to 19.8% according to the SEER database.

Categorization of gliomas previously focused on histological features (Bailey and Cushing, 1927); however, characterization methods have shifted toward high-resolution molecular profiling, including identification of isocitrate dehydrogenase (IDH) mutation, co-deletion of chromosomal arms, O6-methylguanine-DNA methyltransferase (*MGMT*) promoter methylation, and miR-181d expression (Jiang et al., 2016). Additionally, new stratifications have been proposed using gene expression profiles or specific gene mutations (Phillips et al., 2006; Ceccarelli et al., 2016), methylation status (Hegi et al., 2005; Shah et al., 2011), and the presence of neoantigens (Zhang et al., 2019; Sun et al., 2021). Numerous studies have focused on interpreting the RNA-seq profiles of gliomas in an attempt to elucidate their dynamics and mechanisms, with studies on recurrent glioblastoma able to distinguish comprehensive transcriptome profiling in the malignant progression of human gliomas (Zhao et al., 2017) and find critical clues of *MET*-related mutations (Hu et al., 2018) and oncogenic fusions (Bao et al., 2014). The findings of these studies have markedly advanced the investigation of GBM and facilitated prognostic and therapeutic developments, but the highly heterogeneous nature of GBM still often leads to the failure of extensive treatment regimens.

The complexity of GBM components and the immune microenvironment has attracted significant attention in recent years, with categorizations based on molecular profiling revealing tissue similarities between proneural, proliferative, and mesenchymal-type gliomas, respectively (Phillips et al., 2006). Certain immune components, such as tumor-associated macrophages (TAMs), have been identified as regulators of the proneural-to-mesenchymal transition (Bhat et al., 2013) and contributors to immunosuppression (Gabrilovich, 2017), thus leading to poor prognosis. However, a comprehensive characterization of non-tumor cells in the niches of primary glioblastoma has not been fully established. Investigations into the tumor components and immune microenvironment would help unravel the cross-talk between the immune system and cancer cells and allow determination of therapeutic targets for the development of novel cancer treatments.

In this study, we generated a comprehensive non-tumor cell landscape in the microenvironment of GBM by integrating four large-scale gene expression profiling data cohorts of primary glioblastoma with gene signatures covering a total of 64 non-tumor cell types. The GBMs in each cohort are hierarchically clustered into negative or positive immune response classes with significantly different overall survival. Additionally, we investigated the risk levels associated with immune cell types and the enrichment of Gene Ontology (GO) terms. In particular, we confirmed enrichment of a negative prognostic factor (macrophages) in scRNA-seq data of high-grade gliomas and in samples from GBM patients exhibiting short-term survival by immunohistochemical (IHC) staining.

MATERIALS AND METHODS

Gene Expression and Clinical Data

Four cohorts of gene expression profiles of GBM tumor tissues were collected from public domains including Cohort 1 (Wang et al., 2016; Zhang et al., 2019), Cohort 2 (TCGA; RNA sequences; Cancer Genome Atlas Research, 2008), Cohort 3 (REMBRANDT, mRNA microarray; Gusev et al., 2018), and Cohort 4 (TCGA, mRNA microarray; Brennan et al., 2013), respectively. Samples that were not diagnosed as GBM or did not include complete gene expression or clinical data were removed, resulting in 75, 152, 181, and 559 samples in Cohorts 1, 2, 3, and 4, respectively. The single-cell RNAseq data of eight HGGs can be accessed through Gene Expression Omnibus (accession: GSE103224; Yuan et al., 2018). Tumor samples were obtained from 12 glioblastomas, including from six short-term-survival and six long-term-survival patients. All research protocols and ethics comply with the Declaration of Helsinki. Sample collection and data analyses were approved by the Beijing Tiantan Hospital institutional review board (KY 2020–093-02), and written informed consent was obtained from each participant.

Gene Signatures of Immune Cells

Gene signatures ($n = 540$) covering 64 cell types were collected from multiple sources (Bindea et al., 2013; Rooney et al., 2015; Tirosh et al., 2016; Aran et al., 2017; Charoentong et al., 2017). The 64 cell types were further categorized into five groups: hematopoietic stem cells (HSCs) and hematopoietic cells (lymphoid and myeloid lineage), stromal cells, and others, as shown in **Supplementary Material 1A,B**.

Generating a Normalized Enrichment Score for Estimating Cell-Enrichment Status

An normalized enrichment score (NES) for the Mann–Whitney–Wilcoxon gene set test was adapted to evaluate the enrichment status of cells (Frattini et al., 2018). The NES was determined as follows:

$$NES = 1 - \frac{U}{mn}$$

$$U = nm + \frac{m(m+1)}{2} - T$$

where m is the number of genes in a gene set, n is the number of genes outside the gene set, and T is the sum of the ranks of the genes in the gene set (Zhang et al., 2019).

Given a gene signature, the gene expression data of a glioblastoma tumor sample were separated into two sections comprising genes expressed in the gene signature and the rest of the genes, respectively. The Wilcoxon rank-sum test was then applied to calculate the NES. For each cell signature, the NES value was calculated to quantify the probability that the expression of a gene in the gene signature was greater than the expression of a gene outside of the gene signature. The higher the NES value, the more likely that the cell is enriched in the tumor sample.

Risk Level for Gene Signatures

Cox regression (proportional hazards regression) in the R was applied for every gene signature in each cohort. The protective factor was defined when the hazard ratio of a gene signature was <1, and the risk factor was defined when this was >1. Signatures with a $p \leq 0.05$ were defined as significantly associated with survival (addressed as prognostic signatures below), with only prognostic signatures used for further analysis. If all prognostic signatures of one cell type were either protective or risk factors, they were defined as consistent factors, otherwise, inconsistent factors.

Stratification of Glioblastoma Patients

Hierarchical clustering of GBMs was applied to z-score transformed NESs of these signatures using R. Euclidean distance and complete method were used for clustering, and heat maps were drawn using the R: "pheatmap." Kaplan–Meier survival analysis was performed using R: "survival" and "survminer."

Go Enrichment Analysis

Gene Set Enrichment Analysis (GSEA; Subramanian et al., 2005) was performed upon negative and positive immune response clusters using a total of 6,166 GO terms from the Molecular Signatures Database (MSigDB; Liberzon et al., 2011), including cellular component (cc), molecular function (mf), and biological process (bp), followed by visualization through cytoscape (Shannon et al., 2003). The results are shown in **Supplementary Material 2A-D**.

Identification of Non-Transformed Cells From scRNA-Seq Data

For scRNA-seq data, genes expressed in less than or equal to 10 cells were eliminated, followed by a moving average method (Chung et al., 2017) to determine chromosome expression patterns. The number of original molecules per cell was converted to $\log_2(\text{cpm} + 1)$. The moving average used 100 gene lengths as the window, and the value for the gene in the center of the window was considered the average expression of the window. We used the Seurat package (v.3.0; Butler et al., 2018; Stuart et al., 2019) to analyze the screened data according to standard procedures. Amplification of chromosome 7 and loss of chromosome 10 were used to differentiate malignant (transformed) cells from non-malignant (non-transformed) cells (Weller et al., 2015).

Determination of Non-Transformed Cell Types

Scibet (Li et al., 2020) was used to predict the identities of the non-transformed cells in the scRNA-seq data. The trained model "30 major human cell types,"[1] including 30 major human cell types from 42 scRNA-seq datasets, served as the reference for cell type identification.

Stratification of Single-Cell Gene Expression Samples

To determine whether a sample in the scRNA-seq data was positive or negative immune response, Spearman correlation analysis was applied between the sample in the scRNA-seq cohort and the samples in the four gene expression profiling cohorts, respectively. Only positive correlations were retained, and the mean value of the correlation coefficients in each cohort was calculated. The fold change for a sample in the scRNA-seq data was calculated as the mean correlation coefficient of the sample in the scRNA-seq data involving samples in the positive immune response clusters divided by the mean correlation coefficients of the sample in the scRNA-seq data involving samples in the negative immune response clusters. The fold changes in the correlation coefficients calculated for the four cohorts were multiplied to determine the total fold change. A total fold change >1 indicated that the Spearman correlation coefficient was higher in the positive immune response clusters, and thus, the sample in the scRNA-seq data was determined as positive immune response; otherwise, it was designated as negative immune response (**Supplementary Material 3**).

IHC Staining for Macrophage Markers

Tumor samples used for IHC staining were obtained from 12 GBMs, including six short-term-survival and six long-term-survival patients. The surgically removed tumor tissues were stored in formalin immediately after excision and embedded in paraffin within 3 days. IHC staining and image capture were performed as previously described (Hu et al., 2018). The primary antibody for the detection of macrophage marker MS4A4A was obtained from Sigma-Aldrich (HPA029323; St. Louis, MO, United States), with staining was performed according to manufacturer instructions. The proportion of positive cells was counted using ImageJ software (v.1.52; National Institutes of Health, Bethesda, MD, United States). Clinical information and IHC staining results are summarized in **Supplementary Material 4**.

Statistical Analysis

Values of p for NES distributions in negative immune response and positive immune response clusters were calculated using Student's t-test, and those for IHC staining percentages were generated from the Wilcoxon test. All analyses were conducted in R. Values of $p \leq 0.05$ were determined as statistical significance.

[1] http://scibet.cancer-pku.cn/download_references.html

RESULTS

Stratification of Glioblastomas Based on Cell Type-Specific Enrichment Status

Based on a total of 540 gene signatures covering 64 cell types (**Supplementary Material 1A**), we applied the NES algorithm we previously developed (Frattini et al., 2018) to determine the enrichment status of each cell type, followed by filtering the gene signatures with enrichment status correlated with overall survival (prognostic signatures). The workflow for stratifying samples is shown in **Figure 1A**. Unsupervised hierarchical clustering stratified samples into two significantly different prognostic clusters among the four cohorts ($p = 0.025$, $p = 0.015$, $p = 0.0004$, and $p = 0.00056$ for cohort 1–4, respectively; **Figures 1B–E; Supplementary Figures 1A–D**; **Table 1**). Clusters with patients exhibiting long-term overall survival were found universally enriched with CD8 T cells, whereas short-term overall survival clusters were characterized by enrichment of "stromal cells," such as mesenchymal stem cells (MSCs). Therefore, we designated the long- and short-term overall survival clusters as positive and negative immune response, respectively. Additionally, we discovered that the enrichment status calculated from different gene signatures exhibited similar and stable trends for CD8 naïve T cells, common lymphoid progenitors (CLPs), epithelial cells, HSCs, lymphoid endothelial cells, neurons, natural killer T cells (NKTs), and γΔT cells (**Figure 1F**).

The Predicted Risk and Protective Landscape of Non-Tumor Cells in the Glioblastoma Microenvironment

To understand the prognostic effect of different cell types, we estimated associations between the enrichment status of gene signatures and overall survival through Cox regression analysis across four gene expression profiling cohorts. In each cohort, statistically significant gene signatures with a hazard

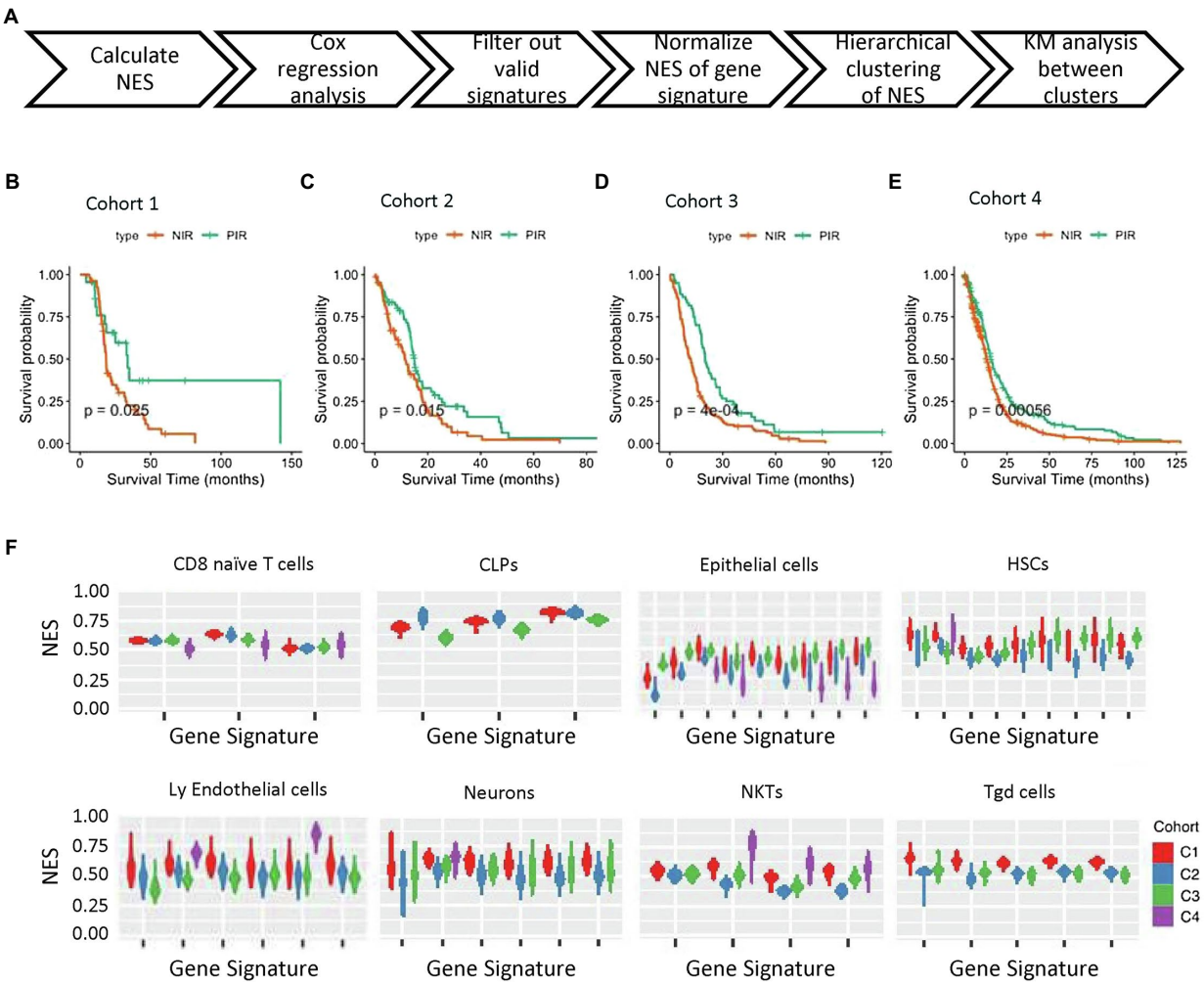

FIGURE 1 | NES-based stratification of patient survival. **(A)** Workflow of NES-based stratification and validation of survival time. **(B–E)** Kaplan–Meier survival curves of the NIR and PIR clusters in the four cohorts (PIR, orange; NIR, green). **(F)** NES distribution of four gene expression profiling cohorts of tumor tissues from GBM patients. Cell types and cohorts are noted. NES, normalized enrichment score; NIR, negative immune response; and PIR, positive immune response.

TABLE 1 | Hierarchical clustering results for the four cohorts.

Cohort	Sample	Signature	Cell type	PIR	NIR	*p*	Data source (References)
1	75	31	18	22	53	0.02499	Wang et al., 2016; Zhang et al., 2019
2	152	51	24	67	85	0.01462	Cancer Genome Atlas Research, 2008
3	181	57	24	60	121	0.0004	Gusev et al., 2018
4	559	138	46	198	361	0.00056	Brennan et al., 2013

NIR, negative immune response; PIR, positive immune response.

ratio > 1 or < 1 were defined as risk or protective factors, respectively. We found that risk effects consistently agreed with statistically significant gene signatures for given cell types, including activated dendritic cells (aDCs), astrocytes, class-switched memory (CSM) B cells, epithelial cells, fibroblasts, macrophages, M2 macrophages, monocytes, MSCs, NKTs, and plasmacytoid (p)DCs. By contrast, CD8 naïve T cells, CD8 T cells, endothelial cells, eosinophils, megakaryocyte–erythroid progenitor cells, plasma cells, and regulatory T cells (Tregs) were consistently estimated as being protective. Additionally, basophils, B cells, CD8 central memory T cells, mast cells, multi-potent progenitor cells, memory B cells, naïve B cell, and T helper 1 (Th1) cells were predicted as being protective according to majority of gene signatures across the four cohorts, whereas CD4 central memory T cells, mesangial cells, and pericytes were predicted as a risk by most of the gene signatures. Interestingly, the risk and protective effects of CD8 effector memory T cells, DCs, myocytes, and NK cells were inconsistent according to the different gene signatures (**Figure 2A**).

Notably, we identified inconsistencies in some estimated risk or protective effects predicted by the gene signatures across the four cohorts. The prognostic effects of enrichment status estimated from one gene signature for basophils, B cells, pericytes, and Th1 cells were inconsistent among the four cohorts (**Figures 2A,B**); however, basophils, B cells, and pericytes were more likely to manifest an enrichment-dependent effect on survival time, with basophils and B cells being protective when highly enriched and pericytes presenting a risk when highly enriched.

Statistically significant signatures showed consistency across risk levels valued from different perspective, i.e., risk level NES distribution, risk factor hazard ratio, and occurrence cohort count. **Figure 2C** shows the hazard ratios for cell types demonstrating consistent agreement in their prognostic effects across all corresponding signatures in at least two cohorts. MSCs, pDCs, CSM B cells, and CLPs were consistent risk factors with relatively high hazard ratios in at least two cohorts. Conversely, common myeloid progenitors, CD4 naïve T cells, plasma cells, and CD4 T cells showed hazard ratios <1, suggesting potentially strong protective effects (**Figure 2C**). **Figure 2D** shows the group count of consistent risk levels. Astrocytes, MSCs, monocytes, pDCs, NKTs, macrophages, M2 macrophages, fibroblasts, epithelial cells, CSM B cells, and aDCs were consistent risk factors appearing in at least two cohorts, with astrocytes being significantly negatively correlated with overall survival in all four cohorts. CD8 T cells, Tregs, plasma cells, MEPs, eosinophils, endothelial cells, and CD8 naïve T cells were also consistent risk factors, with CD8 T cells most frequently identified in three cohorts; however, for risk factors identified in only two cohorts (i.e., Tregs), more evidence is needed to support these findings.

Identification of Immune Dysregulation in the Negative Immune Response Cluster

We then performed GSEA for the four cohorts. Enrichment map analysis of dysregulated GO terms revealed that those related to the immune system, metabolism, and organogenesis were highly enriched in all four cohorts (**Figure 3A; Supplementary Figures 2A–C; Supplementary Material 2**). Specifically, GO terms related to the immune system (defense response, cytokines, myeloid lineage, and lymphoid lineage cell regulation) were enriched in negative immune response clusters, suggesting uniform dysregulation of the immune response in negative immune response clusters. Interferon (IFN)-related GO terms were significantly enriched in the negative immune response group (**Figure 3B**), consistent with constitutive type I IFNs (IFN-α and IFN-β) facilitating glioma-related immune escape (Silginer et al., 2017), unfavorable prognosis, chemotherapy resistance, and more aggressive immune response (Zhu et al., 2019).

Activities associated with several interleukins (ILs), including IL-6, IL-8, and IL-10, were enriched in negative immune response clusters (**Figure 3C**), with IL-8 expression negatively correlated with GBMs survival and positively correlated with the expression of genes associated with the glioblastoma-initiating cell phenotype, as well as the possibility of GBM recurrence (Hasan et al., 2019). Additionally, IL-1β contributes to cancer cell stemness, invasiveness, and drug resistance in glioblastoma (Wang et al., 2012; Yeung et al., 2013).

Moreover, we identified macrophage activation, differentiation, and chemotaxis as enriched activities in negative immune response clusters (**Figure 3D**), consistent with identification of macrophages as risk factors. Downregulation of major histocompatibility complex (MHC)-I and -II molecules is associated with glioma migration and invasion (Zagzag et al., 2005), with their altered expression associated with the negative immune response cluster (**Figure 3E**).

Majority of nervous system-associated GO terms (nervous system organogenesis in G1, nervous system organogenesis, neural function and synaptic in G2, and nervous system organogenesis in G4) was enriched in the positive immune response cluster (**Figure 3A; Supplementary Figures 2B,C**), demonstrating that regulation of the nervous system was a shared feature in the positive immune response cluster. This agrees with the proneural subtype of gliomas categorized by molecular profiling, in that this subtype usually demonstrated tissue similarity with adult and fetal brain and biological processes related to

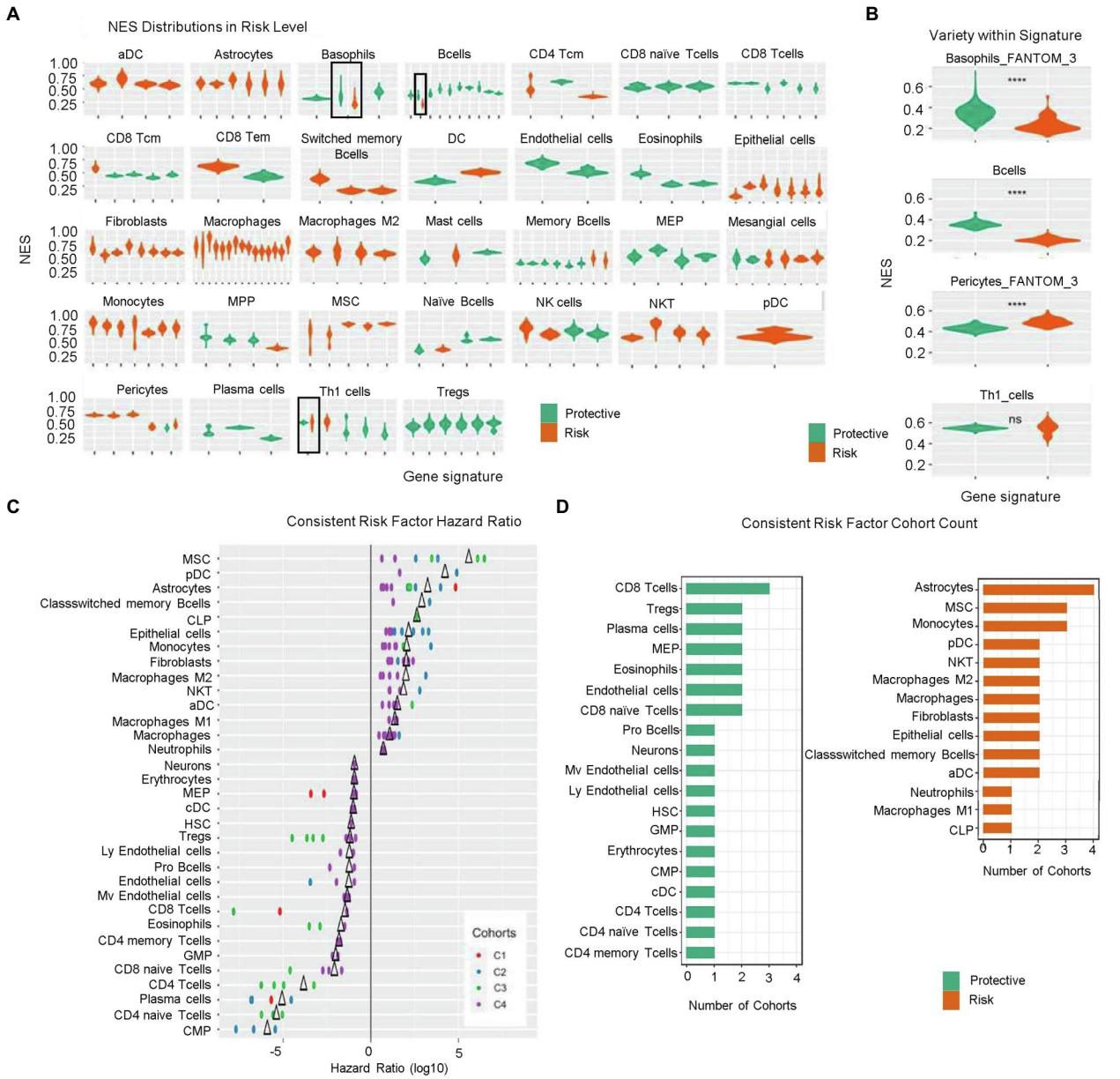

FIGURE 2 | Risk levels according to calculated NESs. **(A)** NES distribution of prognostic signatures as denoted by risk levels (risk factors, orange; protective factors, green). **(B)** Variety of NES distribution within signatures. ns, $p > 0.05$; and ****$p \leq 0.0001$. **(C)** Hazard ratio of consistent risk factors (>1, risk factor; <1, protective factor; and Δ, mean value). **(D)** Group count of consistent risk factors. NES, normalized enrichment score; NIR, negative immune response; ns, not significant; and PIR, positive immune response.

neurogenesis (Phillips et al., 2006). Additionally, this glioma subtype is regarded as less malignant relative to other subtypes (e.g., proliferative and mesenchymal; Phillips et al., 2006).

Mesenchymal Differentiation Characterized in the Negative Immune Response Cluster

Gliomas of the mesenchymal subtype are defined by high expression of chitinase 3-like 1 and MET5, as well as a high frequency of neurofibromatosis type 1 (*NF1*) mutation/deletion and low levels of *NF1* mRNA (Verhaak et al., 2010). The negative immune response clusters defined by cell-enrichment analysis shared an obvious similarity with this glioma subtype. We discovered that five stromal cell types (fibroblasts, pericytes, MSC, mesangial cells, and endothelial cells) exhibited a significantly higher NES value in the negative immune response cluster than in the positive immune response cluster in at least three cohorts (**Figures 4A–E**). Of note, negative immune response clusters with endothelial cells showed higher NESs

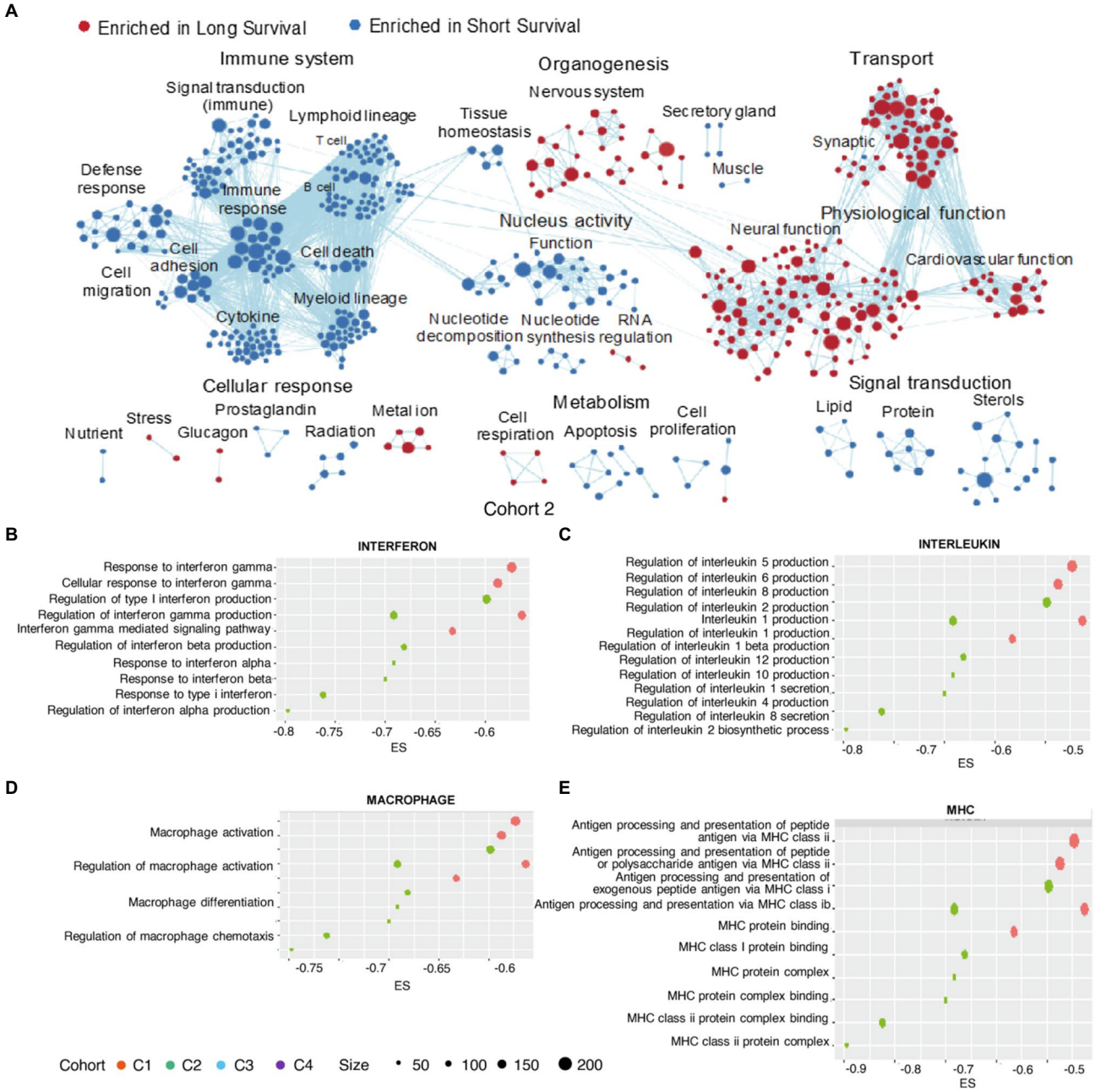

FIGURE 3 | GO enrichment in clusters. Enrichment map of GO terms (selected according to $p < 0.05$) aggregated by functions for Cohort 2 **(A)**. Enrichments scores for GO terms (selected according to $p < 0.05$) associated with the immune-related **(B)** interferons, **(C)** interleukins, **(D)** macrophages, and **(E)** major histocompatibility complex (MHCs). GO, Gene Ontology; MHC, major histocompatibility complex.

in three cohorts but distributed between two different signatures (**Supplementary Figure 2D**). Lymphoid endothelial cells showed higher negative immune response enrichment in one cohort, with no significant differences observed in other cohorts. These results supported tissue similarities between negative immune response clusters and the mesenchymal subtype.

Furthermore, we identified aspects related to mesenchymal differentiation in negative immune response clusters, with enrichment of activities related to tumor necrosis factor (TNF)-α and nuclear factor-kappaB (NF-κB) identified from three cohorts and all four cohorts (**Figures 4F,G**), respectively. Previous studies of glioma sphere cultures indicated that TNF-α promotes mouse embryonic stem cell differentiation accompanied by increased resistance to radiotherapy in an NF-κB-dependent manner (Bhat et al., 2013). Macrophages are also an important source of TNF-α secretion.

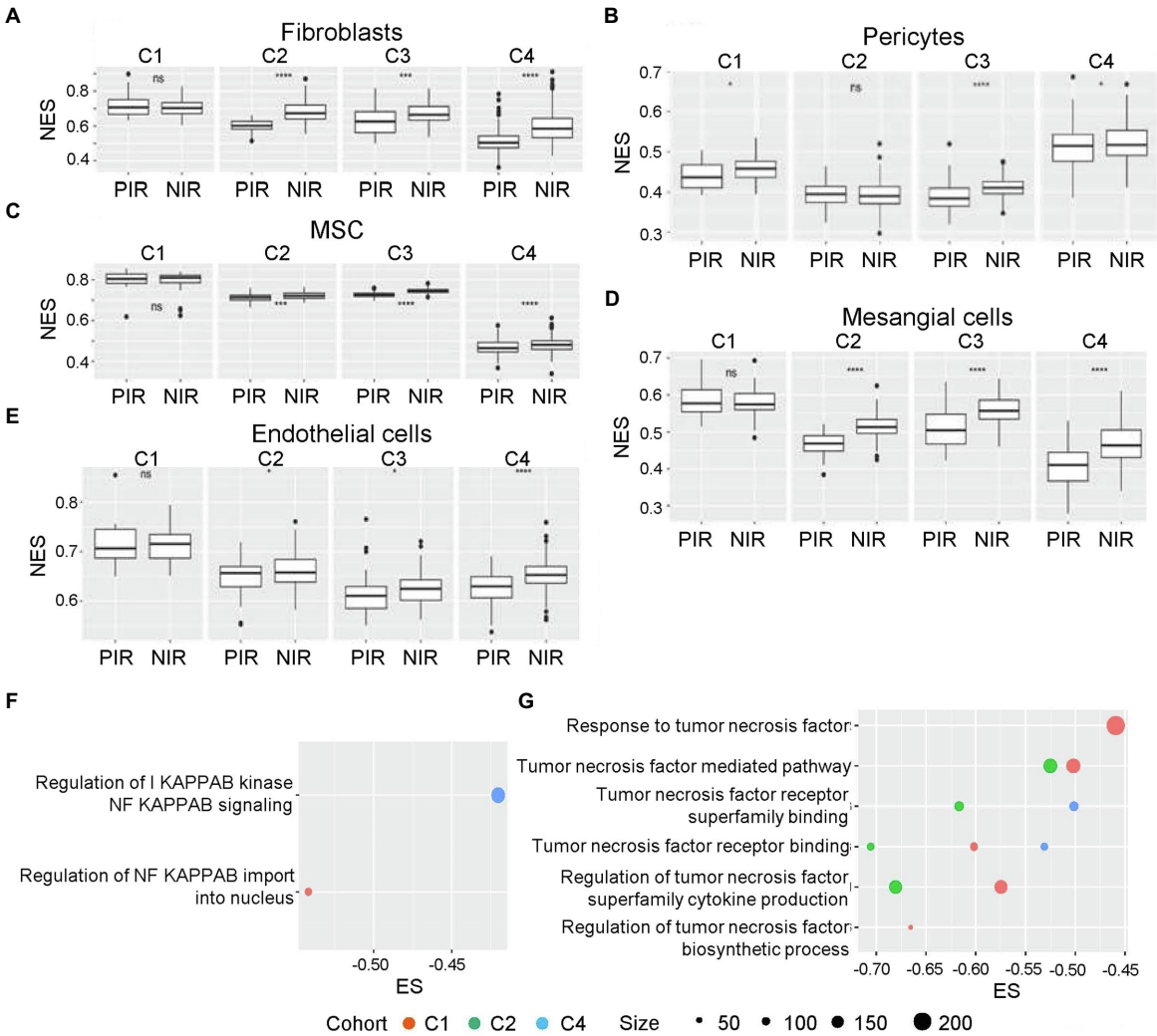

FIGURE 4 | NES distributions in NIR and PIR clusters. **(A)** Fibroblasts, **(B)** pericytes, **(C)** mesenchymal stem cells (MSCs), **(D)** mesangial cells, and **(E)** endothelial cells. ns, $p > 0.05$; *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$; and ****$p \leq 0.0001$. Enrichment scores for GO terms associated with the mesenchymal differentiation-related cytokines (selected according to a $p < 0.05$; **F**) NF-κB and **(G)** tumor necrosis factor (TNF)-α. GO, Gene Ontology; MSC, mesenchymal stem cell; NES, normalized enrichment score; NF-κB, nuclear factor-kappaB; NIR, negative immune response; PIR, positive immune response; and TNF-α, tumor necrosis factor-α.

scRNA-seq and IHC Confirmation of the Negative Prognostic Effects of TAMs

To validate our findings, we collected scRNA-seq data for cell-component analysis. We classified all eight samples with available scRNA-seq data into negative or positive immune response clusters by calculating NES-based Spearman similarity between single-cell samples and bulk tumor samples (**Supplementary Material 3**). The results identified samples PJ016, PJ017, PJ032, and PJ048 as negative immune response and PJ018, PJ025, PJ032, and PJ035 as positive immune response.

We applied Seurat and copy number variation analyses to distinguish non-transformed cells from malignant transformed glioma cells in the scRNA-seq data. All HGGs, except PJ016, harbored clear amplification of chromosome 7 and loss of chromosome 10 (**Supplementary Figures 3A–H**), consistent with transformed tissues demonstrating large-scale copy number alterations and aneuploidies (Venteicher et al., 2017; Taylor et al., 2018), as well as glioblastoma often being accompanied with amplification of chromosome 7 and loss of chromosome 10 (Zagzag et al., 2005). PJ016 was found apparent loss of chromosomes 13 and 19, revealing that the cell population had indeed undergone transformation (Lee et al., 1995; Ritland et al., 1995; Nakamura et al., 2000).

The identities of non-transformed cells in the glioma microenvironment were then determined using Scibet (Li et al., 2020; **Figures 5A–H**). We found no immune cells in PJ016 or PJ048 (**Table 2**), possibly due to the heterogeneity of different sampling areas. Those with a high percentage of macrophages (PJ017 and PJ032; 46.63 and 55.12%, respectively) belonged to the negative immune response cluster (**Table 2**), whereas samples with fewer macrophages (PJ018, PJ025, and PJ035; 2.28, 1.70, and 8.12%, respectively) overlapped with the positive

immune response cluster (**Table 2**), confirming macrophage enrichment as a risk factor.

Moreover, we confirmed the negative prognosis associated with macrophages IHC staining for the macrophage marker MS4A4A in 12 glioblastoma samples, including six from short-term-survival and six from long-term-survival patients (**Figure 5K**; **Supplementary Material 4**). The short-term-survival samples showed a significantly higher percentage of MS4A4A-positive cells relative to the six long-term-survival samples ($p = 0.00051$; **Figures 5I,J**).

DISCUSSION

In this study, we generated a landscape of glioblastoma niches using four gene expression profiling cohorts of tumor tissues from GBMs based on the NES method. The patients in each cohort were divided into two categories (positive or negative immune response) according to hierarchical clustering analysis of cell type-based enrichment status and showing a significantly different survival ($p < 0.05$). The analysis revealed risk factors, including astrocytes, macrophages, monocytes, NKTs, and MSC, as well as protective factors, CD8 T cells, CD8 naive T cells, and plasma cells. Additionally, GSEA demonstrated that immune system- and organogenesis-related GO terms were uniformly enriched in negative immune response clusters, whereas positive immune response clusters were enriched in the nervous system. Moreover, significant signs of mesenchymal differentiation were observed in the negative immune response clusters, and validation using scRNA-seq analysis and IHC staining showed correlations between the presence of macrophages and negative immune response.

Potential mechanisms associated with specific cell types manifested consistent risk levels. Some cell types exhibited identical risk levels across the four cohorts and all gene expression signatures. Specifically, astrocytes were frequently observed as a consistent risk factor with a high hazard ratio. As an important component of the blood–brain barrier and the tripartite synaptic neural network, the normal physiological role of astrocytes involves promoting mutual communication with neurons. However, astrocytes can also develop into tumor cells and form astrocytomas. Given the heterogeneity of gliomas, the high frequency of astrocytes as a risk factor is explainable. Moreover, evidence suggests that tumor-reactive astrocytes can interact with glioma tumor cells and promote the development, invasion, and survival of gliomas by releasing different cytokines or regulating the entry and exit of calcium and hydrogen ions in cell channels (Guan et al., 2018).

NKTs were also a consistent risk factor. miR-92a was reported to induce immune tolerance of NKTs to glioma cells (Tang et al., 2014). Co-culture of glioma cells and NKTs showed miR-92a expressing in glioma cells played a key role in inducing the elevated expression of IL-6 and IL-10 in NKTs (Tang et al., 2014). In the present study, we found IL-6- and IL-10-related GO terms in the negative immune response cluster. Compared with NKTs cultured alone, the expression of antitumor molecules, including perforin, Fas ligand, and IFN-γ, was significantly reduced in NKTs co-cultured with glioma cells (Tang et al., 2014). Moreover, IL-6 + IL-10+ NKTs exhibit a weak ability to induce apoptosis in glioma cells but have an immunosuppressive effect on CD8 T cell activity (Tang et al., 2014).

CD8 T cells play defensive roles against cancer cells, consistent with the risk levels generated in the present analysis. Serologic analysis of antigens using recombinant cDNA expression cloning identified several tumor-associated antigens capable of generating a specific response in a variety of human cancers, including malignant glioma (Struss et al., 2001; Prins et al., 2003). Tumor-related antigens can be recognized by cytotoxic CD8 T cells in the context of tumors expressing MHC-I (Prins and Liau, 2003; Yang et al., 2004), suggesting that a T cell-dependent immune response might improve the outcome of glioma patients through an antigen-mediated immune response. This was supported by a clinical study of newly diagnosed glioblastoma patients that reported significantly attenuated CD8 T cell infiltration in samples from long-survival patients (>403 days) relative to that in samples from short-survival patients (<95 days; Yang et al., 2010). These findings agreed with those of the present study showing that CD8 T cells were categorized as a protective factor.

Some cell types exhibited inconsistent risk levels. In these cases, it is likely that other conditions caused a shift in risk levels (e.g., age, co-existence with other cells, or a combination of other clinical symptoms). Different signatures of the same cell type might display different risk levels, suggesting the impact of cell status. To further investigate this concept, a specific gene in each gene signature should be investigated. Other conditions, such as the presence of neoantigens (Zhang et al., 2019), IDH mutation(s) (Phillips et al., 2006; Parsons et al., 2008), and *MGMT* methylation (Shah et al., 2011), can also provide insight into conditions causing a shift in risk levels. Furthermore, the data used in this study were from primary gliomas; therefore, comparisons between recurrent and primary glioma samples would provide additional information concerning dynamics in the glioma microenvironment.

Myeloid lineage cells, such as monocytes and macrophages, were consistent risk factors in agreement with previously reported results (Hambardzumyan et al., 2016). These cells (i.e., TAMs) account for more than 30% of the total number of solid tumor cells (Boussiotis and Charest, 2018, 1–3). Numerous studies report that the frequency of TAM detection is usually higher in tumors with a mesenchymal subtype and/or recurrent tumors (Wang et al., 2017). Glioma stem cells are recently shown to release periostin, which accumulates in the surrounding environment of blood vessels and acts as an inducer of TAM chemotaxis through signaling *via* the integrin receptor αvβ3 (Zhou et al., 2015). Transforming growth factor (TGF)-β released by TAMs induces matrix metalloprotein-9 expression in glioblastoma stem cells, thereby increasing their invasiveness (Ye et al., 2012). Furthermore, the supernatant from glioma stem cells (GSCs) inhibits the

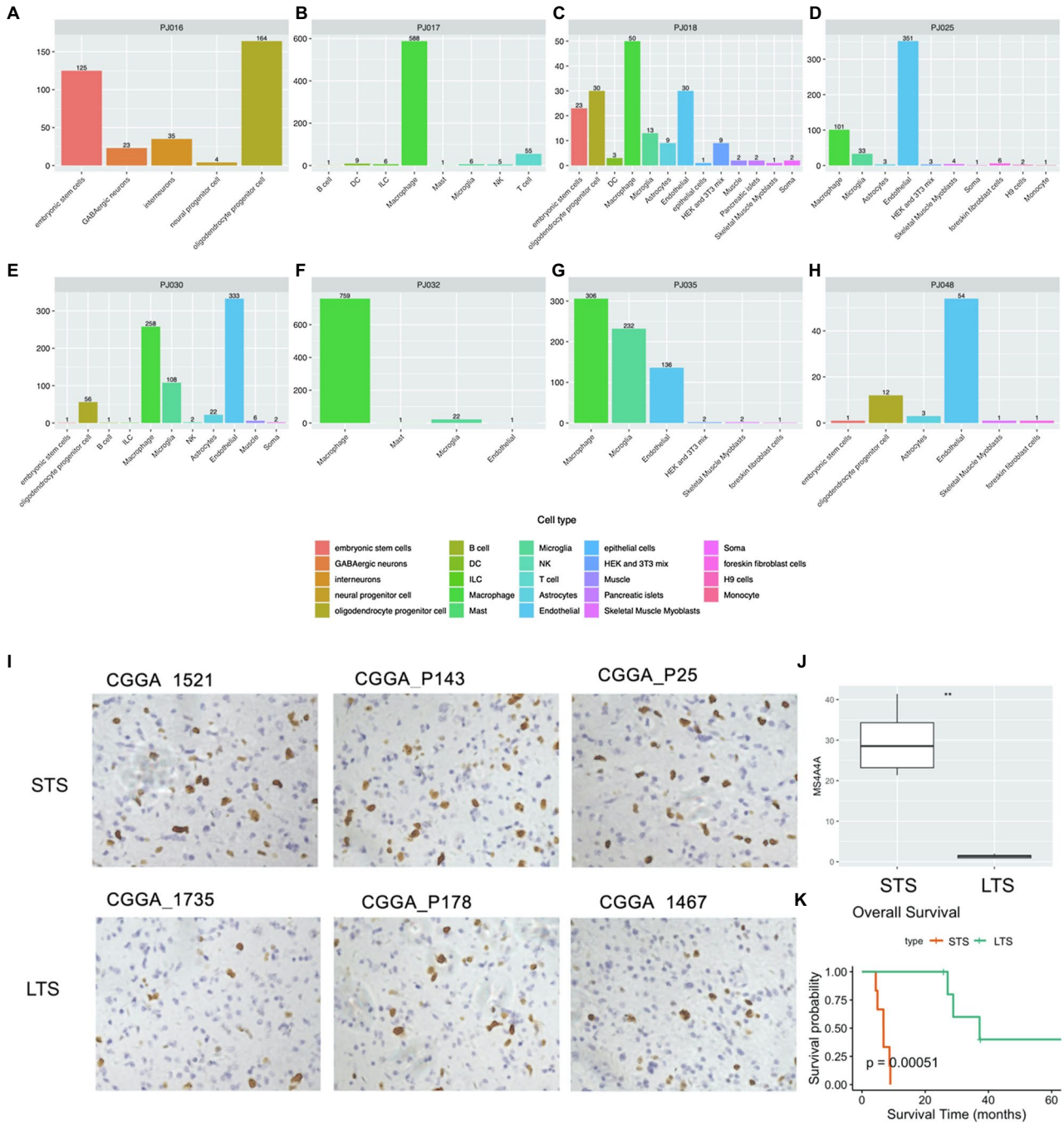

FIGURE 5 | Cell type analysis using scRNA-seq data. **(A–H)** Cell type counts in scRNA-seq samples. **(I)** IHC staining of macrophages (NIR samples, upper; PIR samples, bottom). **(J)** Percentage of macrophages in NIR and PIR samples (according to staining for MS4A4A; scale bar: 100 μm). **(K)** Kaplan–Meier survival curves of NIR and PIR samples. IHC, immunohistochemical; LTS, long-term survival; NIR, negative immune response; PIR, positive immune response; scRNA, single-cell RNA; and STS, short-term survival.

phagocytic activity of TAMs and induces IL-10 and TGF-β secretion (Wu et al., 2010).

Ontogeny analysis revealed that macrophages in human GBM can be divided into either blood-derived or tissue-resident variants (i.e., microglia; Wang et al., 2017). These two ontogenies were also found in other types of cancer and displayed different prognostic effects. In mouse mammary carcinoma, a distinction was made between monocyte-derived TAMs and resident mammary tissue macrophages; it was found that only the former contributes to the suppression of antitumor cytotoxic

TABLE 2 | Summary of scRNA-seq analysis.

Patient	Age	Sex	Diagnosis	Macrophage	Microglial	All	Macrophage: All (%)	Macrophage: Microglia	Cluster
PJ016	49	F	Glioblastoma, WHO grade IV	0	0	3,085	0	—	NIR
PJ017	62	M	Glioblastoma, WHO grade IV	588	6	1,261	46.63	98	NIR
PJ018	65	M	Glioblastoma, WHO grade IV	50	13	2,197	2.28	3.85	PIR
PJ025	74	M	Glioblastoma, WHO grade IV	101	33	5,924	1.70	3.06	PIR
PJ030	56	F	Anaplastic astrocytoma, WHO grade III	258	108	3,097	8.33	2.39	PIR
PJ032	63	F	Glioblastoma, recurrent	759	22	1,377	55.12	34.5	NIR
PJ035	50	M	Glioblastoma, recurrent	306	232	3,768	8.12	1.32	PIR
PJ048	59	M	Glioblastoma, WHO grade IV	0	0	3,084	0	—	NIR

NIR, negative immune response; PIR, positive immune response; scRNA, single-cell RNA; and WHO, World Health Organization.

T cell responses (Franklin et al., 2014; Pombo Antunes et al., 2020). Normal naïve microglial cells can reduce the ability of human stem cells to acquire a spheroid morphology, thereby adversely affecting GSCs and inhibiting the growth of gliomas. However, another study suggested that microglial cells or monocytes derived from gliomas lack such antitumor potential (Sarkar et al., 2013). scRNA-seq analysis of human gliomas showed that blood-derived TAMs upregulate immunosuppressive cytokines and demonstrate an altered metabolism relative to microglial TAMs and that the gene signature of blood-derived TAMs but not microglial TAMs correlates with significantly inferior survival in low-grade glioma (Wu et al., 2010). Signatures of microglial TAMs were not included among the curated markers used for tumor tissue analysis; however, scRNA-seq analysis showed that negative immune response samples comprised a significantly higher macrophage: microglia ratio than positive immune response samples (98 vs. 34.5, respectively; **Table 2**).

CONCLUSION

We present a comprehensive characterization of non-tumor cells in the niches of primary glioblastoma by integrating four large cohorts of GBM gene expression data and 540 gene signatures covering 64 non-tumor cells types. We find that non-tumor cell type enrichment status is useful for stratifying glioblastomas into different prognostic groups (positive or negative immune response clusters). The negative immune response clusters are uniformly enriched with immune system- and organogenesis-related GO terms, whereas positive immune response clusters are enriched with the nervous system. The mesenchymal differentiation is also observed in the negative immune response clusters. Moreover, risk analysis using cell components to determine glioma niches helps interpret the impact of cell type on cancer prognosis. Astrocytes, macrophages, monocytes, NKTs, and MSC are found as risk factors, and CD8 T cells, CD8 naive T cells, and plasma cells are protective factors. Particularly, the high presence of macrophages in the negative immune response clusters is validated using scRNA-seq analysis and IHC staining of GBMs from independent cohorts. Future investigations should focus on cell types with variable risk levels in order to elucidate the potential mechanisms involved in shifts in prognostic effects. Other stratification methods should be established and evaluated for categorizing samples individually rather than as groups.

DATA AVAILABILITY STATEMENT

This data can be found at: The data that support the findings of this study are openly available. The availability of download URL and clinical information for the four datasets was indicated in the original researches including Cohort 1 (Wang et al., 2016; Zhang et al., 2019), Cohort 2 (TCGA; RNA sequences; Cancer Genome Atlas Research, 2008), Cohort 3 (REMBRANDT, mRNA microarray; Gusev et al., 2018), and Cohort 4 (TCGA, mRNA microarray; Brennan et al., 2013). The single-cell RNAseq data of eight HGGs can be accessed through Gene Expression Omnibus (accession: GSE103224; Yuan et al., 2018). All the raw data and original images of IHC were also deposited at github.[2] The code for calculating clustering and survival analysis, cox regression analysis, NES score, and NES distribution, was deposited at github.[3]

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Beijing Tiantan Hospital institutional review board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WZ, YF, and JZ conceived and supervised the study. LL, TS, YH, GL, LW, XH, HW, YL, YC, HYW, and JL curated the data. ZX, GZL, and WL performed the analysis. ZX, GZL, and JZ investigated the results. WZ, GZL, and WL performed the

[2]https://github.com/zhangjbig/xzx/tree/main/Data
[3]https://github.com/zhangjbig/xzx

validation experiments. ZX and WL conducted the visualization. ZX, WZ, YF, and JZ wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the Youth Thousand Scholar Program of China (JZ), Program for High-Level Overseas Talents, Beihang University (JZ), Outstanding and innovative program in medicine and engineering, Beihang University (JZ), National Natural Science Foundation of China (no. 81672479 to WZ, 11421202, and 11827803 to YBF), National Natural Science Foundation of China (NSFC)/Research Grants Council (RGC) Joint Research Scheme (81761168038; WZ), and Beijing Municipal Administration of Hospitals' Mission Plan (SML20180501; WZ).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at https://www.frontiersin.org/articles/10.3389/fgene.2021.741325/full#supplementary-material

REFERENCES

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 18:220. doi: 10.1186/s13059-017-1349-1

Bailey, P., and Cushing, H. (1927). A classification of the tumors of the glioma group on a histo-genetic basis with a correlated study of prognosis. *Arch NeurPsych.* 17:570.

Bao, Z. S., Chen, H. M., Yang, M. Y., Zhang, C. B., Yu, K., Ye, W. L., et al. (2014). RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. *Genome Res.* 24, 1765–1773. doi: 10.1101/gr.165126.113

Bhat, K. P. L., Balasubramaniyan, V., Vaillant, B., Ezhilarasan, R., Hummelink, K., Hollingsworth, F., et al. (2013). Mesenchymal differentiation mediated by NF-kappaB promotes radiation resistance in glioblastoma. *Cancer Cell* 24, 331–346. doi: 10.1016/j.ccr.2013.08.001

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* 39, 782–795. doi: 10.1016/j.immuni.2013.10.003

Boussiotis, V. A., and Charest, A. (2018). Immunotherapies for malignant glioma. *Oncogene* 37, 1121–1141. doi: 10.1038/s41388-017-0024-z

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2013). The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477. doi: 10.1016/j.cell.2013.09.034

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

Cancer Genome Atlas Research, N. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi: 10.1038/nature07385

Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* 164, 550–563. doi: 10.1016/j.cell.2015.12.028

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer Immunogenomic analyses reveal genotype-Immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* 18, 248–262. doi: 10.1016/j.celrep.2016.12.019

Chung, W., Eum, H. H., Lee, H. O., Lee, K. M., Lee, H. B., Kim, K. T., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* 8:15081. doi: 10.1038/ncomms15081

Franklin, R. A., Liao, W., Sarkar, A., Kim, M. V., Bivona, M. R., Liu, K., et al. (2014). The cellular and molecular origin of tumor-associated macrophages. *Science* 344, 921–925. doi: 10.1126/science.1252510

Frattini, V., Pagnotta, S. M., Tala, F., Russo, M. V., Lee, S. B., et al. (2018). A metabolic function of FGFR3-TACC3 gene fusions in cancer. *Nature* 553, 222–227. doi: 10.1038/nature25171

Gabrilovich, D. I. (2017). Myeloid-derived suppressor cells. *Cancer Immunol. Res.* 5, 3–8. doi: 10.1158/2326-6066.CIR-16-0297

Guan, X., Hasan, M. N., Maniar, S., Jia, W., and Sun, D. (2018). Reactive astrocytes in glioblastoma multiforme. *Mol. Neurobiol.* 55, 6927–6938. doi: 10.1007/s12035-018-0880-8

Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J. C., Fine, H., and Madhavan, S. (2018). The REMBRANDT study, a large collection of genomic data from brain cancer patients. *Sci. Data.* 5:180158. doi: 10.1038/sdata.2018.158

Hambardzumyan, D., Gutmann, D. H., and Kettenmann, H. (2016). The role of microglia and macrophages in glioma maintenance and progression. *Nat. Neurosci.* 19, 20–27. doi: 10.1038/nn.4185

Hasan, T., Caragher, S. P., Shireman, J. M., Park, C. H., Atashi, F., Baisiwala, S., et al. (2019). Interleukin-8/CXCR2 signaling regulates therapy-induced plasticity and enhances tumorigenicity in glioblastoma. *Cell Death Dis.* 10:292. doi: 10.1038/s41419-019-1387-6

Hegi, M. E., Diserens, A. C., Gorlia, T., Hamou, M. F., de Tribolet, N., Weller, M., et al. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* 352, 997–1003. doi: 10.1056/NEJMoa043331

Hu, H., Mu, Q., Bao, Z., Chen, Y., Liu, Y., Chen, J., et al. (2018). Mutational landscape of secondary glioblastoma guides MET-targeted trial in brain tumor. *Cell* 175, 1665–1678. doi: 10.1016/j.cell.2018.09.038

Jiang, T., Mao, Y., Ma, W., Mao, Q., You, Y., Yang, X., et al. (2016). CGCG clinical practice guidelines for the management of adult diffuse gliomas. *Cancer Lett.* 375, 263–273. doi: 10.1016/j.canlet.2016.01.024

Kleihues, P., Burger, P. C., and Scheithauer, B. W. (1993). The new WHO classification of brain tumours. *Brain Pathol.* 3, 255–268. doi: 10.1111/j.1750-3639.1993.tb00752.x

Lee, S. H., Kim, J. H., Rhee, C. H., Kang, Y. S., Lee, J. H., Hong, S. I., et al. (1995). Loss of heterozygosity on chromosome 10, 13q(Rb), 17p, and p53 gene mutations in human brain gliomas. *J. Korean Med. Sci.* 10, 442–448. doi: 10.3346/jkms.1995.10.6.442

Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Chen, C., et al. (2020). SciBet as a portable and fast single cell type identifier. *Nat. Commun.* 11:1818. doi: 10.1038/s41467-020-15523-2

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Nakamura, M., Yang, F., Fujisawa, H., Yonekawa, Y., Kleihues, P., and Ohgaki, H. (2000). Loss of heterozygosity on chromosome 19 in secondary glioblastomas. *J. Neuropathol. Exp. Neurol.* 59, 539–543. doi: 10.1093/jnen/59.6.539

Ohgaki, H., and Kleihues, P. (2005). Epidemiology and etiology of gliomas. *Acta Neuropathol.* 109, 93–108. doi: 10.1007/s00401-005-0991-y

Parsons, D. W., Jones, S., Zhang, X., Lin, J. C., Leary, R. J., Angenendt, P., et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science* 321, 1807–1812. doi: 10.1126/science.1164382

Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., et al. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 9, 157–173. doi: 10.1016/j.ccr.2006.02.019

Pombo Antunes, A. R., Scheyltjens, I., Duerinck, J., Neyns, B., Movahedi, K., and Van Ginderachter, J. A. (2020). Understanding the glioblastoma immune microenvironment as basis for the development of new immunotherapeutic strategies. *elife* 9:e52176. doi: 10.7554/eLife.52176

Prins, R. M., and Liau, L. M. (2003). Immunology and immunotherapy in neurosurgical disease. *Neurosurgery* 53, 144–152. doi: 10.1227/01.NEU.0000068865.34216.3A

Prins, R. M., Odesa, S. K., and Liau, L. M. (2003). Immunotherapeutic targeting of shared melanoma-associated antigens in a murine glioma model. *Cancer Res.* 63, 8487–8491.

Ritland, S. R., Ganju, V., and Jenkins, R. B. (1995). Region-specific loss of heterozygosity on chromosome 19 is related to the morphologic type of human glioma. *Genes Chromosomes Cancer* 12, 277–282. doi: 10.1002/gcc.2870120407

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., and Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 160, 48–61. doi: 10.1016/j.cell.2014.12.033

Sarkar, S., Doring, A., Zemp, F. J., Silva, C., Lun, X., Wang, X., et al. (2013). Therapeutic activation of macrophages and microglia to suppress brain tumor-initiating cells. *Ann. Neurosci.* 20, 46–55. doi: 10.5214/ans.0972.7531.200407

Shah, N., Lin, B., Sibenaller, Z., Ryken, T., Lee, H., Yoon, J. G., et al. (2011). Comprehensive analysis of MGMT promoter methylation: correlation with MGMT expression and clinical response in GBM. *PLoS One* 6:e16146. doi: 10.1371/journal.pone.0016146

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

Silginer, M., Nagy, S., Happold, C., Schneider, H., Weller, M., and Roth, P. (2017). Autocrine activation of the IFN signaling pathway may promote immune escape in glioblastoma. *Neuro-Oncology* 19, 1338–1349. doi: 10.1093/neuonc/nox051

Struss, A. K., Romeike, B. F., Munnia, A., Nastainczyk, W., Steudel, W. I., Konig, J., et al. (2001). PHF3-specific antibody responses in over 60% of patients with glioblastoma multiforme. *Oncogene* 20, 4107–4114. doi: 10.1038/sj.onc.1204552

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. 3rd, et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902. doi: 10.1016/j.cell.2019.05.031

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.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Sun, T., He, Y., Li, W., Liu, G., Li, L., Wang, L., et al. (2021). neoDL: a novel neoantigen intrinsic feature-based deep learning model identifies IDH wild-type glioblastomas with the longest survival. *BMC Bioinf.* 22:382. doi: 10.1186/s12859-021-04301-6

Tang, B., Wu, W., Wei, X., Li, Y., Ren, G., and Fan, W. (2014). Activation of glioma cells generates immune tolerant NKT cells. *J. Biol. Chem.* 289, 34595–34600. doi: 10.1074/jbc.M114.614503

Taylor, A. M., Shih, J., Ha, G., Gao, G. F., Zhang, X., Berger, A., et al. (2018). Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* 33, 676–689. doi: 10.1016/j.ccell.2018.03.007

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H. 2nd, Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. doi: 10.1126/science.aad0501

Venteicher, A. S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M. G., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* 355:eaai8478. doi: 10.1126/science.aai8478

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Wang, J., Cazzato, E., Ladewig, E., Frattini, V., Rosenbloom, D. I., Zairis, S., et al. (2016). Clonal evolution of glioblastoma under therapy. *Nat. Genet.* 48, 768–776. doi: 10.1038/ng.3590

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., et al. (2017). Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell* 32, 42–56. doi: 10.1016/j.ccell.2017.06.003

Wang, L., Liu, Z., Balivada, S., Shrestha, T., Bossmann, S., Pyle, M., et al. (2012). Interleukin-1β and transforming growth factor-β cooperate to induce neurosphere formation and increase tumorigenicity of adherent LN-229 glioma cells. *Stem Cell Res Ther* 3, 1–16. doi: 10.1186/scrt96

Weller, M., Weber, R. G., Willscher, E., Riehmer, V., Hentschel, B., Kreuz, M., et al. (2015). Molecular classification of diffuse cerebral WHO grade II/III gliomas using genome-and transcriptome-wide profiling improves stratification of prognostically distinct patient groups. *Acta Neuropathol.* 129, 679–693. doi: 10.1007/s00401-015-1409-0

Wu, A., Wei, J., Kong, L. Y., Wang, Y., Priebe, W., Qiao, W., et al. (2010). Glioma cancer stem cells induce immunosuppressive macrophages/microglia. *Neuro-Oncology* 12, 1113–1125. doi: 10.1093/neuonc/noq082

Yang, I., Kremen, T. J., Giovannone, A. J., Paik, E., Odesa, S. K., Prins, R. M., et al. (2004). Modulation of major histocompatibility complex class I molecules and major histocompatibility complex-bound immunogenic peptides induced by interferon-alpha and interferon-gamma treatment of human glioblastoma multiforme. *J. Neurosurg.* 100, 310–319. doi: 10.3171/jns.2004.100.2.0310

Yang, I., Tihan, T., Han, S. J., Wrensch, M. R., Wiencke, J., Sughrue, M. E., et al. (2010). CD8+ T-cell infiltrate in newly diagnosed glioblastoma is associated with long-term survival. *J. Clin. Neurosci.* 17, 1381–1385. doi: 10.1016/j.jocn.2010.03.031

Ye, X. Z., Xu, S. L., Xin, Y. H., Yu, S. C., Ping, Y. F., Chen, L., et al. (2012). Tumor-associated microglia/macrophages enhance the invasion of glioma stem-like cells via TGF-beta1 signaling pathway. *J. Immunol.* 189, 444–453. doi: 10.4049/jimmunol.1103248

Yeung, Y. T., McDonald, K. L., Grewal, T., and Munoz, L. (2013). Interleukins in glioblastoma pathophysiology: implications for therapy: targeting ILs in glioblastoma. *Br. J. Pharmacol.* 168, 591–606. doi: 10.1111/bph.12008

Yuan, J., Levitin, H. M., Frattini, V., Bush, E. C., Boyett, D. M., Samanamud, J., et al. (2018). Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Med.* 10:57. doi: 10.1186/s13073-018-0567-9

Zagzag, D., Salnikow, K., Chiriboga, L., Yee, H., Lan, L., Ali, M. A., et al. (2005). Downregulation of major histocompatibility complex antigens in invading glioma cells: stealth invasion of the brain. *Lab. Investig.* 85, 328–341. doi: 10.1038/labinvest.3700233

Zhang, J., Caruso, F. P., Sa, J. K., Justesen, S., Nam, D. H., Sims, P., et al. (2019). The combination of neoantigen quality and T lymphocyte infiltrates identifies glioblastomas with the longest survival. *Commun Biol* 2:135. doi: 10.1038/s42003-019-0369-7

Zhao, Z., Meng, F., Wang, W., Wang, Z., Zhang, C., and Jiang, T. (2017). Comprehensive RNA-seq transcriptomic profiling in the malignant progression of gliomas. *Sci Data* 4:170024. doi: 10.1038/sdata.2017.24

Zhou, W., Ke, S. Q., Huang, Z., Flavahan, W., Fang, X., Paul, J., et al. (2015). Periostin secreted by glioblastoma stem cells recruits M2 tumour-associated macrophages and promotes malignant growth. *Nat. Cell Biol.* 17, 170–182. doi: 10.1038/ncb3090

Zhu, C., Zou, C., Guan, G., Guo, Q., Yan, Z., Liu, T., et al. (2019). Development and validation of an interferon signature predicting prognosis and treatment response for glioblastoma. *Onco. Targets. Ther.* 8:e1621677. doi: 10.1080/2162402X.2019.1621677

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 Genetics

ORIGINAL RESEARCH
published: 22 September 2021
doi: 10.3389/fgene.2021.723802

Identification and Validation of EMT-Related lncRNA Prognostic Signature for Colorectal Cancer

Danfeng Li[1], *Xiaosheng Lin*[1], *Binlie Chen*[2,1], *Zhiyan Ma*[2,1], *Yongming Zeng*[1] *and Huaiming Wang*[1]*

[1]Department of Gastrointestinal Surgery, The First Affiliated Hospital of Shantou University Medical College, Shantou, China, [2]Medical College, Shantou University, Shantou, China

Background: This study aimed to explore the biological functions and prognostic role of Epithelial-mesenchymal transition (Epithelial-mesenchymal transition)-related lncRNAs in colorectal cancer (CRC).

Methods: The Cancer Genome Atlas database was applied to retrieve gene expression data and clinical information. An EMT-related lncRNA risk signature was constructed relying on univariate Cox regression, Least Absolute Shrinkage and Selector Operation (LASSO) and multivariate Cox regression analysis of the EMT-related lncRNA expression data and clinical information. Then, an individualized prognostic prediction model based on the nomogram was developed and the predictive accuracy and discriminative ability of the nomogram were determined by the receiver operating characteristic curve and calibration curve. Finally, a series of analyses, such as functional analysis and unsupervised cluster analysis, were conducted to explore the influence of independent lncRNAs on CRC.

Results: A total of 581 patients were enrolled and an eleven-EMT-related lncRNA risk signature was identified relying on the comprehensive analysis of the EMT-related lncRNA expression data and clinical information in the training cohort. Then, risk scores were calculated to divide patients into high and low-risk groups, and the Kaplan-Meier curve analysis showed that low-risk patients tended to have better overall survival (OS). Multivariate Cox regression analysis indicated that the EMT-related lncRNA signature was significantly associated with prognosis. The results were subsequently confirmed in the validation dataset. Then, we constructed and validated a predictive nomogram for overall survival based on the clinical factors and risk signature. Functional characterization confirmed this signature could predict immune-related phenotype and was associated with immune cell infiltration (i.e., macrophages M0, M1, Tregs, CD4 memory resting cells, and neutrophils), tumor mutation burden (TMB).

Conclusions: Our study highlighted the value of the 11-EMT-lncRNA signature as a predictor of prognosis and immunotherapeutic response in CRC.

Keywords: EMT-related lncRNA, nomogram, colorectal cancer, prognostic, signature

OPEN ACCESS

Edited by:
Jiangning Song,
Monash University, Australia

Reviewed by:
Fei Guo,
Tianjin University, China
Bolin Chen,
Northwestern Polytechnical University, China

****Correspondence:***
Huaiming Wang
13750417745@163.com

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

Received: *11 June 2021*
Accepted: *02 September 2021*
Published: *22 September 2021*

Citation:
Li D, Lin X, Chen B, Ma Z, Zeng Y and Wang H (2021) Identification and Validation of EMT-Related lncRNA Prognostic Signature for Colorectal Cancer. Front. Genet. 12:723802. doi: 10.3389/fgene.2021.723802

INTRODUCTION

Colorectal cancer (CRC) is the third most common malignancy with the second-highest cancer-related mortality worldwide. The number of cases is expected to rise by 60% in the year 2030 worldwide. Despite the development of CRC therapies such as surgery, radiotherapy, chemotherapy, targeted therapy, immunotherapy, the 5-years survival rate of late stages CRC still less than 20% (Siegel et al., 2017). Increasing data underline that the tumor microenvironment (TME) contributes a vital role in CRC progression, as well as in the response to therapy. In most types of cancer, the infiltration of CD8 T cells and tumor-infiltrating lymphocytes (TIL) in tumor beds is a biomarker for a good prognosis (Ma et al., 2019). Similarly, the presence of CD8 T cells in the tumor bed and infiltrating margins is strongly associated with prognosis in CRC (Zhang et al., 2018).

Based on the mutation pattern and the ratio of MSI markers, CRC tumors can be divided into the dMMR group and pMMR group. In recent years, studies have shown that dMMR-MSI-H CRC tumors have a high tumor mutation burden and can present new antigens on major histocompatibility complex (MHC) class I molecules, which makes them more sensitive to T cell activation therapy, while the pMMR-MSI-L CRC tumor has a low tumor mutation burden, with a low immune response rate (Ledys et al., 2018). Therefore, in 2017, the Food and Drug Administration (FDA) approved PD-1 drugs for dMMR-MSI-H mCRC patients. Unfortunately, only about 15% of CRC patients with the dMMR-MSI-H phenotype, and among all mCRC patients, the dMMR-MSI-H phenotype only accounts for about 5%. Moreover, not all CRC cases with the dMMR-MSI-H phenotype respond well to immunotherapies (Fabrizio et al., 2018). A series of studies showed that the effective rate of immunotherapy in CRC with dMMR-MSI-H phenotype is only 40%, while in pMMR CRC patients, the effective rate of immunotherapy is very low, and recent biological advances suggest that combination therapy can reverse this resistance (Ghiringhelli and Fumet, 2019). Furthermore, several experimental data have shown that tumor intrinsic factors may also modulate responses to immunotherapy, such as genes participating in cell adhesion, extracellular matrix remodeling, angiogenesis, wound healing, and mesenchymal transformation (Hugo et al., 2016). As such, it is urgent to look for biomarkers or more effective strategies based on tumor gene-expression profiling to treat patients with various subsets of advanced CRC.

Epithelial-mesenchymal transition (EMT) is a process in which epithelial cells lose connection and polarity, and acquire plasticity, migration, invasion ability, stem cell-like characteristics, and resistance to apoptosis. It has been proved that EMT is an important way to promote tumor cell metastasis. Accumulating preclinical researches have confirmed that the level of EMT contributes to the level of immunosuppression, with more mesenchymal tumors being more resistant to immunotherapy, and tumor immunosuppression and immune evasion could be reversed by the EMT progress. (Terry et al., 2017; Dongre and Weinberg, 2019).

Moreover, emerging studies revealed that EMT is regulated by a complex regulatory network, including the typical regulation of EMT transcription factors (EMT-TFs), noncoding RNAs, epigenetic modification, post-translational regulation, and alternative splicing factors (Chaffer et al., 2016; Diepenbruck and Christofori, 2016; Nieto et al., 2016). Thus, EMT-related lncRNAs and genes may be a promising target for future therapeutic interventions.

Currently, accumulating shreds of evidence indicated that lncRNAs play a vital role in the progression of tumors and can be used as robust predictors of the prognosis for cancer patients. It has been well known that lncRNAs involved EMT progression. For instance, in our previous study, we identified that linc00662 was significantly increased in CRC cells and tissues, and significantly stimulating EMT progression and inducing tumor growth both *in vivo* and *in vitro* (Wang et al., 2020a). Moreover, other research revealed that decreasing the expression of linc01133 can inhibit EMT and metastasis in CRC cells (Kong et al., 2016). Nevertheless, a single lncRNA may only explain its partial effect on tumors, so it is very important to comprehensively analyze the expression profile of EMT-related lncRNAs, as well as their different pathological features and prognostic value in CRC, which may lead to a deeper understanding of the effect of EMT-related lncRNAs on tumors and to propose newer treatment strategies.

In the current research, we analyzed the RNAseq data and corresponding clinical information retrieved from the TCGA (N = 581) database to comprehensively explore the prognostic role of EMT-related lncRNA, and an 11-lncRNA signature was constituted and validated in the training and test cohorts. Furthermore, we then characterized the underlying molecular and immune profile of EMT-related lncRNAs signature in CRC. Consequently, we found that this signature could identify different immune infiltration states, TIDE prediction score, and MSI status of each patient, which explains it was a promising prognostic biomarker for CRC patients receiving immunotherapy.

METHODS

Data Acquisition

The RNA-seq reads count and clinical information were obtained from the TCGA database (https://portal.gdc.cancer.gov/). Samples with a survival time≥of 30 days were selected to ensure higher quality analysis. Subsequently, 581 patients with CRC from the TCGA were included for further analysis. Then, we retrieved somatic mutation profiles of all tumor samples in the TCGA database.

Identification of Epithelial-Mesenchymal Transition-Related lncRNAs

200 EMT-related genes were downloaded from the Molecular Signature database v7.1 (MSigDB) (http://www.broad.mit.edu/gsea/msigdb/). To identify EMT-related lncRNAs, firstly, all lncRNAs expression data were extracted from the TCGA database relying on the GENCODE project (http://www.gencodegenes.org). Then, Pearson correlation analysis between EMT-related genes and all lncRNA expression data in samples was performed to identify the EMT-related lncRNA based on $|\text{Cor pearson}| > 0.6$ and $p\text{-value} < 0.01$.

Development and Validation of the Prognostic Signature

CRC patients were randomly divided into training and test cohorts with a 6:4 ratio. In the training cohort, univariate Cox regression analysis was carried out to explore the relationship between each EMT-related lncRNA expression and overall survival (OS). Then, these lncRNAs were further analyzed by utilizing LASSO penalized Cox proportional hazards regression to identify the best risk model in the R package "glmnet". Using the following formula: risk score=(β1*G1+β2*G2+β3*G3+⋯+βn*Gn) to calculate the risk score for each patient, where β is the coefficient of each lncRNA, G represents each lncRNA expression value, and n denotes the number of lncRNAs. Patients were classified into two risk groups depending on the median risk score. Moreover, the survival curve was adopted using the Kaplan-Meier method in the R survminer package, where the differences between the two risk groups were calculated by the log-rank test. Meanwhile, a time-dependent receiver operating characteristic (ROC) curve was determined using R 'survivalROC' package, of which the area under the curve (AUC) was calculated to assess the accuracy of the prognostic risk signature. To further verify the predictive performance of the prognostic signature, the risk scores were also calculated in the testing cohort utilizing the same prognostic formula, and the Kaplan–Meier survival curve and ROC curve were conducted with a cutoff value of the median risk score.

Independence of the Epithelial-Mesenchymal Transition-Related lncRNA Signature

Univariate Cox regression analysis and multivariate Cox regression analysis were used to identify independence by exploiting the lncRNA characteristics of OS and corresponding clinical information. $p < 0.05$ was considered as statistically significant.

Nomogram Construction and Validation

The 'rms' R package was used to establish the nomogram based on all independent prognostic factors (https://cran.r-project.org/web/packages/rms/index.html). Then, Calibration plot curve analysis was applied to evaluate the discrimination and the calibration of the nomogram.

Gene Set Enrichment Analysis Enrichment Analysis

In the signaling pathway analysis, differential expression analysis was first performed on all genes to analyze the samples with the high and low-risk score using the 'DESeq2' package of R. Enrichment analysis to determine the signaling pathways in which the differentially expressed genes are involved was then carried out by using the gene set enrichment analysis (GSEA) method based on the HALLMARK gene sets with the 'clusterProfiler' package of R. When $p < 0.05$ and FDR <0.05, the path ways were considered as statistically significant.

Gene Mutation Analysis

In the gene mutation analysis, information on genetic alterations was obtained from the cBioPortal database, and the quantity and quality of gene mutations in two risk subgroups were analyzed by utilizing the 'Maftools' package of R. Then, we calculated the TMB of each patient and described the difference of TMB in two risk subgroups.

Tumor Microenvironment Analysis

To evaluate the tumor microenvironment in CRC, we identified the infiltration levels of 22 immune cells using the CIBERSORT algorithm based on the expression level of all genes. First, we uploaded the expression data of all genes to the CIBERSORTx web portal. Next, the algorithm was run using the LM22 signature for 1,000 permutations. The CRC samples with an output p-value < 0.05 were selected for further analysis. Moreover, the immune core and the stromal score were calculated using the "estimate" R package. Single sample GSEA (ssGSEA) analysis was then performed with the 'GSVA' package of R, to estimate the abundance of 28 immune infiltrate cells. Additionally, TIMER 2.0 (Tumor Immune Estimation Resource) database was used to explore the correlation of mutation genes and immune infiltration level in CRC.

Immunotherapeutic Sensitivity With Prognostic Signature

To further validate the predictive performance of the given prognostic signature for the ICIs response, the Tumor Immune Dysfunction and Exclusion (TIDE) algorithm was assigned to assess the immunogenicity and immunotherapeutic sensitivity of CRC patients. The results were measured by the TIDE score, which was calculated online (http://tide.dfci.harvard.edu/). According to the default settings, a patient with a TIDE value < 0 was defined as a responder (positive sensitivity to immunotherapy), whereas a patient with a TIDE value > 0 was defined as a non-responder (negative sensitivity to immunotherapy).

Statistical Analysis

R software (R version: 3.6.3) was used to perform all data statistical analyses. Wilcoxon test (Mann-Whitney test) was applied to analyze continuous variables, whereas the Fisher's exact test or chi-square test was used to analyze the categorical data. The survival difference was calculated with the K-M analysis methods and the log-rank test. For all statistical analyses, p-value less than 0.05 indicated statistical significance.

RESULTS

Identified Epithelial-Mesenchymal Transition-Related lncRNA in CRC

To explore EMT-Related genes in CRC, we initially retrieved the data from the MSigDB database, with the hallmark gene sets name: HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION, we collected altogether 200 EMT-related genes (**Supplementary Table S1**). Then, we carried out correlation analysis on EMT-related genes and EMT-related lncRNAs, and the absolute Pearson coefficient >0.6 and p-value < 0.01 was used as the screening criteria, we identified a total of 1381 EMT-related lncRNAs (**Supplementary Table S2**). Finally, we merged the LncRNA expression data and clinical information for further analysis.

Construction and Validation of the EMT-Related lncRNA Signature

In total, 581 eligible patients with integrated information, as well as a survival time≥of 30 days were incorporated in the TCGA-CRC dataset and randomly divided into two independent cohorts at a ratio of 6:4, and 1381 EMT-related lncRNAs were included to identify the prognostic risk model. In the univariate Cox regression analyses, 34 lncRNAs were significantly related to OS, which was considered as potential predictors. Then, a LASSO regression algorithm was applied for feature selection, when the partial likelihood binomial deviation reaches the minimum value, the most suitable tuning parameter λ for LASSO regression is 0.055 (**Figure 1A**), 25 variables with non-zero coefficients retained in the LASSO analysis (**Figure 1B**) were further used for multivariate stepwise Cox regression analysis. Then, we established an 11-lncRNA signature model through multivariate stepwise Cox regression hazards analysis (**Figure 1C**). The risk score of each patient in the training set and validation set is calculated according to the risk formula:

Riskscore=AC010536.3*(0.295166381)+AC026369.1*(0.256505491)+AL391095.2*(0.588758986)+AC018755.4*(-0.363854498)+AC002456.1*(0.449864493)+AC020703.1*(-0.887819474)+AC060234.3*(-0.871113022)+AC079070.1*(1.188015484)+EGFLAM.AS4*(-0.851540675)+LINC01147*(-0.551055647)+PGM5.AS1*(0.160723859).

Taking the median risk score as the cutoff value, we categorized patients into a high-risk group and low-risk group. As depicted in **Figures 2A,B**, our data showed that high-risk group patients had a worse OS than low-risk group patients ($p < 0.0001$ in the Training cohort and $p = 0.024$ in the testing cohort, log-rank test). Additionally, as it showed in **Figure 2C**, the high expression level of AC010536.3, AC026369.1, AL391095.2, AC002456.1, AC079070.1and PGM5. AS1 was reported in the high-risk group, conversely, the expression level of AC018755.4, AC020703.1, EGFLAM. AS4 and LINC01147 were higher in the low-risk group, which was consistent in the test cohort (**Figure 2D**). Besides, it was found the OS patients in the high-risk group have corresponded to more death cases in the training cohort and consistent in the validation cohort (**Figures 2E,F**). By drawing a ROC curve based on the risk model, the AUC value in the training cohort was 0.778, 0.812, 0.825, and 0.655, 0.613, 0.655 in the testing cohort in 1,3,5 year prediction, indicating a good prediction prognostic accuracy (**Figures 2G,F**).

To further explore the prognostic value of EMT-lncRNA markers for CRC patients stratified by clinical variables, we divided patients into different groups according to age, gender, and stage, and our data showed that the risk score of CRC patients was positively associated with the stage, but no significant correlation with age, gender and plasma CEA level, considering the different stratified analysis (**Figures 3A–D**).

Construction of the Nomogram and Performance

To verify whether the EMT-related lncRNA signature can be used as an independent predictor of OS, we used univariate and multivariate Cox regression analyses. The results showed that age, stage, and the lncRNA signature can be used as independent predictors of OS (**Figure 4A**). Then, the EMT-related lncRNA signature, age, and stage were selected for the construction of the nomogram (**Figure 4B**). The AUC was 0.816,0.827,0.834 (**Figure 4C**) and 0.734,0.793,0.819 (**Figure 4D**) in the training cohort and testing cohort in predicting 1 year, 3 years, and 5 years OS in CRC, indicating good discrimination and as shown in **Figure 4E**, the calibration plots also present high performance in predicting 1 year, 3 years and 5 years OS in CRC. These results indicated that the nomogram has high accuracy.

Molecular Characteristics of the Epithelial-Mesenchymal Transition-Related lncRNA Signature

As showed in **Figure 5A**, in total, 58 DEGs were obtained after performing the difference analysis on the mRNA of the high - and low-risk groups, including 24 up-regulated and 34 down-regulated DEGs based on the cut-off criteria ($p < 0.05$ and $|logFC|>1$). Then, GSEA analysis was applied to determine the significant pathway associated with the high- and low-risk group in the training cohort, patients in the high-risk group were mainly enriched in cancer and tumor metastasis-related pathways, such as aptical_junction, coagulation, epithelial-mesenchymal transition, hedgehog_signaling, hypoxia, myog enesis and Wnt_β_catenin_signaling pathways (**Figure 5B**). While patients in the low-risk group were mainly enriched in immune response-related pathways, such as allograft_rejection, complement, IL2_STAT5_signaling, IL_6_JAK_STAT_3 signaling, inflammatory_response, interferon_alpha_response, interferon_gamma_response, and TNFA_signaling_via_NFKB pathways (**Figure 5C**).

Besides, gene mutations were analyzed to gain further biological insight into the high- and low-risk group in the training cohort. The results indicated that there was no significant difference in mutation counts between the two groups, and missense_mutation was the most common type. Then, we selected the top 15 genes with the highest mutation rates in two groups (**Figures 6A,B**), the mutation rates of APC, TP53, TTN, KRAS, SYNE1, MUC16, PIK3CA, FAT4, RYR2, DNAH11 were both higher than 16% in the two groups. What is different is that the mutation rates of APC and TP53 were higher in the high-risk group than that in the low-risk group (81 vs 73% and 64 vs55%, which led to the decreased infiltration of CD4 + and CD8 + T lymphocytes in the high-risk group. **Supplementary Figure S3**), and the mutation of CSMD3, USH2A, and NEB genes were more common in the high-risk group, while the mutation of DNAH5, FAT3, and FBXW7 genes were more common in the low-risk group, Interestingly, in the TIMER 2.0 database, we found that the high mutation rate of USH2A, and NEB genes in the high-risk group resulted in decreased infiltration of CD4 + T lymphocytes and CD8 + T lymphocytes in the tumor center, while increased infiltration of Treg cells. In the low-risk group, the high mutation rates of DNAH5, FAT3, and FBXW7 genes resulted in increased infiltration of central CD4 + T lymphocytes and CD8 + T lymphocytes, while decreased infiltration of Treg cells, as shown in **Supplementary Figure S1** and **Supplementary Figure S2**. We then further explored whether the high - and low-risk groups were associated with TMB, and our data demonstrated that the TMB was

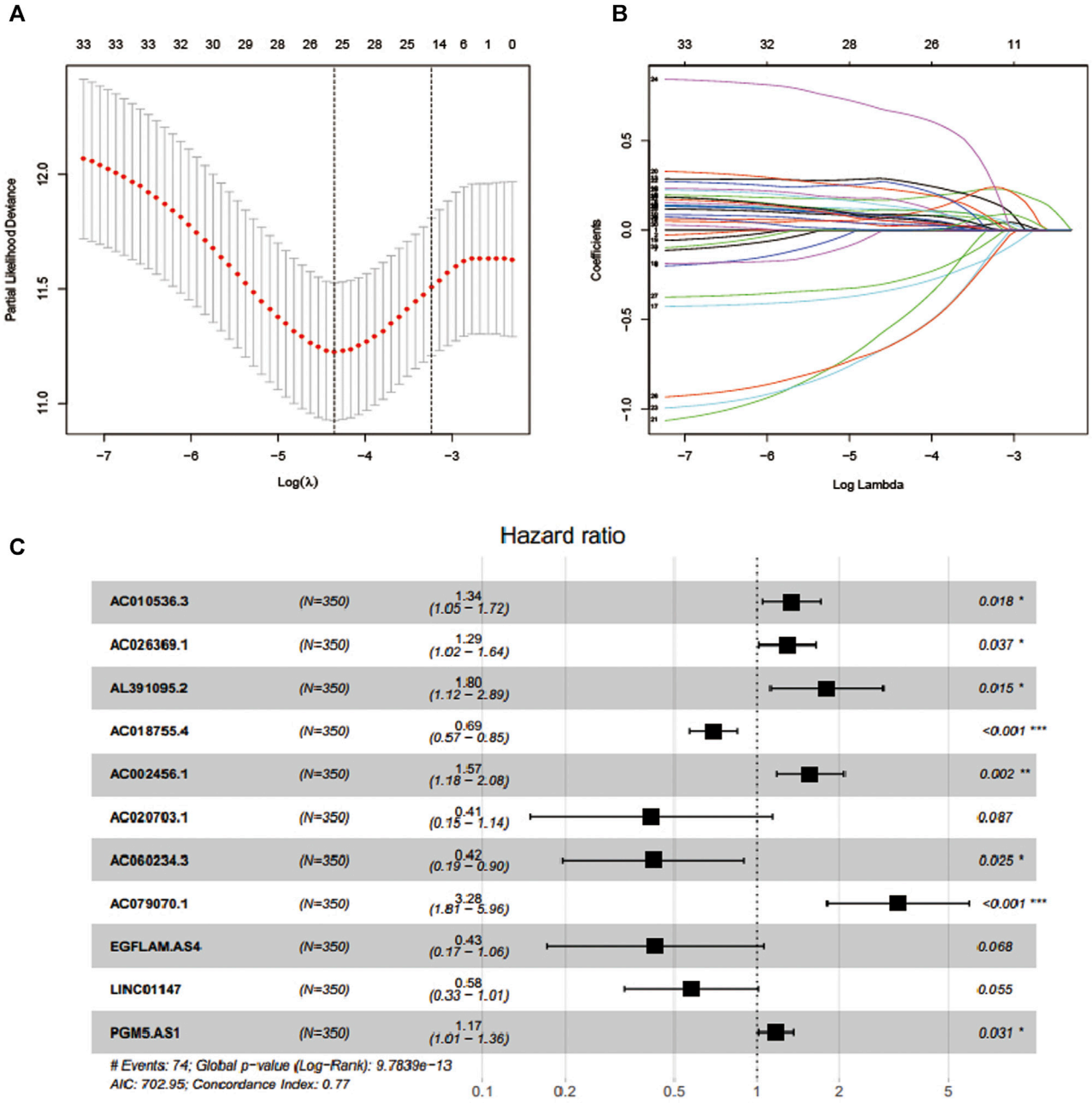

FIGURE 1 | Feature selection using LASSO COX regression and multivariate Cox regression analysis. **(A)** Selection of tuning parameter (λ) in the LASSO regression using 10-fold cross-validation via minimum criteria. **(B)** LASSO coefficient profiles for clinical features and 25 nonzero coefficients are selected. C.11 lncRNAs were selected through multivariate Cox regression analysis.

slightly higher in the low-risk group than that in the high-risk group (**Figure 6C**, $p = 0.059$).

Immune Characteristics of Different Subgroups

We further evaluated the status of immune cell infiltration in TCGA colorectal cancer transcriptome using the ssGSEA approach, and 28 immune-related terms were incorporated to assess the abundance of immune cells in the tumor immune microenvironment. The results showed that 21 immune cell types were significantly different between the two groups (**Figure 7A**). Next, the CIBERSORT algorithm was performed to investigate the immune infiltration in CRC tissues between the high-risk and low-risk group. The results revealed that Neutrophils cells, macrophages M1 cells, T cell CD4 memory resting cells were more abundant in the low-risk group while macrophage M0 cells and T cells regulatory cells were more abundant in the high-risk

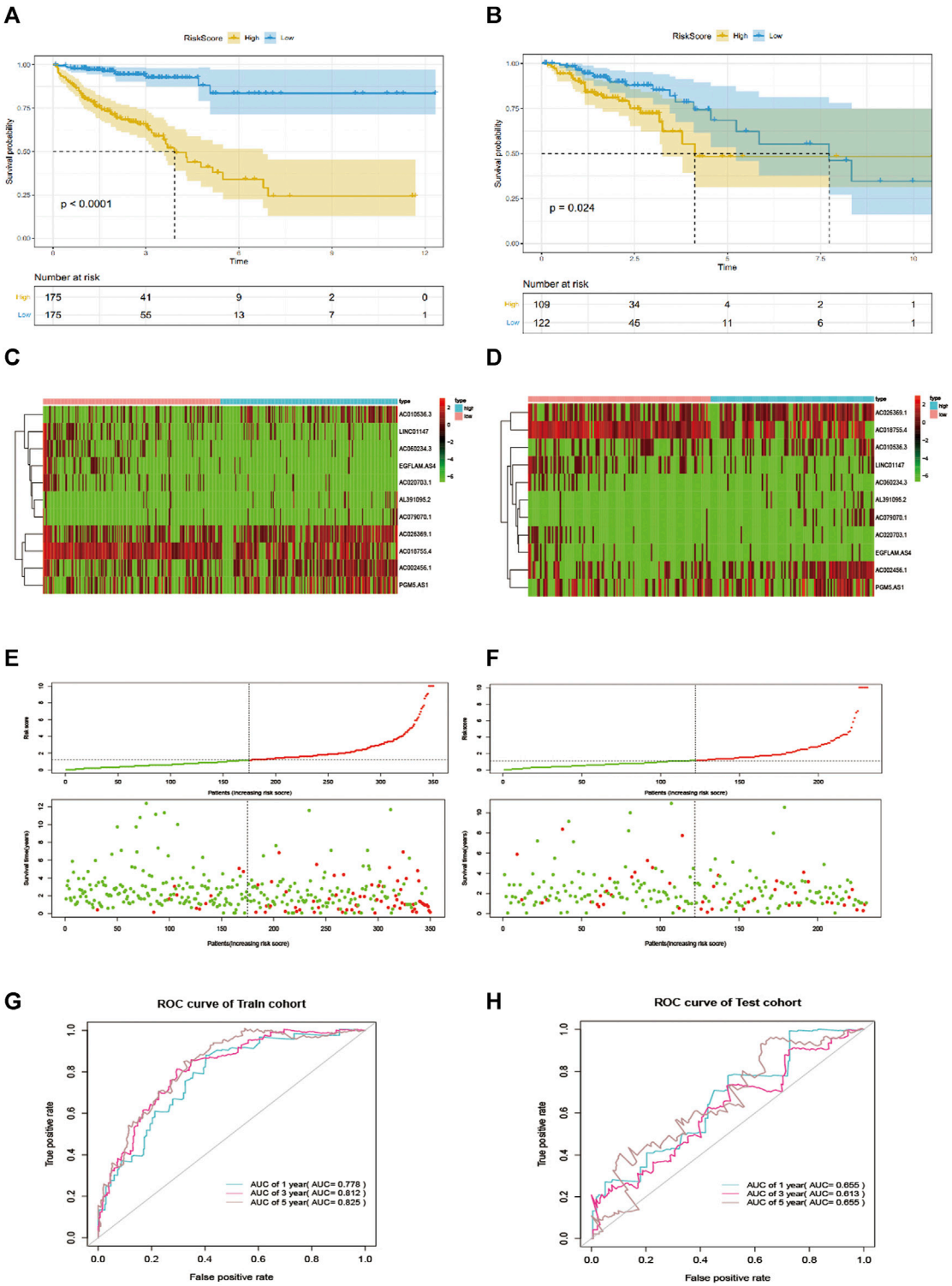

FIGURE 2 | EMT-related lncRNA signature predicts OS in patients with CRC. **(A, B)** Kaplan-Meier curve to verify the predictive effect of the signature in the training and test cohort. **(C, D)** The heatmap of the expression profiles of members in the 11-lncRNA signature. **(E, F)**. Distribution of risk scores per patient in the training and test cohort. **(H, G)**: ROC curve analysis to evaluate the diagnostic efficacy of the signature.

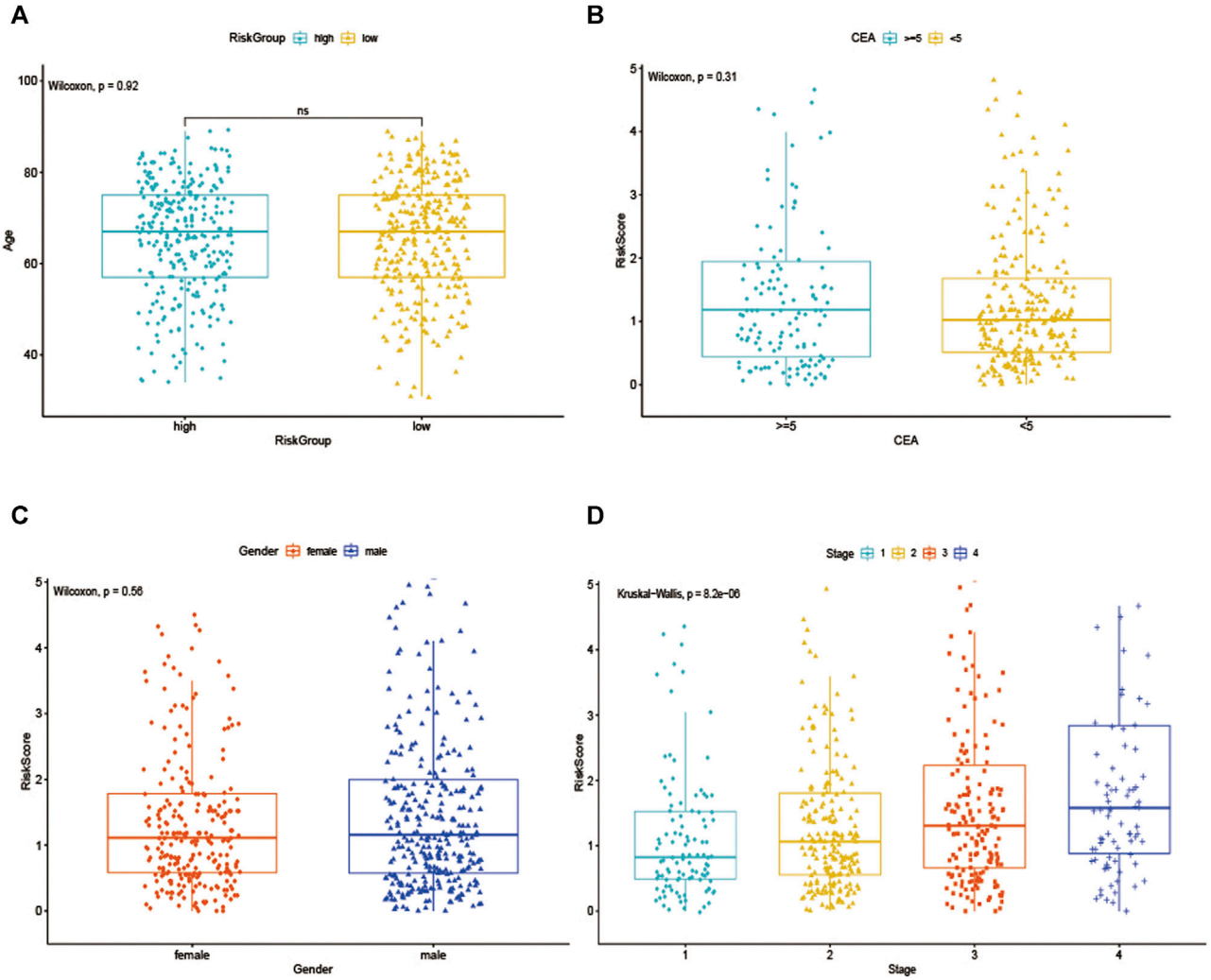

FIGURE 3 | The association between EMT-related signature risk score and clinical factors, including. **(A)** Age, **(B)** CEA, **(C)** Gender, **(D)** Stage.

group (**Figure 7B**). Subsequently, the ESTIMATE algorithm was performed, and we found that estimatescore and immunescore were much higher in the low-risk group than in the high-risk group, while there was no significant difference in stromalscore between the two groups. Therefore, those results indicate that there were more immune components in TME in the low-risk group (**Figure 7C**).

The Benefit of ICI Therapy in Two Different Subgroups

As we knew, the higher TIDE prediction score represented a higher potential for immune evasion, which indicated that the patients were less sensitive to ICI therapy. Then TIDE was used to assess the potential clinical efficacy of immunotherapy in two groups. The results revealed that the low-risk group had a lower TIDE score than the high-risk group, indicating that the low-risk patients could benefit more from ICI therapy than those in the high-risk group (**Figure 8A**). Also, we found the low-risk group had a higher microsatellite instability (MSI) score (**Figure 8B**), while the high-risk group had a higher T cell exclusion score (**Figure 8C**), but there was no difference in T cell dysfunction between the two subgroups (**Figure 8D**).

DISCUSSION

Even with significant advances in screening and treatment strategies, CRC remains the second largest cause of cancer-related death around the world (Siegel et al., 2020). Therefore, a better understanding of CRC pathogenesis and exploring potential biomarkers will likely yield novel insights into the management and prognosis of CRC. In recent years, it has been widely revealed that EMT is closely related to cancer progression and metastasis (Mittal, 2018; Aiello and Kang, 2019), among them, abnormal development of genome, including lncRNA and mRNA, is a typical feature of

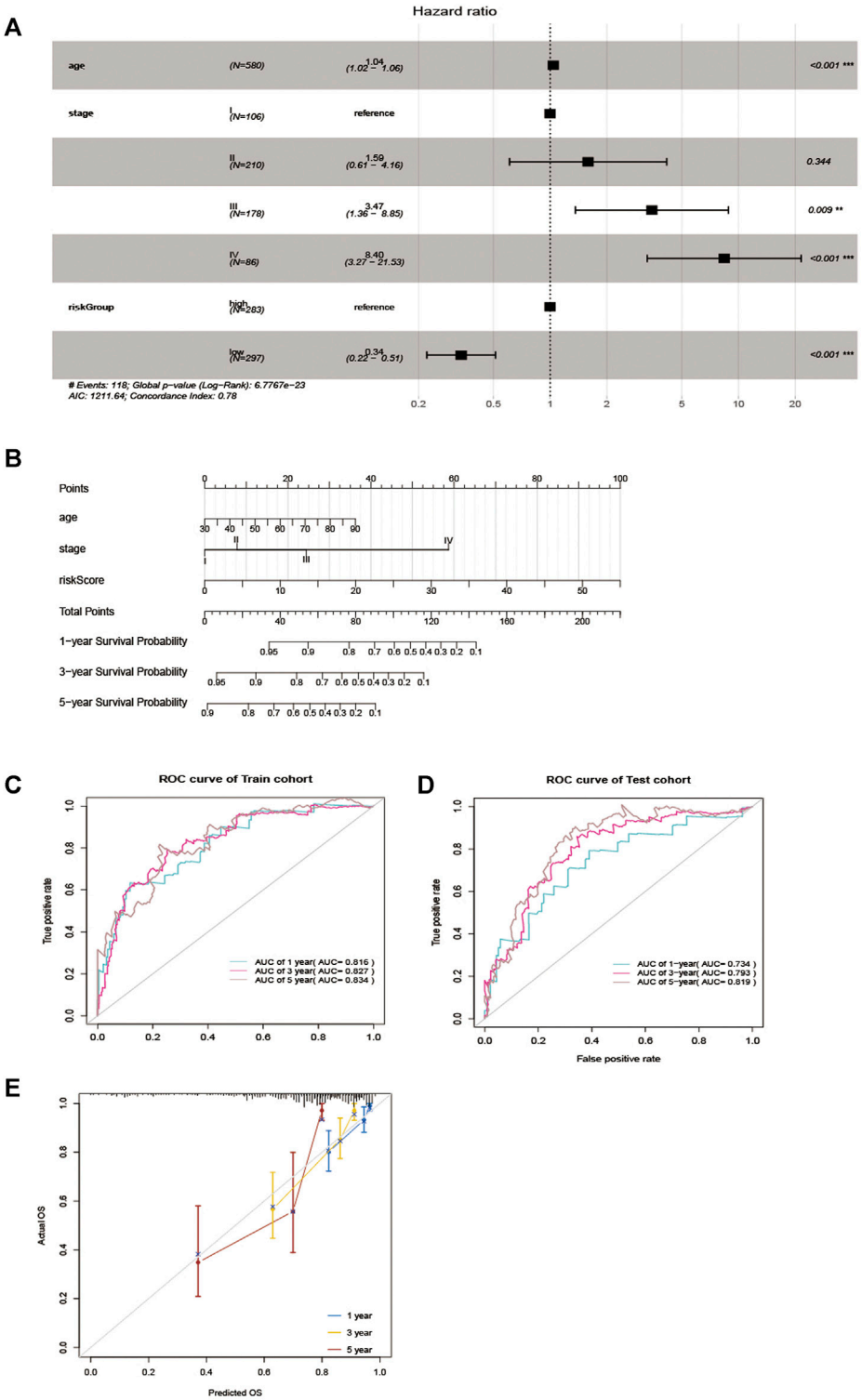

FIGURE 4 | Nomogram for predicting overall survival (OS) of patients with CRC. **(A)** Multivariable analyses for each clinical feature. **(B)** Nomogram construction for the 1-, three- and 5 year OS prediction for the CRC. **(C, D)** Evaluation of the accuracy of the nomogram in 1-, three- and 5 years by using the ROC analysis. **(E)** Calibration curve for the nomogram model for predicting 1-, three- and 5 years OS.

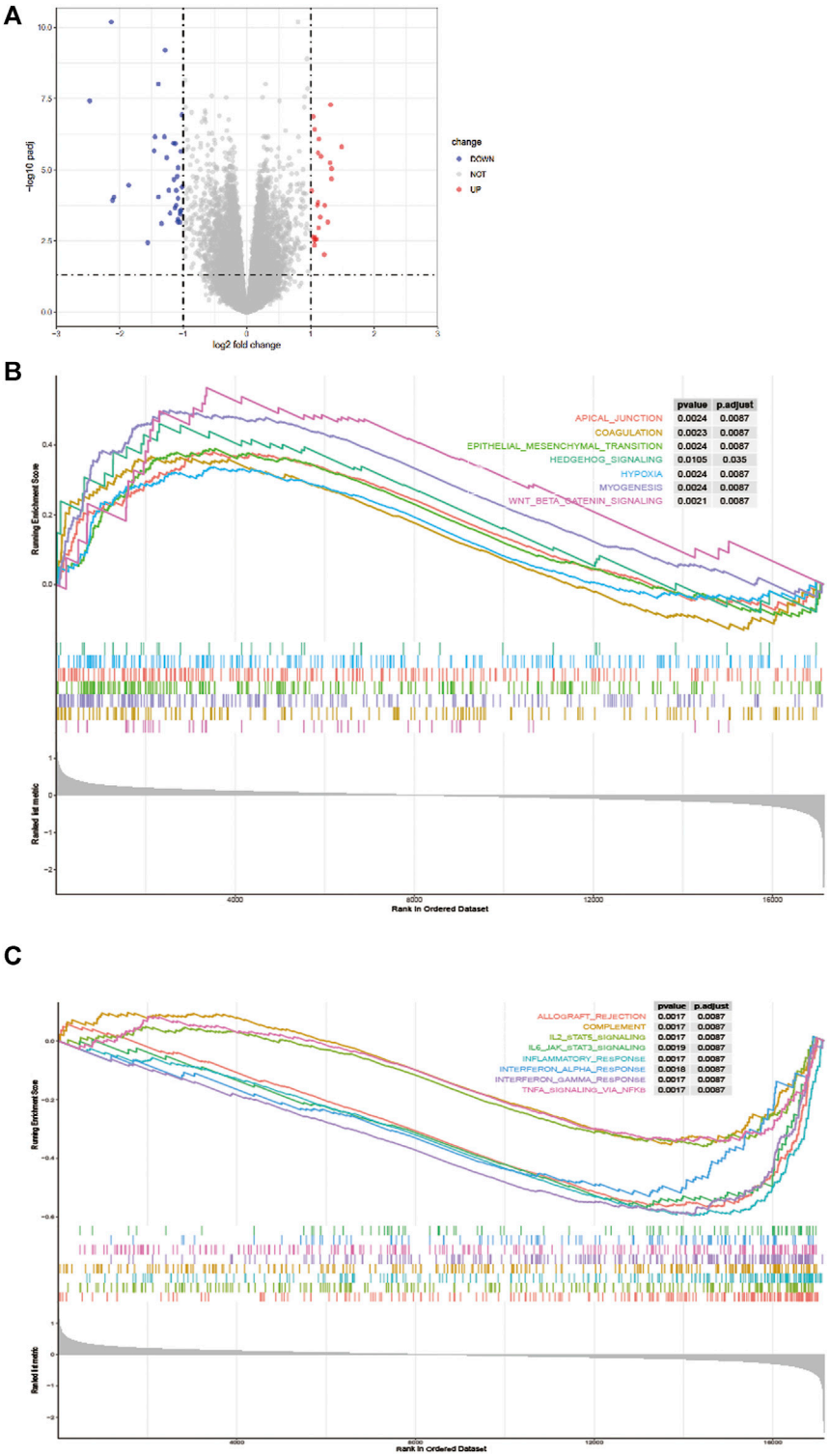

FIGURE 5 | Difference analysis. **(A)** A volcano map shows different EMT-related genes of the high - and low-risk groups. Gene Set Enrichment Analysis (GSEA) for identifying the significant pathway associated with the high-risk group **(B)** and low-risk group **(C)**.

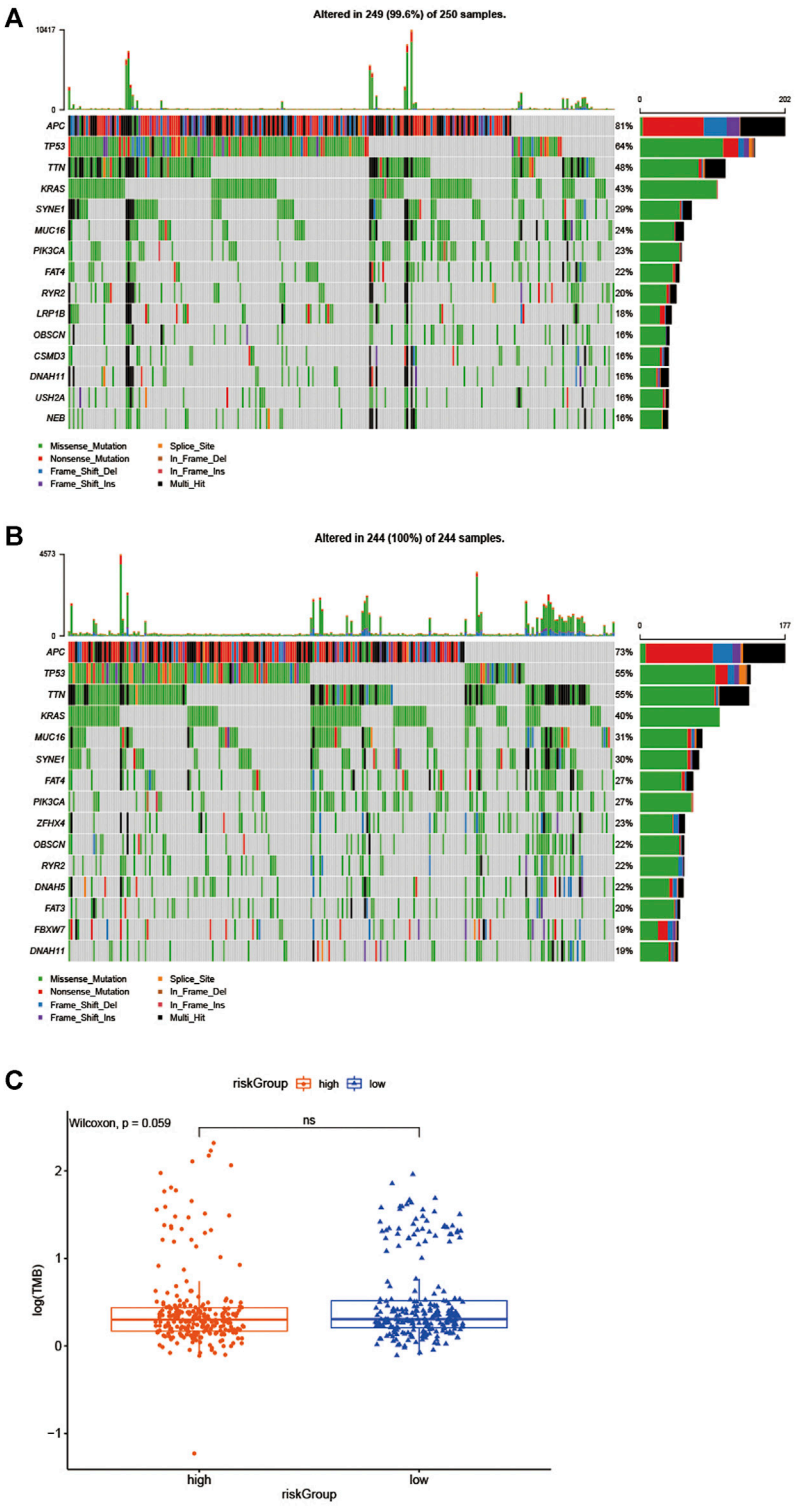

FIGURE 6 | Biological insight into the high- and low-risk group. **(A, B)** significantly mutated genes in the mutated CRC samples of the high - and low-risk groups. **(C)** The proportions of TME cells in the high - and low-risk groups.

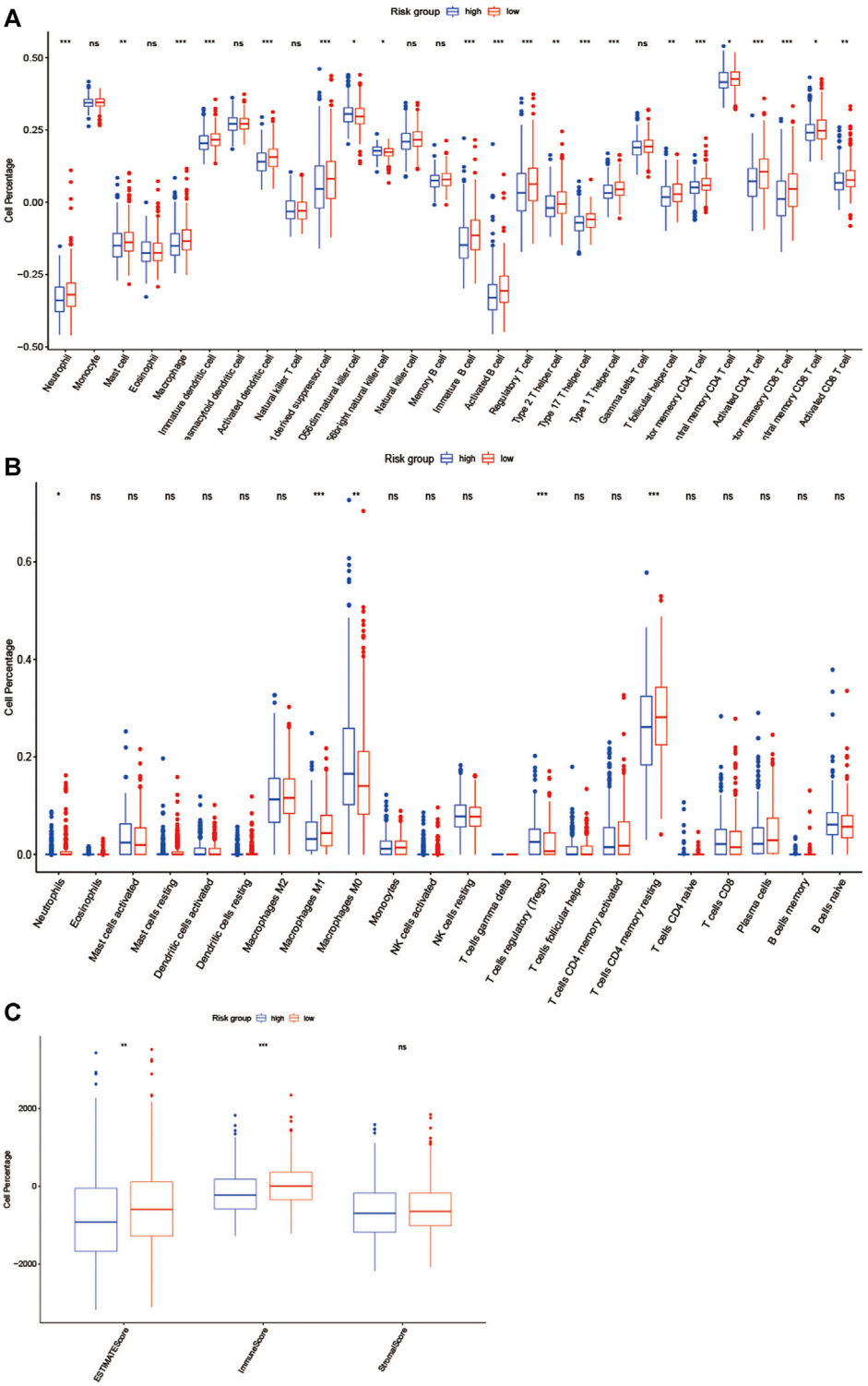

FIGURE 7 | EMT-related lncRNA clusters significantly associated with the immune microenvironment. **(A, B)** Statistical differences in each type of immune cell between high-risk group and low-risk group using ssGSEA approach **(A)** and CIBERSORT algorithm **(B)**. **(C)**Stromal score and immune score were calculated via ESTIMATE method between high-risk group and low-risk group. (ns: not significant, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$).

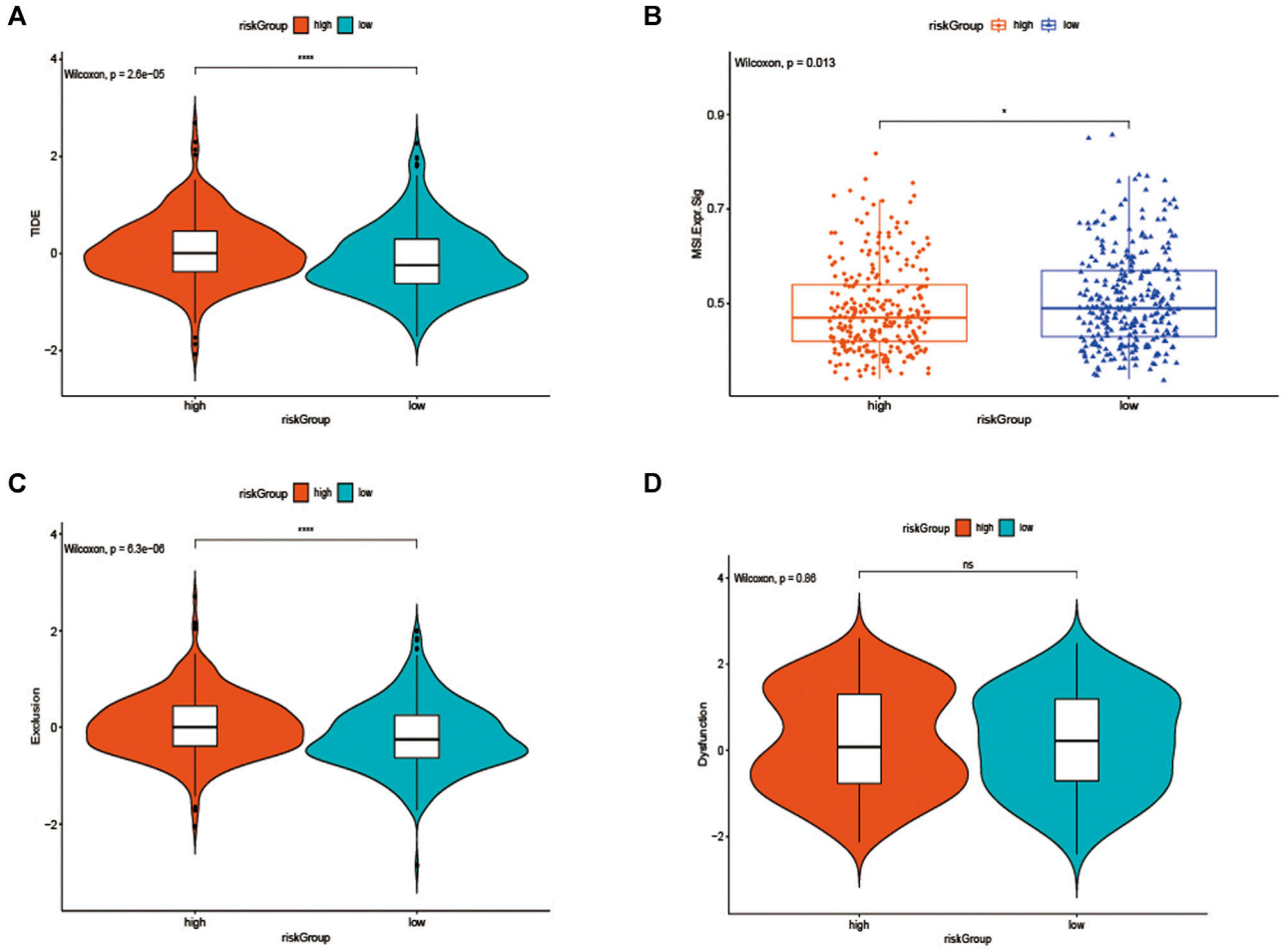

FIGURE 8 | (A) TIDE, **(B)** MSI, and T cell exclusion **(C)** and dysfunction score **(D)** in the high-risk Group and low-risk group. The score between the two subgroups were compared through the Wilcoxon test (ns: not significant, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

regulating the tumor EMT process, in the present study, we aim to address the prognostic value of EMT-related lncRNAs in CRC.

Based on the TCGA dataset, we established an innovative and efficient EMT-related lncRNA signature, and then its validity was verified on the validation set, the ROC analysis results revealed its high prognostic value in two data set, besides, the signature showed a significant correlation with the TNM stage, furthermore, our data showed that low-risk group patients had a better OS. As the EMT-related lncRNAs were considered as a potential predictor for OS, as well as age and stage, a nomogram was constructed based on the above factors and showed promising performance in the 1-, three- and 5 years, which may distinguish individualized treatment of in CRC patients. In brief, our data revealed that a marvelous prognostic value of our EMT-related lncRNA signature, which may provide a theoretical basis for EMT-related targeted therapies for CRC. Moreover, in GSEA analysis, the results indicated that different pathways related to the progression of tumor were significantly enriched in the low-risk group and high-risk group, however, immune response-related pathways mainly enriched in the low-risk group and tumor metastasis-related pathways mainly plays a regulatory role in the high-risk group.

The EMT-related lncRNA signature was made up of 11 lncRNAs, and the molecular mechanism or prognostic value of them has not been exposed, due to their high prognostic value, subsequent experiments are needed to clarify their role in CRC.

Additionally, to explore biological characteristics of the subgroups in the training cohort, we then studied gene mutations of the high- and low-risk group. The results showed that missense variations were the most common type in the two groups, and significant variation differences between the two groups were APC and TP53, which were more common in the high-risk group than low-risk group (81 vs 73% and 64 vs55%), In Michael J et al. study, they revealed that APC and TP53 mutation is the most strongly negatively associated with MSI but positively associated with distant metastasis, which suggested a worse prognosis (Schell et al., 2016). Furthermore, in TIMER 2.0

database, we found that mutations in APC and TP53 genes can reduce the infiltration of CD4 + T lymphocytes and CD8 + T lymphocytes. In addition, we also found that high mutations in USH2A and NEB genes in high-risk groups lead to decreased infiltration of CD4 + T lymphocytes and CD8 + T lymphocytes in tumor centers and increased infiltration of Treg cells, which may be a factor leading to the characteristics of active immune response and low invasive tumor phenotype in patients in the low risk group. In the low-risk group, the high mutation rate of DNAH5, FAT3 and FBXW7 genes also led to the increase of central CD4 + T lymphocyte and CD8 + T lymphocyte infiltration and the decrease of Treg cell infiltration. Those may partly explain that the higher mutation rates of APC, TP53, USH2A and NEB genes lead to a worse survival prognosis in the high-risk group. What's more, it is well known that TMB has been shown to be a potential biomarker for predicting ICI treatment response in many tumor types (Goodman et al., 2017; Jardim et al., 2021), our results revealed that the TMB was slightly higher in the low-risk group than that in the high-risk group, we thought that may partly explain the low-risk group was more sensitive to immunotherapy.

To further understand the immune characteristics of the two groups. The ssGSEA method was used to further evaluate the immune-cell infiltration status of TCGA colorectal cancer transcriptome, and the results suggested that neutrophils, macrophage M1 cells, T cells, and CD4 memory resting cells were enriched in the low-risk group, while M0 cells and T cell regulatory cells were more common in the high-risk group, numerous studies have shown that dense infiltration of T cells, especially cytotoxic CD8 T cells, and high density of M1 macrophages may be associated with acute inflammation, suggesting a good prognosis (Fuchs et al., 2019; Marcelis et al., 2020). In contrast, in many malignancies, M2 macrophages (the major subtype of macrophages) are associated with chronic inflammation and contribute to tumor growth and the development of aggressive phenotypes and have been associated with adverse outcomes (Mantovani et al., 2002; Yamaguchi et al., 2016), and it is noteworthy that our findings support these conclusions. Furthermore, according to the ESTIMATE algorithm, we identified that estimatescore and immunescore were much higher in the low-risk group, which suggests that the low-risk group had more immune components in TME, implying a favorable immunotherapy strategy.

It has been reported that TIDE is used to identify the underlying factors of two mechanisms of tumor immune escape: induction of T cell dysfunction in tumors with high cytotoxic T lymphocyte (CTL) invasion, and prevention of T cell invasion in tumors with low CTL levels (Wang et al., 2020b; Fu et al., 2020; Tsukada et al., 2020), interestingly, in our study, we also discovered that the low-risk group not only had a higher MSI score and lower TIDE score, but also had a lower T cell exclusion score, when compared to the high-risk group, even if there was no difference in T cell dysfunction between the two subgroups, those results suggested that these low-risk group patients had lower levels of immune escape and more MSI, and the higher mutational burden makes the tumor immunogenic and sensitive to PD1 therapy (Lin et al., 2020).

In the current study, we employed Univariate Cox analysis and LASSO algorithms to select significant candidate EMT-related lncRNAs for further multivariate Cox regression to construct the prognostic signature, and stratified analysis revealed that the signature was significantly associated with TNM stages. Furthermore, we used ssGSEA, CIBERSORT algorithm and the ESTIMATE method to assess the relative immune cell infiltrations of each sample. Differentially infiltration of immune cells and diverse tumor mutation burden (TMB) scores might give rise to the efficacy of lncRNA signature for predicting the sensitivity of immunotherapy for CRC patients. The effective signature we constructed was due to the TCGA database with sufficient tumor samples and complete clinical data.

Previous methods to study tumor immune microenvironment include immunohistochemistry and flow cytometry, both of which are inevitably limited to narrow views when comprehensively analyzing the composition of immune cells, and flow cytometry may lead to cytolysis of some cell types. In this study, the gene expression profile and clinical information of colorectal cancer were downloaded from TCGA database, and CIBERPORT, ESTIMATE and ssGSEA algorithm, general gene expression based evolutionary algorithm, are used to quantify cell components from gene expression profiles of large tissues. Therefore, different types of infiltrating immune cells can be quantified at the same time, so that the method avoids the concerns of various surface markers and possible cell separation. Of course, there are some limitations in using public database analysis, for example, in our study, we used the TIDE score to evaluate the potential clinical efficacy of the signature on immunotherapy, our results suggest that the TIDE score of the high-risk group is slightly higher than that of the low-risk group, but there is no significant statistical difference, this may be due to the insufficient number of CRC cases in TCGA database. Moreover, the signature also lacks external clinical samples to verify its effectiveness, which is also the disadvantage of using our method in public database, which depends on further improvement in future work. In one word, although the EMT-related lncRNA signature we developed is somewhat innovative, there are some limitations: the risk signature is established based on the TCGA public database, however, there is no strong external data to verify the effectiveness and practicability. Furthermore, the TCGA database was of limited size, and important clinical information was missing, which can lead to potential basis or errors.

CONCLUSION

Collectively, our study developed and validated an EMT-related lncRNA signature that could be used as a certain reliable tool for predicting individual prognosis and decision-making in the treatment of patients with CRC.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

DL and HW carried out data analysis. XL and BC drafted the manuscript; YZ and ZM participated in study design and data collection. All authors read and approved the final manuscript.

FUNDING

This work was supported by grants from the Shantou Science and Technology Bureau (No. 2006241552607) and "Dengfeng Project" for the construction of high-level hospitals in Guangdong Province–the First Affliated Hospital of Shantou University Medical College Supporting Funding (No. 202003–17).

ACKNOWLEDGMENTS

The authors appreciated TCGA database for providing the original data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.723802/full#supplementary-material

Supplementary Figure 1 | **(A–C)** The high mutation rate of USH2A, NEB genes in the high-risk group related to the decreased infiltration of CD4 + and CD8 + T lymphocytes, while increased infiltration of Treg cells, except CSMD3.

Supplementary Figure 2 | **(A–C)** The high mutation rate of DNAH5, FAT3, and FBXW7 genes in the low-risk group related to the increased infiltration of CD4 + and CD8 + T lymphocytes, while decreased infiltration of Treg cells.

Supplementary Figure 3 | **(A, B)** The higher mutation rate of APC and TP53 genes in the high-risk group led to the decreased infiltration of CD4 + and CD8 + T lymphocytes.

REFERENCES

Aiello, N. M., and Kang, Y. (2019). Context-dependent EMT Programs in Cancer Metastasis. *J. Exp. Med.* 216 (5), 1016–1026. doi:10.1084/jem.20181827

Chaffer, C. L., San Juan, B. P., Lim, E., and Weinberg, R. A. (2016). EMT, Cell Plasticity and Metastasis. *Cancer Metastasis Rev.* 35, 645–654. doi:10.1007/s10555-016-9648-7

Diepenbruck, M., and Christofori, G. (2016). Epithelial-mesenchymal Transition (EMT) and Metastasis: Yes, No, Maybe? *Curr. Opin. Cel Biol.* 43, 7–13. doi:10.1016/j.ceb.2016.06.002

Dongre, A., and Weinberg, R. A. (2019). New Insights into the Mechanisms of Epithelial-Mesenchymal Transition and Implications for Cancer. *Nat. Rev. Mol. Cel Biol* 20 (2), 69–84. doi:10.1038/s41580-018-0080-4

Fabrizio, D. A., George Jr, T. J., Dunne, R. F., Frampton, G., Sun, J., Gowen, K., et al. (2018). Beyond Microsatellite Testing: Assessment of Tumor Mutational burden Identifies Subsets of Colorectal Cancer Who May Respond to Immune Checkpoint Inhibition. *J. Gastrointest. Oncol.* 9, 610–617. doi:10.21037/jgo.2018.05.06

Fu, J., Li, K., Zhang, W., Wan, C., Zhang, J., Jiang, P., et al. (2020). Large-scale Public Data Reuse to Model Immunotherapy Response and Resistance. *Genome Med.* 12, 21. doi:10.1186/s13073-020-0721-z

Fuchs, Y. F., Sharma, V., Eugster, A., Kraus, G., Morgenstern, R., Dahl, A., et al. (2019). Gene Expression-Based Identification of Antigen-Responsive CD8+ T Cells on a Single-Cell Level. *Front. Immunol.* 10, 2568. doi:10.3389/fimmu.2019.02568

Ghiringhelli, F., and Fumet, J. D. (2019). Is There a Place for Immunotherapy for Metastatic Microsatellite Stable Colorectal Cancer? *Front. Immunol.* 10, 1816. doi:10.3389/fimmu.2019.01816

Goodman, A. M., Kato, S., Bazhenova, L., Patel, S. P., Frampton, G. M., Miller, V., et al. (2017). Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol. Cancer Ther.* 16 (11), 2598–2608. doi:10.1158/1535-7163.mct-17-0386

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., et al. (2016). Genomic and Transcriptomic Features of Response to AntiPD-1 Therapy in Metastatic Melanoma. *Cell* 165, 35–44. doi:10.1016/j.cell.2016.02.065

Jardim, D. L., Goodman, A., de Melo Gagliato, D., and Kurzrock, R. (2021). The Challenges of Tumor Mutational Burden as an Immunotherapy Biomarker. *Cancer Cell* 39, 154–173. doi:10.1016/j.ccell.2020.10.001

Kong, J., Sun, W., Li, C., Wan, L., Wang, S., Wu, Y., et al. (2016). Long Non-coding RNA LINC01133 Inhibits Epithelial-Mesenchymal Transition and Metastasis in Colorectal Cancer by Interacting with SRSF6. *Cancer Lett.* 380, 476–484. doi:10.1016/j.canlet.2016.07.015

Ledys, F., Klopfenstein, Q., Truntzer, C., Arnould, L., Vincent, J., Bengrine, L., et al. (2018). RAS Status and Neoadjuvant Chemotherapy Impact CD8+ Cells and Tumor HLA Class I Expression in Liver Metastatic Colorectal Cancer. *J. Immunotherapy Cancer* 6, 123. doi:10.1186/s40425-018-0438-3

Lin, A., Zhang, J., and Luo, P. (2020). Crosstalk between the MSI Status and Tumor Microenvironment in Colorectal Cancer. *Front. Immunol.* 11, 2039. doi:10.3389/fimmu.2020.02039

Ma, X., Bi, E., Lu, Y., Su, P., Huang, C., Liu, L., et al. (2019). Cholesterol Induces CD8+ T Cell Exhaustion in the Tumor Microenvironment. *Cel Metab.* 30 (1), 143–156. Epub 2019 Apr 25. doi:10.1016/j.cmet.2019.04.002

Mantovani, A., Sozzani, S., Locati, M., Allavena, P., and Sica, A. (2002). Macrophage Polarization: Tumor-Associated Macrophages as a Paradigm for Polarized M2 Mononuclear Phagocytes. *Trends Immunol.* 23 (11), 549–555. doi:10.1016/s1471-4906(02)02302-5

Marcelis, L., Antoranz, A., Delsupehe, A.-M., Biesemans, P., Ferreiro, J. F., Debackere, K., et al. (2020). In-depth Characterization of the Tumor Microenvironment in central Nervous System Lymphoma Reveals Implications for Immune-Checkpoint Therapy. *Cancer Immunol. Immunother.* 69 (9), 1751–1766. doi:10.1007/s00262-020-02575-y

Mittal, V. (2018). Epithelial Mesenchymal Transition in Tumor Metastasis. *Annu. Rev. Pathol.* 13, 395–412. doi:10.1146/annurev-pathol-020117-043854

Nieto, M. A., Huang, R. Y.-J., Jackson, R. A., and Thiery, J. P. (2016). Emt: 2016. *Cell* 166, 21–45. doi:10.1016/j.cell.2016.06.028

Schell, M. J., Yang, M., Teer, J. K., Lo, F. Y., Madan, A., Coppola, D., et al. (2016). A Multigene Mutation Classification of 468 Colorectal Cancers Reveals a Prognostic Role for APC. *Nat. Commun.* 7, 11743. doi:10.1038/ncomms11743

Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G. S., Barzi, A., et al. (2017). Colorectal Cancer Statistics, 2017. *CA: A Cancer J. Clinicians* 67 (3), 177–193. doi:10.3322/caac.21395

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., et al. (2020). Colorectal Cancer Statistics, 2020. *CA A. Cancer J. Clin.* 70 (3), 145–164. doi:10.3322/caac.21601

Terry, S., Savagner, P., Ortiz-Cuaran, S., Mahjoubi, L., Saintigny, P., Thiery, J.-P., et al. (2017). New Insights into the Role of EMT in Tumor Immune Escape. *Mol. Oncol.* 11 (7), 824–846. doi:10.1002/1878-0261.12093

Tsukada, T., Kinoshita, J., and Oyama, K. (2020). Dentification and Validation of Stromal-Tumor Microenvironment-Based Subtypes Tightly Associated with

PD-1/pd-L1 Immunotherapy and Outcomes in Patients with Gastric Cancer. *Cancer Cel Int* 20, 92.

Wang, H., Yu, M., Hu, W., Chen, X., Luo, Y., Lin, X., et al. (2020). Linc00662 Promotes Tumorigenesis and Progression by Regulating miR-497-5p/AVL9 axis in Colorectal Cancer. *Front. Genet.* 10, 1385. doi:10.3389/fgene.2019.01385

Wang, Q., Li, M., Yang, M., Yang, Y., Song, F., Zhang, W., et al. (2020). Analysis of Immune-Related Signatures of Lung Adenocarcinoma Identified Two Distinct Subtypes: Implications for Immune Checkpoint Blockade Therapy. *Aging (Albany NY)* 12, 3312–3339. doi:10.18632/aging.102814

Yamaguchi, T., Fushida, S., Yamamoto, Y., Tsukada, T., Kinoshita, J., Oyama, K., et al. (2016). Tumor-associated Macrophages of the M2 Phenotype Contribute to Progression in Gastric Cancer with Peritoneal Dissemination. *Gastric Cancer* 19 (4), 1052–1065. doi:10.1007/s10120-015-0579-8

Zhang, L., Zhao, Y., Dai, Y., Cheng, J.-N., Gong, Z., Feng, Y., et al. (2018). Immune Landscape of Colorectal Cancer Tumor Microenvironment from Different Primary Tumor Location. *Front. Immunol.* 9, 1578. doi:10.3389/fimmu.2018.01578

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 September 2021
doi: 10.3389/fgene.2021.739520

Construction of a Prognostic Model in Lung Adenocarcinoma Based on Ferroptosis-Related Genes

Min Liang[1]†, Mafeng Chen[2], Yinghua Zhong[3], Shivank Singh[4] and Shantanu Singh[5]*

[1]Department of Respiratory and Critical Care Medicine, Maoming People's Hospital, Maoming, China, [2]Department of Otolaryngology, Maoming People's Hospital, Maoming, China, [3]Department of Pediatrics, Fogang County Hospital of Traditional Chinese Medicine, Qingyuan, China, [4]Southern Medical University, Guangzhou, China, [5]Division of Pulmonary, Critical Care and Sleep Medicine, Marshall University, Huntington, WV, United States

Background: Lung adenocarcinoma is one of the most common malignant tumors of the respiratory system, ranking first in morbidity and mortality among all cancers. This study aims to establish a ferroptosis-related gene-based prognostic model to investigate the potential prognosis of lung adenocarcinoma.

Methods: We obtained gene expression data with matching clinical data of lung adenocarcinoma from the The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases. The ferroptosis-related genes (FRGs) were downloaded from three subgroups in the ferroptosis database. Using gene expression differential analysis, univariate Cox regression, and LASSO regression analysis, seven FRGs with prognostic significance were identified. The result of multivariate Cox analysis was utilized to calculate regression coefficients and establish a risk-score formula that divided patients with lung adenocarcinoma into high-risk and low-risk groups. The TCGA results were validated using GEO data sets. Then we observed that patients divided in the low-risk group lived longer than the overall survival (OS) of the other. Then we developed a novel nomogram including age, gender, clinical stage, TNM stage, and risk score.

Results: The areas under the curves (AUCs) for 3- and 5-years OS predicted by the model were 0.823 and 0.852, respectively. Calibration plots and decision curve analysis also confirmed the excellent predictive performance of the model. Subsequently, gene function enrichment analysis revealed that the identified FRGs are important in DNA replication, cell cycle regulation, cell adhesion, chromosomal mutation, oxidative phosphorylation, P53 signaling pathway, and proteasome processes.

Conclusions: Our results verified the prognostic significance of FRGs in patients with lung adenocarcinoma, which may regulate tumor progression in a variety of pathways.

Keywords: lung adenocarcinoma, ferroptosis, gene, prognosis, prognostic

OPEN ACCESS

Edited by:
Jiangning Song,
Monash University, Australia

Reviewed by:
Chen Li,
Monash University, Australia
Yong-Zi Chen,
Tianjin Medical University Cancer Institute and Hospital, China

***Correspondence:**
Min Liang
imtuaska@163.com

†***ORCID:***
Min Liang
orcid.org/0000-0002-6313-2022

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

Received: *11 July 2021*
Accepted: *30 August 2021*
Published: *22 September 2021*

Citation:
Liang M, Chen M, Zhong Y, Singh S and Singh S (2021) Construction of a Prognostic Model in Lung Adenocarcinoma Based on Ferroptosis-Related Genes. Front. Genet. 12:739520. doi: 10.3389/fgene.2021.739520

Abbreviations: AUCs, areas under the curves; DEGs, differentially expressed genes; FDR, false discovery rate; FRGs, ferroptosis-related genes; GEO, Gene Expression Omnibus; GO, Gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomics; NSCLC, non-small cell lung cancer; OS, overall survival; ROC, receiver operating characteristic; TCGA, The Cancer Genome Atlas Program

BACKGROUND

Lung cancer is the most commonly diagnosed malignant tumor worldwide, whose morbidity and mortality rate rank first among all cancers, and the severity is increasing year by year, posing a great threat to human health (Sung et al., 2021). According to pathological classification, the disease can be categorized as small cell lung adenocarcinoma and non-small cell lung cancer (NSCLC), among which, the latter accounts for about 2/3 cases (Rodríguez-Martínez et al., 2018). NSCLC can be divided into three types, including lung adenocarcinoma, squamous cell lung cancer, and non-small cell lung cancer, of which lung adenocarcinoma accounts for about 40% of cases (Bender, 2014). Epidemiological data reveal that the 5-years overall survival rate of lung cancer in all stages is as low as 15.9% (Ettinger et al., 2013). Therefore, it is of great importance to find biomarkers closely associated with the prognostic outcomes of lung cancer, especially lung adenocarcinoma, as well as to evaluate the prognosis outcome of patients with squamous cell lung cancer through these markers, which can improve the prognosis and formulate individualized diagnosis and treatment (Santarpia et al., 2020).

Ferroptosis, a relatively novel kind of cell death discovered recently, is involved in the pathophysiological process of many diseases including tumors (Wu et al., 2019), and it is different from apoptosis, necrosis, and autophagy due to a feature: being iron-dependent. It is caused by the accumulation of toxic lipid reactive oxygen species and the consumption of polyunsaturated fatty acids (Li et al., 2020). Polyunsaturated fatty acid is an important substrate in ferroptosis, and the C–H bond in the diallyl group of polyunsaturated fatty acid is easily attacked by oxidation. Compared with normal cells, cancer cells have the phenomenon of iron ion aggregation, and the regulation of ferroptosis from the perspective of iron homeostasis can effectively kill tumor cells (Lei et al., 2021). In recent years, for the treatment of advanced tumors, especially drug-resistant tumors, inducing the death of cancer cells through ferroptosis has become a very promising option (Xu et al., 2021). In addition to various induction molecules, many genes can also be markers of ferroptosis (Chen et al., 2021a). At present, ferroptosis-related genes have shown good predictive performance in not a few tumors, including glioma (Zhuo et al., 2020), liver cancer (Tang et al., 2020), pancreatic cancer (Jiang et al., 2021a), gastroenteric tumor (Angius et al., 2019), urologic neoplasms (Liu et al., 2021), and thyroid cancer (Ge et al., 2021). However, the relationship between FRGs and prognosis and the outcome of patients with lung adenocarcinoma has not been reported in depth.

Therefore, our study aims to explore the role of FRGs to predict outcome, and on the basis of which, establish a prognostic model to assess the prognostic outcome of patients with lung adenocarcinoma. Based on the differentially expressed genes (DEGs) related to ferroptosis, a prognostic model was constructed according to the training set, and the predictive power of the model was verified in the validation set. Finally, we proceeded with a functional enrichment analysis to investigate the biological mechanism of FRGs in lung adenocarcinoma.

In this study, a prognostic model consisting of seven genes associated with ferroptosis was established with excellent predictive power. Enrichment analysis showed that these genes were associated with the development of lung adenocarcinoma.

METHODS

Resources and Pre-processing

The gene expression data and related clinical information of lung tumors were extracted from the The Cancer Genome Atlas (TCGA) database (https://genome.nih.gov/). Ferroptosis-related gene sets were extracted from three subgroups in the ferroptosis database (http://www.zhounan.org/ferrdb/). Edge R package from R was used to normalize the entire data set, set $|\log2FC| > 0.5$ and false discovery rate (FDR) < 0.05 as the threshold to construct a volcano map, to further obtain differentially expressed ferroptosis-related genes.

Construction of the Prognostic Model

According to downloaded clinical data of lung adenocarcinoma cases from TCGA, patients with an adequate follow-up time (>30 days) were screened and divided into a training set and an internal validation set at a ratio of 2:1. The sets were divided to construct a prognostic model and verify the model, respectively. In addition, two Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) data sets are applied for external validation of the model. Based on the training set, we performed univariate Cox regression analysis on FRGs and survival data and set $p < 0.05$ to identify differential FRGs related to prognosis, and LASSO regression was performed to further screen the genes. After obtaining seven target genes, a multi-factor stepwise Cox regression was performed to analyze their respective coefficients (βi). Finally, a risk-scoring formula consisting of βi and gene expression level (Expi) was constructed as follow:

$$\text{Risk score} = \sum_{i=1}^{i} (\beta i^{*}\text{Exp}_i)$$

According to the model, the risk score of invidual could be acquired. In addition, the median of the risk score was set as a critical value, and all patients included were divided into high- and low-risk groups. To reveal the prognostic outcome difference between the groups Kaplan–Meier survival curve was used. Then, the above results were verified by the validation set. To further assess the ability to predict, we conducted a subgroup analysis to compare the OS between the two groups.

Construction of the Nomogram

We incorporated clinical features including risk score, clinical and TNM stage, age, and gender into the final model to establish a novel nomogram to predict the OS of patients individually. In addition, AUC was obtained through a receiver operating characteristic (ROC) curve to assess the accuracy of the nomogram. Subsequently, we used the calibration chart and decision curve analysis to verify the predictive ability of the

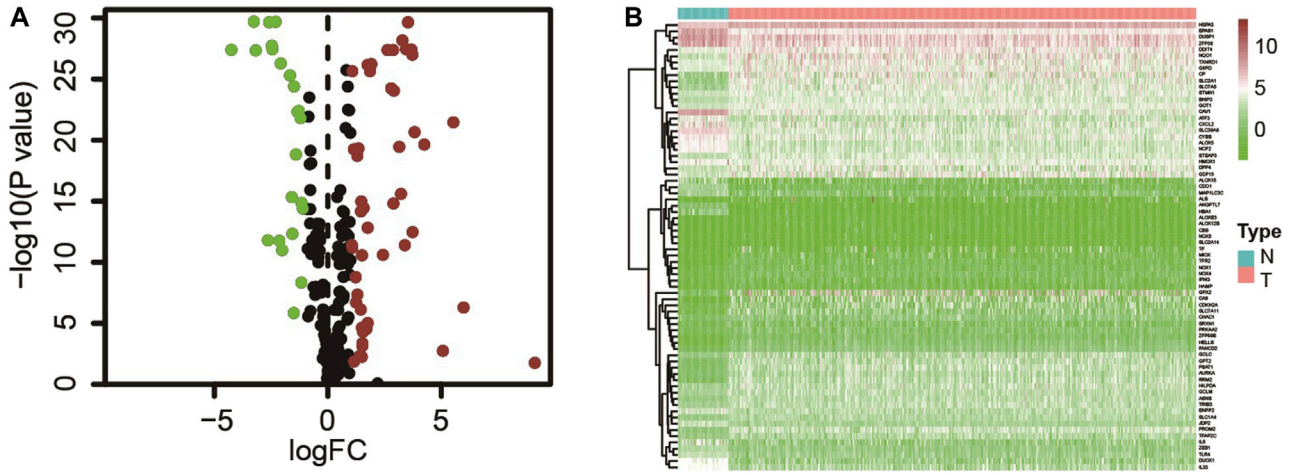

FIGURE 1 | (A) Ferroptosis-related genes differentially expressed in tumor and adjacent tissues; **(B)** Expression of ferroptosis-related genes in tumors and adjacent tissues.

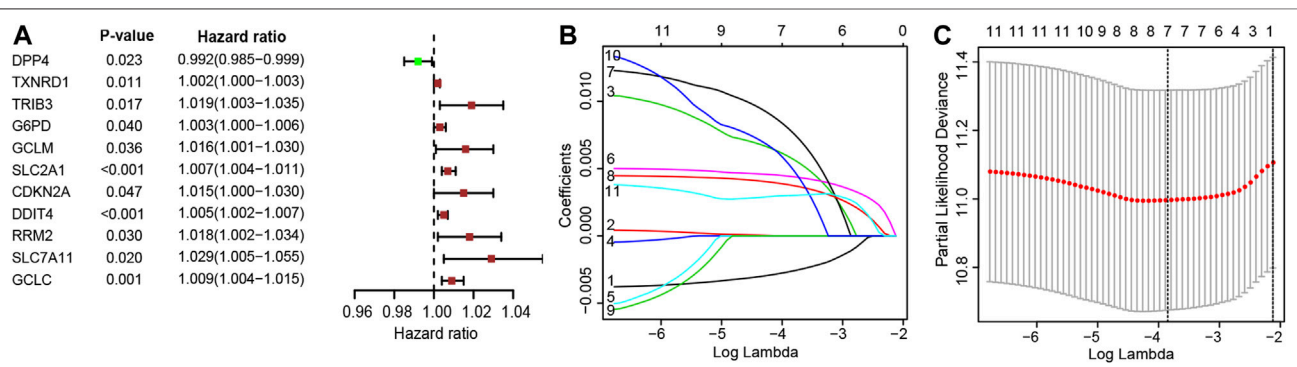

FIGURE 2 | (A) 11 Ferroptosis-related genes significantly associated with overall survival (OS) in lung adenocarcinoma; **(B)** Adjustment parameters in LASSO regression model; **(C)** LASSO coefficient spectrum of ferroptosis-related genes.

model (Ge et al., 2021). The validation set was used to verify the results obtained finally.

Gene Set Enrichment Analysis

The reference set (2. cp.kegg.v6.2. symbols.gmt, c5. all. v6.2. symbols.gmt) was downloaded from the Molecular Signatures Database (http://www.gsea-msigdb.org/gsea/msigdb/index.jsp), and the number of random combinations was set as 1,000 according to the default weighted enrichment method (NPM $p < 0.05$, FDR $p < 0.05$). The study conducted gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomics (KEGG) analysis on the DEGs in the groups and deduced their functions by analyzing gene sets. Therefore, this way it can be used to clarify the question on whether the gene set shows a statistically significant difference between the two biological states. The study explored whether the DEGs between the two groups are enriched during the disease progression as well.

Statistical Analysis

All statistical analyses were conducted using R version 4.1.0 (package: limma, pheatmap, survival, glmnet, survminer, survivalROC, rms, and timeROC). Univariate and multivariate Cox regression were used to analyze the correlation between clinical features, risk scores, and the OS of patients. The ROC curve, C-index, the calibration curve, and DCA curve were used to assess the predictive power of the model. Two-tailed $p < 0.05$ was considered statistically significant.

RESULTS

Extraction of Ferroptosis-Related Genes in Lung Adenocarcinoma

The gene expression result with matching clinical data of lung adenocarcinoma (497 tumor tissues and 54 paracancerous

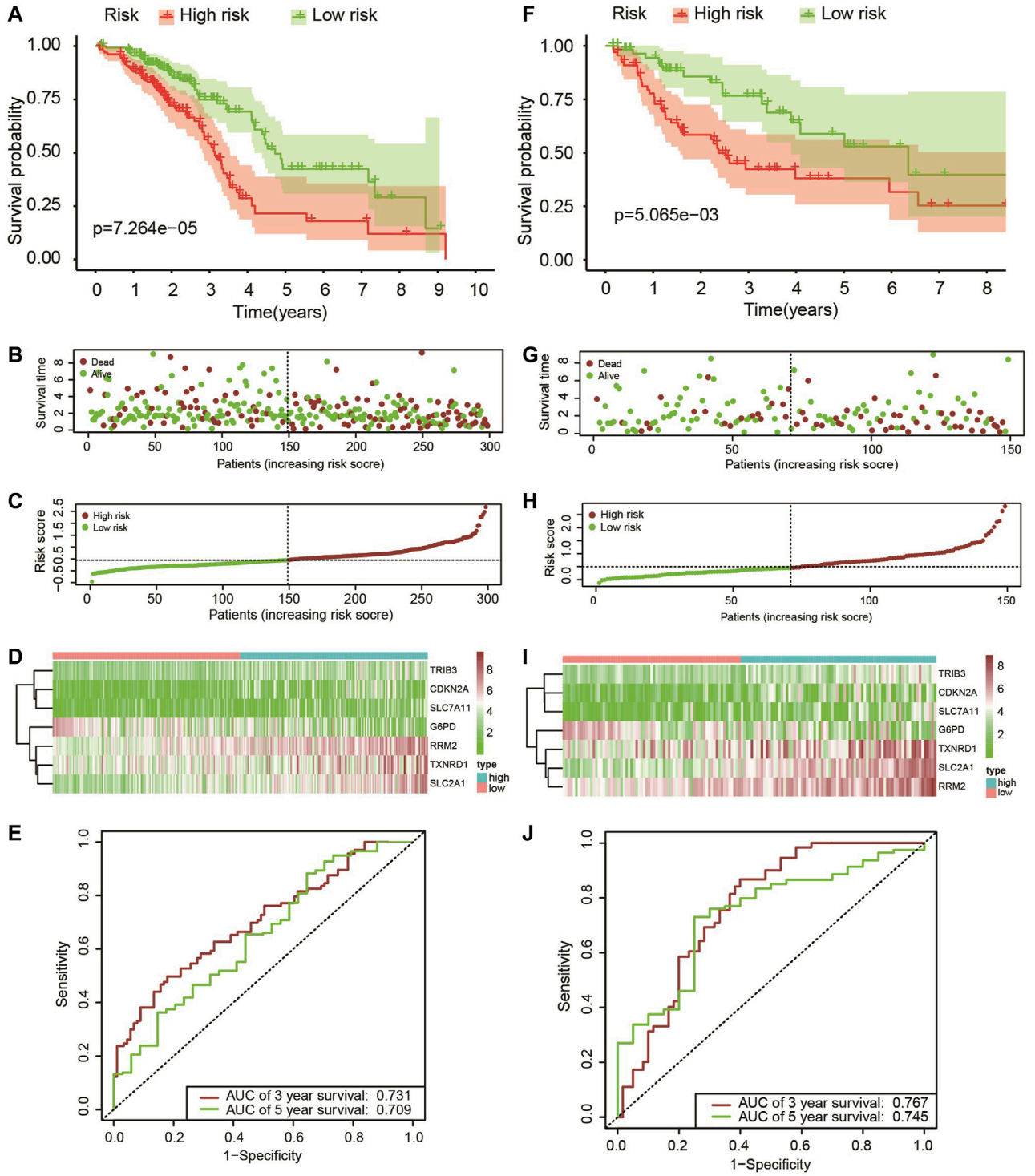

FIGURE 3 | (**A–F**) Kaplan–Meier survival analysis of patients with lung adenocarcinoma in high- and low-risk groups; (**B**,**G**) The overall survival rate and status of patients with lung adenocarcinoma; (**C**,**H**) The distribution of risk scores; (**D**,**I**) The expression level of these seven ferroptosis-related genes in the low-risk group and the high-risk group, the cool color represents low expression, while the warm color represents high expression; (**E**,**J**) 3-years and 5-years ROC curve analysis of the prognostic model.

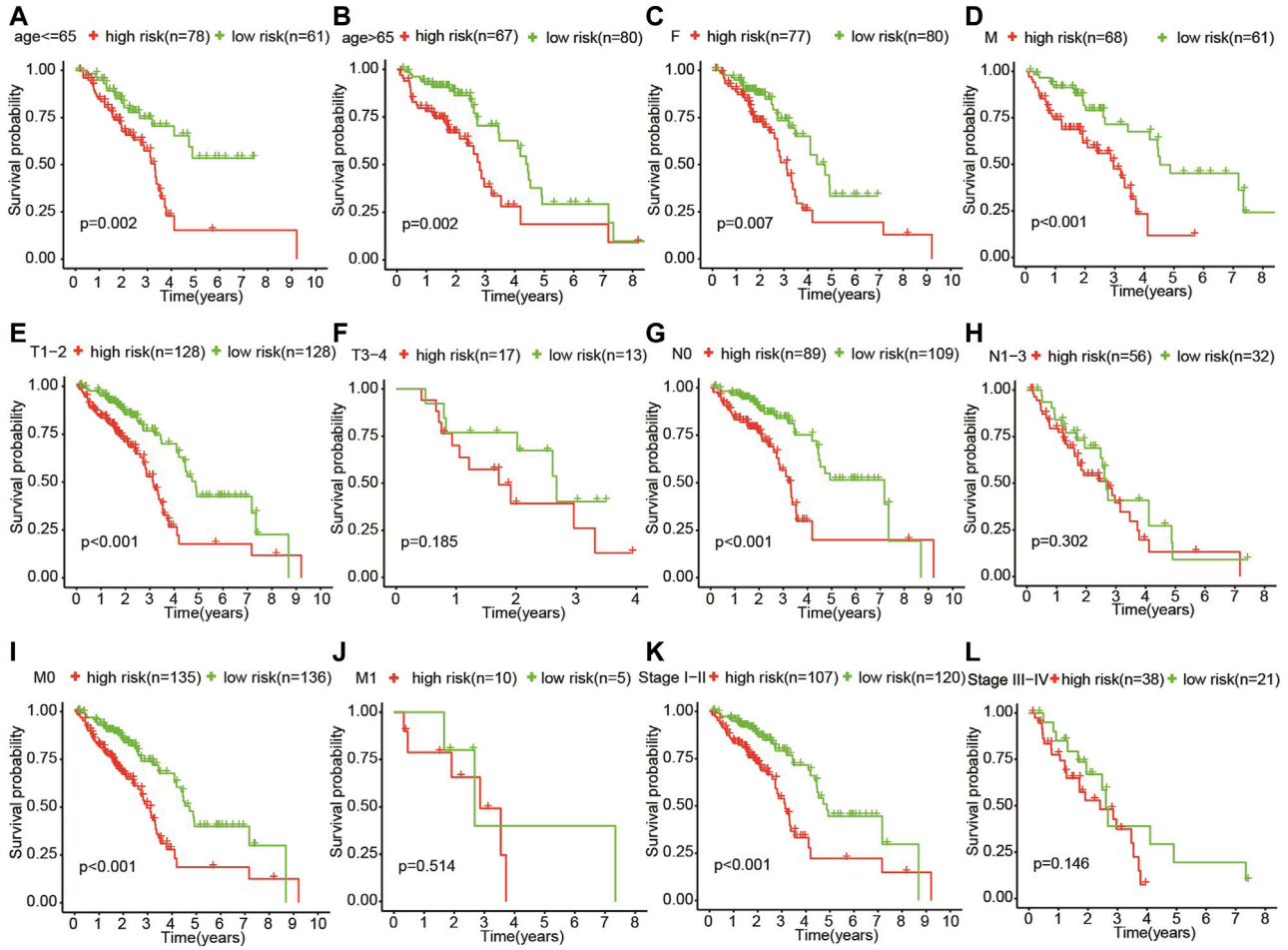

FIGURE 4 | Subgroup analysis showed that according to age (**A**,**B**), gender (**C**,**D**), TNM stage (**E–K**), and clinical stage (**L**,**M**), the OS of patients in the high-risk group was more unfavorable than that in the other group.

tissues) were extracted from the TCGA database, and the ferroptosis-related gene set (259 genes) was obtained from the ferroptosis database. Expression matrices of all ferroptosis-related genes were extracted from the TCGA dataset and differential expression analysis was performed (**Supplementary Table S1**). Seventy-two differential ferroptosis-related genes in lung adenocarcinoma tissues and adjacent tissues were screened out, of which 49 were upregulated, and 23 were downregulated (**Figure 1** and **Supplementary Table S2**).

Construction of the Prognostic Model

The patients who met an adequate follow-up time (>30 days) were divided into training and validation sets at a ratio of 2 to 1. In the training set, we performed univariate Cox regression on FRGs and corresponding clinical survival data, and initially screened 11 FRGs related to prognosis (**Figure 2A**). Seven target genes were further screened by LASSO regression analysis, namely TXNRD1, TRIB3, SLC2A1, CDKN2A, RRM2, SLC7A11, and G6PD (**Figures 2B,C**). The obtained target genes were adapted to calculate the risk score of the individual through the Cox proportional hazard regression model (**Supplementary Table S3**). The median of the risk score was set as a cut-off value, on which basis we divided patients into two groups (**Figure 3**). The Kaplan–Meier curve showed that the OS of the high-risk group was worse than the other group (**Figure 3A**). The risk curves and scatter plots can reveal the risk score and survival status of each patient. As shown in **Figures 3B,C**, the mortality and risk coefficient of the high-risk group were significantly higher than that of the low-risk group. **Figure 3D** displays the expression profile of these seven genes. The ROC curve analysis of the 3- and 5-years OS yielded AUCs of 0.731 and 0.709, respectively (**Figure 3E**). Similar results were observed using the same process in the internal validation set (**Figures 3F–J**) and external validation (GSE37745 and GSE68465) (**Supplementary Figure S1**). Subgroup analysis showed that according to age, gender, TNM stage, or clinical stage, the prognosis of patients in the low-risk group are more favorable (**Figure 4**).

Construction of the Nomogram

We then developed a nomogram, including age, gender, clinical stage, TNM stage, and risk score, as shown in **Figure 5A**. The

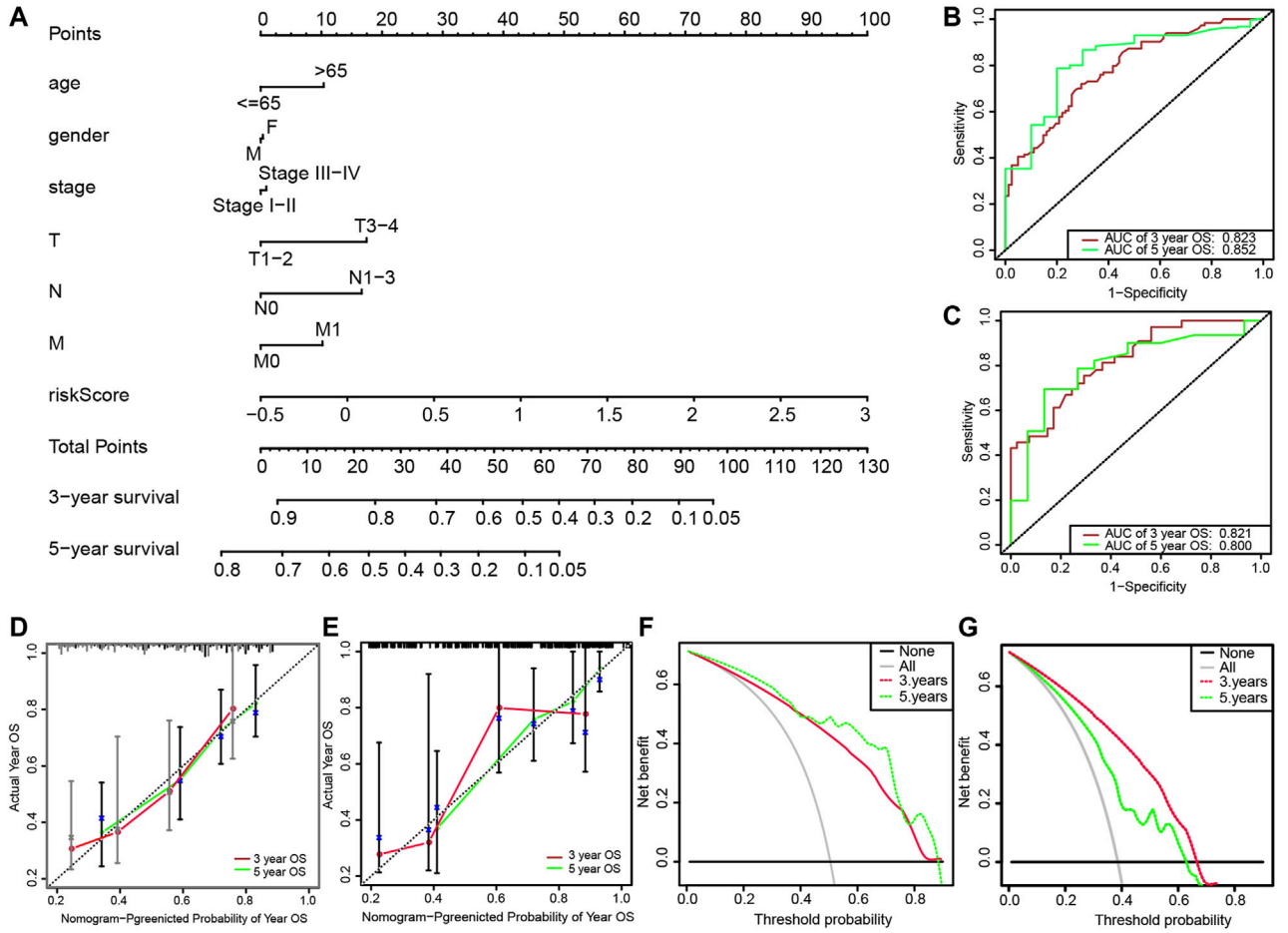

FIGURE 5 | (**A**) Nomogram for predicting 3-years and 5-years survival rates of patients with lung adenocarcinoma; (**B**) ROC curve analysis for predicting 3-years and 5-years OS through nomogram in the training set; (**C**) ROC curve analysis for predicting 3-years and 5-years OS through the nomogram in the validation set; (**D**) Calibration of 3-years survival rate based on the nomogram; (**E**) Calibration of 5-years survival rate based on the nomogram; (**F**) Decision curve analysis of 3-years survival rate based on the nomogram; (**G**) Decision curve analysis of 5-years survival rate based on the nomogram.

ROC curve analysis of the 3- and 5-years OS of the prognostic model yielded AUCs of 0.823 and 0.852, respectively (**Figure 5B**). The established calibration chart and decision curve analysis show that the nomogram has a favorable predictive effect (**Figures 5D–G**). In the validation set, the 3- and 5-years AUCs obtained by analyzing the ROC curve of the novel prediction model were 0.821 and 0.800, respectively, as shown in **Figure 5C**.

Gene Set Enrichment Analysis

The GO and KEGG enrichment analysis were performed on the DEGs of the above-mentioned high-and low-risk groups (**Figure 6**). The result of the analysis disclosed that the gene set was enriched in DNA replication, cell cycle regulation, cell adhesion, and chromosome mutation. As shown by the KEGG pathway enrichment analysis, screened genes were deeply involved in the cell cycle, oxidative phosphorylation, P53 signaling pathway, proteasomes, and so on. These results may provide a direction for researchers to study the mechanism of ferroptosis-related genes on lung adenocarcinoma in the future.

DISCUSSION

Lung adenocarcinoma is a common malignant tumor with a poor prognosis (Devarakonda et al., 2015). Predicting the outcome of tumors is of incontestable clinical significance in the diagnosis and treatment of patients with lung adenocarcinoma (Chen et al., 2021b). Previous studies have shown that chest CT, serum tumor markers, and TNM staging can be used as prognostic indicators of lung cancer (Calvayrac et al., 2017). However, there are certain limitations. In case of a risk of radiation exposure, the sensitivity and specificity are relatively low (Hoseok and Cho, 2015; Welch, 2017). Therefore, it is important to find predictors that can accurately predict the prognosis of patients with lung cancer.

According to new studies, ferroptosis has shown non-negligible potential in cancer treatment, especially for tumors that are not

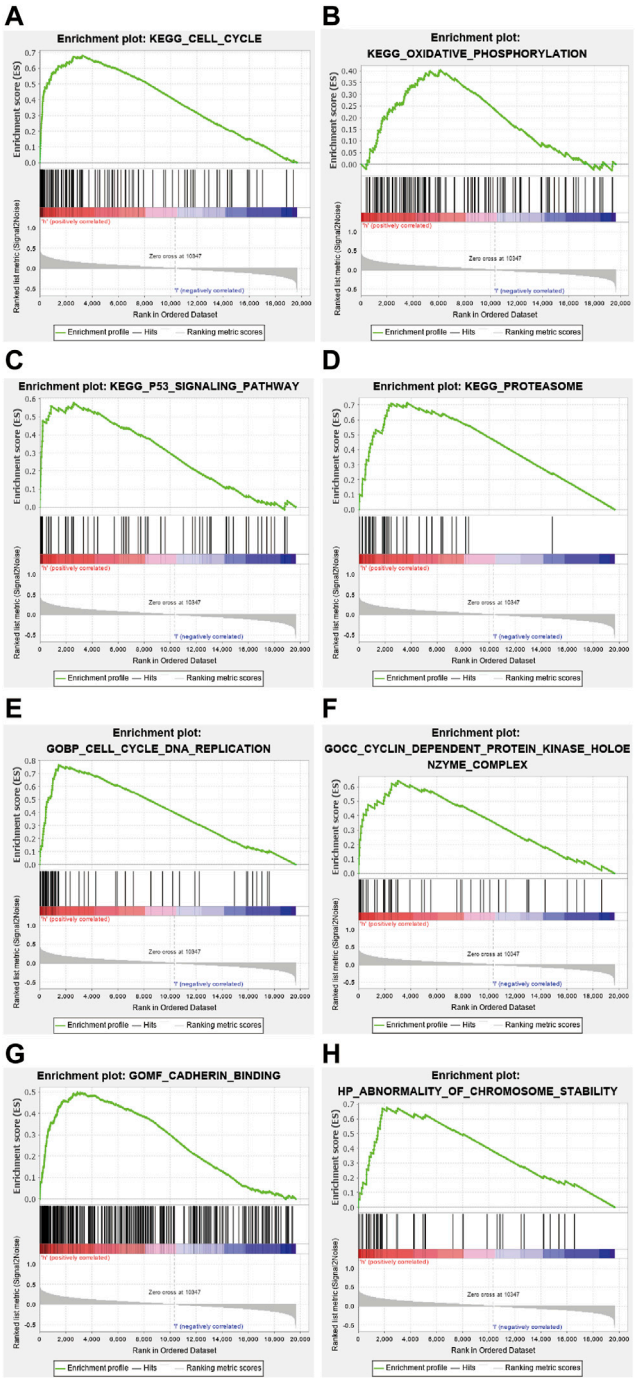

FIGURE 6 | Kyoto Encyclopedia of Genes and Genomics (KEGG) enrichment analysis revealed that genes identified were involved in the following processes (**A**) cell cycle, (**B**) oxidative phosphorylation, (**C**) P53 signaling pathway, and (**D**) proteasome. Gene Ontology (GO) enrichment analysis demonstrated that the gene set was enriched in (**E**) DNA replication, (**F**) cell cycle regulation, (**G**) cell adhesion, and (**H**) chromosome variation.

sensitive to traditional therapies (Liang et al., 2019; Proneth and Conrad, 2019). P53 is a widely studied gene that can suppress tumors, and inhibit the expression of cystine/glutamate antiporter at the transcriptional level to regulate the process of ferroptosis (Xie et al., 2017). In addition, studies have shown that the increase in iron-dependent reactive oxygen species can cause lipid peroxidation outside the mitochondria triggering ferroptosis, thereby inhibiting tumor development (Liang et al., 2019). In lung cancer, due to the upregulation of cystine/glutamate antiporter and the decrease in iron, ferroptosis is usually inhibited, which leads to the relapse and development of tumors (Lai et al., 2019). Therefore, our study aims to explore the relationshipbetween FRGs and the prognosis of patients with lung adenocarcinoma with the described underlying mechanism.

We obtained gene expression data with clinical data of lung adenocarcinoma from the public database. FRGs were extracted from the ferroptosis database. First, we identified seven target genes through DEGs and regression analysis. Multivariate Cox analysis was adapted to calculate regression coefficients and a prognostic model was developed, thereby dividing patients with lung adenocarcinoma into high- and low-risk groups. We observed that patients in the latter group lived longer OS than the other. Furthermore, we developed a nomogram according to the outcomes of multivariate Cox regression. ROC curve, calibration chart, and decision curve confirmed the prediction power of the nomogram. Compared with previous studies, the AUC value of the prognostic model based on ferroptosis-related genes (AUC = 0.823) was higher than that of the prognostic model based on metabolic genes (AUC = 0.767) (Yu et al., 2020), immune genes (AUC = 0.718) (Song et al., 2020), and autophagy genes (AUC = 0.810) (Wang et al., 2020).

A risk scoring model consisting of seven genes (TXNRD1, TRIB3, SLC2A1, CDKN2A, RRM2, SLC7A11, and G6PD) associated with ferroptosis was constructed. Thioredoxin reductase (TXNRD1) is overexpressed in lung cancer cells to maintain tumor survival, and this overexpression has been shown to be associated with clinical outcomes (Zhu et al., 2019). Studies have shown that TRIB3 is significantly upregulated in LUAD cell lines and tissues. TRIB3 gene knockdown significantly inhibited the growth and invasion of LUAD cells (Xing et al., 2020). The progression of lung adenocarcinoma can be inhibited by inhibiting SLC2A1 expression (Wang et al., 2017). CDKN2A is associated with DNA methylation and is closely related to the prognosis of patients (Tsou et al., 2007). Inhibition of RRM2 can activate STING signaling pathway and inhibit the enhancement of radiosensitivity of lung adenocarcinoma (Jiang et al., 2021b). Inhibition of SLC7A11 leads to poor prognosis in KRAS-mutated lung adenocarcinoma (Hu et al., 2020). Previous studies have shown that G6PD is an independent prognostic factor for lung adenocarcinoma (Nagashio et al., 2019).

Subsequently, we conducted a gene function enrichment analysis to reveal the mechanism of ferroptosis genes on lung adenocarcinoma. The results demonstrated that FRGs we screened were involved in DNA replication, cell cycle regulation, cell adhesion, chromosomal mutation, oxidative phosphorylation, P53 signaling pathway, and proteasome processes. Hence, FRGs can be used as predictors of lung adenocarcinoma prognosis and may play a crucial role in lung adenocarcinoma biology.

In conclusion, the study found seven ferroptosis-related genes by searching the database with prognostic value for patients with

lung adenocarcinoma. We constructed a clinical prognostic model of FRGs, which possesses a good effect on predicting the survival rate of patients with lung adenocarcinoma, indicating that FRGs can very well predict the prognosis outcome of patients with lung adenocarcinoma, and play a crucial role in the relapse and development of lung adenocarcinoma.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

ML conceptualized the study. ML and MC provided the sudy design. SvS and StS acquired the data and performed the statistical analyses. YZ and ML performed the data analysis and interpretation. ML and MC prepared the manuscript. SvS edited the manuscript. StS reviewed the manuscript.

FUNDING

The funding for this study was provided by the High-level Hospital Construction Project of Maoming People's Hospital, the Research Project of Maoming Science and Technology Bureau (Grant No. 2021121), and the Outstanding Young Talents Program of Maoming people's hospital. The funders had no role in the design of the study and collection, analysis, and interpretation of the data, and in writing the manuscript.

ACKNOWLEDGMENTS

The authors thank Mrs. Yunru Fan and Dr. Alexandra Lam for providing instructive advice and useful suggestions on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.739520/full#supplementary-material

REFERENCES

Angius, A., Uva, P., Pira, G., Muroni, M. R., Sotgiu, G., Saderi, L., et al. (2019). Integrated Analysis of miRNA and mRNA Endorses a Twenty miRNAs Signature for Colorectal Carcinoma. *Int. J. Mol. Sci.* 20 (16), 4067. doi:10.3390/ijms20164067

Bender, E. (2014). Epidemiology: The Dominant Malignancy. *Nature* 513 (7517), S2–S3. doi:10.1038/513S2a

Calvayrac, O., Pradines, A., Pons, E., Mazières, J., and Guibert, N. (2017). Molecular Biomarkers for Lung Adenocarcinoma. *Eur. Respir. J.* 49 (4), 1601734. doi:10.1183/13993003.01734-2016

Chen, X., Comish, P. B., Tang, D., and Kang, R. (2021). Characteristics and Biomarkers of Ferroptosis. *Front. Cel Dev. Biol.* 9, 637162. doi:10.3389/fcell.2021.637162

Chen, Y., Zitello, E., Guo, R., and Deng, Y. (2021). The Function of LncRNAs and Their Role in the Prediction, Diagnosis, and Prognosis of Lung Cancer. *Clin. Transl. Med.* 11 (4), e367. doi:10.1002/ctm2.367

Devarakonda, S., Morgensztern, D., and Govindan, R. (2015). Genomic Alterations in Lung Adenocarcinoma. *Lancet Oncol.* 16 (7), e342–e351. doi:10.1016/S1470-2045(15)00077-7

Ettinger, D. S., Akerley, W., Borghaei, H., Chang, A. C., Cheney, R. T., Chirieac, L. R., et al. (2013). Non-small Cell Lung Cancer, Version 2.2013. *J. Natl. Compr. Canc Netw.* 11 (6), 645–653. doi:10.6004/jnccn.2013.0084

Ge, M., Niu, J., Hu, P., Tong, A., Dai, Y., Xu, F., et al. (2021). A Ferroptosis-Related Signature Robustly Predicts Clinical Outcomes and Associates with Immune Microenvironment for Thyroid Cancer. *Front. Med.* 8, 637743. doi:10.3389/fmed.2021.637743

Hoseok, I., and Cho, J. Y. (2015). Lung Cancer Biomarkers. *Adv. Clin. Chem.* 72, 107–170. doi:10.1016/bs.acc.2015.07.003

Hu, K., Li, K., Lv, J., Feng, J., Chen, J., Wu, H., et al. (2020). Suppression of the SLC7A11/glutathione axis Causes Synthetic Lethality in KRAS-Mutant Lung Adenocarcinoma. *J. Clin. Invest.* 130 (4), 1752–1766. doi:10.1172/JCI124049

Jiang, P., Yang, F., Zou, C., Bao, T., Wu, M., Yang, D., et al. (2021). The Construction and Analysis of a Ferroptosis-Related Gene Prognostic Signature for Pancreatic Cancer. *Aging* 13 (7), 10396–10414. doi:10.18632/aging.202801

Jiang, X., Li, Y., Zhang, N., Gao, Y., Han, L., Li, S., et al. (2021). RRM2 Silencing Suppresses Malignant Phenotype and Enhances Radiosensitivity via Activating cGAS/STING Signaling Pathway in Lung Adenocarcinoma. *Cell Biosci.* 11 (1), 74. doi:10.1186/s13578-021-00586-5

Lai, Y., Zhang, Z., Li, J., Li, W., Huang, Z., Zhang, C., et al. (2019). STYK1/NOK Correlates with Ferroptosis in Non-small Cell Lung Carcinoma. *Biochem. Biophys. Res. Commun.* 519 (4), 659–666. doi:10.1016/j.bbrc.2019.09.032

Lei, P., Ayton, S., and Bush, A. I. (2021). The Essential Elements of Alzheimer's Disease. *J. Biol. Chem.* 296, 100105. doi:10.1074/jbc.REV120.008207

Li, J., Cao, F., Yin, H.-l., Huang, Z.-j., Lin, Z.-t., Mao, N., et al. (2020). Ferroptosis: Past, Present and Future. *Cell Death Dis.* 11 (2), 88. doi:10.1038/s41419-020-2298-2

Liang, C., Zhang, X., Yang, M., and Dong, X. (2019). Recent Progress in Ferroptosis Inducers for Cancer Therapy. *Adv. Mater.* 31 (51), 1904197. doi:10.1002/adma.201904197

Liu, J., Ma, H., Meng, L., Liu, X., Lv, Z., Zhang, Y., et al. (2021). Construction and External Validation of a Ferroptosis-Related Gene Signature of Predictive Value for the Overall Survival in Bladder Cancer. *Front. Mol. Biosci.* 8, 675651. doi:10.3389/fmolb.2021.675651

Nagashio, R., Oikawa, S., Yanagita, K., Hagiuda, D., Kuchitsu, Y., Igawa, S., et al. (2019). Prognostic Significance of G6PD Expression and Localization in Lung Adenocarcinoma. *Biochim. Biophys. Acta Proteins Proteom.* 1867 (1), 38–46. doi:10.1016/j.bbapap.2018.05.005

Proneth, B., and Conrad, M. (2019). Ferroptosis and Necroinflammation, a yet Poorly Explored Link. *Cell Death Differ.* 26 (1), 14–24. doi:10.1038/s41418-018-0173-9

Rodríguez-Martínez, Á., Torres-Durán, M., Barros-Dios, J. M., and Ruano-Ravina, A. (2018). Residential Radon and Small Cell Lung Cancer. A Systematic Review. *Cancer Lett.* 426, 57–62. doi:10.1016/j.canlet.2018.04.003

Santarpia, M., Aguilar, A., Chaib, I., Cardona, A. F., Fancelli, S., Laguia, F., et al. (2020). Non-Small-Cell Lung Cancer Signaling Pathways, Metabolism, and PD-1/PD-L1 Antibodies. *Cancers* 12 (6), 1475. doi:10.3390/cancers12061475

Song, C., Guo, Z., Yu, D., Wang, Y., Wang, Q., Dong, Z., et al. (2020). A Prognostic Nomogram Combining Immune-Related Gene Signature and Clinical Factors Predicts Survival in Patients with Lung Adenocarcinoma. *Front. Oncol.* 10, 1300. doi:10.3389/fonc.2020.01300

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Tang, B., Zhu, J., Li, J., Fan, K., Gao, Y., Cheng, S., et al. (2020). The Ferroptosis and Iron-Metabolism Signature Robustly Predicts Clinical Diagnosis, Prognosis and

Immune Microenvironment for Hepatocellular Carcinoma. *Cell Commun. Signal* 18 (1), 174. doi:10.1186/s12964-020-00663-1

Tsou, J. A., Galler, J. S., Siegmund, K. D., Laird, P. W., Turla, S., Cozen, W., et al. (2007). Identification of a Panel of Sensitive and Specific DNA Methylation Markers for Lung Adenocarcinoma. *Mol. Cancer* 6, 70. doi:10.1186/1476-4598-6-70

Wang, X., Yao, S., Xiao, Z., Gong, J., Liu, Z., Han, B., et al. (2020). Development and Validation of a Survival Model for Lung Adenocarcinoma Based on Autophagy-Associated Genes. *J. Transl. Med.* 18 (1), 149. doi:10.1186/s12967-020-02321-z

Wang, Y., Shi, S., Ding, Y., Wang, Z., Liu, S., Yang, J., et al. (2017). Metabolic Reprogramming Induced by Inhibition of SLC2A1 Suppresses Tumor Progression in Lung Adenocarcinoma. *Int. J. Clin. Exp. Pathol.* 10 (11), 10759–10769.

Welch, H. G. (2017). Cancer Screening, Overdiagnosis, and Regulatory Capture. *JAMA Intern. Med.* 177 (7), 915–916. doi:10.1001/jamainternmed.2017.1198

Wu, Y., Song, J., Wang, Y., Wang, X., Culmsee, C., and Zhu, C. (2019). The Potential Role of Ferroptosis in Neonatal Brain Injury. *Front. Neurosci.* 13, 115. doi:10.3389/fnins.2019.00115

Xie, Y., Zhu, S., Song, X., Sun, X., Fan, Y., Liu, J., et al. (2017). The Tumor Suppressor P53 Limits Ferroptosis by Blocking DPP4 Activity. *Cel Rep.* 20 (7), 1692–1704. doi:10.1016/j.celrep.2017.07.055

Xing, Y., Luo, P., Hu, R., Wang, D., Zhou, G., and Jiang, J. (2020). TRIB3 Promotes Lung Adenocarcinoma Progression via an Enhanced Warburg Effect. *Cancer Manag. Res.* 12, 13195–13206. doi:10.2147/CMAR.S287956

Xu, G., Wang, H., Li, X., Huang, R., and Luo, L. (2021). Recent Progress on Targeting Ferroptosis for Cancer Therapy. *Biochem. Pharmacol.* 190, 114584. doi:10.1016/j.bcp.2021.114584

Yu, X., Zhang, X., and Zhang, Y. (2020). Identification of a 5-Gene Metabolic Signature for Predicting Prognosis Based on an Integrated Analysis of Tumor Microenvironment in Lung Adenocarcinoma. *J. Oncol.* 2020, 1–12. doi:10.1155/2020/5310793

Zhu, B., Ren, C., Du, K., Zhu, H., Ai, Y., Kang, F., et al. (2019). Olean-28,13b-olide 2 Plays a Role in Cisplatin-Mediated Apoptosis and Reverses Cisplatin Resistance in Human Lung Cancer through Multiple Signaling Pathways. *Biochem. Pharmacol.* 170, 113642. doi:10.1016/j.bcp.2019.113642

Zhuo, S., Chen, Z., Yang, Y., Zhang, J., Tang, J., and Yang, K. (2020). Clinical and Biological Significances of a Ferroptosis-Related Gene Signature in Glioma. *Front. Oncol.* 10, 590861. doi:10.3389/fonc.2020.590861

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 Genetics

ORIGINAL RESEARCH
published: 23 September 2021
doi: 10.3389/fgene.2021.728913

A Unified Framework for Inattention Estimation From Resting State Phase Synchrony Using Machine Learning

Xun-Heng Wang and Lihua Li**

Institute of Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou, China

Inattention is one of the most significant clinical symptoms for evaluating attention deficit hyperactivity disorder (ADHD). Previous inattention estimations were performed using clinical scales. Recently, predictive models for inattention have been established for brain-behavior estimation using neuroimaging features. However, the performance of inattention estimation could be improved for conventional brain-behavior models with additional feature selection, machine learning algorithms, and validation procedures. This paper aimed to propose a unified framework for inattention estimation from resting state fMRI to improve the classical brain-behavior models. Phase synchrony was derived as raw features, which were selected with minimum-redundancy maximum-relevancy (mRMR) method. Six machine learning algorithms were applied as regression methods. 100 runs of 10-fold cross-validations were performed on the ADHD-200 datasets. The relevance vector machines (RVMs) based on the mRMR features for the brain-behavior models significantly improve the performance of inattention estimation. The mRMR-RVM models could achieve a total accuracy of 0.53. Furthermore, predictive patterns for inattention were discovered by the mRMR technique. We found that the bilateral subcortical-cerebellum networks exhibited the most predictive phase synchrony patterns for inattention. Together, an optimized strategy named mRMR-RVM for brain-behavior models was found for inattention estimation. The predictive patterns might help better understand the phase synchrony mechanisms for inattention.

Keywords: predictive models, inattention, feature selection, regression algorithms, phase synchrony

OPEN ACCESS

Edited by:
Zhaowen Qiu,
Northeast Forestry University, China

Reviewed by:
Guoxian Yu,
Shandong University, China
Qiang Li,
Fourth Military Medical University, China

****Correspondence:***
Xun-Heng Wang
xhwang@hdu.edu.cn
Lihua Li
lilh@hdu.edu.cn

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

Received: *22 June 2021*
Accepted: *23 August 2021*
Published: *23 September 2021*

Citation:
Wang X-H and Li L (2021) A Unified Framework for Inattention Estimation From Resting State Phase Synchrony Using Machine Learning. Front. Genet. 12:728913. doi: 10.3389/fgene.2021.728913

INTRODUCTION

Estimating personalized cognitive or behavioral scores from neuroimaging is an interesting yet challenging topic nowadays (Rosenberg et al., 2016; Shen et al., 2017; Yoo et al., 2017; Rosenberg et al., 2018; Sui et al., 2020). The individual brain-age, Intelligence Quotient (IQ), attention, as well as personality can be estimated either from structural or functional MRI using machine learning (Zhao et al., 2019; Cai et al., 2020; Lin et al., 2020; Munsell et al., 2020; Niu et al., 2020). Among those brain-behavior models, predicting individual attention from neuroimaging has drawn a significant amount of research interests (Rosenberg et al., 2016, 2018; Yoo et al., 2017). Attention is a key function in psychology. Attention is also a significant feature for diagnosis of ADHD (Xiao et al., 2016;

Zhao et al., 2018; Wang et al., 2018a,b). Inattention can lead to dysfunction of memory, learning, and other important cognitive tasks (Brown et al., 2009; Fassbender et al., 2011; Vaidya et al., 2020). Before the present time, the inattention scores were always estimated using clinical scales, which were subjective measures reported by participants (Zhang et al., 2005). Furthermore, the neural mechanisms of inattention are still unclear to date. Therefore, it is of great interest to build predictive models for inattention using resting state fMRI.

The predictive models for inattention estimations contain three parts. One important component of a predictive model is the input features. Currently, most of the raw features for inattention estimations were based on linear functional connectivity (Rosenberg et al., 2016; Yoo et al., 2017). The nonlinear complexity (i.e., phase synchrony) remained unknown (Wang et al., 2017). Another important component is the regression algorithms. The well-established connectome-based predictive modeling (CPM) for inattention estimation was based on multi-linear regression (Shen et al., 2017). The comparisons of performance of different regression algorithms remain largely unexplored (Yoo et al., 2017; Sui et al., 2020). The third component is the model validation procedure. So far, most of the predictive models were evaluated using leave-one-out cross validation. Although several studies validated their models using two independent datasets, the N-fold cross validations might also be beneficial for inattention estimation (Scheinost et al., 2019).

In addition, different preprocessing steps (i.e., global signal regression (GSR), data scrubbing) might have impacts on the brain connectivity (Li et al., 2019a). Although the benefits of GSR for resting fMRI are still under debate, previous studies found that GSR might enhance the brain-behavior relationships (Murphy et al., 2009; Wong et al., 2012; Li et al., 2019a). The data scrubbing or volume censoring methods also have impacts on functional connectivity features (Yan et al., 2013; Parkes et al., 2018; Li et al., 2019b; Lindquist et al., 2019). Therefore, different preprocessing steps should be considered in the brain-behavior regression tasks. So far, the effects of different preprocessing procedures on estimation of inattention using phase synchrony remain unclear.

In this paper, we aimed to apply a unified framework to estimate the personalized inattention from resting state phase synchrony. First, a cohort of participants with both inattention scores and resting state fMRI datasets were obtained from the ADHD-200 database. Then, the resting state fMRI datasets were preprocessed using different strategies that were with or without GSR or scrubbing. Third, the regional signals were obtained from the normalized images. Fourth, phase synchrony was derived as input for the regression tasks. Fifth, the inattention scores were estimated using different regression algorithms. Finally, the regression models were analyzed using 100 runs of 10-fold cross validations. The impacts of different preprocessing strategies on the regression tasks are compared in the results section. The predictive patterns are discussed in the discussion section.

MATERIALS AND METHODS

Participants and MRI Protocols

Participants in this study were obtained from the ADHD-200 database. To be consistent with previous studies, the samples from the Peking University were selected as subjects. There were 95 ADHD and 126 healthy controls. Each participant signed the consent form that was approved by the ethics committee of Peking University. The inattention scores were measured using the ADHD rating scales. For each participant, a high-resolution T-1 weighted anatomical MRI and a sequence of resting state fMRI datasets (TR = 2 s, 235 volumes) were acquired using a Siemens 3 T MRI scanner. The detailed information of MRI parameters could be found at the website of ADHD-200.[1]

Data Preprocessing

The anatomical MRI were skull-stripped, segmented, and nonlinearly deformed to standard space. The resting state fMRI was normalized using the following procedures: dropped the first five volumes, slice-timing, motion correction, skull-stripped, nuisance signal regression, temporal filtering (0.01–0.1 Hz), scrubbing, spatial normalization. Specially, an artifactual volume was marked with frame-wise displacement >0.5 mm or DVARS value =1. The forward volume and backward volume were also marked as artifactual scan points. The detailed information of data preprocessing could be found in previous works (Wang et al., 2017, 2018b). After preprocessing, the regional time-courses were extracted using a previously well-established brain atlas that consisted of 268 functional nodes (Shen et al., 2013).

Phase Synchrony

The phase synchrony is a bivariate complexity measure with nonlinear properties. The phase synchrony has been widely applied in neuroscience as an alternative feature for conventional functional connectivity. One advantage of phase synchrony was the nonnegative property. Another advantage was the nonlinear property. The phase synchrony could be obtained using the following steps: (1) get the instantaneous phases of each time-signal using Hilbert transform; (2) unwarp the instantaneous phases; (3) get the instantaneous phase differences between each pair of time-signals; (4) discard the artifactual instantaneous phase differences if scrubbing was applied on preprocessing steps; and (5) compute the mean phase coherence as phase synchrony index (Sun and Small, 2009; Sun et al., 2012).

Regression Models

The minimum-redundancy maximum-relevancy (mRMR) features (Ding and Peng, 2005) were selected using the praznik package.[2] A number of features were detected based on significant correlations with inattention ($p < 0.05$). First,

[1]http://fcon_1000.projects.nitrc.org/indi/adhd200
[2]https://cran.r-project.org/web/packages/praznik

the number of significant inattention-correlated features ($p < 0.05$) was obtained in each cross-validation. Second, the numbers of features were obtained after 100 runs of 10-fold cross-validations. Finally, the mean value of numbers of features was calculated for the mRMR procedure. In addition, the classical correlation coefficients method was also applied to select features ($p < 0.05$). The predictive power of inattention-correlated features with $p < 0.05$ and $r > 0$ was analyzed additionally. The features selected by the covariance between inattention and phase synchrony were analyzed with the number of features the same as that of the mRMR. The regression models were solved using six algorithms: the support vector regression (SVR), the partial least squares (PLS), the relevance vector machine (RVM), the ridge regression (RR), the elastic net (ENET), and the least absolute shrinkage and selection operator (LASSO). In this study, the SVR algorithm was carried out using the svm() function in e1071 package.[3] The PLS algorithm was carried out using the pls () function in the texir package.[4] The RVM algorithm was carried out using the rvm() function in kernlab package,[5] which automatically solved the sigma parameter. The RR, ENET and LASSO algorithms were carried out using the glmnet() function in the glmnet package[6] with alpha = 0, 0.5, and 1, respectively. The six algorithms used their default parameters in the R packages for comparisons of cross-validations. The CPM algorithm was carried out additionally using the MATLAB toolbox.[7] Furthermore, the parameters were fine-tuned for the regression algorithms using the caret package.[8] The RR, lasso, and ENET were analyzed using the glmnet model, which fine-tuned the alpha and lambda parameters. The PLS algorithm was analyzed using the pls model, which fine-tuned the number of component parameter. The support vector machine algorithm was analyzed using the svmLinear model, which fine-tuned the cost parameter.

Evaluations

In this paper, 100 runs of 10-fold cross-validations were applied on the regression tasks. For each run, the original samples were divided into 10 folds. For each fold, nine folds of training samples and a fold of testing samples were applied to build predictive models. The outputs of 10 folds were joined together to match with the original inattention scores. The performance of the regression models was evaluated by correlation coefficients, which were computed using the 1,000 times of permutations test. The values of p were analyzed using the RVAideMemoire package.[9] The pipeline for the feature selection, regression, and validation procedures could be found in **Figure 1**.

[3]https://cran.r-project.org/web/packages/e1071/index.html
[4]https://CRAN.R-project.org/package=textir
[5]https://www.rdocumentation.org/packages/kernlab/versions/0.9-29
[6]https://cran.r-project.org/web/packages/glmnet/index.html
[7]https://github.com/YaleMRRC/CPM
[8]https://topepo.github.io/caret/index.html
[9]https://cran.r-project.org/package=RVAideMemoire

RESULTS

Performance of Predictive Models

Different feature selection methods and regression algorithms have impacts on the performance of the predictive models. **Figure 2** shows the performance of the predictive models based on classical feature selection ($p < 0.05$). **Figure 3** shows the performance of the predictive models based on classical feature selection ($p < 0.05$, $r > 0$). **Figure 4** shows the performance of the predictive models based on covariance feature selection. **Figure 5** shows the performance of the predictive models based on fine-tuning of the regression algorithms. **Figure 6** shows the performance of the predictive models based on mRMR feature selection. **Table 1** shows the performances of predictive models based on classical feature selection with GSR and scrubbing. **Table 2** shows the performances of predictive models based on mRMR with GSR and scrubbing. The CPM-based models with GSR and scrubbing can achieve a mean accuracy of 0.31. The best predictive models can achieve a total accuracy of 0.56 based on mRMR and RVM. The PLS also exhibits predictive powers. The PLS based on mRMR can achieve a total accuracy of 0.34.

The predictive models with GSR outperform that without GSR. **Figures 2A,B, Figures 3A,B, Figures 4A,B, Figures 5A,B** as well as **Figures 6A,B** show the performance of the predictive models with GSR. **Figures 2C,D, Figures 3C,D, Figures 4C,D, Figures 5C,D**, as well as **Figures 6C,D** show the performance of the predictive models without GSR. The performance of the predictive models with GSR is significantly higher than that without GSR.

The predictive models without scrubbing outperform those with scrubbing. **Figures 2A,C, Figures 3A,C, Figures 4A,C, Figures 5A,C** as well as **Figures 6A,C** show the performance of the predictive models with scrubbing. **Figures 2B,D, Figures 3B,D, Figures 4B,D, Figures 5B,D**, as well as **Figures 6B,D** show the performance of the predictive models without scrubbing. The performance of predictive models with scrubbing is a little lower than that without scrubbing.

In addition, the predictive models without fine-tuning (**Figure 6**) outperform that with fine-tuning (**Figure 5**). The positive weighted features significantly improve the performance of the regression models with GSR, but remarkably reduce the performance of the regression models without GSR, as indicated in **Figure 3**.

Predictive Patterns Related to Inattention

Figure 7 shows the predictive patterns related to inattention based on the mRMR feature selection with GSR and scrubbing. The 268 nodes are divided into 8 functional systems according to a previous study (Finn et al., 2015). The 8 functional systems are named as the medial frontal (MF) network, frontoparietal (FP) network, default mode (DM) network, subcortical-cerebellum (SC) network, motor cortex (MC) network, visual I (V1) network, visual II (V2) network, and visual association (VA) network. With 100 runs of 10-fold feature selection procedures, 1,000 arrays of most predictive features are selected

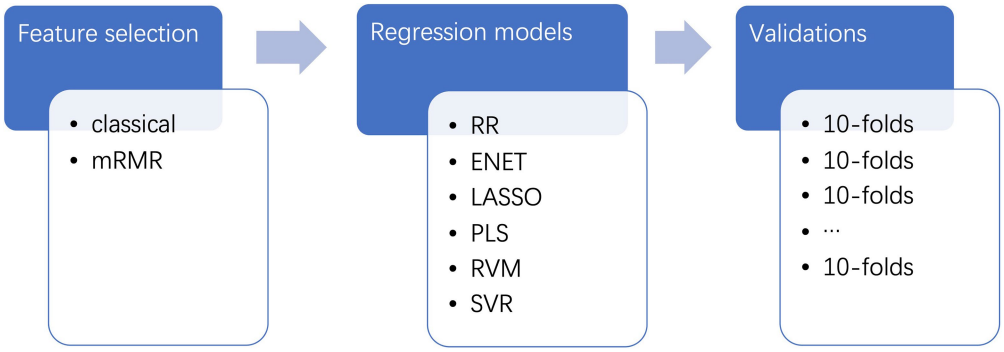

FIGURE 1 | Pipelines for the predictive models. The raw features of phase synchrony are firstly selected by two feature selection methods. Then, the selected features are trained and tested using several regression algorithms. Finally, the predictive models are validated using 100 runs of 10-fold cross-validations.

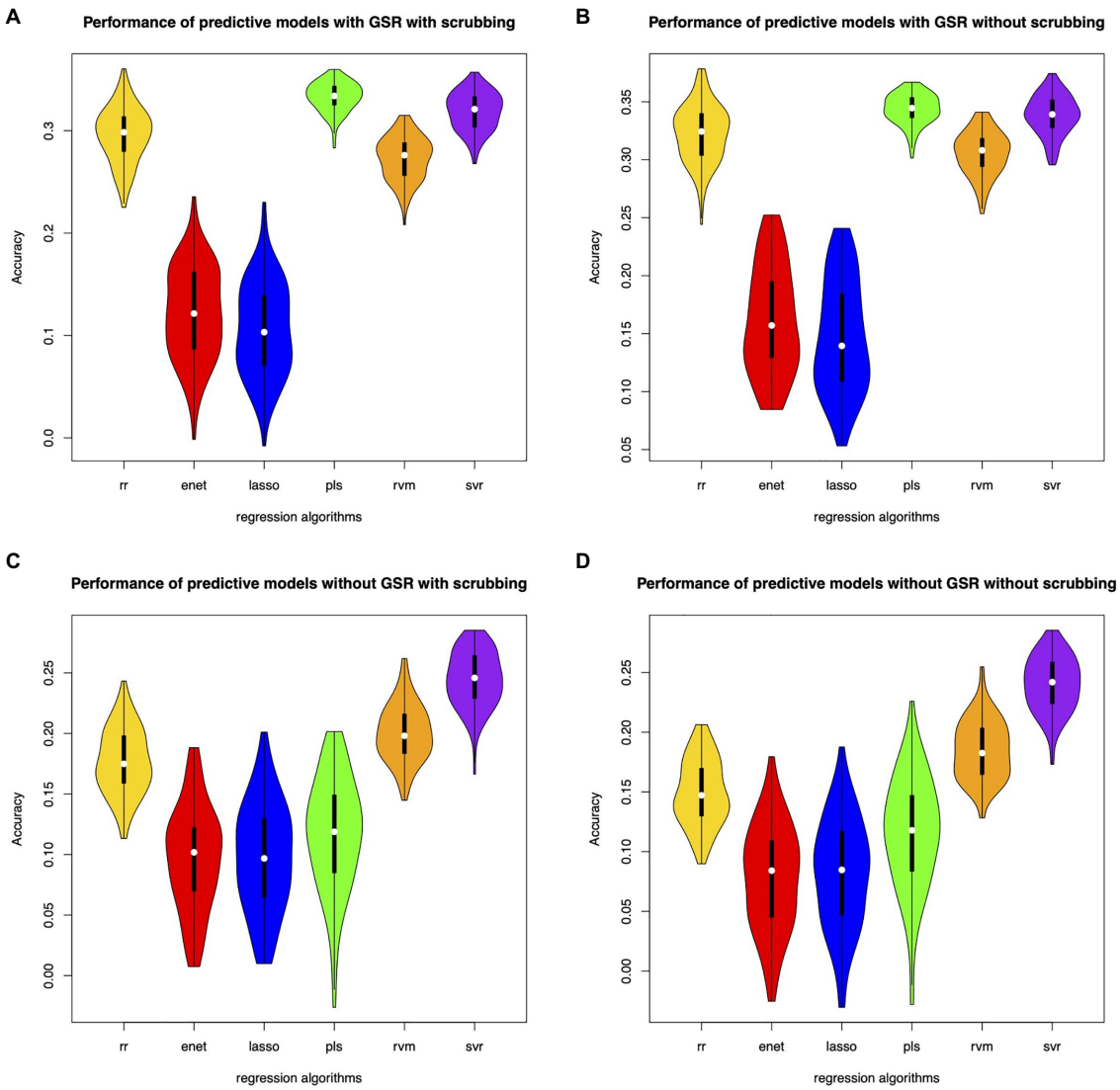

FIGURE 2 | Performance of the predictive models with classical feature selection ($p < 0.05$). **(A)** denotes performance of the predictive models with GSR and scrubbing. **(B)** denotes performance of the predictive models with GSR and without scrubbing. **(C)** denotes performance of the predictive models without GSR and with scrubbing. **(D)** denotes performance of the predictive models without GSR and scrubbing.

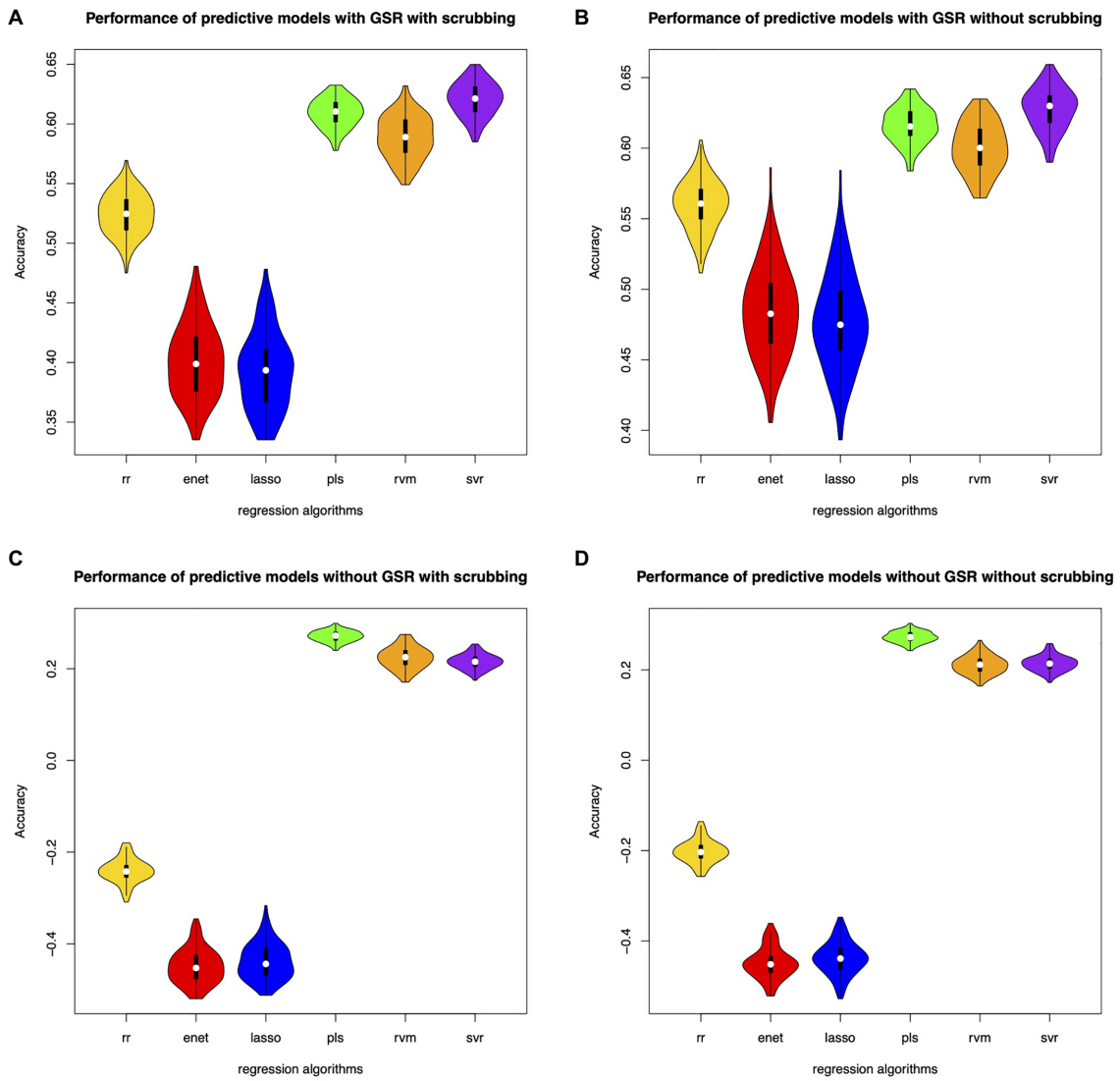

FIGURE 3 | Performance of the predictive models with classical feature selection ($p < 0.05$ and $r > 0$). **(A)** denotes performance of the predictive models with GSR and scrubbing. **(B)** denotes performance of the predictive models with GSR and without scrubbing. **(C)** denotes performance of the predictive models without GSR and with scrubbing. **(D)** denotes performance of the predictive models without GSR and scrubbing.

as important attributes. Only features that appeared more than 900 times are displayed in **Figure 7**. The most predictive brain regions are located in the bilateral SC network. The second predictive brain regions are located in the bilateral MC network. The right MF network is more predictive than the left MF network. The DM network and visual networks are less predictive than other networks. Both intra- and inter-hemisphere connections are found for inattention estimation.

DISCUSSION

In this paper, we applied several feature selection methods and six regression algorithms to build predictive models for inattention estimation using phase synchrony. The effects of different preprocessing steps (i.e., GSR, scrubbing) were considered in computing phase synchrony. We found that the RVMs based on mRMR features significantly improve the performance of inattention estimation from resting state phase synchrony. In addition, we also found that GSR significantly enhanced the relationships between phase synchrony and inattention. Furthermore, the predictive patterns were discovered using mRMR methods. In summary, we proposed a novel framework for inattention estimation from phase synchrony, which could be supplementary biomarkers for predictive models.

The performance of regression models was related to several procedures in inattention estimation. First, the feature selection methods might affect the accuracy of prediction. The features selected by conventional correlation coefficients were univariate attributes, which did not consider the relationships among

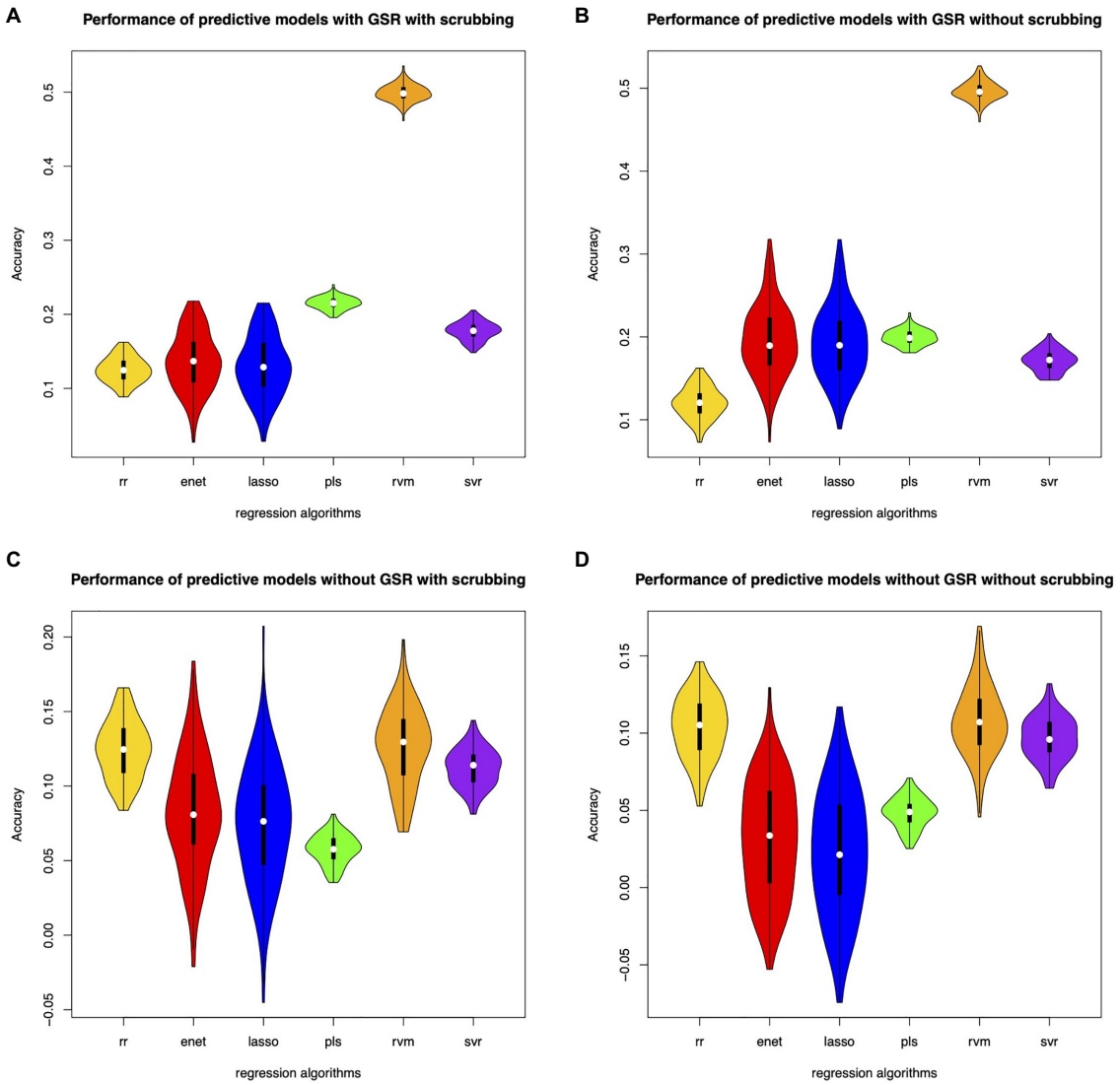

FIGURE 4 | Performance of the predictive models with covariance-based feature selection. **(A)** denotes performance of the predictive models with GSR and scrubbing. **(B)** denotes performance of the predictive models with GSR and without scrubbing. **(C)** denotes performance of the predictive models without GSR and with scrubbing. **(D)** denotes performance of the predictive models without GSR and scrubbing.

the raw features. The significant inattention-correlated features with positive weights ($p < 0.05$ and $r > 0$) can improve the performance of regression models but were dependent on GSR procedures. The performance of covariance-based feature selection was lower than that of conventional correlation-based models, since the covariance-based features might not be the significantly inattention-correlated. To overcome this limitation, mRMR was proposed to select multivariate features (Ding and Peng, 2005). The selected features significantly improved the performance of inattention estimation. Second, the regression algorithms also affect the performance of predictive models. We found that in addition to RVM, the PLS was an alternative algorithm for inattention estimation, which was consistent with previous findings (Yoo et al., 2017). Specially, we found RVMs based on mRMR features outperformed the other methods. The results indicated that the fine-tuning procedure does not improve the performance of the regression models. The poor performance of the fine-tuning might be caused by the 10-fold cross-validation procedures, since the training samples were different among the cross-validations. Of note, the RVM exhibited the best performance using automatic fine-tuning, implying that the sigma parameter for RVM was robust for different datasets. Third, the different preprocessing steps significantly affect the prediction. GSR significantly enhanced the relationships between phase synchrony and inattention. Scrubbing had little effect on the final results. The results suggested that GSR should be considered in brain-behavioral prediction task (Li et al., 2019a). Fourth, the cross-validations might have effect on the performance of prediction tasks. Here, 100 runs of 10-fold cross-validations were performed

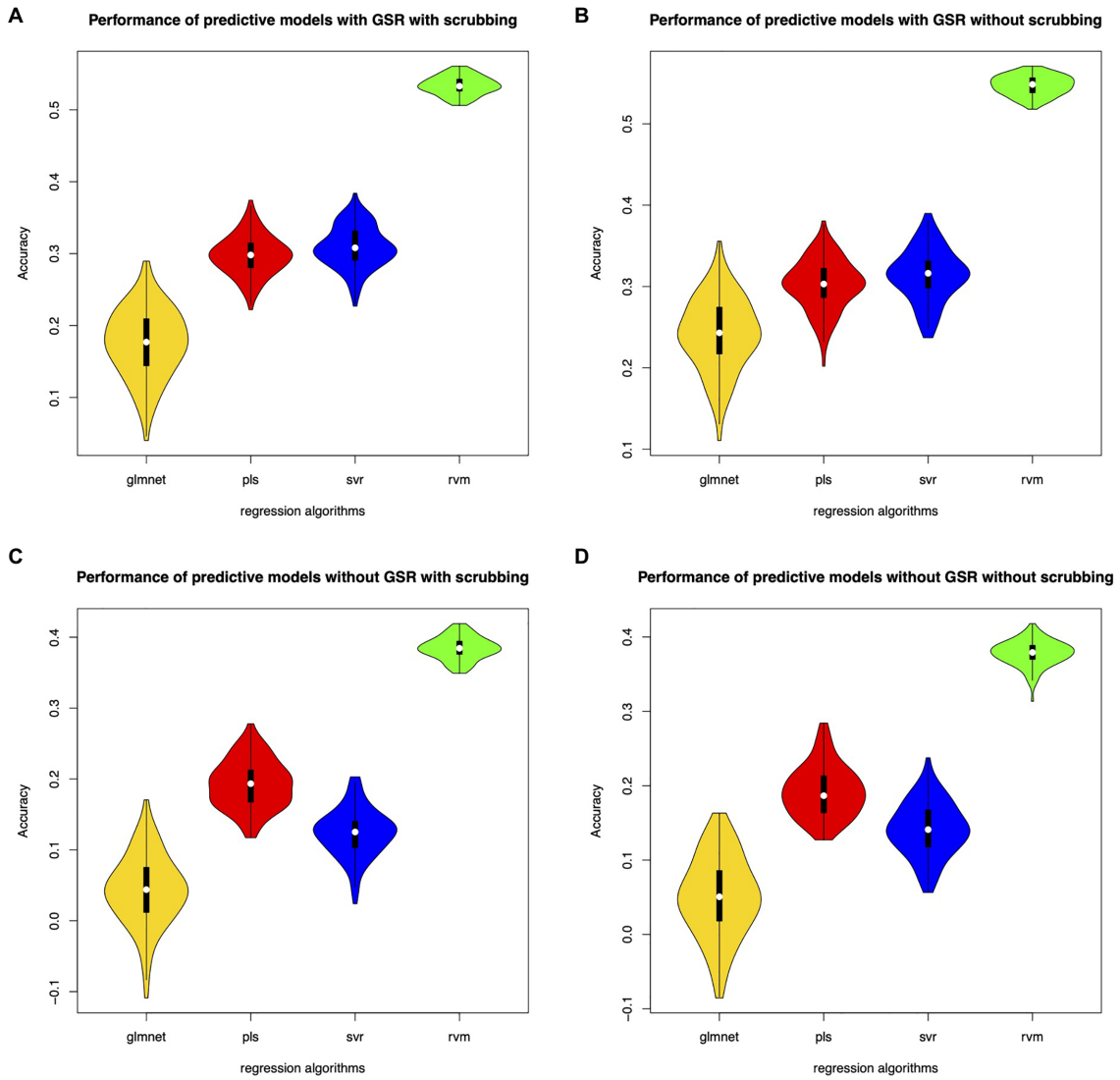

FIGURE 5 | Performance of the predictive models with fine-tuned parameters. **(A)** denotes performance of the predictive models with GSR and scrubbing. **(B)** denotes performance of the predictive models with GSR and without scrubbing. **(C)** denotes performance of the predictive models without GSR and with scrubbing. **(D)** denotes performance of the predictive models without GSR and scrubbing.

to evaluate the predictive models. The correlation coefficients were reliable and the MAE values were also stable, suggesting the robustness of the predictive models. In this paper, we applied different algorithms to build predictive models for inattention. After comparing with different methods, we found that the mRMR-RVM strategy might be beneficial for inattention estimation from neuroimaging features.

Predictive patterns related to inattention were discovered using mRMR feature selection. The visual networks, default mode networks, medial frontal network, frontoparietal network, subcortical-cerebellum network, as well as motor cortex exhibited altered phase synchrony in patients with ADHD. The predictive connections in visual network and motor cortex suggested that the sensorimotor functions might be distinctive in ADHD (Zang et al., 2007). The altered connectivity patterns in medial frontal network and frontoparietal network might reflect the inattention mechanisms in ADHD (Tao et al., 2017). Previous studies found altered functional connectivity in default mode networks in ADHD, suggesting the abnormal resting state baseline activity in patients (Hoekzema et al., 2014). Decreased subcortical volumes were also found in ADHD compared to healthy controls (Lu et al., 2019). In this study, we found that the bilateral subcortical-cerebellum networks exhibited the most predictive phase synchrony patterns. We also found that the motor cortex had the second predictive brain regions. Both inter- and intra-hemisphere synchrony patterns were found to be related to inattention. In addition, the altered phase synchrony exhibited asymmetry patterns. Those findings implied that the whole brain phase synchrony was predictive to inattention estimation. In summary, this study provided a new way to

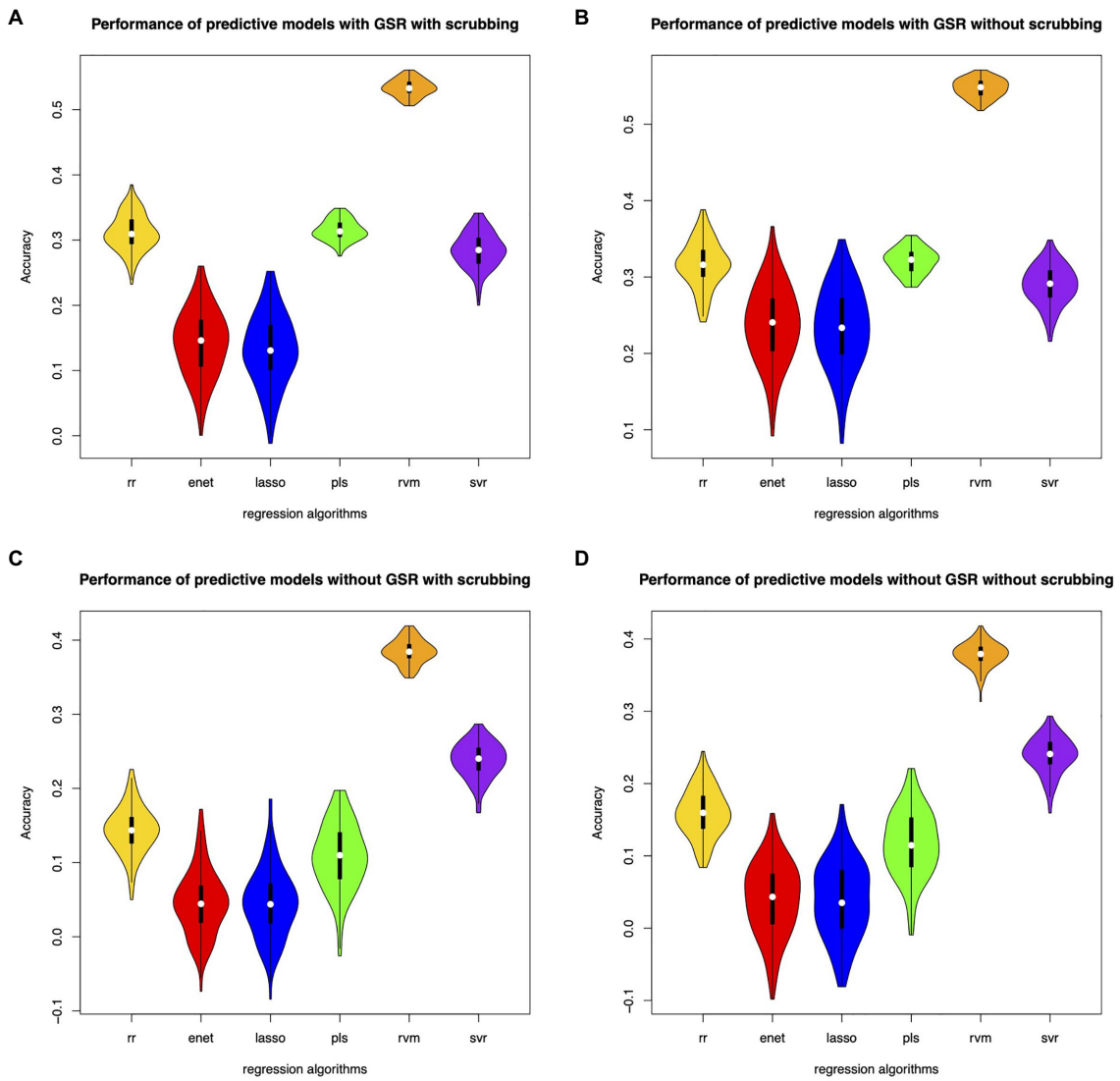

FIGURE 6 | Performance of the predictive models with the mRMR feature selection. **(A)** denotes performance of the predictive models with GSR and scrubbing. **(B)** denotes performance of the predictive models with GSR and without scrubbing. **(C)** denotes performance of the predictive models without GSR and with scrubbing. **(D)** denotes performance of the predictive models without GSR and scrubbing.

TABLE 1 | Performance of predictive models based on classical feature selection with GSR and scrubbing.

Algorithms	*r*	MAE	RMSE
RR	0.3 ± 0.03	5.95 ± 0.06	6.9 ± 0.06
ENET	0.12 ± 0.05	6.5 ± 0.14	7.6 ± 0.15
LASSO	0.11 ± 0.05	6.57 ± 0.14	7.74 ± 0.16
PLS	0.33 ± 0.01	5.83 ± 0.05	6.91 ± 0.05
RVM	0.27 ± 0.02	6.01 ± 0.05	6.97 ± 0.05
SVR	0.32 ± 0.02	5.94 ± 0.04	6.85 ± 0.04

TABLE 2 | Performance of predictive models based on mRMR with GSR and scrubbing.

Algorithms	*r*	MAE	RMSE
RR	0.31 ± 0.03	5.92 ± 0.07	6.87 ± 0.07
ENET	0.14 ± 0.05	6.52 ± 0.19	7.77 ± 0.2
LASSO	0.13 ± 0.05	6.6 ± 0.21	7.88 ± 0.22
PLS	0.32 ± 0.02	5.9 ± 0.06	6.98 ± 0.06
RVM	0.53 ± 0.01	5.42 ± 0.03	6.28 ± 0.04
SVR	0.28 ± 0.03	6.12 ± 0.03	6.98 ± 0.03

decode the inattention using phase synchrony and mRMR feature selection, which might be beneficial for individual prediction of inattention.

This study has several limitations which should be solved in future studies. First, the dynamic properties of functional connectivity remain unexplored for inattention. Novel feature

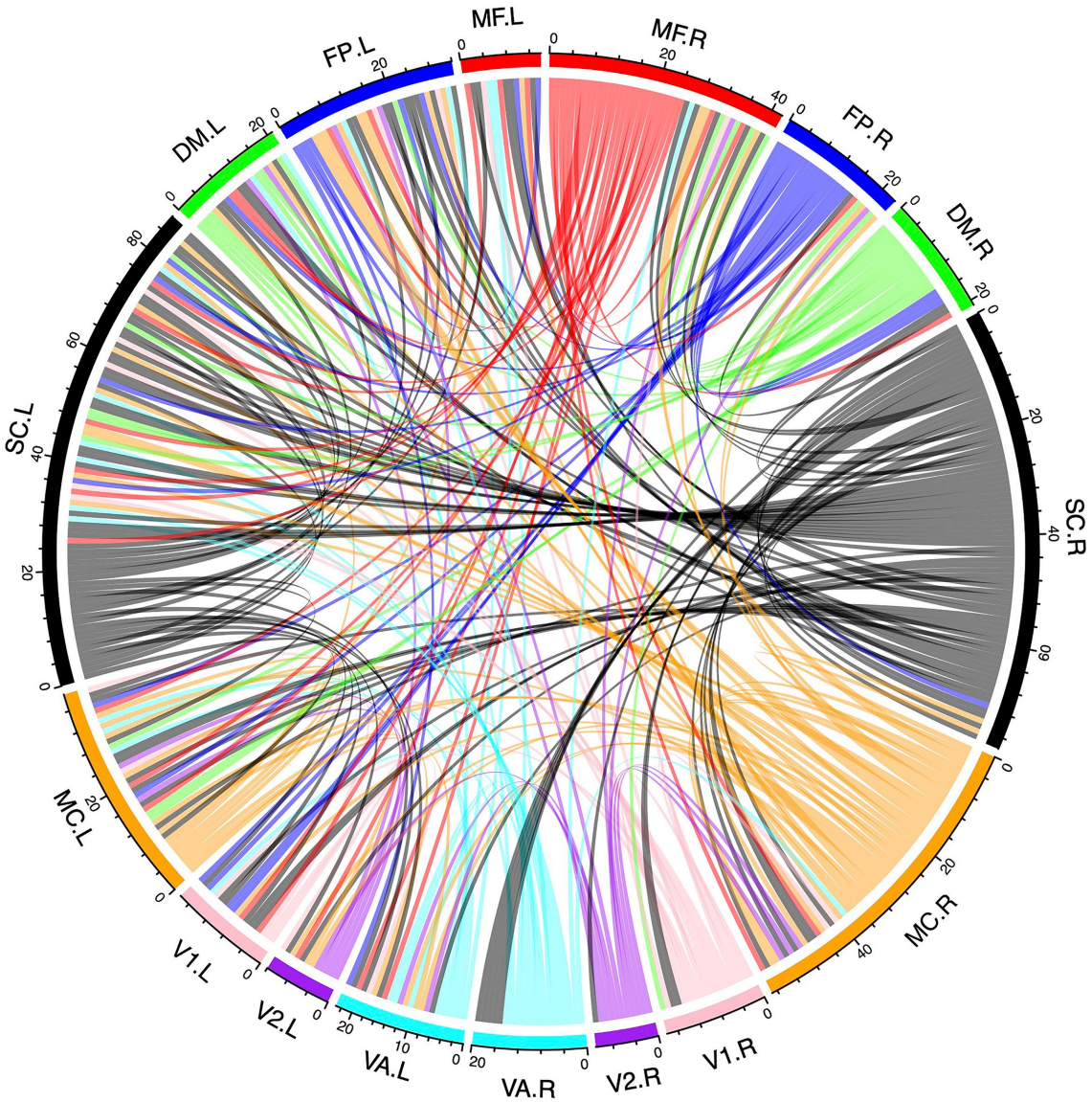

FIGURE 7 | Predictive patterns of phase synchrony for inattention. MF stands for the medial frontal network. FP represents the frontoparietal network. DM means the default mode network. SC denotes the subcortical-cerebellum network. MC represents the motor cortex network. V1 denotes the visual I network. V2 denotes the visual II network. VA stands for the visual association network.

extraction methods for dynamic phase synchrony should be investigated for inattention estimation. Second, the performance of the inattention estimations should be improved with novel feature selection methods and regression algorithms. Third, the mRMR features could not reflect the positive or negative correlations between phase synchrony and inattention. Fourth, the regression models should be tested using an independent dataset, although the regression models were well-validated using 100 runs of 10-fold cross-validations. Fifth, there were different MRI protocols for the samples, which should be scanned with the same MRI scanner and parameters. In summary, the feature extraction models, feature selection methods, regression algorithms, and testing procedures should be improved to enhance the performance and the generalization ability of the regression models for individual inattention estimation.

CONCLUSION

This paper applied different algorithms to build the predictive models for inattention from resting state fMRI. We also analyzed the impacts of different preprocessing steps on the predictive models. The RVMs based on mRMR features significantly improve the performance of inattention estimation from resting state phase synchrony. We also found that PLS might be an

alternative method for brain-behavioral prediction tasks. In addition, the GSR strengthens the relationships between neuroimaging features and behavioral scores. In summary, we proposed a unified framework for brain-behavioral models based on phase synchrony. We also found an optimized strategy named mRMR-RVM for inattention estimation.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be available at: http://fcon_1000.projects.nitrc.org/indi/adhd200.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Peking University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

X-HW and LL contributed to conception and design of the study and wrote the first draft of the manuscript. X-HW performed the statistical analysis. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research was supported in part by the National Key R&D Program of China under grant no. 2018YFA0701702 and the National Natural Science Foundation of China (62071158).

REFERENCES

Brown, T. E., Reichel, P. C., and Quinlan, D. M. (2009). Executive function impairments in high IQ adults with ADHD. *J. Atten. Disord.* 13, 161–167. doi: 10.1177/1087054708326113

Cai, H., Zhu, J., and Yu, Y. (2020). Robust prediction of individual personality from brain functional connectome. *Soc. Cogn. Affect. Neurosci.* 15, 359–369. doi: 10.1093/scan/nsaa044

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinforma. Comput. Biol.* 3, 185–205. doi: 10.1142/S0219720005001004

Fassbender, C., Schweitzer, J. B., Cortes, C. R., Tagamets, M. A., Windsor, T. A., Reeves, G. M., et al. (2011). Working memory in attention deficit/hyperactivity disorder is characterized by a lack of specialization of brain function. *PLoS One* 6:e27240. doi: 10.1371/journal.pone.0027240

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., et al. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* 18, 1664–1671. doi: 10.1038/nn.4135

Hoekzema, E., Carmona, S., Ramos-Quiroga, J. A., Richarte Fernández, V., Bosch, R., Soliva, J. C., et al. (2014). An independent components and functional connectivity analysis of resting state fMRI data points to neural network dysregulation in adult ADHD. *Hum. Brain Mapp.* 35, 1261–1272. doi: 10.1002/hbm.22250

Li, J., Kong, R., Liegeois, R., Orban, C., Tan, Y., Sun, N., et al. (2019a). Global signal regression strengthens association between resting-state functional connectivity and behavior. *NeuroImage* 196, 126–141. doi: 10.1016/j.neuroimage.2019.04.016

Li, W., Qiao, L., Zhang, L., Wang, Z., and Shen, D. (2019b). Functional brain network estimation With time series self-scrubbing. *IEEE J. Biomed. Health Inform.* 23, 2494–2504. doi: 10.1109/JBHI.2019.2893880

Lin, Y. C., Baete, S. H., Wang, X., and Boada, F. E. (2020). Mapping brain-behavior networks using functional and structural connectome fingerprinting in the HCP dataset. *Brain Behav.* 10:e01647. doi: 10.1002/brb3.1647

Lindquist, M. A., Geuter, S., Wager, T. D., and Caffo, B. S. (2019). Modular preprocessing pipelines can reintroduce artifacts into fMRI data. *Hum. Brain Mapp.* 40, 2358–2376. doi: 10.1002/hbm.24528

Lu, L., Zhang, L., Tang, S., Bu, X., Chen, Y., Hu, X., et al. (2019). Characterization of cortical and subcortical abnormalities in drug-naive boys with attention-deficit/hyperactivity disorder. *J. Affect. Disord.* 250, 397–403. doi: 10.1016/j.jad.2019.03.048

Munsell, B. C., Gleichgerrcht, E., Hofesmann, E., Delgaizo, J., Mcdonald, C. R., Marebwa, B., et al. (2020). Personalized connectome fingerprints: their importance in cognition from childhood to adult years. *NeuroImage* 221:117122. doi: 10.1016/j.neuroimage.2020.117122

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., and Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *NeuroImage* 44, 893–905. doi: 10.1016/j.neuroimage.2008.09.036

Niu, X., Zhang, F., Kounios, J., and Liang, H. (2020). Improved prediction of brain age using multimodal neuroimaging data. *Hum. Brain Mapp.* 41, 1626–1643. doi: 10.1002/hbm.24899

Parkes, L., Fulcher, B., Yucel, M., and Fornito, A. (2018). An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage* 171, 415–436. doi: 10.1016/j.neuroimage.2017.12.073

Rosenberg, M. D., Finn, E. S., Scheinost, D., Papademetris, X., Shen, X., Constable, R. T., et al. (2016). A neuromarker of sustained attention from whole-brain functional connectivity. *Nat. Neurosci.* 19, 165–171. doi: 10.1038/nn.4179

Rosenberg, M. D., Hsu, W. T., Scheinost, D., Todd Constable, R., and Chun, M. M. (2018). Connectome-based models predict separable components of attention in novel individuals. *J. Cogn. Neurosci.* 30, 160–173. doi: 10.1162/jocn_a_01197

Scheinost, D., Noble, S., Horien, C., Greene, A. S., Lake, E. M., Salehi, M., et al. (2019). Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage* 193, 35–45. doi: 10.1016/j.neuroimage.2019.02.057

Shen, X., Finn, E. S., Scheinost, D., Rosenberg, M. D., Chun, M. M., Papademetris, X., et al. (2017). Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* 12, 506–518. doi: 10.1038/nprot.2016.178

Shen, X., Tokoglu, F., Papademetris, X., and Constable, R. T. (2013). Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *NeuroImage* 82, 403–415. doi: 10.1016/j.neuroimage.2013.05.081

Sui, J., Jiang, R., Bustillo, J., and Calhoun, V. (2020). Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biol. Psychiatry* 88, 818–828. doi: 10.1016/j.biopsych.2020.02.016

Sun, J., Hong, X., and Tong, S. (2012). Phase synchronization analysis of EEG signals: an evaluation based on surrogate tests. *I.E.E.E. Trans. Biomed. Eng.* 59, 2254–2263. doi: 10.1109/TBME.2012.2199490

Sun, J., and Small, M. (2009). Unified framework for detecting phase synchronization in coupled time series. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* 80:046219. doi: 10.1103/PhysRevE.80.046219

Tao, J., Jiang, X., Wang, X., Liu, H., Qian, A., Yang, C., et al. (2017). Disrupted control-related functional brain networks in drug-naive children with attention-deficit/hyperactivity disorder. *Front. Psych.* 8:246. doi: 10.3389/fpsyt.2017.00246

Vaidya, C. J., You, X., Mostofsky, S., Pereira, F., Berl, M. M., and Kenworthy, L. (2020). Data-driven identification of subtypes of executive function across typical development, attention deficit hyperactivity disorder, and autism spectrum disorders. *J. Child Psychol. Psychiatry* 61, 51–61. doi: 10.1111/jcpp.13114

Wang, X.-H., Jiao, Y., and Li, L. (2017). Predicting clinical symptoms of attention deficit hyperactivity disorder based on temporal patterns between and within intrinsic connectivity networks. *Neuroscience* 362, 60–69. doi: 10.1016/j.neuroscience.2017.08.038

Wang, X. H., Jiao, Y., and Li, L. (2018a). Diagnostic model for attention-deficit hyperactivity disorder based on interregional morphological connectivity. *Neurosci. Lett.* 685, 30–34. doi: 10.1016/j.neulet.2018.07.029

Wang, X. H., Jiao, Y., and Li, L. (2018b). Identifying individuals with attention deficit hyperactivity disorder based on temporal variability of dynamic functional connectivity. *Sci. Rep.* 8:11789. doi: 10.1038/s41598-018-30308-w

Wong, C. W., Olafsson, V., Tal, O., and Liu, T. T. (2012). Anti-correlated networks, global signal regression, and the effects of caffeine in resting-state functional MRI. *NeuroImage* 63, 356–364. doi: 10.1016/j.neuroimage.2012.06.035

Xiao, C., Bledsoe, J., Wang, S., Chaovalitwongse, W. A., Mehta, S., Semrud-Clikeman, M., et al. (2016). An integrated feature ranking and selection framework for ADHD characterization. *Brain Inform.* 3, 145–155. doi: 10.1007/s40708-016-0047-1

Yan, C.-G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di Martino, A., et al. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *NeuroImage* 76, 183–201. doi: 10.1016/j.neuroimage.2013.03.004

Yoo, K., Rosenberg, M. D., Hsu, W. T., Zhang, S., Li, C. R., Scheinost, D., et al. (2017). Connectome-based predictive modeling of attention: comparing different functional connectivity features and prediction methods across datasets. *NeuroImage* 167, 11–22. doi: 10.1016/j.neuroimage.2017.11.010

Zang, Y.-F., He, Y., Zhu, C.-Z., Cao, Q.-J., Sui, M.-Q., Liang, M., et al. (2007). Altered baseline brain activity in children with ADHD revealed by resting-state functional MRI. *Brain and Development* 29, 83–91. doi: 10.1016/j.braindev.2006.07.002

Zhang, S., Faries, D. E., Vowles, M., and Michelson, D. (2005). ADHD rating scale IV: psychometric properties from a multinational study as a clinician-administered instrument. *Int. J. Methods Psychiatr. Res.* 14, 186–201. doi: 10.1002/mpr.7

Zhao, X., Rangaprakash, D., Yuan, B., Denney, T. S. Jr., Katz, J. S., Dretsch, M. N., et al. (2018). Investigating the Correspondence of Clinical Diagnostic Grouping With Underlying Neurobiological and Phenotypic Clusters Using Unsupervised Machine Learning. *Front. Appl. Math. Stat.* 4:25. doi: 10.3389/fams.2018.00025

Zhao, Y., Klein, A., Castellanos, F. X., and Milham, M. P. (2019). Brain age prediction: cortical and subcortical shape covariation in the developing human brain. *NeuroImage* 202:116149. doi: 10.1016/j.neuroimage.2019.116149

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 handling editor declared a past co-authorship with one of the authors LL.

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 September 2021
doi: 10.3389/fgene.2021.753948

Diagnosis of Ovarian Neoplasms Using Nomogram in Combination With Ultrasound Image-Based Radiomics Signature and Clinical Factors

Lisha Qi[1,2,3,4†], Dandan Chen[1,2,3,4†], Chunxiang Li[2,3,4,5], Jinghan Li[6], Jingyi Wang[1,2,3,4], Chao Zhang[2,3,4,7], Xiaofeng Li[2,3,4,8], Ge Qiao[1,2,3,4], Haixiao Wu[2,3,4,7], Xiaofang Zhang[9] and Wenjuan Ma[2,3,4,10]*

[1]Department of Pathology, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China, [2]National Clinical Research Center for Cancer, Tianjin, China, [3]Key Laboratory of Cancer Prevention and Therapy, Tianjin, China, [4]Tianjin's Clinical Research Center for Cancer, Tianjin, China, [5]Department of Ultrasonographic Diagnosis and Therapy, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China, [6]Department of Ultrasonographic Diagnosis and Therapy, Tianjin Ninghe Hospital, Tianjin, China, [7]Department of Bone and Soft Tissue Tumors, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China, [8]Department of Molecular Imaging and Nuclear Medicine, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China, [9]Department of Clinical Laboratory, Tianjin Medical University General Hospital, Tianjin, China, [10]Department of Breast Imaging, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China

OPEN ACCESS

Edited by:
Ming Fan,
Hangzhou Dianzi University, China

Reviewed by:
Jingjing Chen,
The Affiliated Hospital of Qingdao University, China
T Niu,
Georgia Institute of Technology, United States

***Correspondence:**
Wenjuan Ma
mawenjuan2008@163.com

[†]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: 05 August 2021
Accepted: 13 September 2021
Published: 28 September 2021

Citation:
Qi L, Chen D, Li C, Li J, Wang J, Zhang C, Li X, Qiao G, Wu H, Zhang X and Ma W (2021) Diagnosis of Ovarian Neoplasms Using Nomogram in Combination With Ultrasound Image-Based Radiomics Signature and Clinical Factors. Front. Genet. 12:753948. doi: 10.3389/fgene.2021.753948

Objectives: To establish and validate a nomogram integrating radiomics signatures from ultrasound and clinical factors to discriminate between benign, borderline, and malignant serous ovarian tumors.

Materials and methods: In this study, a total of 279 pathology-confirmed serous ovarian tumors collected from 265 patients between March 2013 and December 2016 were used. The training cohort was generated by randomly selecting 70% of each of the three types (benign, borderline, and malignant) of tumors, while the remaining 30% was included in the validation cohort. From the transabdominal ultrasound scanning of ovarian tumors, the radiomics features were extracted, and a score was calculated. The ability of radiomics to differentiate between the grades of ovarian tumors was tested by comparing benign vs borderline and malignant (task 1) and borderline vs malignant (task 2). These results were compared with the diagnostic performance and subjective assessment by junior and senior sonographers. Finally, a clinical-feature alone model and a combined clinical-radiomics (CCR) model were built using predictive nomograms for the two tasks. Receiver operating characteristic (ROC) analysis, calibration curve, and decision curve analysis (DCA) were performed to evaluate the model performance.

Results: The US-based radiomics models performed satisfactorily in both the tasks, showing especially higher accuracy in the second task by successfully discriminating borderline and malignant ovarian serous tumors compared to the evaluations by senior sonographers (AUC = 0.789 for seniors and 0.877 for radiomics models in task one; AUC = 0.612 for senior and 0.839 for radiomics model in task 2). We showed that the CCR model, comprising CA125 level, lesion location, ascites, and radiomics signatures, performed the best (AUC = 0.937, 95%CI 0.905–0.969 in task 1, AUC = 0.924, 95%

CI 0.876–0.971 in task 2) in the training as well as in the validation cohorts (AUC = 0.914, 95%CI 0.851–0.976 in task 1, AUC = 0.890, 95%CI 0.794–0.987 in task 2). The calibration curve and DCA analysis of the CCR model more accurately predicted the classification of the tumors than the clinical features alone.

Conclusion: This study integrates novel radiomics signatures from ultrasound and clinical factors to create a nomogram to provide preoperative diagnostic information for differentiating between benign, borderline, and malignant ovarian serous tumors, thereby reducing unnecessary and risky biopsies and surgeries.

Keywords: radiomics, serous ovarian tumor, ultrasound, classification, nomogram, image analysis

INTRODUCTION

Histologically, serous tumors are the most prevalent ovarian tumors, representing 70% of the cases. (Javadi et al., 2016; Brett et al., 2017; Lheureux et al., 2019; Lisio et al., 2019). Such tumors can be classified into benign, borderline, and malignant lesions that exhibit distinct clinicopathological characteristics owing to which they exhibit differences in terms of therapeutic schemes, and prognoses. Benign tumors, which are usually slow-growing, respond well to conventional treatments. In contrast, the borderline serous ovarian tumors might be malignant potential, necessitating fertility-sparing surgery for fertile women who desire it. (du Bois et al., 2016; Chui et al., 2019). Moreover, therapy for ovarian cancer usually involves surgery and platinum/taxane doublet-based chemotherapy. (Lisio et al., 2019; Kuroki and Guntupalli, 2020). The diagnosis of serous ovarian tumors is difficult without incisional or aspiration biopsy. However, the varied characteristics of the serous ovarian tumors make it challenging to diagnose between borderline and malignant ovarian tumors using fine-needle aspiration. (Kuroki and Guntupalli, 2020). Therefore, it is crucial to develop a non-invasive and accurate preoperative identification technique for ovarian tumors for appropriate treatment planning by avoiding inadequate excision or surgical overtreatment, especially for premenopausal patients wanting to retain their fertility.

Adnexal ultrasound, a non-invasive, low-cost, and safe procedure, is currently the first-line imaging modality for ovarian tumor screening and diagnosis. Even though such pattern-recognition-based classification of ovarian masses into benign or malignant tumors demands much expertise, (Van Holsbeke et al., 2010; Dakhly et al., 2019)' there is a shortage of expert examiners. Radiomics offers automatic extraction of mineable high-dimensional quantitative data from clinical images, thereby bypassing the need for human intervention, and shows great promise in tumor detection, diagnosis, and prognostic evaluation. (Chiappa et al., 2020; Mayerhoefer et al., 2020). Several researchers have recently employed radiomics features based on MRI, CT and ultrasound to evaluate the clinical outcomes of ovarian cancer patients. (Rizzo et al., 2018; Lu et al., 2019; Zhang et al., 2019; Veeraraghavan et al., 2020; Yao et al., 2021).

This study utilizes a two-step radiomics classification of serous ovarian tumors based on the imaging and builds a nomogram combining the clinical factors to distinguish benign, borderline, and malignant ovarian tumors.

MATERIALS AND METHODS

Patients and Study Design

This study was in accordance with the Declaration of Helsinki. The Ethics Committee of Tianjin Medical University Cancer Hospital approved this retrospective study (Approval No. bc2021114), and informed consent was waived. All the clinical and biodatas have been anonymized. We enrolled 412 patients with ovarian tumor from Tianjin Medical University Cancer Institute and Hospital (Tianjin, China). All patients were enrolled between March 2013 to December 2016. Patients with mucinous tumor, endometrioid tumor, clear cell cancer, metastatic cancer and the tumor with poor quality ultrasound images were excluded from the study. In total, 265 patients meeting the inclusion criteria were enrolled consecutively in our study. The samples comprised 106 ovarian cystadenomas, 65 borderline tumors, and 108 ovarian malignancies, all of which were pathologically confirmed to be serous. Of the tumors we eventually included, the ultrasound images of 28 tumors were from 14 patients who had bilateral ovarian tumor (7 patients with bilateral borderline serous tumors, seven patients with bilateral ovarian serous cancer). The patient data included age, age at menarche, CA125 level (range: 5.11–5000 IU/L), location of the lesion (unilateral or bilateral), family history of cancer, and ascites. The inclusion criteria were as follows: 1) histological diagnosis of benign, borderline, and malignant ovarian serous tumors; 2) availability of preoperative US images suitable for diagnostic analysis; 3) US scanning performed before neoadjuvant therapy or surgical resection. The exclusion criteria included the following: 1) no US results or the ovarian mass was not completely visible in the image; 2) mucinous, clear cell, endometrioid, metastatic cancer (**Figure 1**).

In a two-step decision-making approach, two tasks were performed to train and validate the ability to distinguish between benign vs borderline and malignant (task 1) and borderline vs malignant (task 2). A clinical-feature alone model and a combined clinical-radiomic (CCR) model were

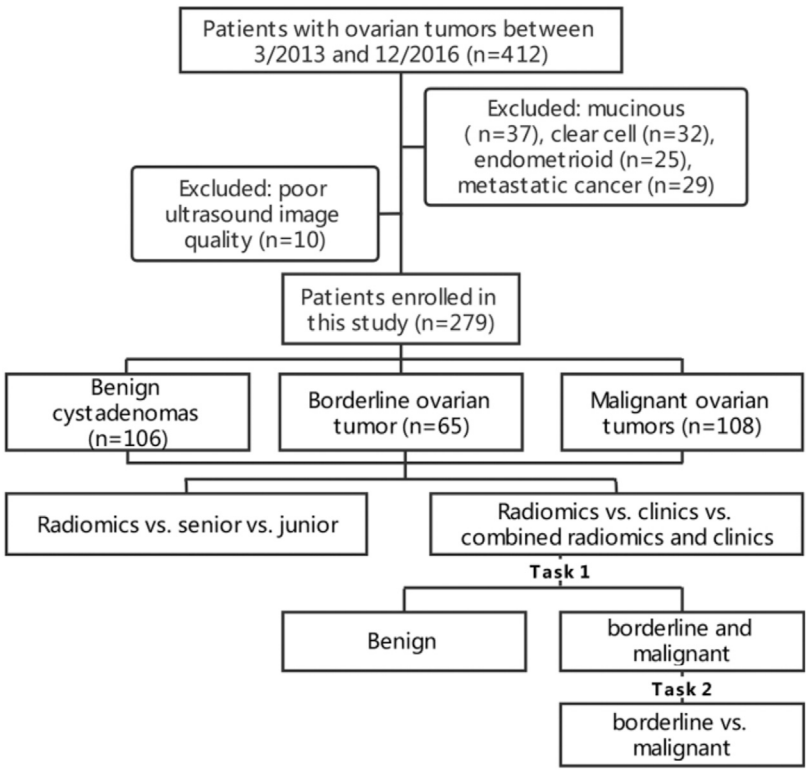

FIGURE 1 | Flowchart of patient recruitment and experiments design.

built using predictive nomograms for each task. During the development of these tasks, we used a fixed 70%/30% training/validation cohort split. A 10-fold cross-validation was done to evaluate the true diagnostic potential of this method.

Ultrasound Imaging and Segmentation

All ultrasound images were acquired using a Philips iU22/HD11 (California, United States) ultrasound machine with a 5–12 MHz probe and retrieved from the picture archiving and communication systems (PACS) for image segmentation and analysis at our institution (**Figure 2A**). The boundary of lesions manually segmented using ImageJ (https://imagej.nih.gov/ij/) by a sonographer with more than 8 years of experience. When the boundary was not determined, another experienced sonographer was consulted for a final opinion. The two sonographers were blind to the pathological and clinical information.

Radiomics Signature Construction

Eight hundred and fifty-five radiomics features, including shape, gray-scale histograms, texture, and wavelet features, were extracted automatically from each segmented region of interest using an in-house software written in MATLAB R2018b (MathWorks, Inc., Natick, Massachusetts). Detailed information on the feature extraction algorithms is provided in **Supplemental Table S1**.

For each task, we followed a three-step procedure to identify the reliable radiomic features. First, the Wilson test was used to identify features highly related to the biomarkers with a significance of less than 0.05 ($p < 0.05$). Pearson correlation matrices were used to assess the correlation between the features where a correlation coefficient greater than 0.8 was considered redundant. One of two features with a lower p-value was excluded. Next, the minor absolute shrinkage and selection operator (LASSO) regression method was used to select the most useful prognostic combination of features followed by the computation of the radiomics score (Radscore) for each patient through a linear combination of selected features weighted by their respective coefficients.

Human Readout

All images from the validation cohort were in random order subjected to critical evaluation by a senior (LCX, with 8 years of working experience) and a junior sonographer (LJH, with 2 years of working experience) in the ultrasound department, where each of them had carried out over 200 scans of ovarian ultrasound images per year. Both readers were blinded to the clinical information, study design, and background.

Nomogram Construction

Clinical factors, including age, CA125 level, lesion location, family history of cancer, ascites, and Radscore, were evaluated using univariate analysis in the training set. Variables with $p <$ 0.05 in the univariate analysis were included in the multivariate

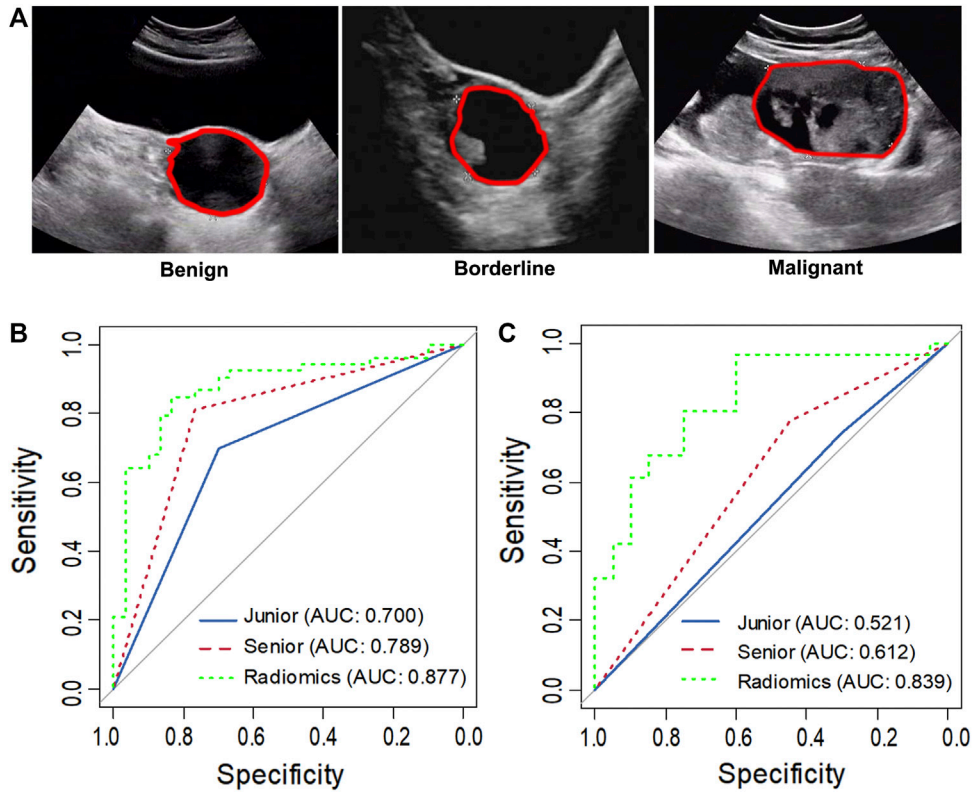

FIGURE 2 | Representative ultrasound images of benign, borderline and malignant ovarian serous tumors **(A)**. The asterisk indicates the tumor boundary. The red marker line indicates the region of interest (ROI). ROC curve analysis comparing the diagnosis of the senior and junior sonographer and radiomics in task 1 **(B)** and task 2 **(C)**.

logistic analysis. The clinical and CCR models were built using these clinical variables with or without a Radscore for each task. These models were presented in the form of a nomogram.

Statistical Methods

In this study, the continuous variables were presented as the mean (± standard deviation), and categorical variables were recorded as numbers and percentages. The chi-square test, Fisher's exact test, or Wilcoxon sum-rank test were used to identify categorical variables for the univariate analysis. Binary logistic regression analysis was used for multivariate analysis. Based on the factors mentioned above, the multivariate logistic regression model was adopted to establish two nomograms for diagnosing ovarian neoplasms: clinical-feature alone model vs CCR model. The performance of the nomogram was evaluated based on diagnostic accuracy, sensitivity, and specificity of receiver operating characteristic (ROC) curves and calibration curves. The difference in the area under the curve (AUC) between the training and validation datasets was tested using the *p*-value of integrated discrimination improvement (IDI) and Delong's (D) test, and the 95% confidence intervals (CI) were calculated.

All statistical analyses were conducted using the R software (version 6.1, R Foundation for Statistical Computing, Vienna, Austria). A two-tailed difference was considered significant at $p < 0.05$.

RESULTS

Evaluation of the Clinical Parameters of the Patients

The clinical features of patients in the training and validation cohorts for the two tasks were summarized in **Tables 1**, **2**. We observed a significant difference in the CA125 level, lesion location, and ascites between benign and non-benign serous ovarian lesions in the training cohort (**Table 1**). As shown in **Table 2**, age, CA125 level, and ascites significantly differed between the borderline and malignant serous ovarian tumors.

A Comparative Analysis of the Diagnostic Performances of the Radiomics Model, the Senior and Junior Sonographer

In task 1, LASSO was used to evaluate the diagnostic capability of 17 potential informative predictors (**Supplementary Figures S1A, C**), and the outputs were to Radscore calculation formula (**Supplemental Material**). We observed that the differences in the Radscore values between the benign and non-benign serous ovarian tumors in the training and validation cohorts were statistically significant ($p < 0.001$, **Supplementary Figures S2A, B**). The ROC curve analysis of

TABLE 1 | Clinical characteristics of patients in training and validation cohorts in task 1

Characteristics	Training cohorts		*p*-value		Validation cohorts		*p*-value
	Benign (n = 76)	Non-benign (n = 120)	Univariate analysis	Multivariate analysis	Benign (n = 30)	Non-benign (n = 53)	
Age[#]	51.2 ± 13.4	48.0 ± 13.5	0.102	—	49.1 ± 16.1	49.7 ± 11.2	0.861
Age at menarche[#]	14.6 ± 1.77	14.6 ± 1.85	0.869	—	14.6 ± 1.52	14.7 ± 1.69	0.119
CA125 level (IU/L), No (%)	—	—	<0.001*	<0.001*	—	—	<0.001*
0	75 (98.7)	73 (0.6)	—	—	0 (0.0)	26 (49.1)	—
1	1 (1.3)	47 (0.4)	—	—	30 (100.0)	27 (50.9)	—
Tumor side, No (%)		—	<0.001*	0.002*	—	—	<0.001*
Bilateral	15 (19.7)	69 (56.7)	—	—	4 (13.3)	34 (64.5)	—
Unilateral	61 (80.3)	52 (43.3)	—	—	26 (86.7)	19 (35.9)	—
Family history of cancer, No (%)		—	0.161	—	—	—	0.789
Yes	14 (18.4)	34 (28.3)	—	—	8 (26.7)	17 (32.1)	—
No	62 (81.7)	86 (71.7)	—	—	22 (73.3)	36 (67.9)	—
Ascites, No (%)	—	—	<0.001*	<0.001*	—	—	0.001*
Yes	0 (0.0)	39 (32.5)	—	—	30 (100.0)	36 (67.9)	—
No	76 (100.0)	81 (67.5)	—	—	0 (0.0)	17 (32.1)	—

*Note: Non-benign, borderline and malignant tumors, # mean ± SD, ≤500 IU/L, 0; >500 IU/L, 1. SD, standard deviation. *p value < 0.05.*

TABLE 2 | Clinical characteristics of patients in training and validation cohorts in task 2

Characteristics	Training cohorts		*p*-value		Validation cohorts		*p*-value
	Borderline (n = 45)	Malignant (n = 77)	Univariate analysis	Multivariate analysis	Borderline (n = 20)	Malignant (n = 31)	
Age[#]	43.8 ± 14.0	52.3 ± 9.06	<0.001*	<0.001*	36.7 ± 13.0	53.7 ± 12.2	<0.001*
Age at menarche[#]	14.2 ± 2.02	14.8 ± 1.49	0.084	—	14.0 ± 1.86	15.2 ± 1.95	0.039*
CA125 level (IU/L), No (%)		—	0.001*	0.003*	—	—	0.312
0	35 (77.8)	35 (45.4)	—	—	6 (30.0)	15 (48.4)	—
1	10 (22.2)	42 (54.5)	—	—	14 (70.0)	16 (51.6)	—
Tumor side, No (%)		—	0.432	—	—	—	1
Bilateral	24 (53.3)	48 (62.3)	—	—	12 (60.0)	18 (58.1)	—
Unilateral	21 (46.7)	29 (37.7)	—	—	8 (40.0)	13 (41.9)	—
Family history of cancer, No (%)		—	0.639	—	—	—	0.201
Yes	12 (26.7)	25 (32.5)	—	—	3 (15.0)	11 (35.5)	—
No	33 (73.3)	52 (67.5)	—	—	17 (85.0)	20 (64.5)	—
Ascites, No (%)	—	—	<0.001*	0.006*	—	—	0.125
Yes	5 (11.1)	33 (42.9)	—	—	16 (80.0)	17 (54.8)	—
No	40 (88.9)	44 (57.1)	—	—	4 (20.0)	14 (45.2)	—

*Note: # mean ± SD. ≤500 IU/L, 0; >500 IU/L, 1. SD, standard deviation. *p value < 0.05.*

the radiomics model showed AUCs of 0.907 (95% CI 0.863–0.950) and 0.877 (95% CI, 0.798–0.957) in the training and validation sets, respectively revealed no significant differences (D = 0.633; p = 0.5278). Next, we evaluated the diagnostic capability of the two sonographers to draw our comparative analysis. **Figure 2B**; **Table 3**; **Supplementary Table S2** showed the diagnostic performance of the junior sonographer, senior sonographer, and radiomics model, respectively. A statistically significant difference between the junior sonographer and the radiomics model (D = 3.611; p < 0.001) was observed. However, there was no statistically significant difference between the performances of the senior sonographer and the radiomics model (D = 1.473; p = 0.141).

In Task 2, 22 potential informative predictors were explored using the LASSO method (**Supplementary Figures S1B, D**). Differences in the Radscore value between the borderline and malignant serous ovarian tumors in the training and validation cohorts were statistically significant (p < 0.001, **Supplementary Figures S2C, D**). The ROC curves of the radiomics model showed AUCs of 0.891 (95% CI 0.833–0.950) and 0.839 (95% CI 0.725–0.952) in the training and validation cohorts, respectively, with no significant difference between them (D = 0.607; p = 0.546). **Figure 2C**; **Table 3**; **Supplementary Table S2** showed the diagnostic performance of the junior sonographer, senior sonographer, and radiomics model, respectively. There was a statistically significant difference between the performances

TABLE 3 | Diagnostic performance comparison among the senior sonologist, the junior sonologist, radiomics, clinics and combination of radiomics and clinics in the validation cohort of each task.

		AUC (95%CI)	ACC (95%CI)	SEN (95%CI)	SPE (95%CI)
Task 1	senior	0.789 (0.695–0.883)	0.795 (0.692–0.876)	0.697 (0.511–0.838)	0.860 (0.726–0.937)
	junior	0.699 (0.595–0.803)	0.699 (0.588–0.795)	0.568 (0.396–0.725)	0.804 (0.656–0.901)
	Radiomics	0.877 (0.798–0.957)	0.843 (0.747–0.914)	0.758 (0.574–0.883)	0.900 (0.774–0.963)
	Clinics	0.855 (0.786–0.924)	0.807 (0.706–0.886)	0.684 (0.512–0.820)	0.911 (0.779–0.971)
	Combination	0.914 (0.851–0.976)	0.880 (0.790–0.941)	0.813 (0.630–0.821)	0.922 (0.803–0.975)
Task 2	senior	0.612 (0.478–0.747)	0.647 (0.501–0.776)	0.563 (0.306–0.792)	0.686 (0.506–0.826)
	junior	0.521 (0.392–0.650)	0.569 (0.423–0.707)	0.429 (0.188–0.703)	0.622 (0.448–0.771)
	Radiomics	0.839 (0.725–0.952)	0.824 (0.691–0.916)	0.923 (0.621–0.996)	0.790 (0.622–0.899)
	Clinics	0.829 (0.706–0.950)	0.784 (0.647–0.887)	0.714 (0.477–0.878)	0.833 (0.645–0.937)
	Combination	0.890 (0.794–0.987)	0.863 (0.737–0.943)	0.842 (0.595–0.958)	0.875 (0.701–0.959)

AUC area under the receiver operator characteristic curves, ACC accuracy, SEN sensitivity, SPE specificity.

of the junior/senior sonographer and the radiomics model (senior: D = 3.5, $p < 0.001$; junior: D = -4.640, $p < 0.001$).

By comparing the results from the 10-fold cross-validation run of the models built above to the results obtained on the fixed training/validation split, we found that the performance estimates were comparable for both the tasks with no indication of substantial overfitting (**Supplementary Table S3**).

Construction and Validation of the Nomogram

Next, we utilized the features mentioned above for each task to perform multivariate logistic regression analysis to construct the two models for diagnosing ovarian neoplasms, thereby leading to the generation of two nomograms, the clinical-feature alone model (**Figure 3D** and **Figure 4D**) and the combined clinical-radiomic (CCR) model (**Figure 3A** and **Figure 4A**).

For task 1, **Figure 3**, **Figure 5** and **Table 3** showed the calibration curve and performance of the clinical-alone and CCR models. The ROC curves of the clinical-alone model showed AUCs of 0.817 (95% CI 0.765–0.868) and 0.855 (95% CI 0.786–0.924) in the training and validation cohorts, respectively (**Figures 5A,C**), with no significant difference between them (D = -0.88079; $p = 0.3796$). The ROC curves of the CCR model showed AUCs of 0.937 (95% CI 0.905–0.969) and 0.914 (95% CI 0.851–0.976) in the training and validation cohorts, respectively (**Figures 5A,C**), with no significant difference between them (D = 0.6394; $p = 0.524$). The calibration curve indicating the prediction from the two models (solid line) closely followed the 45-degree line in the training and validation cohorts, suggesting good diagnostic accuracy (**Figures 3B,C** for the CCR model and **Figures 3E,F** for the clinical alone model).

For task 2, the CCR performed satisfactorily in the training (AUC 0.924 [95% CI 0.876–0.971]) and the validation (AUC 0.890 [95% CI 0.794–0.987]) cohorts, respectively (**Figures 5B,D**), with no significant difference between them (D = 0.607; $p = 0.546$). The ROC curves of the clinical-feature alone model showed AUCs of 0.815 (95% CI 0.740–0.890) and 0.829 (95% CI 0.706–0.950) in the training and validation cohorts, respectively, with no significant difference between them (D = −0.189, $p = 0.85$). The calibration curve suggested good diagnostic accuracy for the CCR model (**Figures 4B,C**), which was slightly worse for the clinical-feature alone model (**Figures 4E,F**).

Difference in the Prediction Performance Between the Clinical Alone Model and Combined Clinical-Radiomic Model

As shown in **Table 3** and DCA curves (**Figure 6**), the CCR model showed a relatively better predictive performance than the clinical-feature alone model for two tasks (task 1: IDI = 0.154, 95% CI: 0.078–0.231, $p < 0.001$; task 2: IDI = 0.815, 95% CI: 0.066–0.303, $p = 0.002$). The decision curves indicated that using the clinical features combined radiomics nomogram to predict types of serous ovarian cancer adds more benefit than the clinical-feature alone model.

DISCUSSION

In this study, we divided the three-class classified (benign vs borderline vs malignant tumors) ovarian neoplasms into two categories, i.e., benign vs borderline and malignant (task 1) and borderline vs malignant (task 2). First, two US-imaging-based radiomics models were established for each task. The diagnostic efficiency of the radiomics models was compared with that of junior and senior sonographers to evaluate their integrity. Both tasks of radiomics analysis showed satisfactory performance, especially in task 2, indicating higher accuracy than the experienced sonographer at identifying borderline ovarian tumors. Then, the combined clinical-radiomics CCR model was established for each task, where the CCR models significantly outperformed the clinical models.

To date, US-based examinations were considered the primary imaging technique for preoperative prediction of ovarian tumors. (Di Legge et al., 2017). Benign serous ovarian tumors are typically simple smooth-walled unilocular or multilocular cystic masses, (Virgilio et al., 2019), whereas serous borderline ovarian tumors tend to form cystic masses with profuse papillary projections. (Timor-Tritsch et al., 2019). Moreover, serous ovarian tumors form large,

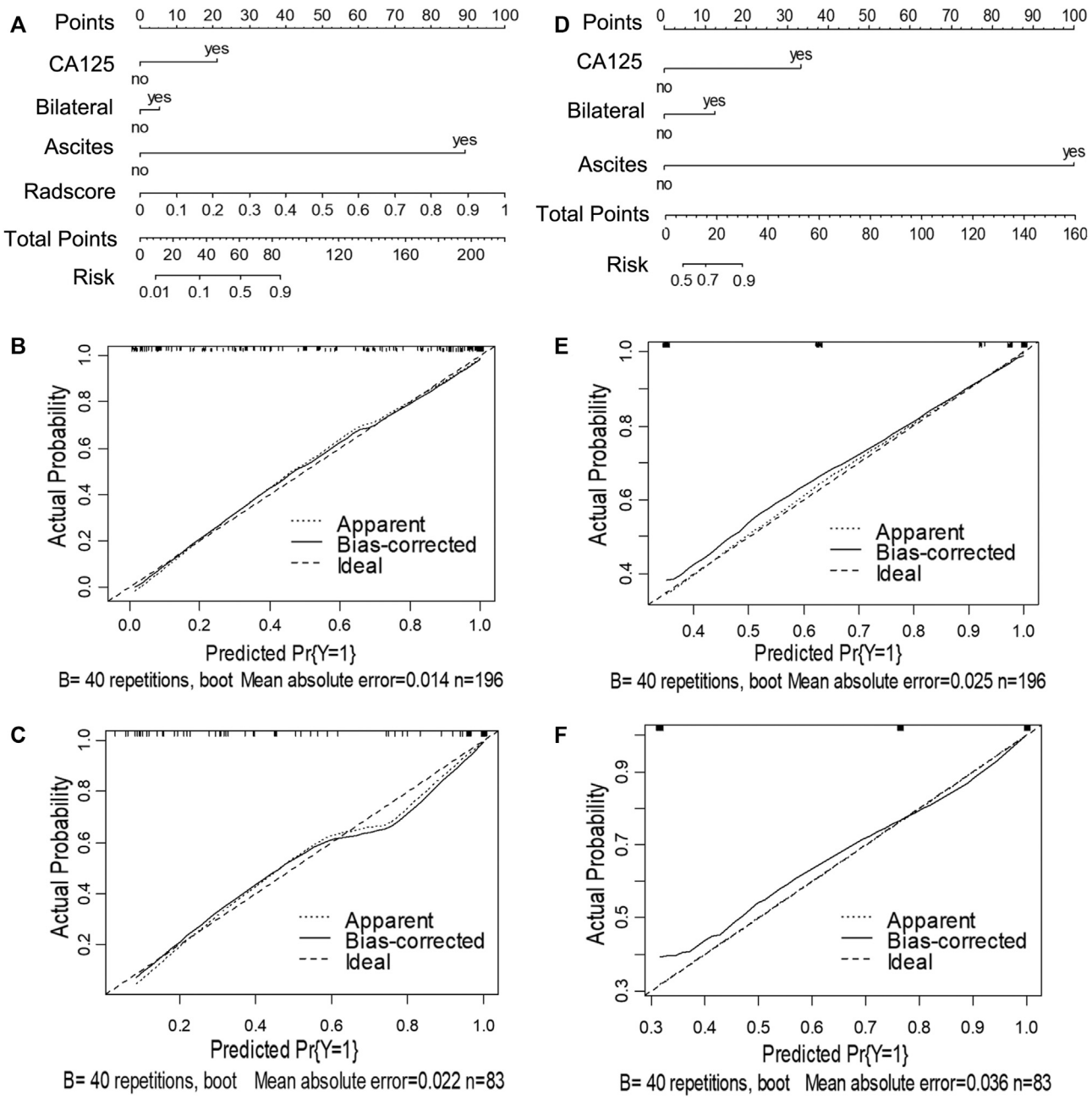

FIGURE 3 | Nomograms and calibration curves of the combined clinical-radiomic (CCR) and clinical alone model in task 1. The nomograms were constructed in the training cohort. **(A)** Nomogram in the CCR model. **(B, C)** Calibration curves in the training and validation cohort of the CCR nomogram, respectively. **(D)** Nomogram in the clinical alone model. **(E, F)** Calibration curves in the training and validation cohort of the clinical alone nomogram, respectively.

complex, solid, and cystic masses irregular, thick cystic walls with septations, necrosis, and solid mural nodules. (Moro et al., 2017). However, these imaging features are not specific and, to a certain extent are subject to the diagnostic experience of the sonographer. Nevertheless, conventional imaging evaluation by manual assessment of lesions by expert sonographers relying on semantic features provides a wealth of information on tumor heterogeneity, despite having a few drawbacks.

In this era of personalized and targeted oncology, radiomics enabled digitally encrypted medical images to be transformed into numerous quantitative features that provide information on tumor pathophysiology. (Bolton et al., 2012; Jiang et al., 2018; Mayerhoefer et al., 2020; Jian et al., 2021). To date, only one study has reported discriminating between benign and malignant ovarian tumors by computerized ultrasound image analysis using deep neural networks (DNNs). (Christiansen et al., 2021). However, distinguishing the borderline tumors using

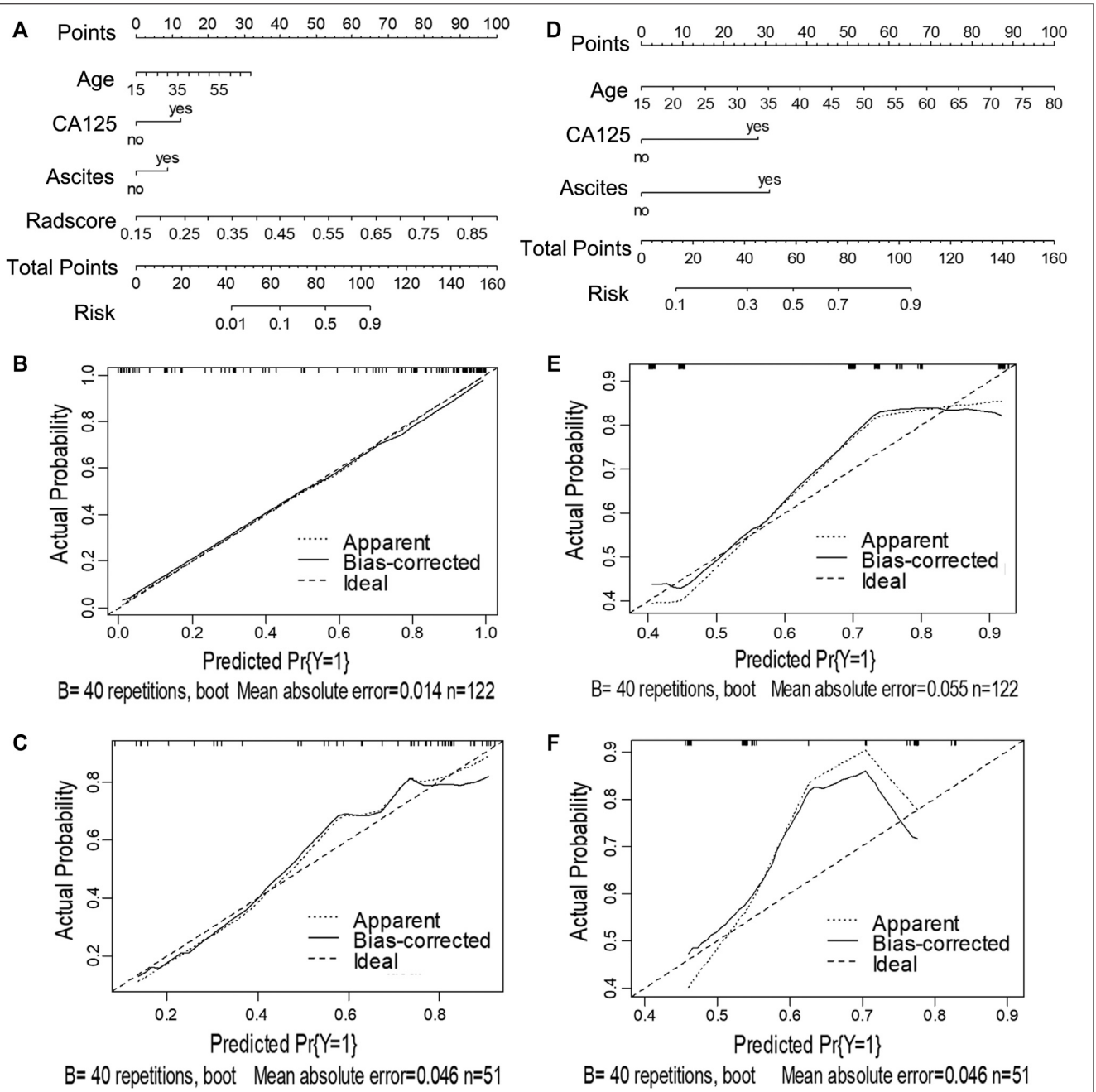

FIGURE 4 | Nomograms and calibration curves of the CCR and clinical alone model in task 2. The nomograms were constructed in the training cohort. **(A)** Nomogram in the CCR model. **(B, C)** Calibration curves in the training and validation sets of the CCR nomogram, respectively. **(D)** Nomogram in the clinical alone model. **(E, F)** Calibration curves in the training and validation cohort of the clinical alone nomogram, respectively.

DNNs remains largely unexplored. Additionally, some reports have indicated that the MRI radiomics model can achieve higher accuracy in discriminating benign ovarian lesions from malignancies and between type I and type II ovarian epithelial cancer. (Zhang et al., 2019; Qian et al., 2020). Pan et al. developed a nomogram model that combined CT radiomics and semantic features, which could be used for imaging biomarkers (radiomic and semantic features) to classify serous and mucinous types of ovarian cystadenomas. (Pan et al., 2020). Song and colleagues established classification predictive tasks constructed from radiomics features extracted from dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) pharmacokinetic protocol from 104 ovarian lesions to discriminate between benign, borderline, and malignant ovarian tumors. In consistence with our results, radiomics analysis based on the DCE-MRI pharmacokinetic protocol demonstrated good

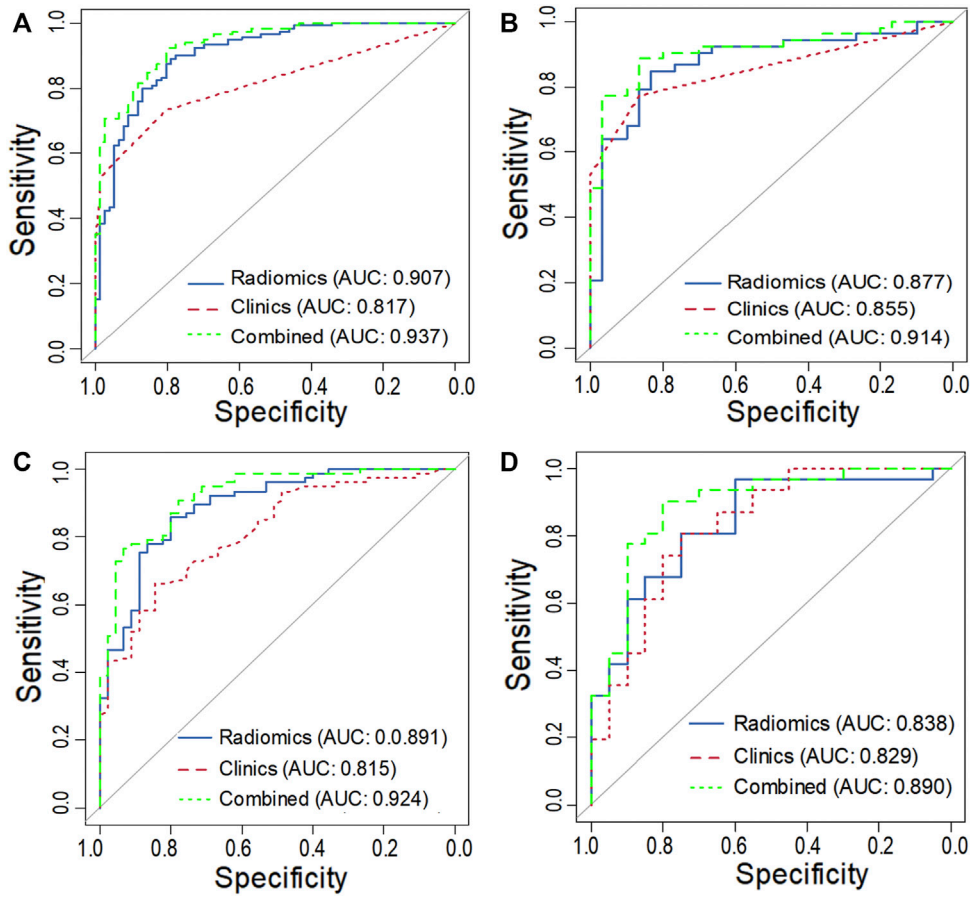

FIGURE 5 | ROC curves of the training and validation cohort in task 1 **(A, C)** and task 2 **(B, D)**.

differentiation between benign, borderline and malignant ovarian tumors in both two- and 3-class classification predictive tasks. (Song et al., 2020). To our knowledge, this is the first attempt to predict benign, borderline, and malignant ovarian serous tumors using radiomics features based on US images. The results of the 10-fold cross-validation confirmed those performance estimates, indicating no substantial overfitting.

Imaging features alone are often insufficient to determine the diagnosis and management of ovarian neoplasms. Hence, clinicians also consider the clinical context, including age, serological indicators, and familial risk factors, to make decisions. CA125 could serve as a critical serum biomarker for diagnosing and monitoring the relapse of serous ovarian cancer. (Matulonis et al., 2016). Ascites contain various cellular and acellular components that are known to facilitate metastasis and contribute to chemoresistance in ovarian serous cancer. (Ford et al., 2020). It is known that age is one of the most important poor prognostic markers for ovarian cancer. The incidence of ovarian cancer in women under 55 years of age is lower than that in women older than 55 years (Ma et al., 2019). Borderline and malignant serous ovarian tumors are more likely to occur in both ovaries. As expected, in this study cohort, the CA125 level was higher in the borderline and malignant serous ovarian tumor group than in the benign group. More borderline and malignant serous ovarian tumor cases were associated with ascites and showed involvement of both ovaries. (Jayson et al., 2014; Gershenson, 2017). We included these easily obtained clinical risk factors and US-based radiologic factors together with CA125 levels in our model development process. The improved nomogram model performed significantly better than the radiomics model or clinical model alone. The success of the nomogram model supported the idea that combining imaging features with complementary information from clinical reports that reflect the global outlook of the tumor is more helpful in the differential diagnosis of benign, borderline, and malignant serous ovarian tumors.

It is worth noting that the associations between the clinical variables and pathological diagnosis were discrepant in the training and validation cohorts. For example, CA125 level and ascites showed p values less than 0.05 in the training cohort, but they were not significantly associated with pathological diagnosis in the validation cohort of task 2. This result shows that clinical factors may be vulnerable to variations in data sets. However, radiomics features were consistently associated with pathological diagnosis and had accurate discriminative ability across all datasets.

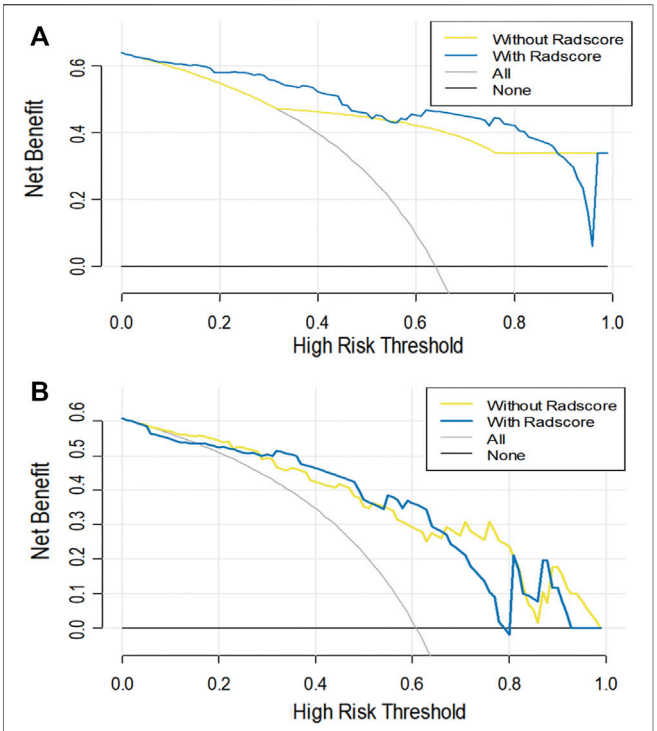

FIGURE 6 | The decision curve analysis for the clinical alone model and the CCR model in task 1 **(A)** and task 2 **(B)**. The *Y*-axis represents the net benefit. The red line represents the clinical alone monogram. The green line represents the CCR model. The grey line represents the assumption that all patients have ovarian tumors. The black line represents the assumption that no patient has ovarian tumors.

However, the present study has some limitations. First, this was a retrospective study conducted in a single hospital with limited sample size. External multi-center validation in a larger cohort is needed in the future to improve the radiomics analysis. Second, because ovarian tumors comprise benign, borderline, and malignant lesions, discrimination results among the three categories need to be obtained directly. Therefore, the need of the hour is a 3-class classification task of radiomics analysis based on US imaging, which will be developed in the future.

CONCLUSIONS

In conclusion, the current study presents a nomogram constructed from the US-based radiomics signature, clinical risk factors, and serum biomarkers. It could provide complementary diagnostic information to differentiate between benign, borderline, and malignant ovarian serous tumors, thereby contributing to reducing the number of unnecessary and risky biopsies and surgeries.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary files**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Tianjin Medical University Cancer Hospital approved this retrospective study (Approval No. bc2021114). Written informed consent from the patients was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

LQ and WM designed the study and wrote the manuscript. WM and DC performed the experiments and analyzed the data. CL and JL evaluated ultrasound images and segmented lesions. CZ and XL reviewed the manuscript. JW, GQ, HW, and XZ performed the experiments.

FUNDING

We are grateful for the support from grants from the National Natural Science Foundation of China (No. 81702161 to ZC, No. 81801781 and 82072004 to MWJ, No. 81402391 to ZXF) and the Science and amp; Technology Development Fund of Tianjin Education Commission for Higher Education (No. 2020KJ131 to LCX).

ACKNOWLEDGMENTS

We are grateful for the support from grants from the National Natural Science Foundation of China (No. 81702161 to ZC, No. 81801781 and 82072004 to MWJ, No. 81402391 to ZXF) and the Science and Technology Development Fund of Tianjin Education Commission for Higher Education (No. 2020KJ131 to LCX).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.753948/full#supplementary-material

REFERENCES

Bolton, K. L., Chenevix-Trench, G., Goh, C., Sadetzki, S., Ramus, S. J., Karlan, B. Y., et al. (2012). Association between BRCA1 and BRCA2 Mutations and Survival in Women with Invasive Epithelial Ovarian Cancer. *JAMA* 307, 382–390. doi:10.1001/jama.2012.20

Brett, M. R., Jennifer, B. P., Thomas, A. S., Jennifer, B. P., and Thomas, A. S. (2017). Epidemiology of Ovarian Cancer: a Review. *Cancer Biol. Med.* 14, 9–32. doi:10.20892/j.issn.2095-3941.2016.0084

Chiappa, V., Bogani, G., Interlenghi, M., Salvatore, C., Bertolina, F., Sarpietro, G., et al. (2020). The Adoption of Radiomics and Machine Learning Improves the Diagnostic Processes of Women with Ovarian MAsses (The AROMA Pilot Study). *J. Ultrasound*. doi:10.1007/s40477-020-00503-5

Christiansen, F., Epstein, E. L., Smedberg, E., Åkerlund, M., Smith, K., and Epstein, E. (2021). Ultrasound Image Analysis Using Deep Neural Networks for Discriminating between Benign and Malignant Ovarian Tumors: Comparison with Expert Subjective Assessment. *Ultrasound Obstet. Gynecol.* 57, 155–163. doi:10.1002/uog.23530

Chui, M. H., Xing, D., Zeppernick, F., Wang, Z. Q., Hannibal, C. G., Frederiksen, K., et al. (2019). Clinicopathologic and Molecular Features of Paired Cases of Metachronous Ovarian Serous Borderline Tumor and Subsequent Serous Carcinoma. *Am. J. Surg. Pathol.* 43, 1462–1472. doi:10.1097/PAS.0000000000001325

Dakhly, D. M. R., Gaafar, H. M., Sediek, M. M., Ibrahim, M. F., and Momtaz, M. (2019). Diagnostic Value of the International Ovarian Tumor Analysis (IOTA) Simple Rules versus Pattern Recognition to Differentiate between Malignant and Benign Ovarian Masses. *Int. J. Gynecol. Obstet.* 147, 344–349. doi:10.1002/ijgo.12970

Di Legge, A., Pollastri, P., Mancari, R., Ludovisi, M., Mascilini, F., Franchi, D., et al. (2017). Clinical and Ultrasound Characteristics of Surgically Removed Adnexal Lesions with Largest Diameter ≤ 2.5 Cm: a Pictorial Essay. *Ultrasound Obstet. Gynecol.* 50, 648–656. doi:10.1002/uog.17392

du Bois, A., Trillsch, F., Mahner, S., Heitz, F., and Harter, P. (2016). Management of Borderline Ovarian Tumors. *Ann. Oncol.* 27 (Suppl. 1), i20–i22. doi:10.1093/annonc/mdw090

Ford, C. E., Werner, B., Hacker, N. F., and Warton, K. (2020). The Untapped Potential of Ascites in Ovarian Cancer Research and Treatment. *Br. J. Cancer* 123, 9–16. doi:10.1038/s41416-020-0875-x

Gershenson, D. M. (2017). Management of Borderline Ovarian Tumours. *Best Pract. Res. Clin. Obstet. Gynaecol.* 41, 49–59. doi:10.1016/j.bpobgyn.2016.09.012

Javadi, S., Ganeshan, D. M., Qayyum, A., Iyer, R. B., and Bhosale, P. (2016). Ovarian Cancer, the Revised FIGO Staging System, and the Role of Imaging. *Am. J. Roentgenology* 206, 1351–1360. doi:10.2214/AJR.15.15199

Jayson, G. C., Kohn, E. C., Kitchener, H. C., and Ledermann, J. A. (2014). Ovarian Cancer. *The Lancet* 384, 1376–1388. doi:10.1016/s0140-6736(13)62146-7

Jian, J., Li, Y. a., Pickhardt, P. J., Xia, W., He, Z., Zhang, R., et al. (2021). MR Image-Based Radiomics to Differentiate Type I and Type II Epithelial Ovarian Cancers. *Eur. Radiol.* 31, 403–410. doi:10.1007/s00330-020-07091-2

Jiang, Y., Chen, C., Xie, J., Wang, W., Zha, X., Lv, W., et al. (2018). Radiomics Signature of Computed Tomography Imaging for Prediction of Survival and Chemotherapeutic Benefits in Gastric Cancer. *EBioMedicine* 36, 171–182. doi:10.1016/j.ebiom.2018.09.007

Kuroki, L., and Guntupalli, S. R. (2020). Treatment of Epithelial Ovarian Cancer. *BMJ* m3773, m3773. doi:10.1136/bmj.m3773

Lheureux, S., Braunstein, M., and Oza, A. M. (2019). Epithelial Ovarian Cancer: Evolution of Management in the Era of Precision Medicine. *CA A. Cancer J. Clin.* 69, 280–304. doi:10.3322/caac.21559

Lisio, M.-A., Fu, L., Goyeneche, A., Gao, Z.-h., and Telleria, C. (2019). High-Grade Serous Ovarian Cancer: Basic Sciences, Clinical and Therapeutic Standpoints. *Ijms* 20, 952. doi:10.3390/ijms20040952

Lu, H., Arshad, M., Thornton, A., Avesani, G., Cunnea, P., Curry, E., et al. (2019). A Mathematical-Descriptor of Tumor-Mesoscopic-Structure from Computed-Tomography Images Annotates Prognostic- and Molecular-Phenotypes of Epithelial Ovarian Cancer. *Nat. Commun.* 10, 764. doi:10.1038/s41467-019-08718-9

Ma, J., Ren, S., Ding, J., Liu, S., Zhu, J., Ma, R., et al. (2019). Expression of RRBP1 in Epithelial Ovarian Cancer and its Clinical Significance. *Biosci. Rep.* 39, BSR20190656. doi:10.1042/BSR20190656

Matulonis, U. A., Sood, A. K., Fallowfield, L., Howitt, B. E., Sehouli, J., and Karlan, B. Y. (2016). Ovarian Cancer. *Nat. Rev. Dis. Primers* 2, 16061. doi:10.1038/nrdp.2016.61

Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., et al. (2020). Introduction to Radiomics. *J. Nucl. Med.* 61, 488–495. doi:10.2967/jnumed.118.222893

Moro, F., Baima Poma, C., Zannoni, G. F., Vidal Urbinati, A., Pasciuto, T., Ludovisi, M., et al. (2017). Imaging in Gynecological Disease (12): Clinical and Ultrasound Features of Invasive and Non-invasive Malignant Serous Ovarian Tumors. *Ultrasound Obstet. Gynecol.* 50, 788–799. doi:10.1002/uog.17414

Pan, S., Ding, Z., Zhang, L., Ruan, M., Shan, Y., Deng, M., et al. (2020). A Nomogram Combined Radiomic and Semantic Features as Imaging Biomarker for Classification of Ovarian Cystadenomas. *Front. Oncol.* 10, 895. doi:10.3389/fonc.2020.00895

Qian, L., Ren, J., Liu, A., Gao, Y., Hao, F., Zhao, L., et al. (2020). MR Imaging of Epithelial Ovarian Cancer: a Combined Model to Predict Histologic Subtypes. *Eur. Radiol.* 30, 5815–5825. doi:10.1007/s00330-020-06993-5

Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Buscarino, V., Colarieti, A., et al. (2018). Radiomics of High-Grade Serous Ovarian Cancer: Association between Quantitative CT Features, Residual Tumour and Disease Progression within 12 Months. *Eur. Radiol.* 28, 4849–4859. doi:10.1007/s00330-018-5389-z

Song, X.-l., Ren, J.-L., Zhao, D., Wang, L., Ren, H., and Niu, J. (2020). Radiomics Derived from Dynamic Contrast-Enhanced MRI Pharmacokinetic Protocol Features: the Value of Precision Diagnosis Ovarian Neoplasms. *Eur. Radiol.* 31, 368–378. doi:10.1007/s00330-020-07112-0

Timor-Tritsch, I. E., Foley, C. E., Brandon, C., Yoon, E., Ciaffarrano, J., Monteagudo, A., et al. (2019). New Sonographic Marker of Borderline Ovarian Tumor: Microcystic Pattern of Papillae and Solid Components. *Ultrasound Obstet. Gynecol.* 54, 395–402. doi:10.1002/uog.20283

Van Holsbeke, C., Daemen, A., Yazbek, J., Holland, T. K., Bourne, T., Mesens, T., et al. (2010). Ultrasound Experience Substantially Impacts on Diagnostic Performance and Confidence when Adnexal Masses Are Classified Using Pattern Recognition. *Gynecol. Obstet. Invest.* 69, 160–168. doi:10.1159/000265012

Veeraraghavan, H., Vargas, H. A., Sánchez, A.-J., Micco, M., Mema, E., Lakhman, Y., et al. (2020). Integrated Multi-Tumor Radio-Genomic Marker of Outcomes in Patients with High Serous Ovarian Carcinoma. *Cancers* 12, 3403. doi:10.3390/cancers12113403

Virgilio, B. A., De Blasis, I., Sladkevicius, P., Moro, F., Zannoni, G. F., Arciuolo, D., et al. (2019). Imaging in Gynecological Disease (16): Clinical and Ultrasound Characteristics of Serous Cystadenofibromas in Adnexa. *Ultrasound Obstet. Gynecol.* 54, 823–830. doi:10.1002/uog.20277

Yao, F., Ding, J., Hu, Z., Cai, M., Liu, J., Huang, X., et al. (2021). Ultrasound-based Radiomics Score: a Potential Biomarker for the Prediction of Progression-free Survival in Ovarian Epithelial Cancer. *Abdom. Radiol.* 46, 4936–4945. doi:10.1007/s00261-021-03163-z

Zhang, H., Mao, Y., Chen, X., Wu, G., Liu, X., Zhang, P., et al. (2019). Magnetic Resonance Imaging Radiomics in Categorizing Ovarian Masses and Predicting Clinical Outcome: a Preliminary Study. *Eur. Radiol.* 29, 3358–3371. doi:10.1007/s00330-019-06124-9

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 October 2021
doi: 10.3389/fgene.2021.771830

The Predictive Role of Immune Related Subgroup Classification in Immune Checkpoint Blockade Therapy for Lung Adenocarcinoma

Xiaozhou Yu[1,2,3†], *Ziyang Wang*[1,2,3,4†], *Yiwen Chen*[1,2,3], *Guotao Yin*[1,2,3], *Jianjing Liu*[1,2,3], *Wei Chen*[1,2,3], *Lei Zhu*[1,2,3], *Wengui Xu*[1,2,3]* *and Xiaofeng Li*[1,2,3]*

[1]*Department of Molecular Imaging and Nuclear Medicine, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin, China,* [2]*Key Laboratory of Cancer Prevention and Therapy, Tianjin, China,* [3]*Tianjin's Clinical Research Center for Cancer, Tianjin, China,* [4]*Department of Molecular Imaging and Nuclear Medicine, Tianjin Cancer Hospital Airport Hospital, Tianjin, China*

OPEN ACCESS

Edited by:
Ming Fan,
Hangzhou Dianzi University, China

Reviewed by:
Chen Liu,
Peking University People's Hospital, China
Mengya Zang,
Southern Medical University, China

***Correspondence:**
Xiaofeng Li
xli03@tmu.edu.cn
Wengui Xu
wenguixy@yeah.net

[†]*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: *07 September 2021*
Accepted: *30 September 2021*
Published: *15 October 2021*

Citation:
Yu X, Wang Z, Chen Y, Yin G, Liu J, Chen W, Zhu L, Xu W and Li X (2021) The Predictive Role of Immune Related Subgroup Classification in Immune Checkpoint Blockade Therapy for Lung Adenocarcinoma. Front. Genet. 12:771830. doi: 10.3389/fgene.2021.771830

Background: In lung adenocarcinoma (LUAD), the predictive role of immune-related subgroup classification in immune checkpoint blockade (ICB) therapy remains largely incomplete.

Methods: Transcriptomics analysis was performed to evaluate the association between immune landscape and ICB therapy in lung adenocarcinoma and the associated underlying mechanism. First, the least absolute shrinkage and selection operator (LASSO) algorithm and K-means algorithm were used to identify immune related subgroups for LUAD cohort from the Cancer Genome Atlas (TCGA) database (n = 572). Second, the immune associated signatures of the identified subgroups were characterized by evaluating the status of immune checkpoint associated genes and the immune cell infiltration. Then, potential responses to ICB therapy based on the aforementioned immune related subgroup classification were evaluated *via* tumor immune dysfunction and exclusion (TIDE) algorithm analysis, and survival analysis and further Cox proportional hazards regression analysis were also performed for LUAD. In the end, gene set enrichment analysis (GSEA) was performed to explore the metabolic mechanism potentially responsible for immune related subgroup clustering. Additionally, two LUAD cohorts from the Gene Expression Omnibus (GEO) database were used as validation cohort.

Results: A total of three immune related subgroups with different immune-associated signatures were identified for LUAD. Among them, subgroup 1 with higher infiltration scores for effector immune cells and immune checkpoint associated genes exhibited a potential response to IBC therapy and a better survival, whereas subgroup 3 with lower scores for immune checkpoint associated genes but higher infiltration scores for suppressive immune cells tended to be insensitive to ICB therapy and have an unfavorable prognosis. GSEA revealed that the status of glucometabolic reprogramming in LUAD was potentially responsible for the immune-related subgroup classification.

Conclusion: In summary, immune related subgroup clustering based on distinct immune associated signatures will enable us to screen potentially responsive LUAD patients for ICB therapy before treatment, and the discovery of metabolism associated mechanism is beneficial to comprehensive therapeutic strategies making involving ICB therapy in combination with metabolism intervention for LUAD.

Keywords: lung adenocarcinoma, immune related subgroups, immune checkpoint blockade therapy, transcriptomics analysis, glucometabolic reprogramming

INTRODUCTION

Lung cancer is one of the most common type of malignancies worldwide, and is the leading cause of cancer-related death among men and women globally (Siegel et al., 2021). Non-small cell lung cancer (NSCLC), which includes squamous cell carcinoma, adenocarcinoma and large cell carcinoma, accounts for more than 80% of all primary lung cancers (Kano et al., 2020). Within NSCLC, adenocarcinoma is the most common histological subtype (Zhang et al., 2020). Despite great improvements in LUAD treatment in recent decades, particularly molecular-targeted therapeutic strategies, such as tyrosine kinase inhibitors (TKIs) treatment targeting epidermal growth factor receptor (EGFR) and/or anaplastic lymphoma kinase (ALK) (Ge and Shi, 2015), the prognosis for LUAD patients remains poor with a 5-years survival rate of only 15% (Siegel et al., 2021). Fortunately, as an emerging therapeutic approach for tumor, immunotherapy, such as immune checkpoint blockade (ICB) therapy, is increasingly approved to be effective for LUAD (Huang et al., 2020a). Cytotoxic T-lymphocyte antigen 4 (CTLA-4) and programmed cell death protein 1/programmed cell death ligand 1 (PD-1/PD-L1) are crucial immune checkpoints to maintain homeostasis for immune response (Meyers and Banerji, 2020). Actually, attenuated anti-tumor immune response or induced immunosuppression in local tumor microenvironment (TME) partially result from excessive negative immune response mediated by immune checkpoints (Anichini et al., 2020). ICB therapy aims to enhance anti-tumor immune response by inhibiting detrimental immunosuppression induced by immune checkpoint in TME.

Owing to heterogeneity existing in LUAD and development of acquired resistance to ICB therapy, the overall performance of ICB therapy in clinical practice for LUAD is far from satisfactory (Pathak et al., 2020). As one of the most immunological cancer type, immunological surveillance, immunoediting and immune escape play a critical role in LUAD development and progression (Song et al., 2020). Screening for potentially responsive LUAD patients to ICB therapy before treatment by using an effective immunoligical biomarker is beneficial to remarkably improve the outcome of LUAD patients with ICB therapy (Wu et al., 2020). Tumor-infiltrating lymphocyte (TIL) score and PD-L1 expression in TME are previously suggested as potential biomarkers to select potentially sensitive subpopulation to ICB therapy prior to treatment and to predict survival for LUAD patients (Gascón et al., 2020; Jin et al., 2020; Hashemi et al., 2021). However, evaluations for the status of TIL and PD-L1 are currently non-standardized and limited by tissue samples availability. A comprehensive analysis of the immune associated signature in TME enable a further understanding of the interplay between local immune status and tumor immunotherapy responsiveness (Park et al., 2020; Wang et al., 2020).

"Omics" techniques which are characterized by high-throughput interfaces are able to investigate complex biological systems in order to identify molecular signatures responsible for the complicated biological phenotype (Gillette et al., 2020; Lazarou et al., 2020). In the present investigation, bioinformatics analyses based on ribonucleic acid (RNA) sequencing (RNA-seq) data and clinical information from Cancer Genome Atlas (TCGA) database were performed to comprehensively explore the predictive role of immune associated signature in therapeutic responsiveness to ICB therapy for LUAD.

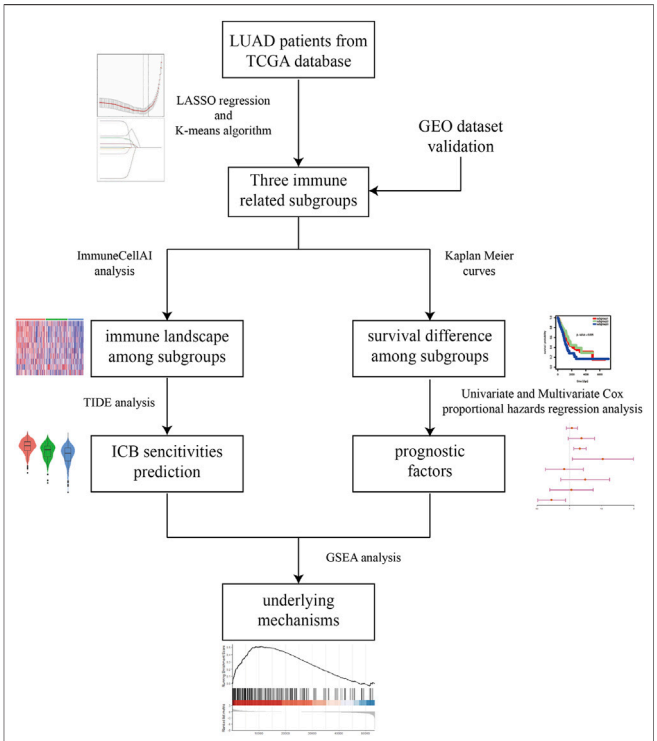

FIGURE 1 | The workflow of this study. Briefly, immune related subgroup clustering was performed by using LASSO algorithm and K-means algorithm. After characterization of the immune associated signatures of the identified subgroups, TIDE algorithm analysis was performed to predict the potential sensitivities to ICB therapy. Meanwhile, survival analysis and further Cox proportional hazards regression analysis were also performed for LUAD. In the end, GSEA was performed to explore the metabolic mechanism potentially responsible for immune related subgroup clustering.

TABLE 1 | Clinicopathological characteristics of LUAD patients from the training and validation sets.

Characteristics	TCGA	GSE68465	GSE72094
	Training set	Validation set	Validation set
Patient numbers	751	443	442
Age	65.2 ± 10.0	64.4 ± 10.1	69.2 ± 9.3
Gender	—	—	—
Male	342	223	202
Female	409	220	240
Tumor stages	—	—	—
Not reported	10	—	28
I	409	—	265
II	176	—	69
III	118	—	63
IV	38	—	17
Race	—	—	—
Not reported	70	129	45
Caucasian	581	295	399
African	84	12	13
Asian	16	7	3
Smoking history	—	—	—
Not reported	22	94	74
Never	108	49	33
Ever	621	300	335

First, immune related subgroup clustering was performed by using the least absolute shrinkage and selection operator (LASSO) algorithm and K-means algorithm. Second, the immune associated signatures of the identified subgroups were characterized by evaluating the status of immune checkpoint associated genes and the immune cells infiltration. Then, potential responses to ICB therapy were predicted via tumor immune dysfunction and exclusion (TIDE) algorithm analysis, and the relationship between the immune associated signature based on the aforementioned immune related subgroup classification and potential sensitivities to ICB therapy were determined. Additionally, survival analysis and further Cox proportional hazards regression analysis were also performed for LUAD, and gene set enrichment analysis (GSEA) was performed to explore the metabolic mechanism potentially responsible for immune related subgroup clustering. In the end, two microarray data sets from the Gene Expression Omnibus (GEO) database were used as validation cohorts in the study. The work flow of this study was shown in **Figure 1**.

Transcriptomics analysis of the association between immune associated signature and ICB therapy in LUAD not only explains for the heterogeneity in the reactivity to ICB therapy partially from an immunological perspective, but also provide potentially promising biomarker or target to direct sensitive LUAD patients screening prior to ICB therapy and combination therapy strategy making involving ICB therapy in combination with metabolism intervention.

MATERIALS AND METHODS

Data Acquisition

The RNA-seq data sequenced on the Illumina RNA sequencing platform for LUAD samples from TCGA samples were download from the Cancer Genomics Browser of the University of California Santa Cruz (UCSC) Xena (https://xena.ucsc.edu/public) (Cline et al., 2013). Then, log2 (x+1) transformed HT-seq counts data and Fragments Per Kilobase Million (FPKM) data were selected for further analysis. The corresponding phenotype and survival information were also downloaded from the UCSC Xena. The latest gene ID annotation file (gencode.v32. annotation.gtf) was downloaded from the GENCODE database (http://www.gencodegenes.org) (Frankish et al., 2019) for Entrez gene ID and Ensembl gene ID transformation. Finally, after matching the TCGA sample ID in RNA-seq with the corresponding phenotype and survival information, a total of 572 LUAD samples in TCGA database were included in the study. Meanwhile, a total of 824 genes directly involved in immunological processes were collected using the Immunome database (Breuer et al., 2013). In addition, Microarray data for 398 LUAD samples in GSE72094 and 442 LUAD samples in GSE68465 were also acquired from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). The corresponding gene chip annotation messages and clinical messages of this two data sets were downloaded using the R package GEOquery (Davis and Meltzer, 2007). The clinicopathological characteristics of LUAD patients from the training and validation sets were summarized in **Table 1**.

Data Preprocessing and Immune Related Subgroup Clustering

RNA-seq data and microarray data for LUAD from public database were first standardized for further analysis. "Combat" algorithm (Johnson et al., 2007) of R package sva (Leek et al., 2012) was employed to reduce the batch effect which may lead to deviations and bias to unrelated biological or scientific differences between subgroups (Leek et al., 2010). To filter out the missing values, intersective genes were selected from the TCGA cohort, GSE72094 cohort, GSE68465 cohort and Immunome database in the current study. Based on the expression of intersective genes for LUAD cohort from the TCGA database, LASSO algorithm and 10-fold cross validation method in R package glmnet (Friedman et al., 2010) were used to select the optimal gene set of the immune associated genes for immune related subgroup clustering. The total within sum of square and average silhouette width were calculated using R package factoextra to identify the best number of clustering. K-means algorithm, a classical unsupervised learning algorithm of artificial intelligence, was used for sample clustering in R software version 3.6.0 (https://www.r-project.org/) by 10 iterations with at least 30 samples for each subgroup. Moreover, consensus matrix analysis was performed in each data set to validate the clustering number, and consensus matrices were generated using the R package ConsensusClusterPlus (Wilkerson and Hayes, 2010). The principal component analysis (PCA) plot of the clustered samples were also drawn in the present study.

Evaluation of Immune Cell Infiltration Scores and Immune Checkpoint Associated Genes Scores in Tumor Microenvironment as the Immune Associated Signature

Immune cell Abundance Identifier (ImmuCellAI) (Miao et al., 2020), a gene set signature-based method, was used to evaluate the

infiltration scores of immune cells in the TME of LUAD. ImmuCellAI is capable of precisely estimating the abundance of 24 types of immune cell, including 18 T-cell subsets ($CD4^+$, $CD8^+$, $CD4^+$ naïve, $CD8^+$ naïve, central memory T (Tcm), effector memory T (Tem), Tr1, induced regulatory T cells (iTreg), natural regulatory T cells (nTreg), Th1, Th2, Th17, Follicular helper T cells (Tfh), cytotoxic T cells (Tc), mucosal-associated invariant T cells (MAIT), exhausted T cells (Tex), gamma delta T ($\gamma\delta$ T), and natural killer T (NKT) cells) and six other important immune cells (B cells, macrophages, monocytes, neutrophils, dendritic cell (DC), and natural killer (NK) cells). In addition, it was reported that ImmuCellAI can estimate the abundance of immune cells with superior accuracy to other methods, especially on many T-cell subsets. Immune checkpoint associated genes, such as CTLA4, CD28, CD80, CD86, CD274 (PD-L1) and PD-1 (PDCD1), were selected from previous relevant studies focusing on the correlation between these genes and LUAD development, progression and prognosis.

Prediction of Potential Sensitivity to Immune Checkpoint Blockades Therapy for Lung Adenocarcinoma Patients Based on Immune Related Subgroup Classification

Tumor immune dysfunction and exclusion (TIDE) algorithm (Fu et al., 2020) was used to calculate the potential possibility to respond to ICB therapy for LUAD patients based on immune related subgroup classification. Generally, TIDE analysis mainly consists of scores for TIDE, immune dysfunction, immune exclusion and several immune associated cells and effector molecules. Among which, negative score for TIDE suggests a lack of immune evasion phenotype. Meanwhile, T dysfunction score shows how a gene interacts with cytotoxic T cells to influence patient survival outcome, and the T cell exclusion score assesses the gene expression levels in immunosuppressive cell types that drive T cell exclusion. Scores for suppressive immune cells, such as cancer associated fibroblasts (CAF), myeloid-derived suppressor cell (MDSC), M2 macrophage indicate immune evasion or immunosuppression, suggesting a low possibility to respond to ICB therapy. Whereas, scores for effector immune cells, associated effector molecular and immune checkpoint associated genes, such as $CD8^+$T cells, interferon-γ (IFN-γ) and PD-L1 (CD274) represent a potential sensitivity to ICB therapy. Additionally, immune related subgroup clustering, immune associated cells infiltration, immune checkpoint associated genes and clinicopathologic parameters, such as age, gender, pathological TNM stages, tumor stages in LUAD were also evaluated and analyzed between different immune related subgroups to perform a Cox proportional hazards regression analysis.

Gene Set Enrichment Analysis (GSEA) to Explore the Underlying Mechanism Responsible for the Immune Related Subgroup Clustering of Lung Adenocarcinoma

GSEA is a bioinformatics analysis to determine whether a prior defined set of genes shows statistically significant and concordant differences between two groups (Sun et al., 2020). GSEA version 4.1.0, was used, the number of permutations was set to 1,000, and FDR <0.05 was the screening threshold. Given a close relationship between glucose metabolism reprogramming in tumor and anti-tumor immunomodulation, glucose metabolism process associated gene signatures, including the process of glycolysis, gluconeogenesis, tricarboxylic acid (TCA) cycle and oxidative phosphorylation (OXPHOS) in mitochondria were compared between the identified immune related subgroups (subgroup 1 vs subgroup 3) to explore the underlying mechanism responsible for the immune related subgroup clustering of LUAD.

Statistical Analysis

The differences of immune associated signatures existed between immune related subgroups, such as the expression of immune check point genes and the infiltration scores of immune associated cells, were evaluated by using Kruskal-Wallis test. Before that, Shapiro-Wilk test and Tukey's test were used to evaluate the status of normal distribution, and F test was used to perform homogeneity tests of variances. In addition, a survival analysis (overall survival) using Kaplan-Meier method was performed for LUAD patients, and the log-rank test was used to compare the differences of survival existed between the immune related subgroups aforementioned. Furthermore, univariate Cox proportional hazards regression analysis was performed to determine the correlation between survival and a variety of factors, including clinicopathologic parameters and immune associated signature factors. Afterwards, significantly associated factors were selected for further multivariate Cox proportional hazards regression analysis to determine independent risk factors. A p-value under 0.05 was considered to indicate a statistically significant difference. Data was analyzed using R software version 3.6.0. Multiple testing was corrected using the Benjamini-Hochberg's false rediscovery rate (FDR).

RESULTS

Immune-Associated Subgroup Clustering for Lung Adenocarcinoma From the Cancer Genome Atlas Database

The LASSO algorithm and 10-fold cross-validation were used to extract the optimal subsets of immune associated genes based on Immunome database for immune related subgroup clustering of LUAD cohort from TCGA database. As shown in **Figure 2A**, the optimal λ which have the minimum mean square error was selected by 10-fold cross validation. LASSO coefficient profile of the selected subsets of immune associated genes (n = 11) at the optimal λ for immune related subgroup clustering of LUAD was depicted in **Figure 2B**. To optimize the average silhouette width and the total within sum of square, the optimal number of clustering was set with k = 3 (**Figures 2C,D**). Based on this clustering, LUAD cohort (n = 572) from TCGA was divided into subgroup 1 (n = 252), subgroup2 (n = 188) and subgroup 3 (n = 132). The consensus matrix and the principal component analysis (PCA) plots of this immune related subgroup classification when

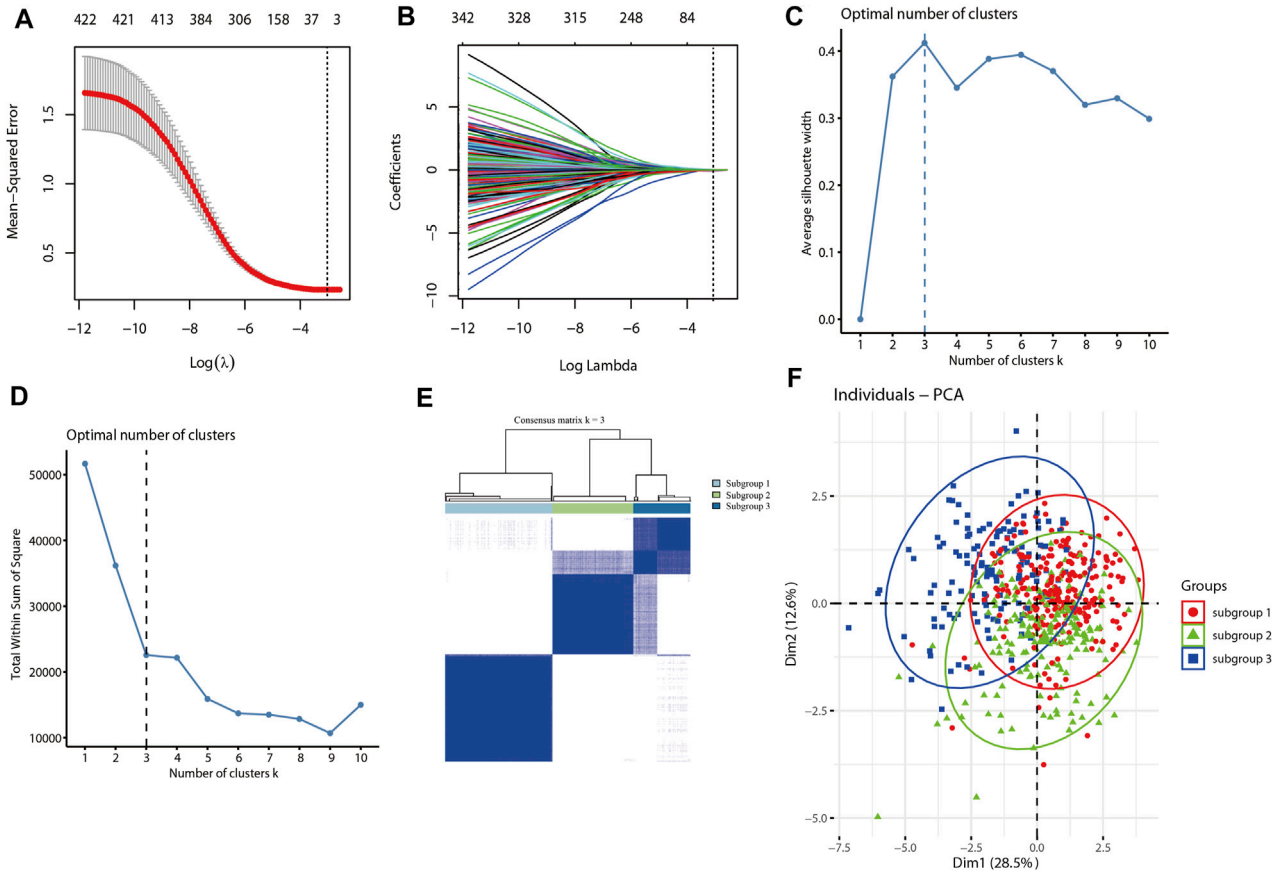

FIGURE 2 | Immune associated subgroup clustering of LUAD by using the LASSO and K-means algorithm. **(A)** LASSO regression model with 10-cross validation was used to select the optimal λ (dash line) which have the minimum mean square error (red dots). **(B)** LASSO coefficient profiles of the selected subsets of immune associated genes at the optimal λ (grey line) for immune related subgroup clustering of LUAD. **(C)** The curve of average silhouette width under corresponding cluster number k, and the maximum of average silhouette width was achieved when k = 3. **(D)** The curve of total within sum of squared error curve under corresponding cluster number k, and it reached the "elbow point" when k = 3. **(E)** The consensus clustering of immune related subgroup of LUAD when k = 3. **(F)** The PCA plot of clustered samples in the LUAD, where samples in subgroup-1 (*n* = 252) are shown in red, subgroup-2 (*n* = 188) in green and subgroup-3 (*n* = 132) in blue.

k = 3 was shown in **Figure 2E** and **Figure 2F**, respectively. The results about this clustering were further validated in GSE72094 and GSE68465 data sets (**Supplementary Figure S1**).

Characterization of Immune Associated Signature Based on Immune Related Subgroup Clustering of Lung Adenocarcinoma

In the current investigation, immune checkpoint associated genes and immune cell infiltration scores were used to represent the immune associated signature of each of the immune related subgroups of LUAD. The status of immune checkpoint associated genes, such as CTLA4, CD28, CD80, CD86, PD-L1 (CD274) and PD-1 (PDCD1), were first evaluated for LUAD based on the aforementioned immune related subgroup clustering. As demonstrated in the heatmap (**Figure 3A**), the levels of these immune checkpoint associated genes were significantly different between the three subgroups ($p < 0.05$). Box plots were also used to show the differences in each of these immune checkpoint associated genes between the three subgroups (**Figure 3B**). Generally, subgroup 1 tended to have significantly higher expression levels of immune checkpoint associated genes in comparison with other subgroups, particularly with subgroup 3. Next, immune cell infiltration estimation was performed by using ImmuCellAI. As shown in **Figure 3C**, the general infiltration score was higher in subgroup1 in contrast with other subgroups, and a total of 16 immune cell infiltration scores were found to be statistically different between the three immune-related subgroups. In detail, the infiltration scores for effector immune cells, such as $CD8^+$ cells and cytotoxic cells, were found to be statistically higher in subgroup 1 than that in other subgroups, whereas CD8 naive cell infiltration score was relatively lower in subgroup 1 compared to other subgroups. Meanwhile, the cell infiltration scores for suppressive immune cells, such as natural regulatory T cells (nTreg) and induced regulatory T cells (iTreg) were significantly higher in subgroup 3 than that in the other subgroups. (**Figure 3D**). Similar results with regard to the characterization of immune associated signature based on immune related subgroup clustering of

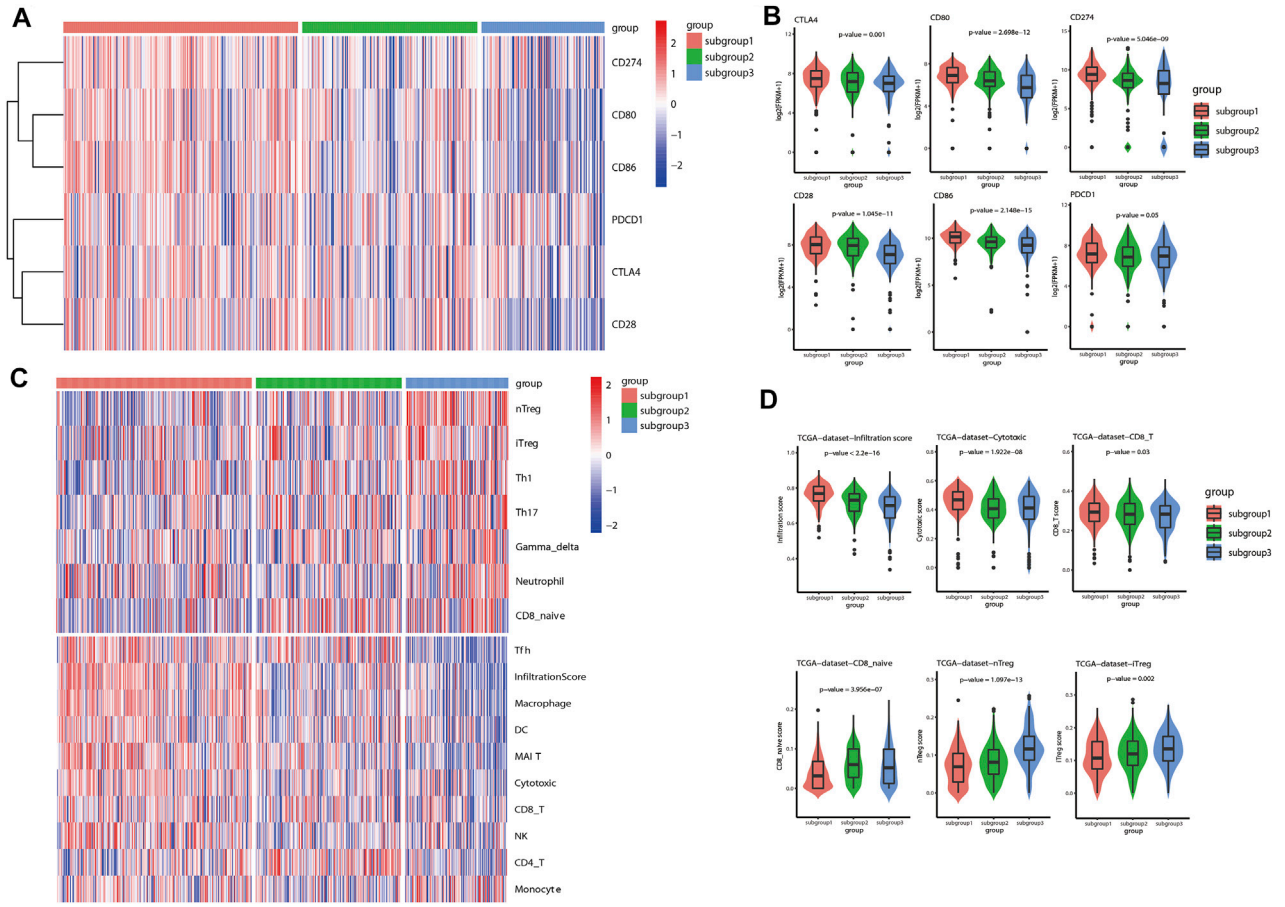

FIGURE 3 | Characterization of immune associated signatures of the identified immune related subgroups of LUAD. **(A)** As shown in the gene expression heatmap, the levels of immune checkpoint associated genes, including CTLA4, CD28, CD80, CD86, CD274 (PD-L1) and PDCD1 (PD-1), were significantly different between the identified immune related subgroups of LUAD ($p < 0.05$). **(B)** Box plots were also shown to demonstrate the differences in each of these included immune checkpoint associated genes between the three subgroups. Generally, subgroup 1 tended to have higher expression levels of immune checkpoint associated genes in comparison with other subgroups ($p < 0.05$). **(C)** ImmuCellAI was used to evaluate the immune cell infiltration scores in the TME of LUAD. As shown in the heatmap, immune cell infiltration scores were found to be statistically different between the three immune related subgroups. The general infiltration score was remarkably higher in subgroup 1 in comparison with other subgroups, particularly with subgroup 3. **(D)** Box plots were also shown to indicate the differences between the three subgroups with regard to the infiltration scores of several representative immune cells. The infiltration scores of positive immune response, such as $CD8^+$ cells and cytotoxic cells, were significantly higher in subgroup 1 than that in the other subgroups. Whereas, the infiltration scores of negative immune response, such as natural regulatory T cells (nTreg) and induced regulatory T cells (iTreg), were found to be statistically higher in subgroup 3 than that in other subgroups.

LUAD were also validated in GSE72094 and GSE68465 data sets (**Supplementary Figure S2** and **Supplementary Figure S3**).

Estimation of Potential Sensitivity to Immune Checkpoint Blockades Therapy for Lung Adenocarcinoma Based on Immune Related Subgroup Clustering

Tumor immune dysfunction and exclusion (TIDE) algorithm was used to evaluate the potential sensitivity to ICB therapy for LUAD patients included in different immune related subgroups. As shown in the heatmap (**Figure 4A**), the TIDE analysis associated scores were significantly different between the three subgroups ($p < 0.05$). Based on the TIDE analysis, a higher potential sensitivity to ICB therapy was suggested for subgroup 1 which was with higher scores for TIDE, dysfunction, $CD8^+$ cells, and interferon-γ (IFN-γ), but with lower scores for exclusion, M2 macrophage and MDSC in comparison with that in subgroup 3 (**Figure 4B**). Similarly, this TIDE analysis results were also validated in GSE72094 data sets (**Supplementary Figure S3**). GSE68465 data set was not used as validation cohort to perform TIDE analysis and survival analysis because of the lack of information for CD274.

Survival Analysis and Cox Proportional Hazards Regression Analysis for Lung Adenocarcinoma

With regard to survival analysis for LUAD, the Kaplan Meier curves were drawn and the log-rank test was performed in

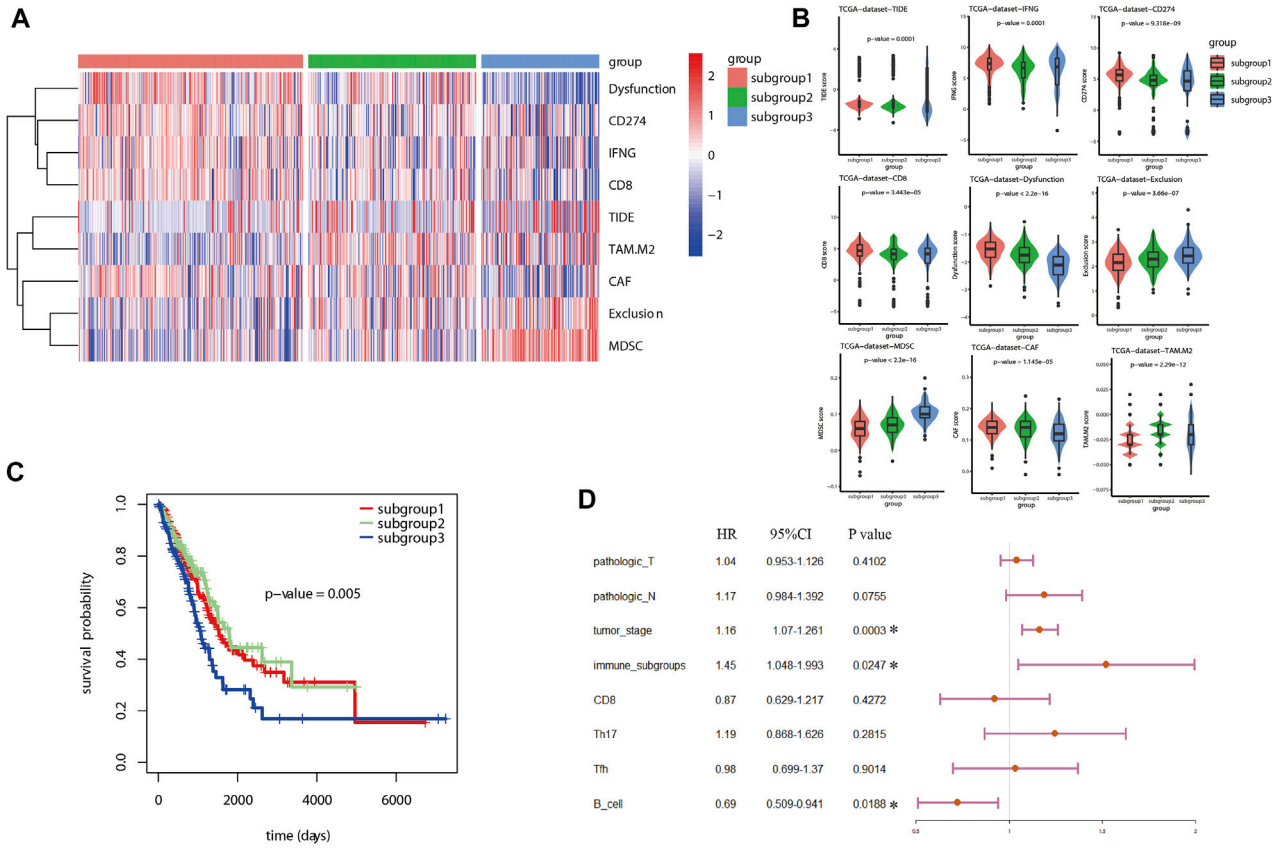

FIGURE 4 | TIDE analysis and survival analysis for LUAD based on immune related subgroup clustering. **(A)** TIDE analysis was used to evaluate the potential sensitivity to ICB therapy for LUAD patients. As shown in the heatmap, the TIDE analysis associated scores were significantly different between the three subgroups ($p < 0.05$). **(B)** Based on the TIDE analysis, a higher potential sensitivity to ICB therapy was suggested for subgroup 1 which was with higher scores for TIDE, dysfunction, $CD8^+$ cells and interferon-γ (IFN-γ), but with lower scores for exclusion, M2 macrophage and MDSC in comparison with that in subgroup 3. **(C)** Kaplan Meier analysis was performed to estimate the survival of LUAD. As shown, subgroup 3 tended to have an unfavorable prognosis in comparison with subgroup 1 and subgroup 2. **(D)** A total of 8 factors were included in further multivariate Cox proportional hazards regression analysis to identify independent risk factors for LUAD after an univariate Cox proportional hazards regression analysis. As shown in the forest plots, immune associated subgroup clustering, tumor stage and B cells infiltration were suggested as potential independent factors influencing overall survival (OS) of LUAD ($p < 0.05$).

this study. As demonstrated in **Figure 4C**, subgroup 3 tended to have an unfavorable prognosis in comparison with that of subgroup 1. Then, univariate Cox proportional hazards regression analysis was performed to identify the significant factors influencing the overall survival (OS) of LUAD. Among all the included factors, including the clinicopathologic parameters, immune checkpoint associated genes, immune cell infiltration scores, TIDE algorithm scores and immune related subgroup classification, a total of eight factors were proved to be significant risk factors influencing survival of LUAD (**Table 2**). Afterwards, all the eight factors were included in further multivariate Cox proportional hazards regression analysis to identify independent risk factors for LUAD. As shown in the forest plots (**Figure 4D**), immune related subgroup clustering, tumor stage and B cell infiltration were suggested as potential independent factors influencing OS of LUAD ($p < 0.05$). The results of survival analysis and Cox proportional hazards regression analysis in validation data set (GSE72094) were also shown in **Supplementary Figure S4**.

Potential Metabolism Associated Mechanism Responsible for Immune Related Subgroup Clustering of Lung Adenocarcinoma

Based on the immune related subgroup clustering (subgroup 1 vs subgroup 3), gene set enrichment analyses (GSEA) was performed on LUAD data set from the TCGA database using the gene sets significantly associated with glucose metabolism, including the process of glycolysis (**Figure 5A**), tricarboxylic acid (TCA) cycle (**Figure 5B**), gluconeogenesis (**Figure 5C**), oxidative phosphorylation (OXPHOS) in mitochondria (**Figure 5D**). FDR (Q value) < 0.05 was set as the screening threshold. As shown, the upward parabolas indicated that all the included processes of glucose metabolism was enhanced in subgroup 1 in contrast with that in subgroup 3. Glucose metabolic reprogramming was

TABLE 2 | Univariate Cox proportional hazards regression analysis of the prognostic factors for overall survival of LUAD.

Characteristics	HR	95% CI	P Value
Clinical features	—	—	—
Gender	1.05	0.79–1.41	0.72
Pathologic_T	1.18	1.09–1.27	< **0.01***
Pathologic_N	1.36	1.2–1.55	< **0.01***
Pathologic_M	0.98	0.9–1.07	0.68
Age	1.01	0.99–1.02	0.30
Tumor_stage	1.24	1.17–1.32	< **0.01***
Immune_subgroups	1.24	1.03–1.48	**0.02***
Gene mutation	—	—	—
TP53	1.21	0.91–1.63	0.19
EGFR	1.4	0.94–2.1	0.10
KRAS	1.13	0.82–1.56	0.46
TIDE	—	—	—
TIDE	1.03	0.74–1.45	0.84
IFNG	1.13	0.84–1.52	0.42
CD274	1.16	0.85–1.57	0.35
CD8	0.72	0.54–0.97	**0.02***
Dysfunction	0.85	0.63–1.14	0.26
Exclusion	1.31	0.98–1.77	0.07
CAF	1.11	0.83–1.49	0.47
TAM.M2	0.97	0.72–1.31	0.86
ImmuCellAI	—	—	—
CD4_naive	0.81	0.52–1.25	0.33
CD8_naive	0.9	0.67–1.21	0.49
Cytotoxic	0.92	0.69–1.23	0.58
Exhausted	0.92	0.68–1.23	0.55
Tr1	0.8	0.59–1.07	0.12
nTreg	1.18	0.88–1.58	0.26
iTreg	1	0.75–1.34	0.99
Th1	7.42	0.69–79.79	0.09
Th2	1.24	0.92–1.65	0.15
Th17	1.36	1.02–1.83	**0.03***
Tfh	0.67	0.5–0.9	< **0.01***
Central_memory	1.11	0.83–1.5	0.47
Effector_memory	1.16	0.71–1.89	0.54
NKT	0.85	0.63–1.14	0.27
MAIT	0.93	0.7–1.25	0.63
DC	0.94	0.7–1.26	0.66
B_cell	0.6	0.45–0.82	< **0.01***
Monocyte	6.25	0.84–46.22	0.07
Macrophage	0.9	0.37–2.18	0.82
NK	1.04	0.78–1.39	0.78
Neutrophil	1.2	0.9–1.6	0.22
Gamma_delta	1	0.74–1.33	0.98
CD4_T	0.8	0.59–1.07	0.12
CD8_T	0.81	0.6–1.08	0.15
InfiltrationScore	0.82	0.62–1.1	0.19

Bold value indicates that the differences between groups were statistically significant.

suggested as one of the underlying mechanisms for immune related subgroup clustering of LUAD. The results of GSEA analysis in validation data set (GSE72094 and GSE68465) were also shown in **Supplementary Figure S5**.

DISCUSSION

As an emerging therapeutic approach for malignancies, tumor immunotherapy, particularly for ICB therapy, is increasingly proved to be effective for LUAD patients (Huang et al., 2020a; Kano et al., 2020). However, a remarkable improvement in overall response rate and prognosis for LUAD patients is still not achieved due to the inherent intertumoral and intratumoral heterogeneity and the development of acquired resistance to ICB therapy (Jin et al., 2020; Song et al., 2020). To address this issue, a promising biomarker which is capable of predicting therapeutic efficiency before treatment is needed to screen potential responsive subpopulation prior to treatment and monitor the therapeutic efficiency during the process of treatment (Meyers and Banerji, 2020). Tumor-immune relationship plays an important role in tumor development and tumor progression, and tumor immune microenvironment (TIM) is widely accepted as a significant factor influencing therapeutic efficiency of ICB therapy (Song et al., 2020; Wang et al., 2020). Specifically, tumor-infiltrating lymphocytes (TILs) score (Gascón et al., 2020; Hashemi et al., 2021) and PD-L1 status (Wu et al., 2020) were previously suggested as potential biomarkers to be applied in clinical practice for LUAD. However, its translation from bench to bedside is largely limited by the dependence on tissue sample availability and the non-standardization for evaluation of TIL score and PD-L1 expression. Though previous studies tried to use immunophenotypic subtype classification based on immune signature to address this issue (Song et al., 2020; Wang et al., 2020; Xu et al., 2020), a systematic and comprehensive analysis (Seo et al., 2018; Zhang et al., 2020) is still required to determine the correlation between immune landscape based on immune related subgroup clustering and therapeutic reactivity to ICB therapy, and the underlying mechanism is of necessity to be explored (Huang et al., 2020b; Giannone et al., 2020). Previous studies from Chen YS. et al. (Xu et al., 2020) and Chen KX. et al. (Seo et al., 2018) performed immune related subgroup classification by using computational algorithms, however, an elaborated immune landscape characterization for distinct immune related subgroups were inadequate. Even though results from Xing Y. et al. (Song et al., 2020) and Kim Y. et al. (Xu et al., 2020) suggested a potential implication of immune subtype classification for ICB immunotherapy in lung cancer, a comprehensive analysis of the potential response to ICB immunotherapy for lung cancer, such as TIDE algorithm, was actually lacked. In the present investigation, we focused on both the elucidation of different immune signatures and prediction of potential response to ICB therapy for lung adenocarcinoma based on immune related subgroup clustering by using K-means algorithm, a classical unsupervised learning algorithm of artificial intelligence. More importantly, we conducted GSEA analysis to explore metabolism associated mechanism potentially responsible for immune related subgroup clustering of LUAD, particularly emphasized on the glucometabolic mechanism to shed light on comprehensive treatment strategy involving ICB immunotherapy in combination with glucose metabolism intervention.

Three distinct immune related subgroups were classified for LUAD in the current study based on RNA-seq data set from TCGA database (n = 572) by using a K-means algorithm. Among the classification, subgroup 1 was characterized by higher levels of immune checkpoint associated genes and higher cell infiltration scores for immune associated effector cells, and tended to be more sensitive to ICB therapy and have a favorable prognosis. Whereas, subgroup 3 with lower levels of immune checkpoint associated genes but higher cell infiltration scores for immune associated suppressive cells was found to be less responsive to ICB therapy and have a poor prognosis.

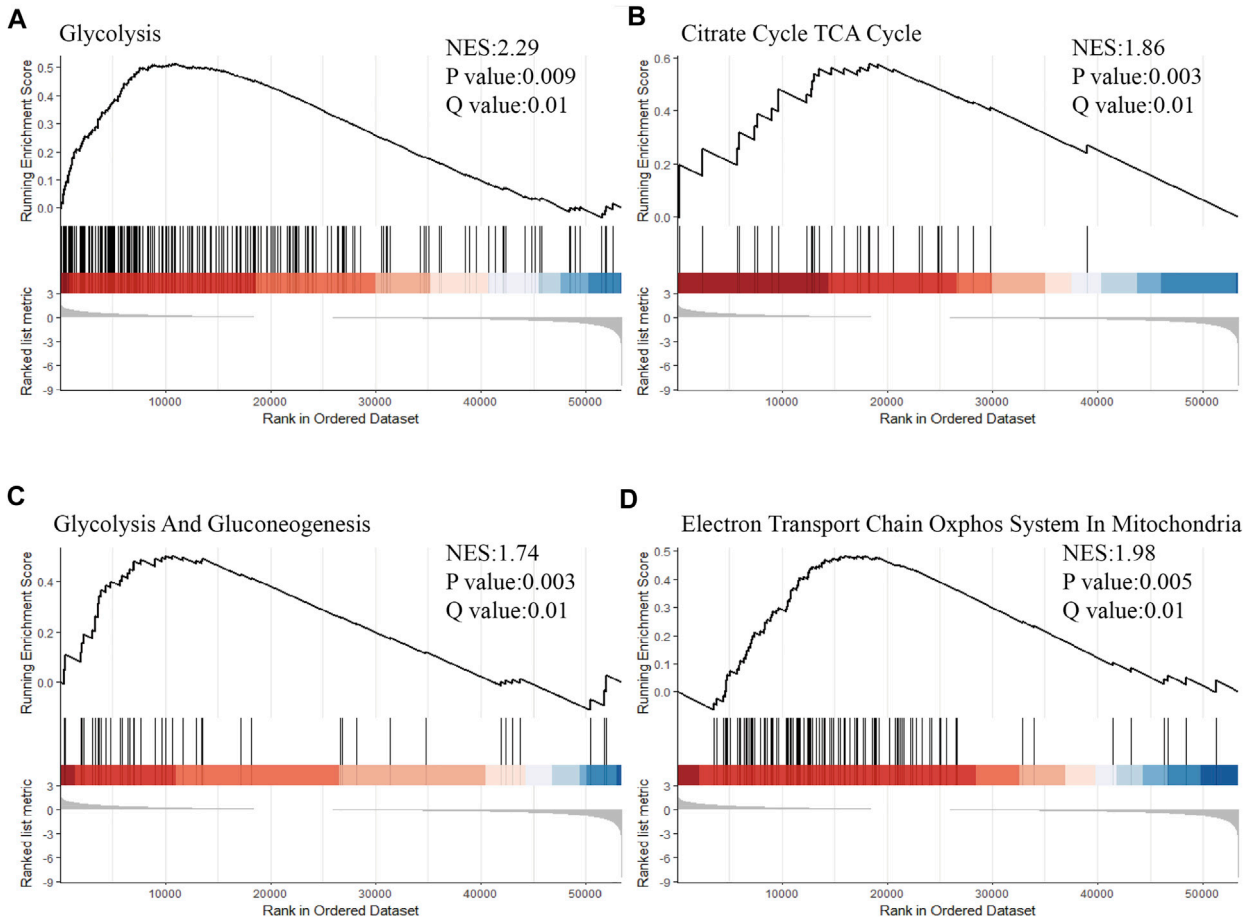

FIGURE 5 | Glucose metabolic reprogramming was suggested as one of the underlying mechanisms for immune related subgroup clustering of LUAD. Based on the immune related subgroup clustering (subgroup 1 vs subgroup 3), GSEA was performed by using the gene sets significantly associated with glucose metabolism, including processes of **(A)** glycolysis (Normalized Enrichment Score (NES) = 2.29, $p < 0.01$, $Q < 0.05$), **(B)** tricarboxylic acid (TCA) cycle (NES = 1.86, $p < 0.01$, $Q < 0.05$), **(C)** gluconeogenesis (NES = 1.74, $p < 0.01$, $Q < 0.05$) and **(D)** oxidative phosphorylation (OXPHOS) in mitochondria (NES = 1.98, $p < 0.01$, $Q < 0.05$) (NES = 1.98, $p < 0.01$, $Q < 0.05$). FDR <0.05 was set as the screening threshold. An upward parabola indicated that the indicated process was enhanced in subgroup 1 in contrast with subgroup 3. The barcode plot indicates the position of the genes in each gene set; red and blue colors represent positive and negative Pearson's correlation with subgroup classification.

Presumedly, subgroup 1 represented an immune-hot or with an immunocompetent TME with a higher infiltration score and an immunocompetent subtype which was possibly associated with a potential response to ICB therapy and a favorable prognosis. Whereas, subgroup 3 was considered as an immunodeficient or immunosuppressive landscape with a lower infiltration score or with an immunosuppressive subtype, suggesting a potential resistance to ICB therapy and an unfavorable prognosis. With respect to subgroup 2, a median subtype with a mixture of characteristics of subgroup 1 and subgroup 3 was considered. After Kaplan Meier analysis and Cox proportional hazards regression analysis, the immune related subgroup clustering was found to be an independent risk factor influencing the OS of LUAD patients. In the end, the GSEA analysis revealed that the metabolic reprogramming status in LUAD is potentially one of the underlying mechanisms for the distinct immune associated signatures based on the immune related subgroup clustering (Hensley et al., 2016; Faubert et al., 2017; Smolle et al., 2020). The enhanced glucose metabolism in subgroup 1 was consistent with the immune-hot landscape and a relatively immunocompetent subtype, whereas the decreased glucose metabolism in subgroup 3 suggested an immunodeficient landscape and/or an immunosuppressive subtype. Validation LUAD cohorts from external GEO database were also used to confirm the aforementioned results. To sum up, the present investigation provided a deep understanding of the interaction between tumor cells and surrounding immune cells (Kareva and Hahnfeldt, 2013; Speiser et al., 2016) and shed light on an improvement in ICB therapy or derived combination treatment for LUAD involving ICB therapy and metabolism intervention treatment.

As we know that, metabolic reprogramming and immunomodulation are two hallmarks of tumor (Hanahan and Weinberg, 2011). From a metabolic perspective, both tumorigenesis and immunoregulation are intricately associated with metabolic reprogramming. Specifically, the metabolic interplay between tumor cells and infiltrating immune cells significantly contributes to tumor

progression and tumor immunosuppression. As reported previously, metabolic competition between tumor cells and surrounding immune cells (Chang et al., 2015) and an accumulation of a variety of metabolite caused by metabolic reprogramming (Feng et al., 2017) in TME are partially responsible for immune landscape remodeling. Even though improvement in ICB therapy for LUAD in recent decades, a potential marker for effective stratification of LUAD patients before treatment and a promising target for associated molecular targeted therapy in combination with ICB therapy are expected to bring out breakthrough to clinical management for LUAD. The heterogeneity in metabolism status of LUAD was previously described (Hensley et al., 2016) and further confirmed by metabonomics analysis by investigation from others (Lazarou et al., 2020; Zhao et al., 2020). Additionally, multi-omics analysis based on single cell sequencing data also recovered a close correlation between immune status and metabolic reprogramming (Kim et al., 2020; Xiao et al., 2020; Zhong et al., 2021). Therefore, ICB therapy combined with metabolism intervention is expected to improve the prospect of LUAD treatment.

In spite of the innovation and valuable results mentioned above with respect to this study, a few limitations existing in the current investigation is noteworthy. First, the TCGA database mainly comprises Caucasian population, while validation cohort from GEO database mostly consists of Asian patients, thus racial bias was not inevitable in this study. To attenuate this bias, two external validation cohorts from GEO database were used to validate the results. Then, as actual sensitivity to ICB therapy for LUAD was not available in this study because the clinical information regarding to ICB therapy was mostly not provided in TCGA and GEO databases, only potential reactivity to ICB therapy for LUAD was evaluated based on TIDE analysis. In the end, the correlation between immune associated signature and sensitivity to ICB therapy and underlying metabolic reprogramming-associated mechanism were not further validated by basic research *in vitro* and clinical investigation *in vivo*, which is what we aim to do in future.

CONCLUSION

In the current investigation, a novel immune related subgroup clustering by an unsupervised learning model was identified for LUAD. Distinct immune associated landscape based on this clustering was significantly correlated with potential sensitivity to ICB therapy and prognosis for LUAD. GSEA analysis revealed that the heterogeneity in metabolic reprogramming is potentially one of the underlying mechanisms responsible for the correlation between immune landscape and potential reactivity to ICB therapy for LUAD. The immune related subgroup clustering based on the transcriptomics analysis will enable us to screen potentially responsive LUAD patients to ICB therapy. Additionally, metabolism intervention is a promising approach to improve the therapeutic efficiency of ICB therapy for LUAD.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

XY and ZW contributed equally to this work. XY, ZW, and XL conceived and designed the study, conducted statistical analysis, and wrote the original draft, WC, LZ, GY, YC, and JL performed the investigation and data interpretation; WX and XL reviewed and revised the manuscript. All authors read and approved the final version of the manuscript for publication.

FUNDING

This study was supported by the Science and Technology Development Fund of Tianjin Education Commission for Higher Education (2018KJ061).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.771830/full#supplementary-material

REFERENCES

Anichini, A., Perotti, V. E., Sgambelluri, F., and Mortarini, R. (2020). Immune Escape Mechanisms in Non Small Cell Lung Cancer. *Cancers* 12 (12), 3605. doi:10.3390/cancers12123605

Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., et al. (2013). InnateDB: Systems Biology of Innate Immunity and Beyond-Rrecent Updates and Continuing Curation. *Nucleic Acids Res.* 41 (Database issue), D1228–D1233. doi:10.1093/nar/gks1147

Chang, C.-H., Qiu, J., O'Sullivan, D., Buck, M. D., Noguchi, T., Curtis, J. D., et al. (2015). Metabolic Competition in the Tumor Microenvironment Is a Driver of Cancer Progression. *Cell* 162 (6), 1229–1241. doi:10.1016/j.cell.2015.08.016

Cline, M. S., Craft, B., Swatloski, T., Goldman, M., Ma, S., Haussler, D., et al. (2013). Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser. *Sci. Rep.* 3, 2652. doi:10.1038/srep02652

Davis, S., and Meltzer, P. S. (2007). GEOquery: a Bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23 (14), 1846–1847. doi:10.1093/bioinformatics/btm254

Faubert, B., Li, K. Y., Cai, L., Hensley, C. T., Kim, J., Zacharias, L. G., et al. (2017). Lactate Metabolism in Human Lung Tumors. *Cell* 171 (2), 358–371. doi:10.1016/j.cell.2017.09.019

Feng, J., Yang, H., Zhang, Y., Wei, H., Zhu, Z., Zhu, B., et al. (2017). Tumor Cell-Derived Lactate Induces TAZ-dependent Upregulation of PD-L1 through GPR81 in Human Lung Cancer Cells. *Oncogene* 36 (42), 5829–5839. doi:10.1038/onc.2017.188

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., et al. (2019). GENCODE Reference Annotation for the Human and Mouse Genomes. *Nucleic Acids Res.* 47 (D1), D766–D773. doi:10.1093/nar/gky955

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33 (1), 1–22. doi:10.18637/jss.v033.i01

Fu, J., Li, K., Zhang, W., Wan, C., Zhang, J., Jiang, P., et al. (2020). Large-scale Public Data Reuse to Model Immunotherapy Response and Resistance. *Genome Med.* 12 (1), 21. doi:10.1186/s13073-020-0721-z

Gascón, M., Isla, D., Cruellas, M., Gálvez, E. M., Lastra, R., Ocáriz, M., et al. (2020). Intratumoral versus Circulating Lymphoid Cells as Predictive Biomarkers in Lung Cancer Patients Treated with Immune Checkpoint Inhibitors: Is the Easiest Path the Best One? *Cells* 9 (6), 1525. doi:10.3390/cells9061525

Ge, L., and Shi, R. (2015). Progress of EGFR-TKI and ALK/ROS1 Inhibitors in Advanced Non-small Cell Lung Cancer. *Int. J. Clin. Exp. Med.* 8 (7), 10330–10339.

Giannone, G., Ghisoni, E., Genta, S., Scotto, G., Tuninetti, V., Turinetto, M., et al. (2020). Immuno-Metabolism and Microenvironment in Cancer: Key Players for Immunotherapy. *Int. J. Mol. Sci.* 21 (12), 4414. doi:10.3390/ijms21124414

Gillette, M. A., Satpathy, S., Cao, S., Dhanasekaran, S. M., Vasaikar, S. V., Krug, K., et al. (2020). Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell* 182 (1), 200–225.e35. doi:10.1016/j.cell.2020.06.013

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: the Next Generation. *Cell* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013

Hashemi, S., Fransen, M. F., Niemeijer, A., Ben Taleb, N., Houda, I., Veltman, J., et al. (2021). Surprising Impact of Stromal TIL's on Immunotherapy Efficacy in a Real-World Lung Cancer Study. *Lung Cancer* 153, 81–89. doi:10.1016/j.lungcan.2021.01.013

Hensley, C. T., Faubert, B., Yuan, Q., Lev-Cohain, N., Jin, E., Kim, J., et al. (2016). Metabolic Heterogeneity in Human Lung Tumors. *Cell* 164 (4), 681–694. doi:10.1016/j.cell.2015.12.034

Huang, J., Li, J., Zheng, S., Lu, Z., Che, Y., Mao, S., et al. (2020a). Tumor Microenvironment Characterization Identifies Two Lung Adenocarcinoma Subtypes with Specific Immune and Metabolic State. *Cancer Sci.* 111 (6), 1876–1886. doi:10.1111/cas.14390

Huang, Z., Su, W., Lu, T., Wang, Y., Dong, Y., Qin, Y., et al. (2020b). First-Line Immune-Checkpoint Inhibitors in Non-small Cell Lung Cancer: Current Landscape and Future Progress. *Front. Pharmacol.* 11, 578091. doi:10.3389/fphar.2020.578091

Jin, R., Liu, C., Zheng, S., Wang, X., Feng, X., Li, H., et al. (2020). Molecular Heterogeneity of Anti-PD-1/pd-L1 Immunotherapy Efficacy Is Correlated with Tumor Immune Microenvironment in East Asian Patients with Non-small Cell Lung Cancer. *Cancer Biol. Med.* 17 (3), 768–781. doi:10.20892/j.issn.2095-3941.2020.0121

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* 8 (1), 118–127. doi:10.1093/biostatistics/kxj037

Kano, H., Ichihara, E., Harada, D., Inoue, K., Kayatani, H., Hosokawa, S., et al. (2020). Utility of Immune Checkpoint Inhibitors in Non-small-cell Lung Cancer Patients with Poor Performance Status. *Cancer Sci.* 111 (10), 3739–3746. doi:10.1111/cas.14590

Kareva, I., and Hahnfeldt, P. (2013). The Emerging "hallmarks" of Metabolic Reprogramming and Immune Evasion: Distinct or Linked? *Cancer Res.* 73 (9), 2737–2742. doi:10.1158/0008-5472.can-12-3696

Kim, N., Kim, H. K., Lee, K., Hong, Y., Cho, J. H., Choi, J. W., et al. (2020). Single-cell RNA Sequencing Demonstrates the Molecular and Cellular Reprogramming of Metastatic Lung Adenocarcinoma. *Nat. Commun.* 11 (1), 2285. doi:10.1038/s41467-020-16164-1

Lazarou, G., Chelliah, V., Small, B. G., Walker, M., Graaf, P. H., and Kierzek, A. M. (2020). Integration of Omics Data Sources to Inform Mechanistic Modeling of Immune-Oncology Therapies: A Tutorial for Clinical Pharmacologists. *Clin. Pharmacol. Ther.* 107 (4), 858–870. doi:10.1002/cpt.1786

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The Sva Package for Removing Batch Effects and Other Unwanted Variation in High-Throughput Experiments. *Bioinformatics* 28 (6), 882–883. doi:10.1093/bioinformatics/bts034

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data. *Nat. Rev. Genet.* 11 (10), 733–739. doi:10.1038/nrg2825

Meyers, D. E., and Banerji, S. (2020). Biomarkers of Immune Checkpoint Inhibitor Efficacy in Cancer. *Curr. Oncol.* 27 (Suppl. 2), S106–S114. doi:10.3747/co.27.5549

Miao, Y. R., Zhang, Q., Lei, Q., Luo, M., Xie, G. Y., Wang, H., et al. (2020). ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy. *Adv. Sci.* 7 (7), 1902880. doi:10.1002/advs.201902880

Park, C., Na, K. J., Choi, H., Ock, C.-Y., Ha, S., Kim, M., et al. (2020). Tumor Immune Profiles Noninvasively Estimated by FDG PET with Deep Learning Correlate with Immunotherapy Response in Lung Adenocarcinoma. *Theranostics* 10 (23), 10838–10848. doi:10.7150/thno.50283

Pathak, R., Pharaon, R. R., Mohanty, A., Villaflor, V. M., Salgia, R., and Massarelli, E. (2020). Acquired Resistance to PD-1/pd-L1 Blockade in Lung Cancer: Mechanisms and Patterns of Failure. *Cancers* 12 (12), 3851. doi:10.3390/cancers12123851

Seo, J.-S., Kim, A., Shin, J.-Y., and Kim, Y. T. (2018). Comprehensive Analysis of the Tumor Immune Micro-environment in Non-small Cell Lung Cancer for Efficacy of Checkpoint Inhibitor. *Sci. Rep.* 8 (1), 14576. doi:10.1038/s41598-018-32855-8

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

Smolle, E., Leko, P., Stacher-Priehse, E., Brcic, L., El-Heliebi, A., Hofmann, L., et al. (2020). Distribution and Prognostic Significance of Gluconeogenesis and Glycolysis in Lung Cancer. *Mol. Oncol.* 14 (11), 2853–2867. doi:10.1002/1878-0261.12780

Song, Y., Yan, S., Fan, W., Zhang, M., Liu, W., Lu, H., et al. (2020). Identification and Validation of the Immune Subtypes of Lung Adenocarcinoma: Implications for Immunotherapy. *Front. Cel Dev. Biol.* 8, 550. doi:10.3389/fcell.2020.00550

Speiser, D. E., Ho, P.-C., and Verdeil, G. (2016). Regulatory Circuits of T Cell Function in Cancer. *Nat. Rev. Immunol.* 16 (10), 599–611. doi:10.1038/nri.2016.80

Sun, S., Guo, W., Wang, Z., Wang, X., Zhang, G., Zhang, H., et al. (2020). Development and Validation of an Immune-related Prognostic Signature in Lung Adenocarcinoma. *Cancer Med.* 9 (16), 5960–5975. doi:10.1002/cam4.3240

Wang, Q., Li, M., Yang, M., Yang, Y., Song, F., Zhang, W., et al. (2020). Analysis of Immune-Related Signatures of Lung Adenocarcinoma Identified Two Distinct Subtypes: Implications for Immune Checkpoint Blockade Therapy. *Aging* 12 (4), 3312–3339. doi:10.18632/aging.102814

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: a Class Discovery Tool with Confidence Assessments and Item Tracking. *Bioinformatics* 26 (12), 1572–1573. doi:10.1093/bioinformatics/btq170

Wu, Y., Lin, L., and Liu, X. (2020). Identification of PDL1-Related Biomarkers to Select Lung Adenocarcinoma Patients for PD1/PDL1 Inhibitors. *Dis. Markers* 2020, 7291586. doi:10.1155/2020/7291586

Xiao, Z., Locasale, J. W., and Dai, Z. (2020). Metabolism in the Tumor Microenvironment: Insights from Single-Cell Analysis. *Oncoimmunology* 9 (1), 1726556. doi:10.1080/2162402x.2020.1726556

Xu, F., Chen, J.-x., Yang, X.-b., Hong, X.-b., Li, Z.-x., Lin, L., et al. (2020). Analysis of Lung Adenocarcinoma Subtypes Based on Immune Signatures Identifies Clinical Implications for Cancer Therapy. *Mol. Ther. - Oncolytics* 17, 241–249. doi:10.1016/j.omto.2020.03.021

Zhang, Y., Yang, M., Ng, D. M., Haleem, M., Yi, T., Hu, S., et al. (2020). Multi-omics Data Analyses Construct TME and Identify the Immune-Related Prognosis Signatures in Human LUAD. *Mol. Ther. - Nucleic Acids* 21, 860–873. doi:10.1016/j.omtn.2020.07.024

Zhao, C., Kong, X., Han, S., Li, X., Wu, T., Zhou, J., et al. (2020). Analysis of Differential Metabolites in Lung Cancer Patients Based on Metabolomics and Bioinformatics. *Future Oncol.* 16 (18), 1269–1287. doi:10.2217/fon-2019-0818

Zhong, R., Chen, D., Cao, S., Li, J., Han, B., and Zhong, H. (2021). Immune Cell Infiltration Features and Related Marker Genes in Lung Cancer Based on Single-Cell RNA-Seq. *Clin. Transl Oncol.* 23 (2), 405–417. doi:10.1007/s12094-020-02435-2

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.

GLOSSARY

ANN artificial neural network

ALK anaplastic lymphoma kinase

CTLA-4 cytotoxic T-lymphocyte antigen 4

CLP common lymphoid progenitor

CAF cancer-associated fibroblast

DC dendritic cell

DBSCAN density-based spatial clustering of applications with noise

EGFR epidermal growth factor receptor

FDR false rediscovery rate

GEO Gene Expression Omnibus

GSEA single-sample gene set enrichment analysis

HR hazard ratio

ICBs immune checkpoint blockades

IFN-γ interferon-γ

ImmuCellAI Immune cell abundance identifier

LASSO least absolute shrinkage and selection operator

LUAD lung adenocarcinoma

MAIT mucosal-associated invariant T

MDSC myeloid-derived suppressor cell

NK natural killer

NKT natural killer T

NSCLC non-small cell lung cancer (NSCLC)

OS Overall survival

OXPHOS oxidative phosphorylation

PD-1/PD-L1 programmed cell death protein 1/programmed cell death ligand 1

PCA principal components analysis

RNA-seq RNA sequencing

SVM support vector machine

TCA tricarboxylic acid

Tc cytotoxic T

TCGA the Cancer Genome Atlas

Tex exhausted T cells

TIDE tumor immune dysfunction and exclusion

TIL tumor-infiltrating lymphocyte

TIM tumor immune microenvironment

TKIs tyrosine kinase inhibitors

TME tumor microenvironment

Treg Regulatory T

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 October 2021
doi: 10.3389/fgene.2021.744557

Interaction-Based Feature Selection Algorithm Outperforms Polygenic Risk Score in Predicting Parkinson's Disease Status

*Justin L. Cope[1], Hannes A. Baukmann[1], Jörn E. Klinger[1], Charles N. J. Ravarani[1], Erwin P. Böttinger[2], Stefan Konigorski[2] and Marco F. Schmidt[1]**

[1]biotx.ai GmbH, Potsdam, Germany, [2]Digital Health Center, Hasso Plattner Institute for Digital Engineering, University of Potsdam, Potsdam, Germany

Polygenic risk scores (PRS) aggregating results from genome-wide association studies are the state of the art in the prediction of susceptibility to complex traits or diseases, yet their predictive performance is limited for various reasons, not least of which is their failure to incorporate the effects of gene-gene interactions. Novel machine learning algorithms that use large amounts of data promise to find gene-gene interactions in order to build models with better predictive performance than PRS. Here, we present a data preprocessing step by using data-mining of contextual information to reduce the number of features, enabling machine learning algorithms to identify gene-gene interactions. We applied our approach to the Parkinson's Progression Markers Initiative (PPMI) dataset, an observational clinical study of 471 genotyped subjects (368 cases and 152 controls). With an AUC of 0.85 (95% CI = [0.72; 0.96]), the interaction-based prediction model outperforms the PRS (AUC of 0.58 (95% CI = [0.42; 0.81])). Furthermore, feature importance analysis of the model provided insights into the mechanism of Parkinson's disease. For instance, the model revealed an interaction of previously described drug target candidate genes *TMEM175* and *GAPDHP25*. These results demonstrate that interaction-based machine learning models can improve genetic prediction models and might provide an answer to the missing heritability problem.

Keywords: epistasis, machine learning, feature selection, parkinson's disease, PPMI (parkinson's progression markers initiative)

OPEN ACCESS

Edited by:
Jiangning Song,
Monash University, Australia

Reviewed by:
Jiancheng Zhong,
Hunan Normal University, China
Zhen Chen,
Qingdao University, China

****Correspondence:***
Marco F. Schmidt
ms@biotx.ai

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

Received: *20 July 2021*
Accepted: *27 September 2021*
Published: *20 October 2021*

Citation:
Cope JL, Baukmann HA, Klinger JE, Ravarani CNJ, Böttinger EP, Konigorski S and Schmidt MF (2021) Interaction-Based Feature Selection Algorithm Outperforms Polygenic Risk Score in Predicting Parkinson's Disease Status. Front. Genet. 12:744557. doi: 10.3389/fgene.2021.744557

INTRODUCTION

The need to understand how to predict phenotypes from genetic data is becoming ever-more important for the prediction of disease risk for individuals and for plant and animal breeding as well as for genome editing. Polygenic risk scores (PRS), simple additive models, are the state of the art in the investigation of the genetic architecture of complex traits or diseases, and, more importantly, in the prediction of disease susceptibility. (Wray et al., 2007; Evans et al., 2009; International Schizophrenia Consortium et al., 2009). A Polygenic Risk Score is calculated for a given individual as the weighted sum of the number of risk allele single nucleotide polymorphisms (SNP) for which the individual was tested. The weights used in this calculation are the regression coefficients from a prior genome-wide association study (GWAS).

Importantly, PRS models are not optimized for predictive performance. (Chatterjee et al., 2013; Dudbridge, 2013). There are three reasons for this:

(1) Due to the current limited sample size of discovery GWAS datasets (<1,000,000 individuals), biologically relevant rare variants with small effect sizes cannot be detected. Additionally, the limited sample sizes of discovery GWAS can lead to biased PRS models that might not perform well in populations with ancestry different to that of the discovery dataset. (Reisberg et al., 2017; Duncan et al., 2019).
(2) It has been shown that statistically significant outcome-associated SNPs are not automatically good predictors of that outcome. (Lo et al., 2015).
(3) It has been reported that genetic effects discovered in genome-wide association studies do not sum to the estimate of the heritability of the trait derived from twin studies. (Yang et al., 2010). This has been called the *missing heritability problem* in GWAS. (Manolio et al., 2009). Besides potentially missing relevant rare variants and suboptimal SNP selection based on *p*-values, classical PRS models ignore complex gene-gene interactions, also known as *epistasis*, of the trait or disease due to their simple additive structure.

The concept of epistasis was first described more than 100 years ago. (Bateson, 1906). Statistical epistasis, as observed in genome-wide association studies, is genetic variance that can be attributed to gene interaction and is defined as a function of the allele frequencies in a population. Detection of epistasis in discovery GWAS and modeling its impact is challenging because of linkage disequilibrium (LD), replication of identified gene-gene interactions in validation datasets, model complexity, and high dimensionality. (Wei et al., 2014).

Machine learning algorithms that improve automatically through the use of data represent an opportunity to find gene-gene interactions in order to build models with better predictive performance than PRS. Nevertheless, in a recent study, a PRS model outperformed five machine learning algorithms (Naïve Bayes classifier, regularized regression, random forest, gradient boosting, and support vector machine) that were used to build predictive models for coronary artery disease status. (Gola et al., 2020).

Here we revisit the potential of machine learning algorithms to predict disease status compared to a PRS model. For this purpose, we adopt the Parkinson's Progression Markers Initiative (PPMI) dataset (Marek et al., 2011, 2018) (https://www.ppmi-info.org) as this dataset has been intensively analyzed and is broadly available for replication studies. We explore two machine learning approaches in particular, which complement those applied by Gola et al.: deep learning and interaction-based feature selection. The first approach, deep learning, employs artificial neural networks to discover automatically from raw data the representations needed for classification. Despite not being widely used in the field of genomics, there is work on applying deep learning to GWAS: Romero et al., 2016 predict genetic ancestry by introducing a multi-task architecture including a parameter prediction network, thereby considerably reducing the feature space under consideration. The second approach, interaction-based feature selection, also drastically reduces the feature space–in this case, by leveraging contextual information obtained via data mining, allowing for the testing of a small set of complex hypotheses containing interactions of multiple variants. Further details concerning these approaches are described in the Methods section, following a presentation of the results of our investigation below.

RESULTS

Data Preparation

For all 471 subjects in the PPMI database (368 cases and 152 controls) subject genotyping information was collected from two complementary genotyping chips (NeuroX and ImmunoChip). After careful quality control and harmonization, we merged that information into a single dataset with 369,036 variants and 436 individuals (296 cases and 140 controls). The data was then split into three disjoint sets: a training set (n = 367) for training predictive models; a validation set (n = 33) for so-called *hyperparameter tuning*, and a test set (n = 36) for model evaluation. Training and validation are described in further detail below for each approach as appropriate. In all cases, evaluation metrics were calculated on the basis of bootstrap resampling with 10^4 iterations.

Genome-Wide Association Study

A genome-wide association (GWA) analysis was performed on the training data. The Manhattan plot of the *p*-values resulting from the analysis is shown in **Figure 1**. Seven single nucleotide polymorphisms (SNPs) showed *p*-values less than 10^{-4} (**Table 1**).

Polygenic Risk Score

To calculate the PRS, seven different *p*-value thresholds (0.001, 0.05, 0.1, 0.2, 0.4, and 0.5) for the subjects in the training, validation and test set were used. The PRS of the subjects in the training set were then used to train a separate logistic regression classifier for each *p*-value threshold. Receiver operating characteristics (ROC) curves were used to evaluate classifier performance relative to the validation data. The classifier with the highest mean area under the curve (AUC) was that which had been trained on the PRS resulting from the 0.05 *p*-value threshold, comprising the weighted sum of 57 different SNPs. This classifier was finally evaluated relative to the test data set, where the mean AUC was 0.58 with a 95% confidence interval from 0.42 to 0.81 **Figure 3**. **Table 2** presents these results, along with the estimates of accuracy, sensitivity, and specificity corresponding to the optimal Youden's index of 0.21.

Deep Learning

We applied Romero *et al.*'s approach to the PPMI dataset, again using the training data set to train competing networks with distinct hyperparameter settings and the validation data set to select between these networks. When evaluated relative to the test data set, the mean AUC of the final deep learning model was 0.67 (95% CI = [0.47; 0.83]) and the optimal Youden index

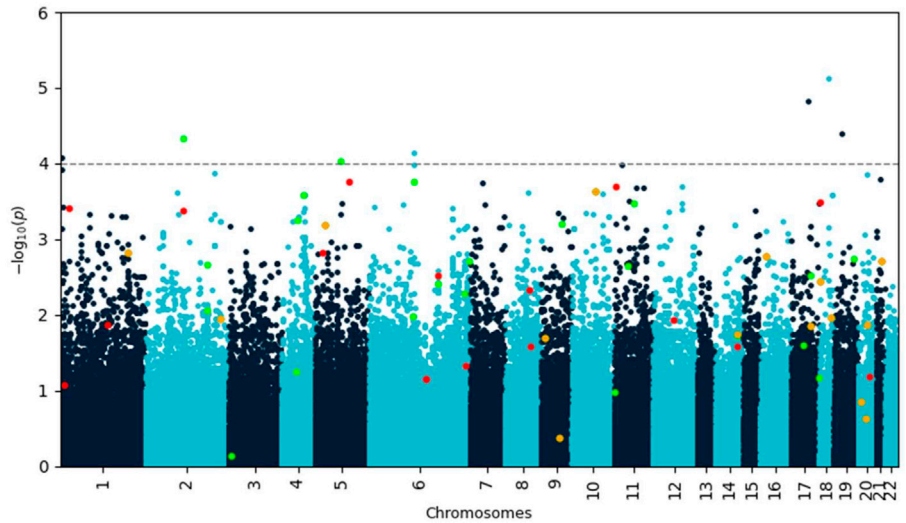

FIGURE 1 | Manhattan plot of negative decadic logarithm of *p*-values for SNPs as determined by SAIGE analysis. Variants identified by Lasso with feature selection are highlighted in red and green if they increase or decrease disease risk, respectively. Variants highlighted in orange occur in both protective and risk-enhancing groups of SNPs, depending on their genotype. Most of these biologically meaningful variants would have been missed by using a simple *p*-value cutoff.

TABLE 1 | PPMI GWAS results identified seven SNPs with a *p*-value < 10^{-4}. Positions and rs IDs according to Human Genome Reference hg19 (GRCh37).

Chr	Pos	SNP Id	rs Id	Gene	*p*-value
1	173,266,578	imm_1_171,533,201	rs4916319	*TNFSF4* (upstream)	0.000083
2	209,087,335	exm2261159	rs4675743		0.000046
5	156,376,703	exm498917	rs6873053	*TIMD4* (downstream)	0.000092
6	133,716,974	rs212805	rs212805	*EYA4*	0.000074
17	25,895,033	imm_17_22,919,160	rs4795747		0.000015
18	5,479,093	rs7238186	rs7238186	*EPB41L3* (downstream)	0.000007
19	57,909,872	exm1513284	rs4801478	*ZNF548*	0.000040

TABLE 2 | Performance comparison of all models.

Method	AUC [95% CI]	Accuracy	Sensitivity	Specificity	Youden's index
PRS	0.56 [0.42; 0.81]	0.60	0.62	0.56	0.21
Deep learning	0.67 [0.47; 0.83]	0.60	0.42	0.88	0.29
LASSO w/feature selection	0.85 [0.72; 0.96]	0.81	0.81	0.80	0.61
LASSO w/o feature selection	0.51 [0.39; 0.63]	0.62	0.87	0.09	0.12

corresponding to the accuracy, sensitivity, and specificity measures reported in **Table 2** was 0.29.

Feature Selection and LASSO Regression

A set of less than 100 polygenic hypotheses were generated using the interaction-based feature selection approach applied to the training data, as described in the Methods section below. (See also an overview of our approach in **Figure 3**.). These hypotheses were summarized in a term that was used to build a LASSO regression model on the basis of the validation data. (Tibshirani, 1996). The predictive performance of this model, based on 47 SNPs in several different interaction terms, was then evaluated relative to the test set **Figure 4**. The mean area under the curve (AUC) for the LASSO model with prior feature selection was 0.85 [95% CI = (0.72; 0.96)] and the optimal Youden index corresponding to the accuracy, sensitivity, and specificity measures reported in **Table 2** was 0.61. A LASSO model without prior feature selection that was built for comparison was evaluated in the same manner but did not deliver outcomes that were significantly better than chance (**Table 2**), in line with Gola et al. (2020) s results for regularized regression.

Exploring the feature selection based model with its interactive terms provides insights about the genes associated with Parkinson's disease. An annotation of all 47 SNPs in our

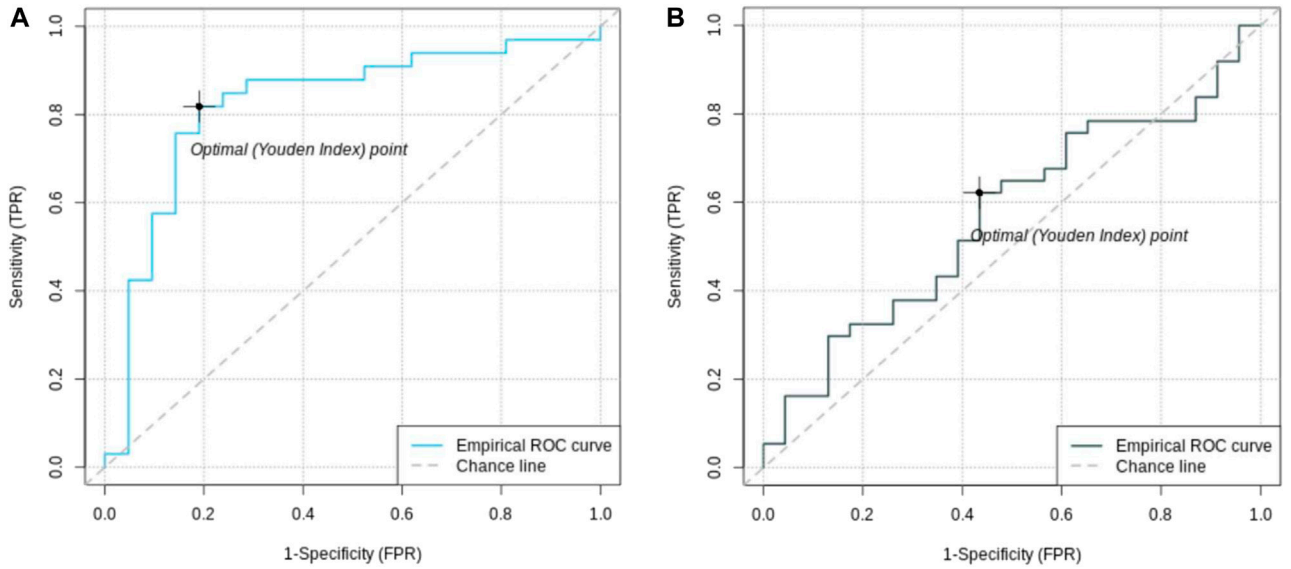

FIGURE 2 | Receiver operating characteristic (ROC) curves of feature selected machine learning model **(A)** and polygenic risk score **(B)**. The AUC of the feature selected model with 0.85 [95% CI = (0.72; 0.96)] is better than the AUC of the PRS with 0.56 [95% CI = (0.42; 0.81)].

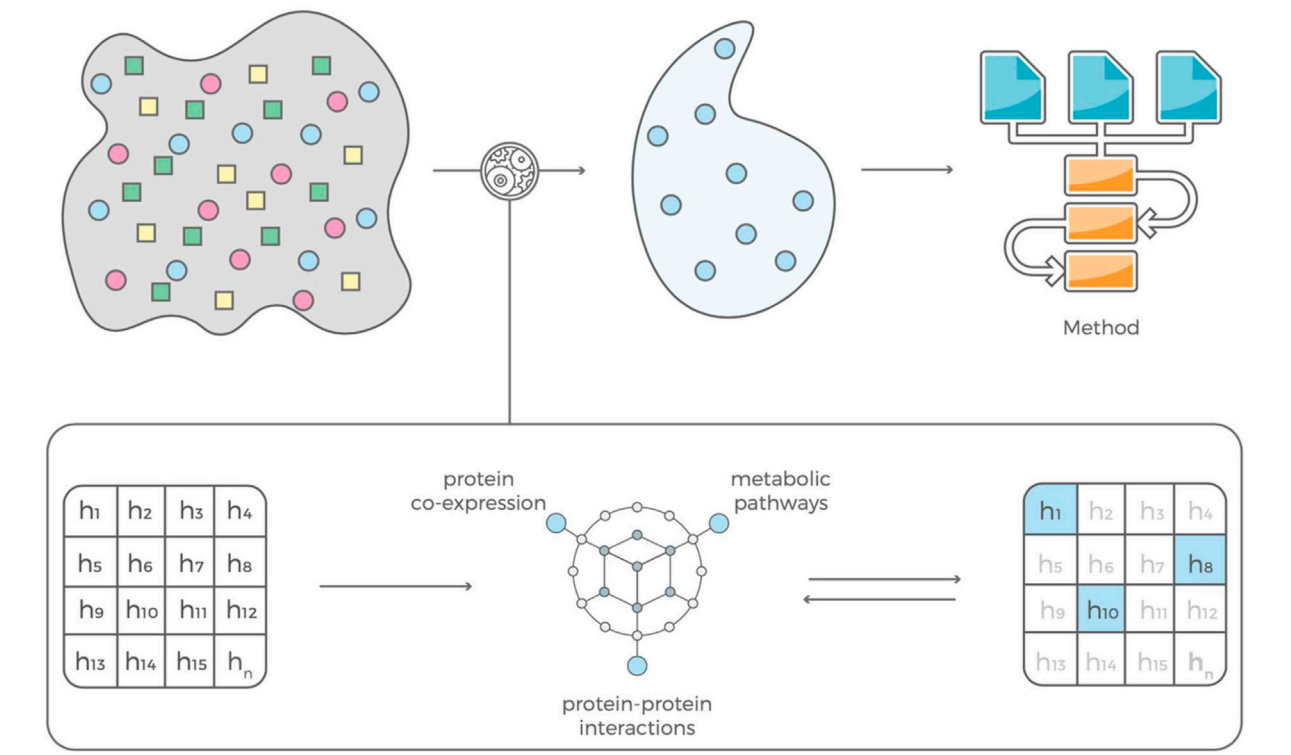

FIGURE 3 | Our feature selection consists of two complementary modules that are in feedback with each other. The contextual module uses information mined from the scientific literature, pathway libraries and protein co-expression data and an evaluation module that estimates predictive power of a feature based on that contextual information. The selected features can be used to build prediction models with standard machine learning algorithms.

model can be found in the Supplementary Information. An exciting result from this analysis of the PPMI dataset is the statistical interaction of variants rs3822019 on chromosome four in gene *TMEM175*, coding for a potassium channel in late endosomes, and rs17022,452 on chromosome 2, close to the coding region of *GAPDHP25*, glyceraldehyde-3 phosphate

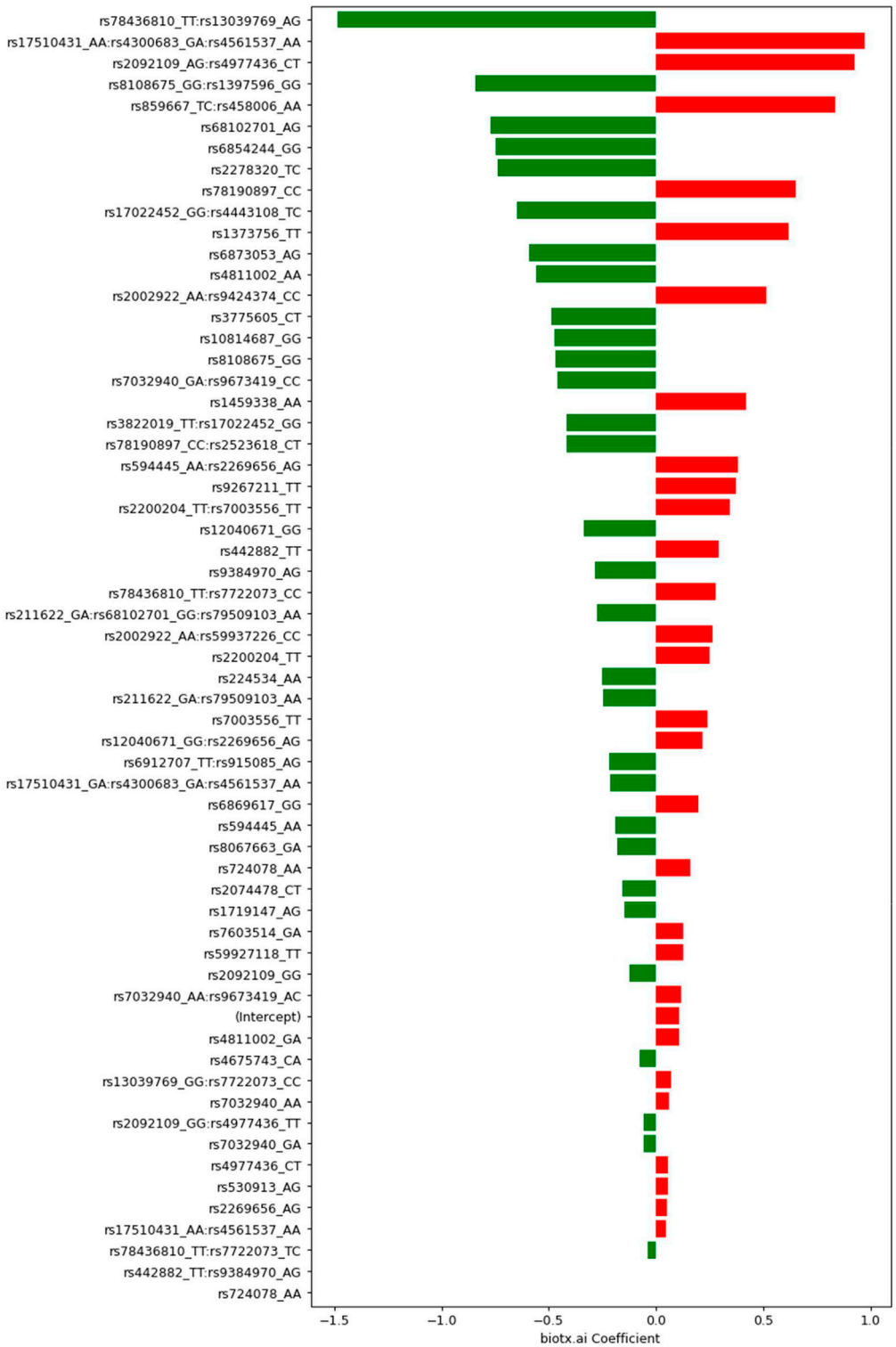

FIGURE 4 | Coefficients determined by Lasso with feature selection for SNPs and groups of SNPs. Negative values (green) indicate protective (combinations of) variants, positive values (red) mark risk variants. The respective genotypes of each variant are indicated by one-letter codes of the bases, where the first letter corresponds to the reference allele, and the second corresponds to the observed, alternative allele.

TABLE 3 | PD cases and controls among bearers of the respective genotype combinations of the identified variants rs3822019 and rs17022,452.

Genotype combination	Cases	Controls
rs3822019_TT/rs17022,452_GG	0	0
rs3822019_TT/rs17022,452_GA	6/100%	0
rs3822019_TC/rs17022,452_GG	2/50%	2/50%
rs3822019_TT/-	7/100%	0
-/rs17022,452_GG	7/87.5%	1/12.5%
rs3822019_TC/rs17022,452_GA	27/87.1%	4/12.9%
rs3822019_TC/-	68/73.9%	24/26.1%
-/rs17022,452_GA	66/75%	22/25%
-/-	113/56.5%	87/43.5%

dehydrogenase pseudogene 25. rs3822019 is an intron variant that has been linked to Parkinson's disease. (Nalls et al., 2014).

DISCUSSION

We analyzed the PPMI dataset and built predictive models using polygenic risk scores, a deep learning algorithm for genomic data (Romero et al., 2016), and LASSO regression with and without interaction-based feature selection to reduce the hypothesis space. The PRS model comprises 57 SNPs and showed an AUC of 0.58 whereas the deep learning model had an AUC of 0.67. Notably, the deep learning model consists of abstract embeddings instead of single SNPs like the PRS. Therefore, identification of disease-associated SNPs and further insights into the disease mechanism are not possible here. The LASSO regression model built on interactions containing only 47 SNPs that were discovered via the use of contextual information outperformed the other predictive models with an AUC of 0.85. Beyond that, the approach was able to associate new variants with the disease that would not have shown up under an additive approach such as PRS.

We investigated how the combinations of the relevant genotypes rs3822019_TT (*TMEM175*) and rs17022,452_GG (*GAPDHP25*) split the individuals into cases and controls (**Table 3**). All subjects that are homozygous for rs3822019_TT are affected by PD. Furthermore, most individuals heterozygous for this variant (rs3822019_TT) or homozygous for rs17022,452_GG are cases (76.4 and 75.0%, respectively). These results support the relevance of the association between these variants and PD status.

The *TMEM175/GAK/DGKQ* locus was the third strongest risk locus in a GWA study of Parkinson's disease (Krohn et al., 2020) and has been described as a potential drug target. (Diogo et al., 2018; Jinn et al., 2019). Deficiency in the potassium channel TMEM175 results in unstable lysosomal pH, which leads to decreased lysosomal catalytic activity and increased α-synuclein aggregation, among other effects. As a potassium channel, TMEM175 has a high potential as a druggable target and a tractable therapeutic strategy has been proposed. (Jinn et al., 2017).

GAPDH has been targeted with the investigational drug Omigapil for prevention of PD, ALS, congenital muscular dystrophy and myopathy. The drug has been shown to protect against behavioural abnormalities and neuro-degeneration in animal models of Parkinson's disease. However, PD development has been terminated due to lack of benefit. (Olanow et al., 2006).

There seem to be various causes of Parkinson's disease, yet the pathogenesis of this disease appears to be converging on common themes—oxidative stress, mitochondrial dysfunction, and protein aggregation—all of which are tightly linked to autophagy. (Lynch-Day et al., 2012). Both TMEM175 (Jinn et al., 2019) and GAPDH (Butera et al., 2019) regulate autophagy. Disturbed expression of autophagy genes in blood of PD patients. (Lynch-Day et al., 2012).

To summarize, we here present an approach to apply machine learning algorithms to high-dimensional genomic data using a contextual knowledge based feature selection. PRS models require a large set of SNPs, which leads to overfitting and limits their use in clinical practice. We generated more parsimonious models overcoming these limitations–with only 47, partly interacting SNPs, our model was able to outperform a PRS model based on 57 SNPs for Parkinson's disease. Analysis of feature importance of our model identified a gene-gene interaction of *TMEM175* and *GAPDHP25*. TMEM175 has been described as a potential drug target and further information on its mechanism of action could be invaluable. A recently discovered interaction with pseudogene *GAPDHP25* could provide helpful insights. In conclusion, applying machine learning algorithms to feature-selected genomic data led to an interaction-based model with better predictive performance than PRS and has paved the way for the generation of new insights into disease mechanisms.

METHODS

Parkinson's Progression Marker Initiative Dataset

The Parkinson's Progression Marker Initiative (PPMI) dataset (https://www.ppmi-info.org) contains 471 subjects (368 cases and 152 controls), and for each subject, genotyping information collected from two complementary chips (NeuroX and ImmunoChip) is available. (Marek et al., 2011). After careful quality control and harmonization (e.g., genome build conversion, strand alignment) as described in the literature (Marees et al., 2018), we merged that information into a single dataset with 380,939 variants in total.

After this initial data harmonization, an additional set of quality control steps were performed on variants and individuals that aimed to remove biases that could affect the downstream analysis. First, SNPs and individuals were filtered based on their missingness in the dataset. This ensured the exclusion of SNPs that had a high proportion of subjects where genotyping information was unavailable or of poor quality. Similarly, individuals where a large proportion of SNPs could not be measured were excluded. This step was achieved by setting the missing call rate threshold to 0.02 (i.e., >2%); as a result, 6,084 variants and 22 people were removed. SNP filtering was performed before individual filtering.

With high missing call rates filtered, all variants not on autosomal chromosomes were removed (5,731 variants). This was followed by the identification and removal of variants deviating from Hardy-Weinberg equilibrium, which can indicate genotyping errors. These variants were identified in a two-stage process whereby we first applied a threshold of 1e-6 exclusively to controls, followed by a threshold of 1e-10 applied to all samples, leading to the removal of 0 and 202 variants, respectively.

Next, individuals were filtered based on their heterozygosity rates, which can indicate sample contamination. Individuals deviating by more than 3 standard deviations from the mean of the rate of all samples (13 individuals) were removed. To assess the heterozygosity rate per sample, variants in linkage disequilibrium were first extracted, scanning the genome at a window size of 50 variants, a step size of 5, and a pairwise correlation threshold of 0.2.

Finally, relatedness between individuals was ascertained through the calculation and assessment of their respective identity by descent coefficients (IBD). Only one individual in a related pair would be kept, although in this case, no related individuals were identified and so none were removed.

The final quality-controlled dataset contained 369,036 variants and 436 individuals passing the various filters.

GENOME-WIDE ASSOCIATION STUDY

As a preliminary step, a genome-wide association (GWA) analysis was performed with the R package SAIGE (Zhou et al., 2018) to test individual variants for their association with Parkinson's disease.

Polygenic Risk Score

The PRS was constructed by using PLINK (Purcell et al., 2007) following the guidelines provided by Choi et al. (Choi et al., 2020) and the accompanying tutorial (https://choishingwan.github.io/PRS-Tutorial/plink/). The clumping cut-off of r^2 was 0.1. For all subjects in the training, validation and test sets, seven distinct risk scores were calculated, corresponding to seven potential *p*-value thresholds (0.001, 0.05, 0.1, 0.2, 0.4, 0.5). The seven risk scores for the subjects in the training set were then used to train seven separate logistic regression classifiers (one for each *p*-value threshold) using the *glm* function in R (www.R-project.org). These classifiers were evaluated relative to the validation data set, leading to the selection of the classifier based on the PRS calculated using the *p*-value threshold of 0.05. The predictions of this final classifier were then evaluated relative to the test set.

Deep Learning

The deep learning prediction model was built using a Diet Network according to the procedure described by Romero et al. (Romero et al., 2016) The model is composed of three networks: one basic and two auxiliary networks. After a basic discriminative network with optional reconstruction path, follows a network that predicts the input fat layer parameters, and finally, a network that predicts the reconstruction fat layer parameters. The official code can be found here: https://github.com/adri-romsor/DietNetworks.

Feature Selection

The interaction-based feature selection approach that we adopt organizes data mined from journal articles, pathway libraries, protein co-expression libraries, and drug candidate libraries (e.g., dbSNP, ClinVar, OMIM, Reactome, STRING database) into a hierarchical knowledge graph, which generates disease-specific hypotheses based on interactions of genetic variants (**Figure 1**). Each interaction's predictive power is determined using the training data set and the *glm* function in R (www.R-project.org). If an interaction predicts disease status well, the graph is incentivized to 'fine-tune' the hypothesis by comparing a set of very similar hypotheses. If a hypothesis has little or no predictive power, the graph is not incentivized to explore it or similar hypotheses further and will instead propose hypotheses containing different variants. (Klinger et al., 2021). This learning process is driven by gradient descent, meaning that it converges when the average performance of the new multi-variant hypothesis does not increase. After convergence, the selected features are used to build prediction models with standard machine learning algorithms, such as LASSO regression (Friedman et al., 2010).

LASSO Regression

LASSO (least absolute shrinkage and selection operator) regression models were computed by using the glmnet package (https://glmnet.stanford.edu/index.html) for R (www.R-project.org) and its function cv.glmnet with five-fold cross-validation in order to avoid overfitting. (Friedman et al., 2010).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ppmi-info.org/.

AUTHOR CONTRIBUTIONS

JK, CR, EB, SK, and MS contributed to conception and design of the study. JC, HB, JK, and CR organized the database and performed the statistical analysis. MS wrote the first draft of the manuscript. JC, HB, JK, CR, and MS wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The research work was supported by the Investitionsbank des Landes Brandenburg (ILB), the European Regional Development Fund (ERDF), and the European Social Fund+ (ESF+).

ACKNOWLEDGMENTS

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI–a public-private partnership–is funded by The Michael J. Fox Foundation for Parkinson's Research and funding partners. List of full names of all the PPMI funding partners can be found at www.ppmi-info.org/fundingpartners. We also thank the program digital solutions made in Brandenburg (digisolBB) for its continued support.

REFERENCES

Bateson, W. (1906). The Progress of Genetics since the Rediscovery of Mendel's Papers. *Prog. Rei Bot.* 1, 368.

Butera, G., Mullappilly, N., Masetto, F., Palmieri, M., Scupoli, M. T., Pacchiana, R., et al. (2019). Regulation of Autophagy by Nuclear GAPDH and its Aggregates in Cancer and Neurodegenerative Disorders. *Int. J. Mol. Sci.* 20, 2062. doi:10.3390/ijms20092062

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J. H. (2013). Projecting the Performance of Risk Prediction Based on Polygenic Analyses of Genome-Wide Association Studies. *Nat. Genet.* 45 (4), 400–405. doi:10.1038/ng.2579

Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: A Guide to Performing Polygenic Risk Score Analyses. *Nat. Protoc.* 15, 2759–2772. doi:10.1038/s41596-020-0353-1

Diogo, D., Tian, C., Franklin, C. S., Alanne-Kinnunen, M., March, M., Spencer, C. C. A., et al. (2018). Phenome-Wide Association Studies across Large Population Cohorts Support Drug Target Validation. *Nat. Commun.* 9, 4285. doi:10.1038/s41467-018-06540-3

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *Plos Genet.* 9, e1003348. doi:10.1371/journal.pgen.1003348

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., et al. (2019). Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations. *Nat. Commun.* 10, 3328. doi:10.1038/s41467-019-11112-0

Evans, D. M., Visscher, P. M., and Wray, N. R. (2009). Harnessing the Information Contained within Genome-wide Association Studies to Improve Individual Prediction of Complex Disease Risk. *Hum. Mol. Genet.* 18, 3525–3531. doi:10.1093/hmg/ddp295

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

Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., and König, I. R. (2020). Polygenic Risk Scores Outperform Machine Learning Methods in Predicting Coronary Artery Disease Status. *Genet. Epidemiol.* 44, 125–138. doi:10.1002/gepi.22279

International Schizophrenia ConsortiumPurcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., et al. (2009). Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder. *Nature* 460, 748–752. doi:10.1038/nature08185

Jinn, S., Drolet, R. E., Cramer, P. E., Wong, A. H.-K., Toolan, D. M., Gretzula, C. A., et al. (2017). TMEM175 Deficiency Impairs Lysosomal and Mitochondrial Function and Increases α-synuclein Aggregation. *Proc. Natl. Acad. Sci. USA* 114, 2389–2394. doi:10.1073/pnas.1616332114

Jinn, S., Blauwendraat, C., Toolan, D., Gretzula, C. A., Drolet, R. E., Smith, S., et al. (2019). Functionalization of the TMEM175 p.M393T Variant as a Risk Factor for Parkinson Disease. *Hum. Mol. Genet.* 28, 3244–3254. doi:10.1093/hmg/ddz136

Klinger, J., Ravarani, C., Bannard, C., Lamparter, M., Schwinges, A., Cope, J., et al. (2021). *Critically Ill COVID-19 Status Associated Trait Genetics Reveals CDK6 Inhibitors as Potential Treatment*. doi:10.21203/rs.3.rs-568366/v1 Available at https://www.medrxiv.org/content/10.1101/2021.05.18.21256584v2

Krohn, L., Öztürk, T. N., Vanderperre, B., Ouled Amar Bencheikh, B., Ruskey, J. A., Laurent, S. B., et al. (2020). Genetic, Structural, and Functional Evidence Link TMEM175 to Synucleinopathies. *Ann. Neurol.* 87, 139–153. doi:10.1002/ana.25629

Lo, A., Chernoff, H., Zheng, T., and Lo, S.-H. (2015). Why Significant Variables Aren't Automatically Good Predictors. *Proc. Natl. Acad. Sci. USA* 112, 13892–13897. doi:10.1073/pnas.1518285112

Lynch-Day, M. A., Mao, K., Wang, K., Zhao, M., and Klionsky, D. J. (2012). The Role of Autophagy in Parkinson's Disease. *Cold Spring Harbor Perspect. Med.* 2, a009357. doi:10.1101/cshperspect.a009357

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the Missing Heritability of Complex Diseases. *Nature* 461, 747–753. doi:10.1038/nature08494

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., et al. (2018). A Tutorial on Conducting Genome-wide Association Studies: Quality Control and Statistical Analysis. *Int. J. Methods Psychiatr. Res.* 27, e1608. doi:10.1002/mpr.1608

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., et al. (2011). The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* 95, 629–635. doi:10.1016/j.pneurobio.2011.09.005

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C., et al. (2018). The Parkinson's Progression Markers Initiative (PPMI) - Establishing a PD Biomarker Cohort. *Ann. Clin. Transl Neurol.* 5, 1460–1477. doi:10.1002/acn3.644

Nalls, M. A., Pankratz, N., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., et al. (2014). Large-Scale Meta-Analysis of Genome-Wide Association Data Identifies Six New Risk Loci for Parkinson's Disease. *Nat. Genet.* 46, 989–993. doi:10.1038/ng.3043

Olanow, C. W., Schapira, A. H., LeWitt, P. A., Kieburtz, K., Sauer, D., Olivieri, G., et al. (2006). TCH346 as a Neuroprotective Drug in Parkinson's Disease: A Double-Blind, Randomised, Controlled Trial. *Lancet Neurol.* 5, 1013–1020. doi:10.1016/s1474-4422(06)70602-0

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., 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

Reisberg, S., Iljasenko, T., Läll, K., Fischer, K., and Vilo, J. (2017). Comparing Distributions of Polygenic Risk Scores of Type 2 Diabetes and Coronary Heart Disease within Different Populations. *PLoS One* 12, e0179238. doi:10.1371/journal.pone.0179238

Romero, A., Carrier, P. L., Erraqabi, A., Sylvain, T., Auvolat, A., Dejoie, E., et al. (2016). Diet Networks: Thin Parameters for Fat Genomics. arXiv [cs.LG]. Available at: http://arxiv.org/abs/1611.09340.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting Epistasis in Human Complex Traits. *Nat. Rev. Genet.* 15, 722–733. doi:10.1038/nrg3747

Wray, N. R., Goddard, M. E., and Visscher, P. M. (2007). Prediction of Individual Genetic Risk to Disease from Genome-wide Association Studies. *Genome Res.* 17, 1520–1528. doi:10.1101/gr.6665407

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs Explain a Large Proportion of the Heritability for Human Height. *Nat. Genet.* 42, 565–569. doi:10.1038/ng.608

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., et al. (2018). Efficiently Controlling for Case-Control Imbalance and Sample

Relatedness in Large-Scale Genetic Association Studies. *Nat. Genet.* 50, 1335–1341. doi:10.1038/s41588-018-0184-y

Conflicts of Interest: JC, HB, JK, CR, and MS are employed by biotx.ai GmbH.

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 November 2021
doi: 10.3389/fgene.2021.783513

A Combined Nomogram Model to Predict Disease-free Survival in Triple-Negative Breast Cancer Patients With Neoadjuvant Chemotherapy

Bingqing Xia[1,2], He Wang[3], Zhe Wang[4], Zhaoxia Qian[1], Qin Xiao[2], Yin Liu[2], Zhimin Shao[2], Shuling Zhou[2], Weimin Chai[5], Chao You[2] and Yajia Gu[2]**

[1]International Peace Maternity and Child Health Hospital, Shanghai, China, [2]Shanghai Cancer Center, Fudan University, Shanghai, China, [3]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China, [4]Shanghai United Imaging Medical Technology Co., Ltd., Shanghai, China, [5]Ruijin Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

OPEN ACCESS

Edited by:
Ming Fan,
Hangzhou Dianzi University, China

Reviewed by:
Zaiyi Liu,
Guangdong Provincial People's Hospital, China
Dengbin Wang,
Shanghai Jiaotong University, China

***Correspondence:**
Chao You
youchao8888@aliyun.com
Yajia Gu
cjr.guyajia@vip.163.com

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

Received: *26 September 2021*
Accepted: *22 October 2021*
Published: *12 November 2021*

Citation:
Xia B, Wang H, Wang Z, Qian Z, Xiao Q, Liu Y, Shao Z, Zhou S, Chai W, You C and Gu Y (2021) A Combined Nomogram Model to Predict Disease-free Survival in Triple-Negative Breast Cancer Patients With Neoadjuvant Chemotherapy. Front. Genet. 12:783513. doi: 10.3389/fgene.2021.783513

Background: To investigate whether the radiomics signature (Rad-score) of DCE-MRI images obtained in triple-negative breast cancer (TNBC) patients before neoadjuvant chemotherapy (NAC) is associated with disease-free survival (DFS). Develop and validate an intuitive nomogram based on radiomics signatures, MRI findings, and clinicopathological variables to predict DFS.

Methods: Patients (n = 150) from two hospitals who received NAC from August 2011 to May 2017 were diagnosed with TNBC by pathological biopsy, and follow-up through May 2020 was retrospectively analysed. Patients from one hospital (n = 109) were used as the training group, and patients from the other hospital (n = 41) were used as the validation group. ROIs were drawn on 1.5 T MRI T1W enhancement images of the whole volume of the tumour obtained with a 3D slicer. Radiomics signatures predicting DFS were identified, optimal cut-off value for Rad-score was determined, and the associations between DFS and radiomics signatures, MRI findings, and clinicopathological variables were analysed. A nomogram was developed and validated for individualized DFS estimation.

Results: The median follow-up time was 53.5 months, and 45 of 150 (30.0%) patients experienced recurrence and metastasis. The optimum cut-off value of the Rad-score was 0.2528, which stratified patients into high- and low-risk groups for DFS in the training group ($p < 0.001$) and was validated in the external validation group. Multivariate analysis identified three independent indicators: multifocal/centric disease status, pCR status, and Rad-score. A nomogram based on these factors showed discriminatory ability, the C-index of the model was 0.834 (95% CI, 0.761–0.907) and 0.868 (95% CI, 0.787–949) in the training and the validation groups, respectively, which is better than clinicoradiological nomogram(training group: C-index = 0.726, 95% CI = 0.709–0.743; validation group: C-index = 0.774,95% CI = 0.743–0.805).

Conclusion: The Rad-score derived from preoperative MRI features is an independent biomarker for DFS prediction in patients with TNBC to NAC, and the combined radiomics nomogram improved individualized DFS estimation.

Keywords: radiomics, neoadjuvant chemotherapy, nomogram, triple-negative breast cancer, disease-free survival

INTRODUCTION

Triple-negative breast cancer (TNBC) is a clinical challenge because of its invasive nature, high risk of distant metastasis, and poor prognosis. Compared with other breast cancer patients, TNBC patients are 2–3.5 times more likely to have distant recurrence (Fatayer et al., 2016). It has been demonstrated that the probability of a pathological complete response (pCR) is higher in TNBC patients who receive neoadjuvant therapy (NAC) (close to 31% at present) than in patients with other molecular subtypes, suggesting that NAC improves DFS in this group of patients (Houssami et al., 2012). However, pCR alone is not enough to predict the long-term recurrence-free survival rate of patients with TNBC, and an efficient prognostic biomarker is urgently needed to help stratify patients and create treatment guidelines.

Recently, some studies have indicated that radiomics can be used to obtain a series of related parameters to quantify the heterogeneity of lesions and shows promise for improving tumour prognosis. In previous studies, the radiomics nomogram provided a promising prediction of neoadjuvant chemotherapy efficacy in breast cancer patients based on pretreatment MRI images (Bian et al., 2020; Chen et al., 2020). Another study reported that the radiomics signature(Rad-score) could be used for DFS prediction in HER-2-positive invasive breast cancer treated with NAC, and the radiomics-clinicoradiologic-based nomogram may potentially be useful for personalized treatment strategies (Li et al., 2020). However, there is no relevant research on TNBC.

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) has excellent sensitivity and good specificity for breast cancer diagnosis and plays an important role in characterizing the heterogeneity of tumours. Most studies involving radiomics analysis only use the initial enhancement phase of DCE-MRI, and the additional value of radiomics calculated from later enhancement images was limited. Nevertheless, the radiomics features derived from the phases of multiple DCE-MRI images cannot be ignored, which may imply more information changing over time points.

The purpose of this study was to investigate whether the radiomics derived from all DCE-MRI phases obtained in TNBC patients before NAC are associated with DFS and to compare the combined radiomics nomogram and the clinicoradiological nomogram for their abilities in predicting DFS in patients with TNBC treated with NAC.

MATERIALS AND METHODS

The institutional review board approved this two-institution study and retrospective radiomics data analysis (approval No: 2004216-14), and the requirement for written informed consent was waived.

Patients

Between August 2011 and May 2017, a total of 150 patients from two hospitals were enrolled according to the inclusion criteria. The inclusion criteria included 1) oestrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) were all negative according to a core-needle biopsy performed before treatment (the HER2 score (2+) obtained based on immunohistochemistry and gene amplification was confirmed with fluorescence *in situ* hybridization), 2) patients who received NAC and underwent a final surgery, and 3) patients who underwent an examination using the same machine (Aurora Dedicated Breast MRI System, USA, Aurora). The exclusion criteria included the following: 1) patients who did not undergo a magnetic resonance examination before treatment, 2) patients whose lesions were hardly identified on breast MR images, 3) patients with confirmed systemic metastasis, 4) patients with no final pathological results after treatment, and 5) patients who were lost to follow-up after operations. Finally, all patients were required to undergo an MR examination within 30 days before neoadjuvant therapy. The following information was also recorded for all patients: age, menopausal status, start date of NAC, clinical stage, pre-NAC-T stage and N stage, tumour histologic type, Ki67, surgery type, and date of progression (local recurrence and distant metastasis) to determine duration (months) of DFS. DFS was calculated from the date of surgery to the date of breast cancer recurrence and metastasis, the last confirmation of no evidence of disease, or the most recent follow-up examination.

Magnetic Resonance Imaging

Before treatment, all MR scans were performed with an AURORA 1.5T breast magnetic resonance machine (Aurora Dedicated Breast MRI System, United States, Aurora). The patients underwent this procedure in the prone position with both breasts naturally suspended in a dedicated breast coil. The scanning range included the bilateral breasts and axillary regions. DCE-MRI was performed using axial T1-weighted fat suppression (TE/TR = 5 ms/29 ms, slice thickness = 1.5 mm with no gap, FOV = 360 mm, matrix = 360 × 360) and consisted of one precontrast and three consecutive postcontrast dynamic series. Gd-DTPA was injected into the dorsal hand vein via a bolus injection (0.1 mmol/kg) at a rate of 2.0 ml/s. The scanning time for each phase was approximately 2 min.

All medical images and clinical records were independently reviewed by two radiologists specializing in breast imaging diagnosis (with 5 and 15 years of experience, respectively). The

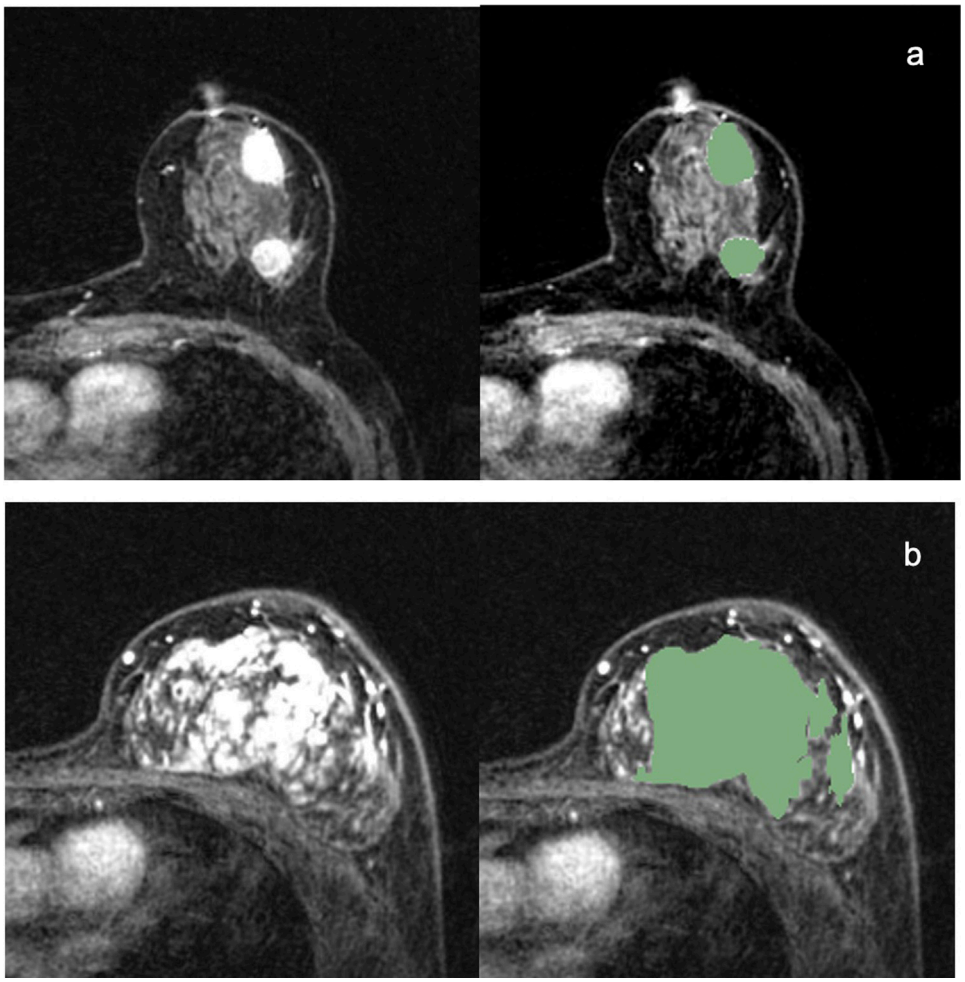

FIGURE 1 | Examples of 3D segmentations of Triple-negative breast cancers.

morphologic manifestations (such as mass or nonmass enhancement and TIC curve) of each lesion were determined according to the 2013 Breast Imaging Reporting and Data System (BI-RADS) MR imaging lexicon standard proposed by the American College of Radiology.

Tumour Masking and Inter-Observer Reproducibility Evaluation

ROIs were manually drawn by the radiologist on the whole volume of the tumours (including the necrotic regions) with 3D Slicer software (https://www.slicer.org). The 3D segmentation ROIs of the whole tumour were first created on the first post-contrast DCE images and then propagated to the pre-contrast and the other two post-contrast series of DCE images. For multifocal/centric and nonmass enhancement tumours, ROIs were drawn over all lesions. Examples of 3D segmentation are shown in **Figure 1**. **Figure 1A** is MR images of TNBC with multifocal/centric masses. The green area represents the scope of ROI delineation, and each lesion is delineated by layers. **Figure 1B** is MR images of TNBC with non-mass lesions.The green area is delineated by ROI and delineated according to the scope of enhancement.

Using 50 randomly selected samples, the interobserver reproducibility of ROI detection and radiomic feature extraction was measured. Two experienced radiologists (BQX and QX) described the ROI independently, and then the radiomic features extracted from the above two ROIs were compared to obtain the interclass correlation coefficient. An ICC score greater than 0.8 was interpreted as satisfactory agreement. The ICC for the radiomic features was defined as high (ICC $\geq$ 0.8), medium (0.8 > ICC $\geq$ 0.5) or low (ICC < 0.5).

Treatment Regimen and Criteria for pCR and Recurrence

All patients received paclitaxel sequential/combined anthracycline neoadjuvant chemotherapy with or without platinum. The median duration of NAC was 4 (range, 4–8) months. pCR was defined as ypT0/is and ypN0, which indicate the absence of residual invasive

TABLE 1 | Three groups of extracted features.

	Group	Number (features)	Description
a	Shape features on DCE (DCEshape)	14	The 14 shape-based features were calculated based on the first postcontrast DCE images
b	Texture features based on DCE images with 4 time series (DCEtexture)	372	The 93 texture features (including 18 first-order features, 24 grey-level co-occurrence matrix (GLCM) features, 16 grey-level run length matrix (GLRLM) features, 16 grey-level size zone matrix (GLSZM) features, 5 neighbouring grey tone difference matrix (NGTDM) features, and 14 grey-level dependence matrix (GLDM) features) were calculated based on these four series image sets to yield 372 features
c	Sequential features based on DCE images (DCEsequential)	930	The first six features, including mean, variance, kurtosis, skewness, energy, and entropy, were extracted for each individual subject. The other four features, including Kendall-tau-b, conservation, stability, and dispersion, were calculated for the interactive information between the current subject and the remainder of the subjects. Therefore, a total of 930 DCEsequential features were extracted from 93 texture features

carcinoma in breast tissues with or without ductal carcinoma *in situ* and the absence of any residual cancer in the sampled axillary lymph nodes. A pathological response was determined by senior breast pathologists. Recurrence was defined as local-regional (confined to the ipsilateral breast or chest wall and/or axillary, infraclavicular or supraclavicular lymph nodes) and distant metastasis (to other parts of the body or the contralateral breast). Breast cancer recurrence was confirmed by biopsy, and metastasis was confirmed by biopsy when appropriate or on the basis of an imaging assessment, including PET/CT and other imaging modalities.

Radiomics Analysis, Feature Selection and Rad-Score

The radiomics signature included 1316 radiomics features that were extracted from the training group by the PyRadiomics package in Python software (v. 3.6, Python Software Foundation, https://www.python.org/). All these features were classified into 3 groups (**Table 1**). To characterize the textural changes observed on DCE images over time series, we measured ten new sequential features for each texture feature described in group b (**Supplementary Table S1**). All these features have been applied in previous radiomics studies (Li et al., 2020). Forward stepwise regression was applied to select features. Rad-score was calculated for each patient via a linear combination of selected features that were weighted by their respective coefficients. Feature selection was achieved using the Statistics Toolbox in MATLAB (v. R2018a; MathWorks, Natick, MA).

STATISTICAL ANALYSIS

We compared patient characteristics using commercially available statistical software (IBM SPSS 24.0). When appropriate, significant differences between the training and validation groups were assessed by the Chi-square test, Fisher's test or *t*-test. A two-sided *p* value of less than 0.05 indicates a significant difference. The Rad-scores were divided into two groups (high-risk vs low-risk) using receiver operating characteristic (ROC) curve analysis according to optimal cut-off value determined by maximizing the Youden index (sensitivity + specificity-1). Significant variables in the univariate Cox proportional hazard model ($p < 0.05$) were included in the multivariate analysis. The combined radiomics nomogram incorporated the radiomics signature and various independent risk factors based on multivariate analysis in the training group and was then validated in the validation group. The predictive ability and discriminatory performance of each established model were evaluated using an index of probability of concordance (C-index), and the C-index between the predicted probability and actual outcome was calculated to evaluate the predictive ability and discrimination of the model (Wolbers et al., 2009). The value of the C-index ranges from 0.5–1.0, with 0.5 indicating random chance and 1.0 indicating perfectly accurate discrimination. The nomograms were subjected to bootstrapping validation (1000 bootstrap resamples) to obtain a relatively corrected C-index.

RESULTS

Patient Characteristics

The clinicopathological and MR imaging characteristics of the training and validation groups with TNBC are listed in **Table 2**. Except for the clinical stage, pre-NAC N stage and pCR status, there were no differences between the training and validation groups. The median follow-up time was 54 months (range, 1–101 months) for the training group and 48 months (range, 1–88 months) for the validation group. There were 45 (30.0%) recurrences, 30 (20.0%) in the training group and 15 (10.0%) in the validation group, including 35 patients with distant metastasis (one also had additional local-regional recurrence), 8 with local-regional recurrence only, and 2 with contralateral breast cancers.

Radimics Analysis, Rad-Score Building and Validation

The ICC for radiomic features between the two radiologists BQX and QX ranged from 0.8732 to 0.9671. Two radiologists generally reached a consensus on the delineations. To verify the importance of the new features, two different Radimics models were delevoped. Model 1 only uses the features derived from the first postcontrast phase, while Model 2 uses the features derived from all dynamic phases, including the new features.

TABLE 2 | Comparison of clinical and pathological and pretreatment MR imaging characteristics between training and validation groups.

Characteristics	Training group (*n* = 109)	Validation group (*n* = 41)	*p*
Age, mean (SD), y	47.3 ± 11.1	48.6 ± 13.3	0.545
Menopausal status			0.322
Premenopausal	63(57.8)	20(48.8)	
Postmenopausal	46(42.2)	21(51.2)	
Clinical Stage			0.007[a]
II	83(76.1)	22(53.7)	
III	26(23.9)	19(46.3)	
Pre-NAC T-stage			0.061
T1	10(9.2)	4(9.8)	
T2	68(62.4)	16(39.0)	
T3	22(20.2)	14(34.1)	
T4	9(8.3)	7(17.1)	
Pre-NAC N-stage			0.032[a]
N0	38(34.9)	10(24.4)	
N1	55(50.5)	21(51.2)	
N2	7(6.4)	9(22.0)	
N3	9(8.3)	1(2.4)	
Pathological type			0.575
IDC	105(96.3)	41(100.0)	
ILC,IMPC	4(3.7)	0(0.0)	
KI-67			0.090
≤14%	6(5.5)	6(14.6)	
> 14%	103(94.5)	35(85.4)	
Surgery type			0.075
Breast conservation	21(19.3)	3(7.3)	
Mastectomy	88(80.7)	38(92.7)	
Features at MR imaging			0.455
Mass	86(78.9)	30(73.2)	
Nonmass	23(21.1)	11(26.8)	
Kinetics			0.684
Washout	104(95.4)	38(92.7)	
Plateau or persistent	5(4.6)	3(7.3)	
Multi-focal/centric disease			0.695
Present	31(28.4)	13(31.7)	
Absent	78(71.6)	28(68.3)	
pCR			0.022[a]
Yes	46(42.2)	9(22.0)	
No	63(57.8)	32(78.0)	
Lymphovascular invasion			0.052
Present	23(21.1)	15(36.6)	
Absent	86(78.9)	26(63.4)	
Disease-free survival			0.280
Yes	79(72.5)	26(63.4)	
No	30(27.5)	15(36.6)	

Data are expressed as n*(%) unless otherwise specified.*
The p *values for age were determined by* t *test, while other* p *values were determined by Chi square or Fisher exact tests, as appropriate.*
[a]*indicate statistical significance (*p < *0.05).*
IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; IMPC, invasive micropapillary carcinoma; pCR, pathological complete response.

The results for the two models are shown in **Table 3**. Model 2 achieved a predictive accuracy of 85.4%, sensitivity of 50.0%, specificity of 97.6%, PPV of 88.0%, and NPV of 85.0%, which was more robust than Model 1. Finally, Model 2 was selected for the following study.

In Model 2, six textural features were selected for predicting DFS after forward stepwise regression selection, and the Rad-score calculation formula is presented:y = 0.25688+(−0.12986)×Skewness_glcm_Imc1+(−0.13965)×Entropy_firstorder_RootMeanSquared+(−0.094626)×Entropy_ngtdm_Busyness+0.10472×Kendall-tau-b_glcm_Idmn+(−0.23802)×Conservation_glcm_DifferenceAverage+0.2713×Conservation_ngtdm_Complexity. The above selected features are all from group c (DCEsequential). There was a significant difference in Rad-scores between the recurrence and no recurrence groups ($p < 0.001$) in the training group. The median Rad-score was 0.2349 (range, –0.3165 to 0.9846; interquartile range, 0.1038–0.3812). The optimum cut-off value generated by the ROC curve was 0.2528, and the AUC was 0.852 (95% CI, 0.773–0.932). Using this threshold value, patients were classified into a high-risk group (Rad-score ≥ 0.2528) and a low-risk group (Rad-score < 0.2528). Kaplan-Meier curves showed that the radiomics signature was associated with DFS in the training group ($p < 0.001$), and this finding was confirmed in the validation group ($p < 0.001$) (**Figure 2**).

Univariate and Multivariate Analyses of the Risk Factors for RFS

The results of the univariate and multivariate analyses of the risk factors for RFS in the training group are shown in **Table 4**. A higher Rad-score, multifocal/centric lesions, nonmass lesions, ILC/MIPC histological type, non-pCR and lymphovascular invasion were associated with worse DFS. Furthermore, in the multivariate Cox analysis, a higher Rad-score (DFS odds ratio 26.685; 95% CI 6.654–107.010; $p = 0.000$), multifocal/centric lesions (DFS odds ratio, 2.522; 95% CI, 1.160–5.481; $p = 0.020$), and pCR status (DFS odds ratio, 0.285; 95% CI, 0.100–0.810; $p = 0.019$) remained independent prognostic factors (**Table 4**).

Radiomics Nomogram Building and Validation

The C-index of the two kinds of nomogram models for the prediction of DFS in the training group and validation group is shown in **Table 5**. A combined radiomics nomogram was developed based on multifocal/centric disease status, pCR

TABLE 3 | Summary of radiomics model1 and model2 results.

	Accuracy	Sensitivity	Specificity	PPV	NPV
Model1 (1st PC phase)	76.6%(74.3–78.0)	17.4%(10.7–21.4)	97.1%(95.1, 98.8)	68.1%(50.0–83.3)	77.3%(76.0–78.2)
Model2 (All phases, 1pre-contrast and 3 PC phases	85.4%(84.4–86.2)	50.0%(46.4–50.0)	97.6%(96.3–98.8)	88.0%(82.4–93.3)	85.0%(84.8–85.1)

Confidence intervals are in parenthesis. Above two models were performed using a fine Gaussian support vector machine and conducted using 5-fold cross validation to overcome overfitting. The procedure was repeated for ten rounds to average the estimates of performance.
PC, post-contrast; PPV, positive predictive value; NPV, negative predictive value.

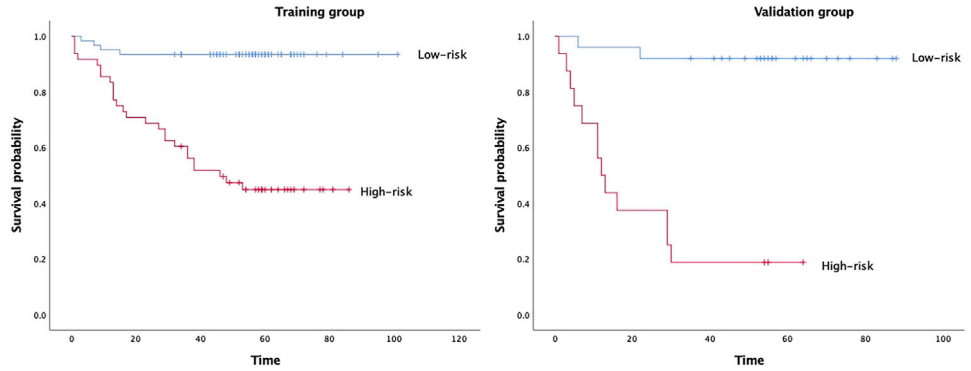

FIGURE 2 | Kaplan–Meier survival analyses according to the radiomics signature with low-risk and high-risk patients in training and validation groups.

TABLE 4 | Univariate and multivariate analysis of disease-free survival in training group.

Characteristics	Univariate analysis			Multivariate analysis		
	OR	95% CI	*p* value	OR	95% CI	*p* value
Age, < 35 years versus ≥35 years	1.775	0.538–5.855	0.346			
Menopausal status, premenopausal versus postmenopausal	1.661	0.810–3.404	0.166			
Clinical Stage, II versus III	1.529	0.700–3.340	0.287			
Pre-NAC Tstage(T1 reference)			0.306			
T2	3	0.354–25.439	0.314			
T3	5.143	0.547–48.365	0.152			
T4	7.2	0.622–83.342	0.114			
Pre-NAC Nstage(N0 reference)			0.248			
N1	1.322	0.512–3.41	0.564			
N2	4.296	0.806–22.9	0.088			
N3	0.403	0.044–3.669	0.42			
Pathologic type, IDC versus ILC, IMPC	5.330	1.602–17.735	0.006[a]	0.851	0.210–3.445	0.821
KI-67, ≤20% versus > 20%	0.452	0.137–1.493	0.193			
Surgery type, Breast conservation versus Mastectomy	2.252	0.683–7.426	0.182			
Features at MR imaging, Mass versus Nonmass	2.454	1.145–5.262	0.021[a]	1.565	0.639–3.832	0.327
Kinetics, Washout versus Plateau or persistent	0.659	0.090–4.84	0.682			
Multi-focal/centric disease, Present versus Absent	3.177	1.549–6.517	0.002[a]	2.522	1.160–5.481	0.020[a]
pCR, Yes versus No	0.232	0.089–0.608	0.003[a]	0.285	0.100–0.810	0.019[a]
Lymphovascular invasion, Present versus Absent	2.254	1.054–4.820	0.036[a]	0.995	0.402–2.461	0.991
Rad-score	52.829	14.821–188.300	0.000[a]	26.685	6.654–107.010	0.000[a]

OR, odds ratio; CI, confidence interval; pCR, pathological complete response.
[a]*indicate statistical significance (p ≤ 0.05).*

TABLE 5 | Performance of the two nomogram for prediction of disease-free survival.

Nomogram	Training		Validation	
	C-index	95%CI	C-index	95%CI
Combined Radiomoics nomogram	0.834	0.761–0.907	0.868	0.787–0.949
Clinicoradiological nomogram	0.726	0.709–0.743	0.774	0.743–0.805

C-index,index of probability of concordance; CI, confidence interval.

status, and Rad-score to predict the DFS rate for NAC among TNBC patients (**Figure 3**). A total score was obtained by adding each single score to estimate the 2-/3-/5-years DFS probability. The C-index was 0.834 (95% CI, 0.761–0.907) and 0.868 (95% CI, 0.787–0.949) in the training and validation groups, respectively, indicating that the combined radiomics nomogram had better discriminatory capability than the clinicoradiological nomogram.

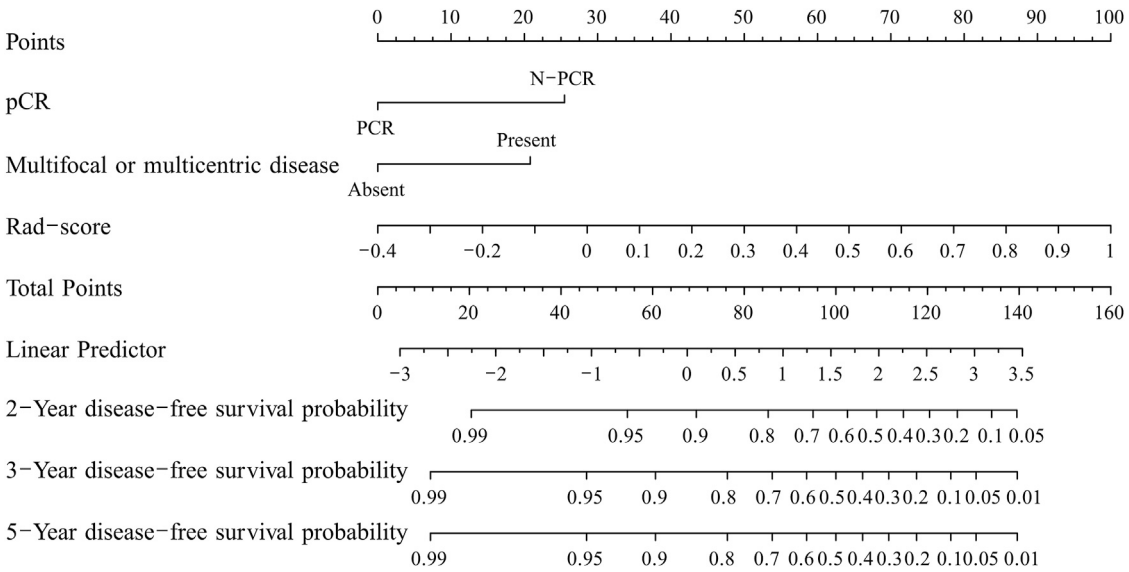

FIGURE 3 | The developed nomogram for predicting disease-free survival in triple-negative breast cancer patients after neoadjuvant chemotherapy.

DISCUSSION

In our study, we demonstrated the prognostic value of multiphase CE-MRI radiomics features for patients with TNBC treated with NAC. In addition, we developed a combined radiomics model that incorporates the radiomics signature and MRI and pathology findings for the individualized prediction of DFS in TNBC patients who underwent NAC. Compared with the clinicoradiological nomogram, the combined radiomics nomogram had superior prognostic performance in DFS estimation.

For feature extraction and selection, we measured ten new sequential features to characterize the textural changes observed on DCE images over time series. These features have not previously been used or described in the domain of breast radiomics except in Li et al. (2020) and Xie et al. (2019) studies, who used new features to differentiate different subtypes of breast cancer and predict DFS in patients with HER2-positive breast cancer treated with NAC. We compared two models to investigate whether the accuracy of the radiomics model was significantly improved after adding new features. Roberto et al. (Lo Gullo et al., 2020) and Gibbs et al. (2019) both demonstrated that delayed postcontrast phases did not add any significant discriminative value to the analysis, which is inconsistent with our research results. The reason may be that we added new sequential features, but they did not include them, and the subjects of their study were subcentimetre masses that were much smaller than ours lesions. Furthermore, the sequential texture features derived from dynamic phases may capture information on both spatial heterogeneity and tumour perfusion, which is more valuable in predicting DFS than differentiating benign and malignant lesions.

In our study, the final Rad-score calculation formula included six potential features all from the new sequential features. The six selected radiomics features comprised one from skewness, two from entropy, one from Kendall-tau-b and two from conservation. Among them, other studies have also emphasized the importance of skewness and entropy in reflecting the heterogeneity of tumours. Kendall-tau-b and conservation were calculated from interactive information between the current subject and the remainder of the subjects, which means that if the changes increased, the Rad-score increased, indicating a worse prognosis. One possible interpretation is that this change may be related to the high perfusion of the tumours, and tumours with abundant blood supply tend to be more heterogeneous and have a worse prognosis. Attentionally, three of the six selected features were GLCM (grey level cooccurrence matrix), and two were NGTDM (neighbourhood grey-tone difference matrix). At present, GLCM is the most widely used texture extraction method, which has also been confirmed in assessing tumour heterogeneity and plays a very important role in various fields. The basic principle of the GLCM is based on spatial correlation between neighbouring pixels. NGTDM represents contrast, which is determined by changes in intensity between a target voxel and the surrounding neighbours and then enables the calculation of the apparent difference between neighbouring regions of voxel intensities. Contrast is also related to tumour heterogeneity; tumours with poor prognosis tend to have higher contrast (Sun and Wee, 1983). Our results also showed that the Rad-score had a promising high value for predicting DFS, which was confirmed by Kaplan–Meier survival curves in the training group ($p < 0.0001$) and in the validation group ($p < 0.0001$). Interestingly, the cut-off value (Rad-score = 0.2528. for predicting DFS was similar to QL's study (Rad-score = 0.2523), regardless of TNBC or HER2-positive breast cancer with NAC.

There were differences in clinical stage, pre-NAC N stage and pCR status between the training and validation groups, which might be associated with differences in study populations with different hospitals. In the validation group, the later the clinical stage, the more difficult it was to achieve pCR. Various previous studies have confirmed that a tumour's response to neoadjuvant

therapy provides prognostic information. The attainment of a pCR after NAC and surgical resection improved the DFS rate of patients (Houssami et al., 2012), (Cortazar et al., 2014; Chen et al., 2017; Symmans et al., 2017), consistent with our study. However, 42.2% of the patients in the training group received pCR after NAC, and this rate is higher than those reported in other studies (Houssami et al., 2012), potentially because we ruled out patients who did not undergo surgery and did not finish a complete NAC regimen. Interestingly, in the training group, multifocal/centric lesions were identified as independent predictors for the DFS of TNBC after NAC. Many studies (Duraker and Çaynak, 2014; Lang et al., 2017) have demonstrated that multifocal/centric foci exhibit more biologically aggressive behaviour than has been observed for unifocal breast cancer, and this could influence DFS and OS. Although the multifocal/centric lexicon was not included in BI-RADS, these patients should receive more attention during postoperative follow-up. While Park et al. (2018a) found that N-stage was a predictor of DFS in breast cancer, our analysis failed to support these findings, possibly due to differences in study populations. In addition, the features at MR imaging(mass vs nonmass) was not associated with DFS in multivariate analysis of variance in our study, which was consistent with the study of Tahmassebi et al. (2019).

The prognostic ability of radiomics signatures has been demonstrated in many studies. For example, Li et al. (2016) suggested that image-based radiomics features may be helpful in assessing the risk of breast cancer recurrence. Park et al. (2018b) demonstrated that Rad-scores generated from radiomics signatures based on preoperative MRI have prognostic value. In our study, we analysed preoperative MRI findings in TNBC, a special pathological type of breast cancer, and supported the notion that the Rad-score helps stratify patients, and patients from high-risk groups need more careful follow-up management.

In this study, we developed a radiomics signature-based nomogram for the individualized prediction of recurrence in patients with TNBC after NAC. The nomogram incorporates three components of a radiomics signature with six selected features, including pCR status and MR findings indicating multifocal/centric lesions, which is promising to facilitate individualized predictions and the prediction of follow-up needs in patients with poor outcomes with regard to DFS.

Our study has several limitations. First, this is a retrospective study. Second, most of the patients were examined using MR after a biopsy, which might have affected assessments. Third, we discuss only DCE images in our study, and further prospective studies should include a variety of breast MR imaging protocols, such as T2W, DWI, and DCE-MRI.

CONCLUSION

In conclusion, the results of our study show that the identified Rad-score has the potential to be used as a biomarker for risk stratification for DFS in patients with TNBC after NAC. In addition, our results show that a radiomics nomogram that incorporates a radiomics signature and MRI and clinicopathological findings can be used to facilitate the individualized prediction of recurrence in patients with TNBC after NAC and surgery. This type of quantitative radiomics prognostic model of breast cancer could be useful for precision medicine and could affect patient follow-up strategies.

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 Fudan University Cancer Hospital Institutional Review Board. This study was a retrospective study and was passed by the Fudan University Cancer Hospital Institutional Review Board (approval No:2004216-14), so the requirement for written informed consent was waived for retrospective data.

AUTHOR CONTRIBUTIONS

BX and HW contributed equally to this work and share first authorship. CY and YG contributed equally to this work and share last authorship. BX: Writing-Original draft preparation, Writing, Conceptualization, Methodology. HW: Data curation, Conceptualization, Validation, Methodology. ZW: Visualization, Investigation, Software. QX: Formal analysis, Investigation. YL and ZS and SZ: Resources. ZQ: Supervision, Software. WC: Validation, Supervision. YG and CY: Writing-Review and Editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.783513/full#supplementary-material

REFERENCES

Bian, T., Wu, Z., Lin, Q., Wang, H., Ge, Y., Duan, S., et al. (2020). Radiomic Signatures Derived from Multiparametric MRI for the Pretreatment Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer. *Br. J. Radiol.*. 93(1115):20200287.doi:10.1259/bjr.20200287

Chen, S., Shu, Z., Li, Y., Chen, B., Tang, L., Mo, W., et al. (2020). Machine Learning-Based Radiomics Nomogram Using Magnetic Resonance Images for Prediction of Neoadjuvant Chemotherapy Efficacy in Breast Cancer Patients. *Front. Oncol.*. 10:1410.doi:10.3389/fonc.2020.01410

Chen, V. E., Gillespie, E. F., Zakeri, K., Murphy, J. D., Yashar, C. M., Lu, S., et al. (2017). Pathologic Response after Neoadjuvant Chemotherapy Predicts Locoregional Control in Patients with Triple Negative Breast Cancer. *Adv. Radiat. Oncol.* 2 (2), 105–109. doi:10.1016/j.adro.2017.01.012

Cortazar, P., Zhang, L., Untch, M., Mehta, K., Costantino, J. P., Wolmark, N., et al. (2014). Pathological Complete Response and Long-Term Clinical Benefit in

Breast Cancer: The CTNeoBC Pooled Analysis. *Lancet* 384 (9938), 164–172. doi:10.1016/s0140-6736(13)62422-8

Duraker, N., and Çaynak, Z. C. (2014). Axillary Lymph Node Status and Prognosis in Multifocal and Multicentric Breast Carcinoma. *Breast J.* 20 (1), 61–68. doi:10.1111/tbj.12205

Fatayer, H., Sharma, N., Manuel, D., Kim, B., Keding, A., Perren, T., et al. (2016). Serial MRI Scans Help in Assessing Early Response to Neoadjuvant Chemotherapy and Tailoring Breast Cancer Treatment. *Eur. J. Surg. Oncol.*. 42(7):965–972.doi:10.1016/j.ejso.2016.03.019

Houssami, N., MacAskill, P., Von Minckwitz, G., Marinovich, M. L., and Mamounas, E. (2012). Meta-analysis of the Association of Breast Cancer Subtype and Pathologic Complete Response to Neoadjuvant Chemotherapy. *Eur. J. Cancer*. 48(18):3342–3354. doi:10.1016/j.ejca.2012.05.023

Lang, Z., Wu, Y., Li, C., Li, X., Wang, X., and Qu, G. (2017). Multifocal and Multicentric Breast Carcinoma: A Significantly More Aggressive Tumor Than Unifocal Breast Cancer. *Anticancer Res.* 37 (8), 4593–4598. doi:10.21873/anticanres.11858

Li, H., Zhu, Y., Burnside, E. S., Drukker, K., Hoadley, K. A., Fan, C., et al. (2016). MR Imaging Radiomics Signatures for Predicting the Risk of Breast Cancer Recurrence as Given by Research Versions of MammaPrint, Oncotype DX, and PAM50 Gene Assays. *Radiology* 281 (2), 382–391. doi:10.1148/radiol.2016152110

Lo Gullo, R., Daimiel, I., Rossi Saccarelli, C., Bitencourt, A., Gibbs, P., Fox, M. J., et al. (2020). Improved Characterization of Sub-centimeter Enhancing Breast Masses on MRI with Radiomics and Machine Learning in BRCA Mutation Carriers. *Eur. Radiol.* 30 (12), 6721–6731. doi:10.1007/s00330-020-06991-7

Park, H., Lim, Y., Ko, E. S., Cho, H.-h., Lee, J. E., Han, B.-K., et al. (2018). Radiomics Signature on Magnetic Resonance Imaging: Association with Disease-free Survival in Patients with Invasive Breast Cancer. *Clin. Cancer Res.* 24, 4705–4714. doi:10.1158/1078-0432.ccr-17-3783

Park, H., Lim, Y., Ko, E. S., Cho, H.-h., Lee, J. E., Han, B.-K., et al. (2018). Radiomics Signature on Magnetic Resonance Imaging: Association with Disease-free Survival in Patients with Invasive Breast Cancer. *Clin. Cancer Res.* 24 (19), 4705–4714. doi:10.1158/1078-0432.ccr-17-3783

Sun, C., and Wee, W. G. (1983). Neighboring gray Level Dependence Matrix for Texture Classification. *Comput. Vision, Graph Image Process.* 23 (3), 341–352. doi:10.1016/0734-189x(83)90032-4

Symmans, W. F., Wei, C., Gould, R., Yu, X., Zhang, Y., Liu, M., et al. (2017). Long-Term Prognostic Risk after Neoadjuvant Chemotherapy Associated with Residual Cancer Burden and Breast Cancer Subtype. *Jco* 35 (10), 1049–1060. doi:10.1200/jco.2015.63.1010

Tahmassebi, A., Wengert, G. J., Helbich, T. H., Bago-Horvath, Z., Alaei, S., Bartsch, R., et al. (2019). Impact of Machine Learning with Multiparametric Magnetic Resonance Imaging of the Breast for Early Prediction of Response to Neoadjuvant Chemotherapy and Survival Outcomes in Breast Cancer Patients. *Invest. Radiol.* 54 (2), 110–117. doi:10.1097/rli.0000000000000518

Wolbers, M., Koller, M. T., Witteman, J. C. M., and Steyerberg, E. W. (2009). Prognostic Models with Competing Risks. *Epidemiology* 20 (4), 555–561. doi:10.1097/ede.0b013e3181a39056

Xie, T., Wang, Z., Zhao, Q., Bai, Q., Zhou, X., Gu, Y., et al. (2019). Machine Learning-Based Analysis of MR Multiparametric Radiomics for the Subtype Classification of Breast Cancer. *Front. Oncol.* 9 (JUN), 1–10. doi:10.3389/fonc.2019.00505

Li, Q, Xiao, Q, Li, J, Duan, S, Wang, H, and Gu, Y. MRI-based radiomic signature as a prognostic biomarker for her2-positive invasive breast cancer treated with NAC. *Cancer Manag Res*. 2020;12:10603-10613. doi:10.2147/CMAR.S271876

Gibbs, P, Onishi, N, Sadinski, M, Gallagher, KM, Hughes, M, Martinez, DF, et al. Characterization of Sub-1 cm Breast Lesions Using Radiomics Analysis. *J Magn Reson Imaging*. 2019;50(5):1468-1477.doi:10.1002/jmri.26732

Conflict of Interest: Author ZW was employed by the company of Shanghai United Imaging Medical Technology.

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

The handling editor declared a past co-authorship with the authors (CY, YG).

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 Genetics

ORIGINAL RESEARCH
published: 30 November 2021
doi: 10.3389/fgene.2021.730732

A Novel Nine-Gene Signature Associated With Immune Infiltration for Predicting Prognosis in Hepatocellular Carcinoma

Rongqiang Liu[1†], ZeKun Jiang[2†], Weihao Kong[3†], Shiyang Zheng[4], Tianxing Dai[5] and Guoying Wang[1*]*

[1]Department of Hepatobiliary Surgery, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China, [2]Department of Gastrointestinal Surgery, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China, [3]Department of Emergency Surgery, The First Affiliated Hospital of Anhui Medical University, Hefei, China, [4]Department of Breast Surgery, The Third Affiliated Hospital of Guangzhou Medical University, Guangzhou, China, [5]Department of Hepatic Surgery and Liver Transplantation Center, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

OPEN ACCESS

Edited by:
Jiangning Song,
Monash University, Australia

Reviewed by:
Jun Liu,
Yuebei People's Hospital, China
Yong-Zi Chen,
Tianjin Medical University Cancer Institute and Hospital, China

***Correspondence:**
Guoying Wang
wanggy3@126.com
Tianxing Dai
daitx1991@126.com

[†]These authors have contributed equally to this work

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

Received: 25 June 2021
Accepted: 08 November 2021
Published: 30 November 2021

Citation:
Liu R, Jiang Z, Kong W, Zheng S, Dai T and Wang G (2021) A Novel Nine-Gene Signature Associated With Immune Infiltration for Predicting Prognosis in Hepatocellular Carcinoma. Front. Genet. 12:730732. doi: 10.3389/fgene.2021.730732

Background: Hepatocellular carcinoma (HCC) is one of the most common malignant tumors worldwide, and its prognosis remains unsatisfactory. The identification of new and effective markers is helpful for better predicting the prognosis of patients with HCC and for conducting individualized management. The oncogene Aurora kinase A (AURKA) is involved in a variety of tumors; however, its role in liver cancer is poorly understood. The aim of this study was to establish AURKA-related gene signatures for predicting the prognosis of patients with HCC.

Methods: We first analyzed the expression of AURKA in liver cancer and its prognostic significance in different data sets. Subsequently, we selected genes with prognostic value related to AURKA and constructed a gene signature based on them. The predictive ability of the gene signature was tested using the HCC cohort development and verification data sets. A nomogram was constructed by integrating the risk score and clinicopathological characteristics. Finally, the influence of the gene signature on the immune microenvironment in HCC was comprehensively analyzed.

Results: We found that AURKA was highly expressed in HCC, and it exhibited prognostic value. We selected eight AURKA-related genes with prognostic value through the protein-protein interaction network and successfully constructed a gene signature. The nine-gene signature could effectively stratify the risk of patients with HCC and demonstrated a good ability in predicting survival. The nomogram showed good discrimination and consistency of risk scores. In addition, the high-risk group showed a higher percentage of immune cell infiltration (i.e., macrophages, myeloid dendritic cells, neutrophils, and CD4+T cells). Moreover, the immune checkpoints SIGLEC15, TIGIT, CD274, HAVCR2, and PDCD1LG2 were also higher in the high-risk group versus the low-risk group.

Conclusions: This gene signature may be useful prognostic markers and therapeutic targets in patients with HCC.

Keywords: AURKA, gene signature, hepatocellular carcinoma, prognosis, immune infiltration, nomogram

INTRODUCTION

According to global data, hepatocellular carcinoma (HCC) is the most common primary liver tumor and the third most common risk factor for cancer-related deaths worldwide (Sung et al., 2021). In 2020, there were approximately 906,000 newly diagnosed patients with liver cancer worldwide and approximately 830,000 liver cancer-related deaths (Sung et al., 2021). The main causes of liver cancer include hepatitis virus infection, smoking, alcoholic cirrhosis, chemical drugs, and aflatoxin infection (Forner et al., 2018). Approximately 400,000 people die annually in China, accounting for >50% of liver cancer-related deaths globally (Chen et al., 2016). Early-stage HCC is insidious and difficult to detect; consequently, a large number of patients already have advanced-stage disease at the time of diagnosis. At present, liver cancer is mainly treated by surgical resection, supplemented by other methods, such as ablation therapy, targeted therapy, and immunotherapy (Hartke et al., 2017). However, even with timely intervention, the recurrence and mortality rates remain high due to the high degree of malignancy in liver cancer, rapid disease progression, and poor prognosis (Li et al., 2015) There are numerous markers used for predicting the prognosis of patients with liver cancer; nevertheless, their effectiveness is currently limited. Therefore, there is an urgent need to identify more effective biomarkers for predicting the prognosis of patients with liver cancer.

Aurora kinase A (AURKA) is a member of the Aurora kinase family, which consists of AURKA A, B, and C (Carmena and Earnshaw, 2003). Human AURKA is located on chromosome 20q13 and encodes a protein of 403 amino acids. It mainly regulates mitotic spindle formation, stability, and chromosome segregation, and plays an important role in cell cycle regulation (Lindon et al., 2016). The abnormal expression of AURKA can lead to chromosomal abnormalities and instability of the cell genome, which is a risk factor for tumor formation (Wu et al., 2018). AURKA is abnormally expressed in a variety of tumors and regulates tumor proliferation, migration, invasion, and metastasis (Yan et al., 2016) In addition, it is involved in multiple signaling pathways, such as the TP53 pathway, Ras/mitogen-activated protein kinase (MAPK) pathway and NFKB pathway (Katayama et al., 2004; Briassouli et al., 2007; Umstead et al., 2017). Previous studies have confirmed that AURKA is related to the prognosis of a variety of cancers (breast, colorectal, pancreatic, gastric, and head and neck) and may be a therapeutic target (Jeng et al., 2004; Reiter et al., 2006; Zhang et al., 2015).

Investigations showed that AURKA played an important role in liver cancer progression. Jeng et al. confirmed that AURKA was overexpressed frequently and correlated with high grade and high stage in HCC (Jeng et al., 2004). Lu et al. reported that AURKA mediated c-Myc's oncogenic effects in HCC (Lu et al., 2015). Zhang et al. revealed AURKA promoted chemoresistance through targeting NF-kappaB/microRNA-21/PTEN signaling pathway in HCC (Zhang et al., 2014). Chen et al. suggested that AURKA promoteed cancer metastasis through regulating epithelial-mesenchymal transition and cancer stem cell properties in HCC (Chen et al., 2017). However, the specific mechanism of AURKA in HCC is still not very clear and needs to be further explored.

Thus far, no study investigated the role of AURKA gene and AURKA-related prognostic genes in liver cancer. It is well established that the tumor immune microenvironment influences tumor progression (Locy et al., 2018). Currently, the immunological value of AURKA in liver cancer has not been reported. In this study, we first analyzed its clinical value in HCC and selected prognostic genes associated with AURKA. Furthermore, we developed an AURKA-related gene signature in HCC. Next, we constructed a nomogram by combining risk scores and clinical characteristics. Finally, we evaluated the relationship between the gene signature and tumor immunity in HCC.

METHODS

Identification of AURKA as a Differentially Expressed Gene (DEG)

The RNA-seq data of LIHC patients from The Cancer Genome Atlas (TCGA, http://gdc.cancer.gov/) database and three datasets, including GSE14323 (HCC, n = 55; normal, n = 60), GSE14520 (HCC, n = 225; normal, n = 220) and GSE25097 (HCC, n = 268; normal, n = 289), from Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) database was downloaded to analysis the AURKA expression in LIHC patients. Gene expression levels were normalized by Robust Multi-Array Average (RMA).

Differential gene expression analysis of mRNAs was performed based on TCGA database. Limma package of R software (R version 3.6.2) was used to conduct differential gene expression analysis. The adjusted p-value was analyzed to correct for false positive results in TCGA or GTEx. "Adjusted p < 0.05 and Log (Fold Change) >1 or Log (Fold Change) <−1" were regarded as the thresholds of differential expression of mRNAs. The study flowchart was presented in **Figure 1**.

Protein-Protein Interaction Network and Gene Enrichment Analysis

STRING (version 11.0; http://string.embl.de/) is an open-access biological database that predicts comprehensive interactions of genes at the protein level from multiple organisms including Homo sapiens (Mering et al., 2003). To screen AURKA-related genes, we used STRING to explore AURKA-related genes as well as conduct PPI network analysis on AURKA-related genes. The protein–protein interactions (PPI) with medium confidence >0.4 were regarded as significant.

To further confirm the underlying function of potential targets, the data were analyzed by functional enrichment. Gene Ontology (GO) is a widely-used tool for annotating genes with functions, especially molecular function (MF), biological pathways (BP), and cellular components (CC). Kyoto Encyclopedia of Genes and Genomes (KEGG) Enrichment Analysis is a practical resource for analytical study of gene functions and associated high-level genome functional

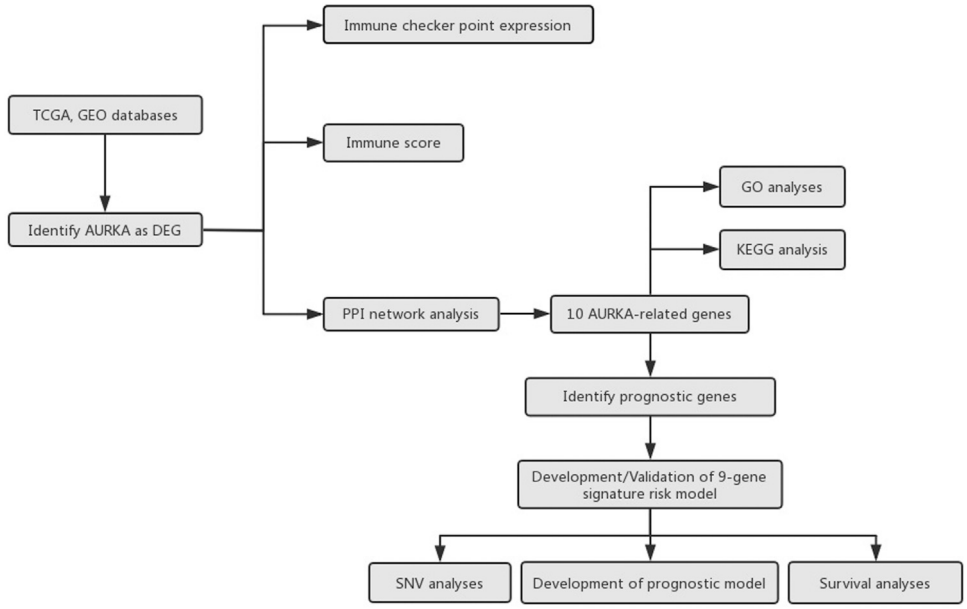

FIGURE 1 | Flow chart of the study.

information. ClusterProfiler package (version: 3.18.0) in R was employed to analyze the GO function of potential targets and enrich the KEGG pathway. $p < 0.05$ was set as the cut-off criterion.

Identification of Prognostic Genes

Gene Expression Profiling Interactive Analysis (GEPIA) is an interactive web application based on The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression databases (Tang et al., 2017). We used the Gene Expression Profiling Interactive Analysis to screen the prognostic value of AURKA and AURKA-related genes. The identification of prognostic genes was based on the following criteria: *1*) significant differences in gene expression levels between normal liver samples and liver tumor samples and *2*) significant association of the gene with both overall survival (OS) and disease-free survival (DFS) of patients with liver hepatocellular carcinoma (LIHC). *p*-values <0.05 denoted statistically significant differences.

Development and Validation of the Gene Signature

Primary screening of the LIHC data from TCGA database was performed for missing data. After deleting the samples with missing data, The available LIHC data from the TCGA database (340 samples) was divided into two subsets: a training set ($n = 240$) and a validation set ($n = 100$) randomly using the random function of Microsoft Excel program. The training set was used to train a predictive model and the test set was applied for validation. Another database International Cancer Genome Consortium (ICGC) (dcc.icgc.org) was also used as validation set.

Identified prognostic genes were submitted to the multivariate Cox regression model to calculate each prognostic gene's coefficient and risk scores. We used X-tile plot (version 3.6.1, http://www.tissuearray.org/rimmlab) to determine the optimum cutoff of AURKA-related gene signature risk score. X-tile which could calculate the best cut-point of sub-populations, is a software developing by team Rimm Laboratory from Yale University (Camp et al., 2004). A train set and two validation sets were divided into three sub-groups using the same best cut-off value. Risk score analysis, including risk score distribution, survival status, and gene expression heatmap among sub-groups were performed. Kaplan–Meier curves were analyzed using Kruskal-Wallis test and visualized through GraphPad Prism (version 8.0). Receiver operating characteristic (ROC) curve analysis was performed to assess the predictive value of AURKA-related gene signature. The clinical characteristics of three cohorts were shown in **Table 1**. The correlation among clinicopathological characteristics and risk groups were analyzed by the chi-square test. For sample sizes of less than 40 or theoretical frequencies (T) of less than 1, Fisher exact probability method was used.

Distribution of Somatic Mutations

To identify the somatic mutations of patients with LIHC, we downloaded single-nucleotide polymorphism (SNV) data and clinical follow-up information from TCGA database. The downloaded single-nucleotide polymorphism data were organized in the multiple alignment (MAF) format and visualized using the "maftools" package in R software. The horizontal histogram showed the genes with the highest frequency of mutation.

TABLE 1 | Clinical characteristics of HCC cohorts.

Clinical features		Training set (n = 240)	Validation set (n = 100)	ICGC-LIRI-JP (n = 243)
Age	<50	45	21	15
	≥50	195	79	228
Gender	Female	80	28	61
	Male	160	72	182
T	Early (T1+T2)	176	76	-
	Late(T3+T4)	63	23	-
	Unknown	1	1	-
N	N0	172	67	-
	N1+N2	2	1	-
	Unknown	66	32	-
M	M0	171	73	-
	M1	1	2	-
	Unknown	68	25	-
Stage	Stage I/II	166	71	146
	Stage III/IV	60	23	97
	Unknown	14	6	0
Grade	G1	38	14	-
	G2	114	47	-
	G3	76	35	-
	G4	9	3	-
	Unknown	3	1	-
Recurrence	yes	134	44	-
	no	106	56	-
RiskGroup	low risk	135	49	199
	mid risk	78	40	37
	high risk	27	11	7
Status	Alive	148	69	199
	Dead	92	31	44

Immune Infiltration Analysis

To explore the associations between different subgroups and immune cells infiltration, we employed Tumor Immune Estimation Resource (TIMER), which is a useful resource for comprehensive analysis of tumor-infiltrating immune cells (Li et al., 2017). The infiltration of six type of immune cell, including B cell, Macrophage, Myeloid dendritic cell, Neutrophil, T cell $CD4^+$ and T cell $CD8^+$, were calculated. SIGLEC15, IDO1, CD274, HAVCR2, PDCD1, CTLA4, LAG3 and PDCD1LG2 were selected to be immune checkpoints and the expression values of these eight immune checkpoints among sub-groups were explored. Differences between the three groups were assessed using the Kruskal-Wallis test. $p < 0.05$ was considered statistically significant. All the above analysis methods and R package were implemented by R foundation for statistical computing (2020) version 4.0.3 and software packages ggplot2 and pheatmap.

Construction and Validation of a Prognostic Model

A nomogram model was constructed to predict the probability of survival at 3 and 5 years for liver cancer patients. Briefly, the prognostic value of the clinicopathological characteristics for OS was estimated through univariate and multivariate Cox regression analyses in both training set and validation set. The performance of the risk model was validated by internal validation and external validation. Internal validation was performed by bootstrap Cox proportional regression analysis based on 1,000 bootstrap samples. Validation set was conducted based on another HCC patients from the TCGA database. Those parameters with p-values <0.05 in both training set and validation set were identified as potential prognostic factors, which were were included in multivariate Cox regression model and visualized using R package "rms." The Calibration curves were plotted to analyze the diagnostic performance of the nomogram. The ROC curve were conducted to determine the clinical value of the nomogram.

RESULTS

AURKA mRNA Level in HCC Samples

The results of the DEG analysis are shown in **Figure 2**. In TCGA database, a total of 421 HCC samples were selected for the DEG analysis. The analysis showed that the AURKA gene was upregulated ($\log_2$ fold change >1.5 and adjusted p-value <0.05), which indicated that the mRNA expression levels of AURKA differ significantly between normal liver tissue and liver cancer tissue. The mRNA expression levels of AURKA were also determined using three Gene Expression Omnibus data sets (GSE14323, n = 115; GSE14520, n = 445; and GSE25097, n = 557). The results also showed that AURKA was significantly highly expressed in HCC.

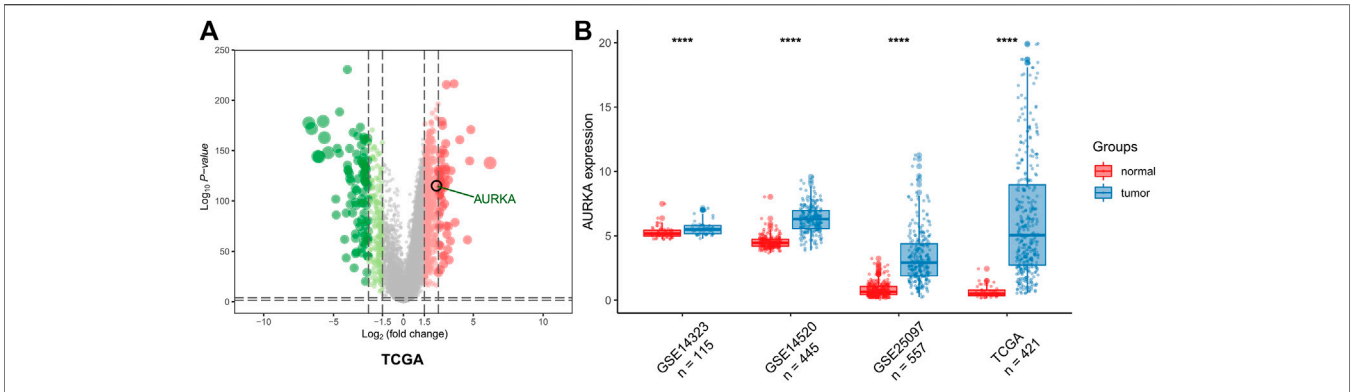

FIGURE 2 | AURKA expression. **(A)** Volcano plot of all DEGs. AURKA gene was identified. **(B)** Histogram of GSE14323, GSE14520 and GSE25097.

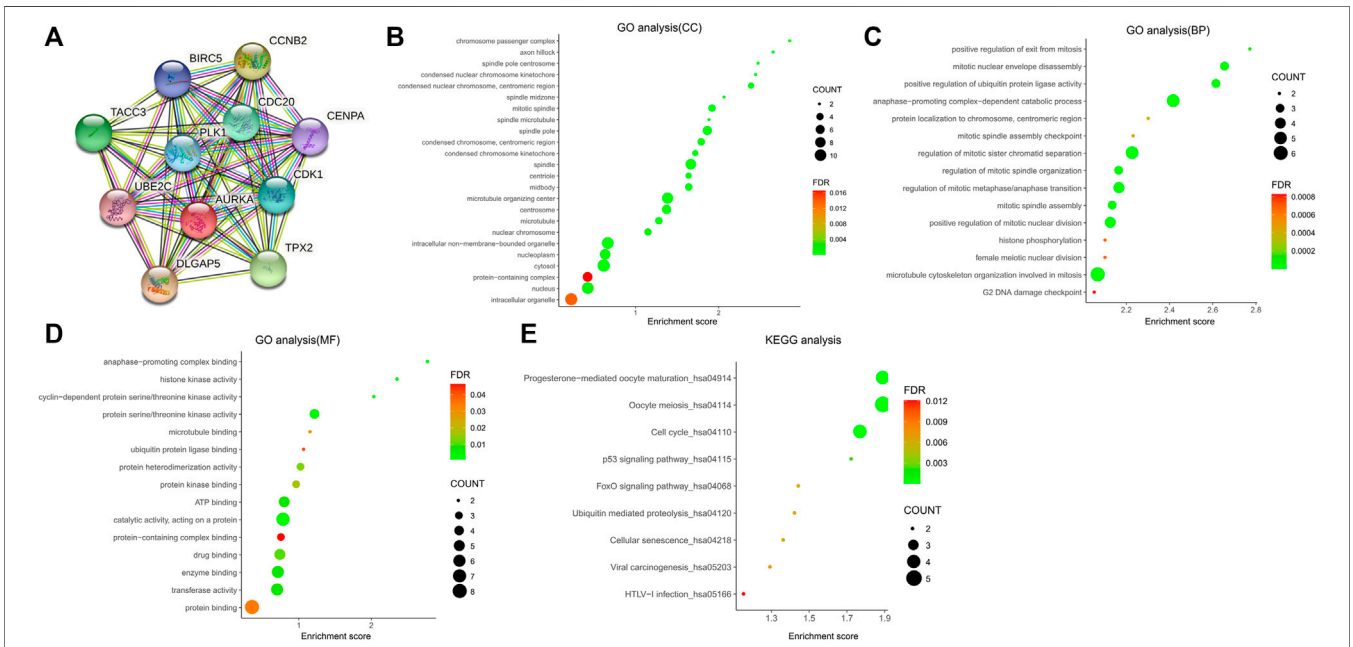

FIGURE 3 | PPI network of AURKA and AURKA-related genes. **(A–D)** GO analysis of AURKA and AURKA-related genes. **(E)** KEGG pathway analysis of AURKA and AURKA-related genes.

PPI Network Analysis

According to the predictive results of the Search Tool for the Retrieval of Interacting Genes (STRING) database, another ten genes were identified as AURKA-related genes with significant interaction, namely targeting protein for Xklp2 (TPX2), cyclin dependent kinase 1 (CDK1), polo like kinase 1 (PLK1), DLG associated protein 5 (DLGAP5), cell division cycle 20 (CDC20), baculoviral IAP repeat containing 5 (BIRC5), transforming acidic coiled-coil containing protein 3 (TACC3), centromere protein A (CENPA), cyclin B2 (CCNB2), and ubiquitin conjugating enzyme E2 C (UBE2C). The protein-protein interaction (PPI) network of AURKA and AURKA-related genes was constructed and visualized using the online STRING database (**Figure 3A**).

The Gene Ontology (GO) enrichment analysis was composed of three parts: GO biological process (GO-BP), GO cellular component (GO-CC), and GO molecular function (GO-MF). In GO-CC (**Figure 3B**), these genes were enriched in condensed nuclear chromosome, centromeric region, mitotic spindle, and spindle pole. In GO-BP (**Figure 3C**), AURKA and its related genes were significantly enriched in mitotic nuclear envelope disassembly, positive regulation of ubiquitin protein ligase activity, and anaphase-promoting complex-dependent catabolic process. In GO-MF (**Figure 3D**), genes were mainly enriched in protein serine/threonine kinase activity, protein heterodimerization activity, and protein kinase binding. For the Kyoto Encyclopedia of Genes and Genomes pathway analysis, nine pathways (**Figure 3E**) were observed, namely

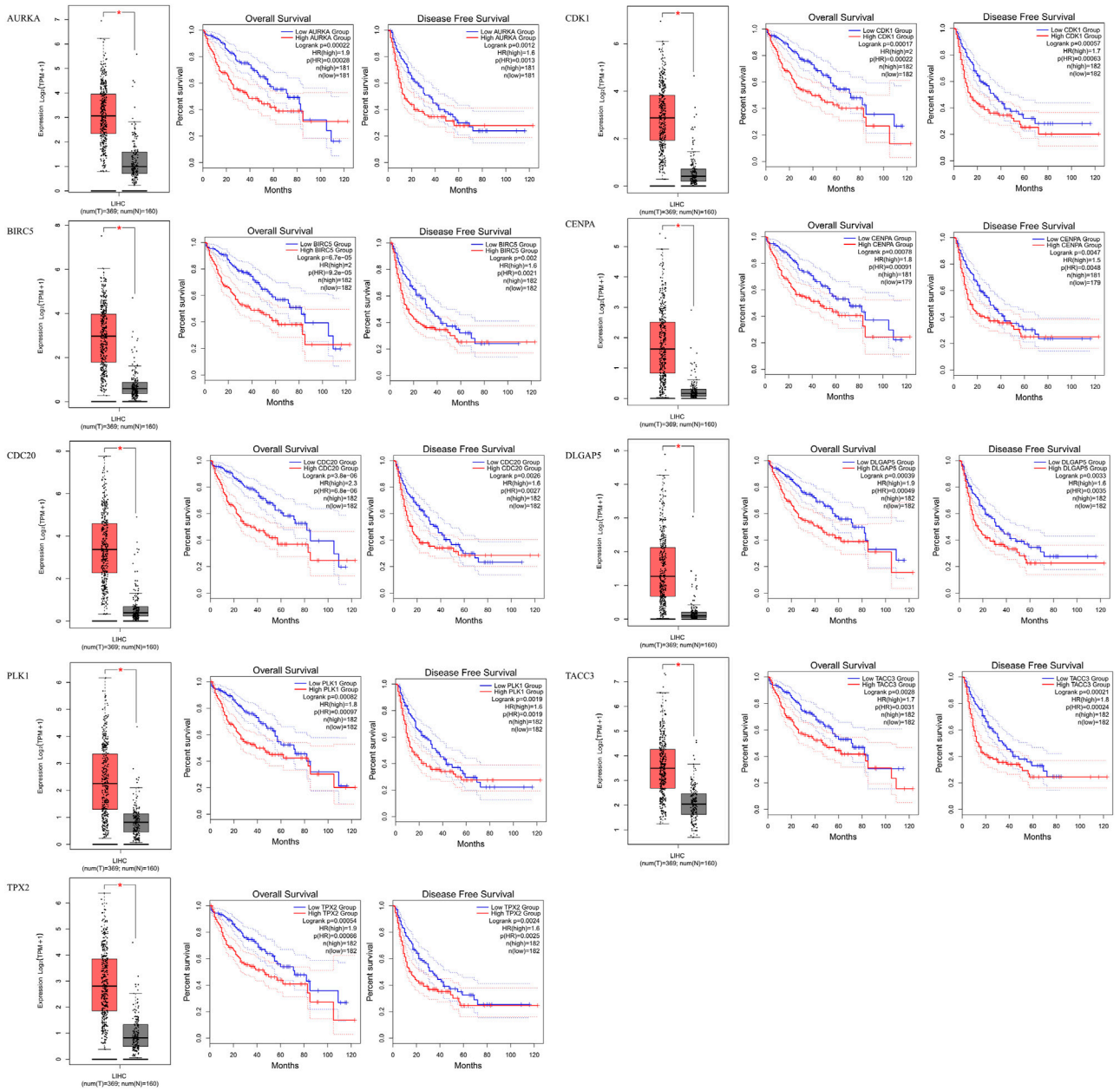

FIGURE 4 | Identification of prognostic genes. AURKA, BIRC5, CDC20, PLK1, TPX2, CDK1, CENPA, DLGAP5 and TACC3 gene expression levels were significantly higher in HCC patients, and these nine genes are related to overall survival and disease-free survival in HCC.

progesterone-mediated oocyte maturation, oocyte meiosis, cell cycle, TP53 signaling pathway, forkhead box O (FOXO) signaling pathway, ubiquitin-mediated proteolysis, cellular senescence, viral carcinogenesis, and human T-lymphotropic virus type I (HTLV–I) infection.

Development and Validation of a Nine-Gene Signature

The results of the identification of prognostic genes are shown in **Figure 4**. According to the screening strategy and criteria described above, two AURKA-related genes, namely CCNB2 (OS Log-rank $p > 0.05$) and UBE2C (OS Log-rank $p > 0.05$), were excluded. Eight other AURKA-related genes were significantly highly expressed in HCC tissue compared with normal liver tissue and correlated with prognosis.

The AURKA and the other eight AURKA-related genes were used to construct a risk model. The risk score was calculated through multivariate regression analyses. The cutoff value of the risk score was identified using the X-tile software. According to the results, the HCC samples were divided into three subgroups. The cutoff values of the

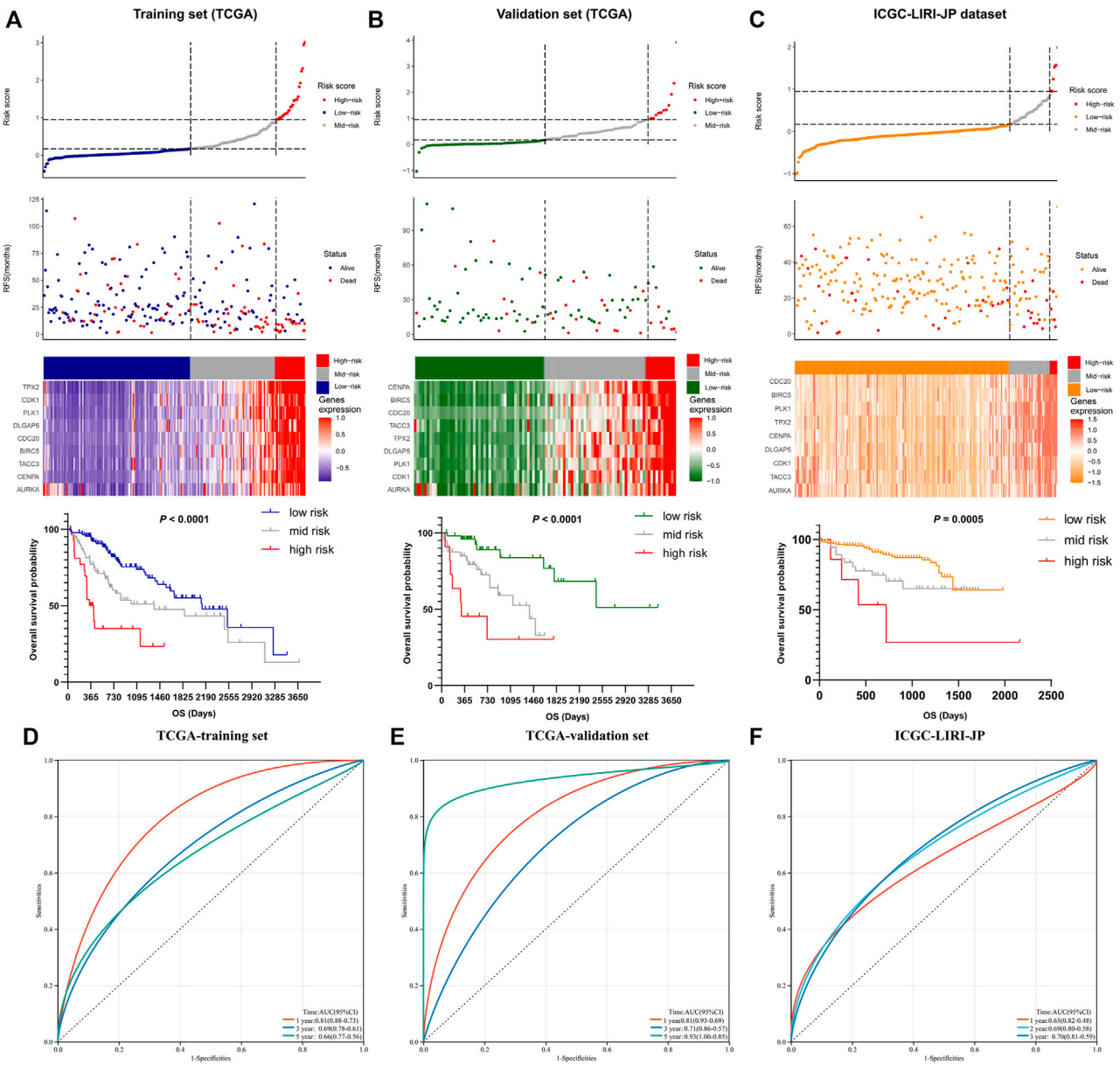

FIGURE 5 | The prognostic significance of the nine-gene risk scoring model. **(A)** The distribution of risk scores and survival status of liver cancer patients in the training set. HCC Patients with a higher risk score have a lower overall survival rate. **(B,C)** The distribution of risk scores and survival status of liver cancer patients in validation set. A higher risk score in HCC patients also have a lower overall survival rate. **(D–F)** The ROC curves of the training set and two validation sets.

risk score were 0.17 and 0.95. HCC samples with a risk score <0.17 and >0.95 classified into the low- and high-risk groups, respectively. The remaining samples were assigned to the moderate-risk group.

The details of the risk model showed in **Figure 5**, which revealed that higher risk scores were associated with higher expression levels of the nine-gene signature. Furthermore, higher risk scores also indicated worse OS. These results were similar in both the training and other two validation sets. Collectively, these results suggest that the risk model had potential value in predicting the prognosis of HCC. Furthermore, as the ROC curves showed, the area under the ROC curve (AUC) of the training set and two validation sets were higher than 0.5, which indicated that AURKA-related gene signature risk model had important value in predicting prognosis (**Figures 5D–F**).

The clinicopathological characteristics among the risk groups in training set were shown in **Table 2**. Age, T stage, pathological TNM (pTNM) stage, grade, recurrence, and survival endpoint were significantly different between the three risk groups ($p < 0.05$).

TABLE 2 | Correlation among risk groups with clinical features in training set.

Clinical features		Risk group			*p* value
		Low	Mid	High	
Age	<50	23	12	10	0.0339
	≥50	112	66	17	
Gender	Female	42	29	9	0.6639
	Male	93	49	18	
T	Early (T1+T2)	113	52	11	<0.001
	Late (T3+T4)	21	26	16	
N	N0	90	59	23	1
	N1+N2	1	1	0	
M	M0	90	58	23	1
	M1	1	0	0	
Stage	Stage I/II	106	49	11	<0.001
	Stage III/IV	21	24	15	
Grade	G1+G2	98	45	9	0.0002
	G3+G4	36	31	18	
Recurrence	Yes	72	23	11	0.0031
	No	63	55	16	
Status	Alive	96	42	10	0.0009
	Dead	39	36	17	

Somatic Mutation Results

Figure 6 illustrates the somatic landscape of the three risk subgroups. Information on the mutation status of each gene in each sample was shown in the waterfall plot, where different colors with specific annotations at the bottom indicated the various types of mutation. The barplot above the legend exhibited the number of mutations. The results showed that CTNNB1 was the most commonly mutated gene in the low-risk group. Tumor protein p53 (TP53) was the most frequently mutated gene in both the moderate- and high-risk groups. Hence, we further grouped the HCC samples into two groups based on the TP53 mutation status. A total of 101 and 241 HCC samples were assigned to the TP53 mutant- and wild-type groups, respectively. The survival analyses of both the TP53 mutant- and wild type cohorts yielded similar results. Higher risk scores were associated with worse prognostic outcome ($p < 0.05$).

Association Between Risk Score and Immune Infiltration

We further investigated differences in the degree of immune infiltration in various risk groups. Using the TIMER database, we evaluated the immune cell infiltration in samples from the three aforementioned groups. As shown in **Figure 7**, we found that the level of immune cell infiltration (macrophages, myeloid dendritic cells, neutrophils, and CD4T cells) was significantly different between the three groups. To further investigate the levels of immune cell infiltration on the gene level, the following eight immune-related genes were selected: sialic acid binding Ig like lectin 15 (SIGLEC15), T cell immunoreceptor with Ig and ITIM domains (TIGIT), CD274, hepatitis A virus cellular receptor 2 (HAVCR2), programmed cell death 1 (PDCD1), cytotoxic T-lymphocyte associated protein 4 (CTLA4), lymphocyte activating 3 (LAG3), and programmed cell death 1 ligand 2 (PDCD1LG2). We compared the expression levels of immune-related genes among the three risk groups. The results are shown in **Figure 8**. Except for CTLA4, LAG3, and PDCD1, the expression levels of the other five immune-related genes differed significantly between the three groups. Overall, the high-risk group showed significantly higher levels of immune gene expression and immune cell infiltration compared with the other groups.

Construction and Evaluation of the Nomogram Model

As shown in **Table 3**, we performed the univariate and multivariate analyses using the SPSS software (version 23.0; IBM Corporation, Armonk, NY, United States) to identify independent prognostic factors predicting OS in patients with HCC. In the training data set, the risk score, T stage, M stage, pTNM stage, and recurrence were identified as independent prognostic factors. The results of the testing and training data sets were similar. Overall, the risk score, T stage, and pTNM stage were identified as independent prognostic factors in both data sets.

Hence, we used age, T stage, N stage, M stage, pTNM stage, recurrence, and risk group as estimated factors in the construction of our model. A nomogram was constructed to estimate the probabilities for three- and 5-year survival. Calibration curves analysis showed that the new nomogram model had good predictive accuracy (**Figure 9B**). Furthermore, the new nomogram model achieved an area under the curve (AUC) of 0.78 at 5 years, which was better than that of a model without the gene signature (AUC = 0.73) or pTNM stage (AUC = 0.70) (**Figure 9C**).

DISCUSSION

The prognosis of patients with HCC varies greatly. The 5-year survival rate after resection of early HCC can be as high as 70%, whereas that of patients with vascular invasion or advanced HCC is markedly lower (Mazzaferro et al., 2011; Grandhi et al., 2016). Therefore, the use of effective prognostic indicators can promptly identify high-risk patients and assist in implementing individualized treatment to improve the prognosis.

In the present study, we found that the AURKA gene was highly expressed in HCC and an independent prognostic risk factor. The GO analysis of AURKA and related genes indicated that these genes are involved in a variety of biological process, including nuclear chromosome condensation, centromeric regions, mitotic spindle and spindle poles, mitotic nuclear envelope disassembly, and ubiquitin protein ligase activity. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis showed that AURKA and related genes regulate the progression of HCC through multiple pathways, such as the TP53 signaling pathway and FOXO signaling pathway. Numerous studies have demonstrated that the TP53 signaling pathway is involved in the development of a variety of tumors and plays a regulatory role in tumor immunity (Stegh, 2012; Blagih et al., 2020; Muñoz-Fontela et al., 2016). The role of the FOXO

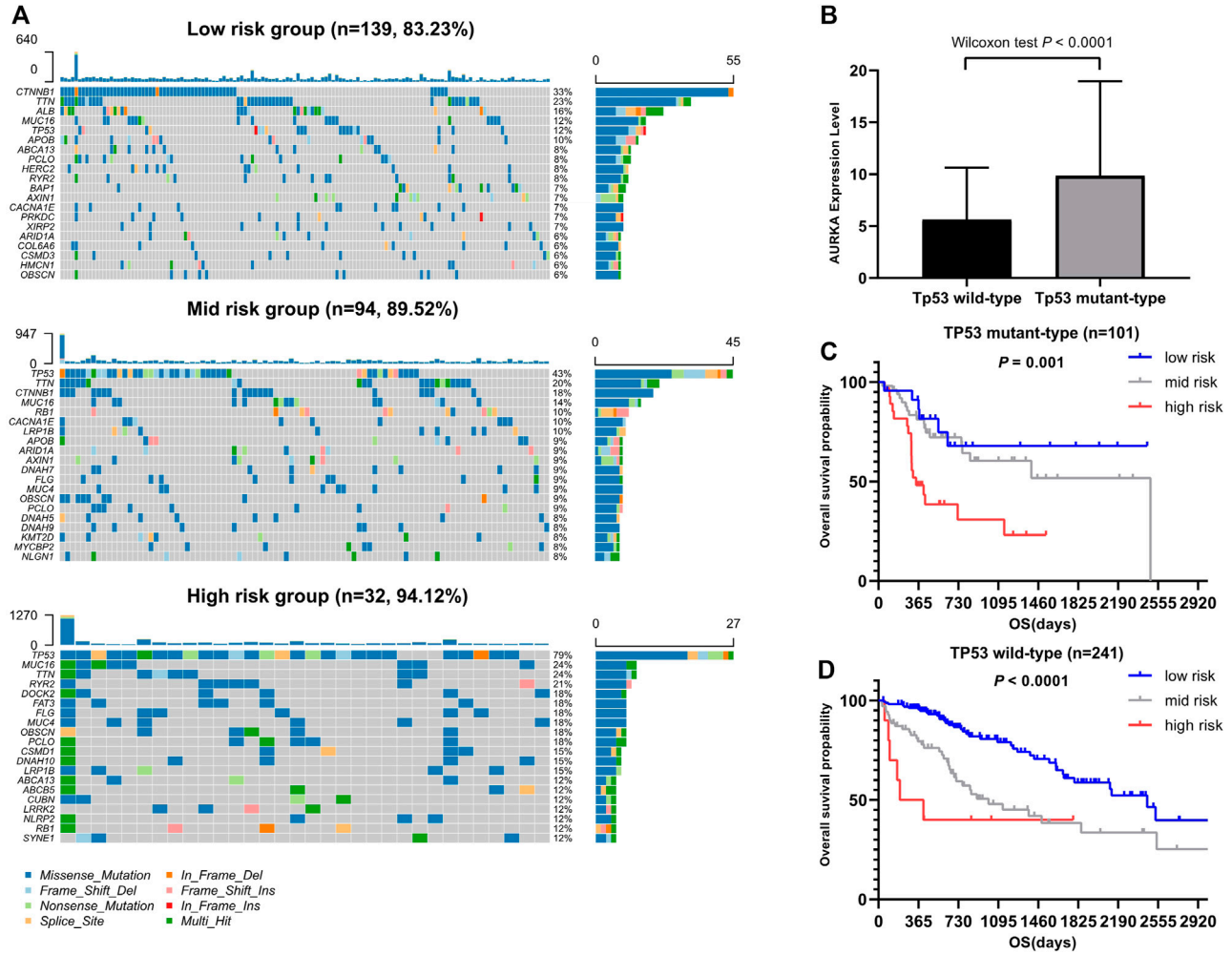

FIGURE 6 | Association between somatic mutation and risk score. **(A)** Overview of somatic mutations in different risk groups. **(B)** Comparison of AURKA expression between TP53 mutation group and Tp53 wild group. **(C)** Kaplan-Meier survival analysis of the different risk groups of TP53 mutation. **(D)** Kaplan-Meier survival analysis of the different risk groups of Tp53 wild.

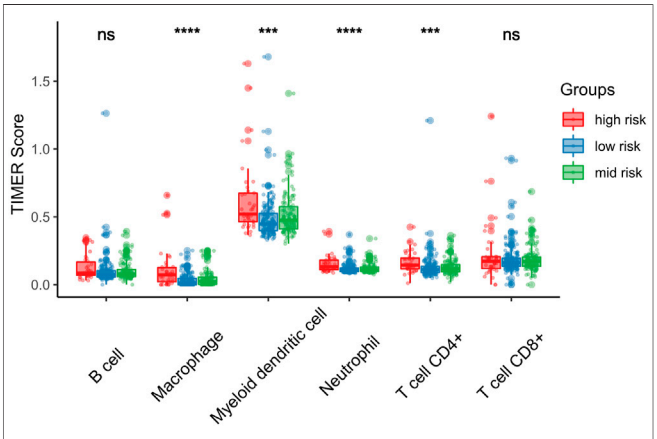

FIGURE 7 | Association between risk score and immune infiltration. Macrophage, Myeloid dendritic cell, Neutrophil and CD4+T cell are significantly high in high risk score groups. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

signaling pathway in tumors has also received extensive attention (Farhan et al., 2017; Coomans de Brachène and Demoulin, 2016). In general, the enrichment analysis revealed some potential mechanisms and possible pathways of AURKA and its related genes in HCC. In addition, it provided some new ideas for the treatment of patients with HCC.

Among the nine genes associated with AURKA, eight genes (BIRC5, CDC20, PLK1, TPX2, CDK1, CENPA, DLGAP5, and TACC3) were identified as prognostic genes in HCC. Therefore, we used these nine genes to construct the AURKA-related gene signature for the prediction of prognosis of patients with liver cancer. In gastric cancer, high expression of BIRC5 promotes gastric cancer metastasis and is associated with poor prognosis (Zou et al., 2019). However, in lung cancer, high expression of BIRC5 may prolong OS and DFS (Vischioni et al., 2004). In pancreatic cancer, CDC20 can promote tumor cell proliferation and affect the progression of pancreatic cancer (Chang et al., 2012). CDK1 is considered a synthetic target for KRAS-mutated

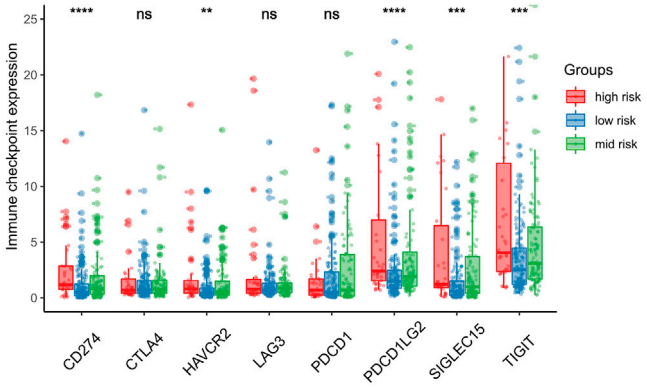

FIGURE 8 | Association between risk score and immune checkpoint. SIGLEC15, TIGIT, CD274, HAVCR2 and PDCD1LG2 expression mainly expressed in the high-risk score group. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

tumors and has been identified as a prognostic marker in numerous types of cancer (Sung et al., 2014; Costa-Cabral et al., 2016). CENPA plays a key role in cell mitosis. Of note, it is abnormally expressed in tumors and regulates tumor cell activity (Valdivia et al., 2009). DLGAP5 expression is regulated by the ubiquitin-proteasome pathway and participates in tumor cell migration and invasion (Hsu et al., 2004). PLK1 is a key regulator of mitosis and is involved in multiple stages of mitosis (De et al., 2018). Downregulation of PLK1 can inhibit the invasion and metastasis of esophageal cancer cells (Li et al., 2014). In colorectal cancer, PLK1 may promote the growth, invasion, and metastasis of colorectal cancer cells through the PDKI-PLK1-MYC signaling pathway (Tan et al., 2013). TACC3 can activate the Akt/RAS/mitogen-activated protein kinase kinase/extracellular signal–regulated kinase (Akt/RAS/MEK/ERK) signaling pathway to promote the malignant transformation of cells (Burgess et al., 2018). High TACC3 expression has been found in a variety of tumors and is closely related to poor prognosis (Wang et al., 2017). TPX2 is mainly involved in centrosomal maturation and spindle formation (Gruss and Vernos, 2004). In gliomas, MiR-1294 can target TPX2 to inhibit tumor cell proliferation and enhance sensitivity to chemotherapy (Chen et al., 2018). Several previous studies have successfully constructed multi-gene signatures for risk stratification and prognosis prediction in HCC (Zhou et al., 2020; Ouyang et al., 2020). In this study, we constructed a new nine-gene signature to predict the prognosis of patients with liver cancer. This gene signature was verified using an internal verification data set. Gene signatures can effectively classify patients into high-, moderate-, and low-risk groups. Higher risk scores in the training and validation data sets indicated a poor prognosis in HCC. Finally, we constructed a personalized nomogram based on the risk scores, with a concordance index of 0.78.

As an important tumor suppressor gene, TP53 plays a vital role in cell cycle regulation. TP53 mutation is a common mutation in tumors and the most important mutation in liver cancer. This mutation can promote the proliferation, migration, and invasion of tumor cells and increase resistance to drugs (Warren et al., 2013). We divided the HCC cohort of TCGA data set into two groups (TP53 mutation and wild type) and investigated the relationship between the gene signature and these two groups. We found that the gene signature could effectively predict the risk of patients in the TP53 mutation

TABLE 3 | Univariate and multivariate Cox regression analyses of risk factors associated with overall survival.

HCC cohorts		Univariable analyses			Multivariable analyses		
		p	HR	95.0% CI	*p*	HR	95.0% CI
Validation set	Age	0.418	1.012	0.984–1.041			
	Gender	0.149	0.592	0.29–1.207			
	T stage	0	4.017	1.925–8.38	0	8.257	2.168–31.451
	N stage	0.04	8.998	1.107–73.152	0.908	1.143	0.118–11.109
	M stage	0.088	3.602	0.825–15.724	0.259	4.284	0.343–53.561
	pTNM_stage	0	4.602	2.114–10.017	0.003	5.318	1.778–15.907
	grade	0.205	0.601	0.273–1.322			
	recurrence	0.635	1.193	0.575–2.476			
	Low risk group	0.001			0.019		
	Mid risk group	0.005	3.936	1.515–10.224	0.033	5.677	1.151–27.994
	High risk group	0	8.064	2.687–24.201	0.005	11.994	2.119–67.908
Training set							
	Age	0.399	1.007	0.991–1.023			
	Gender	0.65	0.906	0.592–1.387			
	T stage	0	2.428	1.595–3.697	0.028	1.968	1.078–3.594
	N stage	0.824	1.253	0.173–9.092	0.517	1.95	0.258–14.738
	M stage	0.012	13.563	1.763–104.315	0.003	26.515	2.979–236.021
	pTNM_stage	0	2.281	1.464–3.554	0	2.281	1.464–3.554
	grade	0.17	1.349	0.88–2.068			
	recurrence	0.016	1.754	1.111–2.769	0.048	1.967	1.005–3.851
	Low risk group	0			0		
	Mid risk group	0.009	1.842	1.169–2.905	0.084	1.854	0.92–3.734
	High risk group	0	4.85	2.702–8.706	0	5.262	2.311–11.981

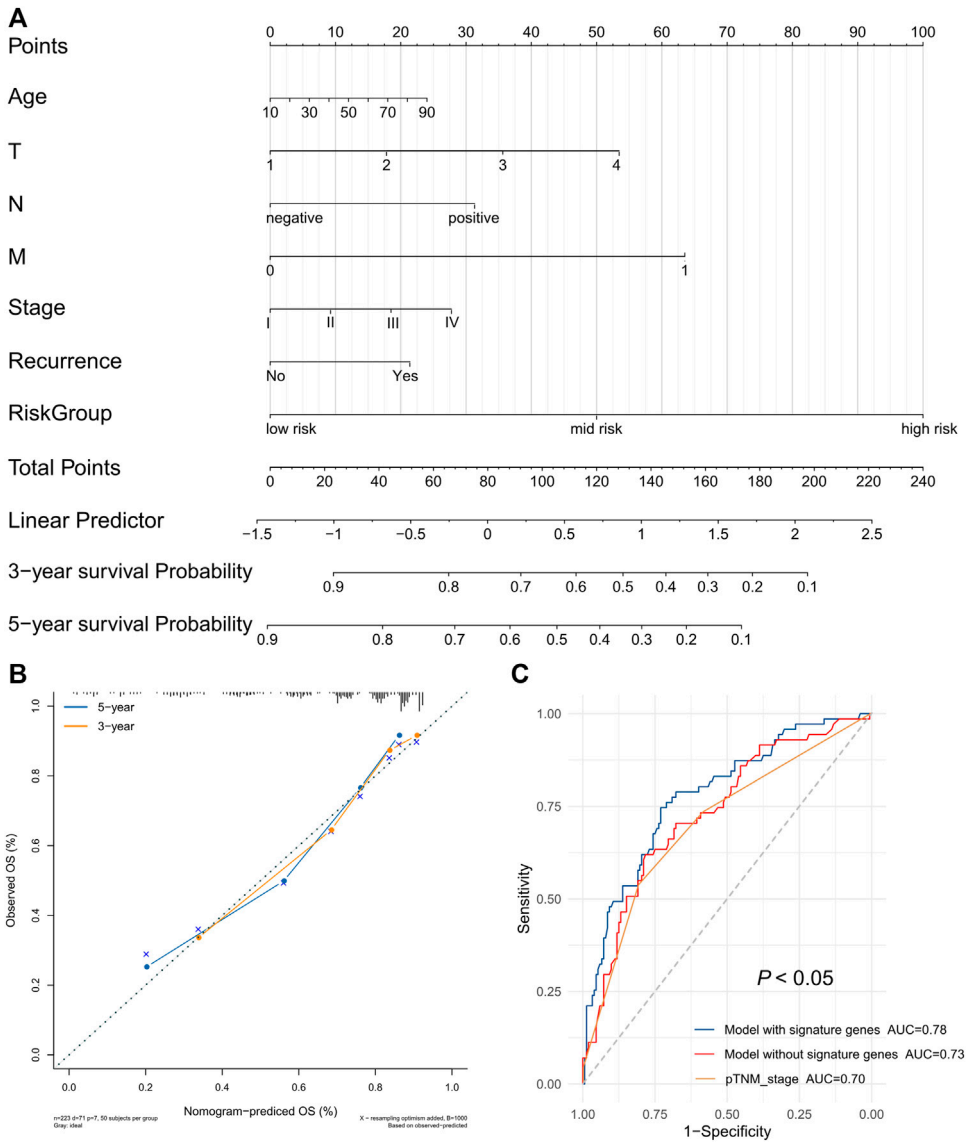

FIGURE 9 | The prognostic nomogram with the risk score in HCC. **(A)** A nomogram for predicting 3- and 5-year survival possibilities of HCC. **(B)** The calibration curve of 3-year and 5-year survival. **(C)** Time-dependent receiver operating characteristic (ROC) curves for gene signature and TNM stage.

group, thereby effectively classifying these patients into low-, moderate-, and high-risk groups. Numerous recent studies have also confirmed the close relationship between TP53 mutation and tumor immunity (Long et al., 2019; Wu et al., 2020; Sun et al., 2020). Based on the immunoprognostic model established by TP53 mutation, Long et al. found that the levels of T cell follicular helper proteins, T cell regulatory proteins, and macrophages M0 were higher in the high-risk HCC group versus the other groups (Long et al., 2019). Notably, the expression of immune checkpoints CTLA4, PDCD1, and T-cell immunoglobulin mucin family member 3 (TIM3) were also higher in the high-risk group. The investigators suggested that TP53 mutations significantly reduced the immune response in liver cancer.

Immune cell infiltration affects tumor progression. Numerous immunotherapies have been used to regulate immune cells in tumors. Therefore, we investigated the immune cell infiltration in different risk groups. We found that the numbers of macrophages, myeloid dendritic cells, neutrophils, and $CD4^+$ T cells differed significantly in different risk score groups. Higher risk scores were linked to higher numbers of these four types of immune cells. Macrophages play a dual role in the tumor microenvironment, promoting tumor formation and development as well as inhibiting tumor growth (Kim and Bae, 2016). It has been confirmed that the degree of macrophage infiltration in the tumor microenvironment is related to prognosis (Conway et al., 2016). It is currently thought that neutrophils in the tumor microenvironment directly kill or

stimulate the immune system to inhibit tumor cells and can also promote immune escape of tumor cells (Kim and Bae, 2016). Myeloid dendritic cells mainly play an antigen-presenting role to activate T cells and induce immune responses (Garris and Luke, 2020). $CD4^+$ T cells mainly support $CD8^+$ T toxic lymphocytes and enhance their anti-tumor immune effect (Ghiringhelli et al., 2006). Our results showed differences in the distribution of immune cells in the tumor microenvironment among the different risk groups and revealed the underlying reason for the poor prognosis observed in the high-risk group. To the best of our knowledge, this is the first study to investigate the relationship between AURKA and related genes and tumor immunity.

Immune checkpoint inhibitors are a new approach to the treatment of tumors. Several immune checkpoint inhibitors have been used effectively in the treatment of liver cancer (Liu and Qin, 2019). In this study, we also assessed the relationship between the risk score and immune checkpoints. The results found that the expression of the five immune checkpoints (SIGLEC15, TIGIT, CD274, HAVCR2, and PDCD1LG2) varied in different risk groups. Higher risk scores were associated with higher expression of the five immune checkpoints. These five immune checkpoints play an important role in the activation of T cells. PDCD1LG1 and PDCD1LG2 are the two ligands of PDCD1. In the tumor microenvironment, PDCD1 on the surface of immune cells binds to the PDCD1LG1 and PDCD1LG2 receptors on the surface of tumor cells to activate a series of signal factors in immune cells. This process initiates a series of signaling factors in immune cells to inhibit T cell activation and promote T cell failure, thus helping tumor cells to evade immunosurveillance (Barclay et al., 2018; Ai et al., 2020). SIGLEC15 is a newly discovered immune checkpoint. Wang et al. reported that high expression of SIGLEC15 in tumors can significantly inhibit the activity of T cells (Wang et al., 2019). Furthermore, the inhibition or knockout of SIGLEC15 expression can improve the anti-tumor ability of T cells in mice (Wang et al., 2019). HAVCR2 is thought to play a dual role, inducing immune tolerance and promoting tumor cell apoptosis (Das et al., 2017). TIGIT is mainly expressed on T cells and natural killer cells. Joller et al. observed significant T cell proliferation in TIGIT-knockout mice (Joller et al., 2011). The expression of these immune checkpoints significantly affects tumor prognosis. At present, the use of single immune checkpoint blockers or combinations of these agents has shown good efficacy in different tumors (Darvin et al., 2018).

This study is characterized by several limitations. Firstly, all analyses were based on public databases. The specific mechanisms of AURKA and related genes in liver cancer have not been thoroughly investigated. Furthermore, the gene signature was associated with immune infiltration and immune checkpoint expression in HCC and affected the prognosis of patients with this disease. Further studies are needed to examine the value of gene signatures in immune invasion and prognosis in HCC.

In summary, we found that nine AURKA-related genes with prognostic value can be used as prognostic markers for liver cancer. The gene signature based on AURKA successfully classified patients with liver cancer into high-, moderate- and low-risk groups. Hence, the gene signature can may be an effective marker for the prognosis of HCC. In addition, the risk score was related to immune cell infiltration and immune checkpoint expression in HCC.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

GW and TD designed the study; RL, ZJ, and WK equally contributed to the literature search, analysis, and writing of the manuscript. SZ supervised this work; All authors read and approved the final manuscript.

REFERENCES

Ai, L., Xu, A., and Xu, J. (2020). Roles of PD-1/pd-L1 Pathway: Signaling, Cancer, and Beyond. *Adv. Exp. Med. Biol.* 1248, 33–59. doi:10.1007/978-981-15-3266-5_3

Barclay, J., Creswell, J., and León, J. (2018). Cancer Immunotherapy and the PD-1/pd-L1 Checkpoint Pathway. *Arch. Esp. Urol.* 71, 393–399.

Blagih, J., Buck, M. D., and Vousden, K. H. (2020). p53, Cancer and the Immune Response. *J. Cel. Sci.* 133, jcs237453. doi:10.1242/jcs.237.45310.1242/jcs.237453

Briassouli, P., Chan, F., Savage, K., Reis-Filho, J. S., and Linardopoulos, S. (2007). Aurora-A Regulation of Nuclear Factor-Kb Signaling by Phosphorylation of IκBα. *Cancer Res.* 67, 1689–1695. doi:10.1158/0008-5472.CAN-06-2272

Burgess, S. G., Mukherjee, M., Sabir, S., Joseph, N., Gutiérrez-Caballero, C., Richards, M. W., et al. (2018). Mitotic Spindle Association of TACC3 Requires Aurora-A-dependent Stabilization of a Cryptic α-helix. *EMBO. J.* 37, e97902. doi:10.15252/embj.201797902

Camp, R. L., Dolled-Filhart, M., and Rimm, D. L. (2004). X-tile. *Clin. Cancer Res.* 10, 7252–7259. doi:10.1158/1078-0432.CCR-04-0713

Carmena, M., and Earnshaw, W. C. (2003). The Cellular Geography of Aurora Kinases. *Nat. Rev. Mol. Cell. Biol.* 4, 842–854. doi:10.1038/nrm1245

Chang, D. Z., Ma, Y., Ji, B., Liu, Y., Hwu, P., Abbruzzese, J. L., et al. (2012). Increased CDC20 Expression Is Associated with Pancreatic Ductal Adenocarcinoma Differentiation and Progression. *J. Hematol. Oncol.* 5, 15. doi:10.1186/1756-8722-5-15

Chen, C., Song, G., Xiang, J., Zhang, H., Zhao, S., and Zhan, Y. (2017). AURKA Promotes Cancer Metastasis by Regulating Epithelial-Mesenchymal Transition and Cancer Stem Cell Properties in Hepatocellular Carcinoma. *Biochem. Biophysical Res. Commun.* 486, 514–520. doi:10.1016/j.bbrc.2017.03.075

Chen, H., Liu, L., Li, X., Shi, Y., and Liu, N. (2018). MicroRNA-1294 Inhibits the Proliferation and Enhances the Chemosensitivity of Glioma to Temozolomide via the Direct Targeting of TPX2. *Am. J. Cancer Res.* 8, 291–301.

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer Statistics in China, 2015. *CA: A Cancer J. Clinicians* 66, 115–132. doi:10.3322/caac.21338

Conway, E. M., Pikor, L. A., Kung, S. H. Y., Hamilton, M. J., Lam, S., Lam, W. L., et al. (2016). Macrophages, Inflammation, and Lung Cancer. *Am. J. Respir. Crit. Care Med.* 193, 116–130. doi:10.1164/rccm.201508-1545CI

Coomans de Brachène, A., and Demoulin, J.-B. (2016). FOXO Transcription Factors in Cancer Development and Therapy. *Cell. Mol. Life Sci.* 73, 1159–1172. doi:10.1007/s00018-015-2112-y

Costa-Cabral, S., Brough, R., Konde, A., Aarts, M., Campbell, J., Marinari, E., et al. (2016). CDK1 Is a Synthetic Lethal Target for KRAS Mutant Tumours. *Plos. One.* 11, e0149099. doi:10.1371/journal.pone.0149099

Darvin, P., Toor, S. M., Sasidharan Nair, V., and Elkord, E. (2018). Immune Checkpoint Inhibitors: Recent Progress and Potential Biomarkers. *Exp. Mol. Med.* 50, 1–11. doi:10.1038/s12276-018-0191-1

Das, M., Zhu, C., and Kuchroo, V. K. (2017). Tim-3 and its Role in Regulating Anti-tumor Immunity. *Immunol. Rev.* 276, 97–111. doi:10.1111/imr.12520

de Cárcer, G., Venkateswaran, S. V., Salgueiro, L., El Bakkali, A., Somogyi, K., Rowald, K., et al. (2018). Plk1 Overexpression Induces Chromosomal Instability and Suppresses Tumor Development. *Nat. Commun.* 9, 3012. doi:10.1038/s41467-018-05429-5

Farhan, M., Wang, H., Gaur, U., Little, P. J., Xu, J., and Zheng, W. (2017). FOXO Signaling Pathways as Therapeutic Targets in Cancer. *Int. J. Biol. Sci.* 13, 815–827. doi:10.7150/ijbs.20052

Forner, A., Reig, M., and Bruix, J. (2018). Hepatocellular Carcinoma. *The Lancet* 391, 1301–1314. doi:10.1016/S0140-6736(18)30010-2

Garris, C. S., and Luke, J. J. (2020). Dendritic Cells, the T-Cell-Inflamed Tumor Microenvironment, and Immunotherapy Treatment Response. *Clin. Cancer Res.* 26, 3901–3907. doi:10.1158/1078-0432.CCR-19-1321

Ghiringhelli, F., Ménard, C., Martin, F., and Zitvogel, L. (2006). The Role of Regulatory T Cells in the Control of Natural Killer Cells: Relevance During Tumor Progression. *Immunol. Rev.* 214, 229–238. doi:10.1111/j.1600-065X.2006.00445.x

Grandhi, M. S., Kim, A. K., Ronnekleiv-Kelly, S. M., Kamel, I. R., Ghasebeh, M. A., and Pawlik, T. M. (2016). Hepatocellular Carcinoma: From Diagnosis to Treatment. *Surg. Oncol.* 25, 74–85. doi:10.1016/j.suronc.2016.03.002

Gruss, O. J., and Vernos, I. (2004). The Mechanism of Spindle Assembly. *J. Cel. Biol.* 166, 949–955. doi:10.1083/jcb.200312112

Hartke, J., Johnson, M., and Ghabril, M. (2017). The Diagnosis and Treatment of Hepatocellular Carcinoma. *Semin. Diagn. Pathol.* 34, 153–159. doi:10.1053/j.semdp.2016.12.011

Hsu, J.-M., Lee, Y.-C. G., Yu, C.-T. R., and Huang, C.-Y. F. (2004). Fbx7 Functions in the SCF Complex Regulating Cdk1-Cyclin B-Phosphorylated Hepatoma Up-Regulated Protein (HURP) Proteolysis by a Proline-Rich Region. *J. Biol. Chem.* 279, 32592–32602. doi:10.1074/jbc.M404950200

Jeng, Y.-M., Peng, S.-Y., Lin, C.-Y., and Hsu, H.-C. (2004). Overexpression and Amplification of Aurora-A in Hepatocellular Carcinoma. *Clin. Cancer Res.* 10, 2065–2071. doi:10.1158/1078-0432.ccr-1057-03

Joller, N., Hafler, J. P., Brynedal, B., Kassam, N., Spoerl, S., Levin, S. D., et al. (2011). Cutting Edge: TIGIT Has T Cell-Intrinsic Inhibitory Functions. *J.I.* 186, 1338–1342. doi:10.4049/jimmunol.1003081

Katayama, H., Sasai, K., Kawai, H., Yuan, Z.-M., Bondaruk, J., Suzuki, F., et al. (2004). Phosphorylation by aurora Kinase A Induces Mdm2-Mediated Destabilization and Inhibition of P53. *Nat. Genet.* 36, 55–62. doi:10.1038/ng1279

Kim, J., and Bae, J.-S. (2016). Tumor-Associated Macrophages and Neutrophils in Tumor Microenvironment. *Mediators Inflamm.* 2016, 1–11. doi:10.1155/2016/6058147

Li, C., Chen, J., Zhang, K., Feng, B., Wang, R., and Chen, L. (2015). Progress and Prospects of Long Noncoding RNAs (lncRNAs) in Hepatocellular Carcinoma. *Cell. Physiol. Biochem.* 36, 423–434. doi:10.1159/000430109

Li, C., Zhou, X., Wang, Y., Jing, S., Yang, C., Sun, G., et al. (2014). miR-210 Regulates Esophageal Cancer Cell Proliferation by Inducing G2/M Phase Cell Cycle Arrest Through Targeting PLK1. *Mol. Med. Rep.* 10, 2099–2104. doi:10.3892/mmr.2014.2416

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer Res.* 77, e108–e110. doi:10.1158/0008-5472.CAN-17-0307

Lindon, C., Grant, R., and Min, M. (2016). Ubiquitin-Mediated Degradation of Aurora Kinases. *Front. Oncol.* 5, 307. doi:10.3389/fonc.2015.00307

Liu, X., and Qin, S. (2019). Immune Checkpoint Inhibitors in Hepatocellular Carcinoma: Opportunities and Challenges. *Oncol.* 24, S3–S10. doi:10.1634/theoncologist.2019-IO-S1-s01

Locy, H., de Mey, S., de Mey, W., De Ridder, M., Thielemans, K., and Maenhout, S. K. (2018). Immunomodulation of the Tumor Microenvironment: Turn Foe into Friend. *Front. Immunol.* 9, 2909. doi:10.3389/fimmu.2018.02909

Long, J., Wang, A., Bai, Y., Lin, J., Yang, X., Wang, D., et al. (2019). Development and Validation of a TP53-Associated Immune Prognostic Model for Hepatocellular Carcinoma. *EBioMedicine* 42, 363–374. doi:10.1016/j.ebiom.2019.03.022

Lu, L., Han, H., Tian, Y., Li, W., Zhang, J., Feng, M., et al. (2015). Aurora Kinase A Mediates C-Myc's Oncogenic Effects in Hepatocellular Carcinoma. *Mol. Carcinog.* 54, 1467–1479. doi:10.1002/mc.22223

Mazzaferro, V., Bhoori, S., Sposito, C., Bongini, M., Langer, M., Miceli, R., et al. (2011). Milan Criteria in Liver Transplantation for Hepatocellular Carcinoma: An Evidence-Based Analysis of 15 Years of Experience. *Liver .Transpl.* 17, S44–S57. doi:10.1002/lt.22365

Mering, C. V., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: A Database of Predicted Functional Associations Between Proteins. *Nucleic Acids Res.* 31, 258–261. doi:10.1093/nar/gkg034

Muñoz-Fontela, C., Mandinova, A., Aaronson, S. A., and Lee, S. W. (2016). Emerging Roles of P53 and Other Tumour-Suppressor Genes in Immune Regulation. *Nat. Rev. Immunol.* 16, 741–750. doi:10.1038/nri.2016.99

Ouyang, G., Yi, B., Pan, G., and Chen, X. (2020). A Robust Twelve-Gene Signature for Prognosis Prediction of Hepatocellular Carcinoma. *Cancer Cel. Int.* 20, 207. doi:10.1186/s12935-020-01294-9

Reiter, R., Gais, P., Jütting, U., Steuer-Vogt, M. K., Pickhard, A., Bink, K., et al. (2006). Aurora Kinase A Messenger RNA Overexpression Is Correlated with Tumor Progression and Shortened Survival in Head and Neck Squamous Cell Carcinoma. *Clin. Cancer Res.* 12, 5136–5141. doi:10.1158/1078-0432.CCR-05-1650

Stegh, A. H. (2012). Targeting the P53 Signaling Pathway in Cancer Therapy - The Promises, Challenges and Perils. *Expert Opin. Ther. Targets* 16, 67–83. doi:10.1517/14728222.2011.643299

Sun, H., Liu, S.-Y., Zhou, J.-Y., Xu, J.-T., Zhang, H.-K., Yan, H.-H., et al. (2020). Specific TP53 Subtype as Biomarker for Immune Checkpoint Inhibitors in Lung Adenocarcinoma. *EBioMedicine* 60, 102990. doi:10.1016/j.ebiom.2020.102990

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Sung, W.-W., Lin, Y.-M., Wu, P.-R., Yen, H.-H., Lai, H.-W., Su, T.-C., et al. (2014). High Nuclear/cytoplasmic Ratio of Cdk1 Expression Predicts Poor Prognosis in Colorectal Cancer Patients. *BMC. Cancer* 14, 951. doi:10.1186/1471-2407-14-951

Tan, J., Li, Z., Lee, P. L., Guan, P., Aau, M. Y., Lee, S. T., et al. (2013). PDK1 Signaling Toward PLK1-MYC Activation Confers Oncogenic Transformation, Tumor-Initiating Cell Activation, and Resistance to mTOR-Targeted Therapy. *Cancer Discov.* 3, 1156–1171. doi:10.1158/2159-8290.CD-12-0595

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: A Web Server for Cancer and normal Gene Expression Profiling and Interactive Analyses. *Nucleic Acids Res.* 45, W98–W102. doi:10.1093/nar/gkx247

Umstead, M., Xiong, J., Qi, Q., Du, Y., and Fu, H. (2017). Aurora Kinase A Interacts with H-Ras and Potentiates Ras-MAPK Signaling. *Oncotarget* 8, 28359–28372. doi:10.18632/oncotarget.15049

Valdivia, M., Hamdouch, K., Ortiz, M., and Astola, A. (2009). CENPA a Genomic Marker for Centromere Activity and Human Diseases. *Cg* 10, 326–335. doi:10.2174/138920209788920985

Vischioni, B., van der Valk, P., Span, S. W., Kruyt, F. A. E., Rodriguez, J. A., and Giaccone, G. (2004). Nuclear Localization of Survivin Is a Positive Prognostic Factor for Survival in Advanced Non-small-cell Lung Cancer. *Ann. Oncol.* 15, 1654–1660. doi:10.1093/annonc/mdh436

Wang, J., Du, S., Fan, W., Wang, P., Yang, W., and Yu, M. (2017). TACC3 as an Independent Prognostic Marker for Solid Tumors: A Systematic Review and Meta-Analysis. *Oncotarget* 8, 75516–75527. doi:10.18632/oncotarget.20466

Wang, J., Sun, J., Liu, L. N., Flies, D. B., Nie, X., Toki, M., et al. (2019). Siglec-15 as an Immune Suppressor and Potential Target for Normalization Cancer Immunotherapy. *Nat. Med.* 25, 656–666. doi:10.1038/s41591-019-0374-x

Warren, R. S., Atreya, C. E., Niedzwiecki, D., Weinberg, V. K., Donner, D. B., Mayer, R. J., et al. (2013). Association of TP53 Mutational Status and Gender with Survival after Adjuvant Treatment for Stage III colon Cancer: Results of

CALGB 89803. *Clin. Cancer Res.* 19, 5777–5787. doi:10.1158/1078-0432.CCR-13-0351

Wu, C., Lyu, J., Yang, E. J., Liu, Y., Zhang, B., and Shim, J. S. (2018). Targeting AURKA-Cdc25c Axis to Induce Synthetic Lethality in ARID1A-Deficient Colorectal Cancer Cells. *Nat. Commun.* 9, 3212. doi:10.1038/s41467-018-05694-4

Wu, X., Lv, D., Cai, C., Zhao, Z., Wang, M., Chen, W., et al. (2020). A TP53-Associated Immune Prognostic Signature for the Prediction of Overall Survival and Therapeutic Responses in Muscle-Invasive Bladder Cancer. *Front. Immunol.* 11, 590618. doi:10.3389/fimmu.2020.590618

Yan, M., Wang, C., He, B., Yang, M., Tong, M., Long, Z., et al. (2016). Aurora-A Kinase: A Potent Oncogene and Target for Cancer Therapy. *Med. Res. Rev.* 36, 1036–1079. doi:10.1002/med.21399

Zhang, J., Li, B., Yang, Q., Zhang, P., and Wang, H. (2015). Prognostic Value of Aurora Kinase A (AURKA) Expression Among Solid Tumor Patients: A Systematic Review and Meta-Analysis. *Jpn. J. Clin. Oncol.* 45, 629–636. doi:10.1093/jjco/hyv058

Zhang, K., Chen, J., Chen, D., Huang, J., Feng, B., Han, S., et al. (2014). Aurora-A Promotes Chemoresistance in Hepatocelluar Carcinoma by Targeting NF-kappaB/microRNA-21/PTEN Signaling Pathway. *Oncotarget* 5, 12916–12935. doi:10.18632/oncotarget.2682

Zhou, W., Zhang, S., Cai, Z., Gao, F., Deng, W., Wen, Y., et al. (2020). A Glycolysis-Related Gene Pairs Signature Predicts Prognosis in Patients with Hepatocellular Carcinoma. *Peer.J.* 8, e9944. doi:10.7717/peerj.9944

Zou, J., Liao, X., Zhang, J., and Wang, L. (2019). Dysregulation of miR-195-5p/-218-5p/BIRC5 Axis Predicts a Poor Prognosis in Patients with Gastric Cancer. *J. Biol. Regul. Homeost .Agents.* 33, 1377–1385. doi:10.23812/19-146-A

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 Genetics

ORIGINAL RESEARCH
published: 26 January 2022
doi: 10.3389/fgene.2022.765569

BGN May be a Potential Prognostic Biomarker and Associated With Immune Cell Enrichment of Gastric Cancer

*Shiyu Zhang, Huiying Yang, Xuelian Xiang, Li Liu, Huali Huang and Guodu Tang**

[1]*Department of Gastroenterology, The First Affiliated Hospital of Guangxi Medical University, Nanning, China*

Background: Biglycan (BGN) plays a role in the occurrence and progression of several malignant tumors, though its role in gastric cancer (GC) remains unclear. The objective of this study was to investigate BGN expression, its role in GC prognosis, and immune infiltration.

Material and Methods: Gene expression data and corresponding clinical information were downloaded from TCGA and GTEx, respectively. We compared the expression of BGN in GC and normal tissues and verified the differential expression via Real-Time PCR and immunohistochemistry. BGN-related differentially expressed genes (DEGs) were identified. Additionally, the relationships between BGN gene expression and clinicopathological variables and survival in patients with GC were also investigated through univariate and multivariate Cox regression analyses. Finally, we established a predictive model that could well predict the probability of 1-, 3-, and 5-years survival in GC.

Results: We found a significantly higher expression of BGN in GC than that in normal tissues ($p < 0.001$), which was verified by Real-Time PCR ($p < 0.01$) and immunohistochemistry ($p < 0.001$). The 492 identified DEGs were primarily enriched in pathways related to tumor genesis and metastasis, including extracellular matrix (ECM)-receptor interaction, focal adhesion pathway, Wnt signaling, and signaling by VEGF. BGN expression was positively correlated with the enrichment of the NK cells ($r = 0.620$, $p < 0.001$) and macrophages ($r = 0.550$, $p < 0.001$), but negatively correlated with the enrichment of Th17 cells ($r = 0.250$, $p < 0.001$). BGN expression was also significantly correlated with histologic grade (GI&G2 *vs.* G3, $p < 0.001$), histologic type (Diffuse type *vs.* Tubular type, $p < 0.001$), histologic stage (stage I *vs.* stage II and stage I *vs.* stage III, $p < 0.001$), T stage (T1 *vs.* T2, T1 *vs.* T3, and T1 *vs.* T4, $p < 0.001$) and *Helicobacter pylori* (HP) infection (yes *vs.* no, $p < 0.05$) in GC. High BGN expression showed significant association with poor overall survival (OS) in GC patients (HR = 1.53 (1.09–2.14), $p = 0.013$). The constructed nomogram can well predict the 1-, 3-, and 5-years overall survival probability of GC patients (C-index = 0.728).

Conclusion: BGN plays an important role in the occurrence and progression of GC and is a potential biomarker for the diagnosis and treatment of GC.

Keywords: biomarker, prognostic index, bioinformatics analysis, gastric cancer, immune infiltration

OPEN ACCESS

Edited by:
Jiangning Song,
Monash University, Australia

Reviewed by:
Poonam Gera,
Advanced Centre for Treatment, India
Shihori Tanabe,
National Institute of Health Sciences (NIHS), Japan

****Correspondence:***
Guodu Tang
tangguodu@stu.gxmu.edu.cn

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

Received: *27 August 2021*
Accepted: *10 January 2022*
Published: *26 January 2022*

Citation:
Zhang S, Yang H, Xiang X, Liu L, Huang H and Tang G (2022) BGN May be a Potential Prognostic Biomarker and Associated With Immune Cell Enrichment of Gastric Cancer. Front. Genet. 13:765569. doi: 10.3389/fgene.2022.765569

INTRODUCTION

Gastric cancer (GC) is considered to be the fifth most common malignancy and the third leading cause of cancer-related deaths (Chen et al., 2016; Bray et al., 2018) worldwide. Disappointingly, most patients with stomach cancer are diagnosed with advanced cancer because they lack specific symptoms (Van Cutsem et al., 2016). Because of the poor prognosis of patients with advanced GC, it is imperative to develop new strategies to improve the survival rate of this disease.

Expression of BGN (Biglycan), the gene as proteoglycan-I, was first detected in bone tissue (Gallagher, 1989). BGN is a member of the small leucine-rich proteoglycans (SLPRs) gene family and encodes a protein core that is modified to form a glycoprotein (Chen et al., 2020). BGN is a key component of the ECM; it participates in scaffolding the collagen fibrils and mediates cell signaling (Appunni et al., 2021). Existing studies have demonstrated the role of BGN in tumor proliferation, adhesion and invasion (Cooper and Giancotti, 2019; Hisamatsu et al., 2020; Moreno-Layseca et al., 2019; Yousefi et al., 2021). BGN could induce the epithelial-mesenchymal transition (EMT) of diverse malignancies and is necessary and sufficient to mediate the pro-EMT effect in pancreatic ductal adenocarcinoma (Thakur et al., 2016). BGN is regulated by the transforming growth factor-beta (TGFB) signaling pathway, a key regulator of the EMT process (Yang et al., 2021). Moreover, BGN is believed to enhance the ability of endometrial cancer cells to migrate and invade tissue (Sun et al., 2016) and is also considered a potential EMT biomarker of colorectal cancer (Li et al., 2017). Existing research findings strongly suggest an important role of BGN in the development of tumors. Immunotherapy of tumors has been one of the hot topics in recent years. Several studies have documented significant effects of immunotherapy on tumors (Zhang et al., 2015; Marrelli et al., 2016; Shitara et al., 2019); however, there is no report on immunotherapy of BGN in GC. Moreover, the role of BGN in the prognosis of GC and how BGN affects the immune infiltration of GC remain poorly understood.

In this study, we analyzed the difference in BGN expression between GC and normal patients in the online database by bioinformatics analysis. Thereafter, differentially expressed genes (DEGs) associated with BGN were identified. DEG-related functional enrichment analysis, Gene Set Enrichment Analysis (GSEA) analysis, and immune infiltration analysis were also carried out. We also explored the relationship between BGN gene expression and clinicopathological variables and survival in patients with GC. Finally, a predictive model that could well predict the probability of 1-, 3-, and 5-years survival in GC was established.

MATERIALS AND METHODS

Data Sources

Gene expression data and corresponding clinical information for GC, which included 375 tumor tissues and 32 normal tissues, were downloaded from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). **Table 1**, **Table 2** shows the characteristics of patients with GC from the TCGA database. The gene expression of 174 normal tissues was downloaded from GTEx through UCSC XENA (http://xena.ucsc.edu). Fragments Per kilobase per Million (FPKM) RNAseq data were converted into transcripts Per Million reads (TPM), and log2 translated for subsequent analysis. All tissue samples with incomplete clinical data were excluded.

BGN Differential Expression in Pan-Cancer and GC Tissues

We downloaded TPM RNAseq data for tumor tissues (TCGA) and normal tissues (TCGA and GTEx) from the UCSC XENA. The differential expression between tumor and normal tissues was tested by Wilcoxon Rank Sum Test and visualized through boxplots and scatter plots. We also used Receiver Operating Characteristic (ROC) curve to determine the diagnostic value of BGN gene expression for GC.

Real-Time PCR of BGN Expressions in GC and Adjacent Tissues

Tumor and para-cancer biopsy tissues were collected from 12 consecutive patients that were diagnosed with GC for the first time from the Endoscopy Center of the First Affiliated Hospital of Guangxi Medical University. The body tissues were immediately immersed in RNA protection solution and rapidly stored in a refrigerator at −80°C. No patient was diagnosed with any other malignancy, nor had they received any treatment for the tumor.

RNA Extraction and Quantitative Real-Time PCR (qRT-PCR)

Total RNA of tissues was extracted using Trizol reagent (R0016, Beyotime Biotechnology Co., Ltd., Shanghai, China, according to the manufacturer's instructions. Complementary DNAs(cDNAs) were generated from 1 μg RNA PrimeScript™ RT Reagent Kit with gDNA Eraser (RR047A, Takara Bio, Inc.). RT-PCR was conducted via the FastStart Universal SYBR Green Master (ROX) (Roche) in the Applied Biosystems QuantStudio TM Real-PCR System (Q6). Human BGN primers were utilized, and the relative mRNA expression was determined using the comparative Ct method with Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) as the reference gene. The primer sequences were as follows:

BGN-forward: 5′-TGACTGGCATCCCCAAAGAC-3′
BGN-reverse: 5′-GAGTAGCGAAGCAGGTCCTC-3′
GAPDH-forward: 5′-GTCAGCCGCATCTTCTTT-3′
GAPDH-reverse: 5′-CGCCCAATACGACCAAAT-3′

Immunohistochemistry

From January 2018 to September 2020, the tumors and adjacent tissues of 80 consecutive patients with GC after surgery in Suqian First People's Hospital were collected. Patients who had received radiation or chemotherapy prior to surgery and had other malignancies were excluded from the study. After dewaxing, hydration, and thermal repair, the primary antibody against BGN

TABLE 1 | The clinical characteristic of Gastric Cancer.

Characteristic	Levels	Overall
N		375
Gender, n (%)	Female	134 (35.7%)
	Male	241 (64.3%)
Age, n (%)	≤ 65	164 (44.2%)
	>65	207 (55.8%)
T stage, n (%)	T1	19 (5.2%)
	T2	80 (21.8%)
	T3	168 (45.8%)
	T4	100 (27.2%)
N stage, n (%)	N0	111 (31.1%)
	N1	97 (27.2%)
	N2	75 (21%)
	N3	74 (20.7%)
M stage, n (%)	M0	330 (93%)
	M1	25 (7%)
Histological type, n (%)	Diffuse Type	63 (16.8%)
	Mucinous Type	19 (5.1%)
	Not Otherwise Specified	207 (55.3%)
	Papillary Type	5 (1.3%)
	Signet Ring Type	11 (2.9%)
	Tubular Type	69 (18.4%)
Pathologic stage, n (%)	Stage I	53 (15.1%)
	Stage II	111 (31.5%)
	Stage III	150 (42.6%)
	Stage IV	38 (10.8%)
Histologic grade, n (%)	G1	10 (2.7%)
	G2	137 (37.4%)
	G3	219 (59.8%)
Residual tumor, n (%)	R0	298 (90.6%)
	R1	15 (4.6%)
	R2	16 (4.9%)
Primary therapy outcome, n (%)	PD	65 (20.5%)
	SD	17 (5.4%)
	PR	4 (1.3%)
	CR	231 (72.9%)
H pylori infection, n (%)	No	145 (89%)
	Yes	18 (11%)
Barretts esophagus, n (%)	No	193 (92.8%)
	Yes	15 (7.2%)
Anatomic neoplasm subdivision, n (%)	Antrum/Distal	138 (38.2%)
	Cardia/Proximal	48 (13.3%)
	Fundus/Body	130 (36%)
	Gastroesophageal Junction	41 (11.4%)
	Other	4 (1.1%)
Age, median (IQR)		67 (58, 73)

R0, No visible or microscopic tumor residue; R1, No visible, but microscopic residual tumor; R2, Visible tumor residue; CR, Complete response; PR, Partial response; SD, Stable disease; PD, Progressive disease.

(ab209234, Abcam, 1:2000) was incubated overnight at 4°C followed by incubation with detection polymer for 40 min at room temperature. 3,3′-Diaminobenzidine DAB (P0202, Beyotime Biotechnology co.) was used for signal detection. The images taken under the microscope were analyzed using the IHC profiler plugin of ImageJ software (Varghese et al., 2014). Finally, SPSS version 23.0 software was used to statistic the results.

Identification of DEGs Between High and Low Expression Groups of BGN

According to the mean value of BGN expression, the data from the TCGA cohort were divided into high expression group and low expression group, and the DESeq2 package (Love et al., 2014) was used for differential analysis. DEGs were defined as having a p.adj <0.05 and |logFC| >1.5. The details of the DEGs were visualized using the volcano map.

Functional Enrichment Analysis of DEGs

After ID conversion of identified DEGs via or.Hs.eg.db package, further functional enrichment analysis was performed through clusterProfiler package (Yu et al., 2012). Enrichments that satisfied the following conditions were considered significant: p.adj<0.05, and q-value<0.2. DEGs results were employed for gene-set enrichment analyses (GSEA) and building gene-set enrichment plots against the Molecular Signatures Database (MSigDB) hallmark gene sets through

TABLE 2 | BGN expression levels in 33 cancers and normal tissues.

Cancers	Groups	Cases (n)	Median	Mean	SD	SE	W value	*p* value
ACC	Normal	128	7.343	7.26	0.923	0.082	8172	< **0.001**
	Tumor	77	6.139	5.99	1.3	0.148		
BLCA	Normal	28	6.128	6.15	0.876	0.166	4289	**0.029**
	Tumor	407	6.801	6.781	1.66	0.082		
BRCA	Normal	292	6.396	6.357	0.944	0.055	26339.5	< **0.001**
	Tumor	1099	8.537	8.397	1.084	0.033		
CESC	Normal	13	8.134	7.765	1.141	0.317	3047	**0.001**
	Tumor	306	6.439	6.421	1.524	0.087		
CHOL	Normal	9	7.301	7.344	0.525	0.175	77	**0.015**
	Tumor	36	8.033	8.093	0.922	0.154		
COAD	Normal	349	4.953	5.008	1.43	0.077	23468.5	< **0.001**
	Tumor	290	6.663	6.57	1.586	0.093		
DLBC	Normal	444	0.692	0.937	0.918	0.044	120	< **0.001**
	Tumor	47	6.668	6.373	1.394	0.203		
ESCA	Normal	666	5.449	5.469	1.17	0.045	16232	< **0.001**
	Tumor	182	7.309	7.478	1.416	0.105		
GBM	Normal	1157	4.252	4.213	0.894	0.026	2708	< **0.001**
	Tumor	166	7.288	7.238	1.027	0.08		
HNSC	Normal	44	5.226	5.365	1.466	0.221	3849.5	< **0.001**
	Tumor	520	7.499	7.392	1.489	0.065		
KICH	Normal	53	7.619	7.31	1.633	0.224	3143	< **0.001**
	Tumor	66	5.115	5.301	1.139	0.14		
KIRC	Normal	100	7.668	7.58	1.351	0.135	13892.5	< **0.001**
	Tumor	531	8.799	8.589	1.356	0.059		
KIRP	Normal	60	7.518	7.339	1.477	0.191	11940.5	< **0.001**
	Tumor	289	6.349	6.42	1.728	0.102		
LAML	Normal	70	0.604	0.714	0.542	0.065	5826.5	0.646
	Tumor	173	0.731	0.942	0.969	0.074		
LGG	Normal	1152	4.249	4.208	0.891	0.026	136652	< **0.001**
	Tumor	523	5.114	5.367	1.222	0.053		
LIHC	Normal	160	7.068	7.082	0.787	0.062	43486	< **0.001**
	Tumor	371	5.844	5.797	1.742	0.09		
LUAD	Normal	347	8.61	8.551	0.992	0.053	118430	< **0.001**
	Tumor	515	8.079	7.978	1.087	0.048		
LUSC	Normal	338	8.673	8.626	0.954	0.052	124402	< **0.001**
	Tumor	498	7.754	7.646	1.283	0.058		
MESO	Tumor	87	9.419	9.347	1.451	0.156	–	–
OV	Normal	88	5.938	5.973	1.227	0.131	10400.5	< **0.001**
	Tumor	427	7.063	7.046	1.443	0.07		
PAAD	Normal	171	4.535	4.645	1.365	0.104	961.5	< **0.001**
	Tumor	179	9.262	8.96	1.173	0.088		
PCPG	Normal	3	7.449	7.465	0.276	0.159	336	0.497
	Tumor	182	7.195	7.231	1.162	0.086		
PRAD	Normal	152	6.951	6.88	1.146	0.093	40731.5	0.133
	Tumor	496	6.777	6.785	1.027	0.046		
READ	Normal	318	5.075	5.096	1.434	0.08	6420.5	< **0.001**
	Tumor	93	6.717	6.745	1.51	0.157		
SARC	Normal	2	6.951	6.951	0.004	0.003	–	–
	Tumor	262	9.046	8.692	1.898	0.117		
SKCM	Normal	813	6.711	6.804	1.121	0.039	178330.5	0.054
	Tumor	469	6.905	6.939	1.355	0.063		
STAD	Normal	206	4.383	4.58	1.398	0.096	5987	< **0.001**
	Tumor	375	7.664	7.601	1.368	0.067		
TGCT	Normal	165	6.33	6.431	0.736	0.057	10324	**0.004**
	Tumor	154	6.82	6.913	1.708	0.138		
THCA	Normal	338	7.84	7.682	1.003	0.055	126239.5	< **0.001**
	Tumor	512	6.909	6.839	1.116	0.049		
THYM	Normal	446	0.696	0.959	0.975	0.046	698.5	< **0.001**
	Tumor	119	6.381	6.193	1.806	0.166		
UCEC	Normal	101	7.483	7.393	0.979	0.097	13701	< **0.001**
	Tumor	181	6.162	6.195	1.531	0.114		
UCS	Normal	78	7.556	7.563	0.821	0.093	1690	**0.018**
	Tumor	57	8.198	7.999	1.436	0.19		
UVM	Tumor	79	6.54	6.415	1.115	0.125	–	–

Bold indicates statistically significant, that is, a p *value less than 0.05.*

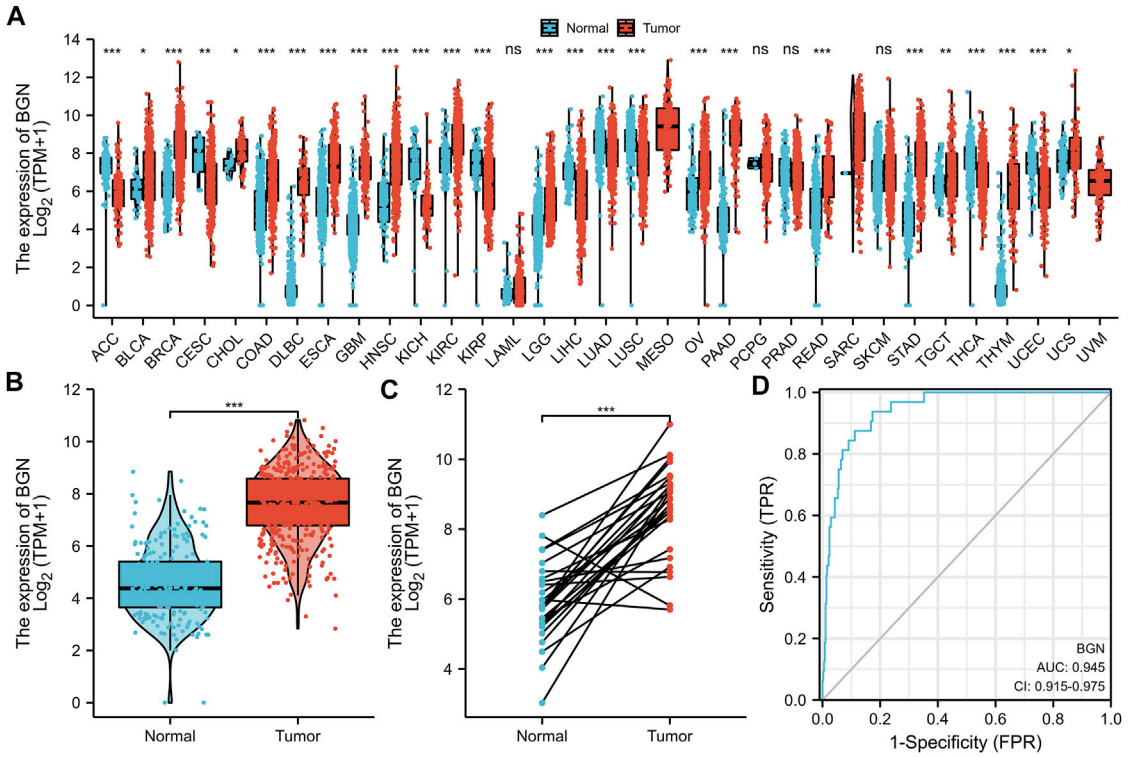

FIGURE 1 | Differential expression of BGN in different tumors and BGN-related differentially expressed genes (DEGs). **(A)** Differential expression of BGN of different cancers compared with normal tissues in the TCGA and GTEx database. **(B,C)** Differential expression of BGN in STAD. **(D)** ROC curve was used to calculate the diagnostic predictive value of BGN expression between STAD and normal tissues. Significance marker: ns, $p \geq 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. The abbreviations for 33 cancers are as follows: Adrenocortical carcinoma (ACC); Bladder Urothelial Carcinoma (BLCA); Breast invasive carcinoma (BRCA); Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC); Cholangiocarcinoma (CHOL); Colon adenocarcinoma (COAD); Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC); Esophageal carcinoma (ESCA); Glioblastoma multiforme (GBM); Head and Neck squamous cell carcinoma (HNSC); Kidney Chromophobe (KICH); Kidney renal clear cell carcinoma (KIRC); Kidney renal papillary cell carcinoma (KIRP); Acute Myeloid Leukemia (LAML); Brain Lower Grade Glioma (LGG); Liver hepatocellular carcinoma (LIHC); Lung adenocarcinoma (LUAD); Mesothelioma (MESO); Ovarian serous cystadenocarcinoma (OV); Pancreatic adenocarcinoma (PAAD); Pheochromocytoma and Paraganglioma (PCPG); Prostate adenocarcinoma (PRAD); Rectum adenocarcinoma (READ); Sarcoma (SARC); Skin Cutaneous Melanoma (SKCM); Testicular Germ Cell Tumors (TGCT); Thyroid carcinoma (THCA); Thymoma (THYM); Uterine Corpus Endometrial Carcinoma (UCEC); Uterine Carcinosarcoma (UCS); Uveal Melanoma (UVM).

the R package, clusterProfiler, and significance was set as an adjusted $p < 0.05$ and FDR<0.25.

Immune Infiltration

After converting the level 3 HTSe1-FPKM format RNAseq data from the stomach adenocarcinoma (STAD) project of TCGA to TPM format, log2 conversion was performed. After normal tissue samples were removed, data from a total of 375 STAD samples were retained for subsequent analysis. The relative tumor infiltration levels of immune cell types were quantified using ssGSEA of clusterProfilerpackage (Yu et al., 2012) to quantify the relative tumor infiltration levels of immune cell types, and the marker genes of immune cell types for single-sample gene-set enrichment analysis (ssGSEA) were obtained from published signature gene lists (Bindea et al., 2013). Spearman's Correlation Test was adopted to determine a correlation between BGN and the immune infiltration levels and the association of immune infiltration with the different expression groups of BGN.

Clinical Correlation Analysis of BGN in Patients With GC

For TCGA data, Wilcoxon signed Rank-Sum test and logistic regression analyses were used to evaluate the relationship between BGN expression and clinicopathological variables. Moreover, univariate and multivariate Cox regression analyses were used to compare the effects of BGN expression and other clinicopathological variables on the overall survival of GC patients. Multivariate Cox regression analysis was used to examine the independent factors affecting the prognosis of GC.

Furthermore, we collected clinicopathological data from 80 patients who underwent immunohistochemistry to evaluate the relationship between BGN expression and clinicopathological

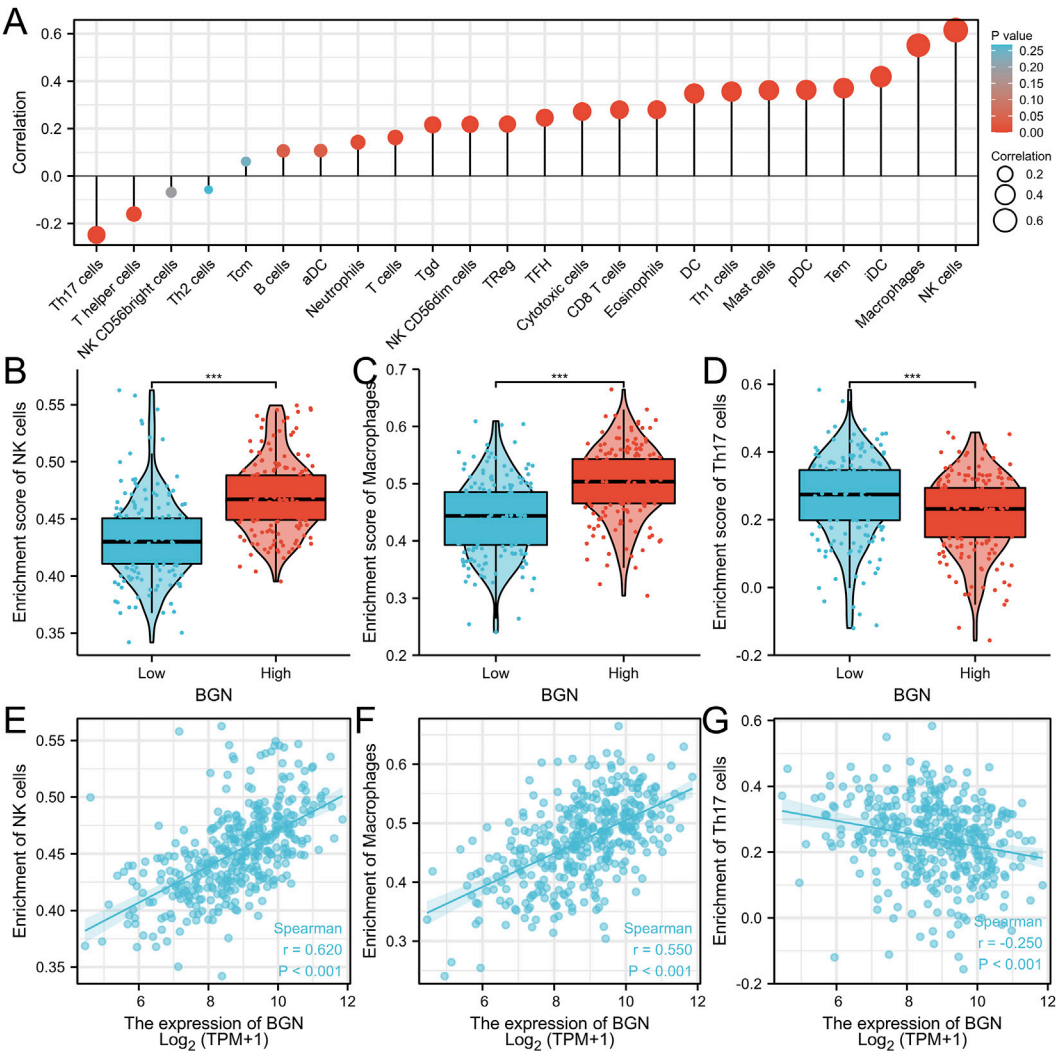

FIGURE 2 | The results of Real-Time PCR and Immunohistochemistry. **(A)** BGN expression in normal tissue (200X). **(B)** BGN expression in gastric cancer tissue (400X). **(C)** BGN expression in normal tissue (200X). **(D)** BGN expression in gastric cancer tissue (400X). **(E)** Relative BGN mRNA level in normal and GC tissues. GC: Gastric cancer. **, $p < 0.01$.

variables. Chi-square tests were used to evaluate the relationship between gender, pathological type, residual tumor status, and BGN expression. Fisher's exact tests were used to evaluate the relationship between pathologic stage, T stage, N stage, primary treatment outcome, and BGN expression. Wilcoxon signed Rank-Sum test was used to evaluate the relationship between age and BGN expression.

Construction and Verification of Nomogram

The identified independent factors associated with GC prognosis were used to construct a nomogram that predicted the probability of 1-, 3-, and 5-years survival in patients with GC. The prognostic data were obtained from a study by Jianfang Liu(Liu et al., 2018). Nomogram was constructed by R package with the survival and rms package. The Harrell's concordance index (C-index) was used to quantify the predictive accuracy, which ranges from 0.5 (no predictive power) to 1 (perfect prediction). Furthermore, calibration plots were generated to examine the performance characteristics of the predictive nomogram.

RESULTS

BGN Differential Expression in Pan-Cancer and GC Tissues

Significant differential expression of BGN was documented in most of the 33 cancers, including in STAD (**Figure 1A**). The expression of BGN in GC (375 cases from TCGA) was significantly higher than in normal tissues (32 para-cancer

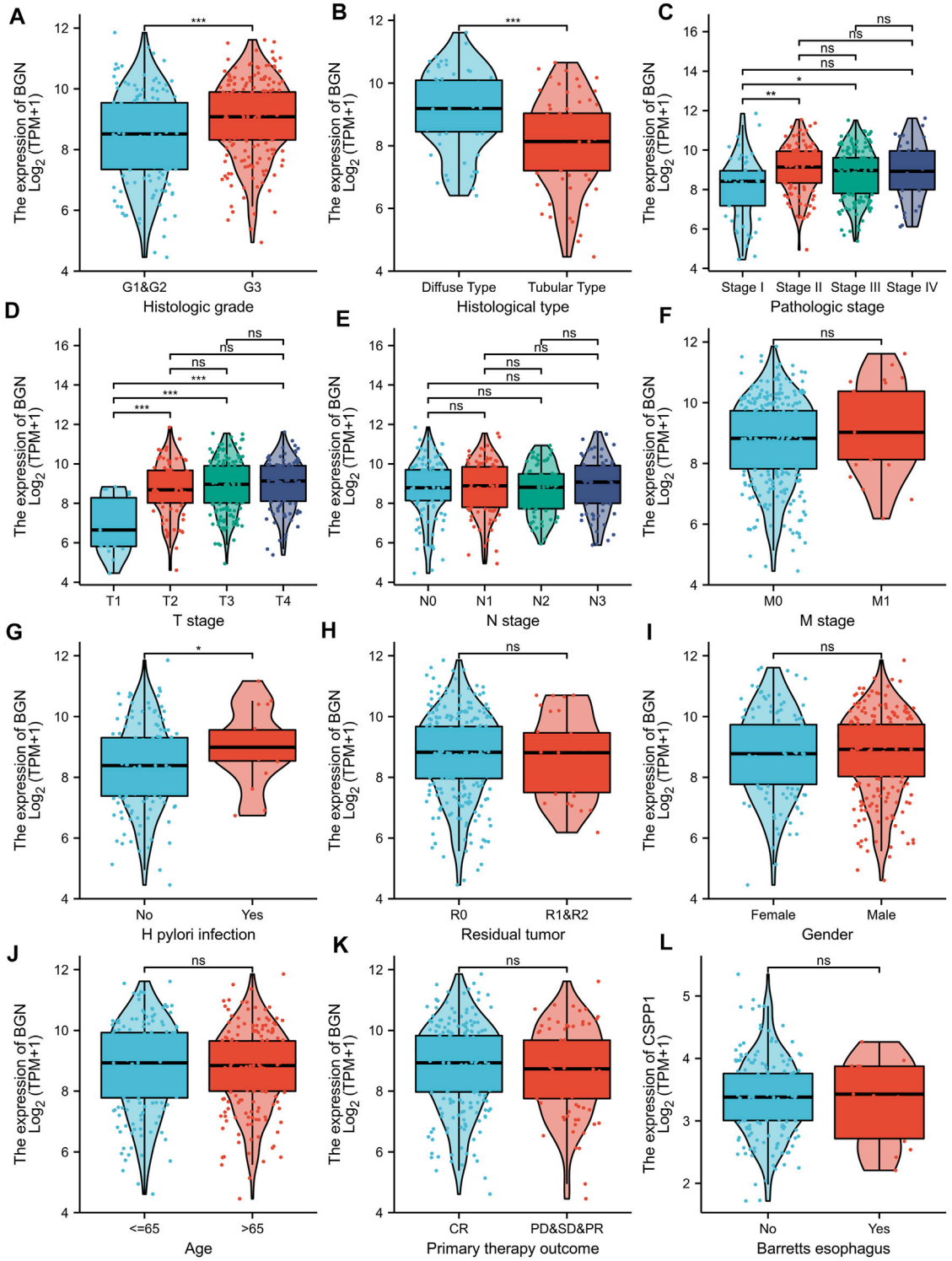

FIGURE 3 | Volcano plot of the DEGs, Functional enrichment analysis and GSEA analysis. **(A)**: **(E)** Volcano plots of the DEGs. Blue represent down-regulated DEGs, red represent up-regulated DEGs. **(B)**: The top three items enriched in biological processes (BP), cellular component (CC), molecular function (MF), and Kyoto Encyclopedia of Genes and Genomes (KEGG) of DEGs. **(C–H)**: Enrichment plots from the gene set enrichment analysis (GSEA). NES, normalized enrichment score; p.adj, adjusted p-value; FDR, false discovery rate.

TABLE 3 | GO and KEGG enrichment analysis.

Ontology	ID	Description	Gene ratio	Bg ratio	*p* Value	p.adjust	q value
BP	GO:0043062	extracellular structure organization	51/305	422/18670	1.31e-29	4.07e-26	3.36e-26
BP	GO:0030198	extracellular matrix organization	47/305	368/18670	2.24e-28	3.47e-25	2.86e-25
BP	GO:0043588	skin development	38/305	419/18670	7.95e-18	8.20e-15	6.77e-15
BP	GO:0070268	cornification	20/305	112/18670	1.66e-15	1.28e-12	1.06e-12
BP	GO:0008544	epidermis development	36/305	464/18670	8.33e-15	5.16e-12	4.26e-12
CC	GO:0062023	collagen-containing extracellular matrix	65/318	406/19717	7.31e-46	1.93e-43	1.72e-43
CC	GO:0005788	endoplasmic reticulum lumen	28/318	309/19717	1.66e-13	2.19e-11	1.96e-11
CC	GO:0044420	extracellular matrix component	12/318	51/19717	2.28e-11	2.00e-09	1.79e-09
CC	GO:0005604	basement membrane	14/318	95/19717	3.83e-10	2.53e-08	2.26e-08
CC	GO:0005581	collagen trimer	13/318	87/19717	1.37e-09	7.25e-08	6.47e-08
MF	GO:0005201	extracellular matrix structural constituent	41/290	163/17697	3.73e-37	1.45e-34	1.21e-34
MF	GO:0048018	receptor ligand activity	36/290	482/17697	2.71e-14	5.28e-12	4.41e-12
MF	GO:0005539	glycosaminoglycan binding	22/290	229/17697	2.93e-11	3.79e-09	3.17e-09
MF	GO:0005518	collagen binding	13/290	67/17697	5.41e-11	5.26e-09	4.40e-09
MF	GO:0061134	Peptidase regulator activity	21/290	219/17697	8.66e-11	6.74e-09	5.63e-09
KEGG	hsa04974	Protein digestion and absorption	17/134	103/8076	6.74e-13	1.31e-10	1.17e-10
KEGG	hsa04512	ECM-receptor interaction	10/134	88/8076	1.70e-06	1.66e-04	1.48e-04
KEGG	hsa04510	Focal adhesion	12/134	201/8076	1.20e-04	0.008	0.007
KEGG	hsa00980	Metabolism of xenobiotics by cytochrome P450	7/134	77/8076	2.71e-04	0.013	0.012
KEGG	hsa05204	Chemical carcinogenesis	7/134	82/8076	4.00e-04	0.016	0.014

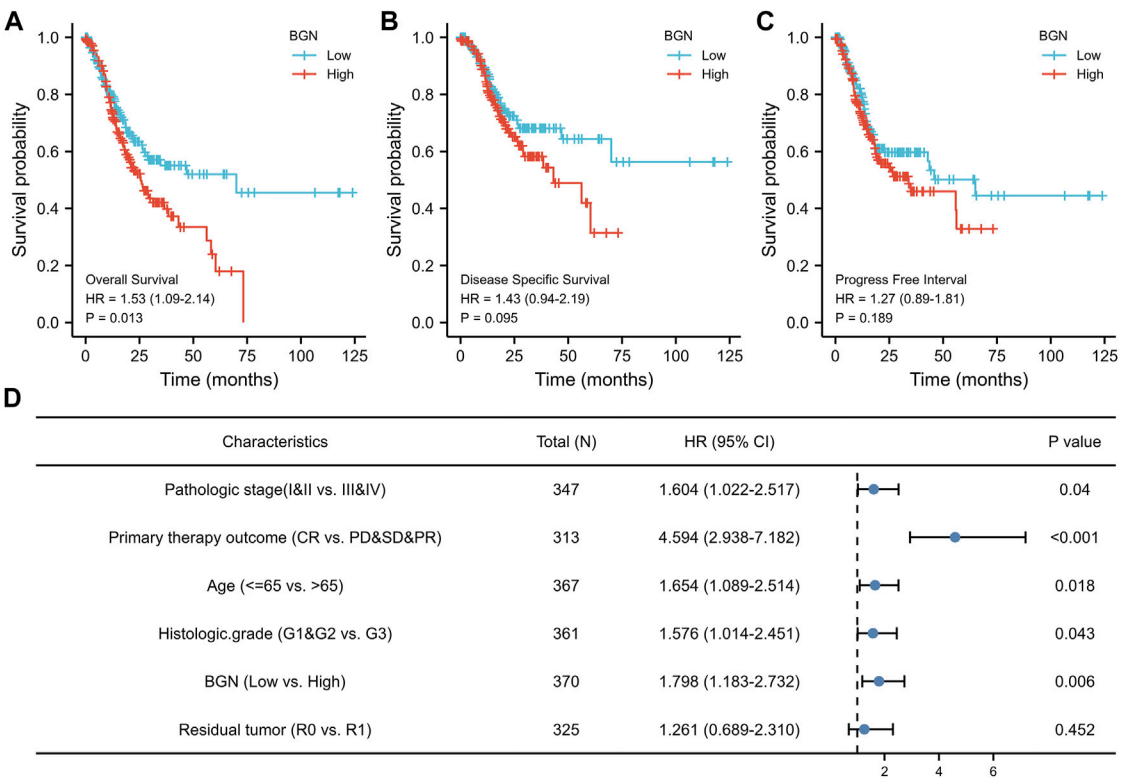

FIGURE 4 | The correlation between BGN expression and immune infiltration. **(A)** Correlation between the relative abundances of immune cells and BGN expression level. The size of dots is positively related to the absolute value of Spearman's R. **(B-D)** The difference of immune cells (Macrophages, NK cells, and Th17 cells) between the high and low expression groups based on the median value of BGN expression. **(E–G)** The correlation of immune cells (Macrophages, NK cells, and Th17 cells) between the high and low expression groups based on median value of BGN expression.

TABLE 4 | Correlation analysis between BGN and immune cells.

Gene	Immune cells	Spearman correlation coefficient	*p* Value
BGN	NK cells	0.620	**<0.001**
BGN	Macrophages	0.550	**<0.001**
BGN	iDC	0.419	**<0.001**
BGN	Tem	0.371	**<0.001**
BGN	pDC	0.363	**<0.001**
BGN	Mast cells	0.362	**<0.001**
BGN	Th1 cells	0.356	**<0.001**
BGN	DC	0.348	**<0.001**
BGN	Eosinophils	0.280	**<0.001**
BGN	CD8 T cells	0.279	**<0.001**
BGN	Cytotoxic cells	0.272	**<0.001**
BGN	Th17 cells	–0.250	**<0.001**
BGN	TFH	0.246	**<0.001**
BGN	TReg	0.219	**<0.001**
BGN	NK CD56dim cells	0.218	**<0.001**
BGN	Tgd	0.216	**<0.001**
BGN	T cells	0.163	**0.002**
BGN	T helper cells	–0.160	**0.002**
BGN	Neutrophils	0.142	**0.006**
BGN	aDC	0.107	**0.038**
BGN	B cells	0.106	**0.040**
BGN	NK CD56bright cells	–0.069	0.185
BGN	Tcm	0.061	0.237
BGN	Th2 cells	–0.057	0.267

Bold indicates statistically significant, that is, a p *value less than 0.05.*

tissues from TCGA and 174 normal tissues from GTEx) ($p < 0.001$) (**Figure 1B**). Similarly, the comparison of 27 tumor tissues in TCGA with the corresponding para-cancer tissues also showed significant expression of BGN in tumor tissues (**Figure 1C**).

Furthermore, based on the expression profile of TCGA in tumor and normal tissues, a ROC curve of BGN for the diagnosis of GC was plotted. **Figure 1D** shows that in the prediction of tumor and normal outcomes, the variable BGN showed high accuracy (AUC = 0.945, CI = 0.915–0.975).

Real-Time PCR and Immunohistochemistry

We further verified the BGN expression level using RT-PCR (**Figure 2E**, $p = 0.0068$) and IHC (**Figures 2A–D**). The results were consistent with those in the TCGA database, indicating significantly higher levels of BGN expression in GC than that in normal tissues.

DEGs Identification, Functional Enrichment Analysis and GSEA Analysis of DEGs

The volcano map shows the expression of identified DEGs between groups with high and low BGN expression (**Figure 3A**). of all the 492 DEGs. Of them, 207 were up-regulated, and 285 were down-regulated genes.

In terms of Biological Process (BP), most of the DEGs were enriched in extracellular structure organization, extracellular matrix (ECM) organization, and skin development. In terms of cellular components (CC), DEGs were mostly enriched in the collagen-containing ECM, endoplasmic reticulum lumen, and ECM components. In terms of molecular functions (MF), the DEGs also showed significant association with ECM structural constituent, receptor-ligand activity, and glycosaminoglycan binding. Furthermore, they were found mainly enriched in three KEGG pathways, including protein digestion and

TABLE 5 | Details of immune cell enrichment score in BGN high expression group and low expression group.

Immune cells	Enrichment scores in high and low expression groups		*p* value
	High (mean ± SD)	Low (mean ± SD)	
Macrophages	0.501 ± 0.061	0.44 ± 0.066	**<0.001**
NK cells	0.47 ± 0.031	0.433 ± 0.036	**<0.001**
Th17 cells	0.218 ± 0.111	0.266 ± 0.12	**<0.001**
aDC	0.394 ± 0.114	0.378 ± 0.119	0.159
B cells	0.231 ± 0.1	0.218 ± 0.112	0.107
CD8 T cells	0.575 ± 0.022	0.564 ± 0.023	**<0.001**
Cytotoxic cells	0.401 ± 0.095	0.36 ± 0.101	**<0.001**
DC	0.36 ± 0.108	0.304 ± 0.102	**<0.001**
Eosinophils	0.391 ± 0.037	0.373 ± 0.039	**<0.001**
iDC	0.433 ± 0.059	0.395 ± 0.054	**<0.001**
Mast cells	0.247 ± 0.087	0.188 ± 0.09	**<0.001**
Neutrophils	0.31 ± 0.092	0.289 ± 0.087	**0.030**
NK CD56bright cells	0.408 ± 0.053	0.412 ± 0.061	0.265
NK CD56dim cells	0.236 ± 0.072	0.208 ± 0.074	**0.001**
pDC	0.544 ± 0.1	0.487 ± 0.103	**<0.001**
T cells	0.392 ± 0.113	0.368 ± 0.114	**0.042**
T helper cells	0.578 ± 0.027	0.587 ± 0.029	**0.004**
Tcm	0.411 ± 0.04	0.406 ± 0.039	0.240
Tem	0.432 ± 0.039	0.406 ± 0.039	**<0.001**
TFH	0.335 ± 0.042	0.316 ± 0.048	**<0.001**
Tgd	0.239 ± 0.041	0.23 ± 0.053	**0.010**
Th1 cells	0.361 ± 0.051	0.328 ± 0.057	**<0.001**
Th2 cells	0.376 ± 0.032	0.375 ± 0.037	0.815
TReg	0.421 ± 0.127	0.376 ± 0.134	**0.002**

Bold indicates statistically significant, that is, a p *value less than 0.05.*

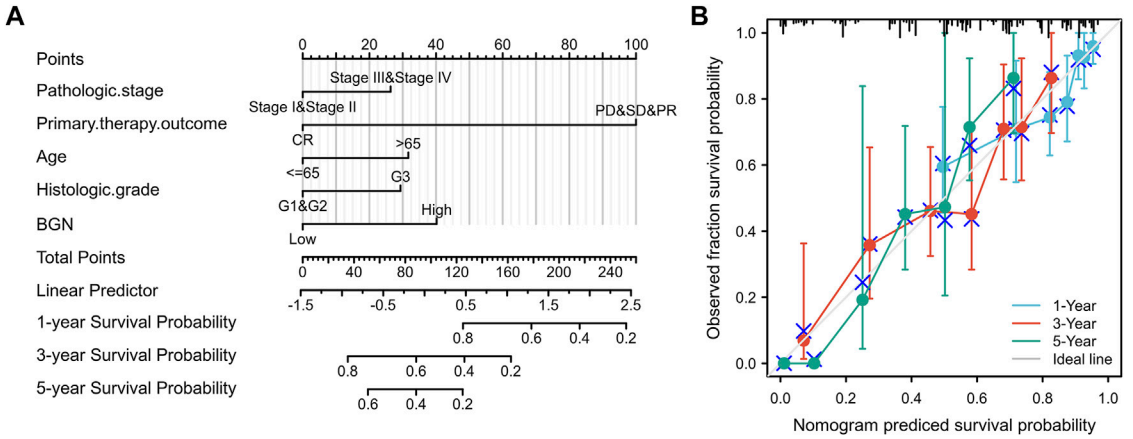

FIGURE 5 | Association with BGN expression and clinicopathological characteristics. **(A)** Histologic grade, **(B)** Histological type, **(C)** Pathologic stage, **(D)** T stage **(E)** N stage, **(F)** M stage, **(G)** *H pylori* infection, **(H)** Residual tumor, **(I)** Gender, **(J)** Age, **(K)** Primary therapy outcome, and **(L)** Barretts esophageal in GC patients in TCGA cohort. TCGA, The Cancer Genome Atlas; GC, gastric cancer.

absorption, ECM-receptor interaction, focal adhesion pathway (**Figure 3B**; **Table 3**). GSEA analysis revealed the following BGN-related enrichment pathways: collagen formulation, immunoregulatory interactions between a lymphoid and a non-lymphoid cell, focal adhesion, ECM glycoproteins, Wnt signaling, and signaling by vascular endothelial growth factor (VEGF), as shown in **Figures 3C–H**.

Correlation Between BGN Expression and Immune Infiltration

The BGN expression showed positive correlation with the enrichment of the NK cells (r = 0.620, $p < 0.001$) and macrophages (r = 0.550, $p < 0.001$) but negative correlation with the enrichment of Th17 cells (r = -0.250, $p < 0.001$) (**Figures 4A–G**; **Table 4**). The enrichment score of macrophages (High: 0.501 ± 0.061, Low: 0.44 ± 0.066, $p < 0.001$) and NK cells (High: 0.47 ± 0.031, Low: 0.433 ± 0.036, $p < 0.001$) in the group with high BGN expression was significantly higher than that in the group with low BGN expression, while the enrichment score of Th17 cells (High: 0.218 ± 0.111, Low: 0.266 ± 0.12, $p < 0.001$) in the group with high BGN expression was significantly lower than that in the group with low BGN expression (**Table 5**). The details of immune cell enrichment score in the BGN high expression group and low expression group are shown in **Table 5**.

Relationship Between BGN Expression and Clinicopathological Variables

BGN expression was remarkably correlated with histologic grade (**Figure 5A**, GI&G2 *vs.* G3, $p < 0.001$), histologic type (**Figure 5B**, Diffuse type *vs.* Tubular type, $p < 0.001$), histologic stage (**Figure 5C**, stage I *vs.* stage II and stage I *vs.* stage III, $p < 0.001$), T stage (**Figure 5D**, T1 *vs.* T2, T1 *vs.* T3, and T1 *vs.* T4, $p < 0.001$) and *Helicobacter pylori* (HP) infection (**Figure 5G**, yes *vs.* no, $p < 0.05$) in gastric cancer (GC). However, the following clinicopathological features showed no significant association with BGN expression: M stage, N stage, residual tumor, gender, age, primary therapy outcome, and Barrett's esophagus (**Figures 5F,H–L**, $p > 0.05$).

The results in **Table 6** showed that BGN expression was remarkably correlated with pathologic stage ($p = 0.008$), T stage ($p = 0.001$), histologic type ($p < 0.001$), and histological grade ($p = 0.025$) in 80 GC patients who underwent immunohistochemistry, but was not significantly associated with gender ($p = 0.802$), N stage ($p = 0.232$), residual tumor ($p = 0.323$), primary therapy outcome ($p = 0.655$), anatomic neoplasm subdivision ($p = 0.905$), and age ($p = 0.600$).

Association With BGN Expression and Prognosis of Patients With GC

The results of survival analysis revealed significant association of greater BGN expression with poor Overall Survival (OS) in GC patients (**Figure 6A**, HR = 1.53 (1.09–2.14), $p = 0.013$), but no significantly association with Disease Specific Survival (DSS) (**Figure 6B**, HR = 1.43 (0.94–2.19), $p = 0.095$), and Progress Free Interval (PFI) (**Figure 6C**, HR = 1.27 (0.89–1.81), $p = 0.189$).

In order to eliminate the influence of other clinicopathological variables on OS of GC, multivariate Cox regression analysis was performed to identify independent factors affecting OS of GC. **Table 7** and **Figure 6D** show that pathologic stage (stage I &II *vs.* stage III &IV, HR (95% CI) = 1.604 (1.022–2.517), $p = 0.040$), primary therapy outcome (CR *vs.* PD &SD &PR, HR (95% CI) = 4.594 (2.938–7.182), $p < 0.001$), age (≤65 *vs.* >65 years, HR (95% CI) = 1.654 (1.089–2.514), $p = 0.018$), histologic grade (G1 & G2 *vs.* G3, HR (95% CI) = 1.576 (1.014–2.451), $p = 0.043$), and BGN (low *vs.* high, HR (95% CI) = 1.798 (1.183–2.732), $p = 0.006$) had significant correlation with OS rates in patients with GC. However, BGN expression showed no association with poor DSS and DSS PFI (**Tables 8**; **Tables 9**).

TABLE 6 | The relationship between BGN expression and clinicopathological variables in 80 patients underwent immunohistochemistry.

Characteristic	Low	High	p
n	40	40	0.802
Gender (M/F), n (%)			
F	10 (12.5%)	12 (15%)	
M	30 (37.5%)	28 (35%)	
Pathologic stage, n (%)			**0.008**
I	11 (13.8%)	2 (2.5%)	
II	16 (20%)	14 (17.5%)	
III	13 (16.2%)	24 (30%)	
T stage, n (%)			**0.001**
T1	12 (15%)	1 (1.2%)	
T2	5 (6.2%)	4 (5%)	
T3	23 (28.7%)	32 (40%)	
T4	0 (0%)	3 (3.8%)	
N stage, n (%)			0.232
N0	10 (12.5%)	11 (13.8%)	
N1	12 (15%)	5 (6.2%)	
N2	7 (8.8%)	12 (15%)	
N3	11 (13.8%)	12 (15%)	
Histological type, n (%)			< **0.001**
Diffuse Type	6 (7.5%)	22 (27.5%)	
Mucinous Type	1 (1.2%)	5 (6.2%)	
Papillary Type	6 (7.5%)	5 (6.2%)	
Signet Ring Type	8 (10%)	6 (7.5%)	
Tubular Type	19 (23.8%)	2 (2.5%)	
Histological grade, n (%)			**0.025**
G1 & G2	24 (30%)	13 (16.2%)	
G3	16 (20%)	27 (33.8%)	
Residual tumor, n (%)			0.323
R0	26 (32.5%)	31 (38.8%)	
R1 & R2	14 (17.5%)	9 (11.2%)	
Primary therapy outcome, n (%)			0.655
CR	27 (33.8%)	32 (40%)	
PD	8 (10%)	5 (6.2%)	
PR	2 (2.5%)	1 (1.2%)	
SD	3 (3.8%)	2 (2.5%)	
Anatomic neoplasm subdivision, n (%)			0.905
Antrum	8 (10%)	8 (10%)	
Cardia	15 (18.8%)	13 (16.2%)	
Fundus/Body	15 (18.8%)	18 (22.5%)	
other	2 (2.5%)	1 (1.2%)	
Age (years), meidan (IQR)	63 (58, 70.5)	66 (58, 71.25)	0.600

R0, No visible or microscopic tumor residue; R1, No visible, but microscopic residual tumor; R2, Visible tumor residue; CR, Complete response; PR, Partial response; SD, Stable disease; PD, Progressive disease.
Bold indicates statistically significant, that is, a p *value less than 0.05.*

Construction and Validation of Nomogram

A nomogram to predict 1-, 3-, and 5-years' OS probability was constructed on the basis of multivariate Cox regression analysis. In it, five variables, namely pathologic stage, primary therapy outcome, age, histologic grade, and BGN expression level, were used. **Figure 7A** depicts 11 rows in the nomogram, with the rows ranging from 2 to 6 representing the above variables. The points of the five variables were added up to the total points, which were displayed in row 7 and corresponded to the linear predictor in the prediction of 1-, 3-, and 5-years survival probability in row 8. The C-index was used to quantify the predictive accuracy, ranging from 0.5 (no predictive power) to 1 (perfect prediction). The C-index of this nomogram was 0.728 (0.705–0.752), indicating that the prediction was in good agreement with the actual survival probability. The nomogram calibration plot (**Figure 7B**) also suggests that the nomogram was well-calibrated, with the mean predicted probabilities close to observed probabilities.

DISCUSSION

In the current study, we compared the expression level of BGN in tumor tissues from TCGA and normal tissues from TCGA and GTEx. The results demonstrated differential expression of BGN in most of the 33 tumors and significant expression in GC tissues. Similar results were obtained on comparison of the GC tissues in TCGA with the matched normal tissues. The expression level of BGN in GC tissues was significantly higher as compared with normal tissues ($p < 0.001$). RT-PCR and IHC also verified this association ($p < 0.01$). The AUC of the ROC curve to predict the diagnostic value of BGN for GC was 0.945 (0.915–0.975), suggesting greater expression of BGN expression in GC diagnosis. The above results suggest that BGN may be a new biomarker for GC.

In addition, 492 BGN-related DEGs, including 207 up-regulated and 285 down-regulated genes, were identified. GO and KEGG enrichment analyses on DEGs were also done. In terms of BP, DEGs were mostly enriched in extracellular structure organization, ECM organization, and skin development. In terms of CC, DEGs were mostly enriched in collagen-containing ECM, endoplasmic reticulum lumen, and ECM components. Also, the DEGs were significantly associated with ECM structural constituent, receptor-ligand activity, and glycosaminoglycan binding in terms of MF. DEGs showed significant enrichment in three KEGG pathways of protein digestion and absorption, ECM-receptor interaction, focal adhesion. ECM plays a key role in the cell microenvironment and in maintaining normal cell activity (Giussani et al., 2019). Recent studies have shown a close correlation of ECM to tumor progression, including in the avoidance of apoptosis, the regulation of cell growth, the promotion of tumor angiogenesis, and the acquisition of invasion and metastasis ability (Pickup et al., 2014; Poltavets et al., 2018; Eble and Niland, 2019). The disorder of collagen, a key component of ECM, correlates with malignant tumor (Levental et al., 2009). Changes in the levels of metabolites related to protein digestion and absorption also have a key role in the development of cancer (Mo et al., 2020). GSEA enrichment analysis revealed that BGN-related DEGs were significantly enriched in collagen formulation (Nissen et al., 2019), immunoregulatory interactions between a lymphoid and a non-lymphoid cell (Sautès-Fridman et al., 2019), focal adhesion (Eke and Cordes, 2015), ECM glycoproteins (Mohan et al., 2020), Wnt signaling (Bugter et al., 2021), and signaling by VEGF (Apte et al., 2019), which were significantly related to the tumor. Considering the above findings, we speculate that BGN-related genes may be involved in the occurrence and progression of GC, and BGN may be a potential therapeutic target for GC.

Immunotherapy of tumors has been one of the hot topics over recent years. The use of Trastuzumab as immunotherapy has been shown to prolong overall survival in patients with HER2-

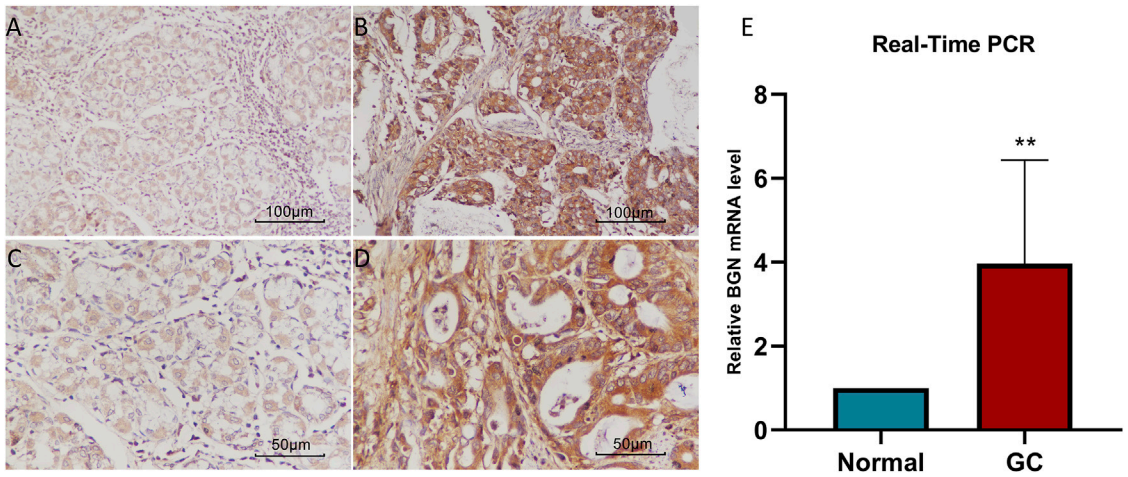

FIGURE 6 | The association between BGN expression and prognosis of patients with Gastric Cancer. **(A)** Overall Survival. **(B)** Disease Specific Survival. **(C)** Progress Free Interval. **(D)** Results of multivariate Cox regression analysis of the relationship between Overall Survival and clinicopathological variables in patients with gastric cancer. HR: Hazard Ratio. CI: Confidence Interval.

TABLE 7 | Univariate regression and multivariate survival method (Overall Survival) of prognostic covariates in patients with Gastric Cancer

Characteristics	Total(N)	Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	*p* Value	Hazard ratio (95% CI)	*p* Value
Pathologic.stage	347				
Stage I&Stage II	164	Reference			
Stage III&Stage IV	188	1.947 (1.358–2.793)	**<0.001**	1.604 (1.022–2.517)	**0.040**
Primary.therapy.outcome	313				
CR	231	Reference			
PD&SD&PR	86	4.228 (2.905–6.152)	**<0.001**	4.594 (2.938–7.182)	**<0.001**
Residual.tumor	325				
R0	298	Reference			
R1&R2	31	3.445 (2.160–5.494)	**<0.001**	1.261 (0.689–2.310)	0.452
Age	367				
≤ 65	164	Reference			
>65	207	1.620 (1.154–2.276)	**0.005**	1.654 (1.089–2.514)	**0.018**
Histologic.grade	361				
G1&G2	147	Reference			
G3	219	1.353 (0.957–1.914)	0.087	1.576 (1.014–2.451)	**0.043**
Gender	370				
Female	134	Reference			
Male	241	1.267 (0.891–1.804)	0.188		
Race	320				
White	238	Reference			
Asian&Black or African American	85	0.801 (0.515–1.247)	0.326		
BGN	370				
Low	188	Reference			
High	187	1.494 (1.070–2.087)	**0.019**	1.798 (1.183–2.732)	**0.006**

R0, No visible or microscopic tumor residue; R1, No visible, but microscopic residual tumor; R2, Visible tumor residue; CR, Complete response; PR, Partial response; SD, Stable disease; PD, Progressive disease.

Bold indicates statistically significant, that is, a p *value less than 0.05.*

positive GC (Shitara et al., 2019). In several clinical trials (Zhang et al., 2015), adoptive cell therapy has also demonstrated promising results against GC. A high incidence of somatic mutations in GC patients suggests ideal candidacy of Trastuzumab for immunotherapy (Marrelli et al., 2016). These results give us more confidence in the treatment of stomach cancer. However, due to the high complexity of the immune microenvironment of GC, the identification of biomarkers associated with GC require greater attention in the future (Zhao et al., 2019). The BGN expression was positively correlated with the enrichment of the NK cells (r = 0.620, $p < 0.001$) and macrophages (r = 0.550, $p < 0.001$) but was

TABLE 8 | Univariate regression and multivariate survival method (Progress Free Interval) of prognostic covariates in patients with Gastric Cancer

Characteristics	Total(N)	Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	*p* Value	Hazard ratio (95% CI)	*p* Value
Pathologic.stage	349				
Stage I&Stage II	164	Reference			
Stage III&Stage IV	188	1.676 (1.154–2.435)	**0.007**	1.202 (0.787–1.834)	0.395
Primary.therapy.outcome	315				
CR	231	Reference			
PD&SD&PR	86	8.041 (5.465–11.832)	<**0.001**	8.297 (5.319–12.941)	<**0.001**
Residual.tumor	326				
R0	298	Reference			
R1&R2	31	3.469 (2.127–5.656)	<**0.001**	1.384 (0.797–2.401)	0.248
Age	369				
≤ 65	164	Reference			
>65	207	0.858 (0.603–1.221)	0.395		
Histologic.grade	363				
G1&G2	147	Reference			
G3	219	1.540 (1.057–2.245)	**0.025**	1.632 (1.064–2.503)	**0.025**
Gender	372				
Female	134	Reference			
Male	241	1.638 (1.099–2.440)	**0.015**	1.404 (0.889–2.217)	0.145
Race	322				
White	238	Reference			
Asian&Black or African American	85	1.061 (0.688–1.637)	0.787		
BGN	372				
Low	188	Reference			
High	187	1.280 (0.897–1.825)	0.174		

R0, No visible or microscopic tumor residue; R1, No visible, but microscopic residual tumor; R2, Visible tumor residue; CR, Complete response; PR, Partial response; SD, Stable disease; PD, Progressive disease.

Bold indicates statistically significant, that is, a p *value less than 0.05.*

TABLE 9 | Univariate regression and multivariate survival method (Disease Specific Survival) of prognostic covariates in patients with Gastric Cancer

Characteristics	Total(N)	Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	*p* Value	Hazard ratio (95% CI)	*p* Value
Pathologic.stage	331				
Stage I&Stage II	164	Reference			
Stage III&Stage IV	188	2.146 (1.352–3.404)	**0.001**	1.500 (0.874–2.575)	0.141
Primary.therapy.outcome	310				
CR	231	Reference			
PD&SD&PR	86	8.697 (5.439–13.908)	<**0.001**	9.129 (5.214–15.984)	<**0.001**
Residual.tumor	314				
R0	298	Reference			
R1&R2	31	5.142 (3.014–8.771)	<**0.001**	1.901 (1.022–3.534)	**0.042**
Age	346				
≤ 65	164	Reference			
>65	207	1.211 (0.797–1.840)	0.371		
Histologic.grade	340				
G1&G2	147	Reference			
G3	219	1.338 (0.862–2.078)	0.194		
Gender	349				
Female	134	Reference			
Male	241	1.573 (0.985–2.514)	0.058	1.338 (0.765–2.341)	0.307
Race	305				
White	238	Reference			
Asian&Black or African American	85	1.097 (0.656–1.836)	0.724		
BGN	349				
Low	188	Reference			
High	187	1.444 (0.945–2.206)	0.089	1.528 (0.931–2.510)	0.094

R0, No visible or microscopic tumor residue; R1, No visible, but microscopic residual tumor; R2, Visible tumor residue; CR, Complete response; PR, Partial response; SD, Stable disease; PD, Progressive disease.

Bold indicates statistically significant, that is, a p *value less than 0.05.*

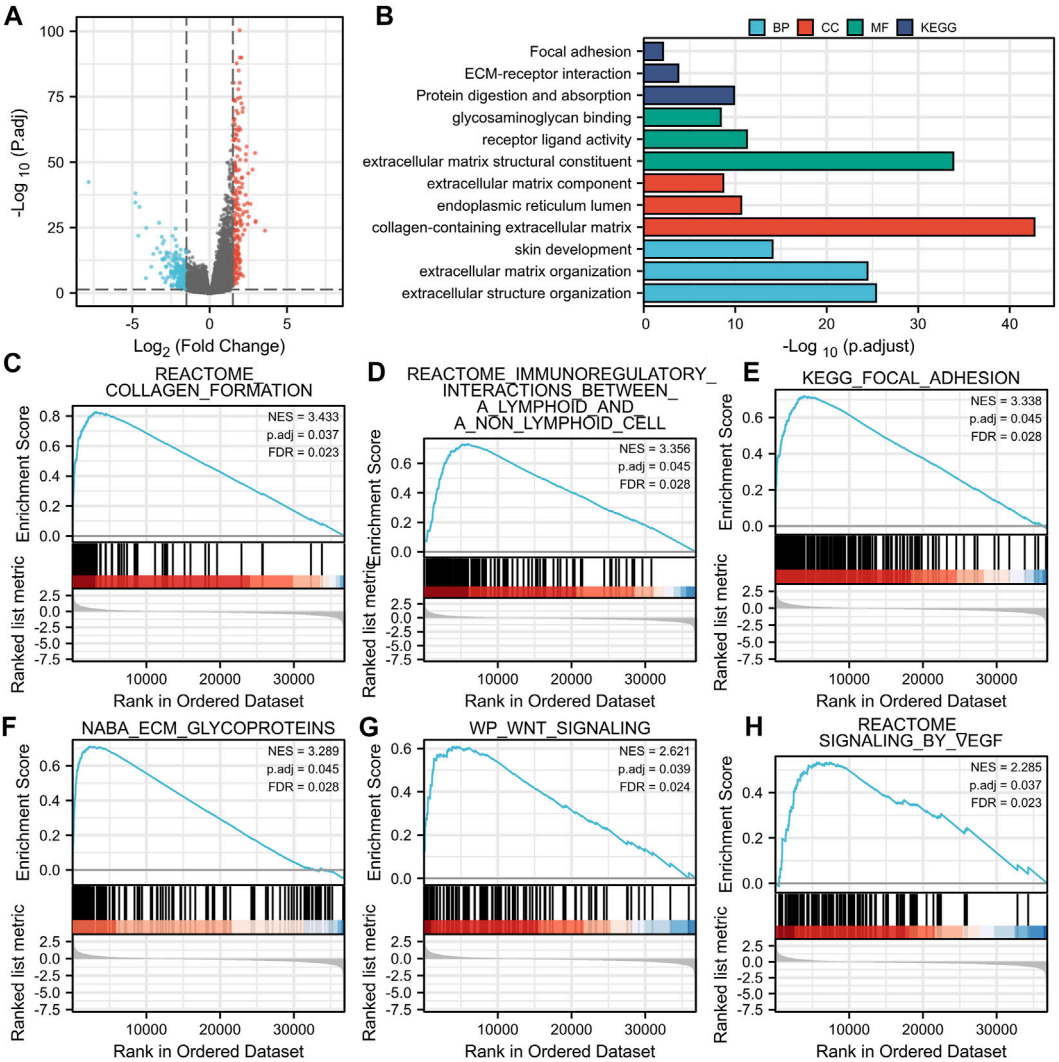

FIGURE 7 | A quantitative method to predict GC patients' probability of 1-, 3-, and 5-years OS. **(A)** A nomogram for predicting the probability of 1-, 3-, and 5-years OS for GC patients. **(B)** Calibration plots of the nomogram for predicting the probability of OS at 1, 3, and 5 years. GC, gastric cancer; OS, overall survival.

negatively correlated with the enrichment of Th17 cells. This indicates that the improvement of innate immunity is accompanied by the decrease of adaptive immunity. Macrophages, a type of immune cell present in large numbers in most tumor types, play an important regulatory role in promoting the development of malignancy (Noy and Pollard, 2014). Macrophages were recruited by inflammatory signals released by cancer cells in primary and metastatic tumors and differentiated into tumor-associated macrophages (TAMs) that promote tumor progression (Qian et al., 2011; Arwert et al., 2018). A large number of Th17 cell infiltrates were reported in different tumor types, including ovarian cancer (Miyahara et al., 2008), hepatocellular carcinoma (Zhang et al., 2009), colorectal cancer (Tosolini et al., 2011), and multiple myeloma (Prabhala et al., 2010). An abundance of Th17 cells in hepatocellular carcinoma and colorectal cancer showed association with poor prognosis (Kryczek et al., 2009). The results indicate that in the occurrence and development of GC, numerous immune cell infiltration changes occur, which may play a certain regulatory role.

BGN expression showed a significant correlation with histologic grade, histologic type, histologic stage, T stage, and *Helicobacter pylori* (HP) infection in patients with GC. Thus, GC patients with high BGN expression may have poorer histological types, lower tumor differentiation, more advanced tumor development, and may show greater association with HP infection. Furthermore, survival analysis suggested a significant correlation of high BGN expression with poor OS. Multivariate Cox regression analysis was conducted to exclude the influence of other variables. This analysis also showed that pathologic stage,

primary therapy outcome, age, histologic grade, and BGN expression level are independent risk factors for OS in GC. These findings strongly suggest the key role of BGN in the development of GC, leading to a poor prognosis of GC.

A nomogram was established to predict 1-, 3-, and 5-years survival probability of GC patients by including the above five independent survivorship risk factors, namely pathologic stage, primary therapy outcome, age, histologic grade, and BGN expression. Our nomogram can predict the OS probability of GC patients very well (C-index = 0.728). The calibration map shows that the nomogram's predicted OS probability matches the actual probability. Because of the very uncertain prognosis of tumor patients, understanding the risk stratification of patients with tumors correctly (Gratian et al., 2014) becomes crucial. Our nomogram based on independent factors related to the survival of GC patients can predict the OS probability of GC patients and can be widely used in clinical practice Cs-Szabó et al., 1995, Vuillermoz et al., 2004.

CONCLUSIONS AND LIMITATIONS

Overall, the findings of the current research are summarized below:

First, we reported and verified the differential expression of BGN in GC and normal tissue and concluded that the occurrence, progression, and prognosis of GC were significantly correlated with BGN. Second, BGN is a good biomarker for the proper diagnosis of GC. Third, BGN-related changes in the tumor microenvironment and immune invasion may play an important role in the occurrence and progression of GC. Finally, as our nomogram could predict the survival probability of GC patients, it may be widely used in clinical practice. Due to the limited conditions, we could not study molecular subtypes. This issue will be addressed in future research.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

ETHICS STATEMENT

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

AUTHOR CONTRIBUTIONS

GT contributed to the conception of the study; SZ performed the experiment and manuscript; HY contributed significantly to analysis and manuscript preparation; XX performed the data analyses and wrote the manuscript; LL and HH helped perform the analysis with constructive discussions.

FUNDING

This research was supported by the National Natural Science Foundation of China (81970558) and Guangxi Natural Science Foundation (2020GXNSFAA259095).

ACKNOWLEDGMENTS

Thanks to Xiaoyang Shi of The First People's Hospital of Suqian for his guidance of experimental technology.

REFERENCES

Appunni, S., Rubens, M., Ramamoorthy, V., Anand, V., Khandelwal, M., and Sharma, A. (2021). Biglycan: an Emerging Small Leucine-Rich Proteoglycan (SLRP) Marker and its Clinicopathological Significance. *Mol. Cel Biochem* 476 (11), 3935–3950. doi:10.1007/s11010-021-04216-z

Apte, R. S., Chen, D. S., and Ferrara, N. (2019). VEGF in Signaling and Disease: Beyond Discovery and Development. *Cell* 176 (6), 1248–1264. doi:10.1016/j.cell.2019.01.021

Arwert, E. N., Harney, A. S., Entenberg, D., Wang, Y., Sahai, E., Pollard, J. W., et al. (2018). A Unidirectional Transition from Migratory to Perivascular Macrophage Is Required for Tumor Cell Intravasation. *Cel Rep.* 23 (5), 1239–1248. doi:10.1016/j.celrep.2018.04.007

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* 39 (4), 782–795. doi:10.1016/j.immuni.2013.10.003

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: A Cancer J. Clinicians* 68 (6), 394–424. doi:10.3322/caac.21492

Bugter, J. M., Fenderico, N., and Maurice, M. M. (2021). Mutations and Mechanisms of WNT Pathway Tumour Suppressors in Cancer. *Nat. Rev. Cancer* 21 (1), 5–21. doi:10.1038/s41568-020-00307-z

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer Statistics in China, 2015. *CA: A Cancer J. Clinicians* 66 (2), 115–132. doi:10.3322/caac.21338

Chen, D., Qin, Y., Dai, M., Li, L., Liu, H., Zhou, Y., et al. (2020). BGN and COL11A1 Regulatory Network Analysis in Colorectal Cancer (CRC) Reveals that BGN Influences CRC Cell Biological Functions and Interacts with miR-6828-5p. *Cancer Manag. Res.* 12, 13051–13069. doi:10.2147/CMAR.S277261

Cooper, J., and Giancotti, F. G. (2019). Integrin Signaling in Cancer: Mechanotransduction, Stemness, Epithelial Plasticity, and Therapeutic Resistance. *Cancer Cell* 35 (3), 347–367. doi:10.1016/j.ccell.2019.01.007

Cs-Szabó, G., Roughley, P. J., Plaas, A. H. K., and Glant, T. T. (1995). Large and Small Proteoglycans of Osteoarthritic and Rheumatoid Articular Cartilage. *Arthritis Rheum.* 38 (5), 660–668. doi:10.1002/art.1780380514

Eble, J. A., and Niland, S. (2019). The Extracellular Matrix in Tumor Progression and Metastasis. *Clin. Exp. Metastasis* 36 (3), 171–198. doi:10.1007/s10585-019-09966-1

Eke, I., and Cordes, N. (2015). Focal Adhesion Signaling and Therapy Resistance in Cancer. *Semin. Cancer Biol.* 31, 65–75. doi:10.1016/j.semcancer.2014.07.009

Gallagher, J. T. (1989). The Extended Family of Proteoglycans: Social Residents of the Pericellular Zone. *Curr. Opin. Cel Biol.* 1 (6), 1201–1218. doi:10.1016/s0955-0674(89)80072-9

Giussani, M., Triulzi, T., Sozzi, G., and Tagliabue, E. (2019). Tumor Extracellular Matrix Remodeling: New Perspectives as a Circulating Tool in the Diagnosis and Prognosis of Solid Tumors. *Cells* 8 (2), 81. doi:10.3390/cells8020081

Gratian, L., Pura, J., Dinan, M., Reed, S., Scheri, R., Roman, S., et al. (2014). Treatment Patterns and Outcomes for Patients with Adrenocortical Carcinoma Associated with Hospital Case Volume in the United States. *Ann. Surg. Oncol.* 21 (11), 3509–3514. doi:10.1245/s10434-014-3931-z

Hisamatsu, E., Nagao, M., Toh, R., Irino, Y., Iino, T., Hara, T., et al. (2020). Fibronectin-containing High-Density Lipoprotein Is Associated with Cancer Cell Adhesion and Proliferation. *Kobe J. Med. Sci.* 66 (1), E40–e48.

Kryczek, I., Banerjee, M., Cheng, P., Vatan, L., Szeliga, W., Wei, S., et al. (2009). Phenotype, Distribution, Generation, and Functional and Clinical Relevance of Th17 Cells in the Human Tumor Environments. *Blood* 114 (6), 1141–1149. doi:10.1182/blood-2009-03-208249

Levental, K. R., Yu, H., Kass, L., Lakins, J. N., Egeblad, M., Erler, J. T., et al. (2009). Matrix Crosslinking Forces Tumor Progression by Enhancing Integrin Signaling. *Cell* 139 (5), 891–906. doi:10.1016/j.cell.2009.10.027

Li, H., Zhong, A., Li, S., Meng, X., Wang, X., Xu, F., et al. (2017). The Integrated Pathway of TGFβ/Snail with TNFα/NFκB May Facilitate the Tumor-Stroma Interaction in the EMT Process and Colorectal Cancer Prognosis. *Sci. Rep.* 7 (1), 4915. doi:10.1038/s41598-017-05280-6

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* 173 (2), 400–e11. e411. doi:10.1016/j.cell.2018.02.052

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Marrelli, D., Polom, K., Pascale, V., Vindigni, C., Piagnerelli, R., De Franco, L., et al. (2016). Strong Prognostic Value of Microsatellite Instability in Intestinal Type Non-cardia Gastric Cancer. *Ann. Surg. Oncol.* 23 (3), 943–950. doi:10.1245/s10434-015-4931-3

Miyahara, Y., Odunsi, K., Chen, W., Peng, G., Matsuzaki, J., and Wang, R.-F. (2008). Generation and Regulation of Human CD4+ IL-17-producing T Cells in Ovarian Cancer. *Proc. Natl. Acad. Sci.* 105 (40), 15505–15510. doi:10.1073/pnas.0710686105

Mo, L., Wei, B., Liang, R., Yang, Z., Xie, S., Wu, S., et al. (2020). Exploring Potential Biomarkers for Lung Adenocarcinoma Using LC-MS/MS Metabolomics. *J. Int. Med. Res.* 48 (4), 030006051989721. doi:10.1177/0300060519897215

Mohan, V., Das, A., and Sagi, I. (2020). Emerging Roles of ECM Remodeling Processes in Cancer. *Semin. Cancer Biol.* 62, 192–200. doi:10.1016/j.semcancer.2019.09.004

Moreno-Layseca, P., Icha, J., Hamidi, H., and Ivaska, J. (2019). Integrin Trafficking in Cells and Tissues. *Nat. Cel Biol* 21 (2), 122–132. doi:10.1038/s41556-018-0223-z

Nissen, N. I., Karsdal, M., and Willumsen, N. (2019). Collagens and Cancer Associated Fibroblasts in the Reactive Stroma and its Relation to Cancer Biology. *J. Exp. Clin. Cancer Res.* 38 (1), 115. doi:10.1186/s13046-019-1110-6

Noy, R., and Pollard, J. W. (2014). Tumor-associated Macrophages: from Mechanisms to Therapy. *Immunity* 41 (1), 49–61. doi:10.1016/j.immuni.2014.06.010

Pickup, M. W., Mouw, J. K., and Weaver, V. M. (2014). The Extracellular Matrix Modulates the Hallmarks of Cancer. *EMBO Rep.* 15 (12), 1243–1253. doi:10.15252/embr.201439246

Poltavets, V., Kochetkova, M., Pitson, S. M., and Samuel, M. S. (2018). The Role of the Extracellular Matrix and its Molecular and Cellular Regulators in Cancer Cell Plasticity. *Front. Oncol.* 8, 431. doi:10.3389/fonc.2018.00431

Prabhala, R. H., Pelluru, D., Fulciniti, M., Prabhala, H. K., Nanjappa, P., Song, W., et al. (2010). Elevated IL-17 Produced by TH17 Cells Promotes Myeloma Cell Growth and Inhibits Immune Function in Multiple Myeloma. *Blood* 115 (26), 5385–5392. doi:10.1182/blood-2009-10-246660

Qian, B.-Z., Li, J., Zhang, H., Kitamura, T., Zhang, J., Campion, L. R., et al. (2011). CCL2 Recruits Inflammatory Monocytes to Facilitate Breast-Tumour Metastasis. *Nature* 475 (7355), 222–225. doi:10.1038/nature10138

Sautès-Fridman, C., Petitprez, F., Calderaro, J., and Fridman, W. H. (2019). Tertiary Lymphoid Structures in the Era of Cancer Immunotherapy. *Nat. Rev. Cancer* 19 (6), 307–325. doi:10.1038/s41568-019-0144-6

Shitara, K., Iwata, H., Takahashi, S., Tamura, K., Park, H., Modi, S., et al. (2019). Trastuzumab Deruxtecan (DS-8201a) in Patients with Advanced HER2-Positive Gastric Cancer: a Dose-Expansion, Phase 1 Study. *Lancet Oncol.* 20 (6), 827–836. doi:10.1016/s1470-2045(19)30088-9

Sun, H., Wang, X., Zhang, Y., Che, X., Liu, Z., Zhang, L., et al. (2016). Biglycan Enhances the Ability of Migration and Invasion in Endometrial Cancer. *Arch. Gynecol. Obstet.* 293 (2), 429–438. doi:10.1007/s00404-015-3844-5

Thakur, A. K., Nigri, J., Lac, S., Leca, J., Bressy, C., Berthezene, P., et al. (2016). TAp73 Loss Favors Smad-independent TGF-β Signaling that Drives EMT in Pancreatic Ductal Adenocarcinoma. *Cell Death Differ* 23 (8), 1358–1370. doi:10.1038/cdd.2016.18

Tosolini, M., Kirilovsky, A., Mlecnik, B., Fredriksen, T., Mauger, S., Bindea, G., et al. (2011). Clinical Impact of Different Classes of Infiltrating T Cytotoxic and Helper Cells (Th1, Th2, Treg, Th17) in Patients with Colorectal Cancer. *Cancer Res.* 71 (4), 1263–1271. doi:10.1158/0008-5472.Can-10-2907

Van Cutsem, E., Sagaert, X., Topal, B., Haustermans, K., and Prenen, H. (2016). Gastric Cancer. *The Lancet* 388 (10060), 2654–2664. doi:10.1016/s0140-6736(16)30354-3

Varghese, F., Bukhari, A. B., Malhotra, R., and De, A. (2014). IHC Profiler: an Open Source Plugin for the Quantitative Evaluation and Automated Scoring of Immunohistochemistry Images of Human Tissue Samples. *PLoS One* 9 (5), e96801. doi:10.1371/journal.pone.0096801

Vuillermoz, B., Khoruzhenko, A., D'Onofrio, M.-F., Ramont, L., Venteo, L., Perreau, C., et al. (2004). The Small Leucine-Rich Proteoglycan Lumican Inhibits Melanoma Progression. *Exp. Cel Res.* 296 (2), 294–306. doi:10.1016/j.yexcr.2004.02.005

Yang, Y., Wang, R., Feng, L., Ma, H., and Fang, J. (2021). LINC00460 Promotes Cell Proliferation, Migration, Invasion, and Epithelial-Mesenchymal Transition of Head and Neck Squamous Cell Carcinoma via miR-320a/BGN Axis. *OncoTargets Ther.* 14, 2279–2291. doi:10.2147/OTT.S282947

Yousefi, H., Vatanmakanian, M., Mahdiannasser, M., Mashouri, L., Alahari, N. V., Monjezi, M. R., et al. (2021). Understanding the Role of Integrins in Breast Cancer Invasion, Metastasis, Angiogenesis, and Drug Resistance. *Oncogene* 40 (6), 1043–1063. doi:10.1038/s41388-020-01588-2

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, J.-P., Yan, J., Xu, J., Pang, X.-H., Chen, M.-S., Li, L., et al. (2009). Increased Intratumoral IL-17-producing Cells Correlate with Poor Survival in Hepatocellular Carcinoma Patients. *J. Hepatol.* 50 (5), 980–989. doi:10.1016/j.jhep.2008.12.033

Zhang, G.-Q., Zhao, H., Wu, J. Y., Li, J. Y., Yan, X., Wang, G., et al. (2015). Prolonged Overall Survival in Gastric Cancer Patients after Adoptive Immunotherapy. *World J. Gastroenterol.* 21 (9), 2777–2785. doi:10.3748/wjg.v21.i9.2777

Zhao, Q., Cao, L., Guan, L., Bie, L., Wang, S., Xie, B., et al. (2019). Immunotherapy for Gastric Cancer: Dilemmas and prospect. *Brief. Funct. Genomics* 18 (2), 107–112. doi:10.1093/bfgp/ely019

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 Genetics

ORIGINAL RESEARCH
published: 02 February 2022
doi: 10.3389/fgene.2022.848205

Predicting Treatment Response in Schizophrenia With Magnetic Resonance Imaging and Polygenic Risk Score

Meng Wang[1,2], *Ke Hu*[1,2], *Lingzhong Fan*[1,2,3], *Hao Yan*[4,5], *Peng Li*[4,5], *Tianzi Jiang*[1,2,3,6,7] *and Bing Liu*[8,9]*

[1]Brainnetome Center and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, [3]Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China, [4]Peking University Sixth Hospital/Institute of Mental Health, Beijing, China, [5]Key Laboratory of Mental Health, Ministry of Health (Peking University), Beijing, China, [6]Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China, [7]Innovation Academy for Artificial Intelligence, Chinese Academy of Sciences, Beijing, China, [8]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China, [9]Chinese Institute for Brain Research, Beijing, China

OPEN ACCESS

Edited by:
Ming Fan,
Hangzhou Dianzi University, China

Reviewed by:
Ming Li,
Kunming Institute of Zoology, China
Lixia Tian,
Beijing Jiaotong University, China

***Correspondence:**
Bing Liu
bing.liu@bnu.edu.cn

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

Received: *04 January 2022*
Accepted: *12 January 2022*
Published: *02 February 2022*

Citation:
Wang M, Hu K, Fan L, Yan H, Li P, Jiang T and Liu B (2022) Predicting Treatment Response in Schizophrenia With Magnetic Resonance Imaging and Polygenic Risk Score. Front. Genet. 13:848205. doi: 10.3389/fgene.2022.848205

Background: Prior studies have separately demonstrated that magnetic resonance imaging (MRI) and schizophrenia polygenic risk score (PRS) are predictive of antipsychotic medication treatment outcomes in schizophrenia. However, it remains unclear whether MRI combined with PRS can provide superior prognostic performance. Besides, the relative importance of these measures in predictions is not investigated.

Methods: We collected 57 patients with schizophrenia, all of which had baseline MRI and genotype data. All these patients received approximately 6 weeks of antipsychotic medication treatment. Psychotic symptom severity was assessed using the Positive and Negative Syndrome Scale (PANSS) at baseline and follow-up. We divided these patients into responders (N = 20) or non-responders (N = 37) based on whether their percentages of PANSS total reduction were above or below 50%. Nine categories of MRI measures and PRSs with 145 different p-value thresholding ranges were calculated. We trained machine learning classifiers with these baseline predictors to identify whether a patient was a responder or non-responder.

Results: The extreme gradient boosting (XGBoost) technique was applied to build binary classifiers. Using a leave-one-out cross-validation scheme, we achieved an accuracy of 86% with all MRI and PRS features. Other metrics were also estimated, including sensitivity (85%), specificity (86%), F1-score (81%), and area under the receiver operating characteristic curve (0.86). We found excluding a single feature category of gray matter volume (GMV), amplitude of low-frequency fluctuation (ALFF), and surface curvature could lead to a maximum accuracy drop of 10.5%. These three categories contributed more than half of the top 10 important features. Besides, removing PRS features caused a modest

accuracy drop (8.8%), which was not the least decrease (1.8%) among all feature categories.

Conclusions: Our classifier using both MRI and PRS features was stable and not biased to predicting either responder or non-responder. Combining with MRI measures, PRS could provide certain extra predictive power of antipsychotic medication treatment outcomes in schizophrenia. PRS exhibited medium importance in predictions, lower than GMV, ALFF, and surface curvature, but higher than measures of cortical thickness, cortical volume, and surface sulcal depth. Our findings inform the contributions of PRS in predictions of treatment outcomes in schizophrenia.

Keywords: schizophrenia, treatment prediction, XGBoost, polygenic risk score, magnetic resonance imaging

1 INTRODUCTION

Pharmacological therapy has long been the cornerstone of schizophrenia management, which aims to relieve psychotic symptoms, such as delusions, hallucinations, and disorganized thinking, et al. (Kane and Correll, 2010; Patel et al., 2014; Tarcijonas and Sarpal, 2019). Whereas, the treatment outcomes of antipsychotic medications generally vary significantly. According to statistics, approximately 10–30% of schizophrenia patients achieve little symptomatic amelioration after receiving multiple trials of typical antipsychotics. Meanwhile, an additional 30–60% of patients with schizophrenia show partial or inadequate improvement in psychotic symptoms (Patel et al., 2014). Further, the long-term disease courses in schizophrenia are even heterogeneous, which are formulated over time (Tarcijonas and Sarpal, 2019). There are twelve treatment trajectories summarized in an over 20-years follow-up study involving more than 500 patients with schizophrenia (Huber et al., 1980). The great variations of treatment outcomes are also confirmed in more recent studies (Carbon and Correll, 2014; Tarcijonas and Sarpal, 2019). Although varying degrees of remission are acquired in a great number of patients with schizophrenia, substantial evidence suggests that antipsychotic medications can lead to various adverse effects (Muench and Hamer, 2010; Patel et al., 2014; Stroup and Gray, 2018). To date, no clinical reliable quantitative markers can be employed to accurately predict the treatment response to antipsychotic medications of a patient with schizophrenia. Therefore, to avert unnecessary side effects, enable early intervention, and adopt appropriate treatments, it is critical to identify prognostic measures that can inform individual treatment outcomes in advance.

Toward this target, considerable efforts are made to identify predictors of antipsychotic treatment outcomes. Recently, magnetic resonance imaging (MRI) has been broadly applied in psychiatry researches, which provides quantitative *in vivo* measures of the brain (Quinlan et al., 2020; Voineskos et al., 2020; Kraguljac et al., 2021). Particularly, one significant area of these applications is the prediction of antipsychotic treatment responses or outcomes in patients with schizophrenia. Overall, a large number of studies focused on structural MRI measures. A longitudinal study of individuals with first-episode schizophrenia reported that the ventricular volume was significantly increased in patients with poor treatment outcomes, which was not observed in better treatment outcome patients and healthy controls (Lieberman et al., 2001). Another independent longitudinal study confirmed this and found schizophrenia patients with poor treatment outcomes had greater lateral ventricular enlargement over time (Ho et al., 2003). In a cross-sectional comparison study, conducted in schizophrenia patients with poor outcomes, favorable outcomes, and healthy individuals, poor outcome patients showed significantly smaller cerebral gray matter particularly in prefrontal regions, and increased volume in the lateral and third ventricles (Staal et al., 2001). A voxel-based comparison analysis of gray matter volume revealed that non-responder schizophrenia patients demonstrated a more severe atrophy pattern than responder patients, particularly in the superior and middle frontal gyri (Quarantelli et al., 2014). Compared with non-resistant schizophrenia patients, treatment-resistant patients showed a significant decrease of thickness in the left dorsolateral prefrontal cortex (Zugman et al., 2013). Cortical gyrification in bilateral insula, left frontal, and right temporal regions were significantly decreased in non-responder patients with first-episode schizophrenia compared with responders (Palaniyappan et al., 2013). Besides, non-responders had smaller thickness in the occipital lobe and smaller asymmetry in the frontal region compared with responders (Szeszko et al., 2012).

In addition to structural MRI, resting-state functional MRI has also been shown to provide prognostic markers. Functional connectivity was one of the most fully investigated measures. Using a seed-based approach, functional connectivity of the striatum with the dorsolateral prefrontal cortex, anterior cingulate, and limbic regions such as the hippocampus and anterior insula, were observed positively correlated with improvement of antipsychotic treatment in patients with first-episode schizophrenia. This relationship was converse when functional connectivity changed to the striatum with the parietal lobe (Sarpal et al., 2015). The prognostic capability of striatal connectivity was also demonstrated in other studies (Sarpal et al., 2016, 2017). Increased functional connectivity in the default mode network (DMN) with the ventromedial prefrontal cortex was found associated with greater efficacy of treatment using olanzapine in schizophrenia (Sambataro et al.,

2010). Besides, functional connectivity of the superior temporal cortex was utilized to successfully predict antipsychotic treatment responses in first-episode drug-naïve schizophrenia patients (Cao et al., 2020). Apart from static functional connectivity, dynamic functional connectivity within DMN regions was proved with the most predictive power of symptom change in schizophrenia compared with other common measures (Kottaram et al., 2020). Several other resting-state functional MRI derived measures were also examined to establish their relationships with treatment outcomes in schizophrenia, such as regional homogeneity (Gao et al., 2018) and amplitude of low-frequency fluctuation (Cui et al., 2019).

Schizophrenia is a highly polygenic disorder with thousands of associated risk loci, with mostly small effects (Smeland et al., 2020). Polygenic risk score (PRS) is a measure to assess an individual's genetic liability to schizophrenia, which is calculated by combining total risk alleles with corresponding weights derived from genome-wide association study results (Choi et al., 2020). In a recent study, PRS was verified as a predictor of antipsychotic efficacy in first-episode schizophrenia. Patients with higher PRS tended to be treatment non-responders than those with lower PRS (Zhang et al., 2019). However, it remains unclear whether PRS can markedly improve prognostication on the basis of MRI-derived predictors. If indeed better prediction performance is acquired when combining PRS and neuroimaging predictors, the precedence of the predictive capability of these predictors requires to be investigated.

In the present study, we worked on the problem and hypothesized that PRS can provide additional prognostic power combined with MRI predictors. We collected a total of 57 patients with schizophrenia, which were divided into responders and non-responders according to their 6 weeks of antipsychotic treatment outcomes. Various neuroimaging and PRS features were calculated. We constructed machine learning classifiers with these baseline features to identify responders or non-responders. Particularly, we concentrated on 1) performance comparison of a classifier trained using a combination of MRI and PRS features with a classifier trained using single MRI features; 2) relative importance or contributions of these features to predictions.

2 MATERIALS AND METHODS

2.1 Participants and Clinical Assessments

Individuals with schizophrenia (N = 97, before screening) were recruited from Peking University Sixth Hospital and Beijing Huilongguan Hospital, whose imaging data were all obtained on a 3.0T Siemens TrioTim MRI scanner. Diagnoses were made by qualified clinicians using the Structured Clinical Interview for DSM-IV. All participants had no history of other DSM-IV Axis I disorders, neurological disorders, cognitive deficits, severe physical diseases, serious head trauma, substance abuse or dependence, and electroconvulsive therapy within the last 6 months. Each individual was treated with only a single second-generation antipsychotic drug, although the specific drug is not totally the same across patients (mainly including risperidone and clozapine). The study was approved by the Medical Research Ethics Committees of the local hospitals. All individuals or their guardians provided written informed consent. Participants were excluded if their clinical assessments at baseline or follow-up were incomplete, or they lacked sMRI, rsfMRI, or genotype data. Quality control (QC) for rsfMRI data was completed by examining the framewise displacement (FD) (Power et al., 2012). Individuals who had a mean FD greater than 0.3 mm were precluded. Besides, subjects were also excluded if they failed to genotyping QC. In total, 57 subjects remained after the screening.

The symptom severity of patients with schizophrenia was evaluated using the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987) by trained clinical psychiatrists. Baseline assessments were completed within 1 week of image acquisition. Follow-up assessments were performed after approximately 6 weeks of antipsychotic treatment. **Table 1** shows demographics and clinical characteristics.

2.2 Image Acquisition and Preprocessing

All images were acquired on a 3.0T Siemens TrioTim scanner. Two-dimension echo-planar imaging (EPI) was used for rsfMRI data with parameters: repetition time (TR) = 2000 ms; echo time (TE) = 30 ms; flip angle (FA) = 90°; field of view (FOV) = 220 × 220 mm^2; matrix size = 64 × 64; voxel dimensions = 3.4375 × 3.4375 × 4.6 mm^3; 240 volumes, and 33 slices. For T1-weighted (T1w) structural images, three-dimension magnetization-prepared rapid gradient-echo (MPRAGE) sequence was performed with parameters: TR = 2,530 ms; TE = 3.5 ms; FA = 7°; inversion time (TI) = 1,100 ms; voxel dimensions = 1 × 1 × 1 mm^3; matrix size = 256 × 256 × 192.

Preprocessing of rsfMRI data was performed using the BRANT toolkit (Xu et al., 2018, https://github.com/kbxu/brant). In brief, several standardized procedures were carried out, including discarding the first ten timepoints, slice timing correction, realignment, coregistration, spatial normalization to Montreal Neurological Institute (MNI) space, resampling, regressing out nuisances of linear trends, global signal as well as head-motion parameters, and performing temporal band-pass filtering at 0.01–0.08 Hz.

2.3 Genotype Data Acquisition and Preprocessing

The procedures of genotype data collection and preprocessing were elaborately described in our previous studies (Liu et al., 2017; Hu et al., 2021). Briefly, for all individuals, ethylene diamine tetraacetic acid (EDTA) anti-coagulated venous blood samples were obtained, from which genomic DNA data were extracted using the EZgene Blood gDNA Miniprep Kit. The whole-genome genotyping was carried out on Illumina Human OmniZhongHua-8 BeadChips with the standard Illumina genotyping protocol.

Genotype processing and QC was implemented using PLINK version 1.07 (Purcell et al., 2007), following the subsequent steps: 1) excluded subjects with missing genotype rates more than 0.05; 2) identified subject pairs with highly similar genotypes and

TABLE 1 | Demographics and clinical information of participants.

	Individuals with schizophrenia (*N* = 57)		
–	**Responder (*N* = 20)**	**Non-responder (*N* = 37)**	***p* value**
Age (years)	25.22 ± 5.4	28.35 ± 7.3	0.10
Sex (male/female)	7/13	20/17	0.27
PANSS total score at baseline	76.90 ± 8.3	79.21 ± 7.8	0.31
PANSS total score at follow-up	44.15 ± 12.4	65.29 ± 8.1	4.30e-10
Percentage reduction of PANSS total score	71.19 ± 27.1%	28.05 ± 13.2%	1.18e-10
Chlorpromazine equivalent dosage (mg/day)	418.42 ± 280.6	531.03 ± 367.9	0.27

PANSS, positive and negative syndrome scale; Data were shown as mean ± standard deviation.

removed the one who had a greater missing genotype rate; 3) removed single nucleotide polymorphisms (SNPs) if their missing genotype rates greater than 0.05, with a minor allele frequency less than 0.01, and significantly deviated from Hardy-Weinberg Equilibrium ($p < 0.001$); 4) used EIGENSTART (Patterson et al., 2006; Price et al., 2006) for principal component analysis (PCA) on linkage disequilibrium (LD) pruned set of autosomal SNPs, which were obtained from LD pruning and removing five long-range LD regions using the HapMap phase three reference data set (Thorisson et al., 2005). Outliers of samples with more than six SD were excluded after achieving 10 principal components; 5) imputation was completed using SHAPEIT (Delaneau et al., 2011) and IMPUTE2(Howie et al., 2009) referred to the 1,000 Genomes phase one dataset. The autosomal SNPs with imputation quality scores greater than 0.8 were further analyzed.

2.4 Predictors and Clinical Outcome

We calculated diverse predictors (features) based on imaging and genotype data and divided subjects into responder and non-responder groups according to clinical outcomes.

2.4.1 Responder and Non-responder

For each individual, the clinical outcome was measured by percentage reduction of PANSS total score relative to baseline, which was calculated as follows:

$$\Delta = \frac{\mathrm{PANSS}_{\mathrm{baseline}} - \mathrm{PANSS}_{\mathrm{followup}}}{\mathrm{PANSS}_{\mathrm{baseline}} - 30} \times 100\%$$

The subtracted value of 30 in the denominator indicates a minimum score of "no symptoms" assessed using PANSS. We defined an individual as a responder in case that the patient achieved a at least 50% reduction of PANSS total score. Subjects not satisfying this criterion were regarded as non-responders. The cut-off threshold was specified at 50%, given that this value roughly reflects a "much improved" condition for acutely ill and non-refractory patients from a clinical perspective (Leucht et al., 2009). Although the statistical power might be reduced when dichotomizing the continuous clinical outcome, it provides a clear and interpretable measure instead (Lewis, 2004; Kottaram et al., 2020).

2.4.2 Gray Matter Volume

Voxel-based morphometry analysis was performed using the VBM8 toolbox (Matsuda et al., 2012, http://dbm.neuro.uni-jena.de/vbm8/), which runs within the SPM8 software (https://www.fil.ion.ucl.ac.uk/spm/software/spm8/). For each subject, the native T1w image was segmented into tissue images of gray matter, white matter, and cerebrospinal fluid, which were then registered to the standard MNI space through non-linear deformation using the high dimensional DARTEL algorithm (Ashburner, 2007). All non-brain tissues were removed in the process. Smoothing was not applied. Each segmented image had a voxel size of 1.5 mm with a resolution of 121 × 145 × 121. Quality control was completed by displaying slices for segmented images and inspecting sample homogeneity. For each gray matter image, we extracted mean gray matter volumes from each of the brain parcellations defined in the Brainnetome atlas (Fan et al., 2016, https://atlas.brainnetome.org/download.html), resulting in a total of 246 regional values.

2.4.3 Cortical Morphologies

Cortical reconstruction was performed on raw T1w images using FreeSurfer version 6.0 (Dale et al., 1999, https://surfer.nmr.mgh.harvard.edu/fswiki/rel6downloads). For each individual, this process estimated various vertex-based cortical surface morphological measures. Quality control was performed by visually examining any errors in the whole reconstruction process. To precisely match cortical locations among subjects, we aligned each reconstructed cortical surface with the fsaverage template, which had 163,842 vertices per hemisphere. We selected five cortical morphologies in the study, including surface area, curvature, sulcal depth, thickness, and volume. As with GMV, we used the Brainnetome parcellations to extract averaged cortical values, resulting in 210 values for each measure. The atlas is already resampled to fsaverage space. Finally, for each individual, we calculated 210 (number of cortical parcellations) × 5 (number of measures) values in total.

2.4.4 Amplitude of Low-Frequency Fluctuation

ALFF is a rsfMRI measure that quantifies the amplitude of spontaneous low-frequency fluctuations of time series signals (Zang et al., 2007). We used the BRANT toolkit to estimate a voxel-based ALFF map for each individual. To be specific, the fast Fourier transform algorithm was first applied to transform time series into the frequency domain and the corresponding power spectrum was achieved. Next, square root values were calculated at each frequency within the spectrum. ALFF was defined as the mean square root across the frequency range of 0.01–0.08 Hz. The rsfMRI data were not performed temporal band-pass

filtering before estimating ALFF maps to avoid possible effects. Finally, each ALFF map was normalized by subtracting the global mean then dividing by the global standard deviation to eliminate inter-subject biases. Likewise, we extracted mean values from ALFF maps based on the Brainnetome atlas and obtained 246 regional values for each individual.

2.4.5 Regional Homogeneity

ReHo measures the similarity between the time series in a given voxel and those in its 26 neighboring voxels based on Kendall's coefficient of concordance (Zang et al., 2004). It is a reflection of synchronization between the time series of a given voxel and its neighbors. We also used the BRANT toolkit to calculate the ReHo map for each subject. Normalization was performed on each ReHo map by dividing the global mean intensity. As with the ALFF map, for each individual, we extracted 246 values from the ReHo map according to the Brainnetome atlas.

2.4.6 Functional Connectivity

For each subject, whole-brain FCs were calculated based on the Brainnetome atlas. We first extracted the mean time series from each of the 246 brain regions defined in the atlas. Then we calculated Pearson's correlations between the extracted time series of each region pair. Particularly, there were $(246 \times 245)/2 = 30{,}135$ unique pairs of regions. We obtained 30,135 FCs for each subject, which was substantially greater than the number of total individuals ($N = 57$). Thus we further performed dimensional reduction by applying PCA on FCs from all subjects and achieved 50 principal components, accounting for 95% amount of variance.

2.4.7 Genetic Characteristics

We calculated step-wise polygenic risk scores (PRSs) for each individual with identical procedures in our prior study (Hu et al., 2021). The PRSs were computed using PLINK version 1.07 (Purcell et al., 2007) and genome-wide association study (GWAS) data from a large number of Chinese individuals (Li et al., 2017). Of note, our study cohort was independent of subjects from the GWAS study, despite they matched in ancestries. We established a list of separate p-value threshold ranges to aggregate SNPs. Specifically, we set step lengths of 0.001 and 0.01 for [0, 0.05) and [0.05, 1) intervals, respectively. The left square bracket and the right parenthesis denoted inclusion and exclusion cut-off values, separately. Consequently, there were 145 PRSs computed for each individual with distinct SNP inclusion thresholds: [0, 0.001), [0.001, 0.002), . . ., [0.049, 0.05), [0.05, 0.06), [0.06, 0.07), . . ., [0.99, 1).

2.5 Classification

We sought to build classification models from a combination of features derived from imaging and genotype data to predict whether a patient with schizophrenia was a responder or a non-responder after receiving 6 weeks of antipsychotic treatment.

2.5.1 Model Building, Training, and Testing

To deal with this prediction problem, we employed extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016) to build binary classifiers to predict individual treatment outcomes. XGBoost is a scalable machine learning system for tree boosting and is publicly available as an open-source package (https://github.com/dmlc/xgboost). We chose the XGBoost method mainly for its significant and broadly recognized impact on various machine learning and data mining challenges (Chen and Guestrin, 2016), as well as its successful applications in brain imaging prediction tasks (Torlay et al., 2017; Sharma and Verbeke, 2020).

We calculated several categories of predictors (features): 1) GMV with 246 regional values, 2) cortical morphologies of surface area, curvature, sulcal depth, thickness, and volume, each of which had 210 values, 3) rsfMRI measures of ALFF (246 values), ReHo (246 values), and FC (50 values), as well as 4) 145 genetic features of PRS. In total, 1983 features were computed. All these categories of features were combined to train XGBoost classifiers. Given the modest sample size of the studied cohort, we applied a leave-one-out cross-validation (LOOCV) strategy to validate classifier performance, which is supposed appropriate for small datasets and used in similar tasks (Cao et al., 2020; Kottaram et al., 2020). Specifically, iteratively held out one subject for validation, and used the rest to train the model until all the subjects were validated once. The eventual result was computed by taking the mean of all the subject validations. Several established measures were calculated for evaluations of classification performance, including accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (ROC-AUC).

It is known XGBoost models tend to contain larger hyperparameter sets compared with basic machine learning classifiers, such as logistic regression, support vector machine, et al. Thus hyperparameter tuning is of great importance to leverage the maximum power of this method. Originally, all parameters were assigned to default values. We tuned one parameter each time and kept the others constant to examine changes in classifier performance as the variation of the specified parameter by performing repetitive LOOCV procedures. In this way, we identified which parameters were relatively important that significantly influenced classifier performance, and which parameters had minor impacts on model performance. We also estimated certain value ranges for each of these crucial parameters. Of note, these value ranges were determined separately, which we considered might constitute a possible optimal searching space. Finally, we concentrated on these significant parameters and performed a fine-grained grid search on the estimated value ranges. Besides, due to the imbalanced sample sizes between responders and non-responders, we calculated the sample weights that were inversely proportional to class frequencies and applied them when fitted models.

2.5.2 Feature Importance

A valuable benefit of using the XGBoost method is that it automatically provides estimates of feature importance from a trained predictive model. Generally, we can directly retrieve importance scores for each feature, which measure how useful or valuable each feature is in the construction of the boosting tree

TABLE 2 | Optimal hyperparameters set of XGBoost classifier for leave-one-out cross-validation.

Parameters	Description	Value
n_estimators	Number of boosting rounds	50
max_depth	Maximum tree depth for base learners	2
learning_rate	Boosting learning rate	0.12
booster	Specify which booster to use: gbtree, gblinear, or dart	gbtree
gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree	0.01
subsample	Subsample ratio of the training instance	0.90
colsample_bytree	Subsample ratio of columns when constructing each tree	0.30
colsample_bylevel	Subsample ratio of columns for each level	0.50
colsample_bynode	Subsample ratio of columns for each split	0.30
reg_alpha	L1 regularization term on weights	0.10
reg_lambda	L2 regularization term on weights	1.65
scale_pos_weight	Balancing of positive and negative weights	2.50

Other hyperparameters not listed in the table were set to default values. The description referred to the XGBoost documentation at https://xgboost.readthedocs.io/en/latest/index.html.

model. The importance can be quantified using several metrics provided by XGBoost, such as gain, coverage, weight, total gain, total coverage. We specified the gain metric for our models, which is supposed as the most relevant attribute to interpret the relative importance of each corresponding feature. A feature is considered more important for generating a prediction if its gain value is higher compared to another feature.

In addition to estimating feature importance through the trained classifier itself, we also evaluated the contributions of feature categories. Specifically, we removed one feature category, such as GMV or cortical thickness, and used all the remaining features to reconstruct predictive models with identical procedures as our main analysis in which all feature categories were used. We determined the contribution of each feature category by evaluating performance change (e.g., accuracy) between each newly built classifier and our main model. If removing a feature category led to a maximum decrease in performance, then this feature category was considered to contribute most to predictions.

3 RESULTS

3.1 Predicting Treatment Response in Schizophrenia

Individuals with schizophrenia were reasonably defined as responders (N = 20) or non-responders (N = 37) according to their amelioration degrees of overall symptom severity, which was assessed using PANSS total score, after accepting 6 weeks of antipsychotic medications treatment. The responders and non-responders were matched in age and sex. There were also no significant differences between the two groups in baseline PANSS total score and chlorpromazine equivalent dosage (**Table 1**). We calculated a multitude of predictors (features), spanning categories of 1) structural imaging (GMV; cortical morphologies of surface area, curvature, sulcal depth, thickness, and volume), 2) functional imaging (ALFF; ReHo; FC), and 3) genetic characteristics (step-wise PRS). Combined with both imaging and genetic features, we constructed binary machine learning classifiers using the XGBoost method to predict individual treatment outcomes (i.e., responder or non-responder). We applied a leave-one-out cross-validation (LOOCV) scheme to validate model performance, and reported several estimated classification metrics to provide a comprehensive evaluation. The XGBoost classifiers were trained with carefully hyperparameters fine-tuning processes. **Table 2** shows the optimal hyperparameters set for LOOCV.

We observed the classification accuracy reached a relatively high score of 86% (**Table 3**). There were eight misclassified individuals altogether, of which 4 subjects were near the cut-off boundary of treatment outcomes (i.e., the 50% threshold). The corresponding percentage reductions of PANSS total score of these four subjects were 45, 43, 40, and 48%. Particularly, several additional metrics that quantify model performance exceeded 80% (**Table 3**), including sensitivity (85%), specificity (86%), F1-score (81%), ROC-AUC (0.86). Meanwhile, the ROC curve demonstrated our classification results were far higher than the chance level (**Figure 1**). Taken together, our classifiers had high predictive power and were not biased to a certain class.

3.2 Evaluating Feature Contributions

To quantify feature importance, we selected the classifier that performed the best on the LOOCV procedure (hyperparameter values of this model were given in **Table 2**). After retraining the classifier on the whole dataset, we directly obtained the importance score of each feature from the 'feature_importances_" attribute in the fitted model. Typically, a higher importance score implied the corresponding feature was relatively more important in predictions. Among the top 10 important features, nine features were derived from structural imaging, which involved categories of GMV, cortical thickness, cortical volume, surface sulcal depth, and surface curvature. There was only one functional imaging feature (i.e., ALFF), and no genetic features existed (**Table 4**). Particularly, the GMV in a certain region of the left inferior frontal gyrus (labeled 31 corresponded to the Brainnetome atlas) ranked the first important. When examining the top 100 important features, all the 10 feature categories were involved (**Figure 2**). More than half of these 100 features belonged to three categories, which were GMV, ALFF, and cortical thickness containing 27, 14, and 13 features respectively.

Besides, we further evaluated the prediction contributions of each feature category. In brief, after iteratively removing one feature category, we built XGBoost classifiers with the remaining

TABLE 3 | Performance of predicting individual treatment outcomes with all imaging and genetic features.

Performance metrics	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	ROC-AUC
Classification results	85.96	85	86.49	80.95	0.86

ROC-AUC, area under the receiver operating characteristic curve. Responder/non-responder = 20/37.

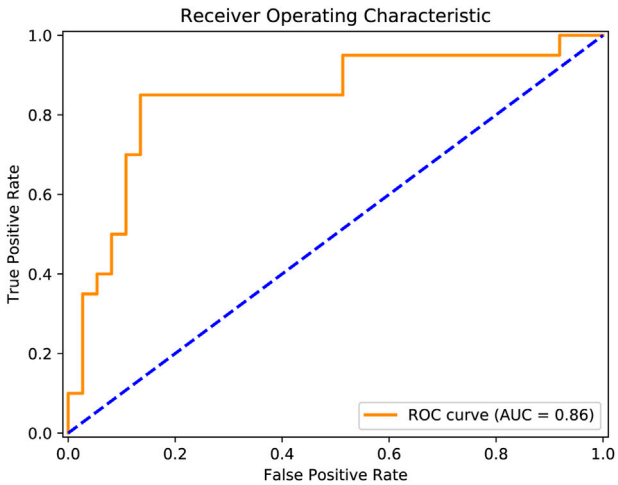

FIGURE 1 | Prediction performance was quantified using the receiver operating characteristic curve. The orange solid line reflected actual classification results, and the blue dashed line indicated the chance level.

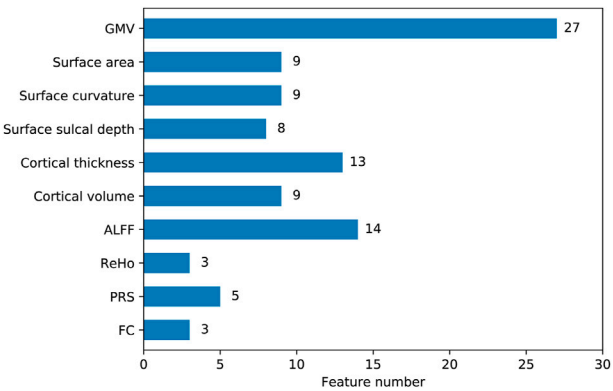

FIGURE 2 | The number of features belonged to each category among the top 100 important features. The y axis represented feature categories. The values labeled on the right of the bars were actual feature numbers.

TABLE 4 | Top 10 important features obtained from the XGBoost classifier trained on the whole dataset.

Rank	Feature category	Atlas region number	Description	Importance score
1	GMV	31	IFG_L_6_2	0.04138
2	Cortical thickness	157	PoG_L_4_2	0.03584
3	GMV	14	SFG_R_7_7	0.03205
4	ALFF	119	PhG_L_6_6	0.03048
5	Cortical thickness	42	OrG_R_6_1	0.03028
6	Cortical volume	189	MVOcC _L_5_1	0.02930
7	GMV	15	MFG_L_7_1	0.02723
8	Surface sulcal depth	210	LOcC _R_2_2	0.02637
9	Surface curvature	152	PCun_R_4_3	0.02594
10	Surface curvature	169	INS_L_6_4	0.02591

IFG, inferior frontal gyrus; PoG, postcentral gyrus; SFG, superior frontal gyrus; PhG, parahippocampal gyrus; OrG, orbital gyrus; MVOcC, medioventral occipital cortex; MFG, middle frontal gyrus; LOcC, lateral occipital cortex; Pcun, precuneus; INS, insular gyrus. L (R), left (right) hemisphere. The atlas region number corresponded to the Brainnetome parcellation (Fan et al., 2016).

features following the main analyses to investigate how the performance changed. We found removing any one of these 10 feature categories could lead to a performance drop (**Table 5**). Specifically, four quantitative metrics including accuracy, sensitivity, F1-score, ROC-AUC decreased consistently, in which the sensitivity measure dropped the most with an average of 21.5%. The specificity had a slight increase (at most 5.4%) in three of the 10 classifiers, indicating a higher bias existed in the three models. In terms of accuracy, the categories of GMV, ALFF, and surface curvature contributed the most to predictions, given removing one of these three categories led to a maximum drop in accuracy score (10.5%). The cortical volume was the least important, since removing this category caused a minimal accuracy decrease (1.8%). PRS exhibited medium importance, excluding of which led to a modest accuracy drop (8.8%).

4 DISCUSSION

Tremendous evidence has suggested that neuroimaging data coupled with machine learning techniques can provide favorable utilities of prognostic predictions in psychiatric

TABLE 5 | Prediction performance of classifiers trained with features after removing certain categories.

Feature categories used	Number of features	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	ROC-AUC
No GMV	1737	75.44	65	81.08	65	0.73
No surface area	1773	77.19	65	83.78	66.67	0.74
No surface curvature	1773	75.44	60	83.78	63.16	0.72
No surface sulcal depth	1773	78.95	65	86.49	68.42	0.76
No cortical thickness	1773	80.70	65	89.19	70.27	0.77
No cortical volume	1773	84.21	70	91.89	75.68	0.81
No ALFF	1737	75.44	60	83.78	63.16	0.72
No ReHo	1737	77.19	55	89.19	62.86	0.72
No FC	1933	77.19	60	86.49	64.86	0.73
No PRS	1838	77.19	70	81.08	68.29	0.76

disorders, including schizophrenia (Janssen et al., 2018). A recent relevant study investigated the ranking of predictive capabilities of multiple neuroimaging and clinical measures when predicting the relative change of symptom severity in schizophrenia at 1-year follow-up (Kottaram et al., 2020). The evaluated neuroimaging predictors included structural imaging measures of cortical thickness and gray matter volume as well as functional imaging derived measures of static and dynamic resting-state connectivity. From the aspect of genetic factors, another recent study examined the relationship between polygenic risk scores (PRSs) and antipsychotic drug treatment outcomes in patients with schizophrenia (Zhang et al., 2019). However, it remains unclear 1) whether neuroimaging combined with PRS can provide better prognostic performance than merely using neuroimaging features; and 2) which category of neuroimaging predictors or PRS provides the most accurate prognostic power, and what is the ranking of their importance or contributions. To address these issues, we collected a cohort of patients with schizophrenia (N = 57), all of which had baseline neuroimaging and genotype data. All these patients received about 6 weeks of antipsychotic medication treatment. Psychotic symptoms were assessed using PANSS at baseline and follow-up. The patients were grouped into responders or non-responders according to their percentages of PANSS total reduction. We calculated various predictors, including 1) six structural imaging measures (GMV; cortical morphologies of surface area, curvature, sulcal depth, thickness, and volume); 2) three resting-state functional imaging measures (ALFF; ReHo; FC), and 3) step-wise PRS. We trained binary machine learning classifiers with these baseline features to identify whether a patient with schizophrenia was a responder or non-responder.

Overall, we achieved an accuracy of 86% when predicting antipsychotic drug treatment outcomes (i.e., responders or non-responders) of patients with schizophrenia using all feature categories (**Table 3**). As far as we know, this performance exceeds the vast majority of results in previous studies and is also more than reported in a recent study (Kottaram et al., 2020). The performance was evaluated using a LOOCV procedure, considering our modest sample size (N = 57). Although this scheme is supposed to yield unstable estimates of predictive performance (Varoquaux et al., 2017), it is frequently employed in numerous neuroimaging studies, especially in those with relatively small sample sizes (Cao et al., 2020; Kottaram et al., 2020). Specifically, in small datasets, LOOCV can provide sufficient data for training compared with other k-fold cross-validation schemes. In addition to accuracy, we found all other estimated classifier metrics were also at a relatively higher level (**Table 3**), such as sensitivity (85%), specificity (86%), F1-score (81%), ROC-AUC (0.86) (**Figure 1**). These extra quantifications further demonstrated our classifier was stable and not biased to predicting either responder (N = 20) or non-responder (N = 37).

We examined the top 10 important features in predictions and found nine of them were structural imaging measures, including three GMV, two cortical thickness, two surface curvature, one cortical volume, and one surface sulcal depth, one was functional imaging measure of ALFF (**Table 4**). PRS features were not of top 10 importance. The three GMV features were all extracted from the frontal lobe regions, including inferior, superior, and middle frontal gyri. Particularly, GMV in the inferior frontal gyrus was the most prominent predictor. Previous studies have revealed GMV reductions in the frontal lobe regions were associated with poor antipsychotic medication treatment in patients with schizophrenia (Staal et al., 2001; Quarantelli et al., 2014; Tarcijonas and Sarpal, 2019). Consistently, significant reductions of GMV in the superior and middle frontal gyri were observed in non-responders (Quarantelli et al., 2014). The two cortical thickness features were estimated from the postcentral gyrus in the parietal lobe and the orbital gyrus in the frontal lobe. However, these two regions were discrepant with prior reported regions of the occipital gyrus (Szeszko et al., 2012) and the dorsolateral prefrontal cortex (Zugman et al., 2013). The remaining five features were barely investigated in similar studies, which covered regions of the left parahippocampal gyrus (ALFF), left medioventral occipital cortex (cortical volume), right lateral occipital cortex (surface sulcal depth), right precuneus, and left insular gyrus (surface curvature). When focusing on the top 100 significant predictors, we found all feature categories were involved (**Figure 2**). Particularly, the top three categories that contained the most features were GMV, ALFF, and cortical thickness, comprising 27, 14, and 13 features respectively. Thus it was straightforward to explain the results that excluding GMV or ALFF features caused the most performance drop of accuracy (10.5%; **Table 5**). Notably, removing surface curvature features also led to the maximum decrease of accuracy (i.e., 10.5%). Collectively, we considered that GMV, ALFF, and surface curvature features had

relatively higher prognostic utilities compared to other feature categories. We observed that removing PRS features gave rise to a modest accuracy drop (8.8%), which was not the least decrease (1.8%) among all categories. This pointed out that PRS features can provide extra prognostic power combined with MRI features, and yet their importance or contributions were between minimum and maximum, inferior to certain MRI measures such as GMV, ALFF, and surface curvature.

There were a few considerations when dealing with predictors and clinical outcomes. We prepared various MRI features, aiming to cover as many measures as possible that were reported in prior relevant studies. We assumed that combining these features would be of great benefit to prognostication since each identified measure could provide certain prognostic information. In our study, although nine MRI measures were computed, more than any previous study used, some were still needed to be examined. For example, the dynamic resting-state functional connectivity measure within the default mode network was demonstrated as the most single accurate predictor of symptom severity change in schizophrenia (Kottaram et al., 2020). As for PRS calculation, it is known that the *p*-value threshold is critical given that only those SNPs with a GWAS association *p*-value below the threshold are included in the procedure (Choi et al., 2020). To avoid potential thresholding effects and duplication of SNPs, 145 step-wise PRSs were calculated as in our previous study (Hu et al., 2021). We defined patients with schizophrenia as responders or non-responders based on their reductions of PANSS total score, which is commonly applied in current practice (Leucht et al., 2009; Cao et al., 2020). However, this approach only focuses on the relative change of PANSS total scores between follow-up and baseline but ignores the actual symptom severity, which can not reflect a clinically significant change. For example, a patient remains highly symptomatic even achieving a 50% reduction of PANSS total score from 120 to 60. Thus it is necessary to further assess whether our features are prognostic of symptom severity (above or below a clinically meaningful cut-off) at follow-up. Another problem is the selection of threshold values, which determines whether a patient is a responder or non-responder. We chose a threshold of 50% in the study, which indicates a much-improved condition for acute patients (Leucht et al., 2009). Different thresholds were proved crucial to clinical trials (Leucht et al., 2007). Therefore, future studies should evaluate prognostications for non-thresholded (i.e., regression analyses) or various fine-step thresholds of PANSS total reductions.

Several limitations need to be considered. First, our sample of patients with schizophrenia was limited for machine learning algorithms, especially for the powerful XGBoost technique (Chen and Guestrin, 2016), which contains more hyperparameters than simple methods such as support vector machines. Although we applied a rational cross-validation strategy, the danger of overfitting can not be eliminated (Varoquaux et al., 2017; Varoquaux, 2018). Larger independent sample replication is required to evaluate the generalizability of our methods. Second, our MRI measures were all calculated based on the Brainnetome atlas (Fan et al., 2016). The choice of brain atlases should not be arbitrary, since it could lead to different results such as in discrimination analysis (Zang et al., 2021). Although we employed a fine-grained parcellation, which contains information on both anatomical and functional connections, comparisons between various brain atlases need to be accomplished. Third, our prediction study just focused on PANSS total reduction, however, it is essential to investigate whether reductions of PANSS subscales (i.e., positive, negative, and general psychopathology) or even specific symptom dimensions could be predicted.

5 CONCLUSION

Polygenic risk score for schizophrenia can provide certain prognostic power when combined with neuroimaging features to predict 6 weeks of antipsychotic medication treatment outcomes in patients with schizophrenia. The relative importance of the polygenic risk score in predictions is between maximum and minimum, lagging behind some neuroimaging measures such as gray matter volume, the amplitude of low-frequency fluctuation, and surface curvature. Overall, our findings inform contributions of the polygenic risk score in machine learning studies that aim to predict treatment outcomes in schizophrenia.

DATA AVAILABILITY STATEMENT

All the neuroimaging features, polygenic risk scores, and codes used in the study are publicly available at https://github.com/BingLiu-Lab/predict_treatment_outcome_schizophrenia. The raw genotype data are not publicly available but can be obtained by interested researchers upon official request and ethical approval by contacting the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Research Ethics Committees of Peking University Sixth and Beijing Huilongguan Hospitals. All participants or their guardians provided written informed consent.

AUTHOR CONTRIBUTIONS

BL and TJ led the project. MW and BL performed data analyses and prepared the manuscript. KH and LF participated in the discussions of the results and the manuscript. HY and PL contributed to the data acquisition.

FUNDING

This work was supported by the National Key Basic Research and Development Program (973) (Grant 2011CB707800) and the Natural Science Foundation of China (Grant Numbers 81771451 and 82071505).

REFERENCES

Ashburner, J. (2007). A Fast Diffeomorphic Image Registration Algorithm. *NeuroImage* 38, 95–113. doi:10.1016/j.neuroimage.2007.07.007

Cao, B., Cho, R. Y., Chen, D., Xiu, M., Wang, L., Soares, J. C., et al. (2020). Treatment Response Prediction and Individualized Identification of First-Episode Drug-Naïve Schizophrenia Using Brain Functional Connectivity. *Mol. Psychiatry* 25, 906–913. doi:10.1038/s41380-018-0106-5

Carbon, M., and Correll, C. U. (2014). Clinical Predictors of Therapeutic Response to Antipsychotics in Schizophrenia. *Dialogues Clin. Neurosci.* 16, 505–524. doi:10.31887/dcns.2014.16.4/mcarbon

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 (San Francisco California USA: ACM), 785–794. doi:10.1145/2939672.2939785

Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: a Guide to Performing Polygenic Risk Score Analyses. *Nat. Protoc.* 15, 2759–2772. doi:10.1038/s41596-020-0353-1

Cui, L.-B., Cai, M., Wang, X.-R., Zhu, Y.-Q., Wang, L.-X., Xi, Y.-B., et al. (2019). Prediction of Early Response to Overall Treatment for Schizophrenia: A Functional Magnetic Resonance Imaging Study. *Brain Behav.* 9, e01211. doi:10.1002/brb3.1211

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage* 9, 179–194. doi:10.1006/nimg.1998.0395

Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A Linear Complexity Phasing Method for Thousands of Genomes. *Nat. Methods* 9, 179–181. doi:10.1038/nmeth.1785

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., et al. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cereb. Cortex* 26, 3508–3526. doi:10.1093/cercor/bhw157

Gao, S., Lu, S., Shi, X., Ming, Y., Xiao, C., Sun, J., et al. (2018). Distinguishing between Treatment-Resistant and Non-Treatment-Resistant Schizophrenia Using Regional Homogeneity. *Front. Psychiatry* 9, 282. doi:10.3389/fpsyt.2018.00282

Ho, B.-C., Andreasen, N. C., Nopoulos, P., Arndt, S., Magnotta, V., and Flaum, M. (2003). Progressive Structural Brain Abnormalities and Their Relationship to Clinical Outcome: A Longitudinal Magnetic Resonance Imaging Study Early in Schizophrenia. *Arch. Gen. Psychiatry* 60, 585–594. doi:10.1001/archpsyc.60.6.585

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genet.* 5, e1000529. doi:10.1371/journal.pgen.1000529

Hu, K., Wang, M., Liu, Y., Yan, H., Song, M., Chen, J., et al. (2021). Multisite Schizophrenia Classification by Integrating Structural Magnetic Resonance Imaging Data with Polygenic Risk Score. *NeuroImage: Clin.* 32, 102860. doi:10.1016/j.nicl.2021.102860

Huber, G., Gross, G., Schuttler, R., Linz, M., and Clemens, S. (1980). Longitudinal Studies of Schizophrenic Patients. *Schizophr. Bull.* 6, 592–605. doi:10.1093/schbul/6.4.592

Janssen, R. J., Mourão-Miranda, J., and Schnack, H. G. (2018). Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 3, 798–808. doi:10.1016/j.bpsc.2018.04.004

Kane, J. M., and Correll, C. U. (2010). Past and Present Progress in the Pharmacologic Treatment of Schizophrenia. *J. Clin. Psychiatry* 71, 1115–1124. doi:10.4088/JCP.10r06264yel

Kay, S. R., Fiszbein, A., and Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophr. Bull.* 13, 261–276. doi:10.1093/schbul/13.2.261

Kottaram, A., Johnston, L. A., Tian, Y., Ganella, E. P., Laskaris, L., Cocchi, L., et al. (2020). Predicting Individual Improvement in Schizophrenia Symptom Severity at 1-year Follow-up: Comparison of Connectomic, Structural, and Clinical Predictors. *Hum. Brain Mapp.* 41, 3342–3357. doi:10.1002/hbm.25020

Kraguljac, N. V., McDonald, W. M., Widge, A. S., Rodriguez, C. I., Tohen, M., and Nemeroff, C. B. (2021). Neuroimaging Biomarkers in Schizophrenia. *Ajp* 178, 509–521. doi:10.1176/appi.ajp.2020.20030340

Leucht, S., Davis, J. M., Engel, R. R., Kane, J. M., and Wagenpfeil, S. (2007). Defining 'Response' in Antipsychotic Drug Trials: Recommendations for the Use of Scale-Derived Cutoffs. *Neuropsychopharmacol* 32, 1903–1910. doi:10.1038/sj.npp.1301325

Leucht, S., Davis, J. M., Engel, R. R., Kissling, W., and Kane, J. M. (2009). Definitions of Response and Remission in Schizophrenia: Recommendations for Their Use and Their Presentation. *Acta Psychiatr. Scand.* 119, 7–14. doi:10.1111/j.1600-0447.2008.01308.x

Lewis, J. A. (2004). In Defence of the Dichotomy. *Pharmaceut. Statist.* 3, 77–79. doi:10.1002/pst.107

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., et al. (2017). Genome-wide Association Analysis Identifies 30 New Susceptibility Loci for Schizophrenia. *Nat. Genet.* 49, 1576–1583. doi:10.1038/ng.3973

Lieberman, J., Chakos, M., Wu, H., Alvir, J., Hoffman, E., Robinson, D., et al. (2001). Longitudinal Study of Brain Morphology in First Episode Schizophrenia. *Biol. Psychiatry* 49, 487–499. doi:10.1016/s0006-3223(01)01067-8

Liu, B., Zhang, X., Cui, Y., Qin, W., Tao, Y., Li, J., et al. (2017). Polygenic Risk for Schizophrenia Influences Cortical Gyrification in 2 Independent General Populations. *Schbul* 43, sbw051–680. doi:10.1093/schbul/sbw051

Matsuda, H., Mizumura, S., Nemoto, K., Yamashita, F., Imabayashi, E., Sato, N., et al. (2012). Automatic Voxel-Based Morphometry of Structural MRI by SPM8 Plus Diffeomorphic Anatomic Registration through Exponentiated Lie Algebra Improves the Diagnosis of Probable Alzheimer Disease. *AJNR Am. J. Neuroradiol.* 33, 1109–1114. doi:10.3174/ajnr.A2935

Muench, J., and Hamer, A. M. (2010). Adverse Effects of Antipsychotic Medications. *Am. Fam. Physician* 81, 617–622.

Palaniyappan, L., Marques, T. R., Taylor, H., Handley, R., Mondelli, V., Bonaccorso, S., et al. (2013). Cortical Folding Defects as Markers of Poor Treatment Response in First-Episode Psychosis. *JAMA Psychiatry* 70, 1031–1040. doi:10.1001/jamapsychiatry.2013.203

Patel, K. R., Cherian, J., Gohil, K., and Atkinson, D. (2014). Schizophrenia: Overview and Treatment Options. *P T* 39, 638–645.

Patterson, N., Price, A. L., and Reich, D. (2006). Population Structure and Eigenanalysis. *PLOS Genet.* 2, e190. doi:10.1371/journal.pgen.0020190

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but Systematic Correlations in Functional Connectivity MRI Networks Arise from Subject Motion. *NeuroImage* 59, 2142–2154. doi:10.1016/j.neuroimage.2011.10.018

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal Components Analysis Corrects for Stratification in Genome-wide Association Studies. *Nat. Genet.* 38, 904–909. doi:10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., 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

Quarantelli, M., Palladino, O., Prinster, A., Schiavone, V., Carotenuto, B., Brunetti, A., et al. (2014). Patients with Poor Response to Antipsychotics Have a More Severe Pattern of Frontal Atrophy: A Voxel-Based Morphometry Study of Treatment Resistance in Schizophrenia. *Biomed. Res. Int.* 2014, 1–9. doi:10.1155/2014/325052

Quinlan, E. B., Banaschewski, T., Banaschewski, T., Barker, G. J., Bokde, A. L. W., Bromberg, U., et al. (2020). Identifying Biological Markers for Improved Precision Medicine in Psychiatry. *Mol. Psychiatry* 25, 243–253. doi:10.1038/s41380-019-0555-5

Sambataro, F., Blasi, G., Fazio, L., Caforio, G., Taurisano, P., Romano, R., et al. (2010). Treatment with Olanzapine Is Associated with Modulation of the Default Mode Network in Patients with Schizophrenia. *Neuropsychopharmacol* 35, 904–912. doi:10.1038/npp.2009.192

Sarpal, D. K., Robinson, D. G., Lencz, T., Argyelan, M., Ikuta, T., Karlsgodt, K., et al. (2015). Antipsychotic Treatment and Functional Connectivity of the Striatum in First-Episode Schizophrenia. *JAMA Psychiatry* 72, 5–13. doi:10.1001/jamapsychiatry.2014.1734

Sarpal, D. K., Argyelan, M., Robinson, D. G., Szeszko, P. R., Karlsgodt, K. H., John, M., et al. (2016). Baseline Striatal Functional Connectivity as a Predictor of Response to Antipsychotic Drug Treatment. *Ajp* 173, 69–77. doi:10.1176/appi.ajp.2015.14121571

Sarpal, D. K., Robinson, D. G., Fales, C., Lencz, T., Argyelan, M., Karlsgodt, K. H., et al. (2017). Relationship between Duration of Untreated Psychosis and Intrinsic Corticostriatal Connectivity in Patients with Early Phase Schizophrenia. *Neuropsychopharmacol.* 42, 2214–2221. doi:10.1038/npp.2017.55

Sharma, A., and Verbeke, W. J. M. I. (2020). Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (N = 11,081). *Front. Big Data* 3, 15. doi:10.3389/fdata.2020.00015

Smeland, O. B., Frei, O., Dale, A. M., and Andreassen, O. A. (2020). The Polygenic Architecture of Schizophrenia - Rethinking Pathogenesis and Nosology. *Nat. Rev. Neurol.* 16, 366–379. doi:10.1038/s41582-020-0364-0

Staal, W. G., Hulshoff Pol, H. E., Schnack, H. G., van Haren, N. E. M., Seifert, N., and Kahn, R. S. (2001). Structural Brain Abnormalities in Chronic Schizophrenia at the Extremes of the Outcome Spectrum. *Ajp* 158, 1140–1142. doi:10.1176/appi.ajp.158.7.1140

Stroup, T. S., and Gray, N. (2018). Management of Common Adverse Effects of Antipsychotic Medications. *World Psychiatry* 17, 341–356. doi:10.1002/wps.20567

Szeszko, P. R., Narr, K. L., Phillips, O. R., McCormack, J., Sevy, S., Gunduz-Bruce, H., et al. (2012). Magnetic Resonance Imaging Predictors of Treatment Response in First-Episode Schizophrenia. *Schizophr. Bull.* 38, 569–578. doi:10.1093/schbul/sbq126

Tarcijonas, G., and Sarpal, D. K. (2019). Neuroimaging Markers of Antipsychotic Treatment Response in Schizophrenia: An Overview of Magnetic Resonance Imaging Studies. *Neurobiol. Dis.* 131, 104209. doi:10.1016/j.nbd.2018.06.021

Thorisson, G. A., Smith, A. V., Krishnan, L., and Stein, L. D. (2005). The International HapMap Project Web Site: Figure 1. *Genome Res.* 15, 1592–1593. doi:10.1101/gr.4413105

Torlay, L., Perrone-Bertolotti, M., Thomas, E., and Baciu, M. (2017). Machine Learning-XGBoost Analysis of Language Networks to Classify Patients with Epilepsy. *Brain Inf.* 4, 159–169. doi:10.1007/s40708-017-0065-7

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and Tuning Brain Decoders: Cross-Validation, Caveats, and Guidelines. *NeuroImage* 145, 166–179. doi:10.1016/j.neuroimage.2016.10.038

Varoquaux, G. (2018). Cross-validation Failure: Small Sample Sizes lead to Large Error Bars. *NeuroImage* 180, 68–77. doi:10.1016/j.neuroimage.2017.06.061

Voineskos, A. N., Jacobs, G. R., and Ameis, S. H. (2020). Neuroimaging Heterogeneity in Psychosis: Neurobiological Underpinnings and Opportunities for Prognostic and Therapeutic Innovation. *Biol. Psychiatry* 88, 95–102. doi:10.1016/j.biopsych.2019.09.004

Xu, K., Liu, Y., Zhan, Y., Ren, J., and Jiang, T. (2018). BRANT: A Versatile and Extendable Resting-State fMRI Toolkit. *Front. Neuroinform.* 12, 52. doi:10.3389/fninf.2018.00052

Zang, Y., Jiang, T., Lu, Y., He, Y., and Tian, L. (2004). Regional Homogeneity Approach to fMRI Data Analysis. *NeuroImage* 22, 394–400. doi:10.1016/j.neuroimage.2003.12.030

Zang, Y. F., He, Y., Zhu, C. Z., Cao, Q. J., Sui, M. Q., Liang, M., et al. (2007). Altered Baseline Brain Activity in Children with ADHD Revealed by Resting-State Functional MRI. *Brain Dev.* 29, 83–91. doi:10.1016/j.braindev.2006.07.002

Zang, J., Huang, Y., Kong, L., Lei, B., Ke, P., Li, H., et al. (2021). Effects of Brain Atlases and Machine Learning Methods on the Discrimination of Schizophrenia Patients: A Multimodal MRI Study. *Front. Neurosci.* 15, 944. doi:10.3389/fnins.2021.697168

Zhang, J.-P., Robinson, D., Yu, J., Gallego, J., Fleischhacker, W. W., Kahn, R. S., et al. (2019). Schizophrenia Polygenic Risk Score as a Predictor of Antipsychotic Efficacy in First-Episode Psychosis. *Ajp* 176, 21–28. doi:10.1176/appi.ajp.2018.17121363

Zugman, A., Gadelha, A., Assunção, I., Sato, J., Ota, V. K., Rocha, D. L., et al. (2013). Reduced Dorso-Lateral Prefrontal Cortex in Treatment Resistant Schizophrenia. *Schizophr. Res.* 148, 81–86. doi:10.1016/j.schres.2013.05.002

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 Genetics

ORIGINAL RESEARCH
published: 25 February 2022
doi: 10.3389/fgene.2022.772090

Pretreatment Thoracic CT Radiomic Features to Predict Brain Metastases in Patients With *ALK*-Rearranged Non-Small Cell Lung Cancer

Hua Wang[1†], *Yong-Zi Chen*[2†], *Wan-Hu Li*[3], *Ying Han*[4], *Qi Li*[5] *and Zhaoxiang Ye*[1*]

[1]Department of Radiology, Key Laboratory of Cancer Prevention and Therapy, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin Clinical Research Center for Cancer, Tianjin, China, [2]Laboratory of Tumor Cell Biology, Key Laboratory of Cancer Prevention and Therapy, Tianjin Clinical Research Center for Cancer, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China, [3]Department of Radiology, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan, China, [4]Department of Biotherapy, Key Laboratory of Cancer Prevention and Therapy, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin Clinical Research Center for Cancer, Tianjin, China, [5]Department of Pathology, Key Laboratory of Cancer Prevention and Therapy, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin Clinical Research Center for Cancer, Tianjin, China

OPEN ACCESS

Edited by:
Jiangning Song,
Monash University, Australia

Reviewed by:
Jiazhou Chen,
South China University of Technology, China
Zhitong Bing,
Institute of Modern Physics (CAS), China

***Correspondence:**
Zhaoxiang Ye
yezhaoxiang@163.com

[†]*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: *07 September 2021*
Accepted: *17 January 2022*
Published: *25 February 2022*

Citation:
Wang H, Chen Y-Z, Li W-H, Han Y, Li Q and Ye Z (2022) Pretreatment Thoracic CT Radiomic Features to Predict Brain Metastases in Patients With ALK-Rearranged Non-Small Cell Lung Cancer. Front. Genet. 13:772090. doi: 10.3389/fgene.2022.772090

Objective: To identify CT imaging biomarkers based on radiomic features for predicting brain metastases (BM) in patients with *ALK*-rearranged non-small cell lung cancer (NSCLC).

Methods: NSCLC patients with pathologically confirmed *ALK* rearrangement from January 2014 to December 2020 in our hospital were enrolled retrospectively in this study. Finally, 77 patients were included according to the inclusion and exclusion criteria. Patients were divided into two groups: BM+ were those patients who were diagnosed with BM at baseline examination (n = 16) or within 1 year's follow-up (n = 14), and BM− were those without BM followed up for at least 1 year (n = 47). Radiomic features were extracted from the pretreatment thoracic CT images. Sequential univariate logistic regression, LASSO regression, and backward stepwise logistic regression were used to select radiomic features and develop a BM-predicting model.

Results: Five robust radiomic features were found to be independent predictors of BM. AUC for radiomics model was 0.828 (95% CI: 0.736–0.921), and when combined with clinical features, the AUC was increased (p = 0.017) to 0.909 (95% CI: 0.845–0.972). The individualized BM-predicting model incorporated with clinical features was visualized by the nomogram.

Conclusion: Radiomic features extracted from pretreatment thoracic CT images have the potential to predict BM within 1 year after detection of the primary tumor in patients with *ALK*-rearranged NSCLC. The radiomics model incorporated with clinical features shows improved risk stratification for such patients.

Abbreviations: ALK, anaplastic lymphoma kinase; AUC, area under the curve; BM, brain metastasis; GLCM, gray-level co-occurrence matrix; GLDM, gray-level dependence matrix; GLRLM, gray-level run length matrix; GLSZM, gray-level size zone matrix; LASSO, least absolute shrinkage and selection operator; NGTDM, neighborhood gray tone difference matrix; NSCLC, non-small cell lung cancer.

Keywords: radiomics, computed tomography, anaplastic lymphoma kinase, lung cancer, brain metastases

INTRODUCTION

Lung cancer is the leading cause of cancer-related mortality worldwide. Non-small cell lung cancer (NSCLC) accounts for 85% of all lung cancer incidence (Molina et al., 2008). Approximately 10%–20% of NSCLC patients have brain metastases (BMs) at initial presentation (Schuette, 2004; Khalifa et al., 2016). Another 25%–50% will develop BMs during the course of their disease (Langer and Mehta, 2005). It has been reported that 91% of BMs were diagnosed within 1 year of initial diagnosis of the primary tumor for patients with lung cancer (Schouten et al., 2002). For stage I–III NSCLC patients, the median time from treatment to onset of BMs as the first site of progression was 12 months (Bajard et al., 2004). NSCLC patients with BMs traditionally have a poor prognosis with a median survival of 7 months (Sperduto et al., 2010).

Anaplastic lymphoma kinase (*ALK*) rearrangements are driver mutations seen in about 3%–5% NSCLC (Gainor et al., 2013). The incidence of BMs is higher in patients with *ALK*-rearranged NSCLC: among those patients, up to 50%–60% will develop BMs during the course of their disease (Zhang et al., 2015). Crizotinib was the first *ALK* inhibitor developed and has demonstrated improved outcomes in patients with *ALK*-positive advanced NSCLC in comparison with chemotherapy (Solomon et al., 2014). However, the intracranial efficacy of crizotinib is poor, due to poor blood–brain barrier penetration (Costa et al., 2011). Second- and third-generation *ALK* inhibitors have shown better but variable intracranial control. Besides, prophylactic cranial irradiation has been discussed as a strategy to reduce the incidence of BM in NSCLC (Carolan et al., 2005; Pechoux et al., 2016). Therefore, developing biomarkers to predict patients at higher risk of BM might be significant in helping identify subgroups who need early detection of BM by close observation and benefit from intensification of systemic therapy, which is crucial for improving outcomes.

Tumor phenotypic differences can be quantified in CT images using radiomic features. Radiomics refers to high-throughput extraction of quantitative image features, which provide a comprehensive description of tumor phenotypes and heterogeneity (Kumar et al., 2012; Lambin et al., 2012). Biomarkers based on radiomic features have been reported to be associated with clinical outcomes and underlying genomic patterns (Chen et al., 2017). In recent years, studies have been performed on the predictive value of radiomic features for tumor progression and distant metastases in NSCLC (Fried et al., 2014; Coroller et al., 2015; Fan et al., 2019; Xu et al., 2019; Kakino et al., 2020; Sun et al., 2021). However, to date, research using a radiomics approach based on thoracic CT images to predict BM for *ALK*-rearranged NSCLC has been rarely reported (Xu et al., 2019). The purpose of this study was to identify CT imaging biomarkers using radiomic features extracted from pretreatment thoracic CT images for predicting BM in patients with *ALK*-rearranged NSCLC, focused on BM within 1 year after initial detection of the primary tumor.

MATERIALS AND METHODS

Study Population and Clinical Data

NSCLC patients with pathologically confirmed *ALK* rearrangement from January 2014 to December 2020 in our hospital were enrolled retrospectively in this study. Patients were consecutively included according to the following inclusion criteria: (1) pathologically confirmed NSCLC with *ALK* rearrangement; (2) available pretreatment thoracic CT images on picture archiving and communication system (PACS) performed less than 1 month before the pathologic sampling were collected; and (3) available brain MRI/PETCT/CT examination data at diagnosis of NSCLC and during follow-up to confirm the status of BMs. Patients who met any of the following criteria were excluded: (1) with other malignant neoplasms; (2) unsatisfactory CT image quality such as severe respiratory motion artifacts; and (3) loss to follow-up within 1 year and without BM at the last follow-up.

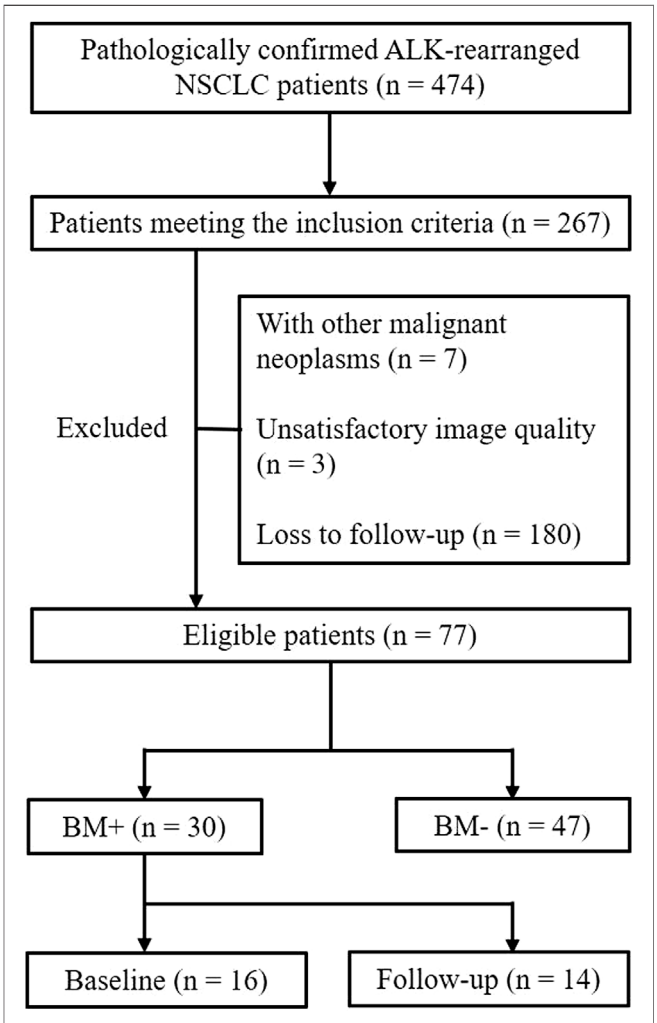

FIGURE 1 | Flowchart of the patient selection process.

TABLE 1 | Demographic and clinical features of the patients.

Clinical features	BM+	BM−	Total	*p*-value
Age, mean ± SD, years	52.23 ± 12.85	55.68 ± 9.19	54.34 ± 10.81	0.208
Age distribution				0.743
≤60	20 (66.7)	33 (70.2)	53 (68.8)	
>60	10 (33.3)	14 (29.8)	24 (31.2)	
Sex, *N* (%)				0.522
Female	15 (50.0)	27 (57.4)	42 (54.5)	
Male	15 (50.0)	20 (42.6)	35 (45.5)	
Smoking status, *N* (%)				0.217
Never	22 (73.3)	28 (59.6)	50 (64.9)	
Ever	8 (26.7)	19 (40.4)	27 (35.1)	
Pathology, *N* (%)				0.140
Adenocarcinoma	29 (96.7)	40 (85.1)	69 (89.6)	
Other	1 (3.3)	7 (14.9)	8 (10.4)	
T stage, *N* (%)				**0.001**
T1/T2	12 (40.0)	37 (78.7)	49 (63.6)	
T3/T4	18 (60.0)	10 (21.3)	28 (36.4)	
N stage, *N* (%)				**<0.001**
N0/N1	3 (10.0)	27 (57.4)	30 (39.0)	
N2/N3	27 (90.0)	20 (42.6)	47 (61.0)	

Abbreviations: BM, brain metastases.
Bolded values indicate a statistically significant result.

Finally, 77 patients were included in the study. Patients were divided into two groups: BM+ were those patients who diagnosed BM at baseline examination ($n = 16$) or within 1 year's follow-up ($n = 14$), and BM− were those without BM followed up for at least 1 year ($n = 47$) (**Figure 1**).

Clinicopathologic features were extracted from patient medical records, including age at diagnosis, sex, smoking status, pathological type, and TNM stage. Tumors were staged according to the new eighth edition of the Union for International Cancer Control and American Joint Committee on Cancer TNM classification system (Detterbeck et al., 2017).

CT Acquisition, Image Segmentation, and Feature Extraction

Pretreatment chest CT examinations were performed using one of the three multi-detector CT systems: Somatom Definition AS+ (Siemens Medical Solutions), Light speed 16 (GE Healthcare), or Discovery CT750 HD (GE Healthcare) scanner. Scanning parameters were as follows: tube voltage, 120 kVp; tube current, 150–200 mA with automatic exposure control; reconstruction thicknesses and intervals were 1.5 mm or 1.25 mm; reconstruction kernel was B30f/Standard for mediastinal window, and B70f/Lung for lung window.

The tumors were segmented using a semi-automatic approach by one radiologist and reviewed by another one, both of whom had experience in thoracic CT diagnosis for more than 10 years. They were both blinded to the clinical data and pathologic information except for lung cancer diagnosis. 3D Slicer V4.11.0[1] (Fedorov et al., 2012), an open-source image processing software, was used to segment the tumors on the images with reconstruction kernel of B70f/Lung and extract three-dimensional (3D) Radiomic features.

Features are grouped as follows: (1) First-order features: These describe the voxel intensity distribution in the delineated ROI. They are usually calculated based on the intensity histogram, including energy, entropy, skewness, kurtosis, uniformity, mean, minimum, and maximum intensity values. (2) Shape features: descriptors of the two- and three-dimensional shape and size of the ROI. (3) Textural features: These contain gray-level co-occurrence matrix (GLCM), gray-level dependence matrix (GLDM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLSZM), and neighborhood gray tone difference matrix (NGTDM). They are computed on the analysis of the three-dimensional directions within the tumor and the consideration of the spatial location of each voxel in the ROI (Shafiq-Ul-Hassan et al., 2017; Xu et al., 2020). (4) Wavelet-based features: These are extracted after applying a series of wavelet filtration to the images. The wavelet transform decomposes the original image into low-and high-frequencies, focusing the features on different frequency ranges within the tumor volume (Rios Velazquez et al., 2017). Finally, a total of 851 features were extracted, including 14 shape features, 18 first-order features, 75 texture features (24 GLCM, 14 GLDM, 16 GLRLM, 16 GLSZM, and 5 NGTDM), and 744 wavelet-based features (**Supplementary Table S1**).

Feature Selection, Radiomic Signature Building, and Development of Prediction Model

Univariate logistic regression analysis was preliminarily used to screen and identify potential predictors from radiomic features. Then, radiomic features with $p < 0.05$ in univariate analysis were further screened by the least absolute shrinkage and selection operator (LASSO) regression method. Tenfold cross-validation was used for selecting features in the LASSO model *via*

[1]http://www.slicer.org.

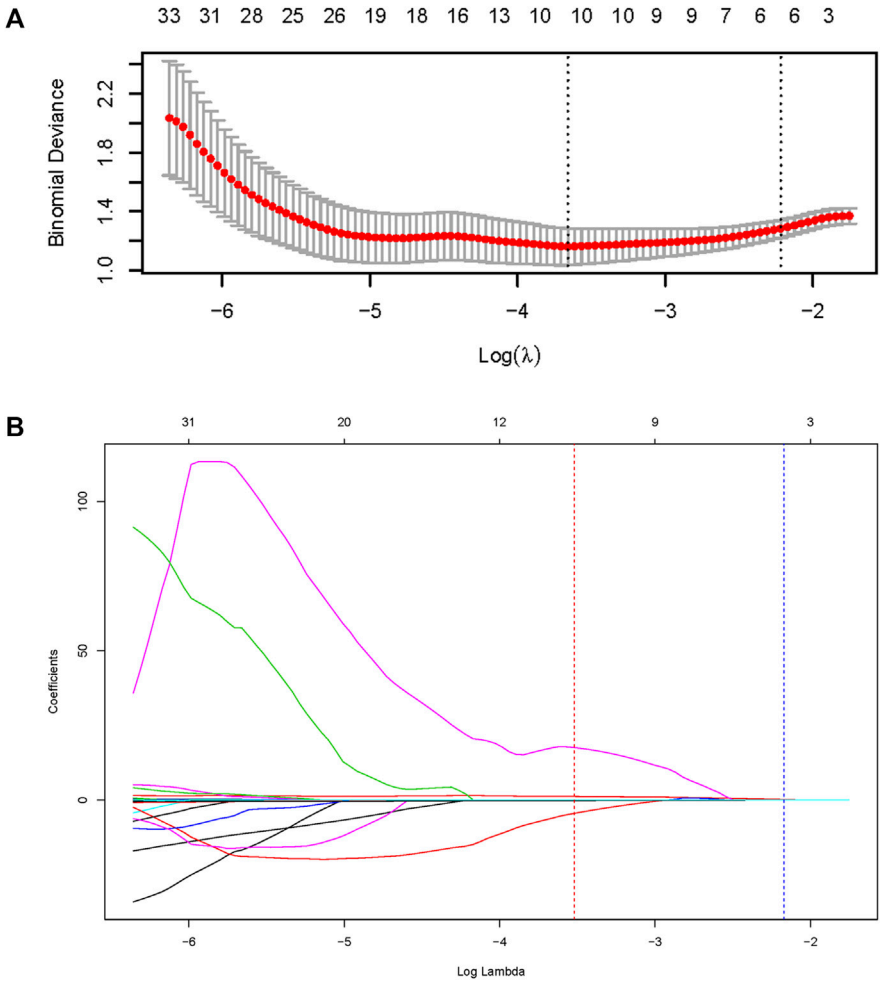

FIGURE 2 | Feature selection using the least absolute shrinkage and selection operator (LASSO) regression method. **(A)** The dotted vertical line was plotted at the value selected by the 10-fold cross-validation *via* minimum criteria (the value of lambda with the lowest partial likelihood deviance). **(B)** Selection of the tuning parameter (lambda) in the LASSO regression using 10-fold cross-validation *via* minimum criteria.

TABLE 2 | Multivariate logistic regression analyses of radiomic features.

Radiomic features	Beta value	Odds ratio (95% CI)	*p*-value	AUC
Original.GLCM.contrast	−0.027	0.973 (0.942–1.006)	0.109	0.600
Wavelet_LHH.GLCM.clusterShade	0.046	1.047 (1.012–1.083)	0.009	0.666
Wavelet_LLH.GLSZM.smallAreaEmphasis	−30.675	0 (0.000–0.045)	0.014	0.632
Wavelet_HLH.firstorder.maximum	0.004	1.004 (1.000–1.007)	0.071	0.657
Wavelet_LLL.firstorder.skewness	−0.355	0.701 (0.498–0.985)	0.041	0.656

Abbreviations: CI, confidence interval; AUC, area under the receiver operating characteristic curve.

minimum criteria. In addition, multivariate logistic regression using a backward elimination strategy was performed to eliminate the redundant features. Finally, the prediction model was established based on the simplified radiomic features with beta values included in the backward stepwise regression as the standardized regression coefficients. A radiomics score (Rad_score) was calculated for each patient *via* a linear combination of selected features weighted by their regression coefficients. To provide the clinician with a quantitative tool to predict the individual probability of BM within 1 year after detection of NSCLC, we also built a nomogram incorporated with clinical features.

Statistical Analyses

Statistical analyses were conducted by R software (V3.6.2)[2]. For the potential clinical prognostic factors, the Student's

[2]http://www.r-project.org.

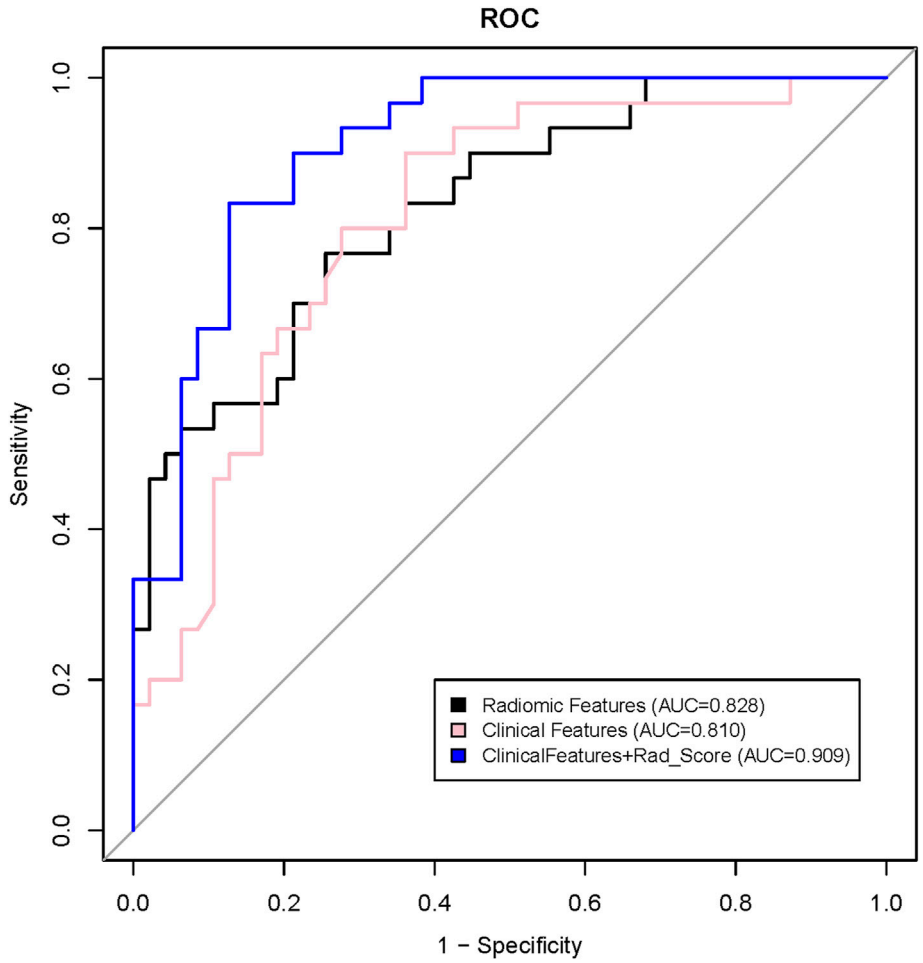

FIGURE 3 | Receiver operating characteristic (ROC) curves for prediction of brain metastases using a clinical model (pink line), a radiomic model (black line), and a model that combined Radiomics score (Rad_score) and clinical features (blue line).

t-test was used to compare the age of the two groups, and the other clinical features were compared using chi-square or Fisher's exact test, where appropriate. The diagnostic efficacy of the clinical, radiomic, and the combined model were analyzed by the receiver operating characteristic (ROC) curve of the subjects, and the differences between the area under the curve (AUC) were compared using DeLong's test. All tests were two-sided. A p-value < 0.05 was defined as significant for all the tests, except that in multivariate logistic regression with backward elimination strategy, a p-value < 0.1 was considered significant so that potential predictors were less likely to be eliminated from the prediction model.

RESULTS

Clinical Features

The patients' clinical data are presented in **Table 1**. There were significant differences in T stage (p = 0.001) and N stage (p < 0.001) between the two groups. Those patients with a higher T or N stage tend to have BM within 1 year after detection of NSCLC.

Radiomic Signature Building

Radiomic signature was built *via* three sequential steps. Firstly, a total of 112 radiomic features associated with BM (p < 0.05) were preliminarily identified by univariate logistic regression analysis (**Supplementary Table S2**). Then, ten radiomic features remained after conducting LASSO regression (**Figure 2**). Eventually, five robust radiomic features were found to be independent predictors of BM by using a backward stepwise logistic regression (**Table 2**). A detailed description of the features is presented in **Supplementary S1**. The prediction model based on the five radiomic features was built, and Rad_score was calculated for each patient. The Rad_score calculation formula was as follows:

Rad_score = Wavelet_LHH.GLCM.ClusterShade * 0.0459–Original. GLCM.Contrast * 0.0270–Wavelet_LLH.GLSZM. SmallAreaEmphasis * 3.6752 + Wavelet_HLH.Firstorder. Maximum * 0.0036–Wavelet_LLL.Firstorder.Skewness * 0.3551.

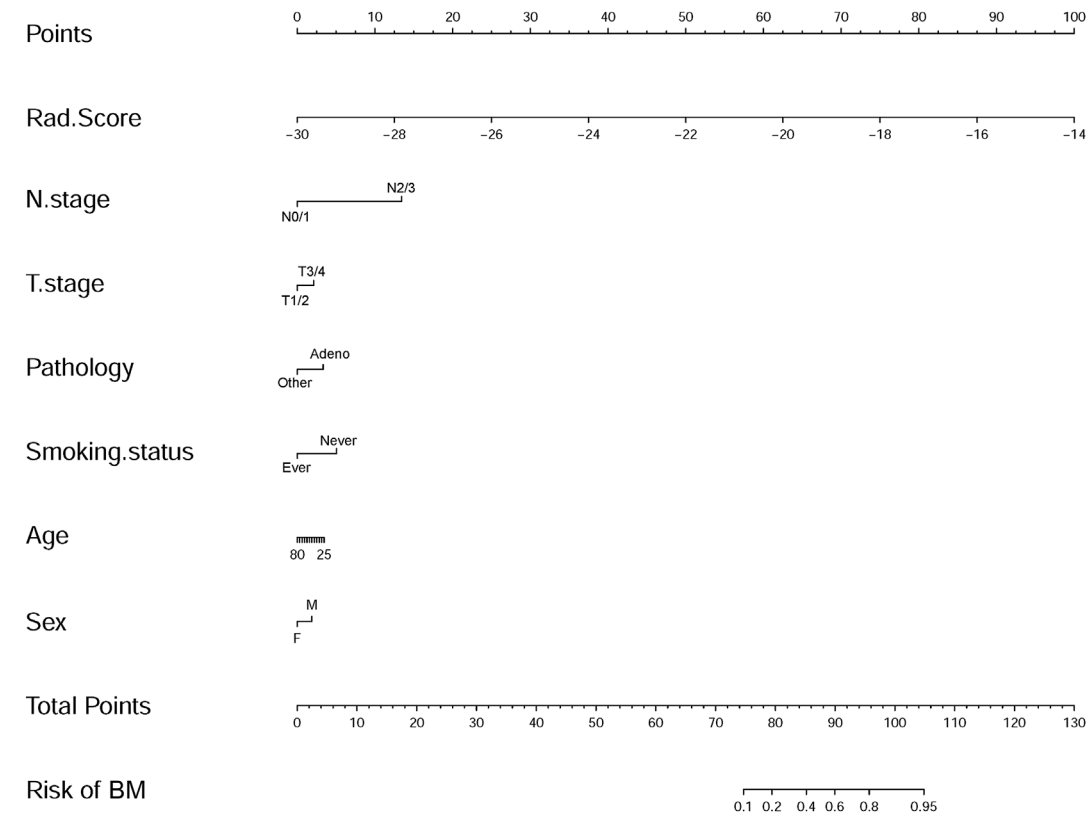

FIGURE 4 | Nomogram developed with the radiomics score (Rad_score) and clinical features incorporated.

Development of an Individualized Prediction Model

To illustrate the potential ability for BM prediction, we compared the models developed by radiomic features, clinical variables, and a combination of them. As shown in **Figure 3**, AUC for the radiomics model was 0.828 (95% CI: 0.736–0.921), which showed no significant difference ($p = 0.785$) with the clinical model (AUC = 0.810, 95% CI: 0.712–0.908), and when combined with clinical features, the AUC of the radiomics model was increased ($p = 0.017$) to 0.909 (95% CI: 0.845–0.972). The combined model was also superior to the clinical model alone ($p = 0.028$). The individualized BM-predicting model incorporated with clinical features is visualized by the nomogram (**Figure 4**). The process involved in the development of the prediction model is shown with a flowchart (**Figure 5**).

DISCUSSION

In this study, we developed a radiomics model with five independent predictors out of 851 candidate radiomic features extracted from pretreatment thoracic CT images for predicting BM within 1 year after detection of the primary tumor in patients with *ALK*-rearranged NSCLC, which showed a good performance with an AUC of 0.828. Furthermore, incorporating the radiomics signature with clinical features resulted in a significant improvement of predictive power with an excellent model performance (AUC = 0.909). We also built an easy-to-use nomogram that facilitates the individualized prediction of BM.

Age, T/N stage, pathological type, tumor genes, and other clinical features have been reported as risk factors or potential predictors of BMs. Patients with younger age (≤60 years), later T/N stage, adenocarcinoma, or non-squamous NSCLC are associated with a higher risk of BM (Robnett et al., 2001; Bajard et al., 2004; Carolan et al., 2005; Shi et al., 2006; Ji et al., 2014; Won et al., 2015). Epidermal growth factor receptor (*EGFR*) mutation was also reported to be a potential risk factor of BM (Shin et al., 2014; Shin et al., 2016). Compared with *EGFR* mutant patients, BMs were more common in patients with *ALK* rearrangement (Kang et al., 2014). Published data on risk factors of BM concerning the clinical features of *ALK*-rearranged NSCLC are minimal (Costa et al., 2015; Johung et al., 2016). Patients of this molecular subtype of NSCLC are relatively young (Yamamoto et al., 2014). While Costa et al. found younger age was associated with BM (Costa et al., 2015), no significant association between age and BM was found by Johung et al. (2016). In our study, most of the patients were younger than 60 years (68.8%); though patients in the BM+ group appeared younger than those in the BM− group (52.23 vs. 55.68 years), no significant difference was presented. Like previous studies (Bajard et al., 2004; Ji et al., 2014; Won et al., 2015), we also found that later T/N stage was associated with a higher risk of BM. Though

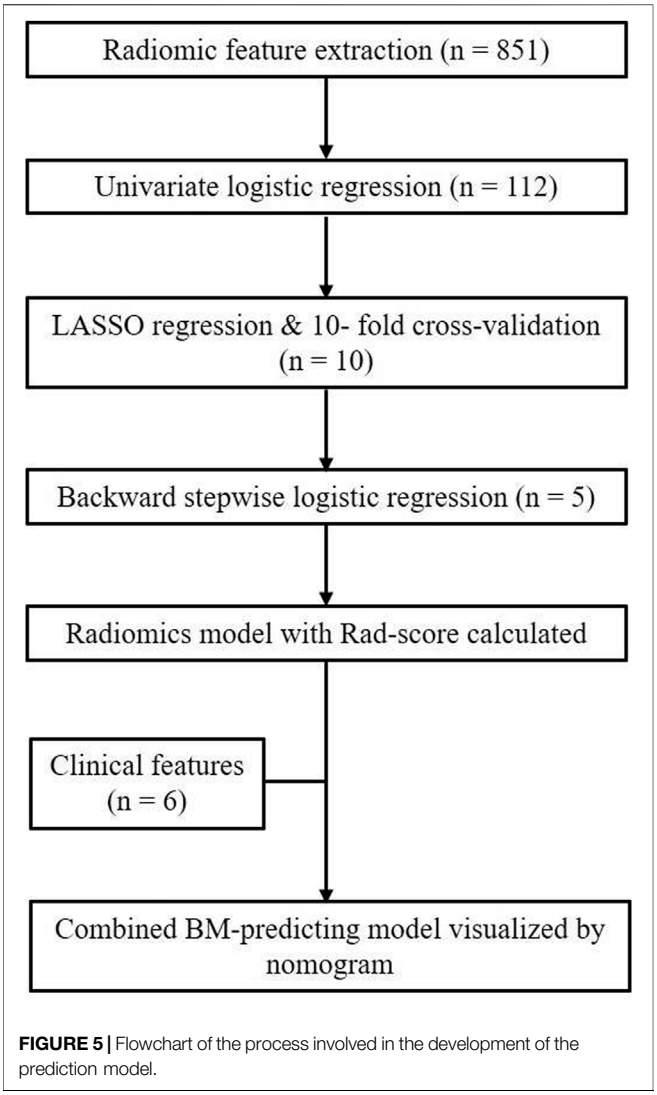

FIGURE 5 | Flowchart of the process involved in the development of the prediction model.

up to 96.7% of patients in BM+ group were adenocarcinoma, no significant association was found between pathological type and BM. It might due to the high prevalence of adenocarcinoma (89.6%) in this cohort, which is consistent with a previous report where adenocarcinoma accounts for most cases (85.3%) of *ALK*-rearranged NSCLC (Barlesi et al., 2016).

On account of the limited value of the clinical prognostic factors in predicting BM in this specific patient subset with a high incidence of BM, developing other biomarkers to build an optimal prediction model is necessary. Radiomics, as a non-invasive method developed in recent years, may potentially improve predictive accuracy in oncology. We found that five radiomic features, including one texture feature (Original.GLCM.Contrast), two wavelet-transformed texture features (Wavelet_LHH.GLCM.ClusterShade and Wavelet_LLH.GLSZM.SmallAreaEmphasis), and two wavelet-transformed first-order features (Wavelet_HLH.Firstorder.Maximum and Wavelet_LLL.Firstorder.Skewness), were independent predictors of BM in patients with *ALK*-rearranged NSCLC. The radiomics signature incorporated with clinical features yielded significantly improved predictive performance compared to both the radiomics model and the clinical model alone. Maximum and Skewness measure the maximal intensity of the histogram and the asymmetry of the histogram from the mean, respectively. Texture features are known to be most closely correlated with tumor heterogeneity and prognosis among all radiomic features, while wavelet-based features are the results of filter transformation of intensity and texture features (Chen et al., 2017). GLCM, ClusterShade, and Skewness (original or filtered) have been reported to be predictors of distant metastases in NSCLC (Coroller et al., 2015; Huynh et al., 2017; Kakino et al., 2020). Sun et al. (2021) also found that GLCM and GLSZM features were predictors of BM as the first failure in patients with curatively resected locally advanced NSCLC. Although differences exist in study objective and implementation, it implies that such features may serve as a risk factor of distant metastases, including BM for NSCLC. Further investigation is needed to explore the extensibility and universal applicability of these radiomic features for NSCLC with other driver gene mutations or distant metastases of other sites.

Recently, Xu et al. (2019) tried to build a radiomic signature to predict pretreatment BM for stage III/IV *ALK*-positive NSCLC patients and found that only one radiomic feature (W_GLCM_LH_Correlation) was an independent predictor (training set: AUC = 0.687, test set: AUC = 0.642), which also exhibited reposeful performance in predicting BM during follow-up (stage III: AUC = 0.682, stage IV: AUC = 0.653). However, due to the low positive rate (27 patients with pretreatment BM out of 132 patients) in their research, splitting data to the training set and test set and further dividing patients without BM at baseline examination into groups of different stages subsequently reduced sample size, which would mitigate statistical power compared to the initial cohort. To overcome this, we combined the patients with BM at baseline examination and within 1 year's follow-up into the BM+ group. We then used a cross-validation approach, which employs repeated data-splitting to prevent overfitting while simultaneously generating estimates of the model coefficients. This process is almost equivalent to data-splitting in producing validated model coefficients. Still, its use of data is more efficient than a dichotomous split into training and test sets (Harrell et al., 1996). However, there remains a high risk of a false-positive result due to the multiplicity of testing with the number of features tested (Fried et al., 2014). Additionally, recent studies have revealed that BM can occur even in patients with early-stage NSCLC or in those without any symptoms (Shi et al., 2006; Ando et al., 2018). Therefore, we did not intentionally exclude the patients with early stage. Actually, in the BM+ group, 40% were T0/1 stage, and 10% were N1/2 stage at the initial diagnosis.

There are several limitations to this study. First, due to the low incidence of *ALK* rearrangement and the high proportion of loss to follow-up, the sample size of our study was relatively small. Therefore, we only performed internal cross-validation, and the independent model assessment could not be committed to avoid overfitting. Expanded sample size and external multicenter validation are necessary for further investigation to confirm our findings. Second, the CT acquisition and reconstruction

parameters were not consistent for all the cases due to the different CT scanners we used. However, radiomics was able to detect a solid signal to predict BM despite the variability. In addition, because some patients did not undergo enhanced CT in the present study, we used plain CT images to extract the radiomic features to keep the sample size as large as possible, which may have an effect on the segmentation of the tumor.

In conclusion, our preliminary study indicates that radiomic features derived from pretreatment thoracic CT images may function as non-invasive biomarkers for predicting BM in patients with *ALK*-rearranged NSCLC. Furthermore, the radiomics model incorporated with clinical features shows improved risk stratification for such patients, allowing individualized treatment to reduce the risk of BM and improve survival.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the institutional review board of Tianjin Medical University Cancer Institute and Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HW and ZY contributed to conception and design of the study. HW, YH, and QL collected the data. HW, WL, and YC analyzed the data. YC performed the statistical analysis. HW wrote the first draft of the manuscript. HW and YC wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant numbers 81601492 and 81702268) and Shandong Cancer Hospital and Institute (clinical research cultivation project-19).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.772090/full#supplementary-material

REFERENCES

Ando, T., Kage, H., Saito, M., Amano, Y., Goto, Y., Nakajima, J., et al. (2018). Early Stage Non-small Cell Lung Cancer Patients Need Brain Imaging Regardless of Symptoms. *Int. J. Clin. Oncol.* 23, 641–646. doi:10.1007/s10147-018-1254-y

Bajard, A., Westeel, V., Dubiez, A., Jacoulet, P., Pernet, D., Dalphin, J. C., et al. (2004). Multivariate Analysis of Factors Predictive of Brain Metastases in Localised Non-small Cell Lung Carcinoma. *Lung Cancer* 45, 317–323. doi:10.1016/j.lungcan.2004.01.025

Barlesi, F., Mazieres, J., Merlio, J.-P., Debieuvre, D., Mosser, J., Lena, H., et al. (2016). Routine Molecular Profiling of Patients with Advanced Non-small-cell Lung Cancer: Results of a 1-year Nationwide Programme of the French Cooperative Thoracic Intergroup (IFCT). *The Lancet* 387, 1415–1426. doi:10.1016/S0140-6736(16)00004-0

Carolan, H., Sun, A. Y., Bezjak, A., Yi, Q.-L., Payne, D., Kane, G., et al. (2005). Does the Incidence and Outcome of Brain Metastases in Locally Advanced Non-small Cell Lung Cancer Justify Prophylactic Cranial Irradiation or Early Detection? *Lung Cancer* 49, 109–115. doi:10.1016/j.lungcan.2004.12.004

Chen, B., Zhang, R., Gan, Y., Yang, L., and Li, W. (2017). Development and Clinical Application of Radiomics in Lung Cancer. *Radiat. Oncol.* 12, 154. doi:10.1186/s13014-017-0885-x

Coroller, T. P., Grossmann, P., Hou, Y., Rios Velazquez, E., Leijenaar, R. T. H., Hermann, G., et al. (2015). CT-based Radiomic Signature Predicts Distant Metastasis in Lung Adenocarcinoma. *Radiother. Oncol.* 114, 345–350. doi:10.1016/j.radonc.2015.02.015

Costa, D. B., Kobayashi, S., Pandya, S. S., Yeo, W.-L., Shen, Z., Tan, W., et al. (2011). CSF Concentration of the Anaplastic Lymphoma Kinase Inhibitor Crizotinib. *Jco* 29, e443–e445. doi:10.1200/JCO.2010.34.1313

Costa, D. B., Shaw, A. T., Ou, S.-H. I., Solomon, B. J., Riely, G. J., Ahn, M.-J., et al. (2015). Clinical Experience with Crizotinib in Patients with Advanced ALK-Rearranged Non-small-cell Lung Cancer and Brain Metastases. *Jco* 33, 1881–1888. doi:10.1200/JCO.2014.59.0539

Detterbeck, F. C., Boffa, D. J., Kim, A. W., and Tanoue, L. T. (2017). The Eighth Edition Lung Cancer Stage Classification. *Chest* 151, 193–203. doi:10.1016/j.chest.2016.10.010

Fan, L., Fang, M., Tu, W., Zhang, D., Wang, Y., Zhou, X., et al. (2019). Radiomics Signature: A Biomarker for the Preoperative Distant Metastatic Prediction of Stage I Nonsmall Cell Lung Cancer. *Acad. Radiol.* 26, 1253–1261. doi:10.1016/j.acra.2018.11.004

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., et al. (2012). 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* 30, 1323–1341. doi:10.1016/j.mri.2012.05.001

Fried, D. V., Tucker, S. L., Zhou, S., Liao, Z., Mawlawi, O., Ibbott, G., et al. (2014). Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-small Cell Lung Cancer. *Int. J. Radiat. Oncology*Biology*Physics* 90, 834–842. doi:10.1016/j.ijrobp.2014.07.020

Gainor, J. F., Varghese, A. M., Ou, S.-H. I., Kabraji, S., Awad, M. M., Katayama, R., et al. (2013). ALK Rearrangements Are Mutually Exclusive with Mutations in EGFR or KRAS: an Analysis of 1,683 Patients with Non-small Cell Lung Cancer. *Clin. Cancer Res.* 19, 4273–4281. doi:10.1158/1078-0432.CCR-13-0318

Harrell, F. E., Jr., Lee, K. L., and Mark, D. B. (1996). Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statist. Med.* 15, 361–387. doi:10.1002/(sici)1097-0258(19960229)15:4<361:aid-sim168>3.0.co;2-4

Huynh, E., Coroller, T. P., Narayan, V., Agrawal, V., Romano, J., Franco, I., et al. (2017). Associations of Radiomic Data Extracted from Static and Respiratory-Gated CT Scans with Disease Recurrence in Lung Cancer Patients Treated with SBRT. *PLoS One* 12, e0169172. doi:10.1371/journal.pone.0169172

Ji, Z., Bi, N., Wang, J., Hui, Z., Xiao, Z., Feng, Q., et al. (2014). Risk Factors for Brain Metastases in Locally Advanced Non-small Cell Lung Cancer with Definitive

Chest Radiation. *Int. J. Radiat. Oncology*Biology*Physics* 89, 330–337. doi:10.1016/j.ijrobp.2014.02.025

Johung, K. L., Yeh, N., Desai, N. B., Williams, T. M., Lautenschlaeger, T., Arvold, N. D., et al. (2016). Extended Survival and Prognostic Factors for Patients with ALK-Rearranged Non-small-cell Lung Cancer and Brain Metastasis. *Jco* 34, 123–129. doi:10.1200/JCO.2015.62.0138

Kakino, R., Nakamura, M., Mitsuyoshi, T., Shintani, T., Kokubo, M., Negoro, Y., et al. (2020). Application and Limitation of Radiomics Approach to Prognostic Prediction for Lung Stereotactic Body Radiotherapy Using Breath-hold CT Images with Random Survival forest: A Multi-institutional Study. *Med. Phys.* 47, 4634–4643. doi:10.1002/mp.14380

Kang, H. J., Lim, H.-J., Park, J. S., Cho, Y.-J., Yoon, H.-I., Chung, J.-H., et al. (2014). Comparison of Clinical Characteristics between Patients with ALK-Positive and EGFR-Positive Lung Adenocarcinoma. *Respir. Med.* 108, 388–394. doi:10.1016/j.rmed.2013.11.020

Khalifa, J., Amini, A., Popat, S., Gaspar, L. E., and Faivre-Finn, C.International Association for the Study of Lung Cancer Advanced Radiation Technology, C, ommittee (2016). Brain Metastases from NSCLC: Radiation Therapy in the Era of Targeted Therapies. *J. Thorac. Oncol.* 11, 1627–1643. doi:10.1016/j.jtho.2016.06.002

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., et al. (2012). Radiomics: the Process and the Challenges. *Magn. Reson. Imaging* 30, 1234–1248. doi:10.1016/j.mri.2012.06.010

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R. G. P. M., Granton, P., et al. (2012). Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis. *Eur. J. Cancer* 48, 441–446. doi:10.1016/j.ejca.2011.11.036

Langer, C. J., and Mehta, M. P. (2005). Current Management of Brain Metastases, with a Focus on Systemic Options. *Jco* 23, 6207–6219. doi:10.1200/JCO.2005.03.145

Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E., and Adjei, A. A. (2008). Non-small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. *Mayo Clinic Proc.* 83, 584–594. doi:10.1016/s0025-6196(11)60735-0

Na, I., Shin, D.-Y., Lee, D., Kim, C., Koh, J., Lee, J., et al. (2016). Epidermal Growth Factor Receptor Mutations and Brain Metastasis in Patients with Nonadenocarcinoma of the Lung. *J. Can. Res. Ther.* 12, 318–322. doi:10.4103/0973-1482.154024

Péchoux, C. L., Sun, A., Slotman, B. J., De Ruysscher, D., Belderbos, J., and Gore, E. M. (2016). Prophylactic Cranial Irradiation for Patients with Lung Cancer. *Lancet Oncol.* 17, e277–e293. doi:10.1016/S1470-2045(16)30065-1

Rios Velazquez, E., Parmar, C., Liu, Y., Coroller, T. P., Cruz, G., Stringfield, O., et al. (2017). Somatic Mutations Drive Distinct Imaging Phenotypes in Lung Cancer. *Cancer Res.* 77, 3922–3930. doi:10.1158/0008-5472.CAN-17-0122

Robnett, T. J., Machtay, M., Stevenson, J. P., Algazy, K. M., and Hahn, S. M. (2001). Factors Affecting the Risk of Brain Metastases after Definitive Chemoradiation for Locally Advanced Non-small-cell Lung Carcinoma. *Jco* 19, 1344–1349. doi:10.1200/JCO.2001.19.5.1344

Schouten, L. J., Rutten, J., Huveneers, H. A. M., and Twijnstra, A. (2002). Incidence of Brain Metastases in a Cohort of Patients with Carcinoma of the Breast, colon, Kidney, and Lung and Melanoma. *Cancer* 94, 2698–2705. doi:10.1002/cncr.10541

Schuette, W. (2004). Treatment of Brain Metastases from Lung Cancer: Chemotherapy. *Lung Cancer* 45 (Suppl. 2), S253–S257. doi:10.1016/j.lungcan.2004.07.967

Shafiq-Ul-Hassan, M., Zhang, G. G., Latifi, K., Ullah, G., Hunt, D. C., Balagurunathan, Y., et al. (2017). Intrinsic Dependencies of CT Radiomic Features on Voxel Size and Number of gray Levels. *Med. Phys.* 44, 1050–1062. doi:10.1002/mp.12123

Shi, A. A., Digumarthy, S. R., Temel, J. S., Halpern, E. F., Kuester, L. B., and Aquino, S. L. (2006). Does Initial Staging or Tumor Histology Better Identify Asymptomatic Brain Metastases in Patients with Non-small Cell Lung Cancer? *J. Thorac. Oncol.* 1, 205–210. doi:10.1016/s1556-0864(15)31569-0

Shin, D.-Y., Na, I. I., Kim, C. H., Park, S., Baek, H., and Yang, S. H. (2014). EGFR Mutation and Brain Metastasis in Pulmonary Adenocarcinomas. *J. Thorac. Oncol.* 9, 195–199. doi:10.1097/JTO.0000000000000069

Solomon, B. J., Mok, T., Kim, D.-W., Wu, Y.-L., Nakagawa, K., Mekhail, T., et al. (2014). First-line Crizotinib versus Chemotherapy in ALK-Positive Lung Cancer. *N. Engl. J. Med.* 371, 2167–2177. doi:10.1056/NEJMoa1408440

Sperduto, P. W., Chao, S. T., Sneed, P. K., Luo, X., Suh, J., Roberge, D., et al. (2010). Diagnosis-specific Prognostic Factors, Indexes, and Treatment Outcomes for Patients with Newly Diagnosed Brain Metastases: a Multi-Institutional Analysis of 4,259 Patients. *Int. J. Radiat. Oncology*Biology*Physics* 77, 655–661. doi:10.1016/j.ijrobp.2009.08.025

Sun, F., Chen, Y., Chen, X., Sun, X., and Xing, L. (2021). CT-based Radiomics for Predicting Brain Metastases as the First Failure in Patients with Curatively Resected Locally Advanced Non-small Cell Lung Cancer. *Eur. J. Radiol.* 134, 109411. doi:10.1016/j.ejrad.2020.109411

Won, Y.-W., Joo, J., Yun, T., Lee, G.-K., Han, J.-Y., Kim, H. T., et al. (2015). A Nomogram to Predict Brain Metastasis as the First Relapse in Curatively Resected Non-small Cell Lung Cancer Patients. *Lung Cancer* 88, 201–207. doi:10.1016/j.lungcan.2015.02.006

Xu, F., Zhu, W., Shen, Y., Wang, J., Xu, R., Outesh, C., et al. (2020). Radiomic-Based Quantitative CT Analysis of Pure Ground-Glass Nodules to Predict the Invasiveness of Lung Adenocarcinoma. *Front. Oncol.* 10, 872. doi:10.3389/fonc.2020.00872

Xu, X., Huang, L., Chen, J., Wen, J., Liu, D., Cao, J., et al. (2019). Application of Radiomics Signature Captured from Pretreatment Thoracic CT to Predict Brain Metastases in Stage III/IV ALK-Positive Non-small Cell Lung Cancer Patients. *J. Thorac. Dis.* 11, 4516–4528. doi:10.21037/jtd.2019.11.01

Yamamoto, S., Korn, R. L., Oklu, R., Migdal, C., Gotway, M. B., Weiss, G. J., et al. (2014). ALKMolecular Phenotype in Non-small Cell Lung Cancer: CT Radiogenomic Characterization. *Radiology* 272, 568–576. doi:10.1148/radiol.14140789

Zhang, I., Zaorsky, N. G., Palmer, J. D., Mehra, R., and Lu, B. (2015). Targeting Brain Metastases in ALK-Rearranged Non-small-cell Lung Cancer. *Lancet Oncol.* 16, e510–e521. doi:10.1016/S1470-2045(15)00013-3

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.

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 | Frontiers in Genetics

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fgene.2021.799663

Pathway-Based Analysis Revealed the Role of Keap1-Nrf2 Pathway and PI3K-Akt Pathway in Chinese Esophageal Squamous Cell Carcinoma Patients With Definitive Chemoradiotherapy

Honghai Dai[1], Yanjun Wei[2], Yunxia Liu[1], Jingwen Liu[3], Ruoying Yu[3], Junli Zhang[3], Jiaohui Pang[3], Yang Shao[3,4], Qiang Li[1] and Zhe Yang[1]**

[1]Tumor Research and Therapy Center, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, China, [2]Tumor Research and Therapy Center, Shandong Provincial Hospital Affiliated to Shandong University, Jinan, China, [3]Nanjing Geneseeq Technology Inc, Nanjing, China, [4]School of Public Health, Nanjing Medical University, Nanjing, China

OPEN ACCESS

Edited by:
Ming Fan,
Hangzhou Dianzi University, China

Reviewed by:
Abhinita S. Mohanty,
Memorial Sloan Kettering Cancer Center, United States
Shuanghu Yuan,
Shandong University, China

***Correspondence:**
Qiang Li
lq1211@126.com
Zhe Yang
sdslyyyz@sina.com

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

Received: *21 October 2021*
Accepted: *13 December 2021*
Published: *25 April 2022*

Citation:
Dai H, Wei Y, Liu Y, Liu J, Yu R, Zhang J, Pang J, Shao Y, Li Q and Yang Z (2022) Pathway-Based Analysis Revealed the Role of Keap1-Nrf2 Pathway and PI3K-Akt Pathway in Chinese Esophageal Squamous Cell Carcinoma Patients With Definitive Chemoradiotherapy. Front. Genet. 12:799663. doi: 10.3389/fgene.2021.799663

Esophageal squamous cell carcinoma (ESCC) is the major type of EC in China. Chemoradiotherapy is a standard definitive treatment for early-stage EC and significantly improves local control and overall survival for late-stage patients. However, chemoradiotherapy resistance, which limits therapeutic efficacy and treatment-induced toxicity, is still a leading problem for treatment break. To optimize the selection of ESCC patients for chemoradiotherapy, we retrospectively analyzed the clinical features and genome landscape of a Chinese ESCC cohort of 58 patients. *TP53* was the most frequent mutation gene, followed by *NOTCH1*. Frequently, copy number variants were found in *MCL1* (24/58, 41.4%), *FGF19* (23/58, 39.7%), *CCND1* (22/58, 37.9%), and *MYC* (20/58, 34.5%). *YAP1* and *SOX2* amplifications were mutually exclusive in this cohort. Using univariate and multivariate analyses, the *YAP1* variant and *BRIP1* mutant were identified as adverse factors for OS. Patients with *PI3K-Akt* pathway alterations displayed longer PFS and OS than patients with an intact *PI3K-Akt* pathway. On the contrary, two patients with *Keap1-Nrf2* pathway alterations displayed significantly shortened PFS and OS, which may be associated with dCRT resistance. Our data highlighted the prognostic value of aberrant cancer pathways in ESCC patients, which may provide guidance for better chemoradiotherapy management.

Keywords: ESCC (Esophageal squamous cell carcinoma), *Keap1-Nrf2* pathway, *PI3K-Akt* pathway, chemoradiotherapy, pathway-based analysis

INTRODUCTION

Esophageal carcinoma (EC) is the ninth most common cancer and remains the sixth leading cause of cancer death worldwide (Bray et al., 2018). Esophageal squamous cell carcinoma (ESCC) and Esophageal adenocarcinoma (EAC) are two major subtypes of EC and account for 90% of EC cases worldwide. On the other hand, different histological types of EC distributed varied around the world. ESCC contributes to 90% of all esophagus carcinomas each year in China, whereas EAC is mainly reported in North America

and Europe (Abnet et al., 2018). Frequent consumption of hot beverages, a common lifestyle in China, results in a higher potential of ESCC, whereas people with gastroesophageal reflux, following a Western pattern diet, and with smoking behavior often have a higher risk of ADC (Dent et al., 2005). The 5-year survival rate of EC patients with only esophagus cancer is 47%, while the rate decreases to 25% if the tumor has spread to the surrounding organs or lymph nodes (Viale, 2020). Due to the poor prognosis and survival in EC, there is a strong demand for studying prognosis-related factors and seeking better treatment for patients with EC (Tustumi et al., 2016). The pathological pattern of Chinese EC provided us a unique opportunity to study the molecular mechanism underlying ESCC pathogenesis and disease outcomes.

Definitive chemoradiation therapy has been employed as the standard first-line therapy for ESCC patients. However, intolerance to radiotherapy and/or resistance to chemoradiotherapy was frequently observed with a high possibility of recurrence. The target therapy drug trastuzumab is the only HER2 monoclonal antibody approved by the FDA as a first-line drug along with chemotherapy for ESCC patients. Ramucirumab, an angiogenesis inhibitor that targets the VEGF/VEGFR2 pathway, has also been approved for EAC therapy (Yang et al., 2020). In addition, immunotherapy has been extensively evaluated in esophageal cancer. Nivolumab and pembrolizumab are two immune checkpoint inhibitors that target the PD-1/PD-L1 pathway approved by the FDA. Nivolumab (mOS = 10.9) has been confirmed to reduce the risk of death by 23% compared to chemotherapy alone (mOS = 8.4) in the phase 3 ATTRACTION-3 trial (mOS = 10.9) (Takahashi et al., 2021). These novel treatments have brought tremendous benefits to patients with a much longer survival time and better prognosis. Hence, the field of research on finding more targets for immune pharmaceuticals and targeted therapy is well worth exploring, and thus increasing the beneficial population.

It is well known that some signaling pathways altered across various tumor types, while others were highly associated with certain types of cancer, such as the oxidative stress response pathway in squamous cell carcinoma (Choe et al., 2021). For ESCC patients, definitive chemoradiotherapy is a standard therapy for non-resectable tumors. Pathways related to oxidative/electrophilic stress, like the cell cycle and *Keap1-Nrf*2 pathways, are therefore highly important for these patients to regulate exogenous stress from reactive oxygen species (ROS)/electrophiles induced by chemotherapy and radiotherapy. Here, we analyzed the alterations of ten canonical cancer-related pathways in this Chinese ESCC cohort (Sanchez-Vega et al., 2018). The ten pathways are cell cycle, *PI3 k*inase/Akt, *Keap1-Nrf2*, *Notch*, *p53*, *Myc*, *Hippo*, b-catenin/*Wnt*, RTK-RAS, and TGFβ signaling. Some pathways significantly correlated with the prognosis, which might aid in stratifying patients for better treatment management.

MATERIALS AND METHODS

Patients and Sample Collection

A total of 65 patients with ESCC were enrolled from the Tumour Research and Therapy Center, Shandong Provincial Hospital Affiliated to Shandong First Medical University, from 2016 to 2020 for retrospective analysis. Six patients were excluded from this study owing to their low-quality tissue samples, and one patient was excluded because no detectable mutation was found in this patient's sample (**Supplementary Figure S1**). Eventually, 58 patients were included in the study. All patients were diagnosed with unresectable locally advanced ESCC or advanced ESCC (stages II-IV, American Joint Committee on Cancer, seventh edition) and underwent standard definitive chemoradiotherapy (dCRT). For each patient, a somatic formalin-fixed paraffin-embedded (FFPE) tissue biopsy was performed before definitive chemoradiotherapy. All tumor tissue samples with at least 10% tumor cell content were subjected to targeted panel sequencing using a 422-gene panel. This study was approved by the Ethical Review Board of the Shandong Provincial Hospital Affiliated to Shandong First Medical University.

DNA Extraction and Library Preparation

The process from DNA extraction to library construction to target enrichment was performed in a CLIA-certified and CAP-accredited laboratory as previously described (Fang et al., 2019; Dai et al., 2020). In brief, genomic DNA from FFPE tissue was extracted using a QIAamp DNA FFPE Tissue Kit (Qiagen). DNA quantitation was then performed by using a QubitTM dsDNA HS Assay Kit for each sample, with its quality been identified by a NanoDropTM 2000 Spectrophotometer. Then we constructed the library for Illumina sequencing from fragmented dsDNA, using a KAPA HyperPrep kit (KAPA BIOSYSTEMS). The main steps of library preparation include end-repair and A-tailing, adapter ligation, and library amplification. The end-repair and A-tailing steps prepare end-repaired DNA, and 3' A-tailing prepares double-stranded DNA. Adapter ligation attaches synthesized oligonucleotides as adapters to one or both ends of targeted DNA fragments. The final step of library preparation performs a low-bias and high-fidelity polymerase chain reaction (PCR) to amplify the targeted sequences carrying proper adapters, accompanied with an AMPure XP agent (Beckman Coulter) for purification. The customed xGen lockdown probes panel, containing 422 refined cancer-related genes, was further used to enrich the targeted genes. Subsequently, the prepared library was quantified using a KAPA Library Quantification Kit (KAPA BIOSYSTEMS), and the size distribution of each sample was calculated by Bioanalyzer 2100 (Agilent Technologies).

DNA Sequencing With Quality Control

Targeted enriched libraries from the last step were sequenced using the Illumina HiSeq4000 Sequencing System to a mean coverage depth of at least 250×. The output BCL files (image data) from sequencing system were then demultiplexed and converted into readable FASTQ files by BCL2Fastq Conversion (version 1.8.4) from Illumina. Fastp (0.20.0; https://github.com/OpenGene/fastp/) was responsible for removing low-quality bases (base quality score Q30 < 30), trimming adapters, and read pruning. Qualified data were then mapped to the reference human genome (hg19 37d5) using a Burrows–Wheeler Aligner (BWA-mem, v0.7.12; https://github.com/lh3/bwa/) to produce

bam files. The bam files were further sorted and then filtered into the final mapped file through the process of reads deduplication, local realignment, and base quality recalibration using Sambamba (v1.3; https://lomereiter.github.io/sambamba/) software. By comparing the consistency of SNP-associated signatures between tissue cell-free DNA and negative control in the Genome Analysis Toolkit (GATK 4.0.0; "https://software.broadinstitute.org/gatk/") contamination module, the samples were matched to each patient, as well as the DNA contamination score was estimate.

Mutation Calling and Annotations

The fully qualified sequencing data were then processed to a series of software for single-nucleotide variations (SNVs), insertion/deletion mutations, fusion, and copy number variation (CNV) detection. VarScan2 (Koboldt et al., 2012) was performed for detecting somatic mutations. Calls with a threshold of ≥1% mutant allele frequency (MAF) and ≥3 reads from both directions were retained. From these variant calls, SNPs in normal samples were filtered based on a list of sources, including dbSNP (Sherry et al., 2001), ClinVAR (Landrum et al., 2016), 1,000 Genome database (Genomes Project et al., 2015), 65,000 exomes project (ExAC) (Karczewski et al., 2017), COSMIC (v70) (Forbes et al., 2015), SIFT (Ng and Henikoff, 2003), and the laboratory's SNP database of pre-existing population. ANNOVAR (Wang et al., 2010) was used to annotate all these SNVs. For somatic mutations, calls were removed if they were present in >1% populations in 1,000 Genome database or in ExAC. The resulting list was further filtered through an in-house mutation list of common sequencing errors. Additionally, a variant with >20% abundance in the normal sample, likely an artifact, was also removed from the mutation list. Structural variants were detected using FACTERA with default parameters (Newman et al., 2014). And the CNVs were detected by ADTEx (GPLv3; http://adtex.sourceforge.net/), both with default parameters. The threshold for CNV loss was 0.65 and 2.0 for the CNV gain.

Mutation Description and Statistical Analysis

Oncoplots, constructed by R (4.0.3), were used to view the overall mutation landscape of ESCC patients in this study. Progression-free survival (PFS) was defined from the date of pathological diagnosis of esophageal carcinoma (EC) to the time of disease progression, worsening, or the last follow-up before progression. Overall survival (OS) started from EC diagnosis to the date of death or the last follow-up. The Kaplan–Meier method was used to estimate these two outcome measures among different genetic groups, different physiological populations, and selected pathways, followed by a stratified log-rank test for evaluating any differences. Subsequently, univariate Cox hazard models were further performed to define any prognostic factors affecting PFS and OS in this cohort. Statistically significant factors (p-value ≤0.1) defined in the single factor analysis were reviewed in detail. The beta coefficient in the pathway-related univariate analysis was the degree of change in the outcome (PFS or OS) for every 1-unit change in the number of pathway gene expression.

RESULTS

Clinical Characteristics and Mutation Landscape of ESCC Patients

The basic characteristic of 58 enrolled ESCC patients is shown in **Supplementary Table S1**. More than half of patients in the cohort were older than 60 years (55.17%), with a median age of 63 (range: 41–83) years. Forty-six patients (79.31%) were male, and only 12 (20.69%) were female. Around sixty-eight percent (39/58) of the patients were smokers, and fifty percentage had a history of alcohol consumption (29/58). More than half of the patients were diagnosed with stage III (36/58, 62.07%) ESCC, and 16 patients (16/58, 27.59%) were in stage II, with additional six patients (6/58, 10.34%) in stage IV.

In these Asian ESCC patients, *TP53* (54/58, 93.1%) was the most frequent mutation gene, followed by *NOTCH1* (30/58, 51.7%) (**Figure 1**). Amplification of *MCL1* (24/58, 41.4%), *FGF19* (23/58, 39.7%), *CCND1* (22/58, 37.9%), and *MYC* (20/58, 34.5%) was the four dominant types of CNV identified in this ESCC cohort. As previously mentioned, *FGF19* and *CCND1* were often co-amplified since they were both at adjacent locations on chromosome 11q13. Interestingly, *YAP1* and *SOX2* amplifications were mutually exclusive to each other in these ESCC patients (**Figure 1**). A similar negative correlation of the protein expression level in *YAP1* and *SOX2* was also found *in vivo* and *in vitro* of pancreatic neoplastic cells (Seo et al., 2013; Murakami et al., 2019).

Gene Alterations Associated With Disease Outcomes in ESCC Patients

In this cohort, the *YAP1* variant and *BRIP1* mutant were identified as adverse factors for PFS and OS in univariate analysis. In multivariate analysis, the *YAP1* variant and *BRIP1* mutant were significantly associated with OS but not with PFS (**Supplementary Table S2**). The Kaplan–Meier plot revealed that median PFS (mPFS) and median OS (mOS) of patients with the *YAP1* variant was 8.61 and 12.55 months, respectively, which were significantly shorter than that of *YAP1* wild-type patients (**Figures 2A,B**). ESCC patients with the *BRIP1* mutant also displayed worse outcomes than ones with the *BRIP1* wild type, achieving an mPFS of 5.87 months and an mOS of 11.38 months (**Figures 2C,D**). *SOX2* amplification, which was mutually exclusive to *YAP1* in this cohort, did not reach statistical significance in univariate analysis (**Figures 2E,F**).

Prognosis Value of Cancer-Associated Pathways in ESCC

Pathway analysis was performed according to the genes in ten cancer-associated pathways in the literature (**Supplementary Table S3**) (Sanchez-Vega et al., 2018). The individual genes in

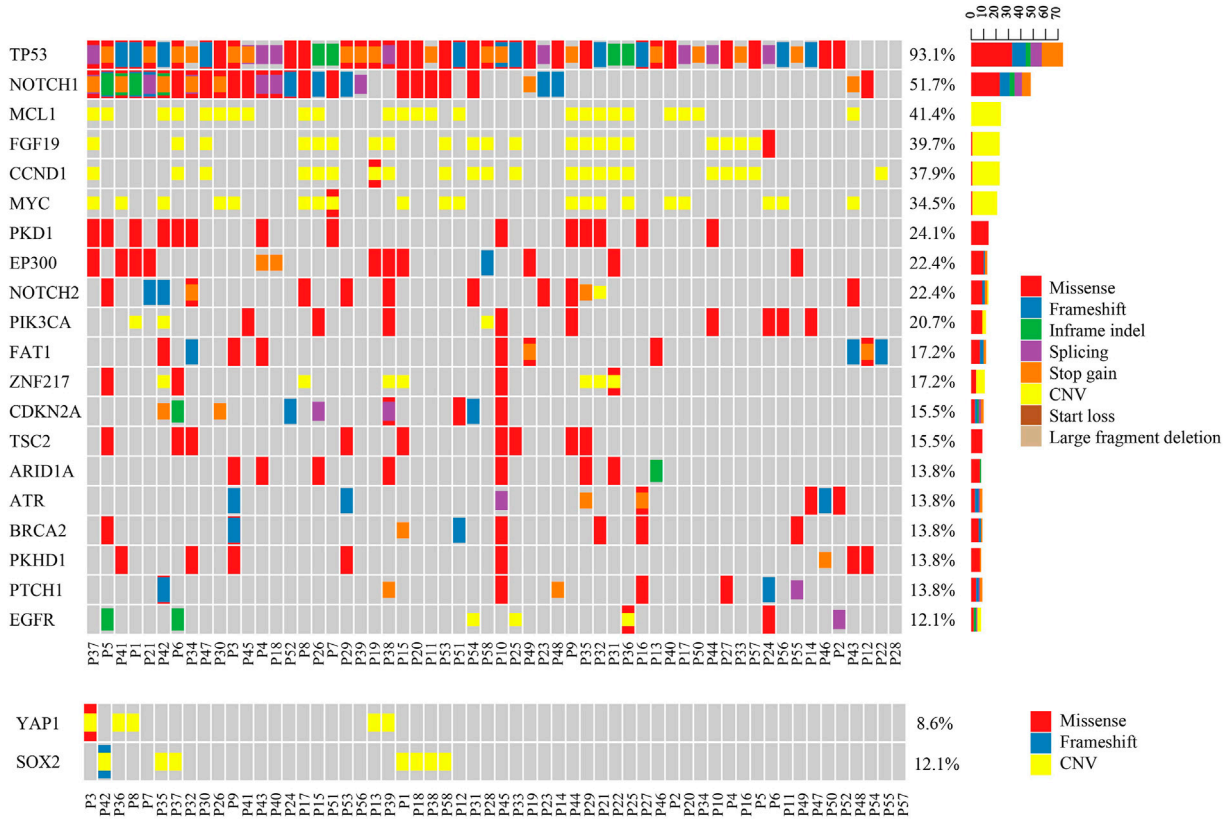

FIGURE 1 | Mutational pattern in Chinese ESCC patients. The upper oncoplot shows the mutational landscape of patients in this cohort. The lower oncoplot shows that *YAP1* gain and *SOX2* gain are mutually exclusive to each other. No patient had both amplifications at the same time.

each included pathway are listed in **Supplementary Table S1**. Around 93% of this EC cohort harbored TP53 signaling pathway alterations. Altered NOTCH (72.41%), RTK-RAS (68.97%), and cell cycle (53.45%) pathway genes were identified in more than 50% of the total cases (**Supplementary Figure S2** and **Supplementary Table S3**). EC patients with mutations in the *Keap1-Nrf2* pathway had much shorter (n = 2, mPFS = 2.75, beta = 3.48, $p < 0.0001$, HR (95% CI = 32.5 (4.48–235)) PFS than wild-type patients (n = 56, mPFS = 16.2) (**Figure 3A**). Similarly, mutations in this pathway also increased the risk of unfavorable OS (n = 2, beta coefficient = 29, $p < 0.0001$) compared to the wild-type counterpart (n = 56, mOS = 26.0) (**Figure 3B**). This observation was also validated using an independent cohort of 88 ESCC patients with OS information (Song et al., 2014). As shown in **Supplementary Figure S3**, seven patients had the altered *Keap1-Nrf2* pathway with a significantly shortened OS compared to patients with the intact *Keap1-Nrf2* pathway (p = 0.039).

In contrast to *Keap1-Nrf2* pathway aberrations, patients with mutations in the *PI3K-Akt* pathway displayed a longer PFS (n = 26, mPFS = 22, beta = 0.74, p = 0.0337, and HR (95% CI) = 0.48 (0.24–0.96) and longer OS (n = 26, mOS = 34.69, beta = 0.71, p = 0.0495, and HR (95% CI) = 32.5 (0.24–1.01). Comparatively, wild-type patients achieved a shorter PFS (n = 32, mPFS = 9.8) and OS (n = 32, mOS = 17.68) (**Figures 3C,D**). In patients with *PI3K-Akt* pathway alterations, three were found with altered *PTEN* and seven were found with altered *PIK3CA*. Patients with *PIK3CA* mutation tend to have longer PFS and OS than patients with wild-type *PIK3CA*. The altered *PTEN* did not show association with PFS or OS in this cohort (**Supplementary Figure S4**).

A representative case of an ESCC patient with *NFE2L2* mutation is shown in **Figure 3E**. The patient was a 49-year-old male diagnosed with stage IV ESCC. He was identified with *NFE2L2* D29G mutation at an allele frequency (AF) of 48.19% before treatment. The *RB1* frameshift mutation and *TP53* G262V were identified at an AF of 52.13 and 37.25%, respectively, at the same time. The tumor quickly progressed after 2.89 months of dCRT and metastasized to distant lymph nodes. Eventually, the patient died after 5.91 months of chemoradiotherapy and chemotherapy.

DISCUSSION

In this study, we retrospectively studied the clinical features and cancer genomes of 58 patients with inoperable ESCC tumors, intending to identify prognostic biomarkers for Chinese ESCC patients. Among all the baseline clinical characteristics, gender appeared to be an independent

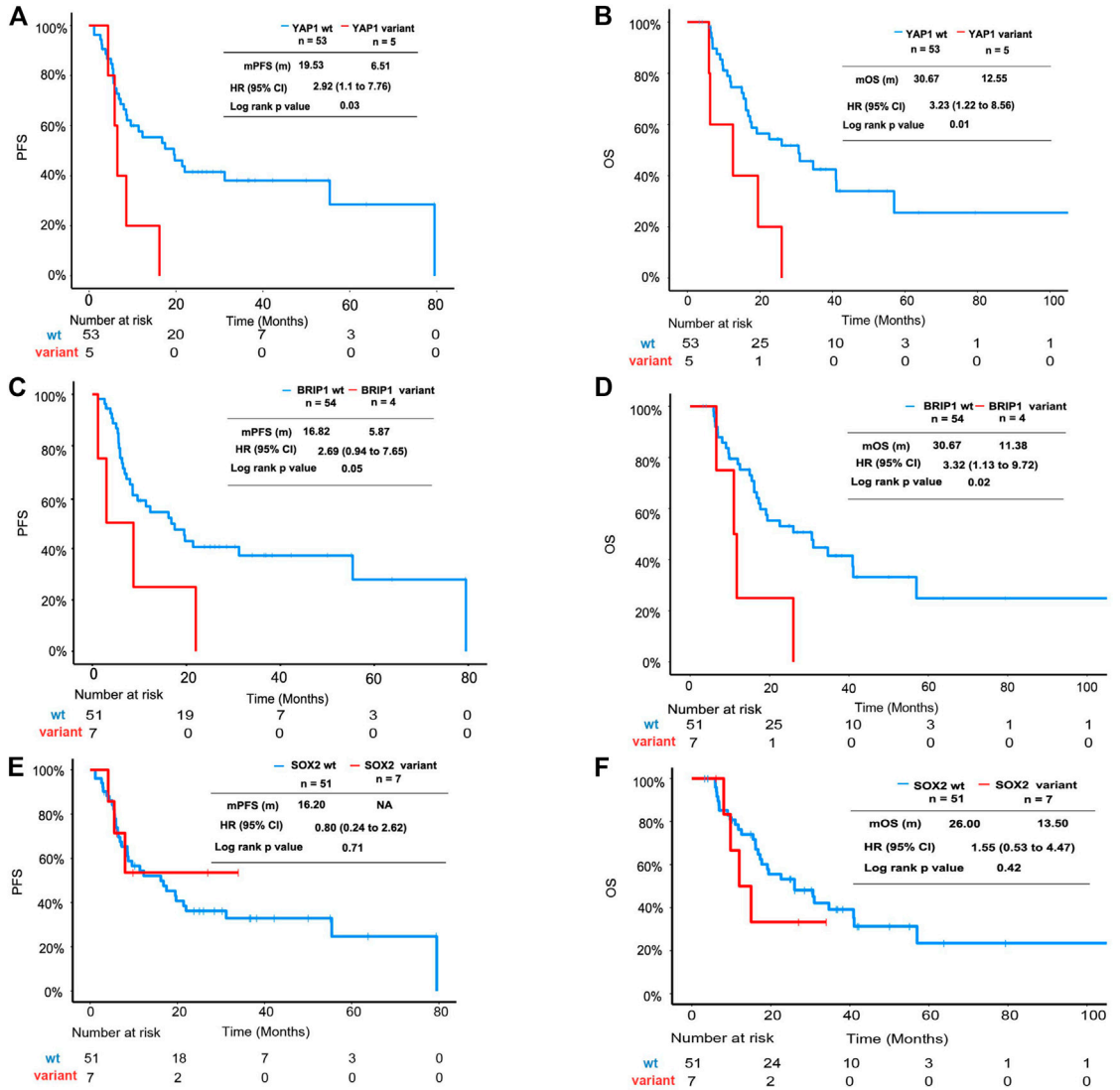

FIGURE 2 | Survival analysis of ESCC patients with *YAP1* mutation, *BRIP1* variation, and *SOX2* mutation. **(A)** Kaplan–Meier plot showing PFS of the subgroup patients with *YAP1* mutation versus patients without *YAP1* mutation. **(B)** Kaplan–Meier plot showing OS of the subgroup patients with *YAP1* mutation versus patients without *YAP1* mutation. **(C)** Kaplan–Meier plot showing PFS of the subgroup patients with *BRIP1* mutation versus patients without *BRIP1* variation. **(D)** Kaplan–Meier plot showing OS of the subgroup patients with *BRIP1* mutation versus patients without *BRIP1* variation. **(E)** Kaplan–Meier plot showing PFS of the subgroup patients with *SOX2* amplification versus patients without *SOX2* mutation. **(F)** Kaplan–Meier plot showing OS of the subgroup patients with *SOX2* amplification versus patients without *SOX2* mutation.

prognostic factor, which was in accord with the previous study (Pandeya et al., 2013). The high frequency of gene amplification was another genetic feature observed in esophageal squamous cell carcinoma. In our cohort, 75.9% (44/58) patients had at least one gene amplified. *MCL1* (24/58, 41.4%), *FGF19* (22/58, 37.9%), *CCND1* (22/58, 37.9%), and *MYC* (20/58, 34.5%) were the four dominant amplified genes. Besides, *YAP1* and *SOX2* were found to be exclusively amplified in different patients in this cohort. By further reviewing the prognosis of patients with/without *YAP1* and *SOX2* amplification, patients without double amplification were found to have the best PFS and OS. The group of patients with *SOX2* amplification and the group with *YAP1* amplification both obtained shorter PFS and OS, which consistent with the previous study (Dai et al., 2020).

Interestingly, the exclusion of *YAP1* amplification and *SOX2* amplification was only reported in one mouse model study that Yap loss intended to induce acute metabolic stress, leading to epigenetic reprogramming with *SOX2* upregulation (Murakami et al., 2019). Most other studies showed that *YAP1* is co-amplified with *SOX2* by *YAP1* binding to *SOX2*'s enhancer region, and *SOX2* may in turn restore *YAP1* by antagonizing the Hippo pathway in maintaining cell stemness and leading to poor prognosis. The cooperation of *YAP1* and *SOX2* was detected

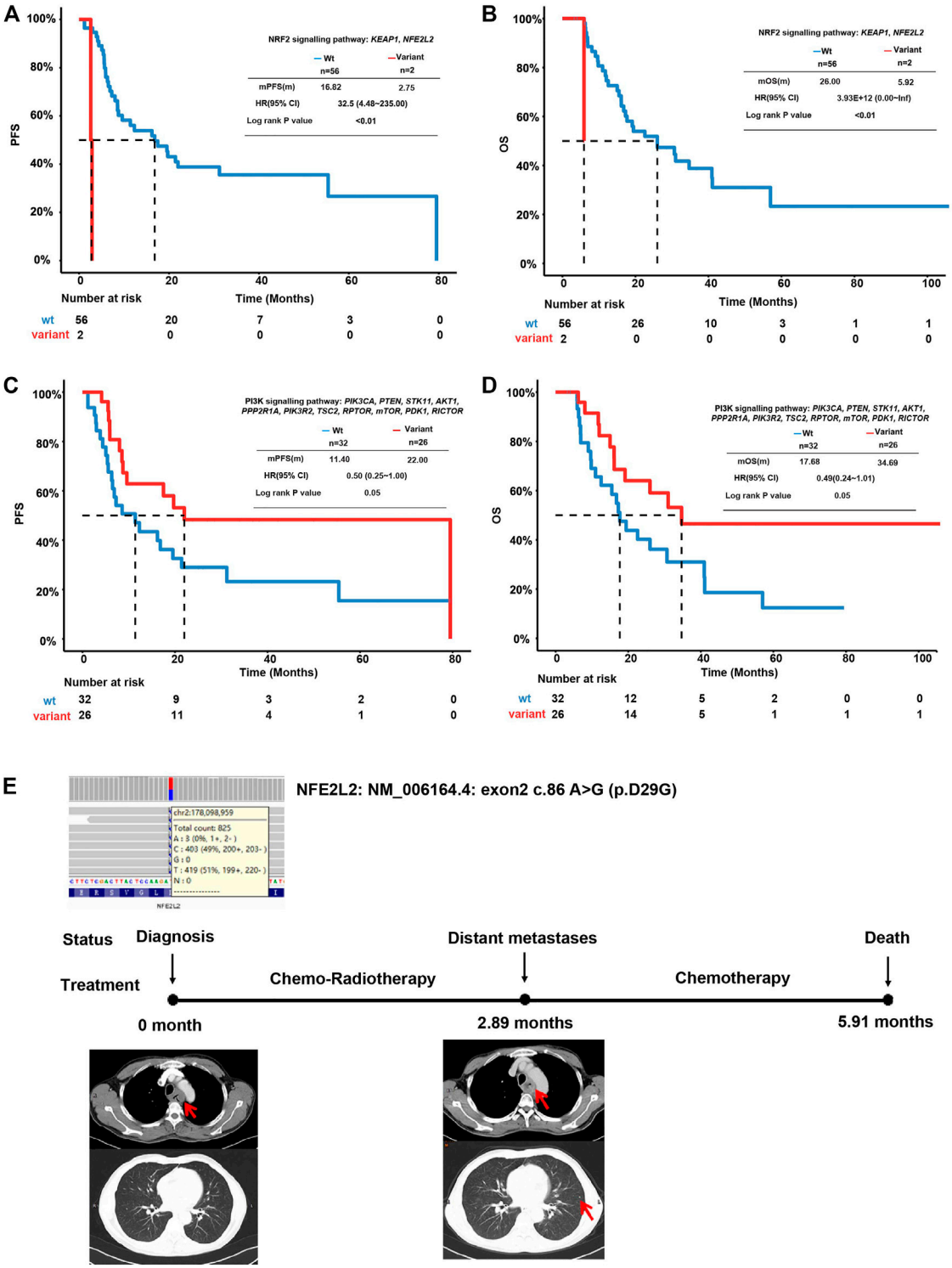

FIGURE 3 | Survival analysis of ESCC patients with altered oncogenic pathways. **(A)** Kaplan–Meier plot for PFS of ESCC patients with an intact or altered *Keap1-Nrf2* pathway. **(B)** Kaplan–Meier plot for OS of ESCC patients with an intact or altered Keap1-Nrf2 pathway. **(C)** Kaplan–Meier plot for PFS of ESCC patients with an intact or altered PI3K-Akt pathway. **(D)** Kaplan–Meier plot for OS of ESCC patients with an intact or altered PI3K-Akt pathway. **(E)** Representative case of a patient with *Keap1-Nrf2* pathway alterations.

in various cancer types, including osteosarcoma, urothelial cancer, and HNSCC (head and neck squamous cell carcinoma) (Murakami et al., 2019; Omori et al., 2019). Thus, behind the scenes of mutual exclusion for *SOX2* amplification and *YAP1* amplification of these patients in this study, there lies a unique unknown molecular mechanism of ESCC tumorigenesis, distinguished from other cancer types, which needs further investigation.

Of the two pathways identified as potential prognostic biomarkers of ESCC, the *Keap1-Nrf2* pathway is known for inducing chemoradioresistance (Taguchi and Yamamoto, 2017; Zhang et al., 2018). One of the major roles of *Nrf2* is to initiate cytoprotective responses under oxidant stress by binding to and activating the antioxidant response element (ARE) in the modular regions of its downstream targets (Kansanen et al., 2013). In addition, *Nrf2* promotes cell proliferation and metabolic reformation by triggering metabolic genes. On the other hand, *Keap1* can inhibit the *Keap1-Nrf2* pathway by suppressing the expression of *Nrf2*. Under oxidative stress and electrophilic stress, the confirmation of *Keap1* is reconstructed due to alterations in its cysteine residues. Newly synthesized *Nrf2* can bypass *Keap1* and translocate into the nucleus by *Keap1* protein inactivation or *Keap1-Nrf2* complex disruption (Kansanen et al., 2013). Here, the two patients carrying mutations in the *Keap1-Nrf2* pathway exhibit poor disease outcomes with shorter PFS and OS compared to *Keap1-Nrf2* pathway wild-type patients. The rapid progression of patients carrying abnormalities in the *Keap1-Nrf2* pathway in other cancer types was reported in several studies (Zoja et al., 2014; Goeman et al., 2019). Due to the limited number of patients with altered *Keap1-Nrf2* pathways in this study, further research is needed to identify whether activating mutations of the *Keap1-Nrf2* pathway is a potential chemo-radioresistance–related biomarker for patients receiving dCRT therapy.

PIK3CA mutation was a commonly reported factor for treatment and prognosis in ESCC patients, but conflicting conclusions were drawn across studies (Wada et al., 2006; Shigaki et al., 2013; Wang et al., 2014). Our studies showed a favorable prognosis among the patients with muted *PI3K* pathways. The PI3K-AKT pathway is considered one of the master regulators for cancer and ideal targets for anticancer drugs (Yang et al., 2019). It is known to play an important role in the development and progression of many solid cancers (Song et al., 2011; Jiao and Nan, 2012; Vredeveld et al., 2012). Further *in vivo* study or expansion of cohort size was needed to confirm our results.

DATA AVAILABILITY STATEMENT

The data has been uploaded to the GSA database. The accession ID is HRA002194. https://ngdc.cncb.ac.cn/gsa-human/s/4I2LAq98.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethical Review Board of Shandong Provincial Hospital Affiliated to Shandong First Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

QL, ZY, HD designed the study. HD, YW, and YL acquired data. RY, JL, JP, and JZ analyzed the data. HD, YW, RY, JL, JP, and JZ wrote the manuscript. Xue Wu, YS, QL, ZY, and HD supervised the study.

FUNDING

This study was supported in part by the Natural Science Foundation of Shandong (Grant No. ZR2020MH229), as well as the special foundation for CSCO Cancer Research (Grant No. Y-QL2019-0149 and Y-2019AZMS-0522) and the project of Shandong University (Grant No. 199/2019 heng).

ACKNOWLEDGMENTS

We would like to thank the patients who participated in this study and their family, as well as the investigators and research staff involved.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.799663/full#supplementary-material

REFERENCES

Abnet, C. C., Arnold, M., and Wei, W.-Q. (2018). Epidemiology of Esophageal Squamous Cell Carcinoma. *Gastroenterology* 154, 360–373. doi:10.1053/j.gastro.2017.08.023

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: a Cancer J. clinicians* 68, 394–424. doi:10.3322/caac.21492

Choe, J. H., Mazambani, S., Kim, T. H., and Kim, J. W. (2021). Oxidative Stress and the Intersection of Oncogenic Signaling and Metabolism in Squamous Cell Carcinomas. *Cells* 10, 606. doi:10.3390/cells10030606

Dai, H., Shao, Y. W., Tong, X., Wu, X., Pang, J., Feng, A., et al. (2020). YAP1 Amplification as a Prognostic Factor of Definitive Chemoradiotherapy in

Nonsurgical Esophageal Squamous Cell Carcinoma. *Cancer Med.* 9, 1628–1637. doi:10.1002/cam4.2761

Dent, J., El-Serag, H. B., Wallander, M. A., and Johansson, S. (2005). Epidemiology of Gastro-Oesophageal Reflux Disease: a Systematic Review. *Gut* 54, 710–717. doi:10.1136/gut.2004.051821

Fang, W., Ma, Y., Yin, J. C., Hong, S., Zhou, H., Wang, A., et al. (2019). Comprehensive Genomic Profiling Identifies Novel Genetic Predictors of Response to Anti-PD-(L)1 Therapies in Non-small Cell Lung Cancer. *Clin. Cancer Res.* 25, 5015–5026. doi:10.1158/1078-0432.ccr-19-0585

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., et al. (2015). COSMIC: Exploring the World's Knowledge of Somatic Mutations in Human Cancer. *Nucleic Acids Res.* 43, D805–D811. doi:10.1093/nar/gku1075

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A Global Reference for Human Genetic Variation. *Nature* 526, 68–74. doi:10.1038/nature15393

Goeman, F., De Nicola, F., Scalera, S., Sperati, F., Gallo, E., Ciuffreda, L., et al. (2019). Mutations in the KEAP1-Nfe2l2 Pathway Define a Molecular Subset of Rapidly Progressing Lung Adenocarcinoma. *J. Thorac. Oncol.* 14, 1924–1934. doi:10.1016/j.jtho.2019.07.003

Jiao, M., and Nan, K. J. (2012). Activation of PI3 kinase/Akt/HIF-1α Pathway Contributes to Hypoxia-Induced Epithelial-Mesenchymal Transition and Chemoresistance in Hepatocellular Carcinoma. *Int. J. Oncol.* 40, 461–468. doi:10.3892/ijo.2011.1197

Kansanen, E., Kuosmanen, S. M., Leinonen, H., and Levonen, A.-L. (2013). The Keap1-Nrf2 Pathway: Mechanisms of Activation and Dysregulation in Cancer. *Redox Biol.* 1, 45–49. doi:10.1016/j.redox.2012.10.001

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., et al. (2017). The ExAC Browser: Displaying Reference Data Information from over 60 000 Exomes. *Nucleic Acids Res.* 45, D840–D845. doi:10.1093/nar/gkw971

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing. *Genome Res.* 22, 568–576. doi:10.1101/gr.129684.111

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., et al. (2016). ClinVar: Public Archive of Interpretations of Clinically Relevant Variants. *Nucleic Acids Res.* 44, D862–D868. doi:10.1093/nar/gkv1222

Murakami, S., Nemazanyy, I., White, S. M., Chen, H., Nguyen, C. D. K., Graham, G. T., et al. (2019). A Yap-Myc-Sox2-P53 Regulatory Network Dictates Metabolic Homeostasis and Differentiation in Kras-Driven Pancreatic Ductal Adenocarcinomas. *Dev. Cel* 51, 113–128. doi:10.1016/j.devcel.2019.07.022

Newman, A. M., Bratman, S. V., Stehr, H., Lee, L. J., Liu, C. L., Diehn, M., et al. (2014). FACTERA: a Practical Method for the Discovery of Genomic Rearrangements at Breakpoint Resolution. *Bioinformatics* 30, 3390–3393. doi:10.1093/bioinformatics/btu549

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting Amino Acid Changes that Affect Protein Function. *Nucleic Acids Res.* 31, 3812–3814. doi:10.1093/nar/gkg509

Omori, H., Sato, K., Nakano, T., Wakasaki, T., Toh, S., Taguchi, K., et al. (2019). Stress-triggered YAP1/SOX2 Activation Transcriptionally Reprograms Head and Neck Squamous Cell Carcinoma for the Acquisition of Stemness. *J. Cancer Res. Clin. Oncol.* 145, 2433–2444. doi:10.1007/s00432-019-02995-z

Pandeya, N., Olsen, C. M., and Whiteman, D. C. (2013). Sex Differences in the Proportion of Esophageal Squamous Cell Carcinoma Cases Attributable to Tobacco Smoking and Alcohol Consumption. *Cancer Epidemiol.* 37, 579–584. doi:10.1016/j.canep.2013.05.011

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–e10. doi:10.1016/j.cell.2018.03.035

Seo, E., Basu-Roy, U., Gunaratne, P. H., Coarfa, C., Lim, D.-S., Basilico, C., et al. (2013). SOX2 Regulates YAP1 to Maintain Stemness and Determine Cell Fate in the Osteo-Adipo Lineage. *Cel Rep.* 3, 2075–2087. doi:10.1016/j.celrep.2013.05.029

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI Database of Genetic Variation. *Nucleic Acids Res.* 29, 308–311. doi:10.1093/nar/29.1.308

Shigaki, H., Baba, Y., Watanabe, M., Murata, A., Ishimoto, T., Iwatsuki, M., et al. (2013). PIK3CA Mutation Is Associated with a Favorable Prognosis Among Patients with Curatively Resected Esophageal Squamous Cell Carcinoma. *Clin. Cancer Res.* 19, 2451–2459. doi:10.1158/1078-0432.ccr-12-3559

Song, L., Xiong, H., Li, J., Liao, W., Wang, L., Wu, J., et al. (2011). Sphingosine Kinase-1 Enhances Resistance to Apoptosis through Activation of PI3K/Akt/NF-Kb Pathway in Human Non-small Cell Lung Cancer. *Clin. Cancer Res.* 17, 1839–1849. doi:10.1158/1078-0432.ccr-10-0720

Song, Y., Li, L., Ou, Y., Gao, Z., Li, E., Li, X., et al. (2014). Identification of Genomic Alterations in Oesophageal Squamous Cell Cancer. *Nature* 509, 91–95. doi:10.1038/nature13176

Taguchi, K., and Yamamoto, M. (2017). The KEAP1-NRF2 System in Cancer. *Front. Oncol.* 7, 85. doi:10.3389/fonc.2017.00085

Takahashi, M., Kato, K., Okada, M., Chin, K., Kadowaki, S., Hamamoto, Y., et al. (2021). Nivolumab versus Chemotherapy in Japanese Patients with Advanced Esophageal Squamous Cell Carcinoma: a Subgroup Analysis of a Multicenter, Randomized, Open-Label, Phase 3 Trial (ATTRACTION-3). *Esophagus* 18, 90–99. doi:10.1007/s10388-020-00794-x

Tustumi, F., Kimura, C. M. S., Takeda, F. R., Uema, R. H., Salum, R. A. A., Ribeiro-Junior, U., et al. (2016). Prognostic Factors and Survival Analysis in Esophageal Carcinoma. *Abcd, Arq. Bras. Cir. Dig.* 29, 138–141. doi:10.1590/0102-6720201600030003

Viale, P. H. (2020). The American Cancer Society's Facts & Figures: 2020 Edition. *J. Adv. Pract. Oncol.* 11, 135–136. doi:10.6004/jadpro.2020.11.2.1

Vredeveld, L. C. W., Possik, P. A., Smit, M. A., Meissl, K., Michaloglou, C., Horlings, H. M., et al. (2012). Abrogation of BRAFV600E-Induced Senescence by PI3K Pathway Activation Contributes to Melanomagenesis. *Genes Dev.* 26, 1055–1069. doi:10.1101/gad.187252.112

Wada, S., Noguchi, T., Takeno, S., and Kawahara, K. (2006). PIK3CA and TFRC Located in 3q Are New Prognostic Factors in Esophageal Squamous Cell Carcinoma. *Ann. Surg. Oncol.* 13, 961–966. doi:10.1245/aso.2006.08.006

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Res.* 38, e164. doi:10.1093/nar/gkq603

Wang, L., Shan, L., Zhang, S., Ying, J., Xue, L., Yuan, Y., et al. (2014). PIK3CA Gene Mutations and Overexpression: Implications for Prognostic Biomarker and Therapeutic Target in Chinese Esophageal Squamous Cell Carcinoma. *PloS one* 9, e103021. doi:10.1371/journal.pone.0103021

Yang, J., Nie, J., Ma, X., Wei, Y., Peng, Y., and Wei, X. (2019). Targeting PI3K in Cancer: Mechanisms and Advances in Clinical Trials. *Mol. Cancer* 18, 26. doi:10.1186/s12943-019-0954-x

Yang, Y.-M., Hong, P., Xu, W. W., He, Q.-Y., and Li, B. (2020). Advances in Targeted Therapy for Esophageal Cancer. *Sig Transduct Target. Ther.* 5, 229. doi:10.1038/s41392-020-00323-3

Zhang, J., Jiao, Q., Kong, L., Yu, J., Fang, A., Li, M., et al. (2018). Nrf2 and Keap1 Abnormalities in Esophageal Squamous Cell Carcinoma and Association with the Effect of Chemoradiotherapy. *Thorac. Cancer* 9, 726–735. doi:10.1111/1759-7714.12640

Zoja, C., Benigni, A., and Remuzzi, G. (2014). The Nrf2 Pathway in the Progression of Renal Disease. *Nephrol. Dial. Transpl.* 29 (Suppl. 1), i19–i24. doi:10.1093/ndt/gft224

Conflict of Interest: RY, JL, JP, JZ, Xue Wu, and YS are shareholders or employees of Nanjing Geneseeq Technology Inc.

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

The reviewer SY declared a shared affiliation with the authors HD, YW, YL, QL, and ZY to the handling editor at time of review.

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

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 28 June 2022
doi: 10.3389/fgene.2022.812105

Time Course Analysis of Transcriptome in Human Myometrium Depending on Labor Duration and Correlating With Postpartum Blood Loss

Lina Chen†, Yihong Luo†, Yunshan Chen, Lele Wang, Xiaodi Wang, Guozheng Zhang, Kaiyuan Ji and Huishu Liu**

Guangzhou Key Laboratory of Maternal-Fetal Medicine, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China

OPEN ACCESS

Edited by:
Ming Fan,
Hangzhou Dianzi University, China

Reviewed by:
Stephen Beesley,
Florida State University, United States
Jun He,
Fujian Medical University, China

***Correspondence:**
Huishu Liu
huishuliu@hotmail.com
Kaiyuan Ji
369027938@qq.com

†These authors have contributed equally to this work

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

Received: *09 November 2021*
Accepted: *12 May 2022*
Published: *28 June 2022*

Citation:
Chen L, Luo Y, Chen Y, Wang L, Wang X, Zhang G, Ji K and Liu H (2022) Time Course Analysis of Transcriptome in Human Myometrium Depending on Labor Duration and Correlating With Postpartum Blood Loss. Front. Genet. 13:812105. doi: 10.3389/fgene.2022.812105

The maintenance of coordinated powerful episodic contractions of the uterus is the crucial factor for normal labor. The uterine contractility is gradually enhanced with the progression of labor, which is related to the gene expression of the myometrium. Competing endogenous RNA (ceRNA) can also regulate the gene expression. To better understand the role of ceRNA network in labor, transcriptome sequencing was performed on the myometrium of 17 parturients at different labor durations (0–24 h). From this, expression levels of mRNA, long non-coding RNA (lncRNA), circular RNA (circRNA), and microRNA (miRNA) were correlated with labor duration. Then, targeting relationships between mRNAs, lncRNAs, circRNAs, and miRNAs were predicted, and the ceRNA regulatory network was established. The mRNA expression patterns associated with cervical dilation and postpartum bleeding were further investigated. This analysis identified 932 RNAs positively correlated with labor duration (859 mRNAs, 28 lncRNAs, and 45 circRNAs) and 153 RNAs negatively correlated with labor duration (122 mRNAs, 28 lncRNAs, and 3 miRNAs). These mRNAs were involved in protein metabolism, transport, and cytoskeleton functions. According to the targeting relationship among these ceRNAs and mRNAs, a ceRNA network consisting of 3 miRNAs, 72 mRNAs, 2 circRNAs, and 1 lncRNA was established. In addition, two mRNA expression patterns were established using time-series analysis of mRNA expression in different phases of cervical dilation. A ceRNA network analysis for blood loss was performed; postpartum bleeding was closely related to inflammatory response, angiogenesis, and hemostasis. This study identified human myometrial transcriptome and established the ceRNA regulatory network depending on labor duration and highlighted the dynamic changes that occur at ceRNAs during parturition, which need to be considered more in the future to better understand how changes in gene expression are relevant to functional changes in human myometrium at labor.

Keywords: labor duration, cervical dilation, postpartum blood loss, myometrium, transcriptome, expression regulation

INTRODUCTION

During labor, the myometrium undergoes a series of sustained and powerful contractions to deliver the baby, a process that produces biochemical and structural changes in the myometrium. The first stage of labor contains the latent phase and the active phase. The latent phase is characterized by slow cervical dilation and varies in duration. As labor progresses, the cervix dilates more rapidly, commonly commences from 4 cm dilation, and the intensity of uterine contractions increases which leads into the active phase, with regular and strong uterine contractions (Friedman, 1955; Krapohl et al., 1970; Liao et al., 2005). Changes in uterine myometrial contractility are underpinned by complex and highly regulated processes, cell structure, and signaling of the myometrium (Li et al., 2021), such as an increase in contractile proteins and changes in glycolytic and oxidative enzymes (Breuiller-Fouche et al., 2006; Wray et al., 2019).

RNA sequencing (RNA-seq) is currently one of the most commonly used methods for high-throughput analysis of gene expression. In previous studies, the complete transcriptome profiles of human myometrium in both quiescent and active states have been sequenced (Ackerman et al., 2021; Chan et al., 2014). Differential analyses of the myometrial transcriptome profiles at different states of cervical dilatation and fetal membrane rupture (ROM) emphasized that a single state of the myometrial transcriptome was unable to represent the physiological dynamic process of labor, and that the different stages of labor are needed to be characterized (Lai et al., 2021). The gene expression in the myometrium may also be influenced by the duration of labor. However, the issue has received little attention and the evidence is inadequate.

In addition, mRNA expression alone is insufficient to elucidate the effects of labor duration on gene expression, as the translation of functional proteins is prone to post-transcriptional regulation. Non-coding RNAs (ncRNAs), including lncRNA, circRNA, and miRNA, play an important role in the post-transcriptional regulation of mRNA. miRNAs function by binding to target mRNA, thus degrading mRNA or inhibiting its translation (Ambros, 2004). lncRNA and circRNA can competitively bind to miRNA through their miRNA responsive elements, thereby effectively controlling the subsequent post-transcriptional regulation of miRNA, reducing the inhibition of miRNA on mRNA expression, and acting as competing endogenous RNA (ceRNA) (Kopp et al., 2018; Zhang et al., 2020). ncRNA has key roles in the governance of myometrial contractility. Previous studies have shown that miRNAs, such as miR-200 family and miR-199a/214 cluster, participated in the hormonal regulation of myometrial quiescence and contractility during pregnancy and labor through the regulation of inflammation- and contraction-associated gene expression (Renthal et al., 2013). lncRNA has been considered to be the most frequent, prevalent, and abundant novel class of human genes (Guttman et al., 2012). Illumina® microarray of myometrium identified 1,692 lncRNAs, of which 13 were differentially expressed (Romero et al., 2014). Despite the importance of ncRNA in the myometrium during labor, there is little information about ncRNA at different phases of labor and lack of regulatory analysis of ncRNA and mRNA.

Cesarean section increases the amount of bleeding after delivery, compared to vaginal delivery (Misme et al., 2016). The most common cause of bleeding is uterine atony (Breathnach et al., 2009), and the main mechanism for preventing excessive bleeding is uterine myometrial contraction and thrombosis. Myometrial muscle fibers stretch in different directions during uterine contractions, squeezing the large blood vessels and therefore controlling bleeding. Pregnancy is primarily a hypercoagulable state to prevent postpartum hemorrhage, and defects in coagulation pathways may also lead to excessive bleeding (Oyelese et al., 2010). The function of the uterine myometrium is closely related to bleeding after delivery. Studies on the correlation between the functions of the myometrium and bleeding at the transcriptome level would help to find the molecular mechanisms of the myometrium in regulating bleeding after delivery.

This study aims to provide a comprehensive workflow and analysis of the expression of ncRNA and mRNA in human term gestation pregnancy myometrium. A correlation analysis approach was used to screen RNAs associated with labor duration or postpartum bleeding, and an analysis of their functions and regulations was performed to reveal the biochemical and structural dynamics of the myometrium in labor, so as to identify effective targets for real-time monitoring of labor.

MATERIALS AND METHODS

Subjects and Tissue Collection

A total of 17 lower uterine segment samples were collected from singleton, nulliparous women undergoing cesarean deliveries at different labor durations, including 2 non-labor (labor duration of 0 h) and 15 spontaneous term in labor (labor duration of 5–24 h). The tissue samples in this study overlapped with those in our previous publication (Chen et al., 2021), a study analyzing mRNA differences between non-laboring (n = 10) and laboring (n = 10) myometrial samples, in which mRNA data from 2 randomly selected non-laboring and 10 laboring samples were used. Five additional myometrial samples at different labor durations were collected, forming a cohort with a labor duration of 0–24 h. The clinical details of patients gathered using clinical phenotype and statistical testing are presented in **Supplementary Table S1**. This research was approved by the Ethics Committee of Guangzhou Women and Children Medical Center (No. 201915401), and the informed consent form was signed by every participant.

The participants underwent a cesarean section for indications of breech, fetal distress, or cephalopelvic disproportion, with no pregnancy (pre-labor) complications, placenta previa, or uterine fibroids. Labor was defined as regular palpable contractions (assessed using cardiotocography) and cervical dilation (assessed by digital examination). Each patient's labor duration was documented from labor start to cesarean section. The starting point was determined using cardiotocography and by digital examination after self-reporting regular contractions. Postpartum blood loss included bleeding from fetal delivery to

2 h postoperation. The quantitative postpartum bleeding was calculated by measuring the blood in the aspirator during operation and the blood in the blood-soaked gauze or nursing pad postoperation (Wilcox et al., 1959). Vaginal speculum examination was used for determining fetal membrane rupture status (ROM).

Myometrial tissue samples were obtained from the lower uterine segment during cesarean section after delivery of the fetus and placenta. Tissue samples were immediately washed with phosphate-buffered saline (Sigma) to reduce the amount of blood, and the attached decidua and adipose tissue were removed using surgical scissors and then dissected into pieces of approximately 100 mg and immersed in RNAlater solution (Sigma) to be snap frozen in liquid nitrogen and stored at −80°C.

Total RNA Extraction

For each sample, total RNAs were extracted from a minimum of 60 mg myometrium tissues, and detailed steps were as stated in our previous study (Chen et al., 2021). All samples (including those from previous study) were sequenced for mRNA and non-coding RNA.

mRNA and Long Non-coding RNA Library Construction and Sequencing

Methods for library construction are as stated in our previous study (Chen et al., 2021). The qualified libraries were pair end sequenced on the BGISEQ-500 System (BGI-Shenzhen, China). HISAT2 (v2.0.4) and RSEM (v1.2.12) were used to map and count the reads of mRNA and lncRNA with the reference of human genome (H. sapiens, GRCh38) and transcriptome (Ensembl, release 84) (Kim et al., 2015; Li et al., 2011). All the datasets presented in this study were deposited in Genome Sequence Archive (GSA) repository with accession number PRJCA009585.

miRNA Library Construction and Sequencing

Total RNAs were separated using polyacrylamide gel electrophoresis (PAGE). The 15% TBE-urea gel was compounded and pre-runned for 15–30 min at 200 V, RNA ladder and total RNA sample were mixed with gel loading dye, respectively, and then heated at 65°C for 5 min. The entire RNA ladder and total RNA sample were loaded onto the gel, and the gel was run at 200 V for 1 h. Small RNA regions corresponding to the 18–30 nt bands in the marker lane (14–30 ssRNA Ladder Marker, TAKARA) were excised and recovered. Then the small RNAs were ligated to adenylated 3' adapters which were annealed to unique molecular identifiers (UMI), followed by the ligation of 5' adapters. The adapter-ligated small RNAs were transcribed into cDNA and subsequently enriched using PCR. The target fragments of 110–13 bp were selected using agarose gel electrophoresis and purified using a QIAquick Gel Extraction Kit (QIAGEN, Valencia, CA). The library was checked for the distribution of fragments size using the Agilent 2100 bioanalyzer and quantified using real-time quantitative PCR (qPCR) (TaqMan Probe). The final ligated PCR products were sequenced using the BGISEQ-500 platform (BGI-Shenzhen, China). The cleaned reads were mapped to the miRBase with Bowtie2 (Langmead et al., 2009), and cmsearch (Nawrocki et al., 2013) was performed for Rfam mapping. The miRDeep2 software was used to predict novel miRNA by exploring the secondary structure (Friedländer et al., 2008). All the datasets presented in this study were deposited in Genome Sequence Archive (GSA) repository with accession number PRJCA009585.

circRNA Library Construction and Sequencing

Total RNAs were treated with DNase I and a Ribo-off rRNA Depletion Kit (Vazyme, Inc.) to degrade DNA and ribosomal RNA, respectively. Linear RNA was removed using RNase R (Epicentre, lnc). Purification was performed using Agencourt RNAClean XP magnetic beads. A tailing mix and RNA index adapters were added to perform end repair. The PCR products were denatured and circularized using the splint oligo sequence. Single-strand circular DNA was formatted as the final library. The library was checked for the distribution of fragments size using the Agilent 2100 bioanalyzer and quantified using BMG microplate reader (OMEGA). Finally, the qualified libraries were pair end sequenced on the BGISEQ-500 (BGI-Shenzhen, China). The software CIRI and find_circ is used to predict circRNA (Gao et al., 2015; Memczak et al., 2013). All the datasets presented in this study were deposited in Genome Sequence Archive (GSA) repository with accession number PRJCA009585.

Identification of Labor Duration or Blood Loss–Correlated RNAs

In RNA-seq data analysis, fragments per kilo base per million mapped reads (FPKM) and reads per million mapped reads (RPM) are two kinds of normalized expression units to remove technical biases such as the depth of sequencing and gene length. FPKM considers the sequencing depth and gene length for normalization and is suitable for paired-end RNA-seq protocols where gene length fluctuates greatly, such as mRNA, lncRNA, and circRNA sequencing (Conesa et al., 2016; Trapnell et al., 2010). RPM considers the sequencing depth but not the transcript length normalization and is suitable for sequencing protocols where reads are generated irrespective of gene length, such as miRNA-seq, as miRNA lengths are typically between 20 and 24 bp (Campbell et al., 2015). In this study, the expression levels of mRNAs, lncRNAs, and circRNAs were presented as FPKM values; the expression levels of miRNAs were presented as RPM values. RNAs with extremely low abundance (average FPKM/RPM of 17 samples <1) were excluded.

The correlation between RNA expression levels and labor duration or blood loss was calculated using the "cor.test" function of Pearson correlation in R (v4.1.1), with p-value < 0.01 as the threshold for statistical significance. The "pheatmap" and "ggplot2" R packages were used to draw heat map (SCR_016418 and SCR_014601; https://scicrunch.org/

resources). A scatter plot of RNA expression values was drawn using Microsoft Excel.

Gene Function Annotation

Gene ontology (GO) and signaling pathway analysis were conducted on the significantly correlated mRNAs using DAVID v6.8 and KOBAS-i online tools, respectively (Bu et al., 2021; Huang et al., 2009), to annotate the biological processes (BP), cellular component (CC), molecular function (MF), and the signaling pathways. GO enrichment analysis and the network construction were performed using ClueGO plug-in of Cytoscape v3.7.2 software (Shannon et al., 2003). Fisher's exact test is adopted to measure the gene enrichment in annotation terms; a p-value < 0.05 was considered to be significantly GO enriched. A corrected p-value < 0.05, corrected by Benjamini–Hochberg method, was considered to be significantly enriched.

circRNA/lncRNA-miRNA-mRNA ceRNA Network Construction

The three correlated miRNAs were selected as the hub components, and the interaction relationships between miRNAs and mRNAs were predicted using microT-CDS (Paraskevopoulou et al., 2013), with a threshold of 0.8. miRNAs that interacted with circRNAs were predicted using the circBank Database (Liu et al., 2019). miRNAs that interacted with lncRNAs were predicted using LncBase v2 of Experimental module (Paraskevopoulou et al., 2016). The three miRNAs targeting mRNAs, circRNAs, and lncRNAs were then selected by overlapping with correlated RNAs. The circRNA/lncRNA–miRNA–mRNA ceRNA network was visualized using Cytoscape v3.7.2 software based on the targeting relationships.

mRNA Expression Profile Time Series Clustering

The degree of cervical dilation was divided into three phases: cervical dilation = 0 cm (non-labor), cervical dilation <4 cm (the latent phase of the first stage of labor), and cervical dilation ≥4 cm (the active phase of the first stage of labor) (Friedman, 1996). All identified mRNAs were clustered using Short Time-Series Expression Miner (STEM) v1.3.13 (Ernst et al., 2006). Expression profiles of mRNAs were clustered based on FPKM value changes over different cervical dilation phases; the maximum number of model profiles was set to 50, and the maximum unit change in model profiles between time points was set to two. A corrected p-value < 0.01 was considered to be significantly enriched. The mRNA relative expression values can be exported from the STEM. The gene expression at the first time point was set to zero, representing the baseline of gene expression.

RESULTS

Clinical Characteristics of the Participants

Our study recruited 17 primigravida women with singleton pregnancies. The median and range of labor duration, cervical dilation, and postpartum blood loss were 12 (0–24) h, 3 (0–10) cm, and 330 (250–620) ml, respectively. The indications for caesarean section included the following: fetal distress (n = 9), breech (n = 2), and cephalopelvic disproportion (failure to progress) (n = 6). The clinical details of patients were gathered by the clinician.

Identification of Labor Duration–Correlated mRNAs

Myometria were collected from 17 parturients undergoing different phases of labor. To identify genes whose expression levels were gradually up- or down-regulated following the duration of labor, the correlation between mRNA expression and labor duration was calculated. The results showed that 859 mRNAs were positively correlated with labor duration and 122 mRNAs were negatively correlated (**Figure 1B**, **Supplementary Table S2**). There were some known labor-associated players including (but not limited to) mRNA-encoding proteins involved in the breakdown of extracellular matrix (matrix metalloproteinase 25, MMP25) (Flores-Pliego et al., 2015), cell extracellular matrix interactions (collagens type IV alpha 6 COL4A6) (Shchuka et al., 2020), and calcium signaling regulation (calcium/calmodulin-dependent protein kinase I, CAMK1, and ID, CAMK1D) (Papandreou et al., 2004). The top two mRNAs with the highest positive correlation coefficient (R) with labor duration were glutaredoxin-3 (GLRX3) and CTD nuclear envelope phosphatase 1 regulatory subunit 1 (CNEP1R1). While the top two mRNAs negatively correlated with labor duration were membrane-associated guanylate kinase, WW And PDZ domain containing 2 (MAGI2) and myocyte enhancer factor 2D (MEF2D) (**Figure 1C**). These findings suggested that substantial transcriptional changes occurred in the myometrium during labor and gene expression changes depended on labor duration.

To better comprehend the functions of these labor duration–correlated mRNAs, we carried out GO and KEGG pathway enrichment analyses. The results of GO enrichment in positively correlated mRNAs revealed that the most enriched BP GO terms were associated with protein metabolic process, especially protein ubiquitination. In addition, vesicle-mediated transport and exosomal secretion were also significantly enriched. The CC GO enriched terms showed that these genes were mostly involved in the composition of nucleoplasm and cytoplasm. In MF GO terms, these genes were significantly enriched for the binding of protein, ribosome, and RNA (**Figure 2A**, **Supplementary Table S3**). Meanwhile, in GO analysis of mRNAs negatively correlated with labor duration, the mRNAs were mainly concentrated in the BP of ion transport and actin regulation, and in CC constituting cytoskeleton components such as T-tubule, costamere, cortical actin cytoskeleton, adherens junction, and sarcolemma, the MF terms were enriched in calmodulin and actin binding (**Figure 2B**, **Supplementary Table S4**). The two groups of mRNAs with expression levels positively or negatively correlated with labor duration were enriched in different GO terms, indicating that biological processes such as material transportation, metabolism, and cell

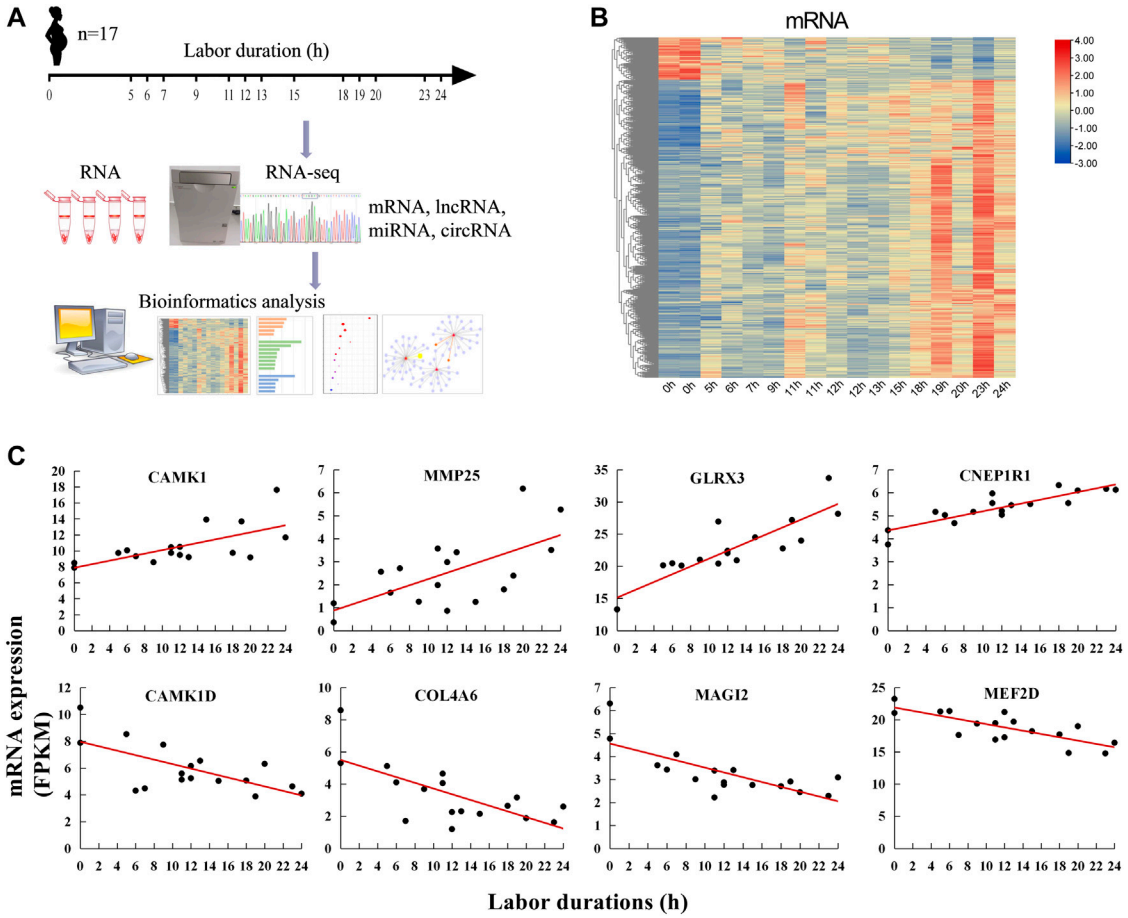

FIGURE 1 | Labor duration–correlated mRNAs expression. **(A)** Overall study design and workflow. **(B)** Heat map of expression of labor duration–correlated mRNAs. Data associated with this figure can be found in **Supplementary Table S2**. **(C)** The expression trend of mRNAs at different labor durations.

structure were dynamically changing during labor. The KEGG enrichment analysis of the labor duration–correlated mRNAs showed that multiple pathways were associated with metabolism, immune, and hypoxia response (**Figure 2C**, **Supplementary Table S5**), indicating that the gene expression was constantly being regulated during labor, as detected.

Labor Duration–Related ceRNA Regulatory Network Construction

ncRNAs including lncRNA, circRNA, and miRNA play important roles in post-transcriptional regulation, which ultimately affects mRNA translation (Panni et al., 2020). To explore the regulatory effect of ncRNAs on mRNA during labor, we identified lncRNA, circRNA, and miRNA expression levels that were significantly correlated with labor duration using high-throughput sequencing and correlation analysis. The lncRNA/circRNA-miRNA-mRNA regulatory network was further established.

A total of 56 lncRNAs correlated with labor duration (28 positively correlated and 28 negatively correlated), 45 circRNAs positively correlated with labor duration, and three miRNAs negatively correlated with labor duration (**Figure 3A**, **Supplementary Table S6**). lncRNA, circRNA, and mRNA all have response elements to bind miRNA directly, which enables them to communicate with and co-regulate each other by competing for binding to the shared miRNAs. Based on the ceRNA theory (Panni et al., 2020), the three miRNAs were defined as the core nodes of the regulatory network, and their target lncRNA, circRNA, and mRNA were predicted through the database. These predicted targets then overlapped with the identified labor duration–correlated lncRNAs, circRNAs, and mRNAs, which were regarded as elements of the ceRNA regulatory network. Finally, there were 75 interaction pairs predicted between three miRNAs and 72 mRNAs, two interaction pairs predicted between one miRNA (hsa-miR-146a-5p) and two circRNAs (hsa_circ_0000897 and hsa_circ_0085849), and one interaction pair predicted between miRNA and lncRNA (hsa-miR-206 and SNHG1). lncRNA/circRNA–miRNA–mRNA ceRNA regulatory networks were constructed based on their targeting relationships. There were three mRNAs (SLC8A1, ZNF207, and GUCY1A2) that can be

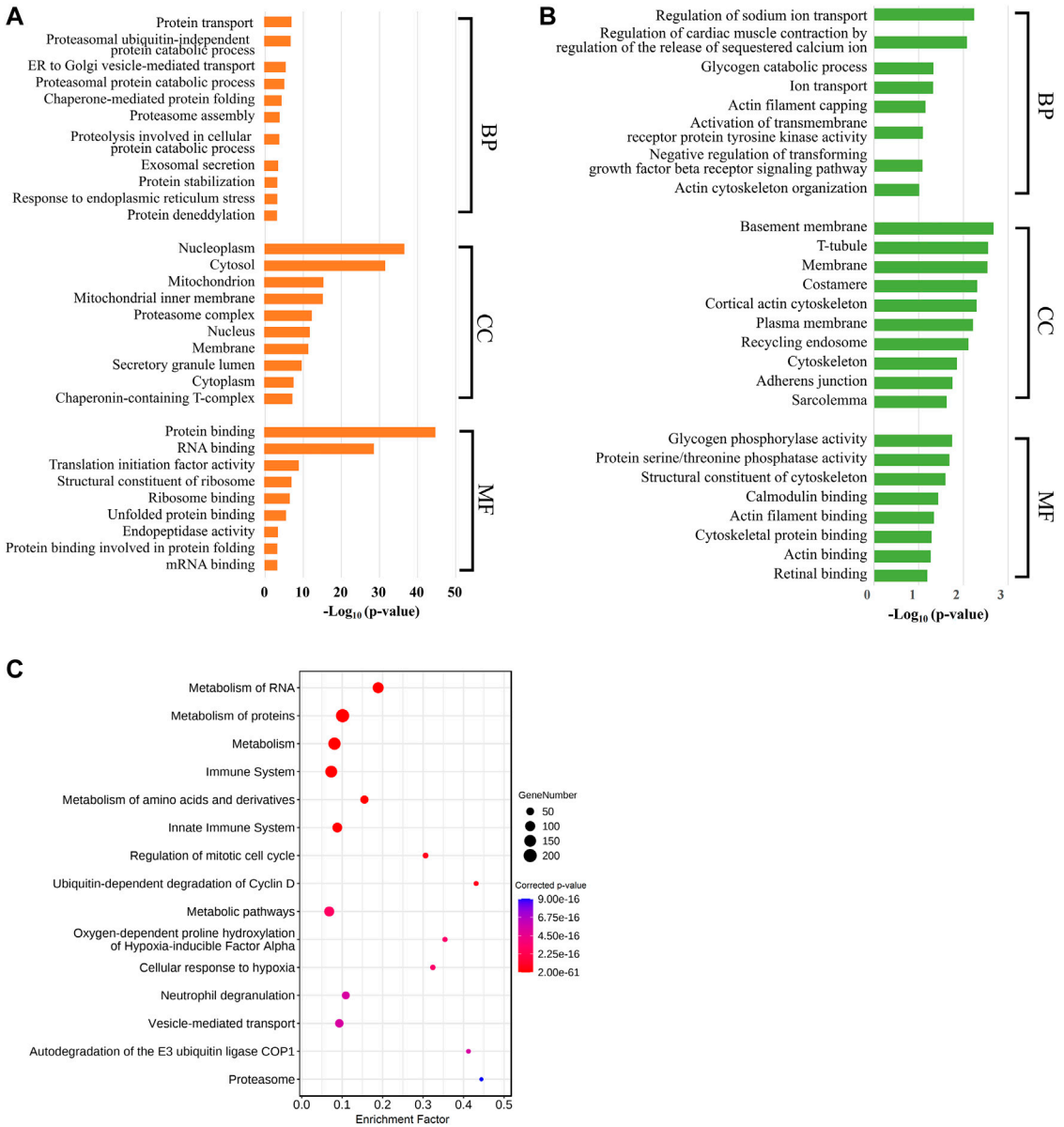

FIGURE 2 | Labor duration–correlated mRNAs function annotation. **(A)** Bar diagram of partial GO term clusters of positively correlated labor duration mRNAs. **(B)** Bar diagram of partial GO term clusters of negatively correlated labor duration mRNAs. **(C)** Bubble diagram of significantly enriched pathways for labor duration–correlated mRNAs.

targeted by two miRNAs (**Figure 3B**), which were important nodes of the network. The two circRNAs and one lncRNA of the regulatory networks were all positively correlated with labor duration (**Figure 3C**), which might be key regulatory ncRNAs for parturition.

Time-Series Analysis of mRNA With Cervical Dilation

Uterine muscle contraction during labor is closely related to cervical dilation. In the latent phase, uterine muscle contractions progress slowly. Then, the speed of dilation accelerates after 4 cm dilation, and when the myometrium reaches an active phase, the contractions become stronger and more regular. In order to identify mRNAs with significant changes in expression between latent and active phases in labor, we analyzed the time-series characteristics of RNA expression in the myometrium during labor with a cutoff of 4 cm cervical dilation. STEM, a tool for the analysis of short-time series gene expression data, was used to cluster and visualize possible changes in the profiles of all 19094 detected mRNAs at three cervical dilation phases: non-labor (n = 2), in labor dilation <4 cm (n = 8), and in labor dilation ≥4 cm (n = 7). There were six significant cluster expression patterns (red and green profiles) in the classified analysis results. In most of the significant

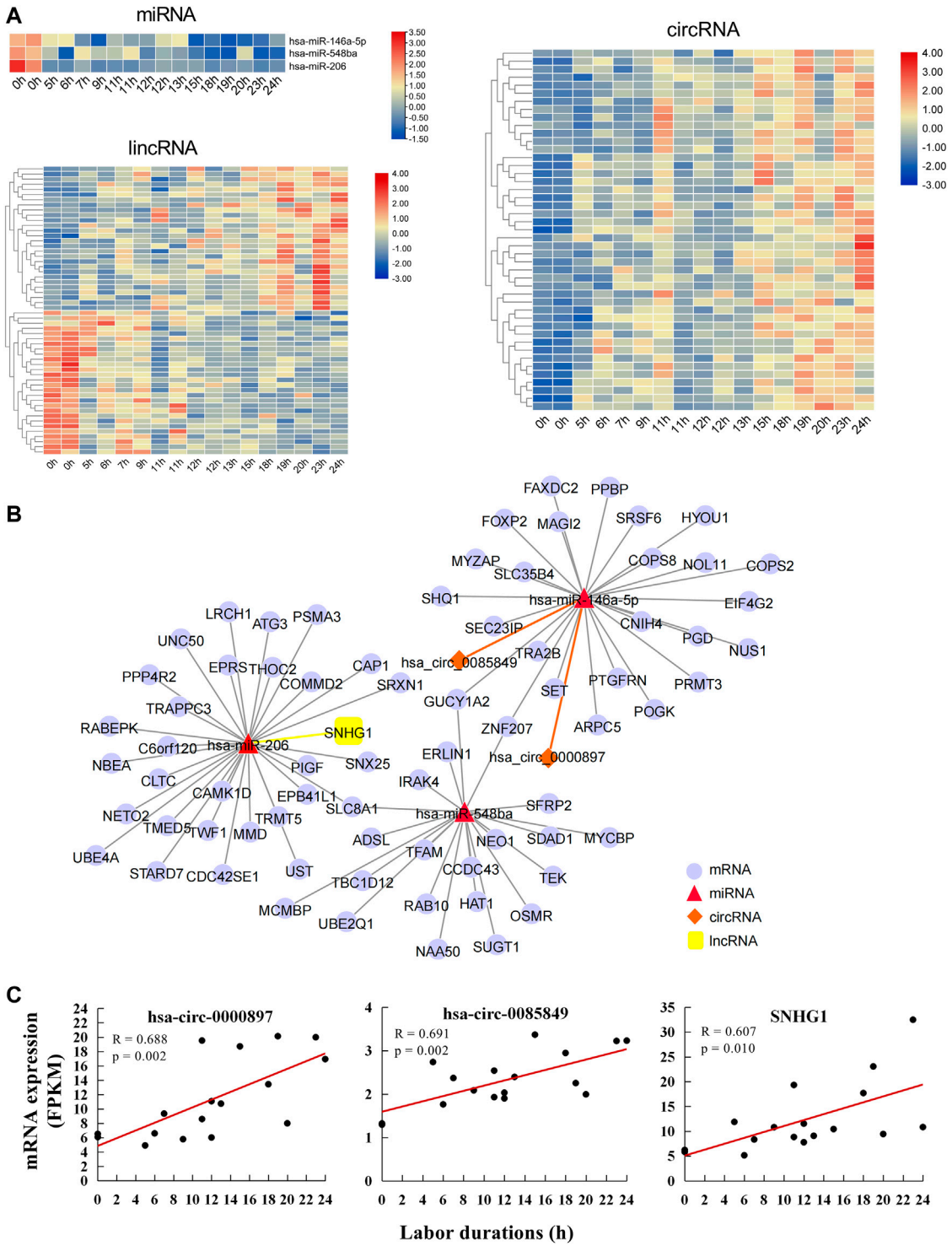

FIGURE 3 | Labor duration–correlated circRNA/lncRNA–miRNA–mRNA ceRNA network. **(A)** Heat map of expression of labor duration correlated miRNAs, lncRNAs, and circRNAs. Data associated with this figure can be found in **Supplementary Table S6**. **(B)** Labor duration correlated ceRNA network. **(C)** The expression trend of ncRNAs in ceRNA network at different labor durations.

profiles, mRNA expressions followed the same trend, either up- or down-regulated, into the active phase. It is worth noting that the mRNA expression in profile 1 did not follow the same trend, but was down-regulated at dilation <4 cm phase followed by returning to the baseline at dilation ≥4 cm phase (**Figure 4A**). Considering the latent phase (labor dilation <4 cm) which was the beginning of the

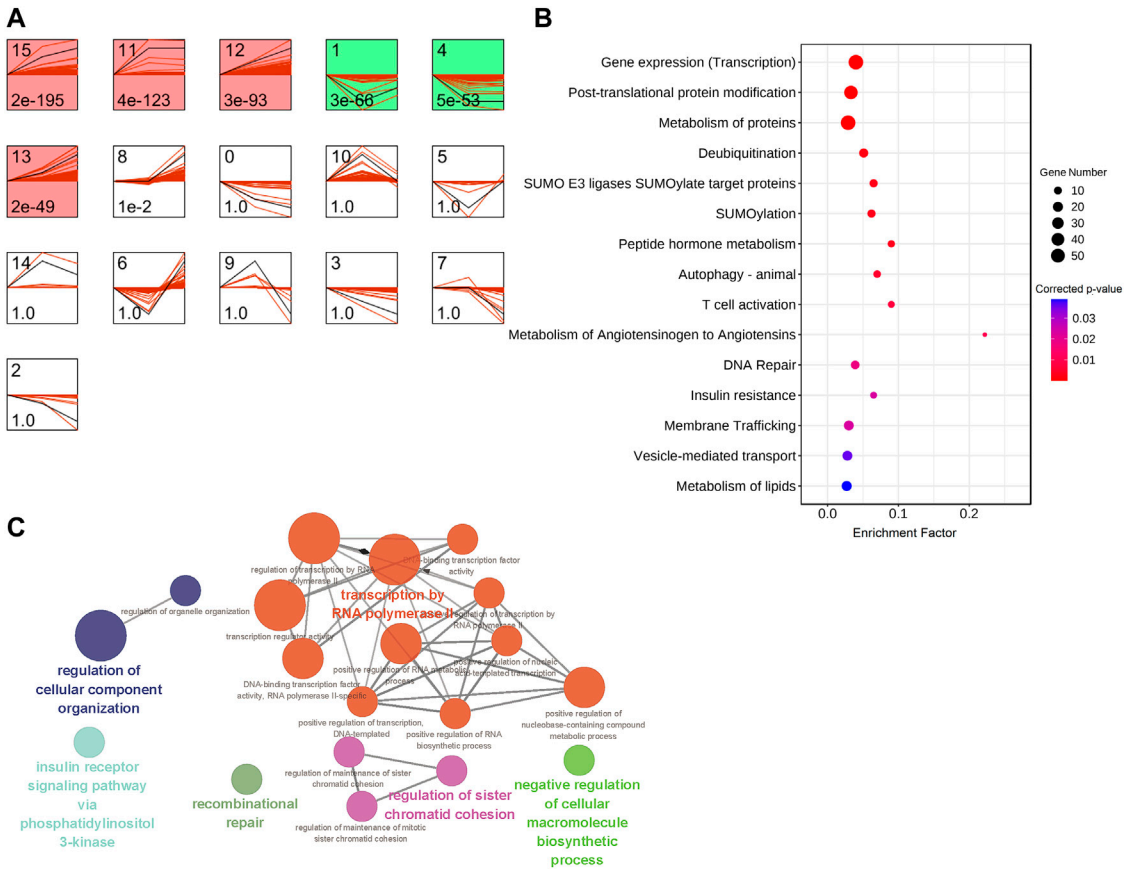

FIGURE 4 | Analysis of mRNAs expression temporal trends and functions in different phases of cervical dilatation. **(A)** STEM analysis of the mRNAs expression temporal trends pattern. The top number indicates the serial number of patterns and the bottom number indicates *p*-value. Colored profiles indicate significant clustered ($p < 0.01$) and the same color represents a similar expression pattern. Colored models meant a statistically significant number of genes were enriched. **(B)** Bubble diagram of partial significantly enriched pathways for mRNAs with more than two-fold change in expression in STEM profile 1. **(C)** ClueGO plot of enriched BP GO annotation for mRNAs with more than two-fold change in expression in STEM profile 1.

labor (the first stage of labor), the mRNAs transiently down-regulated at this phase in profile 1 and raised the possibility that some mRNAs might function specifically in triggering labor onset. STEM analysis was performed only on those cases that did not exhibit failure to progress or only on those failed to progress, and the results showed similar mRNA expression temporal trends. Specifically, for profile 1 in **Figure 4A**, the "up-down" pattern of change in gene expression was also observed in the samples from other two groups (**Supplementary Figure S1A,B**).

There were 493 mRNAs in profile 1 which were expressed more than two-fold changes in the two phases during labor, and the mRNA relative expression values (compared with group A) were exported from the STEM analysis tool (**Supplementary Table S7**). Further pathway analyses indicated that mRNAs which specifically down-regulated at dilation <4 cm phase were mainly enriched in transcription, metabolism of proteins and lipids, SUMOylation, and transportation (**Figure 4B**, **Supplementary Table S8**). GO enrichment analysis showed these mRNAs were enriched in transcription, chromosome repair, regulation of cellular component organization and macromolecule biosynthetic process, and insulin receptor signaling pathway via phosphatidylinositol 3-kinase (**Figure 4C**). These findings suggested that metabolic and transcriptional regulations were important for the initiation of labor.

Identification of Blood Loss–Correlated RNAs

The total amount of postpartum bleeding was quantified, and the correlation between bleeding volume and RNAs expression was analyzed. A total of 120 mRNAs were positively correlated and 19 mRNAs were negatively correlated with blood loss (**Figure 5A**, **Supplementary Table S9**). Functional enrichment analysis of the mRNAs correlated with postpartum bleeding revealed that these genes were associated with multiple coagulation-related functions and pathways. The mRNAs positively correlated with blood loss were enriched in cell migration, cellular response to fibroblast growth factor stimulus, and inflammatory response. CC GO terms focused on blood microparticle and haptoglobin–hemoglobin complex. In MF GO terms, protein binding, interleukin-8 receptor activity, and ubiquitin protein ligase activity were significantly enriched

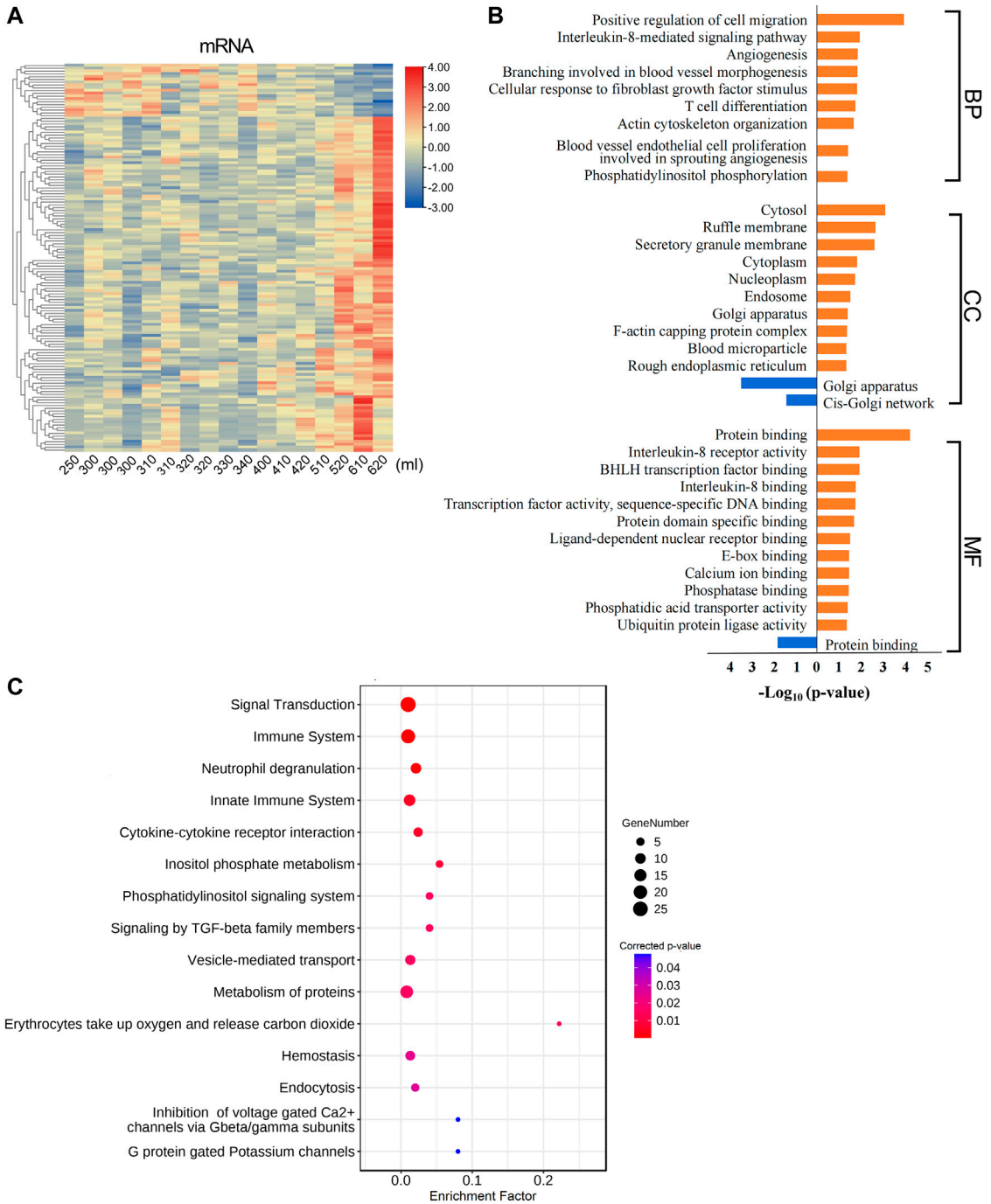

FIGURE 5 | Postpartum bleeding correlated mRNAs function annotation. **(A)** Heat map of expression of postpartum bleeding–correlated mRNAs. Data associated with this figure can be found in **Supplementary Table S9**. **(B)** Bar diagram of partial GO term clusters of positively (orange) and negatively (blue) correlated postpartum bleeding mRNAs. **(C)** Bubble diagram of significantly enriched pathways for postpartum bleeding–correlated mRNAs.

(**Figure 5B**, **Supplementary Table S10**). The pathways were enriched in hemostasis, voltage-gated Ca^{2+} channels, and immune system (**Figure 5C**, **Supplementary Table S11**). In addition, it was found that there were multiple functions related to immune and inflammatory responses, such as T cell differentiation, chemotaxis, and cytokine–cytokine receptor interaction. The functions of ferrous iron transmembrane transport and magnesium ion binding were also enriched significantly (**Figures 5B,C**). These results highlighted the

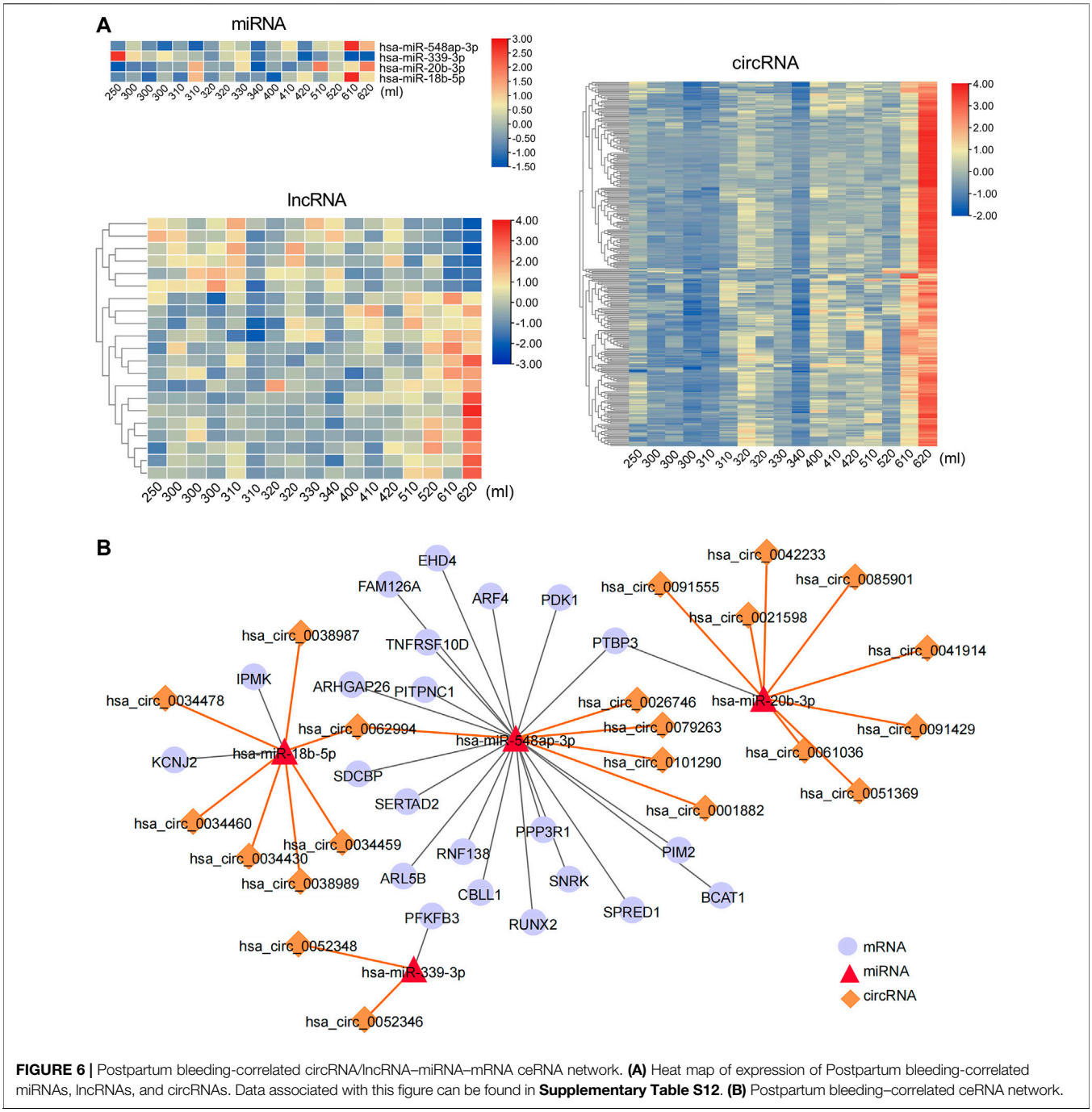

FIGURE 6 | Postpartum bleeding-correlated circRNA/lncRNA–miRNA–mRNA ceRNA network. **(A)** Heat map of expression of Postpartum bleeding-correlated miRNAs, lncRNAs, and circRNAs. Data associated with this figure can be found in **Supplementary Table S12**. **(B)** Postpartum bleeding–correlated ceRNA network.

important roles of coagulation and contraction in blood loss during labor.

To further explore the regulatory effect of ncRNAs on mRNAs related to postpartum blood loss, ncRNAs that correlated with blood loss were identified. A total of 297 ncRNAs were positively correlated with blood loss, including three miRNAs, 15 lncRNAs, and 279 circRNAs. Seven RNAs were negatively correlated with blood loss, including one miRNA and six lncRNAs (**Figure 6A**, **Supplementary Table S12**). Based on the predicted interaction relationships of the blood loss–correlated RNAs, a regulatory network consisting of four miRNAs, 22 mRNAs, and 21 circRNAs was constructed. The mRNA PTBP3 and circRNA hsa_circ_0062994 were the nodes of the network targeting two miRNAs, respectively (**Figure 6B**), suggesting their role as possible key regulators for postpartum blood loss.

DISCUSSION

Labor, a physiologic and continuous process, is traditionally divided into three stages. The first stage refers to the interval

between the onset of labor and full cervical dilatation. In the first stage, the myometrial contraction gradually becomes intense and regular, and the cervix gradually dilates (Liao et al., 2005). In this study, the transcriptome (both mRNA and ncRNA) of the myometrium from different labor durations and cervical dilation was sequenced. It was a time-course analysis that detailed a novel workflow for observing time-dependent changes in myometrial tissue gene expression for ceRNA network analysis during first stage labor over the course of 24 h, and a separate analysis was conducted for cervical dilation and postpartum blood loss status with the same study cohort. Our study presented both mRNA and ncRNA transcriptome data for each study participant, and found postpartum blood loss was correlated with changes in myometrial gene expression. Some genes and pathways closely related to labor durations and cervical dilation may be important targets for regulating myometrial contraction.

A variety of differentially expressed genes between the in labor and not in labor myometrium have been identified in previous studies. The roles of extracellular matrix interaction and calcium signal regulation in the myometrium during labor have been confirmed (Liao et al., 2005). In our analysis, mRNA-encoding proteins known to be involved in functions such as MMP25, COL4A6, CAMK1, and CAMK1D were also found to be associated with labor duration. In correlation analysis, a total of 981 mRNAs were identified to have expression levels significantly varied with the time of labor duration. In STEM analysis for the same samples, there were 493 mRNAs down-regulated in the latent phase and then up-regulated in the active phase, as shown in profile 1 derived from the STEM analysis. Many of these mRNAs were enriched in gene expression–related pathways or biological processes such as RNA degradation and transport, transcription, and protein ubiquitination, indicating that during labor, the genes expressed in the myometrium were constantly being regulated along with the process of labor.

The constant regulation of gene expression in the myometrium during labor resulted in the changes of numerous biological functions. According to the results of functional enrichment analysis of the labor duration–correlated mRNAs, metabolic process was found to be the most prominent enrichment. The myometrium undergoes hypertrophy during pregnancy, storing large amounts of glycogen, lipid, and protein in preparation for labor (Scheepers et al., 2001). The biological processes of glucose and lipid metabolism, which were the main energy supplies, were significantly enriched in mRNAs positively correlated with labor duration. Labor is an energy-intensive process, and an up-regulation of glucose and lipid metabolism can support the intense contraction of the myometrium. Our previous metabolomic profile analysis of myometrium showed that metabolism increased during labor, especially lipolysis and fatty acid oxidation (Qian et al., 2021).

As for protein metabolism, the most significant functions and pathways were protein ubiquitination and deubiquitination, which were enriched by mRNAs positively correlated with labor duration and mRNAs in STEM profile 1, suggesting a potential association to autophagy, as the results showed that the expression of autophagy-related mRNAs also increased during the active phase. Our previous studies showed that autophagy was activated in human myometrium during labor (Wang et al., 2020). Autophagy may serve as protection during transient ischemia and hypoxia caused by uterine contraction (Yan et al., 2013). Studies have shown that there was a complex cross-talk between ubiquitin-proteasome system and autophagy. Protein ubiquitination mediates autophagy and controls the initiation, execution, and termination of autophagy along with deubiquitination (Chen et al., 2019; Shaid et al., 2013). Thus, we speculate that myometrial autophagy may begin to prepare in the latent phase and occur in the active phase. The functions enriched by the mRNAs transiently down-regulated at the latent phase, such as deubiquitination, autophagy, and vesicle-mediated transport, might participate in triggering labor onset.

ceRNA network, which has been proved to be ubiquitous in post-transcriptional regulation of gene expression, interconnects encoding and non-encoding RNAs regulation and works with other cellular and molecular regulation mechanisms (Tay et al., 2014). ceRNA networks were previously reported in carcinogenesis (Ala, 2021), yet it is unclear in the myometrium during labor. The ceRNA network constructed in this study demonstrated the regulatory relationship among different kinds of transcripts correlated with labor duration. We identified three mRNAs that formed the connection points of the whole network; specifically SLC8A1, GUCY1A2, and ZNF207. SLC8A1 encodes sodium/calcium exchanger protein, which contributes to Ca^{2+} transport during excitation–contraction coupling in muscle (Shattock et al., 2015). GUCY1A2 encodes the α2 subunit of soluble guanosine cyclase (sGC), and the sGC-catalyzed production of cyclic guanosine phosphate (cGMP) is involved in the relaxation of smooth muscle in human vas deferens and airways (Britt et al., 2015; Da et al., 2012). ZNF207 encodes kinetochore- and microtubule-binding protein that participates in spindle assembly by blocking ubiquitination and proteasomal degradation of mitotic checkpoint protein BUB3 (Jiang et al., 2014). These mRNAs and the other RNAs communicating with them may regulate the contraction and metabolism of the myometrium during labor, though further investigation is still needed.

Postpartum hemorrhage, a complicated multifactorial process, remains a leading cause of maternal morbidity and mortality. The common causes of excessive bleeding are uterine atony (70%), retained placenta, genital tract injuries, and coagulopathy (Oyelese et al., 2010). According to the results of our transcriptome profiles in different amounts of bleeding after delivery, a number of differentially expressed RNAs were identified. Most of these mRNAs were positively correlated with bleeding volume and markedly enriched functions and pathways of coagulation, inflammatory response, and blood vessel endothelial cell proliferation. Inflammation and wound healing are closely related to the vascular endothelium, and vascular endothelial growth factor is a key factor that regulates this process (Stanley et al., 2015). Our results showed a positive correlation between

inflammatory response and blood vessel endothelial cell alteration and the amount of postpartum blood loss. Eight genes were associated with hemostatic pathways (SELL, CAPZB, SLC16A3, PDPN, TNFRSF10B, SH2B2, NFE2, and TNFRSF10D), which provided novel insights for the prevention and management of postpartum hemorrhage. Case–control studies including patients with postpartum hemorrhage are needed to confirm our findings. The mechanism through which these genes regulate hemostasis in labor requires further studies in animal experiments.

There are several limitations in this study that should be noted. Due to the difficulty of myometrium tissue collection, the transcriptome data were derived from a limited number of samples, and these myometrium variables (labor duration and blood loss) were not evenly distributed for each time point. The study will benefit from validation of results with larger sample cohorts, which requires further research. In this study, six participants underwent failure to progress. Failure to progress was due to many reasons. In our study, the cases of "failure to progress" were all clinically considered to be caused by cephalopelvic disproportion instead of primary uterine atony. The assessment of the labor start time primarily relied on routine clinical method (determined using cardiotocography and by digital examination after self-reporting regular contractions); therefore, it was nearly impossible to record the precise time point when the labor started, even though all the participants were hospitalized before labor started. Similarly, blood loss was estimated using routine measurement methods, which were a mix of uterine bleeding from caesarean incision and postoperative vaginal bleeding. In addition, transcriptome data were derived from total RNA of the whole myometrial tissues, and we were unable to determine which type of cell in the human myometrium contributed to the significant changes in gene expression identified from its data sets. Single-cell omics can provide a better investigation of the cell-specific changes in the myometrium, which could help to identify appropriate targets for future clinical interventions.

By utilizing RNA-seq with advanced bioinformatics techniques, we have shown that there were significant changes in the transcription levels in the myometrium at different phases of labor. Then we analyzed the functions and pathway alterations and constructed a regulatory network of parturition. Our study presented a method of participants' selection criteria of labor duration and cervical dilation status. Transcriptome and its ceRNA network correlated with labor duration and blood loss provided certain potential key RNAs for subsequent molecular mechanism research, which could help determine the causes of changes in human myometrium function during physiological and pathological labor.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in Genome Sequence Archive (GSA) repository with accession number PRJCA009585.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Guangzhou Women and Children Medical Center. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

HL and KJ designed research; LC, KJ, and LW analyzed data and drafted the article; YL, YC, XW, and GZ collected samples and performed research. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This study was funded by the National Natural Science Foundation of China (81871181), Foundation of Guangzhou Municipal Science and Technology Bureau (202102010016), and High-tech Major Featured Technology Program of Guangzhou Municipal Health Commission (2019GX07).

ACKNOWLEDGMENTS

We thank the obstetrics and midwifery staff of the Guangzhou Women and Children's Medical Center for their clinical sample collection cooperation, and Ming Lei for his contribution to the construction of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.812105/full#supplementary-material.

REFERENCES

Ackerman IV, W. E., Buhimschi, C. S., Snedden, A., Summerfield, T. L., Zhao, G., and Buhimschi, I. A. (2021). Molecular Signatures of Labor and Non-labor Myometrium with Parsimonious Classification from Two Calcium Transporter Genes. *JCI Insight* 6 (11), 148425. doi:10.1172/jci.insight.148425

Ala, U. (2021). Competing Endogenous RNAs and Cancer: How Coding and Non-coding Molecules Cross-Talk Can Impinge on Disease. *Int. J. Biochem. Cell. Biol.* 130, 105874. doi:10.1016/j.biocel.2020.105874

Ambros, V. (2004). The Functions of Animal microRNAs. *Nature* 431 (7006), 350–355. doi:10.1038/nature02871

Breathnach, F., and Geary, M. (2009). Uterine Atony: Definition, Prevention, Nonsurgical Management, and Uterine Tamponade. *Seminars Perinatology* 33 (2), 82–87. doi:10.1053/j.semperi.2008.12.001

Breuiller-Fouche, M., and Germain, G. (2006). Gene and Protein Expression in the Myometrium in Pregnancy and Labor. *Reproduction* 131 (5), 837–850. doi:10.1530/rep.1.00725

Britt, R. D., Thompson, M. A., Kuipers, I., Stewart, A., Vogel, E. R., Thu, J., et al. (2015). Soluble Guanylate Cyclase Modulators Blunt Hyperoxia Effects on Calcium Responses of Developing Human Airway Smooth Muscle. *Am. J. Physiology-Lung Cell. Mol. Physiology* 309 (6), L537–L542. doi:10.1152/ajplung.00232.2015

Bu, D., Luo, H., Huo, P., Wang, Z., Zhang, S., He, Z., et al. (2021). KOBAS-I: Intelligent Prioritization and Exploratory Visualization of Biological Functions for Gene Enrichment Analysis. *Nucleic Acids Res.* 49 (W1), W317–W325. doi:10.1093/nar/gkab447

Campbell, J. D., Liu, G., Luo, L., Xiao, J., Gerrein, J., and Juan-Guardela, B. (2015). Assessment of MicroRNA Differential Expression and Detection in Multiplexed Small RNA Sequencing Data. *RNA*. 21(2), 164–171. doi:10.1261/rna.046060.114

Chan, Y. W., van den Berg, H. A., Moore, J. D., Quenby, S., and Blanks, A. M. (2014). Assessment of Myometrial Transcriptome Changes Associated with Spontaneous Human Labour by High-Throughput RNA-Seq. *Exp. Physiol.* 99 (3), 510–524. doi:10.1113/expphysiol.2013.072868

Chen, L., Wang, L., Luo, Y., Huang, Q., Ji, K., Bao, J., et al. (2021). Integrated Proteotranscriptomics of Human Myometrium in Labor Landscape Reveals the Increased Molecular Associated with Inflammation under Hypoxia Stress. *Front. Immunol.* 12, 722816. doi:10.3389/fimmu.2021.722816

Chen, R. H., Chen, Y. H., and Huang, T. Y. (2019). Ubiquitin-mediated Regulation of Autophagy. *J. Biomed. Sci.* 26 (1), 80. doi:10.1186/s12929-019-0569-y

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A Survey of Best Practices for RNA-Seq Data Analysis. *Genome Biol.* 17, 13. doi:10.1186/s13059-016-0881-8

Da, S. F., Claudino, M. A., Báu, F. R., Rojas-Moscoso, J. A., Mónica, F. Z., De Nucci, G., et al. (2012). Vas Deferens Smooth Muscle Responses to the Nitric Oxide-independent Soluble Guanylate Cyclase Stimulator BAY 41-2272. *Eur. J. Pharmacol.* 688 (1-3), 49–55. doi:10.1016/j.ejphar.2012.05.009

Ernst, J., and Bar-Joseph, Z. (2006). STEM: A Tool for the Analysis of Short Time Series Gene Expression Data. *BMC Bioinforma.* 7 (1), 191. doi:10.1186/1471-2105-7-191

Flores-Pliego, A., Espejel-Nuñez, A., Castillo-Castrejon, M., Meraz-Cruz, N., Beltran-Montoya, J., Zaga-Clavellina, V., et al. (2015). Matrix Metalloproteinase-3 (MMP-3) Is an Endogenous Activator of the MMP-9 Secreted by Placental Leukocytes: Implication in Human Labor. *PLoS One* 10 (12), e145366. doi:10.1371/journal.pone.0145366

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., et al. (2008). Discovering microRNAs from Deep Sequencing Data Using miRDeep. *Nat. Biotechnol.* 26 (4), 407–415. doi:10.1038/nbt1394

Friedman, E. A. (1955). Primigravid Labor; A Graphicostatistical Analysis. *Obstet. Gynecol.* 6 (6), 567–589. doi:10.1097/00006250-195512000-00001

Friedman, E. A. (1996). The Length of Active Labor in Normal Pregnancies. *Obstet. Gynecol.* 88 (2), 319–320. doi:10.1016/s0029-7844(96)80258-4

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: An Efficient and Unbiased Algorithm for De Novo Circular RNA Identification. *Genome Biol.* 16 (1), 4. doi:10.1186/s13059-014-0571-3

Guttman, M., and Rinn, J. L. (2012). Modular Regulatory Principles of Large Non-coding RNAs. *Nature* 482 (7385), 339–346. doi:10.1038/nature10887

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4 (1), 44–57. doi:10.1038/nprot.2008.211

Jiang, H., He, X., Wang, S., Jia, J., Wan, Y., Wang, Y., et al. (2014). A Microtubule-Associated Zinc Finger Protein, BuGZ, Regulates Mitotic Chromosome Alignment by Ensuring Bub3 Stability and Kinetochore Targeting. *Dev. Cell.* 28 (3), 268–281. doi:10.1016/j.devcel.2013.12.013

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: A Fast Spliced Aligner with Low Memory Requirements. *Nat. Methods.* 12 (4), 357–360. doi:10.1038/nmeth.3317

Kopp, F., and Mendell, J. T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell.* 172 (3), 393–407. doi:10.1016/j.cell.2018.01.011

Krapohl, A. J., Myers, G. G., and Caldeyro-Barcia, R. (1970). Uterine Contractions in Spontaneous Labor. A Quantitative Study. *Am. J. Obstet. Gynecol.* 106 (3), 378–387. doi:10.1016/0002-9378(70)90363-7

Lai, P. F., Lei, K., Zhan, X., Sooranna, G., Li, J., Georgiou, E. X., et al. (2021). Labour Classified by Cervical Dilatation & Fetal Membrane Rupture Demonstrates Differential Impact on RNA-Seq Data for Human Myometrium Tissues. *PLoS One* 16 (11), e260119. doi:10.1371/journal.pone.0260119

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biol.* 10 (3), R25. doi:10.1186/gb-2009-10-3-r25

Li, B., and Dewey, C. N. (2011). RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome. *BMC Bioinforma.* 12, 323. doi:10.1186/1471-2105-12-323

Li, J. K. H., Lai, P. F., Tribe, R. M., and Johnson, M. R. (2021). Transcription Factors Regulated by cAMP in Smooth Muscle of the Myometrium at Human Parturition. *Biochem. Soc. Trans.* 49 (2), 997–1011. doi:10.1042/BST20201173

Liao, J. B., Buhimschi, C. S., and Norwitz, E. R. (2005). Normal Labor: Mechanism and Duration. *Obstet. Gyn. Clin. N. Am.* 32 (2), 145–164. doi:10.1016/j.ogc.2005.01.001

Liu, M., Wang, Q., Shen, J., Yang, B. B., and Ding, X. (2019). Circbank: A Comprehensive Database for circRNA with Standard Nomenclature. *RNA Biol.* 16 (7), 899–905. doi:10.1080/15476286.2019.1600395

Lombardi, A., Makieva, S., Rinaldi, S. F., Arcuri, F., Petraglia, F., and Norman, J. E. (2017). Expression of Matrix Metalloproteinases in the Mouse Uterus and Human Myometrium during Pregnancy, Labor, and Preterm Labor. *Reprod. Sci.* 25 (6), 938–949. doi:10.1177/1933719117732158

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs Are a Large Class of Animal RNAs with Regulatory Potency. *Nature* 495 (7441), 333–338. doi:10.1038/nature11928

Misme, H., Dupont, C., Cortet, M., Rudigoz, R. C., and Huissoud, C. (2016). Distribution of Blood Loss during Vaginal Delivery and Cesarean Section. *J. Gynecol. Obstet. Biol. Reprod. Paris.* 45 (1), 71–79. doi:10.1016/j.jgyn.2015.01.004

Nawrocki, E., and Eddy, S. (2013). Infernal 1.1: 100-Fold Faster RNA Homology Searches. *Bioinformatics* 29 (22), 2933–2935. doi:10.1093/bioinformatics/btt509

Oyelese, Y., and Ananth, C. V. (2010). Postpartum Hemorrhage: Epidemiology, Risk Factors, and Causes. *Clin. Obstet. Gynecol.* 53 (1), 147–156. doi:10.1097/GRF.0b013e3181cc406d

Panni, S., Lovering, R. C., Porras, P., and Orchard, S. (2020). Non-coding RNA Regulatory Networks. *Biochim. Biophys. Acta Gene Regul. Mech.* 1863 (6), 194417. doi:10.1016/j.bbagrm.2019.194417

Papandreou, L., Chasiotis, G., Seferiadis, K., Thanasoulias, N. C., Dousias, V., Tsanadis, G., et al. (2004). Calcium Levels during the Initiation of Labor. *Eur. J. Obstet. Gynecol. Reprod. Biol.* 115 (1), 17–22. doi:10.1016/j.ejogrb.2003.11.032

Paraskevopoulou, M. D., Georgakilas, G., Kostoulas, N., Vlachos, I. S., Vergoulis, T., Reczko, M., et al. (2013). DIANA-microT Web Server v5.0: Service Integration into miRNA Functional Analysis Workflows. *Nucleic Acids Res.* 41, W169–W173. doi:10.1093/nar/gkt393

Paraskevopoulou, M. D., Vlachos, I. S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., et al. (2016). DIANA-LncBase V2: Indexing microRNA Targets on Non-coding Transcripts. *Nucleic Acids Res.* 44 (D1), D231–D238. doi:10.1093/nar/gkv1270

Pehlivanoğlu, B., Bayrak, S., and Doğan, M. (2013). A Close Look at the Contraction and Relaxation of the Myometrium; the Role of Calcium. *J. Turk Ger. Gynecol. Assoc.* 14 (4), 230–234. doi:10.5152/jtgga.2013.67763

Qian, X., Wang, L., Lin, B., Luo, Y., Chen, Y., and Liu, H. (2021). Maternal Myometrium Metabolomic Profiles in Labor: Preliminary Results. *Gynecol. Obstet. Invest.* 86, 88–93. doi:10.1159/000512460

Renthal, N. E., Williams, K. C., and Mendelson, C. R. (2013). MicroRNAs--mediators of Myometrial Contractility during Pregnancy and Labour. *Nat. Rev. Endocrinol.* 9 (7), 391–401. doi:10.1038/nrendo.2013.96

Romero, R., Tarca, A. L., Chaemsaithong, P., Miranda, J., Chaiworapongsa, T., Jia, H., et al. (2014). Transcriptome Interrogation of Human Myometrium Identifies Differentially Expressed Sense-Antisense Pairs of Protein-Coding and Long Non-coding RNA Genes in Spontaneous Labor at Term. *J. Maternal-Fetal Neonatal Med.* 27 (14), 1397–1408. doi:10.3109/14767058.2013.860963

Scheepers, H. C., de Jong, P. A., Essed, G. G., and Kanhai, H. H. (2001). Fetal and Maternal Energy Metabolism during Labor in Relation to the

Available Caloric Substrate. *J. Perinat. Med.* 29 (6), 457–464. doi:10.1515/JPM.2001.064

Shaid, S., Brandts, C. H., Serve, H., and Dikic, I. (2013). Ubiquitination and Selective Autophagy. *Cell. Death Differ.* 20 (1), 21–30. doi:10.1038/cdd.2012.72

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Shattock, M. J., Ottolia, M., Bers, D. M., Blaustein, M. P., Boguslavskyi, A., Bossuyt, J., et al. (2015). Na+/Ca2+ Exchange and Na+/K+-ATPase in the Heart. *J. Physiol.* 593 (6), 1361–1382. doi:10.1113/jphysiol.2014.282319

Shchuka, V. M., Abatti, L. E., Hou, H., Khader, N., Dorogin, A., Wilson, M. D., et al. (2020). The Pregnant Myometrium Is Epigenetically Activated at Contractility-Driving Gene Loci Prior to the Onset of Labor in Mice. *PLoS Biol.* 18 (7), e3000710. doi:10.1371/journal.pbio.3000710

Stanley, R., Ohashi, T., and Mowa, C. (2015). Postpartum Cervical Repair in Mice: A Morphological Characterization and Potential Role for Angiogenic Factors. *Cell. Tissue Res.* 362 (1), 253–263. doi:10.1007/s00441-015-2184-x

Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The Multilayered Complexity of ceRNA Crosstalk and competition[Journal Article; Research Support. *Nature* 505 (7483), 344–352. doi:10.1038/nature12986

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation. *Nat. Biotechnol.* 28 (5), 511–515. doi:10.1038/nbt.1621

Wang, L., Hu, H., Morse, A. N., Han, X., Bao, J., Yang, J., et al. (2020). Activation of Autophagy in Human Uterine Myometrium during Labor. *Reprod. Sci.* 27 (8), 1665–1672. doi:10.1007/s43032-020-00198-3

Wilcox, C. R., Hunt, A. B., and Owens, C. J. (1959). The Measurement of Blood Lost during Cesarean Section. *Am. J. Obstet. Gynecol.* 77 (4), 772–779. doi:10.1016/s0002-9378(16)36792-8

Wray, S., and Prendergast, C. (2019). The Myometrium: From Excitation to Contractions and Labour. *Adv. Exp. Med. Biol.* 1124, 233–263. doi:10.1007/978-981-13-5895-1_10

Yan, W. J., Dong, H. L., and Xiong, L. Z. (2013). The Protective Roles of Autophagy in Ischemic Preconditioning. *Acta Pharmacol. Sin.* 34 (5), 636–643. doi:10.1038/aps.2013.18

Zhang, C., Huo, S. T., Wu, Z., Chen, L., Wen, C., Chen, H., et al. (2020). Rapid Development of Targeting circRNAs in Cardiovascular Diseases. *Mol. Ther. - Nucleic Acids.* 21, 568–576. doi:10.1016/j.omtn.2020.06.022

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 | Frontiers in Genetics

ORIGINAL RESEARCH
published: 22 July 2022
doi: 10.3389/fgene.2022.833362

Construction of a Novel Prognostic Signature in Lung Adenocarcinoma Based on Necroptosis-Related lncRNAs

*Xiayao Diao, Chao Guo and Shanqing Li**

Department of Thoracic Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Background: Long non-coding RNAs (lncRNAs) are drawing increasing attention as promising predictors of prognosis for lung adenocarcinoma (LUAD) patients. Necroptosis, a novel regulated mechanism of necrotic cell death, plays an important role in the biological process of cancer. The aim of this study was to identify the necroptosis-related lncRNAs (NRLRs) in a LUAD cohort and establish a necroptosis-related lncRNA signature (NRLSig) to stratify LUAD patients.

Methods: NRLRs were identified in LUAD patients from The Cancer Genome Atlas (TCGA) database using Pearson correlation analysis between necroptosis-related genes and lncRNAs. Then the NRLSig was identified using univariate Cox regression analysis and LASSO regression analysis. Assessments of the signature were performed based on survival analysis, receiver operating characteristic (ROC) curve analysis and clustering analysis. Next, a nomogram containing the NRLSig and clinical information was developed through univariate and multivariate Cox regression analysis. Further, functional enrichment analysis of the selected lncRNAs in NRLSig and the association between NRLSig and the immune infiltration were also evaluated.

Results: A 4-lncRNA signature, incorporating LINC00941, AP001453.2, AC026368.1, and AC236972.3, was identified to predict overall survival (OS) and stratify LUAD patients into different groups. Survival analysis, ROC curve analysis and clustering analysis showed good performance in the prognostic prediction of the lncRNA signature. Then, a nomogram containing the NRLSig was developed and showed satisfactory predictive accuracy, calibration and clinical usefulness. The co-expressed genes of selected NRLRs were enriched in several biological functions and signaling pathways. Finally, differences in the abundance of immune cells were investigated among the high-risk group and low-risk group divided by the NRLSig.

Conclusion: The proposed NRLSig may provide promising therapeutic targets or prognostic predictors for LUAD patients.

Keywords: necroptosis, long non-coding RNA, prognostic signature, lung adenocarcinoma, overall survival

OPEN ACCESS

Edited by:
Jiangning Song,
Monash University, Australia

Reviewed by:
Yong-Zi Chen,
Tianjin Medical University Cancer Institute and Hospital, China
Yanhong Zhou,
Central South University, China
Song Xu,
Tianjin Medical University General Hospital, China

****Correspondence:***
Shanqing Li
lishanqing@pumch.cn

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

Received: *11 December 2021*
Accepted: *18 May 2022*
Published: *22 July 2022*

Citation:
Diao X, Guo C and Li S (2022) Construction of a Novel Prognostic Signature in Lung Adenocarcinoma Based on Necroptosis-Related lncRNAs. Front. Genet. 13:833362. doi: 10.3389/fgene.2022.833362

INTRODUCTION

Lung cancer is one of the most common malignancies and the leading cause of cancer-associated deaths worldwide (Siegel et al., 2020). Non-small cell lung cancer (NSCLC) is the most frequently reported subtype, accounting for approximately 85% of all lung cancer cases (Meza et al., 2015). NSCLC includes three main histological subtypes: adenocarcinoma, squamous cell carcinoma and large cell carcinoma (Zappa and Mousa, 2016). Among them, Lung adenocarcinoma (LUAD) is the most prevalent histotype. Although diagnostic techniques and therapeutic strategies have been developed for LUAD patients, the 5-year overall survival (OS) rate of them remains only 15% (Miller et al., 2012). Therefore, it is urgent to identify some novel effective diagnostic markers, therapeutic targets, and prognostic factors to offer early diagnosis, timely treatment, and precise prediction for LUAD patients.

Tumorigenesis and drug resistance are often attribute to resistance to apoptosis in most tumors (Johnstone et al., 2002; Pan et al., 2021). This phenomenon calls for identifying strategies to induce non-apoptotic approaches of programmed cell death as promising novel therapeutics in cancer (Tang R. et al., 2020). Necrosis used to be recognized as a completely opposite form of cell death compared to apoptosis (Linkermann and Green, 2014). However, necroptosis, an alternative regulated cell death, can be elicited by the activation of various signaling pathways, tumor microenvironmental stresses, or multiple chemotherapeutic drugs (Huang et al., 2013; Lalaoui et al., 2015; Galluzzi et al., 2018). Emerging evidence illustrates that necroptosis act as a crucial approach in the regulation of biological processes of tumor, including oncogenesis, progression, metastasis, cancer immunity, and cancer subtypes (Stoll et al., 2017; Seehawer et al., 2018). Manipulating or targeting the necroptotic pathway may also play an important role for bypassing resistance of apoptosis, supporting anti-cancer immunity in cancer therapy and predicting prognosis for cancer patients (Gong et al., 2019).

Long non-coding RNAs (lncRNAs), non-protein-coding transcripts longer than 200 ribonucleotides, play a pivotal role in gene regulation (Agostini et al., 2020). LncRNAs participate in various biological processes, such as immune, metabolism, infection, and more (Gibb et al., 2011; Tan et al., 2021). In addition, lncRNAs exert these functions by interacting with other molecules such as RNA, DNA, and proteins (Md Yusof et al., 2020). In recent years, with the development of high-throughput sequencing techniques, increasing studies have demonstrated many non-coding genes play an important role in the development and progression of tumors. LncRNAs have also been revealed to function as regulators in cancer biology, including proliferation, invasion, and metastasis (Hung et al., 2014; Kim et al., 2014), as well as tumor angiogenesis or lymphangiogenesis (Prensner et al., 2014; He et al., 2018). Notably, it has been revealed that several lncRNAs may act as mediators regulating necroptosis in different tumors. For example, lncRNA H19, as a precursor of miR-675, regulates necroptosis *via* miR-675 in hepatocellular carcinoma (Harari-Steinfeld et al., 2021). Moreover, 16 lncRNAs associated with necroptosis were also identified in gastric cancer patients through bioinformatic analysis (Zhao et al., 2021). However, only small amount of lncRNAs, especially the NRLRs, have been functionally or prognostically well-characterized. Therefore, it is valuable to identify key lncRNAs closely related to necroptosis with prognosis significance in LUAD.

In present study, the lncRNAs expression profiles of LUAD patients were collected from public database. We then developed a necroptosis-related lncRNA signature (NRLSig) and systematically evaluated the associations of necroptosis-related lncRNAs (NRLRs) with the prognosis and clinical or pathological characteristics of LUAD patients. Moreover, we established a nomogram that incorporates the NRLSig and clinical factors to further stratify these patients. The high-risk group and low-risk group identified by NRLSig were compared based on various factors, including tumor-infiltrating immune cells and principal component analysis (PCA). Finally, functional enrichment analysis was also conducted to explore the potential mechanism of the selected lncRNAs. This study revealed the prognostic value of NRLRs in LUAD and constructed a prognostic model to evaluate prognosis of LUAD patients.

MATERIALS AND METHODS

Data Collection

The RNA transcriptome datasets of 535 LUAD patients, including 535 tumor samples and 59 adjacent normal samples, were obtained from The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov). The detailed clinicopathological information, including survival status, survival time, age, gender, TNM stage, T stage, N stage, and M stage, were also downloaded from the above dataset. Only LUAD patients with clear survival time and survival status were included in the study. Patients whose OS was less than 30 days in the TCGA-LUAD database were excluded to reduce statistical bias in this analysis. With corresponding clinical information, the LUAD patients who fit the criteria above were divided into a training set and validation set randomly in a 1:1 ratio by using the "caret" R package.

Identification of Necroptosis-Related lncRNAs in Lung Adenocarcinoma

According to the lncRNAs annotation file obtained from the GENCODE (https://www.gencodegenes.org/) (Derrien et al., 2012), 14,128 lncRNAs were acquired from the RNA transcriptome datasets of the TCGA-LUAD. The differentially expressed lncRNAs between LUAD and normal tissues were identified in the TCGA cohort with false discovery rate (FDR) < 0.05 and $|\text{Log}_2$ fold change (FC)$| \geq 1$ (Glickman et al., 2014). The differential expression analysis was conducted using the "limma" package. To visualize the screening results for differentially expressed lncRNAs, we also plotted the heatmap and volcano plot using the "pheatmap" R package. Moreover, necroptosis-related genes (NRGs) were identified from two sources. 159 NRGs were extracted in the necroptosis pathway (hsa04217) from the KEGG PATHWAY database (https://www.kegg.jp/). 67 NRGs were obtained from literature research (Zhao et al., 2021). Finally, a total of 204 NRGs were included for

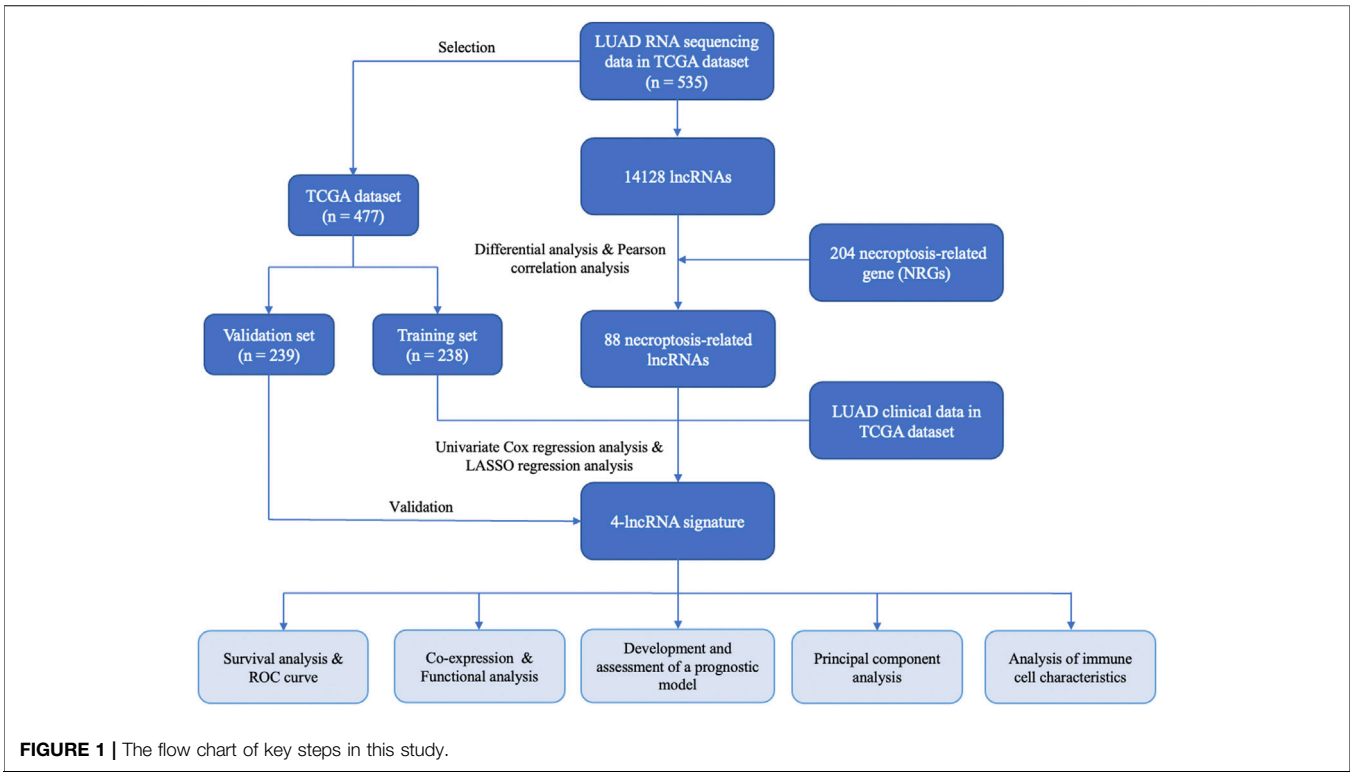

FIGURE 1 | The flow chart of key steps in this study.

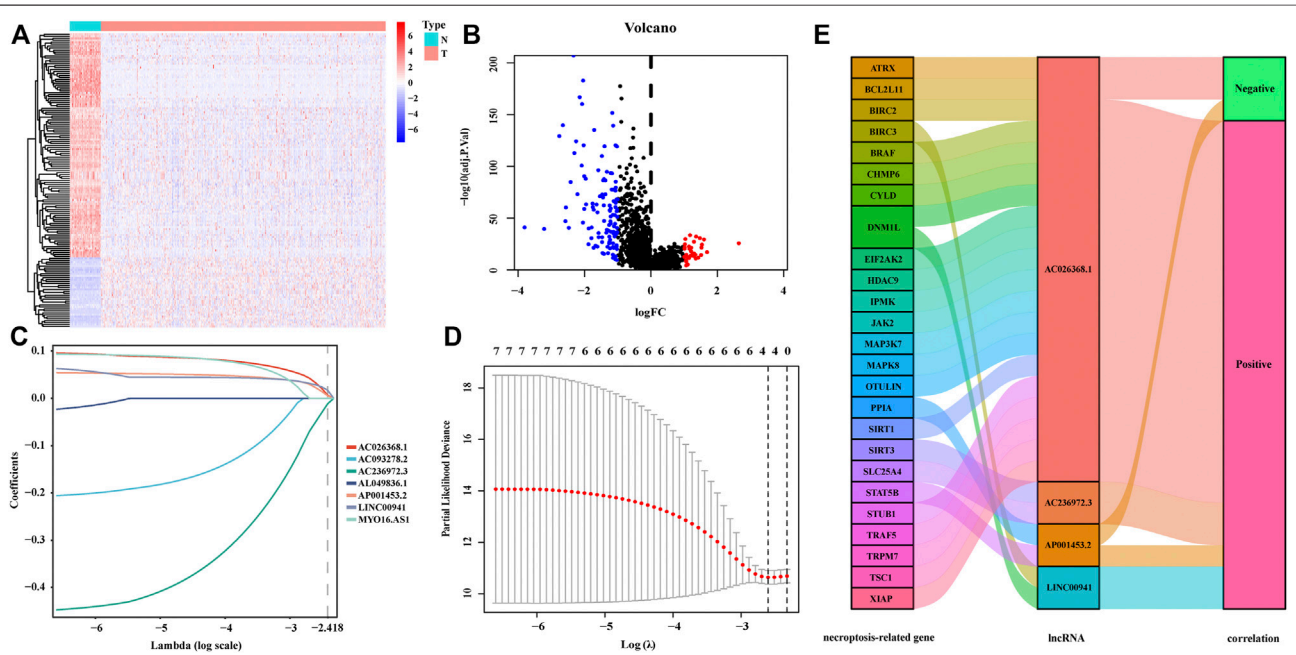

FIGURE 2 | Construction of the signature incorporating necroptosis-related lncRNAs. The heatmap **(A)** and the volcano plot **(B)** of differentially expressed lncRNAs in LUAD patients. **(C,D)** The LASSO coefficient profiles of the 7 lncRNAs at different tuning parameters (λ), 10-fold cross-validation to filtrate candidate necroptosis-related lncRNAs in LASSO regression analysis. **(E)** The Sankey diagram of the correlation between necroptosis-related genes (NRGs) and lncRNAs. The left bar represents the NRGs. The median bar represents the selected lncRNAs. The right bar represents the correlations between NRGs and lncRNAs generated from the Pearson correlation analysis.

TABLE 1 | Clinical features of selected lung adenocarcinoma (LUAD) patients in TCGA dataset. TCGA, The Cancer Genome Atlas.

Characteristic N (477) %	N (477)	%
Gender		
Male	220	46.1
Female	257	53.9
Age, years		
≤65	230	48.2
>65	247	51.8
TNM stage		
Stage I	253	53.0
Stage II	113	23.7
Stage III	78	16.4
Stage IV	25	5.2
Unknown	8	1.7
T stage		
T_1	159	33.3
T_2	254	53.2
T_3	43	9.0
T_4	18	3.8
Unknown	3	0.6
N stage		
N_0	307	64.4
N_1	90	18.9
N_2	67	14.0
N_3	2	0.4
Unknown	11	2.3
M stage		
M_0	447	93,7
M_1	26	5.5
Unknown	4	0.8
Survival status		
Alive	320	67.1
Dead	157	32.8

subsequent research after integrating intersection from these two gene sets (**Supplementary Table S1**). Pearson correlation analysis was conducted between the differentially expressed lncRNAs and 204 NRGs (with the |Correlation Coefficient| > 0.3 and $p < 0.001$) to identify NRLRs using the "limma" package.

Establishment of the Prognostic Necroptosis-Related lncRNA Signature for Lung Adenocarcinoma

According to the corresponding survival information of LUAD cases in the training set, univariate Cox analysis for association with OS was conducted to identify the prognostic NRLRs. Finally, the lncRNAs with p value < 0.05 were selected to further establish the NRLSig through the least absolute shrinkage and selection operator (LASSO) regression algorithm using the "glmnet" package in R software (Tibshirani, 1997; Simon et al., 2011), and 10-fold cross-validation was utilized to filtrate candidate NRLRs and identify the penalty parameter (λ), corresponding to the minimum value of partial likelihood deviance. A risk signature was then developed based on the risk coefficients and the expression levels of optimal prognostic lncRNAs. The prognostic risk score formula was constructed as follows:

$$\text{Risk score} = \sum_{i=1}^{n} \text{coefficients} * \text{Expression of NRLRs}(i)$$

Assessment of the Prognostic Signature Incorporating Necroptosis-Related lncRNAs

LUAD patients in the training set were identified into high-risk group and low-risk group according to the median value of the risk score. Kaplan–Meier survival analysis was performed to evaluate the survival difference between these two groups using the "survival" and "survminer" R packages. The discrimination performance of the NRLSig was also evaluated through the time-dependent receiver operating characteristic (ROC) curve analysis using "timeROC" R package. The consistent formula and cutoff point (the median of risk scores in the training set) were also used to calculate the risk score of each patient in internal validation set and divided into high-risk group and low-risk group. Then, survival analyses and ROC curve analyses were conducted in the validation set and the entire TCGA-LUAD dataset. In addition, principal component analysis (PCA) was performed using "limma" and "scatterplot3d" packages to estimate the clustering ability of prognostic signature. Besides, Kaplan–Meier survival analysis was conducted to examine prognostic significance in each subgroup categorized by clinicopathological features.

Development and Assessment of the Nomogram Containing Necroptosis-Related lncRNA Signature

We further identified whether the risk score generated from the NRLSig and clinicopathological predictors, including age, gender, TNM stage, T stage, N stage, and M stage, were independent prognostic predictors of OS in the entire set through univariate and multivariate Cox regression analysis. Then, we formulated a nomogram based on identified independent variable factors using the "rms" R package. Moreover, the prognostic value of the nomogram was evaluated by the Kaplan–Meier survival analysis based on the high-risk group and low-risk group divided by the median value of the risk score, generated from the nomogram. The discrimination and calibration of the nomogram was estimated in the entire TCGA-LUAD dataset by the ROC curves and calibration curves. Besides, the decision curve analysis (DCA) was utilized to evaluate the clinical usefulness of the model through calculating the net benefits at different threshold probabilities.

Functional Enrichment Analysis and Immune Cell Characteristic Analysis

To explore the biological functions of the selected lncRNAs in NRLSig, we identified the protein-coding genes significantly associated with these lncRNAs from the TCGA dataset through co-expression network analysis. The |Pearson correlation coefficients| > 0.5 and $p < 0.001$ were considered

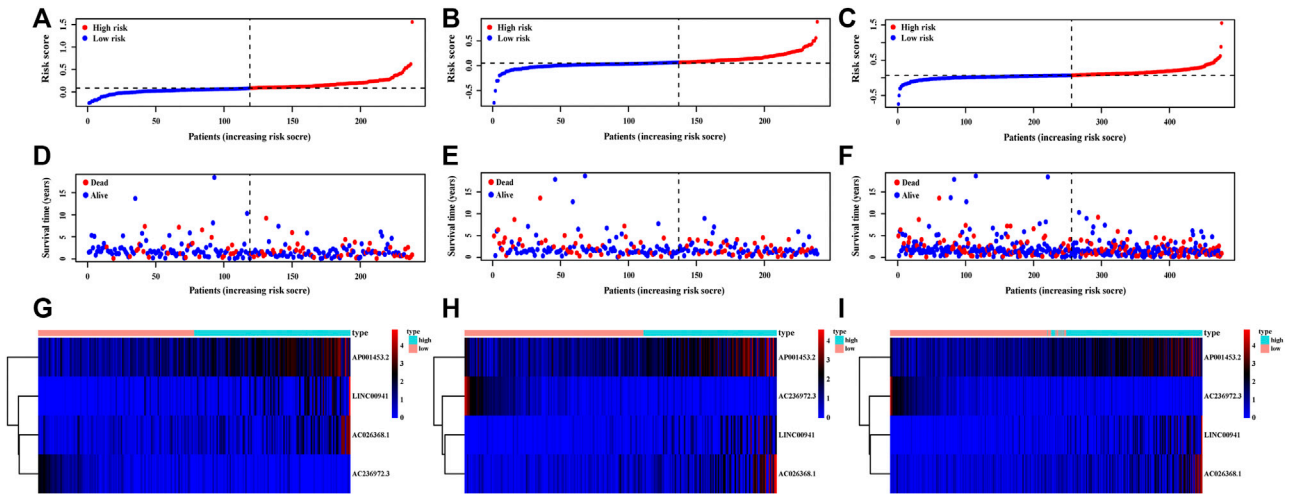

FIGURE 3 | Distribution of LUAD patients stratified by the necroptosis-related lncRNA signature (NRLsig). **(A–C)** The distribution of low-risk group and high-risk group LUAD patients divided by the NRLsig in the training, validation, and entire set, respectively. **(D–F)** Survival statuses of patients in different groups stratified by the NRLsig in the training, validation, and entire set, respectively. **(G–I)** Heatmap of expression statuses of the selected necroptosis-related lncRNAs in the training, validation, and entire set, respectively.

as criteria for significantly correlation. We further performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of the NRGs to investigate the functions of the genes selected above. The functions or pathways with p-Value < 0.05 were regarded as significantly enriched. Functional enrichment analysis was performed using the "clusterProfiler" R package.

The CIBERSORT (Newman et al., 2015; Chen et al., 2018) and TIMER (Li J. et al., 2020; Li T. et al., 2020) algorithms were utilized to analyse the abundances of tumor-infiltrating immune cells among the each LUAD patients in the TCGA dataset. Moreover, the abundances for 22 types of immune cells of the patients in the high-risk group and low-risk group stratified by the NRLSig, including naive $CD4^+$ T cells, resting memory $CD4^+$ T cells, activated memory $CD4^+$ T cells, naive B cells, memory B cells, plasma cells, $CD8^+$ T cells, follicular helper T cells, regulatory T cells, gamma delta T cells, M0 macrophages, M1 macrophages, M2 macrophages, resting natural killer cells, activated natural killer cells, monocytes, resting dendritic cells, activated dendritic cells, resting mast cells, activated mast cells, eosinophils, and neutrophils, were compared and visualized using the CIBERSORT algorithm. In addition, the association between the NRLSig and immune infiltration cells, including B cells, $CD4^+$ T cells, $CD8^+$ T cells, dendritic cells, macrophages, and neutrophils, were also analyzed using the TIMER algorithm.

Tissue Sample Collection and Lung Adenocarcinoma Cell Culture

A total of 12 pairs of LUAD tissues and noncancerous adjacent tissues (NAT) were collected from patients who had undergone surgical resection at the Department of Thoracic Surgery, Peking Union Medical College Hospital (Beijing, China). Written informed consent was obtained from all patients before collection. This study was approved by the Institutional Ethics Review Committee at Peking Union Medical College Hospital and was conducted in accordance with recognized ethical guidelines. All samples were stored at −80°C.

All human LUAD cell lines (A549, H1299, and PC9) and human bronchial epithelial cell line (BEAS-2B) were purchased from the American Type Culture Collection (ATCC). A549 and BEAS-2B cells were cultured in DMEM medium (Gibco). H1299 and PC9 cells were cultured in RPMI 1640 medium (Gibco). All medium was supplemented with 10% fetal bovine serum (BI). All cells were maintained in a humidified incubator with 5% CO_2 at 37°C.

RNA Extraction and qRT-PCR Analysis

Total cellular and tissue RNA was extracted using Trizol reagent (Takara Bio, Japan) following the manufacturer's protocols. Then, RNA samples were reverse transcribed by Hiscript III Reverse Transcriptase kit (Vazyme, Nanjing, China) and corresponding RNA expression was evaluated by qRT-PCR with ChamQ^{TM} Universal SYBR qPCR Master Mix kit (Vazyme). GAPDH acted as the internal reference for normalization. The detailed sequence of primers used were listed in **Supplementary Table S2**.

Statistical Analysis

All statistical analyses were conducted using the R software, version 4.0.2 (https://www.r-project.org). Pearson correlation analysis was used to analyze the correlation between NRGs and NRLRs. Differences in the proportions of clinical characteristics, such as age, gender, and T stage, were analyzed by the chi-squared test. The Mann-Whitney U test was implemented to compare the expression of genes or lncRNA, and abundance of tumor-infiltrating immune cells. Univariate Cox regression analysis and LASSO regression analysis or multivariate Cox regression were conducted to define the

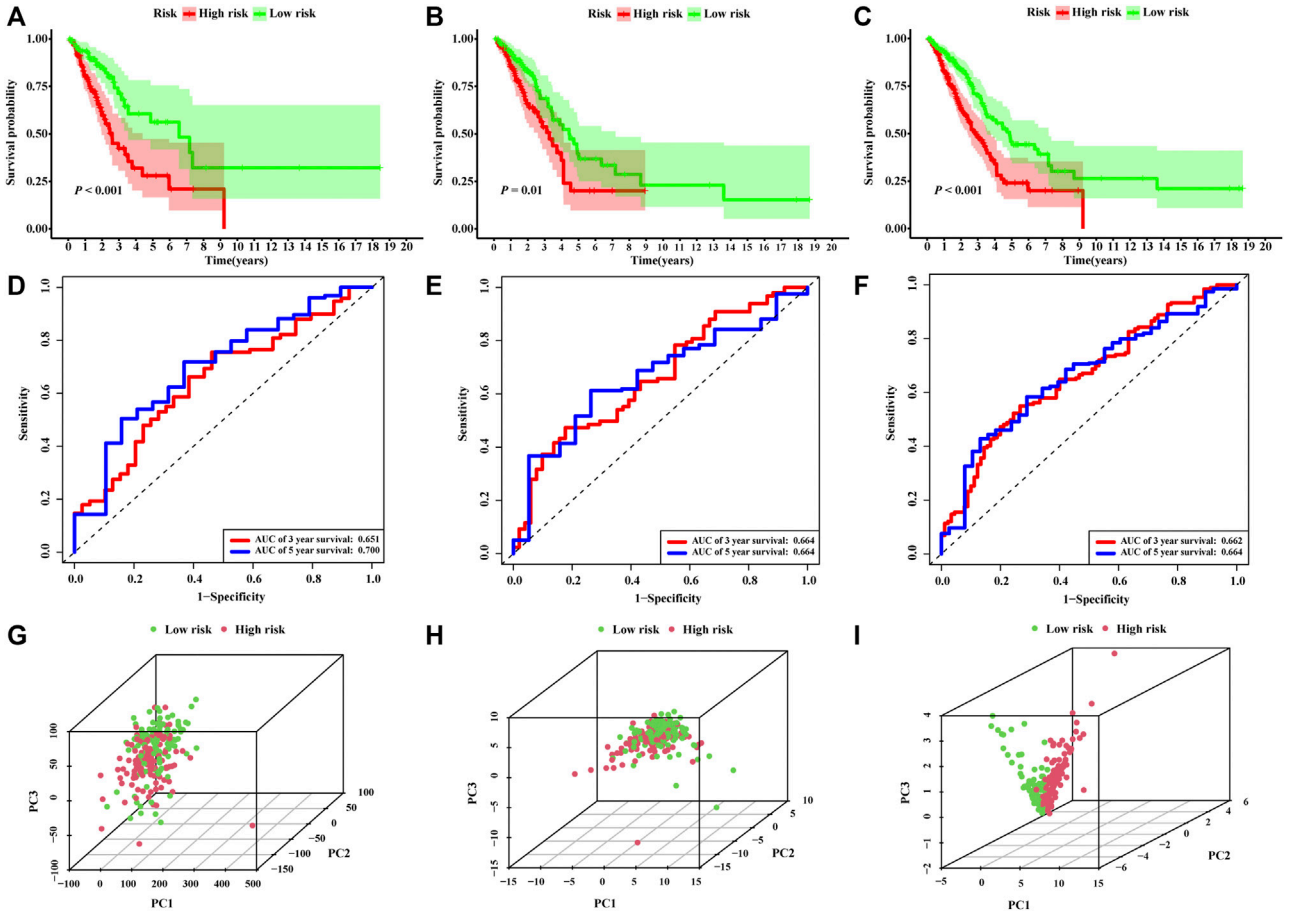

FIGURE 4 | Assessments of the necroptosis-related lncRNA signature (NRLsig). **(A–C)** Kaplan–Meier survival analysis curves of the high-risk group and low-risk group stratified by NRLsig in the training, validation, and entire TCGA-LUAD set, respectively. **(D–F)** Time-dependent ROC curves at 3- and 5-year follow-up in the training, validation and entire TCGA-LUAD set, respectively. Principal component analysis (PCA) of the low-risk group and high-risk group stratified by the whole-genome **(G)**, necroptosis-related genes **(H)**, and the NRLsig **(I)**.

optimal prognostic factor for OS. The OS between high-risk group and low-risk group was compared using the Kaplan-Meier analysis with the log-rank test. All statistical tests were two-tailed, and $p < 0.05$ was considered statistically significant.

RESULTS

Identification of Necroptosis-Related lncRNAs in Lung Adenocarcinoma Patients

The flow chart for the risk signature development and subsequent analyses is illustrated in **Figure 1**. A total of 535 LUAD patients with RNA sequencing data were included in present study. Among 14,128 lncRNAs identified, 696 differentially expressed lncRNAs were significant between tumor and adjacent normal tissues (**Figures 2A,B**). According to these lncRNAs and 204 NRGs, NRLRs were identified through Pearson correlation analysis (|Correlation Coefficient| > 0.3 and $p < 0.001$). Finally, 88 NRLRs were selected for subsequent analyses (**Supplementary Figure S1**).

Construction of the Prognostic Necroptosis-Related lncRNA Signature for Lung Adenocarcinoma Patients

To develop the NRLSig for predicting the survival status of LUAD patients, a total of 477 patients, who meet the inclusion and exclusion criteria, were randomly grouped into a training set (238 patients) and a validation set (239 patients) in a 1:1 ratio. The baseline characteristics of the entire TCGA-LUAD patients are summarized in **Table 1**. Based on the transcription profile of NRLRs, 7 NRLRs were found associated with the OS of LUAD patients in univariate Cox proportional hazards regression analysis. The lncRNAs with p-Value < 0.05 were selected for LASSO regression analysis to further identify optimal prognostic lncRNAs. Finally, a 4-lncRNA signature was constructed based on the optimal value of λ (**Figures 2C,D**). According to the coefficient values, the formula of the NRLsig was presented as follows: $risk\,score = (0.0286 \times LINC00941) + (0.0226 \times AP001453.2) + (0.0328 \times AC026368.1) + (-0.0499 \times AC236972.3)$. Besides, in the Sankey diagram, we identified 22 NRGs were positively

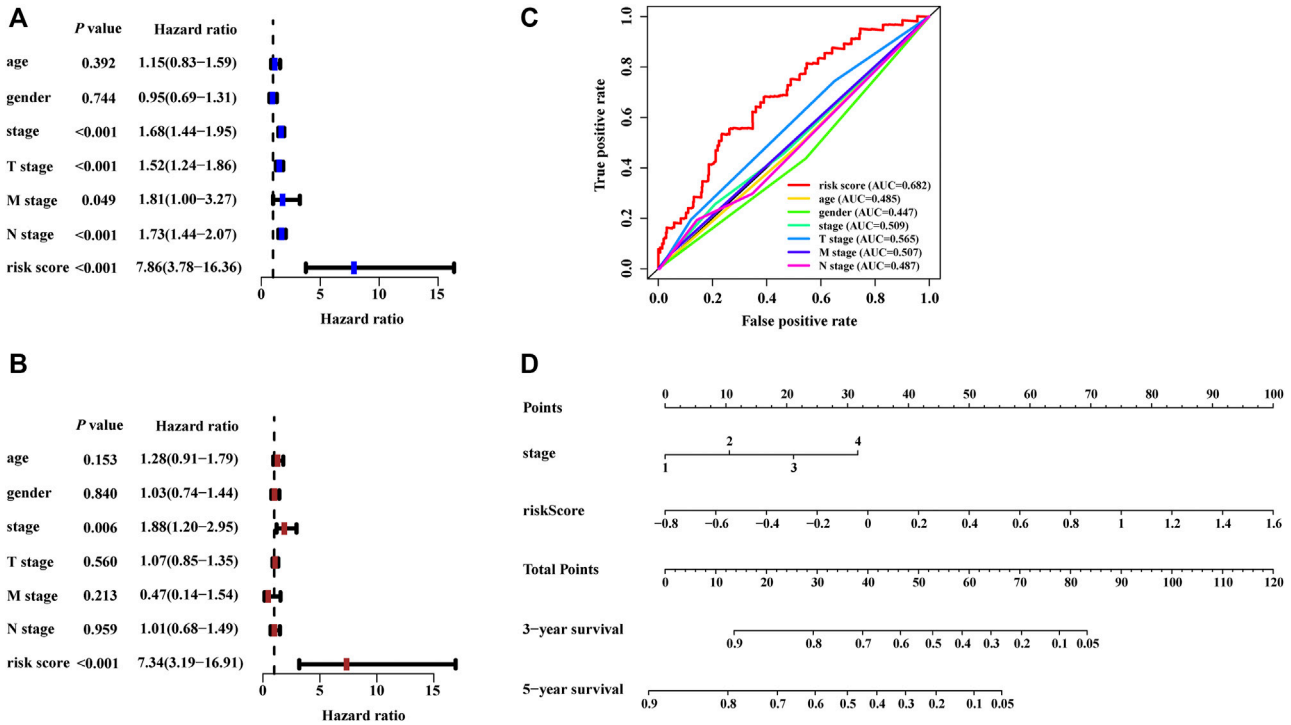

FIGURE 5 | Development of the nomogram based on the risk scores calculated from the signature. **(A,B)** The predictors extracted by univariate and multivariate Cox analyses. **(C)** Time-dependent ROC curves for predicting overall survival at 1 year by risk score, age, gender, stage, T stage, M stage, and N stage. **(D)** A prognostic nomogram was developed to predict the 3- and 5-year survival for LUAD patients.

correlated to the selected NRLRs, while STAT5B, CHMP6 and STUB1 were negatively associated with the lncRNAs (**Figure 2E**).

The risk score of each patient was calculated based on the formula in the training set, validation set, and entire set, and the median of risk scores, as the determined cutoff value, were used to classify patients into a low-risk group or high-risk group (**Figures 3A–C**). The distribution of survival status in each set was plotted in **Figures 3D–F**. These figures illustrated that increasing risk score was positively associated with accumulating number of patients with poor prognoses. The expression levels of the lncRNAs selected in the signature were also showed in **Figures 3G–I**. Survival analyses illustrated that the patients in high-risk group possessed significantly lower survival rate compared to patients in the low-risk group in all three sets ($p < 0.001$, $p = 0.01$ and $p < 0.001$, respectively, **Figures 4A–C**). The AUC of the NRLSig at 3- and 5-year also showed a good discriminative performance in the training set, validation set, and entire set (**Figures 4D–F**). As depicted in **Figures 4G,H**, the high-risk group and low-risk group could not be effectively identified using the whole genome or necroptosis-related genes; however, LUAD patients could be clearly classified into high-risk or low-risk group using NRLSig (**Figure 4I**), further supporting the performance of the lncRNA signature. Survival analysis in subgroups was also conducted and showed significant differences in prognosis among the low-risk group and high-risk group, except for TNM stage III-IV and M1 stage patients, which suggested that the prognostic signature was applicable to different subtypes of LUAD patients (**Supplementary Figures S2A–L**). All these assessments indicated that NRLSig is a reliable independent prognostic risk factor for patients with LUAD.

Development and Performance Assessment of the Nomogram Incorporating the Necroptosis-Related lncRNA Signature

The risk score calculated from the NRLSig and several clinical candidate factors were evaluated by the univariate and multivariate Cox regression algorithm in the entire LUAD set. Univariate Cox regression analysis revealed that the risk score of the signature was correlated with the OS of LUAD patients ($p < 0.001$, **Figure 5A**). Multivariate Cox regression analysis further demonstrated that the risk signature was an independent prognostic factor for predicting the OS of LUAD patients ($p < 0.001$, **Figure 5B**). We also performed time-dependent ROC curves of 1-year OS, and the AUC value for the risk score generated from the NRLSig was 0.682, which was higher than other clinical predictors, further supporting the discriminative power of NRLSig for predicting survival status in LUAD (**Figure 5C**). All variables which were significant ($p < 0.05$) in the multivariate Cox regression analysis were included in the predictive model. Finally, a

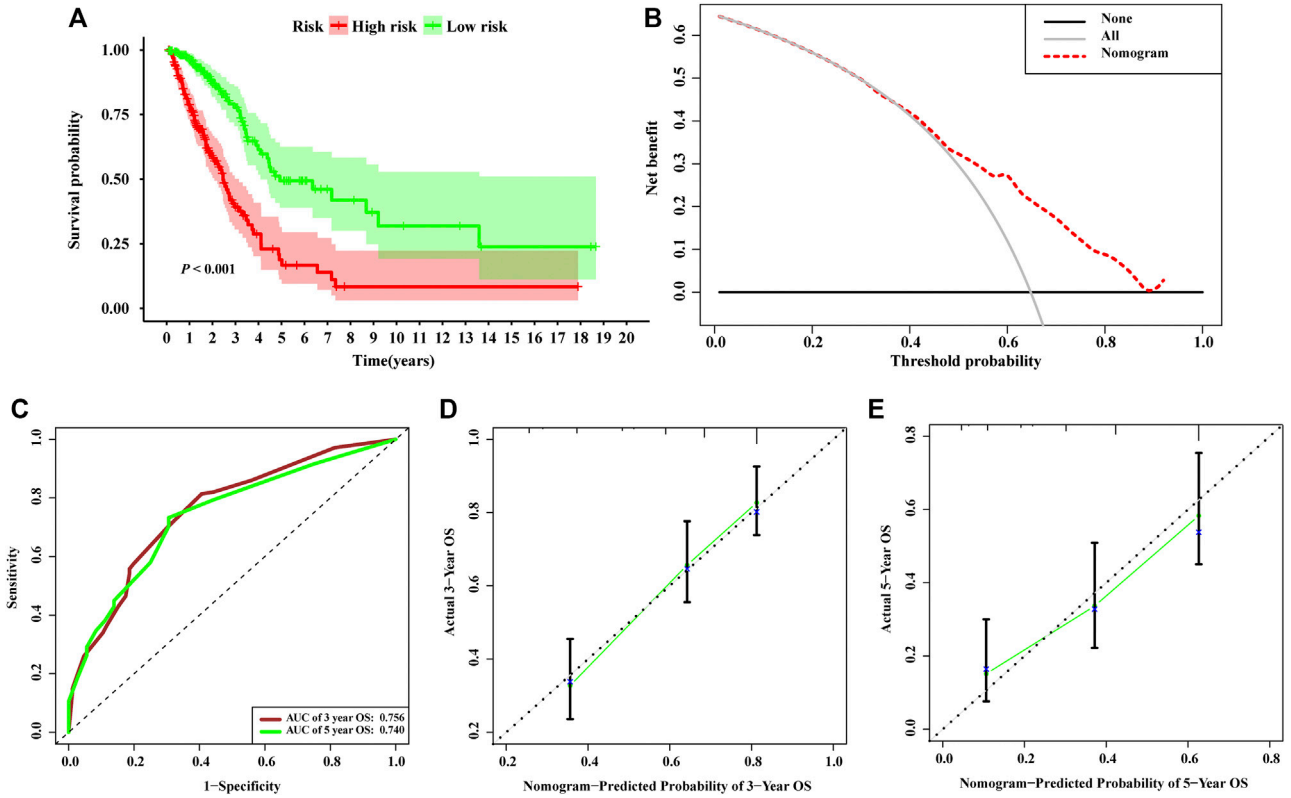

FIGURE 6 | Assessment of the nomogram incorporating the risk scores generated from the signature. **(A)** Kaplan–Meier survival analysis curves of the high-risk group and low-risk group divided by the nomogram in TCGA-LUAD set. **(B)** Decision curve analysis evaluating the clinical usefulness of the nomogram for LUAD patients **(C)** Time-dependent ROC curves at 3- and 5-year follow-up. The calibration curves show the prediction of the nomogram for 3-year **(D)** and 5-year **(E)**.

nomogram to predict the 3- and 5-year OS was constructed incorporating the risk score generated from the NRLSig and the TNM stage (**Figure 5D**).

Based on the risk scores calculated by the nomogram, LUAD patients in the entire set were classified into different risk groups by the median value of the risk score. **Figure 6A** shows that low-risk group patients possessed significantly better prognoses than those in the high-risk group ($p < 0.001$). The result of DCA demonstrated that majority of the red dashed curve was in the area above the gray and the black solid lines, illustrating a higher net benefits could be acquired by using the nomogram to make decision (**Figure 6B**). In addition, the ROC analyses showed satisfactory discrimination performance of the model with an AUC of 0.756, and 0.740 at 3- and 5-year follow-up (**Figure 6C**). Further, a good agreement between the nomogram prediction and actual observation was illustrated *via* calibration curves (**Figures 6D,E**). Consequently, promising predictive value was revealed for this prognostic integrated nomogram.

Functional Enrichment Analysis and Immune Infiltration Analysis

To investigate the potential biological functions and the immune infiltration status associated with NRLRs, we performed co-expression analysis to screen out the NRLRs-related protein-coding genes. Only LINC00941 has been investigated in previous studies (Wang et al., 2019; Wu et al., 2021). Therefore, the functional enrichment analysis focused on this necroptosis-related lncRNA, LINC00941. |Pearson correlation coefficients| > 0.5 and $p < 0.001$ as the criteria selected 12 protein-coding genes from the RNA transcriptome data of the TCGA-LUAD. Among these protein-coding genes, the expression of these genes was positively associated with the expression of LINC00941, except for TMEM125 and NKX2-1 (**Supplementary Figure S3A**). As shown in **Figure 7A**, the GO functional enrichment analysis demonstrated that the correlated genes were mainly clustered in several biological processes or molecular functions such as extracellular matrix organization, extracellular structure organization, transcription regulator complex, and intronic transcription regulatory region sequence-specific DNA binding. At the same time, KEGG terms of correlated genes were significantly enriched in several signaling pathway such as extracellular matrix-receptor (ECM-receptor) interaction, toxoplasmosis, and focal adhesion pathway (**Figure 7B**). Altogether, these analyses suggested that the NRLRs-related protein-coding genes may be mainly correlated with tumor migration and metastasis in LUAD.

The abundances of tumor-infiltrating immune cells were estimated by TIMER and CIBERSORT algorithms. The results generated from the TIMER algorithm demonstrated that the six

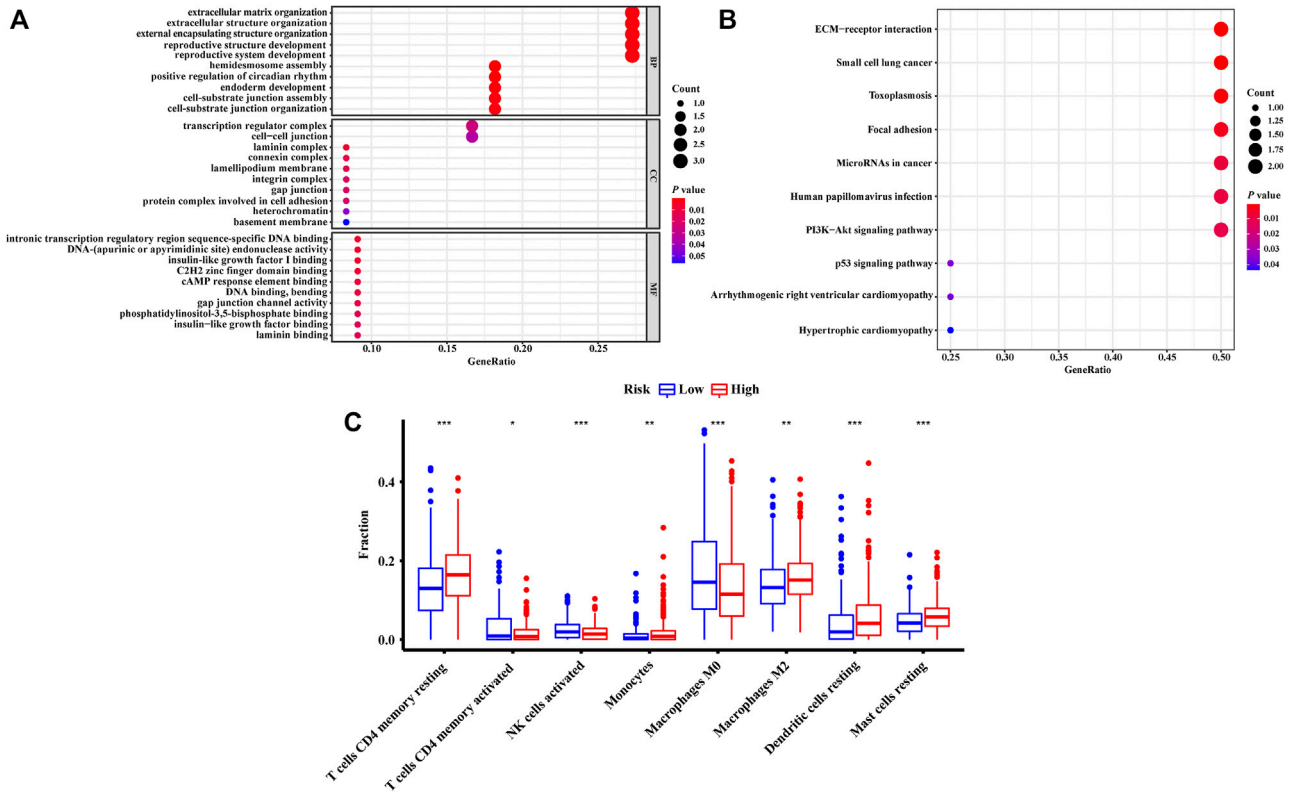

FIGURE 7 | Functional enrichment analysis and the characteristics of immune infiltration of different groups defined by the signature. The GO terms **(A)** and KEGG terms **(B)** enriched by the co-expressed genes related to necroptosis-related lncRNA. **(C)** Differences in abundance of tumor-infiltrating immune cells among two risk groups stratified by the necroptosis-related lncRNA signature.

kinds of immune cell infiltration were negatively correlated with the risk score calculated by the NRLSig, though only B cells and dendritic cells showed significant association with the prognosis of LUAD patients (**Supplementary Figures S4A–F**). Moreover, the boxplots from the CIBERSORT algorithm showed that abundances of resting memory $CD4^+$ T cells, monocytes, M2 macrophages, resting dendritic cells, and resting mast cells were markedly enriched in the high-risk group compared to the low-risk group. On the contrary, the abundances of activated memory $CD4^+$ T cells, activated natural killer (NK) cells, and M0 macrophages in the high-risk group were significantly lower than in the low-risk group (**Figure 7C**). In summary, the association between the risk scores generated from the NRLSig and tumor-infiltrating immune cells were assessed, and the results demonstrated that the risk level of LUAD patients was related to the distribution difference of immune infiltration cells.

Validation of Necroptosis-Related lncRNAs Expression in Cell Lines and Tissue Samples

The expression levels of selected NRLRs were further evaluated and validated in cell lines and tissues. As illustrated in **Figure 8A**, the expression levels of LINC00941, AP001453.2, and AC026368.1 were significantly higher in LUAD cell lines, including A549, H1299, and PC9, than those in human normal lung epithelial cells (BEAS-2B), while AC236972.3 exhibited the opposite trend. We also evaluated the expression level of these 4 lncRNAs in 12 pairs of LUAD tissues and NAT. Consistent expression trends were observed in these tissue samples. LINC00941, AP001453.2, and AC026368.1 showed higher expression levels in LUAD tissues than in NAT, but the expression of AC236972.3 was significantly lower in tumor tissues than in NAT (**Figures 8B–E**). These results further confirmed the correctness of the above bioinformatics analyses.

DISCUSSION

In recent years, with the development of next-generation sequencing, accumulating non-coding RNAs and protein-coding genes have been identified as prognostic predictor for cancer patients (Borad et al., 2016; Lagana et al., 2016; Li N. et al., 2020). In current clinical practice, the traditional staging system may not be optimal for individualized prognostic prediction for LUAD patients (Yao et al., 2021). Thus, it is urgently needed to investigate biomarkers related to tumor diagnosis and prognosis. LncRNAs, a kind of non-protein-coding RNAs, are widely expressed in different tissues and participate in various kinds of biological processes in malignant tumors. Necroptosis, a novel

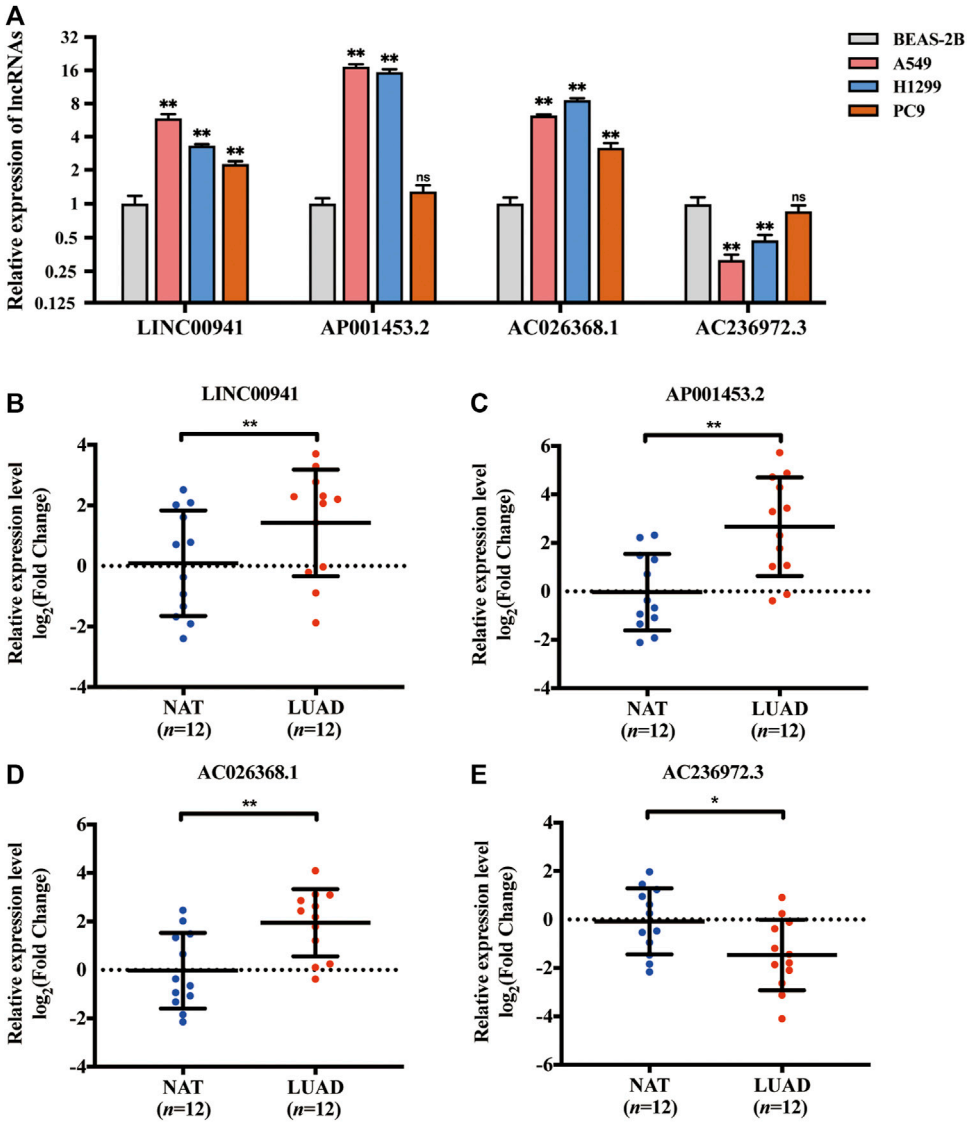

FIGURE 8 | Validation of the expression of the selected necroptosis-related lncRNA in cell lines and tissues. **(A)** The relative expression of 4 necroptosis-related lncRNAs in lung adenocarcinoma cell lines (A549, H1299, and PC9) with human bronchial epithelial cell line (BEAS-2B). **(B–E)** The relative expression of LINC00941, AP001453.2, AC026368.1, and AC236972.3 in 12 pairs of lung adenocarcinoma tissue samples. *$p < 0.05$; **$p < 0.01$.

form of regulated cell death, possesses a mechanistic resemblance to apoptosis and a morphological resemblance to necrosis. Several potential lncRNAs have been identified as the regulators for necroptosis. Therefore, NRLRs have also attracted plenty of attention for promising prognostic value in LUAD.

To the best of our knowledge, this is the first study to identify and comprehensively analyze prognostic NRLRs in LUAD. A signature based on 4 NRLRs and a predictive model incorporating this signature were developed in the present study, and this nomogram showed higher discriminatory accuracy for predicting OS of LUAD patients compared to models constructed in previous studies (Liu and Yang, 2021; Yao et al., 2021). Additionally, we also investigate the enriched biological functions and immune infiltration status related to NRLRs in LUAD cohort.

According to the Sankey diagram, we identified 4 lncRNAs that were related to 25 NRGs. Among these NRGs, dynamin 1-like (DNM1L), a regulator of necroptosis by activating mitochondrial fission, was correlated with LINC00941 and AC026368.1 (Remijsen et al., 2014). It also suggests a potential pivotal role in tumorigenesis and progression of NSCLC (Furukawa et al., 2005). Peptidylprolyl isomerase A (PPIA), associated with AP001453.2, is an intracellular protein released early in the process of necroptosis and has been identified to be a biomarker for this form of cell death (Cabello et al., 2021). Sirtuin 3 (SIRT3), which is related to AC236972.3, can inhibit the proliferation of human small-cell lung cancer cells by promoting apoptosis and necroptosis (Tang X. et al., 2020). Moreover,

LINC00941 was reported that its overexpression could accelerate tumor progression in NSCLC *via* miR-877-3p/VEGFA axis (Ren et al., 2021). The biological function or mechanism of other lncRNAs are still unclear. The understanding of these newly identified lncRNAs needs further mechanistic study.

To explore the potential functions or mechanisms of the lncRNAs in NRLSig, co-expression network analysis and functional enrichment analysis were conducted. The GO enrichment analysis illustrated that the co-expressed genes significantly enriched in several functions. First, a large amount of co-expressed genes was associated with organization of extracellular matrix (ECM). Genetic and epigenetic changes in lung cancer may lead to the conversion of ECM, such as misexpression of collagens, proteases and integrins in the tumor microenvironment, which could consequently cause tumor progression (Götte and Kovalszky, 2018; Paolillo and Schinelli, 2019). In addition, NK2 homeobox 1 (NKX2-1) and aryl hydrocarbon receptor nuclear translocator-like (ARNTL2) both act as a transcription regulator. Loss of NKX2-1 could lead to the recruitment of tumor-associated neutrophils which promote the proliferation of lung squamous cell (Mollaoglu et al., 2018), and high expression of ARNTL2, which could drive metastatic self-sufficiency and predict poor prognosis for LUAD patients (Brady et al., 2016). Moreover, integrin subunit alpha 6 (ITGA6) could regulate lung differentiation in stress response by mediating cell adhesions to laminin (Sanchez-Esteban et al., 2006). Furthermore, several KEGG terms were also enriched. Laminin subunit gamma 2 (LAMC2), enriched in most of the KEGG signaling pathways in present study, was found to promote tumor proliferation, metastasis, and vascular regeneration through ECM-receptor interaction and focal adhesion (Wang et al., 2020).

We found that the abundances of activated memory $CD4^+$ T cells and activated NK cells were significantly lower in the high-risk group compared to the low-risk group, while M2 macrophages were enriched in the high-risk group. This phenomenon suggested that the high-risk group patients may possess deteriorated immune status and immune function. A previous study revealed that interleukin 12 (IL-12) could promote the proliferation and tumor suppression of memory $CD4^+$ T cells presenting in the tumor microenvironment (TME) of lung cancer (Broderick et al., 2005). NK cells, an effecter lymphocyte of the innate immune system, could control tumor proliferation and metastatic spread (Sivori et al., 2021). Further, since M2 macrophages could secrete a series of anti-inflammatory molecules to function as pro-tumoral factors, high abundance of tumor-infiltrating M2 macrophages was associated with unfavorable prognosis of NSCLC patients (Jackute et al., 2018). In general, the dysregulation of the immune status of TME may lead to a discrepancy in survival prognosis among the high-risk group and low-risk group stratified by the prognostic NRLSig.

Several limitations in our study still need to be considered, though we applied many methods to adjust and validate our signature. First, as this was a retrospective study based on public databases, some information related to lung cancer may be unavailable, such as smoking status. Second, we used a single data source in this study. Though we applied internal validation to test our findings, whether the performance of the model in an external cohort would be similarly satisfactory still require further validation. In this study, the RNA expression data and survival information of 163 LUAD patients had been retrieved from the GSE3141 series and GSE37745 series matrices from Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). However, we could not acquire sufficient information of lncRNAs in the GEO cohort because the scale of commercial sequencing data from the GEO dataset was much smaller compared to the size of RNA sequencing data from the TCGA dataset. Third, the *in vitro* and *in vivo* experiments will be required to further elucidate the biological mechanism or prognostic value of NRLRs in LUAD.

In conclusion, we proposed a signature, constructed based on 4 lncRNAs biomarkers, that could independently predict the prognosis of LUAD patients. Moreover, the possible biological functions and immune status of the 4 NRLRs could provide novel insights for further research on the molecular mechanisms of tumorigenesis and progression of LUAD. In all, the NRLRs identified in this study may offer promising therapeutic targets or prognostic predictors for LUAD patients.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: TCGA-LUAD, The Cancer Genome Altas (https://portal.gdc.cancer.gov).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Ethics Review Committee at Peking Union Medical College Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SL had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: SL and XD. Acquisition and analysis of data: XD and CG. Drafting the manuscript: XD. Proofread the manuscript for important intellectual content: all authors. Approval of the final version manuscript: all authors. Accountable for all aspects of the work: all authors.

FUNDING

This research was supported by the CAMS Innovation Fund for Medical Sciences (Grant Number 2021-I2M-1-022).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.833362/full#supplementary-material

REFERENCES

Agostini, M., Ganini, C., Candi, E., and Melino, G. (2020). The Role of Noncoding RNAs in Epithelial Cancer. *Cell Death Discov.* 6, 13. doi:10.1038/s41420-020-0247-6

Borad, M. J., Egan, J. B., Condjella, R. M., Liang, W. S., Fonseca, R., Ritacca, N. R., et al. (2016). Clinical Implementation of Integrated Genomic Profiling in Patients with Advanced Cancers. *Sci. Rep.* 6 (1), 25. doi:10.1038/s41598-016-0021-4

Brady, J. J., Chuang, C.-H., Greenside, P. G., Rogers, Z. N., Murray, C. W., Caswell, D. R., et al. (2016). An Arntl2-Driven Secretome Enables Lung Adenocarcinoma Metastatic Self-Sufficiency. *Cancer Cell* 29 (5), 697–710. doi:10.1016/j.ccell.2016.03.003

Broderick, L., Yokota, S. J., Reineke, J., Mathiowitz, E., Stewart, C. C., Barcos, M., et al. (2005). Human CD4+ Effector Memory T Cells Persisting in the Microenvironment of Lung Cancer Xenografts Are Activated by Local Delivery of IL-12 to Proliferate, Produce IFN-γ, and Eradicate Tumor Cells. *J. Immunol.* 174 (2), 898–906. doi:10.4049/jimmunol.174.2.898

Cabello, R., Fontecha-Barriuso, M., Martin-Sanchez, D., Lopez-Diaz, A. M., Carrasco, S., Mahillo, I., et al. (2021). Urinary Cyclophilin A as Marker of Tubular Cell Death and Kidney Injury. *Biomedicines* 9 (2), 217. doi:10.3390/biomedicines9020217

Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M., and Alizadeh, A. A. (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol.* 1711, 243–259. doi:10.1007/978-1-4939-7493-1_12

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE V7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression. *Genome Res.* 22 (9), 1775–1789. doi:10.1101/gr.132159.111

Furukawa, C., Daigo, Y., Ishikawa, N., Kato, T., Ito, T., Tsuchiya, E., et al. (2005). Plakophilin 3 Oncogene as Prognostic Marker and Therapeutic Target for Lung Cancer. *Cancer Res.* 65 (16), 7102–7110. doi:10.1158/0008-5472.Can-04-1877

Galluzzi, L., Vitale, I., Aaronson, S. A., Abrams, J. M., Adam, D., Agostinis, P., et al. (2018). Molecular Mechanisms of Cell Death: Recommendations of the Nomenclature Committee on Cell Death 2018. *Cell Death Differ.* 25 (3), 486–541. doi:10.1038/s41418-017-0012-4

Gibb, E. A., Brown, C. J., and Lam, W. L. (2011). The Functional Role of Long Non-coding RNA in Human Carcinomas. *Mol. Cancer* 10, 38. doi:10.1186/1476-4598-10-38

Glickman, M. E., Rao, S. R., and Schultz, M. R. (2014). False Discovery Rate Control Is a Recommended Alternative to Bonferroni-type Adjustments in Health Studies. *J. Clin. Epidemiol.* 67 (8), 850–857. doi:10.1016/j.jclinepi.2014.03.012

Götte, M., and Kovalszky, I. (2018). Extracellular Matrix Functions in Lung Cancer. *Matrix Biol.* 73, 105–121. doi:10.1016/j.matbio.2018.02.018

Gong, Y., Fan, Z., Luo, G., Yang, C., Huang, Q., Fan, K., et al. (2019). The Role of Necroptosis in Cancer Biology and Therapy. *Mol. Cancer* 18 (1), 100. doi:10.1186/s12943-019-1029-8

Harari-Steinfeld, R., Gefen, M., Simerzin, A., Zorde-Khvalevsky, E., Rivkin, M., Ella, E., et al. (2021). The lncRNA H19-Derived MicroRNA-675 Promotes Liver Necroptosis by Targeting FADD. *Cancers* 13 (3), 411. doi:10.3390/cancers13030411

He, W., Zhong, G., Jiang, N., Wang, B., Fan, X., Chen, C., et al. (2018). Long Noncoding RNA BLACAT2 Promotes Bladder Cancer-Associated Lymphangiogenesis and Lymphatic Metastasis. *J. Clin. Invest.* 128 (2), 861–875. doi:10.1172/jci96218

Huang, C.-Y., Kuo, W.-T., Huang, Y.-C., Lee, T.-C., and Yu, L. C. H. (2013). Resistance to Hypoxia-Induced Necroptosis Is Conferred by Glycolytic Pyruvate Scavenging of Mitochondrial Superoxide in Colorectal Cancer Cells. *Cell Death Dis.* 4 (5), e622. doi:10.1038/cddis.2013.149

Hung, C.-L., Wang, L.-Y., Yu, Y.-L., Chen, H.-W., Srivastava, S., Petrovics, G., et al. (2014). A Long Noncoding RNA Connects C-Myc to Tumor Metabolism. *Proc. Natl. Acad. Sci. U.S.A.* 111 (52), 18697–18702. doi:10.1073/pnas.1415669112

Jackute, J., Zemaitis, M., Pranys, D., Sitkauskiene, B., Miliauskas, S., Vaitkiene, S., et al. (2018). Distribution of M1 and M2 Macrophages in Tumor Islets and Stroma in Relation to Prognosis of Non-small Cell Lung Cancer. *BMC Immunol.* 19 (1), 3. doi:10.1186/s12865-018-0241-4

Johnstone, R. W., Ruefli, A. A., and Lowe, S. W. (2002). Apoptosis. *Cell* 108 (2), 153–164. doi:10.1016/s0092-8674(02)00625-6

Kim, T., Cui, R., Jeon, Y.-J., Lee, J.-H., Lee, J. H., Sim, H., et al. (2014). Long-range Interaction and Correlation between MYC Enhancer and Oncogenic Long Noncoding RNA CARLo-5. *Proc. Natl. Acad. Sci. U.S.A.* 111 (11), 4173–4178. doi:10.1073/pnas.1400350111

Lagana, A., Perumal, D., Melnekoff, D., Readhead, B., Kidd, B., Leshchenko, V. V., et al. (2016). Integrative Network Analysis of Newly Diagnosed Multiple Myeloma Identifies a Novel RNA-Seq Based High Riskgene Signature. *Blood* 128 (22), 3285. doi:10.1182/blood.V128.22.3285.3285

Lalaoui, N., Lindqvist, L. M., Sandow, J. J., and Ekert, P. G. (2015). The Molecular Relationships between Apoptosis, Autophagy and Necroptosis. *Seminars Cell & Dev. Biol.* 39, 63–69. doi:10.1016/j.semcdb.2015.02.003

Li, J., Cao, F., Yin, H.-l., Huang, Z.-j., Lin, Z.-t., Mao, N., et al. (2020a). Ferroptosis: Past, Present and Future. *Cell Death Dis.* 11 (2), 88. doi:10.1038/s41419-020-2298-2

Li, N., Yuan, J., Tian, W., Meng, L., and Liu, Y. (2020b). T-cell Receptor Repertoire Analysis for the Diagnosis and Treatment of Solid Tumor: A Methodology and Clinical Applications. *Cancer Commun.* 40 (10), 473–483. doi:10.1002/cac2.12074

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020c). TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells. *Nucleic Acids Res.* 48 (W1), W509–w514. doi:10.1093/nar/gkaa407

Linkermann, A., and Green, D. R. (2014). Necroptosis. *N. Engl. J. Med.* 370 (5), 455–465. doi:10.1056/NEJMra1310050

Liu, B., and Yang, S. (2021). A Five Autophagy-Related Long Non-coding RNA Prognostic Model for Patients with Lung Adenocarcinoma. *Ijgm* Vol. 14, 7145–7158. doi:10.2147/ijgm.S334601

Md Yusof, K., Rosli, R., Abdullah, M., and A. Avery-Kiejda, K. (2020). The Roles of Non-coding RNAs in Tumor-Associated Lymphangiogenesis. *Cancers* 12 (11), 3290. doi:10.3390/cancers12113290

Meza, R., Meernik, C., Jeon, J., and Cote, M. L. (2015). Lung Cancer Incidence Trends by Gender, Race and Histology in the United States, 1973-2010. *PLoS One* 10 (3), e0121323. doi:10.1371/journal.pone.0121323

Miller, V. A., Hirsh, V., Cadranel, J., Chen, Y.-M., Park, K., Kim, S.-W., et al. (2012). Afatinib versus Placebo for Patients with Advanced, Metastatic Non-small-cell Lung Cancer after Failure of Erlotinib, Gefitinib, or Both, and One or Two Lines of Chemotherapy (LUX-Lung 1): a Phase 2b/3 Randomised Trial. *Lancet Oncol.* 13 (5), 528–538. doi:10.1016/s1470-2045(12)70087-6

Mollaoglu, G., Jones, A., Wait, S. J., Mukhopadhyay, A., Jeong, S., Arya, R., et al. (2018). The Lineage-Defining Transcription Factors SOX2 and NKX2-1 Determine Lung Cancer Cell Fate and Shape the Tumor Immune Microenvironment. *Immunity* 49 (4), 764–779. e769. doi:10.1016/j.immuni.2018.09.020

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Pan, G., Liu, Y., Shang, L., Zhou, F., and Yang, S. (2021). EMT-associated microRNAs and Their Roles in Cancer Stemness and Drug Resistance. *Cancer Commun.* 41 (3), 199–217. doi:10.1002/cac2.12138

Paolillo, M., and Schinelli, S. (2019). Extracellular Matrix Alterations in Metastatic Processes. *Ijms* 20 (19), 4947. doi:10.3390/ijms20194947

Prensner, J. R., Zhao, S., Erho, N., Schipper, M., Iyer, M. K., Dhanasekaran, S. M., et al. (2014). RNA Biomarkers Associated with Metastatic Progression in Prostate Cancer: a Multi-Institutional High-Throughput Analysis of SChLAP1. *Lancet Oncol.* 15 (13), 1469–1480. doi:10.1016/s1470-2045(14)71113-1

Remijsen, Q., Goossens, V., Grootjans, S., Van den Haute, C., Vanlangenakker, N., Dondelinger, Y., et al. (2014). Depletion of RIPK3 or MLKL Blocks TNF-Driven Necroptosis and Switches towards a Delayed RIPK1 Kinase-dependent Apoptosis. *Cell Death Dis.* 5 (1), e1004. doi:10.1038/cddis.2013.531

Ren, M.-H., Chen, S., Wang, L.-G., Rui, W.-X., and Li, P. (2021). LINC00941 Promotes Progression of Non-small Cell Lung Cancer by Sponging miR-877-3p to Regulate VEGFA Expression. *Front. Oncol.* 11, 650037. doi:10.3389/fonc.2021.650037

Sanchez-Esteban, J., Wang, Y., Filardo, E. J., Rubin, L. P., and Ingber, D. E. (2006). Integrins β1, α6, and α3contribute to Mechanical Strain-Induced Differentiation of Fetal Lung Type II Epithelial Cells via Distinct

Mechanisms. *Am. J. Physiology-Lung Cell. Mol. Physiology* 290 (2), L343–L350. doi:10.1152/ajplung.00189.2005

Seehawer, M., Heinzmann, F., D'Artista, L., Harbig, J., Roux, P.-F., Hoenicke, L., et al. (2018). Necroptosis Microenvironment Directs Lineage Commitment in Liver Cancer. *Nature* 562 (7725), 69–75. doi:10.1038/s41586-018-0519-y

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

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Soft.* 39 (5), 1–13. doi:10.18637/jss.v039.i05

Sivori, S., Pende, D., Quatrini, L., Pietra, G., Della Chiesa, M., Vacca, P., et al. (2021). NK Cells and ILCs in Tumor Immunotherapy. *Mol. Aspects Med.* 80, 100870. doi:10.1016/j.mam.2020.100870

Stoll, G., Ma, Y., Yang, H., Kepp, O., Zitvogel, L., and Kroemer, G. (2017). Pro-necrotic Molecules Impact Local Immunosurveillance in Human Breast Cancer. *Oncoimmunology* 6 (4), e1299302. doi:10.1080/2162402x.2017.1299302

Tan, Y. T., Lin, J. F., Li, T., Li, J. J., Xu, R. H., and Ju, H. Q. (2021). LncRNA-mediated Posttranslational Modifications and Reprogramming of Energy Metabolism in Cancer. *Cancer Commun.* 41 (2), 109–120. doi:10.1002/cac2.12108

Tang, R., Xu, J., Zhang, B., Liu, J., Liang, C., Hua, J., et al. (2020a). Ferroptosis, Necroptosis, and Pyroptosis in Anticancer Immunity. *J. Hematol. Oncol.* 13 (1), 110. doi:10.1186/s13045-020-00946-7

Tang, X., Li, Y., Liu, L., Guo, R., Zhang, P., Zhang, Y., et al. (2020b). Sirtuin 3 Induces Apoptosis and Necroptosis by Regulating Mutant P53 Expression in Small-cell L-ung C-ancer. *Oncol. Rep.* 43 (2), 591–600. doi:10.3892/or.2019.7439

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Stat. Med.* 16 (4), 385–395. doi:10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co

Wang, L., Zhao, H., Xu, Y., Li, J., Deng, C., Deng, Y., et al. (2019). Systematic Identification of lincRNA-based Prognostic Biomarkers by Integrating lincRNA Expression and Copy Number Variation in Lung Adenocarcinoma. *Int. J. Cancer* 144 (7), 1723–1734. doi:10.1002/ijc.31865

Wang, Y., Shi, M., Yang, N., Zhou, X., and Xu, L. (2020). GPR115 Contributes to Lung Adenocarcinoma Metastasis Associated with LAMC2 and Predicts a Poor Prognosis. *Front. Oncol.* 10, 577530. doi:10.3389/fonc.2020.577530

Wu, N., Jiang, M., Liu, H., Chu, Y., Wang, D., Cao, J., et al. (2021). LINC00941 Promotes CRC Metastasis through Preventing SMAD4 Protein Degradation and Activating the TGF-B/smad2/3 Signaling Pathway. *Cell Death Differ*, 28 (1), 219–232. doi:10.1038/s41418-020-0596-y

Yao, J., Chen, X., Liu, X., Li, R., Zhou, X., and Qu, Y. (2021). Characterization of a Ferroptosis and Iron-Metabolism Related lncRNA Signature in Lung Adenocarcinoma. *Cancer Cell Int.* 21 (1), 340. doi:10.1186/s12935-021-02027-2

Zappa, C., and Mousa, S. A. (2016). Non-small Cell Lung Cancer: Current Treatment and Future Advances. *Transl. Lung Cancer Res.* 5 (3), 288–300. doi:10.21037/tlcr.2016.06.07

Zhao, Z., Liu, H., Zhou, X., Fang, D., Ou, X., Ye, J., et al. (2021). Necroptosis-Related lncRNAs: Predicting Prognosis and the Distinction between the Cold and Hot Tumors in Gastric Cancer. *J. Oncol.* 2021, 1–16. doi:10.1155/2021/6718443

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.

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